SYSTEMS MODELING OF THE ORAL METABIOME

by

Robert E. Brown

A Dissertation
Submitted to the Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics

Committee:

Dr. Patrick M. Gillevet, Dissertation Director

Dr. Dmitri Klimov, Committee Member

Dr. Don Seto, Committee Member

Dr. Jeffrey Solka, Committee Member

Dr. Iosif Vaisman, Graduate Program Director

Dr. James Willett, Director, School of Systems Biology

Dr. Richard Diecchio, Associate Dean for Academic and Student Affairs, College of Science

Dr. Vikas Chandhoke, Dean, College of Science

Date: April 7, 2011

Spring Semester 2011
George Mason University
Fairfax, VA

# SYSTEMS MODELING OF THE ORAL METABIOME

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy Bioinformatics at George Mason University

By

Robert E. Brown
Bachelor of General Studies
University of Michigan, 1975
Master of Science
University of Southern California, 1983

Director: Patrick M. Gillevet, Professor
Department of Environmental Science & Policy

Spring Semester 2011
George Mason University
Fairfax, VA

# Dedication

I dedicate this dissertation to my children; Alex, Douglas, Jeffrey, Kelly, and Patricia.

# Acknowledgments

I would like to thank my advisor, Dr. Patrick M. Gillevet, for his support and guidance during this journey. He believed in me and gave me encouragement when I really needed it. I thank Dr. Klimov, Dr. Seto, and Dr. Solka for serving on my committee. I thank Joan Drury for her recommendations, understanding, and support. My research work was sponsored in part by the Case Western Grant NIH/NIDCR AI-U01-68636. I would like to thank Dr. Mahmoud Ghannoum and Pranab K. Mukherjee, the primary investigators on this grant for their collaboration and support. I thank Huan Li for the motivation to continue. Most of all, I thank my family for their love, support, and encouragement.

# Table of Contents

# List of Tables

# List of Figures

**Figure** **Page**

# Abstract

SYSTEMS MODELING OF THE ORAL METABIOME

Robert E. Brown, PhD

George Mason University, 2011

Dissertation Director: Dr. Patrick M. Gillevet

Deciphering the underlying biological processes comprising the Human Oral Metabiome is important to the understanding of Human Immunodeficiency Virus (HIV) disease. The National Institute of Health has launched the Human Microbiome Project (HMP) to accelerate research and discovery techniques for five microbiome sites on the human body. Knowledge discovery techniques are needed to point researchers to follow-on hypotheses to quickly pinpoint areas of great promise. We developed the Differential Correlation Network (DCN) as a technique for researcher's to perform knowledge discovery in the oral mycobiome field. Using data from the Oral Microbiome, Differential Correlation Networks were applied to metabolites, bacteria, and fungi sampled from 12 Controls and 12 HIV Patients. By analyzing 100's of features across disease vs. control classes, statistically significant feature pair differences are captured and presented in Cytoscape. Several interesting differences are discovered and their possible biological significance is presented. The Systems model in conjunction with known biological metadata can identify promising difference networks and direct follow-on research based on DCN generated hypothesis.

# Chapter 1: Introduction

Research into the Human Metabiome introduces new complexity with the breadth of experimental design. These experiments focus on multiple features – metabolites, bacteria, fungi, and host cells interactions, in a variety of ecological niches on the human body. The ability to decipher the myriad of potential pathways and interrelationships is critical to the success of these studies. The goal is the interpretation of the underlying biological processes to specify the cause of disease. Frequently the volumes of data are difficult to interpret and require novel methods to assist the researcher. The National Institute of Health has launched the Human Microbiome Project (HMP) (Peterson, Garges et al. 2009) to accelerate research and discovery of novel techniques for five microbiome sites on the human body. Once data is consolidated from a study, knowledge discovery techniques are needed to point researchers to significant findings that will support the development of follow-on hypotheses. With these follow-on hypotheses, we can focus on key biological processes leading to development of beneficial treatments or diagnostics.

Issues relating to Metabiome studies involve microbial species that are frequently unculturable (Kurokawa, Itoh et al. 2007), sampling environments that are not controlled, and underlying biological functions that may be performed by differing “equivalent” biomic units across experimental samples (Gillevet 2011). Key to this process is the need for novel analytical methods to shift through the experimental data, focus on

significant differences between disease and healthy states, and generate high-value targeted follow-on hypotheses.

## *1.1 Problem Statement*

The rapid scope change of certain biological studies to include an entire environment involving hundreds of species and metabolites requires corresponding updates to the statistical tools used to scrutinize the data. Factors to be considered in metabiome studies include; not all samples having values for all features measured (sparse matrices), abundance data limitations, and especially in knowledge discovery experiments, the use very few samples. With all the algorithms and studies performed to date, the scientific community is still not able to clearly identify underlying biological changes that would explain the difference between a healthy metabiome from diseased. Specifically, data sets with a number of metabolites, bacterial taxa and fungal taxa abundance and small sample sizes have prevented successful biological interpretation of many study results.

The Problem Statement follows:

The ability to distinguish metabiomic communities is challenged by experimental results involving sparse parametric and non-parametric data, the number of features involved, and their associated measurement limitations. These issues impact the ability of researchers to identify significant underlying biological differences explaining metabolomic community conversion from a healthy to disease state.

The purpose of these large scale studies is to capture data from in-vivo environments that are not impacted by subset of environments established in the laboratory. The

metabiomic analysis approach circumvents the issue of unculturable organisms. The biological hypothesis for this study follows.

Examining the change in the relationship between pairs of features, one pair from a control class and the same features paired in the disease class may elucidate healthy versus disease underlying biological function changes allowing for knowledge discovery. This hypothesis will challenge the null hypothesis that states: there will be no significance difference in the correlations for a pair of features in healthy class samples compared to a disease class samples.

## *1.2 Advancing Knowledge Discovery of the Oral Metabolome*

The knowledge discovery analysis being presented in this paper is based on data from an Oral Rinse Metabiomic clinical study performed by Case Western Reserve (Mahmoud A. Ghannoum1* 2011). In this study, as others similar to it, the data is frequently non-parametric. Dhanoa and Fatt (Dhanoa and Fatt 2009) addressed the issue of normalcy in distribution of the data with the Mann-Whitney test, requiring limits being placed on the techniques that can be used to interpret the experimental results. An additional complication is, many of the features will not be present or are immeasurables (below instrumentation thresholds) for many samples or features, creating a sparse matrix of results. This presents additional challenges in interpreting the data and in using approaches similar to Principal Component Analysis (PCA) {Nyamundanda} to locate the major sources of variability in the results.

## *1.3 Glossary*

The following are terms used repeatedly in this paper. To ensure an agreed upon definition the specific meaning of the following terms follow:

**Class** – The subset of oral rinse study samples that were identified to be considered as a unique group, e.g. control class.

**Data** – The universal set of all feature data values for all samples.

**Feature** -- A single identified bacterial genus, fungal genus, or metabolite that was present in at least one oral rinse participant's sample.

**Immeasurable –** Feature data values that are either: below instrument detection level, or that is below researcher defined lower limits, or not present.

**Metabolites** – Biologically active molecules identified within one or more of the Oral Rinse samples.

**Metabiome** – The combination of the microbiome, mycobiome, and metabolites for a given class, e.g. control.

**Metabolome** – The analyzed metabolites for a given class e.g. control.

**Microbiome** -- The collection of identified bacterial taxa, where each taxa's abundance is minimally 1% of the community abundance in a specific sample as specified in the study.

**Mycobiome** – The collection of identified fungal, where each taxa's abundance is equal, or greater than, 1% of the fungal community abundance of a sample as specified by the study.

**Sample** – One participant's oral rinse metabiome feature data results.

## *1.4 Limitations, Assumptions, and Design Controls*

The following are the basis of parameters that have been investigated and limited the analysis of this study.

The number of measurable data values for each class in the experiment must be large enough to result in significant findings. This is true of any research project and is also true with this approach. A correlation difference hard minimum of 4 samples for one feature is the lower limit (Morgenthal, Weckwerth et al. 2006).

An issue to be addressed in large feature-set clinical trails is how the researcher chooses to address immeasurable data values that could indicate, there was no feature present in that sample versus a feature's value was below the equipments threshold. This distinction could be relevant if one considers the possibility of a feature's abundance fluxuating over short time intervals. Chan et. al. (Chan, Rowe et al.) investigated the metabolome differences in the type and amount of metabolites present and concluded that result from either a common metabolic network was being differentially regulated under the two conditions, or possibly there differences in the structure of metabolite relationships between the two experiments.

Database evolution could impact a study by the evolving naming conventions or definitions in the referenced repositories. This is especially problematic when a follow up study may attempt to validate the results of the current study. In a rapidly changing field these issues will be with used for the foreseeable future.

Extensive testing was required to ensure there are no algorithm flaws in the difference correlation methodology. The coupling of internal processing ids to various feature types used to ensure the results are correct for the sample sizes and feature values presented.

The pipeline used in this paper is semi-automated and requires careful monitoring to protect against human error. Moving the process to a completely automated pipeline via a workflow tool such as Galaxy or QIIME (Caporaso, Kuczynski et al.) in the future will reduce potential variability and improve tracking.

The data supplied by the clinical and 'wet lab' Oral Rinse HIV study has many non-unique areas for potential problems. Laboratory Information Management System (LIMS) assist, but solid manual procedures are required, for sample collection, labeled, and then to identification of metabolites, bacteria, and fungi. Standard clinical protocols for sample collection were used in this study. Quality checks involving periodic calibration of the equipment was imperative to accurate reading taken over days. Ensuring result parameters were consistent with other related studies was another approach to validation.

The Oral Metabiome results presented in this paper are from cross sectional studies and not part of a longitudinal study. The results include significant variability across samples. By performing a follow-on study the research may be able to get a better picture of the significance of the feature relationships per sample.

## *1.5 Summary*

Chapter 1 introduced the topic of analysis and corresponding issues in large scale Metabiome studies. These studies present many factors impacting the statistical significance of findings including; data that is often non-parametric, sparse datasets from immeasurable values, and taxa abundance threshold decisions. To date, it is difficult to assign biological interpretation to Metabiomic experiment results. A review of techniques in use today for metabolic result analysis will be addressed in Chapter 2. Chapter 3 will

introduce the Differential Correlation Network (DCN) approach to interpretation of Metabiome data. In Chapter 4 the DCN technique was applied to the Oral Rinse metabolite study data (Mahmoud A. Ghannoum1* 2011), then the full set of bacterial, fungal, and metabolite data features are analyzed using DCN in Chapter 5. The issues and researcher decisions are discussed in Chapter 6. The final Chapter captures the conclusions regarding the DCN approach, if the study hypothesis appears valid, suggestions on future endeavors, and ties the DCN solution back to the problem statement.

# Chapter 2: Literature Review

The explosion of the field known as Metabiomics has gone hand-in-hand with a corresponding evolution of the tools mining the experimental results. The recognition of the many advances in the field of genetics since the Human Genome Project (Watson 1990) has laid the foundation for the new large scale environmental studies.

Prior to the genetic approach to taxon identification, many taxa were never successfully isolated as viable specimens for analysis, presumably because their growth is dependant upon a specific microenvironment that has not been, or cannot be, reproduced experimentally (unculturable) (Huson, Richter et al. 2009).

Among those species that have been isolated, analyses of genetic makeup, gene expression patterns, and metabolic physiologies have rarely extended to inter-species interactions or microbe-host interactions.

With the realization that microbial cells may out-number human cells by a factor of ten ((HMP) 2011) we must ask the question; what is the relationship between ourselves and the microbial world around and within us? These microbial communities are largely unstudied and have poorly understood influence upon our development, physiology, immunity, and nutrition.

For completeness, subsection 2.1 give an overview of sample feature identification based on the advances in phylogeny identification for bacterial species moving straight

from experimental samples directly to quantitative measurement. Analogously, there is a quick review of advances for fungi phylogeny categorization and measurement. Lastly, we present a quick overview of the methods used to identify and quantify metabolites. Subsection 2.2 discusses several algorithms have been created by the statistical and bioinformatics communities to address the issue of determining statistically significant differences between classes based on small sample sizes. Small sample sizes present additionally limitations in which to analyze the results. First, a review of algorithms that, one, apply to single feature at a time within a given class, or two, determine statistical significance of samples across two classes. Subsection 2.3 is a synopsis of Oral Microbiome research to date and a chapter summation is presented in Subsection 2.4.

## *2.1 Bacterial, Fungal, and Metabolite Identification*

As so eloquently stated by Schloss (Schloss and Handelsman 2006) the phylogenetic and ecological complexity of microbial communities necessitates the development of new methods to determine whether two or more communities have the same structure even though it is not possible to sample the communities exhaustively.

Advances in DNA sequencing technologies have created a new field of research; National Research Council report in 2007 coined the term “metagenomics” defined as the in-situ extraction of DNA sequence information from entire microbial communities, including unculturable organisms.

Instead of examining the genome of an individual bacterial strain that has been grown in a laboratory, the metagenomic approach allows analysis of genetic material snapshot derived from complete microbial communities harvested from natural environments.

With the advent of technologies including Polymerase Chain Reaction (PCR), and pyrosequencing, scientists have accelerated the analysis of DNA with reduced cost and increasing speed (Sahota and Stormo; Diggle and Clarke 2004). The commercial launch of 454 Pyrosequencer in 2005 (Ronaghi 2001) (Fakhrai-Rad, Pourmand et al. 2002) (Franca, Carrilho et al. 2002) (Ronaghi 2001) was a milestone in genome sequencing in terms of performance and cost. Average read lengths have increased to 500 base pairs and are thus approaching read lengths obtained from traditional Sanger sequencing 88 (Franca, Carrilho et al. 2002).

The microbiome techniques (Nossa, Oberdorf et al.) are advancing bacterial identification and abundance determination, while side-stepping the issue of unculturable bacterial species. Without a priori information about the taxa present and avoiding in vitro culturing of samples, recent application of molecular biological approaches has led to the detection of many anaerobic species (Gillevet 2011) (Wu, Lewis et al.). The techniques go from clinical samples directly to bacterial rRNA gene amplification. The ability to cost-effectively and quickly sequence environmental in situ bacterial sample 16s ribosome sequences have been an essential protocol for microbiome studies. Using a mixed pool of phylogenetically informative ribosomal 16S subunit DNA sequences to frequently achieve specific species level identification via the length of the amplicon between the two conserved flanking sequences (Wu, Lewis et al.) using terminal restriction fragment length polymorphism (T-RFLP) profiling (Liu, Marsh et al. 1997) or Length Heterogeneity Polymerase Chain Reaction (LH-PCR) (Gillevet 2011) to categorize bacterial samples. The ability to run multiple clinical samples together greatly enhances speed while reducing cost including an approach based on Multi-tagged Pyrosequencing (MTPS) methodology (Gillevet 2011) (Ghannoum, Jurevic et al.) (Naqvi,

Rangwala et al.). Each of theses protocols from clinical capture to wet laboratory processing and beyond has it limitations and assumptions that can adversely impact the reported findings (Wu, Lewis et al.).

With specific reference rRNA sequences available that validate species identification we can compare unknown sequences via search tools such as BLAST to bacteria sequence repositories such as Ribosomal Database (RDB). (Maidak, Cole et al. 2001)

The identification and quantification of Fungal clinical samples follows a similar track to identification of bacterial but uses the Fungal Ribosomal Inter Spatial Region length between the two conserved flanking sequences (Gillevet 2011) to identify fungi via fungal sequence repositories.

In the metabiome extracellular metabolites are a key component to any cross-species interaction. The oral rinse samples were profiled and quantified for excellular metabolites using both Liquid Chromatograph (LC) (Martin, Dumas et al. 2007) or Gas Chromatograph (GC) (Major, Williams et al. 2006) then fed into a Mass Spectrophotometer (MS) (Martin, Dumas et al. 2007) (Oberg and Vitek 2009). These metabolite screenings can distinguish 100s of metabolites in a single sample and offer a multi-dimensional view of many molecular classes simultaneously (Roessner, Wagner et al. 2000).

## *2.2* Non-Parametric and Sparse Data Analysis Techniques

The lessons learned from the Human Genome Project, and the advent of a myriad of large scale cost effective biological and metabolite evaluation techniques, present new opportunities and analysis issues. The problem of being able to identify the underlying

biological causes of disease based on metabiome studies still eludes the research community. Often in knowledge discovery experiments there are only a few samples per class (control vs. disease) that presents challenges for statistical significance determination. The statistical mean and variance of the population is unknown. The ability to compare disease versus control microbiome classes, with small sample sizes, will require non-parametric statistical approaches, not based on the form or the parameters of their distribution, such as variance assumptions.

### 2.2.1 Single Feature Statistical Analysis

There are a multitude of feature statistical analysis algorithms. Wilcoxon signed-rank test (Wilcoxon 1945) is a single feature two class analysis test. The Wilcoxon signed rank test statistically measures if a single feature is significantly different between two classes and is a non-parametric test that doesn't assume a normal distribution of the sample data. Based on another single feature approach, White et. al. used MetaStats (White, Nagarajan et al. 2009 ) to assess the probability that a feature's abundance in one class of samples is statistically significantly different compared to its counterpart in another class. Evaluating obese versus lean subjects, they extracted feature differences not detected in analyses by other papers. MetaStats uses either the false discovery rate, unless the sample features are sparsely populated, then it employs the Fisher's exact test.

### 2.2.2 Multivariant Non-Statistical Analysis

Several non-statistical algorithms exist that will use the entire dataset to determine an

ordered set of variance determinations.

Principal Coordinate Analysis (PCoA) is an unsupervised technique that helps to extract and visualize a few highly informative directions of variation from complex, multidimensional data. PCoA is a vector transformation that maps the distance matrix to a new set of orthogonal axes such that a maximum amount of variation is explained by the first principal coordinate, the second largest amount of variation is explained by the second principal coordinate, etc. An improved version of the PCoA allowing integration and visualization of large dataset for microbial analysis is Fast Unifrac (Hamady, Lozupone et al.).

PCA was invented by Karl Pearson in 1901 and only allows rotations and projections. With PCA, and PCoA, the 2D view is a projection of the 3D view on a 2D space that keeps as much variance as possible. PCA is identical to PCoA if the distance metric is Euclidean (Camilolab.slu 2011) The principal coordinates can be plotted in two or three dimensions to provide an intuitive visualization of the data structure and look at differences between the samples, and look for similarities by sample category as well as hypothesis tests designed for immeasurable data.

The key issue is both PCA and PCoA are not based on a statistical model. In studies with hundreds of features, other approaches are applicable to discern if there is a significant statistical probability of overall class differences.

Modified PCA techniques have attempted to address the issue of sparse datasets, a frequent occurrence in Metabiomic studies that can impact the results of PCA and PCoA. One technique called Sparse PCA (Hui Zou! 2004) allows the analysis to only include a subset of the original values by employing an ‘elastic net’ around the regression coefficients. Another technique for improved results from multivariate metabiomic data is

Probabilistic PCA (PPCA) (Nyamundanda, Brennan et al.) that supports jointly modeling both metabiomic data and additional covariate information.

### 2.2.3 Statistical Multivariate Analysis

Statistical approaches to multivariate analysis are required if one is attempting to determine significance. Libshuff developed by Schloss et. al. (Schloss, Larget et al. 2004) determines statistical significance between sets of clonal libraries of environmental rRNA gene sequences based on the Cramer-von Mises-type statistic that tests for bivariate independence. A study by Singleton et. al. used Libshuff to distinguished rRNA gene sequence libraries from soil and bioreactors (Singleton, Furlong et al. 2001) and correctly failed to find differences between libraries of the same composition. However, issues with the false discovery rate (FDR), defined as the expected proportion of incorrectly rejected null hypotheses, in Libshuff (Schloss, Larget et al. 2004) pushed the creation of TreeClimber (Schloss and Handelsman 2006) by adapting population genetics methods, based on the parsimony test, to determine the relatedness of communities. Their paper indicates TreeClimber was more accurate than Libshuff. Another microbiomic statistical package for comparing microbial communities is Distance-based OTU and Richness (DOTUR) (Schloss and Handelsman 2005). DOTUR assigns bacterial 16S rRNA ISR sequences to OTUs based on the genetic distances between the sequences, combined with bacteria abundance data; it determines if the two communities are significantly different. It can also estimate the minimum number of sequence reads necessary make a determination. DOTUR was used by Baati et. al. to distinguish Mediterranean sea salt crystal bacterial community difference significances (Baati, Guermazi et al.). Similarly, Rattray et. al. used DOTUR for comparative microbial

analysis of earthworm digestive tracts (Rattray, Perumbakkam et al.).

Approaches that use Support vector machine's (SVM) supervised learning framework can build binary classifiers (Gillevet, Sikaroodi et al.) which determine disease versus healthy based on a set of features, after training based on supplied sample features.

Another technique, comparative correlation analysis, as been deployed in an effort to zoom in on specific biological processes underlying non-targeted Metabiomic community experimental profiles. The comparative correlation analysis technique (Morgenthal, Weckwerth et al. 2006) was used to characterize the physiological states of diverse plant species to elucidate participation of metabolites in different reaction networks.

Metabiomic studies entail sampling that is always measuring the sample metric not the population variation.

If the sample sizes are small, then other non-parametric statistical approaches are required to interpret the data. Working with very small sample sizes introduces more risk, often interpreted as Confidence Intervals (CI) (Rosner 2006) of the sample mean, which factors into the tabulation of statistically significant results. The key approach to addressing small sample size data is using results ranks, instead of actual data values, in the statistical computations. This is how the Spearman Rank correlation differs from the more general Pearson correlation coefficient.

As important to improved classification algorithms, is the requirement to tie study results, with large number of findings, to underlying biological process repositories, as demonstrated in a Multiple Factor Analysis (MFA) approach (de Tayrac, Le et al. 2009). MFA was used to link genomics microarray results via Gene Ontology to data-mine for supporting biological information to create gene modules.

All of these algorithms, and associated studies, indicate that a best solution to analyzing

metabiomic data has not been achieved, and we need to continue exploring novel approaches to the problem statement reiterated below:

The ability to distinguish Metabiomic communities is challenged by experimental results involving sparse parametric and non-parametric data, the number of features involved, and their associated measurement limitations. These issues impact the ability of researchers to identify significant underlying biological differences explaining metabiomic community conversion from a healthy to disease state.

## *2.3 Metabiomic Statistical Analysis Pipelines*

In any study, being able to reproduce the results is an absolute necessity. The requirement for traceability and logging of experimental steps has long been a component of wet lab work and attributed to the growth in LIMS capabilities. Metabiomic study research reaches beyond the wet lab and into a computation realm once the resulting raw abundance values and metabolite quantities are made available for algorithm analysis. This has fueled the trend for the development of in-house and web enabled algorithm processing workflow management capabilities.

One such systems, QIIME (Caporaso, Kuczynski et al.), supports end-to-end analysis based on many existing bioinformatics algorithms. QIIME performs many Operational Taxonomic Unit (OTU) and representative set choice algorithms, taxonomy assignment, sequence alignment, and phylogeny construction. Another key component in QIIME is result visualization.

Another open source web enabled workflow environment, with multiple embedded analysis tools, is Galaxy (Giardine, Riemer et al. 2005). It too supports a multitude of in-house analysis tools, while allowing integration of tools accessible via the internet to

support sequence analysis, plus statistical algorithms to determine significant metabiomic community differences.

The ideal conclusion to a researcher's effort is to apply the most effective way to communicate the study results. Many tools have been developed to address large microbiome and metabolite datasets (Wishart 2007). Critical to analysis is the ability to succinctly visualize the results (Kohl, Wiese et al.). The approach chosen for this paper, and that succinctly addresses large metabiomic dataset analysis, is to display the results via a network model (Loscalzo, Kohane et al. 2007). Network models are a user-friendly approach to presenting interconnected features with defined relationships. The network model is capable of conveying many results attributes in a single map. The specific tool chosen to present the data in this paper is Cytoscape (Shannon, Markiel et al. 2003). Cytoscape's ability to map nodes and edges (node connections) using colors or shapes mapped to various metadata e.g. structure, probabilities, pathways, organism type, or metabolite type, while also allowing user access to underlying additional attributes, lends significant power to conveying many types of information in one map.

## *2.4 Overview of Human Oral Microbiome Studies*

The Human Microbiome Project (HMP) was initiated by the NIH Roadmap in 2006. Its mission is to compile a comprehensive characterization of the human microbiome, and determine its role in human health and disease. The HMP is focused on five symbiotic environments of the human body - skin, mouth, nasal cavity, gut, and vagina (Turnbaugh, Ley et al. 2007; Peterson, Garges et al. 2009) (Hsiao and Fraser-Liggett 2009). Similar metabiomic approaches have been applied to delineating prostate cancer

from healthy individuals (Sreekumar, Poisson et al. 2009). The oral microbiome has been studied by many researchers in an attempt to characterize the non-disease state (Bik, Long et al.; Nasidze, Li et al. 2009; Zaura, Keijser et al. 2009; Ghannoum, Jurevic et al.). There is likely only one healthy individual core oral mycobiome (Ghannoum, Jurevic et al.) (Zaura, Keijser et al. 2009), and a longitudinal study by Lazarevic et. al. has inferred that the oral microbiome may be stable over a period of days (Lazarevic, Whiteson et al.). Interestingly, the oral microbiome has been implicated in downstream intestinal inflammatory bowel disease (Singhal, Dian et al.). Studies have attempted to understand the interactions within the oral microbiome and its impact on Candida Albicans (Thein, Samaranayake et al. 2006). Candida Albicans is a significant health issue for HIV Highly Active Anti-Retroviral Treated (HAART) patients, causing a infection named thrush, and is the basis of the Oral Rinse study in which this paper is based (M. Ghannoum 2011). The entire human metabiome plays a significant role in healthy individuals (Ghannoum, Jurevic et al.) (Dewhirst, Chen et al.).

## *2.5 Summary*

There are a myriad of methods to investigate statistically significant differences in metabiomes. Each has its strengths and weaknesses; however, to date techniques have not proven very successful in assigning biological interpretation to metabiomic experimental results. We reviewed different classes of algorithms and how they have been applied to address large and sometimes incomplete datasets using statistical and non-statistical approaches. The next section gave a quick overview of bioinformatic workflow tools, and then techniques for visualizing the results for maximum information conveyance, leading to one visualization approach, network diagrams. Finally, we

reviewed the current research involving the Human Oral Metabolome as the launching point for this study.

Chapter 3 introduces a novel approach to interpreting metabiomic data by focusing on the significant changes in correlations between measured biological components in one community versus a second community. We will discuss the algorithm, the process flow, and then present sample results based on the technique. This is followed by the determination of Confidence Intervals corresponding to the results. Chapter 4 applies the approach presented in Chapter 3 to two human oral rinse microbiome communities - Control and HIV.

# Chapter 3: Research Design and Methodology

“Many experiments focus on the identification of a disease or other significant biological topic. Another pathway is knowledge discovery where the goal is defining the basis for new hypothesis versus validating a stated hypothesis (Mahmoud A. Ghannoum1* 2011). “Principled hypothesis generation is clearly at least as important as hypothesis testing, and appropriate experimental designs ensure that the search for good candidate data is not an aimless fishing expedition but one which is likely to find novel answers in unexpected places” (Kell 2005).

The goal is tying together multiple technologies, data mining the repositories, and transforming or filtering reams of data to gather significant information that could lead to new hypothesis. By processing hundreds of internal or external cellular features and performing statistical analysis can lead to findings that are a great source of new hypotheses (Costanzo, Baryshnikova et al.)

Knowledge discovery is critical to ongoing research, with the dramatic increase in capabilities coupled with the rapid decrease in associated costs have combined to deluge researchers with raw data. Bioinformatics continues to define novel ways of shifting through the reams of data to locate nuggets of knowledge. The oral microbiome is one system that presents an opportunity to apply new techniques to tackle research questions on a large scale and implement knowledge discovery for hypothesis

generation.

Several studies have addressed the composition of the oral metabolome (Bik, Long et al.; Nasidze, Li et al. 2009; Zaura, Keijser et al. 2009; Ghannoum, Jurevic et al.). Other studies have focused on particular healthy versus disease oral metabolome states. As stated in Chapter 2 several techniques address single feature changes and whole class perturbation analysis.

We have recently applied a novel differential correlation network approach for knowledge discovery regarding differences in HIV versus Control oral metabolomes, and define this approach as Differential Correlation Network (DCN), By comparing correlations of feature-pairs between two different classes and identifying statistically significantly different pairings will aide the researcher in identifying specific hypothesis of the underlying biological process change. Combined with supplemental information from other bioinformatic repositories, one can pursue the most intriguing features to be examined in further experimentation. Assessing large volumes of data requires advanced multi-faceted visualization techniques. The final result of the Differential Correlation Network (DCN) is a network model with feature (nodes) annotated with metabolome superpathway, node shapes reflecting feature type, edges (connector) style reflects the encoding of the range of probability statistics, as well as key attributes describing the edge labels. Multiple DCN maps can be displayed together for a side-by-side view that simplifies the researcher initial overview of the relationships. Underlying attributes related to the features, e.g. for metabolites KEGG Pathway ID, CAS ID, Ion Mass, etc. are available via the network diagram attributed database. A reiteration of the hypotheses will be a good reference for the DCN method introduction in this chapter.

(H1) Examining inter-class differential feature pair correlations will differentiate healthy

versus disease classes.

(H2) The application of Differential Correlation Network analysis on experimental data support knowledge discovery related to underlying biological process variation between healthy and disease states.

Thus the goal of the DCN algorithm being presented is knowledge discovery leading to follow-on ‘‘hypothesis-driven’’ studies based on the Oral metabiome. By evaluating the oral metabiome in terms of individual variability, longitudinal trends, and the effect of diet and geography, plus previous studies focused on determining the association of oral mycobiome with health and disease (Ghannoum, Jurevic et al.), we can formulate new hypothesis’s based on analyzed clinical data results. One can ask whether it is possible to detect correlation shifts between classes. Do these shifts reflect underlying biological differences? What types of relationships can be experimentally detected between metabolites, microbiome, and the mycobiome? Thus, this research focuses on the differences between oral metabiome of healthy and HIV samples.

With the most interesting and statistically significant results available in a highly attributed network model, and coupled with known biological metadata, one quickly recognizes the promising biological function perturbations. Armed with these insights, knowledge discovery is enhanced driving hypothesis based experimental research. DCN, coupled with knowledge from existing Microbiome, Genetic, Pathway and other repositories, shall demonstrate if there is significance, and potentially causality, based on the significantly different feature-set relationships that appear between classes. This technique explored relationships of human oral features across -- bacteria, fungi, and metabolites.

## *3.1 Data Requirements*

The Differential Correlation Network (DCN) algorithm is dependent on solid sample processing protocols. Analyzing the feature data requires sufficient sample sizes per class to enable the determination of statistical significant. Outside the scope of this research are issues relating Chain of Custody demarcation, interpreting data within the accuracy limits of the instrumentation, clinical study design and protocols, and wet-lab processing to identify features. However, it is imperative to have accurate, unbiased, and uncontaminated sample feature values to make any valid interpretation of the results. The technique is based on numerical data values, with the samples presented in rows, and the features in columns with the sample-feature value at the cell (column, row) position as represented in Table 3.1-1. Researchers from varied backgrounds use the same word with different meanings. For the sake of clarity this is the interpretation of the following words in the document.

**Class** – The subset of Oral Rinse Study samples that were identified to be considered as a unique group, e.g. Control Class.

**Data** – The universal set of all feature data values for all samples.

**Feature** -- A single identified bacterial genus, fungal genus, or metabolite that was present in at least one Oral Rinse participant's sample.

**Immeasurable –** Feature data values that are below instrument, or researcher defined lower limits, or not present at all.

**Sample** is the clinical raw data capture from one participant in the study. The sample will contain the features being studied.

The samples for each class must be contiguous rows in the input file. Data values that

are listed as zero or blank are considered immeasurable. A single user defined column will contain the display identifications for all samples in all classes. One user defined row will contain the displayed feature identification of the feature corresponding to the data values in that column.

Table 3.1-1 A Three Class, 24 Sample, and 7 Feature Input Example

| Sample_ID | Age | Gender | 1,3-diamino-propane | 1,5-anhydroglucitol (1,5-AG) | 1,6-anhydro-glucose | 1-kestose | 2-amino-butyrate |
|---|---|---|---|---|---|---|---|
| Sample-1a | 34 | m | 200148 | 522941 | | | 594356 |
| Sample-2a | 46 | m | | 651178 | | | 334311 |
| Sample-3a | 59 | m | 403827 | 541635 | | 36812 | 825630 |
| Sample-4a | 22 | m | 113612 | 395420 | | | 384047 |
| Sample-5a | 37 | m | | 393063 | | | |
| Sample-6a | 34 | f | 116760 | 878860 | | | 292762 |
| Sample-7a | 40 | m | | 508757 | | | 447653 |
| Sample-8a | 27 | m | 82073 | 863937 | 167666 | | 446889 |
| Sample-9a | 53 | m | 569651 | 462349 | | | 1289929 |
| Sample-10a | 44 | m | | | 91606 | 722177 | 460779 |
| Sample-11a | 22 | m | | 515482 | 190203 | | 544216 |
| Sample-12a | 47 | m | 44165 | 115169 | | | 280167 |
| *Sample-1b* | *56* | *m* | | *217961* | | | |
| *Sample-2b* | *52* | *m* | *40413* | *575534* | *85467* | | *309029* |
| *Sample-3b* | *40* | *m* | | *704465* | *179152* | | *433659* |
| *Sample-4b* | *40* | *m* | | *134231* | | | *406964* |
| *Sample-5b* | *31* | *f* | *40221* | *150590* | | | |
| *Sample-6b* | *42* | *m* | *37123* | *861600* | | | *557703* |
| Sample-1c | 31 | m | 134173 | 589750 | | 32530 | 558450 |
| Sample-2c | 45 | m | | 478537 | 91239 | | 604543 |
| Sample-3c | 31 | m | 698982 | 433073 | | | 462482 |
| Sample-4c | 22 | m | | 1752329 | 191991 | | 941096 |

## *3.2 Data Analysis Parameters – Immeasurable and Minimum Pairing*

Often exploratory clinical experiments, with the experimental benefit unclear, will usually have a smaller number of samples than one with a focused clinical objective such as the efficacy of a drug. This leads to greater statistical uncertainty and larger confidence intervals for the results. Correlations performed on small sample sizes are best calculated with the ranking of the data values versus the using raw values. Ranking removes the parametric assumptions regarding the underlying population data variance. This issue is addressed in the next section. One DCN option allows the researcher to determine what role immeasurable values will play in the analysis. It is not intuitively obvious if and when immeasurable data values should be included in the calculations. Excluding immeasurable data allows one to focus the results only on features where there is sufficient representation in the samples. Is it valid to ignore features when they are not present in some samples? The voluminous number of immeasurable values will reduce variances and mask potentially valid correlations for when the features values are actually present. This decision is left to the researcher for when many viewpoints need to be explored.

Secondly, another important correlation analysis consideration is at what minimum percent of measurable features for a correlation be deemed significant? If one excludes immeasurable values it is possible that the number of feature pairs left to calculate a correlation may lead to false correlations. One cause for sparse feature matrices relates to the measurement of Operational Taxonomic Units (OTUs) from samples collected across a multitude of ages, sex, race, and geographic locations. One potential reason for many different OTUs present across subsets of samples may be the ability of unique combinations of bacteria or fungi to work together as “functional units” (Fischbach and Krogan) (Lau and Liu 2007) that are biologically functionally equivalent but the

underlying OTUs for the bundled function unit involve different unique bacteria or fungi. Another consideration for DCN is how far to “roll-up” the taxonomic labeling for the analysis (i.e. family designation versus genus). For example, a trade-off may involve using genus versus species level abundances when some OTUs cannot achieve sufficient abundances at the species level. However collapsing taxonomic levels to increase individual abundances causes a corresponding reduction in specificity. A DCN parameter option allows the researcher to require a minimum number of feature-pairs, selected as a percentage of the total class sample count. This value is the minimum measurable values required, for one of the two features of the pair being correlated. One or the other of the features must be present above that minimum percentage to calculate a correlation for that pair of features. For statistical calculation purposes, a minimum of four feature pairs is required to determine a correlation. By increasing the percentage of measurable values necessary, the researcher has another method to restrict their study to the more significant features. The impact to this study’s results regarding the parameter choices for the Oral Metabiome DCN analysis is discussed in Chapter 6.

## *3.3 Class-Feature Pairing Analysis Pipeline*

There are several statistical techniques that give a significance probability for a single feature for many samples across two classes such as Metastat (White, Nagarajan et al. 2009), Wilcoxon (Wilcoxon 1945). DCN gives a similar probability result to pairs of features across classes. This analysis requires the correlation of all possible valid values, depending on researcher chosen parameters as discussed in Section 3.2. It begins by analyzing one of the sample classes – control, disease state 1, disease state

2, etc. Each class will have a number of samples. For each sample, each feature will have a value ranging from a measurable value down to immeasurable. Measurable implies the feature abundance exceeds the minimum threshold that is considered valid for the analytical technique being used to interrogate that feature.

The first stage of the analysis is to rank order the data values for each feature across all Class 1's samples if the number of samples N < 26. If a sample's feature value was immeasurable, and it is to be included in the analysis, it is assigned the lowest rank. In a given feature, if multiple samples have immeasurable values, they each are assigned the averaged lowest rank value. Therefore if two samples for a feature in a class are both immeasurable the rank assigned is (1+2)/2 = rank value 1.5. The other DCN input parameter when applied (and always must equate to at least 4 valid features pairs) is to require one feature set {A} or feature set {B} to contain a minimum number of measurable values. The default is 30% measurable feature pairs, compared to the number of samples in the Class, e.g 12 samples x 30% = 3.6 rounded up = 4 or more pairs. This is to ensure a feature was present in sufficient enough samples to produce significant correlations.

Once all allowed feature values have been successfully ranked, the Correlation is applied to all same Class samples. Sample Features A $\{a_1, a_2, a_n\}$ are paired with every other Feature B $\{b_1, b_2, b_n\}$ creating feature-pair correlation values rho(A-B).

Spearman Correlation was chosen, versus Pearson correlation, to support small sample sizes typical in knowledge discovery experiments, where one cannot assume the data is parametric. The Spearman Correlation technique {Rosner, 2006} non-parametric and is not based on assumptions regarding the parametric nature of the underlying data distribution. The Spearman Correlation is effectively the Pearson correlation using rank

data values versus raw data values. As noted in the Spearman Correlation in equation Eq. 3.3-1, each individual feature must have a non-zero variance; otherwise no correlation can be calculated between it and any other feature.

Each Class maintains its own feature correlations results. The result is for each feature-pair A-B in Class C will have a single correlation value. With N features for the samples in a Class, the correlation process is calculated is attempted on all N features versus all the other N-1 features within the class to create a maximum of (N*(N-1))/2 feature pair correlations such that:

N is the number of Classes ( **C** )

M is the number of samples in $\mathbf{C}_n$

t is the total number of features (F) measured for every sample regardless of Class.

$F_{t,n,m}$ is the 't'th feature for Sample m in Class n

**Cn** = {p / q : p >=0 /q>=0: p, q $\in$ **R**} All real valued feature values in Class n

$\mathbf{F}_{t,n} \subset \mathbf{C}n$ is the ranked set of values in Class n for feature **F**t

$\boldsymbol{u}\mathbf{F}_{t,n}$ = the mean of the ranked values for $\mathbf{F}_{t,n}$

$$Rho_{fafb} = \frac{\sum_{1}^{n}(f_{a,i} - uf_a)(f_{b,i} - uf_b)}{\sqrt{\sum_i (f_{a,i} - uf_{a,i})^2 \sum_i (f_{b,i} - uf_{b,i})^2}} \qquad \text{Eq. 3.3-1}$$

Correlations values are bounded, -1.0 =< rho =<1.0. The correlation process is repeated for all features $F_{a,n}$ $F_{b,n}$ where a <> b. An example of a class feature-pair correlation table is in Table 3.3-1. The programs implementing the Differential Correlation Network are written in Python v2.6.4.

Table 3.3-1In-Class Feature Pair Correlation example

| Feature 1 | Feature 2 | Number sample pairs | Feature 1-2 correlation | Run Info |
|---|---|---|---|---|
| 10 | 138 | 8 | 1.00 | Run-ID |
| 8 | 16 | 8 | 0.99 | Run-ID |
| 16 | 17 | 8 | 0.98 | Run-ID |
| 8 | 17 | 8 | 0.6 | Run-ID |
| 8 | 165 | 8 | 0.55 | Run-ID |
| 8 | 188 | 8 | -0.01 | Run-ID |
| 15 | 110 | 8 | -0.35 | Run-ID |
| 95 | 136 | 8 | -0.99 | Run-ID |
| 43 | 123 | 8 | -0.22 | Run-ID |
| 17 | 165 | 8 | -0.55 | Run-ID |

## *3.4 Differential Correlation Network (DCN) Analysis*

The second phase of the analysis pipeline uses as input the feature-pairs $Rho_{fafb}$ correlation values from two classes. The goal is to determine if there as been a significant change in the correlation of the two features that may lead to knowledge discovery about a biological function variance. The significance of the correlation difference has two components: the actual difference of the two correlations, and the probability of that difference based on the sample size being significant. The input feature-pair set of values from each class was the outcome from the Section 3.3 analysis. To determine the probability significance, the actual number of sample pairs, $N_1$ and $N_2$, involved in calculating each class feature-pair correlation is used. N forms the basis of the standard error calculation to determine the probability assigned to the difference of Class 1's feature-pair A-B correlation and Class 2's feature-pair A-B

correlation value. The difference of these two class feature A-B correlation values is labeled the Differential Correlation (DC). To determine the probability of the Differential Correlation we need to perform an analogous Fisher Transformation modification for each DC calculation. Morgenthal et. al. created a transformation for two correlation values $C_1$ and $C_2$.in Eq. 3.4-1 (Morgenthal, Weckwerth et al. 2006).

$$z = \frac{1}{2} \frac{\log\left(\frac{1+C_1}{1-C_1}\right) - \log\left(\frac{1+C_2}{1-C_2}\right)}{\sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}}$$

Eq. 3.4-1

The z-value significance level between the two sets of correlations is based on the individual class's feature pair correlations rho(A-B), and the actual number of class sample values, $N_1$ and $N_2$, that were used to determine each feature pair's correlation. Using this approach the DCN calculates the statistical significance probability (z-value) of the difference correlation. As shown in Eq. 3.4-1, the minimum threshold for the probability determination is 4 samples, N, from each Class If either class's feature pair correlations didn't contain enough samples, it didn't comply, and no difference correlation is calculated.

Spearman Rank correlation, intended to address small sample sizes, allows correlations to more easily achieve the rho limits of -1.0 and 1.0 since it is using integer values (ranks) instead of actual real number values. In Eq. 3.4-1 achieving the limit values for rho, -1 or 1, will cause the z-value from the equation to go to infinity which only means a probability of 0% or 100%. To address this, the input correlations were artificially limited to a maximum value of -0.99 or 0.99 for use in Eq. 3.4-1 to prevent the calculation from exceeding the underflow, or overflow, limits of the Python language. This is caused by the artificial correlation increase from using ranks to address having only very few samples.

The z-value calculation follows the normal distribution with mean = 0 and Standard Error (S.E.) of 1. Therefore the probability percentage can be determined from the z-value in Excel function Probability = NormDist(z-value,0,1, “TRUE”).

The null hypothesis states there will be no significance difference ($p < 5\%$) in the pair of feature correlations measured in Class 1 as compared to the other Class 2.

Therefore, if we determine a statistically significant difference in correlations we could disprove the null hypothesis. Sample output results of a second stage of the DCN pipeline are in Table 3.4-1. The figure shows an example of the difference correlation result file generated for all features meeting the criteria from all samples in Classes 1 and 2. The resulting output map is displayed in figure 3.4-1 has been formatted to make the presentation clearer.

Table 3.4-1 A sample output from the Differential Correlation Network (DCN) analysis

| Feature 1 | Feature 2 | Class 1 vs. Class2 Feature 1-2 Difference correlation probability | Class 1 Feature A-B Pair Corre-lation value | Class 1 Number Feature A-B Paired Samples | Class 2 Feature A-B Pair Correlation value | Class 2 Number Feature A-B Paired Samples | Run Info |
|---|---|---|---|---|---|---|---|
| Feature A | Feature B | 99% | 0.51 | 12 | -0.99 | 15 | Run-ID |
| Feature A | Feature C | 99% | 0.9 | 12 | -0.88 | 15 | Run-ID |
| Feature A | Feature D | 98% | 0.57 | 12 | -0.97 | 15 | Run-ID |
| Feature A | Feature F | 98% | 0.38 | 12 | 1.00 | 15 | Run-ID |
| Feature B | Feature F | 98% | 0.89 | 12 | -0.81 | 15 | Run-ID |
| Feature C | Feature D | 97% | 0.36 | 12 | -0.97 | 15 | Run-ID |
| Feature C | Feature E | 97% | 0.75 | 12 | -0.9 | 15 | Run-ID |
| Feature D | Feature E | 97% | 0.89 | 12 | -0.76 | 15 | Run-ID |
| Feature D | Feature F | 97% | 0.81 | 12 | -0.86 | 15 | Run-ID |
| Feature E | Feature F | 97% | -0.08 | 12 | 0.98 | 15 | Run-ID |

Below is a simplified combined diagram, figure 3.4-1, of the input and results for a Differential Correlation Network. The key point is only significantly different Class 1 & 2 feature correlations will appear in a DCN. That does not indicate there are significant correlations within a Class, or that both sides of the correlation difference are significant, only the difference is. The number of samples has a large impact on the sample error, and therefore, the significance.

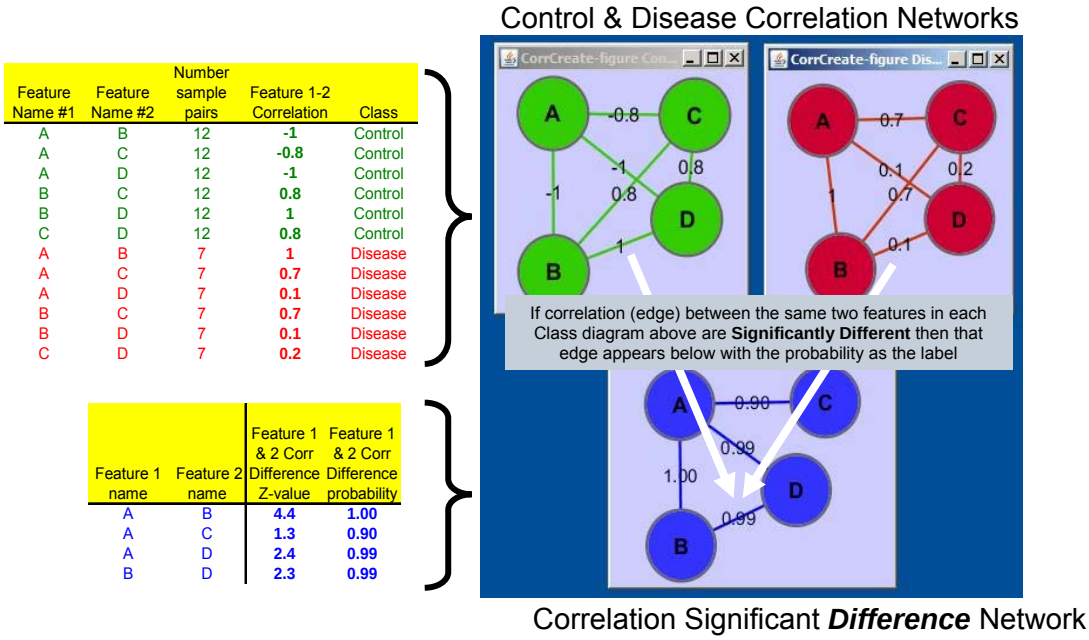



Figure 3.4-1 Overview of Correlations results from two classes map into a Differential Correlation Network

In Figure 3.4-1, the two classes, one control (green) and one disease (red), result in 4 significantly different feature pairs (blue) between them. The upper right network diagram side shows both the control correlation network and disease correlation networks corresponding to their upper left side table. The bottom blue network is the DCN, with edges labeled with probabilities, corresponding to the lower left table.

## *3.5 Feature Pair Differential Correlation 95% Confidence Interval Determination*

To identify the statistical significance of the resulting differential correlation the Confidence Interval (CI) statistical measure is used. The confidence interval gives a

minimum and maximum range of values to include the unknown population differential correlation probability mean. The CI range is calculated for each set clinical experimental sample data. If we reran the clinical experiment with different sample populations, each experiment will have its own sample mean. By using a CI of 95% we are indicating if 100 experiments were run, with each experiment using the same number of random samples from the population, one would expect 95 out of the 100 experimental results would contain the differential correlation population probability mean value (ρ) within their 95% CI probability range. To be able to determine differential correlation probability statistical significance we need to convert the Eq 3.4-1 differential correlation z-value to a probability, by performing an inverse Fisher transformation on (z) in eq. 3.5-1 {Rosner, 2006}.

$$p = \tanh(z) \qquad \text{0-2 Eq 3.5-1}$$

First we need to determine the 95% upper and lower CI around z. This will be 1.96 (95% z) times the Standard Error (SE). The z transform's approximate variance (Standard Error) Eq 3.5-2 is dependent on the samples sizes of the underlying differential correlation.

$$S.E. = \frac{1}{\sqrt{(N1-3)+(N2-3)}} \qquad \text{Eq. 3.5-2}$$

Using the S.E. from Eq. 3.5-2, we obtain the 95% CI = z +- 1.96*S.E., covert the differential correlation z value, and the two CI z values, to probabilities. The Sample Error (SE) calculation, the CI generation and converting all z values is implemented in the 'Calc_95_CI' Python routine.

These calculations are based on RANK values, not actual data values. The small

sample sizes could support using actual values but that isn't statistically valid per our upper sample size of n=12, hence the use of ranks versus the true data values.

## *3.6 Differential Correlation Network Visualization*

Frequently results from knowledge discovery experiments include several hundred features being measured making it difficult to interpret the results without a visualization tool. The network visualization tool use in this study for DCN mapping are based on Cytoscape (Kohl, Wiese et al.). Cytoscape displays nodes (metabolome features) and edges (difference correlations for feature-pairs) and allows for many visual clues to be embedded in the network diagram as defined in the legend. Therefore, many feature and feature-pair attributes can be displayed in one diagram greatly aiding in knowledge discovery, as shown in Figure 3.6-1. Two main files support the network creation in Cytoscape, the Node Attribute file, and the Network Attribute file that defines the connections (edges) between nodes.

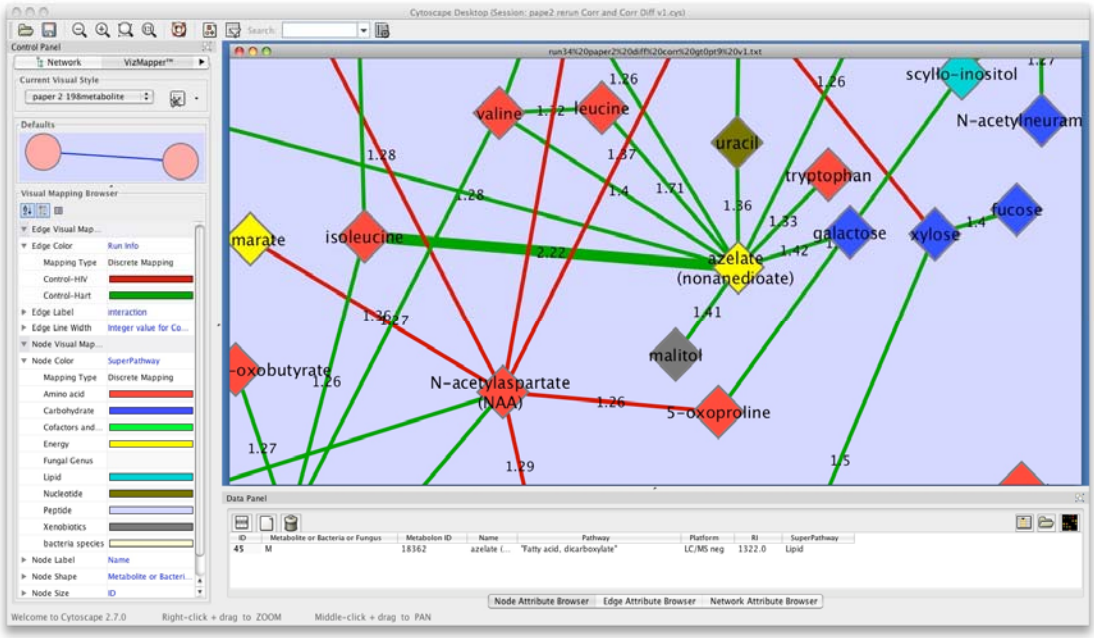



Figure 3.6-1 A sample Oral Rinse DCN displayed in Cytoscape.

Figure 3.6-1 is an example of a network map including several attributes associated with the underlying data. The attributes and usage of these maps is explained in detail later.

## *3.7 Conclusion*

The methodology of the Differential Correlation Network was delineated in this Chapter. Starting off with the feature data and attributes, we explored the issue of what is valid data and how the DCN will support knowledge discovery changes to input parameters. The first processing step of the feature correlation analysis was addressed, as well is the determination of the results and their Confidence Intervals. The second portion of the pipeline specified the actual DCN analysis, followed by the discussion of how to determine statistical significance. Lastly, a presentation of the visualization tool used to

display the DCN network diagrams and some the approaches to conveying multiple attribute details in one map.

The foundation has been laid for using the DCN for knowledge discovery. The next Chapter will apply the technique on the data from a Case Western led study of Oral Rinse Control and HIV samples. Chapter 4 will apply the DCN technique to the metabolites measured in the study and then bacteria and fungi, with their directly linked metabolites (metabiome), are covered in Chapter 5.

# Chapter 4: Analysis of the Oral Metabolites

This Chapter will examine data produced by the Case Western (CW) HIV Oral Metabiome Study (M. Ghannoum 2011). Standard clinical protocols were instituted to define, collect sample, and quantify 198 metabolite features, as discussed in Chapter Two. Using the methodology outlined in Chapter 3, the Differential Correlation Network analysis was applied against the four classes of Case Western Study Oral Rinse Metabolite samples. The analytic results were then integrated into Differential Correlation Networks maps.

The biological significance of this study is to identify easier and less invasive mechanisms to detect HIV disease. Visible signs of HIV infection include the occurrence of oral lesions, oral candidasis, and hairy leukoplakia, that are strongly associated with a low CD4 cell counts and increased plasma viral loads (Shiboski 2002). The oral metabolites are produced by both the host and microbes, and are likely to contribute to health and disease, and may be effective at indicating disease status.

The chapter organization follows: The hypothesis is restated, followed by section 4.1 that describes the oral rinse metabolite sample data and its attributes. Correlation analysis within each class is addressed in Section 4.2, including analysis parameter decisions and Confidence Intervals pertaining to these results. Section 4.3 constructs, from the results in the previous section, Differential Correlation Networks (DCN)

displaying the oral rinse metabolite pair statistically significant results. The last section summarizes the chapter's contributions.

The focus of the Differential Correlation Network pipeline is knowledge discovery, leading to hypotheses creation. In combination with other biological repositories one can postulate extra-cellular underlying biological pathways from the results. This chapter will define oral rinse metabolite DCN data that will be input into the findings for Chapter 7 in support of the hypotheses (reiterated below):

(H1) Examining inter-class differential feature pair correlations will differentiate healthy versus disease classes.

(H2) The application of Differential Correlation Network analysis on experimental data support knowledge discovery related to underlying biological process variation between healthy and disease states.

## *4.1 Oral Rinse Metabolite Samples and Classes*

The oral metabolome plays an important role in providing insight into the progression of HIV disease. Visible signs of HIV include the occurrence of lesions, oral candidasis, and hairy leukoplakia, that are strongly associated with a low CD4 cell counts and increased plasma viral loads (Shiboski 2002). Oral Metabolites are produced by both the host and microbes, and are likely to contribute to health and disease, and may be effective at indicating disease status. Profiling the oral metabolites should lead to an improved understanding of how the oral metabolome influences and is impacted by the underlying disease {M. Ghannoum, 2011}.

This study is scrutinizing the oral metabolome data to determine the biological

metamorphosis from healthy to HIV infected individuals. First, it analyzes variations in oral rinse metabolites, then later bacteria and fungi in Chapter 5. We are looking for candidate indicators to distinguish healthy from HIV disease states, that may lead to less expensive and less intrusive diagnostic techniques. A subset of the oral rinse metabolites and their attributes are listed in Table 4.1-1.

Table 4.1-1 A Subset of the Oral Rinse Sample Metabolites and Data Values

| SAMPLE_ID | 476893 | 476894 | 476895 | 476896 | 476897 | 476898 |
|---|---|---|---|---|---|---|
| SAMPLE_DESCRIPTION | Oral Rinse samples concentrated | Oral Rinse samples concentrated | Oral Rinse samples concentrated | Oral Rinse samples concentrated | Oral Rinse samples concentrated | Oral Rinse samples concentrated |
| CLIENT_ID | HIV-11 | HIV-12 | HEALTHY-1 | HEALTHY-2 | HEALTHY-3 | HEALTHY-4 |
| GROUP | HIV | HIV | CONTROL | CONTROL | CONTROL | CONTROL |
| | Hart | Hart | n/a | n/a | n/a | n/a |
| | | | | | | |
| **Name** | **CW-00143** | **CW-00144** | **CW-00145** | **CW-00146** | **CW-00147** | **CW-00148** |
| alanine | 5,232,252 | 9,166,877 | 9,735,187 | 4,100,405 | 13,976,460 | 3,561,050 |
| aspartate | 1,843,533 | 2,086,362 | 4,813,395 | 1,950,426 | 5,713,138 | 1,772,000 |
| beta-alanine | 189,667 | 64,390 | 88,823 | 61,348 | 177,354 | 47,168 |
| N-acetylaspartate (NAA) | 32,252 | 51,683 | 54,546 | 98,675 | 138,066 | |
| 2-aminobutyrate | 558,450 | 604,543 | 594,356 | 334,311 | 825,630 | 384,047 |
| creatine | 1,199,821 | 1,254,237 | 1,722,453 | 974,433 | 1,725,323 | 717,771 |
| creatinine | 573,411 | 1,062,530 | 818,327 | 662,681 | 804,583 | 670,325 |
| 2-hydroxybutyrate (AHB) | 305,524 | 953,780 | 605,533 | 302,355 | 269,221 | 288,347 |
| cysteine | 110,074 | 102,068 | 87,339 | | 193,893 | |
| taurine | 529,882 | 184,673 | 410,948 | 120,164 | 481,701 | 100,806 |
| gamma-aminobutyrate (GABA) | 49,073 | 13,464 | 70,840 | 43,737 | 95,923 | 8,708 |
| glutamate | 2,612,753 | 7,048,188 | 8,053,112 | 3,155,530 | 35,276,765 | 2,232,347 |
| glutamine | 2,280,154 | 2,159,608 | 1,709,897 | 845,368 | 11,231,434 | 494,426 |
| pyroglutamine* | 145,383 | | 176,316 | 120,977 | 266,057 | 119,406 |
| 5-oxoproline | 165,852 | 260,177 | 275,591 | 135,168 | 471,086 | 112,213 |
| glutathione, oxidized (GSSG) | 82,197 | 59,565 | 163,730 | 51,919 | 95,654 | 18,615 |

The first phase of human oral metabolome processing is to process data from the 24 oral rinse samples (12 Control and 12 HIV). Experimentally, 198 metabolites were measured as being present in one or more samples. The CW Oral Rinse sample data, Table 4.1-

2, was categorized into four Classes; the Control Class is the 12 control samples, Combined HIV Class is the combination of all 12 HIV samples, both Highly Active Anti Retroviral Therapy (HAART) and untreated HIV.

Table 4.1-2 The 24 Sample Metadata for the CW Oral Rinse Study

| CLIENT_ID | HART | Status | Subject # | Age | Gender | Race | CD4 | VL | SAMPLING DATE |
|---|---|---|---|---|---|---|---|---|---|
| HIV-2 | Hart | HIV | 2 | 56 | m | aa | 639 | 75 | 11/9/2005 |
| HIV-3 | Hart | HIV | 3 | 52 | m | aa | 800 | 48 | 11/9/2005 |
| HIV-4 | Hart | HIV | 4 | 40 | m | c | 947 | 48 | 11/9/2005 |
| HIV-5 | Hart | HIV | 5 | 40 | m | c | 280 | 48 | 11/9/2005 |
| HIV-7 | Hart | HIV | 7 | 31 | F | h | 1029 | 48 | 11/9/2005 |
| HIV-8 | Hart | HIV | 8 | 42 | m | aa | 814 | 53 | 11/9/2005 |
| HIV-11 | Hart | HIV | 11 | 31 | m | c | 670 | 68 | 11/10/2005 |
| HIV-12 | Hart | HIV | 12 | 45 | m | aa | 899 | 48 | 11/10/2005 |
| HIV-1 | Untreated | HIV | 1 | 31 | m | h | 380 | 158000 | 11/9/2005 |
| HIV-6 | Untreated | HIV | 6 | 22 | m | aa | 966 | 1100 | 11/9/2005 |
| HIV-9 | Untreated | HIV | 9 | 22 | m | c | 581 | 115000 | 11/9/2005 |
| HIV-10 | Untreated | HIV | 10 | 52 | m | aa | 5 | 185000 | 11/10/2005 |
| HEALTHY-1 | n/a | Normal | 1C | 34 | m | h | | | 11/18/2005 |
| HEALTHY-2 | n/a | Normal | 2C | 46 | m | c | | | 11/19/2005 |
| HEALTHY-3 | n/a | Normal | 3C | 59 | m | aa | | | 12/2/2005 |
| HEALTHY-4 | n/a | Normal | 4C | 22 | m | c | | | 11/19/2005 |
| HEALTHY-5 | n/a | Normal | 5C | 37 | m | aa | | | 11/18/2005 |
| HEALTHY-6 | n/a | Normal | 6C | 34 | F | h | | | 11/18/2005 |
| HEALTHY-7 | n/a | Normal | 7C | 40 | m | c | | | 11/19/2005 |
| HEALTHY-8 | n/a | Normal | 8C | 27 | m | c | | | 11/19/2005 |
| HEALTHY-9 | n/a | Normal | 9C | 53 | m | aa | | | 12/3/2005 |
| HEALTHY-10 | n/a | Normal | 10C | 44 | m | aa | | | 12/8/2005 |
| HEALTHY-11 | n/a | Normal | 11C | 22 | m | aa | | | 11/19/2005 |
| HEALTHY-12 | n/a | Normal | 12C | 47 | m | aa | | | 12/10/2005 |

The 12 Combined HIV samples in Table 4.1-2 are also divided into 2 separate non-overlapping and complete subclasses, HAART HIV (blue) and untreated HIV classes (red). The HAART Class contains the 8 oral rinse samples. Lastly, the untreated HIV Class is the remaining 4 HIV samples comprising the untreated HIV, non-Anti Retroviral Therapy (untreated) samples.

## *4.2 Oral Rinse Metabolite Class Feature Analysis*

The combination of all identified metabolites across all 24 samples yielded a total of 198 metabolites to be quantified for each sample per each Class per Table 4.1-1 and Table 4.1-2. This study specified the inclusion of immeasurable metabolite values, and required at least 30% measurable values in one feature of each feature pair to perform a correlation calculation. A discussion on parameter selection options was covered section 3.2. With this experiment only having 12 class samples maximum, the 30% cutoff (3.6 samples) was increased to 4 pairs where at least one feature of the pair is measureable. Four samples is the hard minimum for the statistical probability calculation, Eq. 3.5-1, so in this study that parameter did not have an impact. When this criteria is not met, that feature pair will be been excluded from the correlation calculation, leaving the other N-1 Class feature pairs to be investigated. Class correlations were the basis to determine if there is a significant difference in each metabolite pair correlation. With 198 metabolites being analyzed, there is a theoretical maximum of 19,503 = (198*(198-1))/2 pair-wise metabolite pair correlations. As stated earlier this total is reduced if the calculation, or minimum feature pair count requirements isn't satisfied. The Python language 'Spearman_Correlation' routine was executed to create each class correlation output file. The file contains all possible correlations for use as input into the DCN algorithm later. The actual correlation pair totals for each class was; 283 for the Control Class, 257 for the Combined HIV Class, 302 for the HAART Class, and 1,237 for the untreated HIV Class. The data was visualized as network diagrams. The network diagrams use a common legend listed below in Figure 4.2-1 and again in the beginning of Chapter 5.

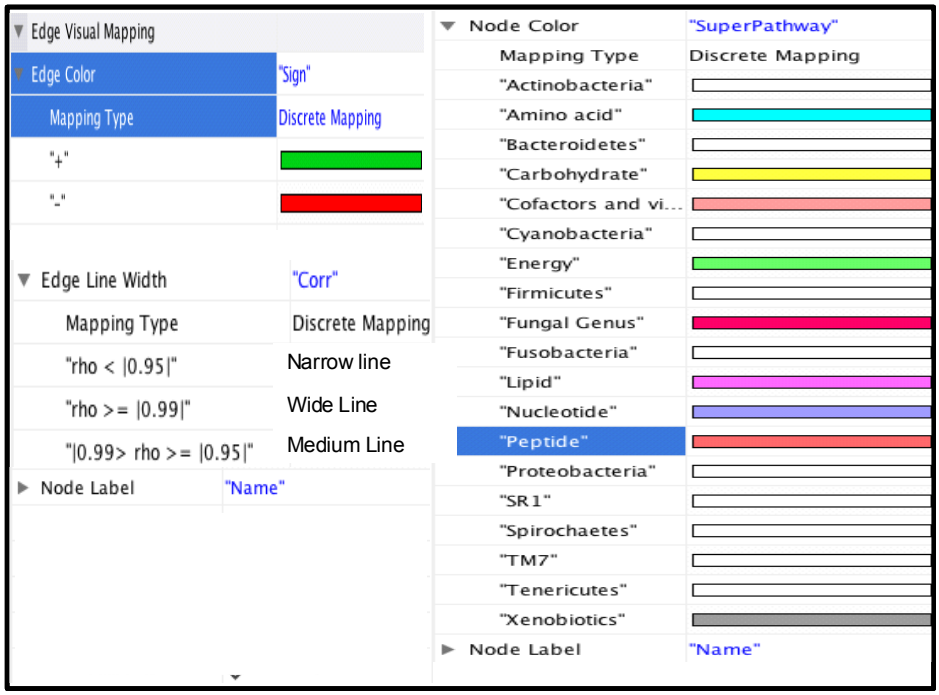



Figure 4.2-1 The Correlation Network Diagram Legend for interpreting attributes to the diagram.

The Network diagram in Chapter 4 contain only metabolites (bacteria and fungus are in Chapter 5), the edges are correlation connections between two metabolites -- if their correlation value, rho, was greater than the cutoff value. The cutoff applies to |rho| so both large positive and negative correlations are displayed. The cutoff minimum varies slightly with each class because of the variation in the number of correlation pairs. To keep the diagram comprehensible the cutoff display value is raised to reduce, or increase, the number for nodes and connections displayed to a visually optimal number. The legend presented in Figure 4.2-1 has 4 main legend components. The first top left component is the color coding for the edges (correlations). If the correlation, rho,

between two metabolites is positive it will display as a green line. If the metabolite pair correlation is negative the edge color is red. Second, is the middle left component of legend is the node pair correlation edge thickness guide, the greater the correlation, the thicker the edge. If the correlation, rho >= 0.99 the edge is thickness, if rho >=0.95 it is the medium thickness, and rho < 0.95 represents the thinness edge. Lastly, on the right side of the legend figure 4.2-1, is the mapping of the KEGG biological superpathway for the metabolite (node). The nodes in the network map figures will display in the color indicated in the Figure 4.2-1 based on the node metabolite superpathway attribute value. For each of the following sections the figures are based on the sections results data being imported into Cytoscape, along with the Metabolite (Node) Metadata.

### 4.2.1 Oral Rinse Metabolite Control Feature Analysis

The Spearman Rank Correlation on all 198 metabolites for the Control Class samples generated a the correlation data results. The results were filtered to only list correlations where rho is greater or equal to |0.84|. There are 281 positive correlations out of a total of 283 correlations. The correlation network created from the results is in Figure 4.2.1-1. The Figure Legend is in Figure 4.2-1.

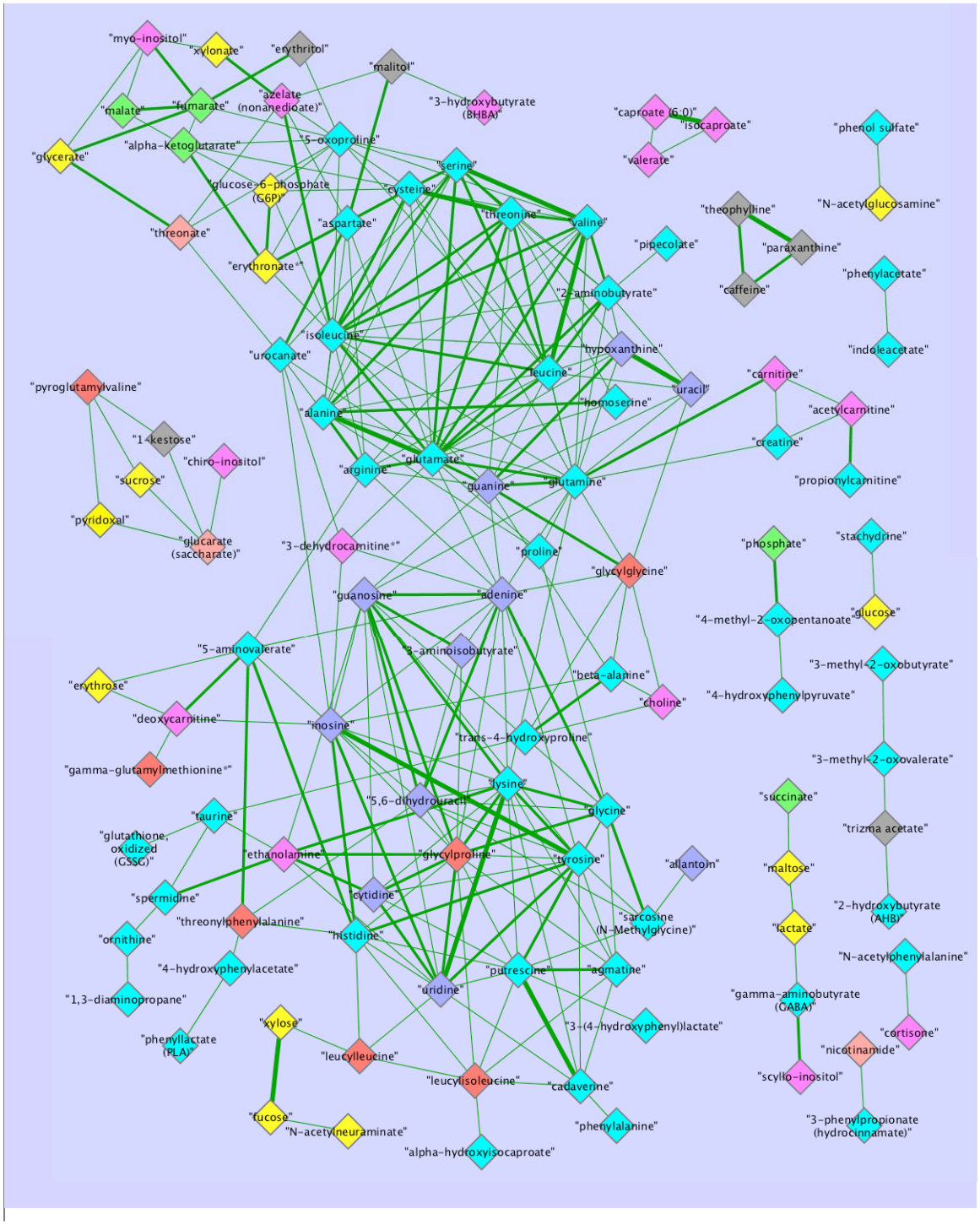



Figure 4.2.1-1 Control Class Metabolite Correlation Network with rho >= |0.84|. Legend is in Figure 4.2-1

### 4.2.2 Oral Rinse Metabolite Combined HIV Feature Analysis

The Spearman Rank Correlation on all 198 metabolites for the Combined HIV Class samples generated correlation results that were filtered to only list correlations where rho is greater or equal to |0.84|. There are 257 correlations, where 254 are positive correlations, and 3 are negative correlations. The correlation network created from the results is in Figure 4.2.2-1. The Figure Legend is in Figure 4.2-1.

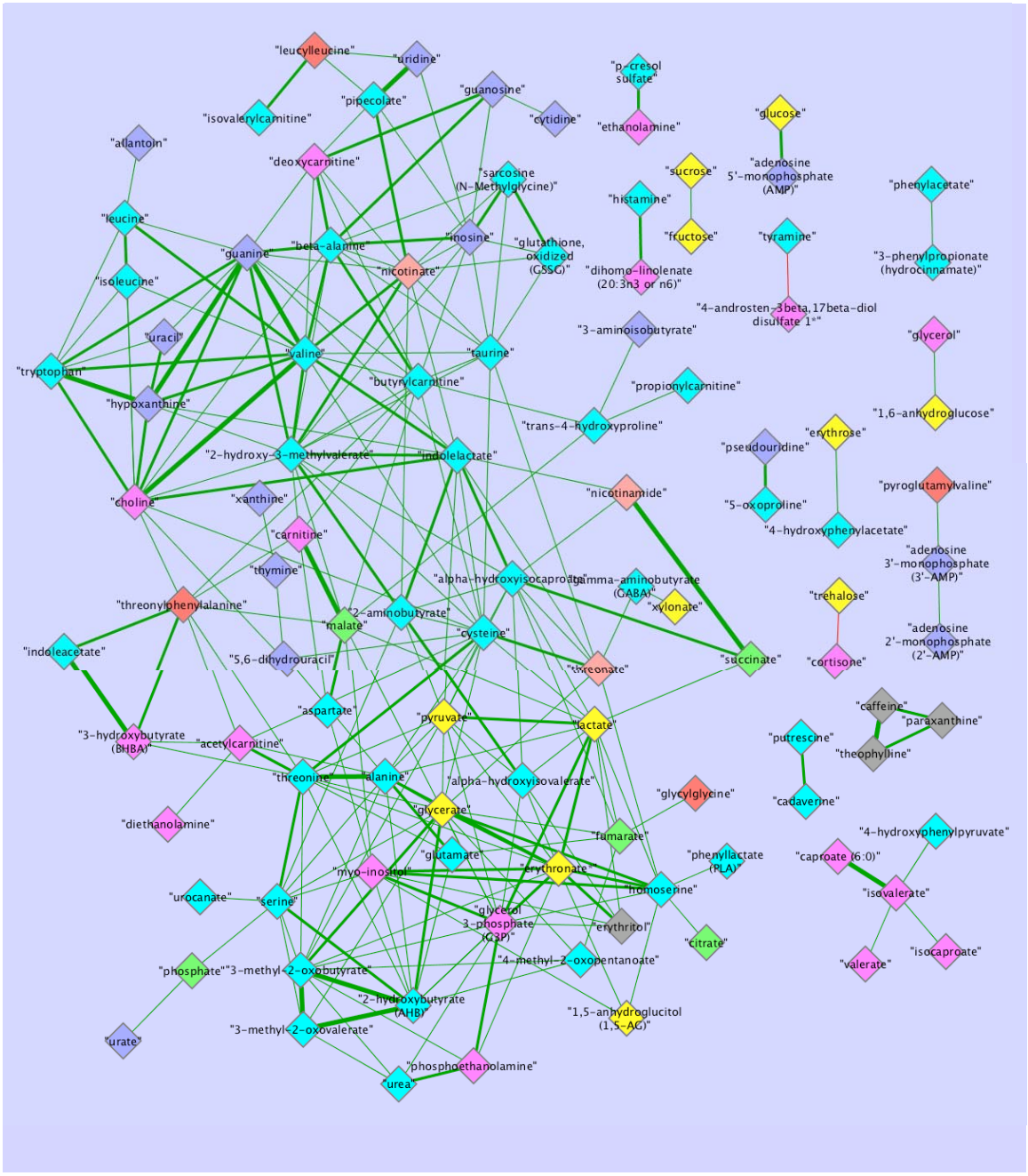



Figure 4.2.2-1 Combined HIV Class Metabolite Correlation Network with rho > |0.83|

Legend is in Figure 4.2-1

### 4.2.3 Oral Rinse Metabolite HIV HAART Feature Analysis

The Spearman Rank Correlation on all 198 metabolites for the Combined HIV Class samples generated correlation results that were filtered to only list correlations where rho is greater or equal to |0.87|. There are 302 correlations, where 272 are positive correlations, and 29 are negative correlations. The correlation network created from the results is in Figure 4.2.3-1. The Figure Legend is in Figure 4.2-1.

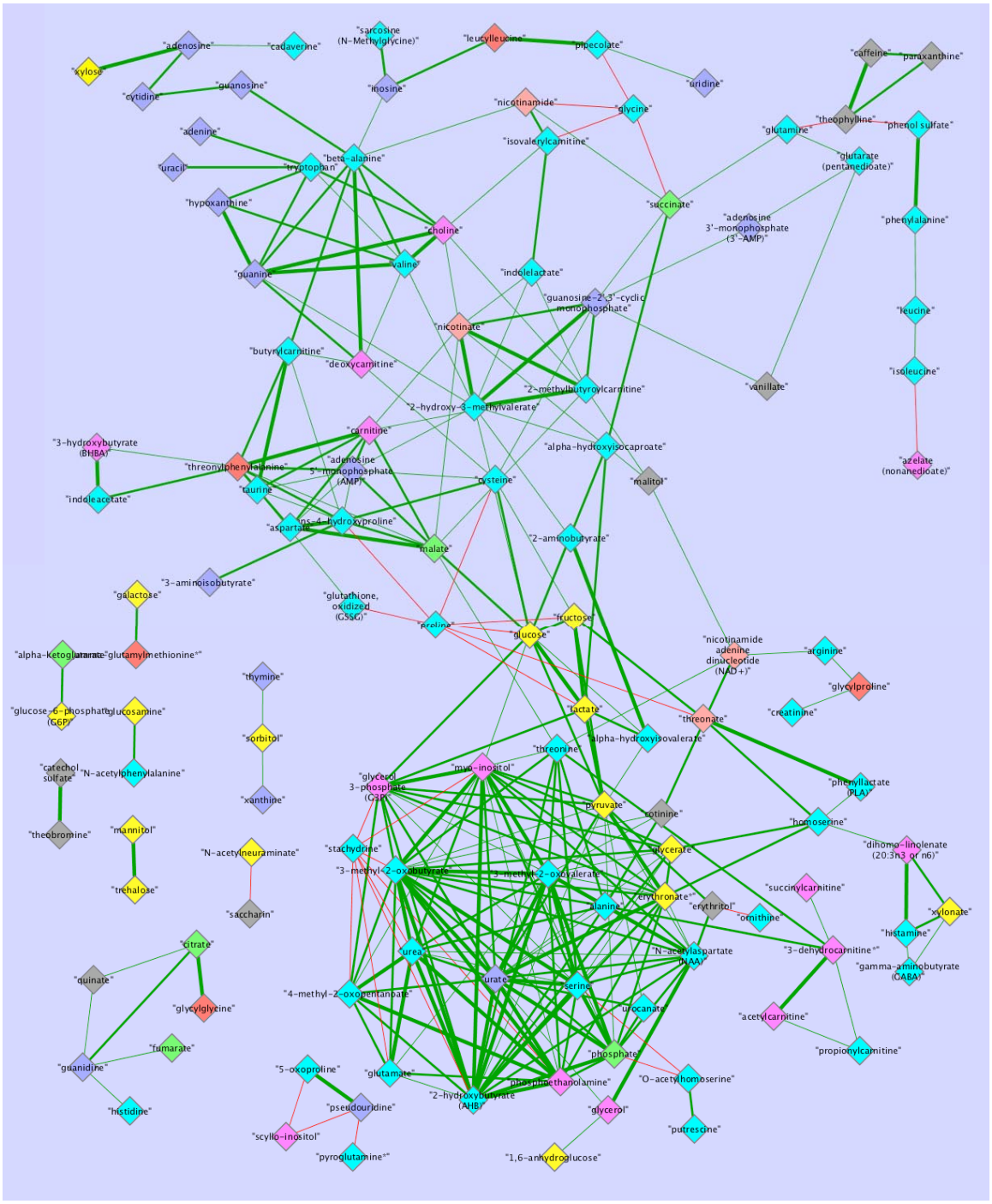



Figure 4.2.3-1 HAART HIV Class Metabolite Correlation Network with rho >= |0.87|
Legend is in Figure 4.2-1

### 4.2.4 Oral Rinse Metabolite Untreated HIV Feature Analysis

The Spearman Rank Correlation on all 198 metabolites for the untreated HIV Class samples generated correlation results that were filtered to only list correlations where rho is equal |1.0|. There are 1,237 correlations, where 1,157 are positive correlations, and 86 are negative correlations. The correlation network created from the results is in Figure 4.2.4-1. One can see the impact of excessive correlations by reducing the number of correlation pairs use to compute the correlation to the minimum allowed. The Figure Legend is in Figure 4.2-1.

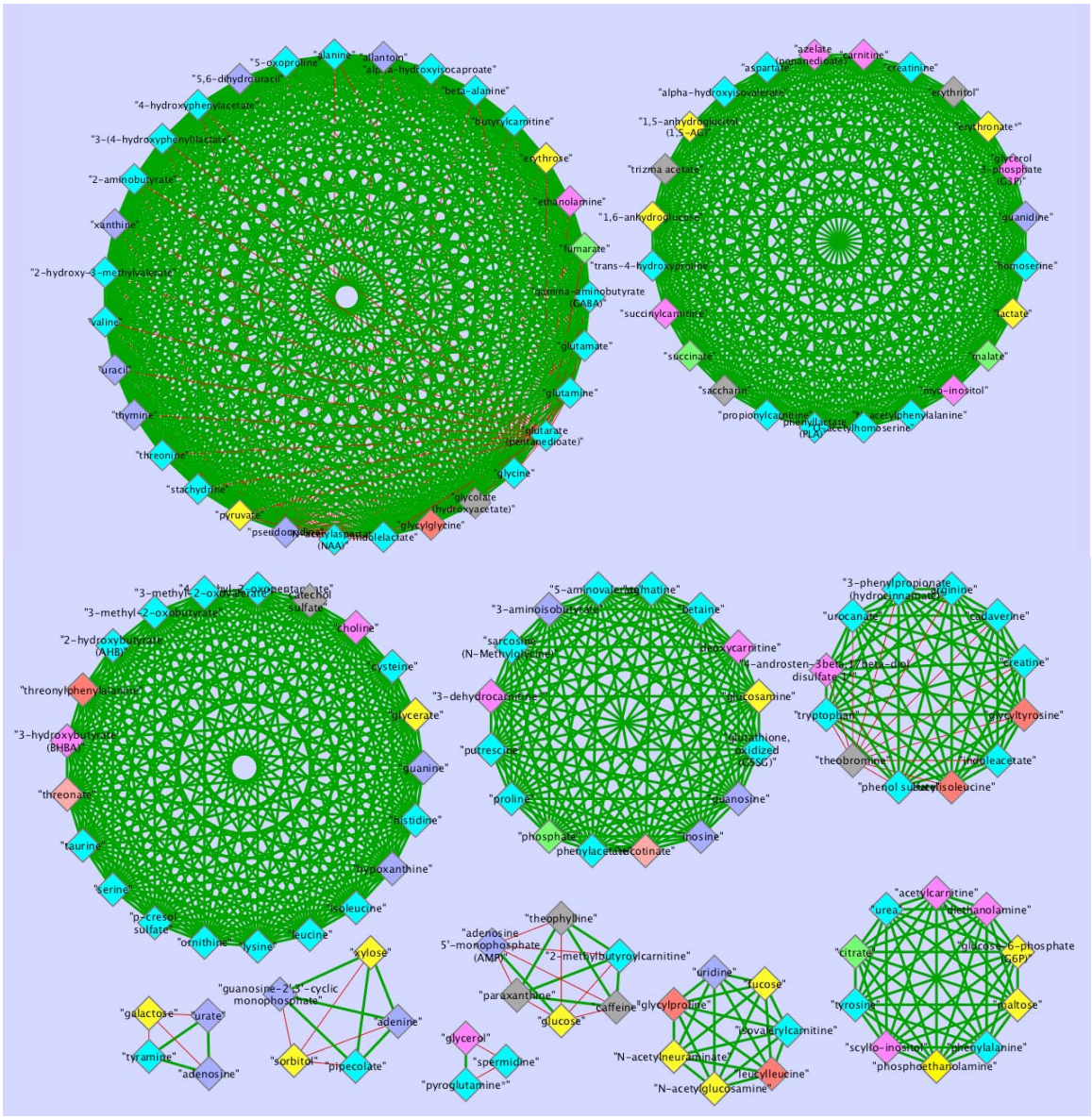



Figure 4.2.4-1 Untreated HIV Class Metabolite Correlation Network with rho = |1.0|.

Legend is in Figure 4.2-1

## *4.3 Oral Rinse Metabolite Differential Correlation Network Analysis*

Section 4.2 covered the analysis performed *within* each class as the basis to determine if there is a significant difference in metabolite pair correlations *across* classes. In this section, each of the Section 4.2 correlation datasets are paired for Differential Correlation Network (DCN) analysis, where the search for significantly different correlations for each Feature pair A-B across classes begins. Using the two class output data files created for Section 4.2, the probability assessment is based on the algorithm defined in Section 3.5 and is calculated via the 'Find_Corr_Significant_Diff' routine written in the Python language.

The null hypothesis states there would be no significance difference in the correlations for a pair of metabolites measured in Class 1 compared to Class 2. There is a theoretical maximum of 19,503 pair-wise metabolite pair significance comparisons to be reviewed. The minimum of 4 sample pairs per class is required for the probability Eq 3.5-1 to be determined. Therefore, if either class' metabolite pair's correlations didn't contain enough samples values, then the correlation probability cannot be computed. The three pertinent class comparisons presented in the following subsections are: Control vs. Combined HIV; Control vs. HAART HIV; and Control vs. untreated HIV. The DCN diagrams use a common legend, and to give the largest footprint to the data maps, the legend is listed below as Figure 4.3-1. The legend applies to all Figures in section 4.3.


Figure 4.3-1 DCN diagram legend for interpreting attributes to Section 4.3 diagrams.

The legend is similar to the one presented in Section 4.2, except the first two components have a different meaning here. The first component is the color coding for the edges (correlations *differences*). If the difference between the Control Class metabolites A-B correlation and the correlation for the same metabolites, A-B, in the HIV Class being compared exceeds the significant probability cutoff, the edge will appear in the DCN Figure. The color of the edge is green if the correlation value for the Control Class features A-B is lower than the correlation value for the HIV Class features. The edge is red if the Control Class value is greater than the HIV Class correlation value. The middle left portion of the legend is different also. The node pair edge thickness

guide now indicates the significance of the probability of the correlation difference, the thicker the edge the higher the significance. If the correlation difference probability, p <= 5% the edge is thickest, if p <= 10% the edge is medium thickness, and p < 90%, but over the probability significant lower limit, the edge is thinnest.

The last legend item is the same as the Figure 4.2-1 legend, but restated for completeness. The right side of the legend, is the mapping of the KEGG biological superpathway for the metabolite (node). The node will display the color indicated in the Figure 4.2-1 based on the metabolite superpathway attribute.

In each of the following subsections the difference correlation analysis was performed for all 198 metabolites listed. Two requirements determine if a metabolite pair will become a DCN statistically significant result. First, is that metabolites had sufficient metabolite value pairs to perform their initial intra-class correlation calculation. The second requirement is there must be a minimum of 4 sample pairs in each class to ascertain the significance difference. By including immeasurable values these two requirements were always satisfied. Metabolite pairs had a probability assigned to their two-class correlation difference and are included in the resulting output data if their probability was below the p =<16% cutoff. The significance standard was relaxed for the small sample sizes and for inclusion in the knowledge discovery process.

### 4.3.1 Metabolite Control vs. Combined HIV Differential Correlation Network Analysis

This subsection addresses the significant correlation difference calculation and presentation for the Control Class versus the Combined HIV Class. The Control Class data values are from samples 13-24. The Combined HIV Class data values are from

samples 1-12. There are 243 correlation differences meeting the above criteria, with 140 showing an higher correlation in the Combined HIV Class as compared to the Control Class, and 103 correlation differences were the reverse is true. To create a network map the results data was imported into Cytoscape, along with the Metabolite (Node) Metadata. The resulting DCN map is in Figure 4.3.1-1. The figure legend is in Figure 4.3-1.

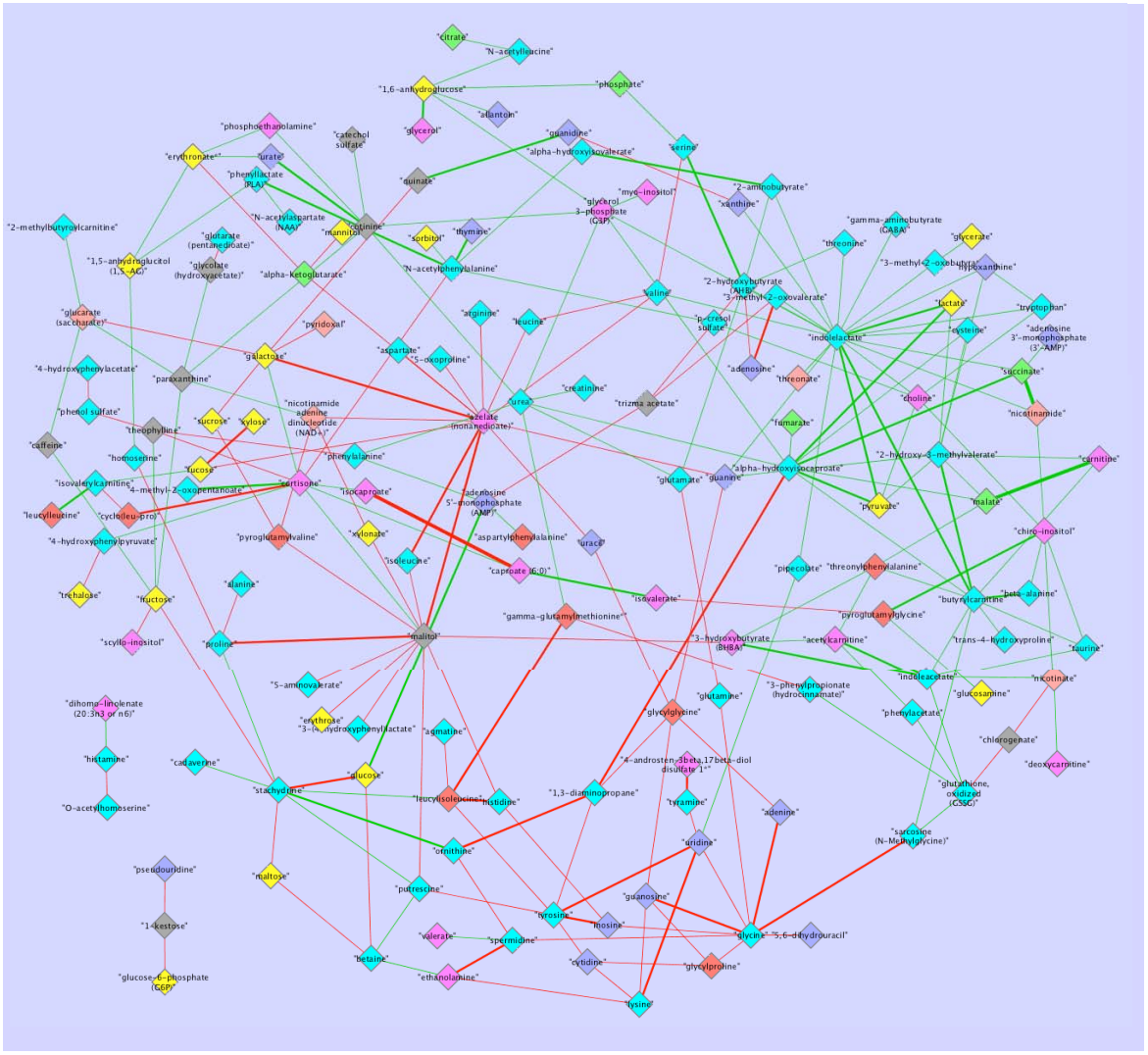



Figure 4.3.1-1 Control Class versus Combined HIV Class Metabolite DCN Diagram with p < 15%. The figure legend is in Figure 4.3-1.

## 4.3.2 Oral Rinse Control vs. HAART HIV Metabolite Differential Correlation Network Analysis

This subsection addresses the significant correlation difference calculation and presentation for the Control Class versus the HAART HIV Class. The Control Class data

values are from samples 13-24. The HAART HIV Class data values are from samples 1-8. There are 323 correlation differences meeting the above criteria, with 107 showing an higher correlation in the HAART HIV Class as compared to the HAART Class, and 216 correlation differences were the reverse is true. To create a network map the results data was imported into Cytoscape, along with the Metabolite (Node) Metadata. The overview of the resulting DCN map is in Figure 4.3.2-1. The figure legend is in Figure 4.3-1.

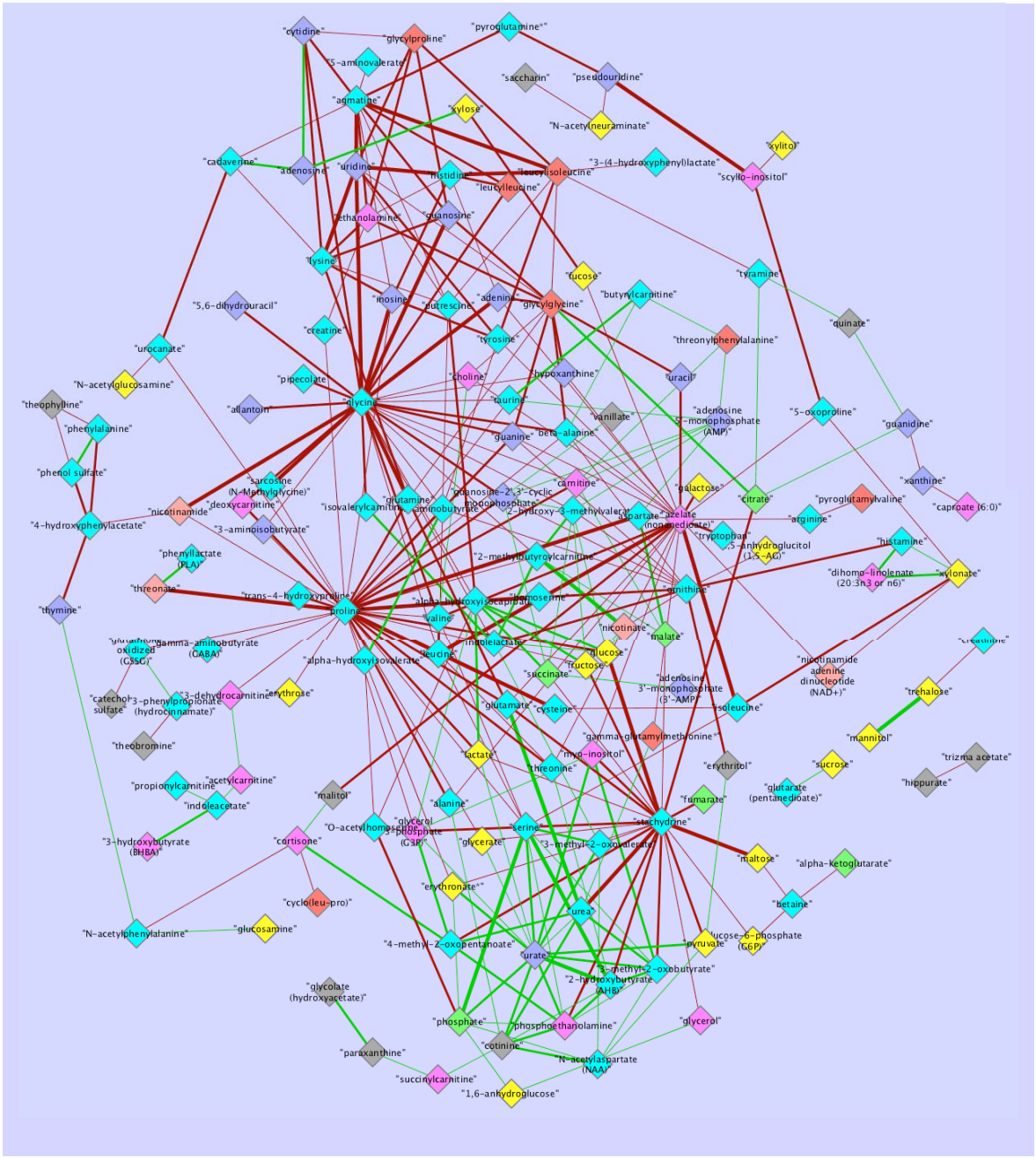



Figure 4.3.2-1 Control Class versus HAART HIV Class Metabolite DCN Diagram with p <17%. The figure legend is in Figure 4.3-1.

### 4.3.3 Oral Rinse Control vs. Untreated HIV Metabolite Network Analysis

This subsection addresses the significant correlation difference calculation and presentation for the Control Class versus the untreated HIV Class. The Control Class data values are from samples 13-24. The untreated HIV Class data values are from samples 9-12. There are 453 correlation differences meeting the above criteria, with 368 showing a higher correlation in the untreated HIV Class as compared to the untreated Class, and 83 correlation differences were lower. To create a network map the Results data was imported into Cytoscape, along with the Metabolite (Node) Metadata. The resulting DCN map is in Figure 4.3.3-1. The figure legend is in Figure 4.3-1.

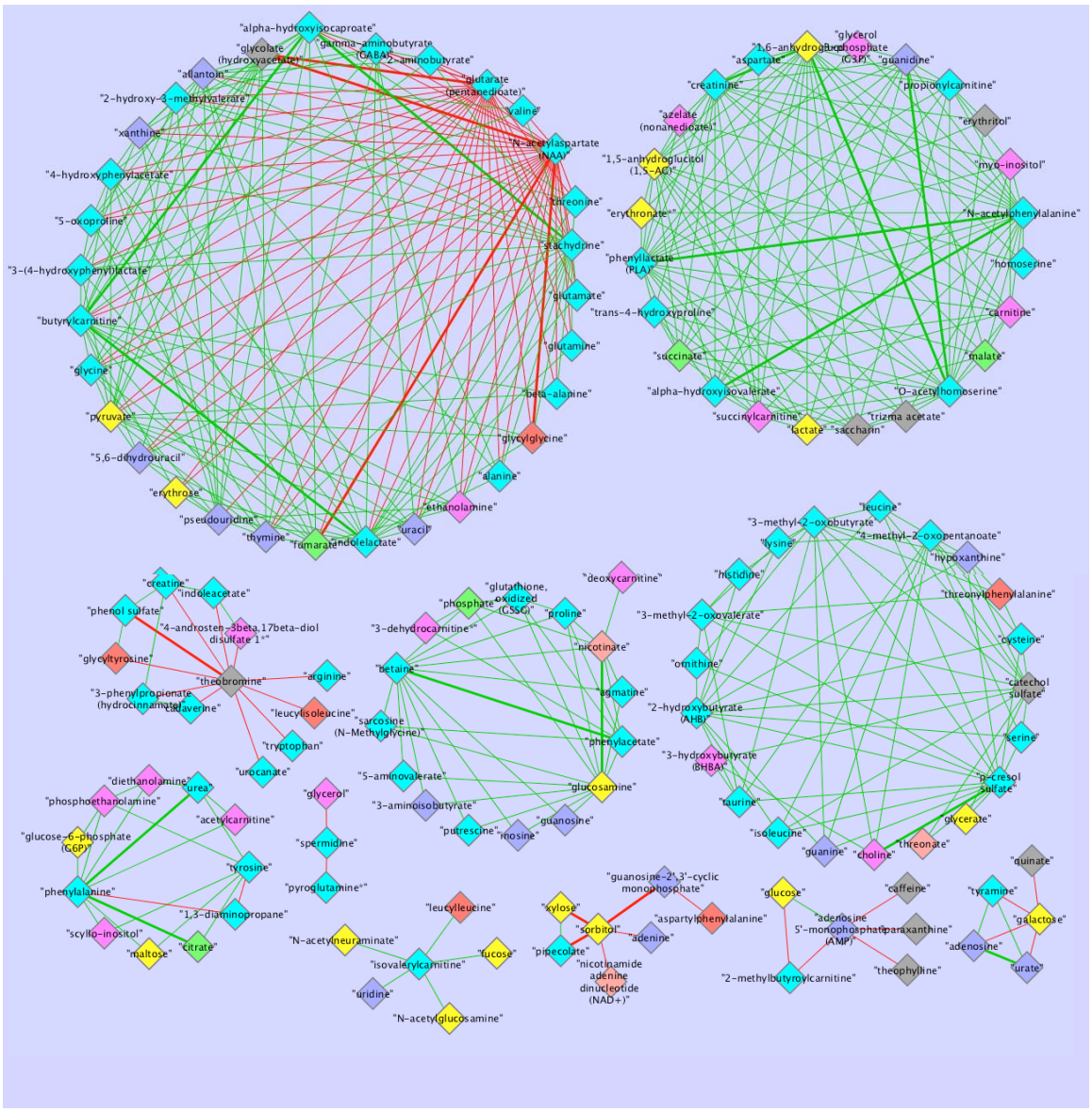



Figure 4.3.3-1 Control Class versus Untreated HIV Class Metabolite DCN Diagram with p<17%. The figure legend is in Figure 4.3-1.

## *4.4 Summary*

Charter 4 began an overview the Oral Rinse clinical trial metabolite only data from the 24

samples that defined the study. The oral rinse samples were comprised of 12 Control and 12 HIV samples, with the HIV being subdivided into 8 on Highly Active Anti-Retroviral Therapy (HAART) and 4 samples that were untreated. The samples had variable number of measurable and immeasurable (undetectable values) for each of 198 metabolites. The analysis performed specified all immeasurable values to be included and that each feature pair must have a least 30% measurable values for one of the metabolite pairs. In Section 4.2 the correlation analysis was performed on all 4 classes – the Control Class, Combined HIV, both HAART, and untreated HIV. The results were explained and network diagrams displayed the pertinent results. Then an analysis of the Confidence Intervals for the correlations derived was presented. Section 4.3 addressed the Differential Correlation Networks (DCN) and the significant correlation differences that comprise it. DCNs were produced, the results presented, and DCN maps displayed for – Control vs. combined HIV, Control vs. HAART, and Control vs. untreated HIV pairings. The oral rinse metabolite profile for the one control class versus the three HIV disease classes presents significantly different metabolite correlations. Chapter 5 will purse the same line of analysis for the complete oral metabiome – i.e. metabolites, bacteria, and fungi.

# Chapter 5: Analysis of the Oral Metabiome

This Chapter will examine data produced by the Case Western (CW) HIV Oral Rinse Metabolome Study (M. Ghannoum 2011). This chapter extends the metabolite analysis of Chapter 4 to include the entire oral rinse metabiome. Based on the methodology outlined in Chapter 3 the Differential Correlation Network analysis was applied against the four classes of Case Western Study Oral Rinse metabiome samples and the analytic results were integrated for display into Differential Correlation Network diagrams.

The chapter is organized as follows. The hypothesis is restated, followed by section 5.1 that describes the oral rinse metabiome sample data, focusing on the bacterial and fungal components. Metabiome correlation analysis within each class is addressed in Section 5.2. Section 5.3 constructs, from the results of Section 5.2, Differential Correlation Networks (DCN) displaying the oral rinse metabiome pair statistically significant results. The last section summarizes the chapter's contributions. The Confidence Intervals associated with the difference correlations is calculated in Section 5.4.

The focus of the Differential Correlation Network pipeline is knowledge discovery leading to hypotheses generation. In combination with other biological repositories one can postulate extra-cellular underlying biological pathways from the results. Chapter 7 will present the findings, and conclusions, based on the results in Chapters 4 & 5. The

hypothesis follow:

(H1) Examining inter-class differential feature pair correlations will differentiate healthy versus disease classes.

(H2) The application of Differential Correlation Network analysis on experimental data support knowledge discovery related to underlying biological process variation between healthy and disease states.

## *5.1 Oral Rinse Bacterial and Fungal Genus Identification*

The metabiome analysis is comprised of three data types being analyzed concurrently – bacteria, fungi, and metabolites. The metabolite component was identified in Section 4.1. The bacterial and fungal components were analyzed and their taxa identified by the analysis of rRNA sequence data and corresponding databases as overviewed in Chapter 2. Due to the technical limitation of the sequencing process, only taxa abundances greater than 1% were used for the bacterial or fungal samples. Oral rinse analysis identified 295 features comprised of 58 bacteria genus, and 39 fungal genus, and 198 metabolites, across all 24 samples in the study. The following Figures 5.1-1 and 5.1-2 give a statistical overview of the genus abundances.

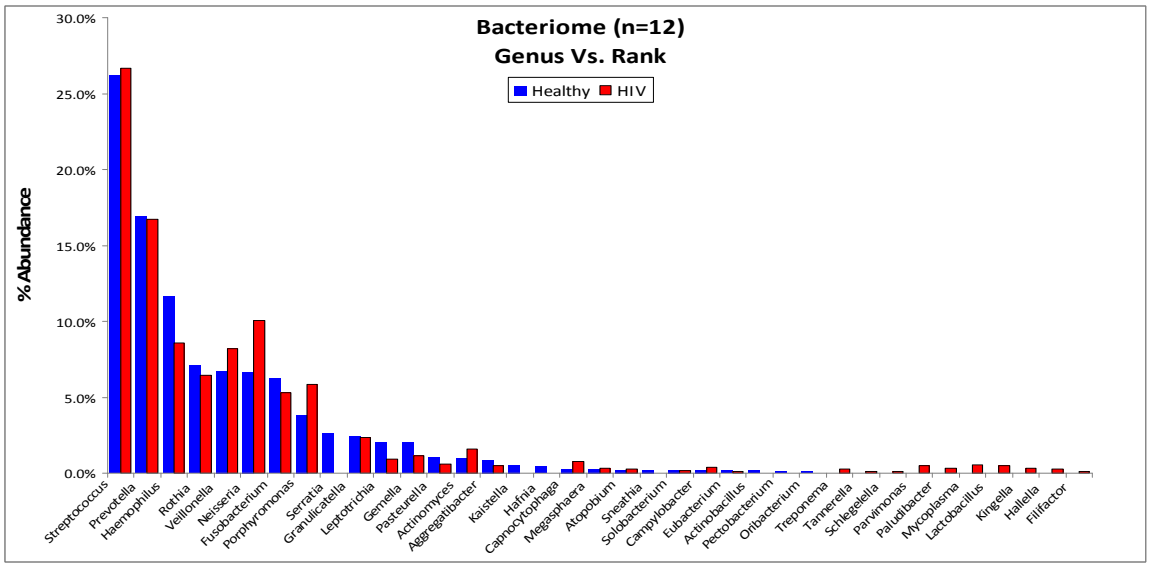
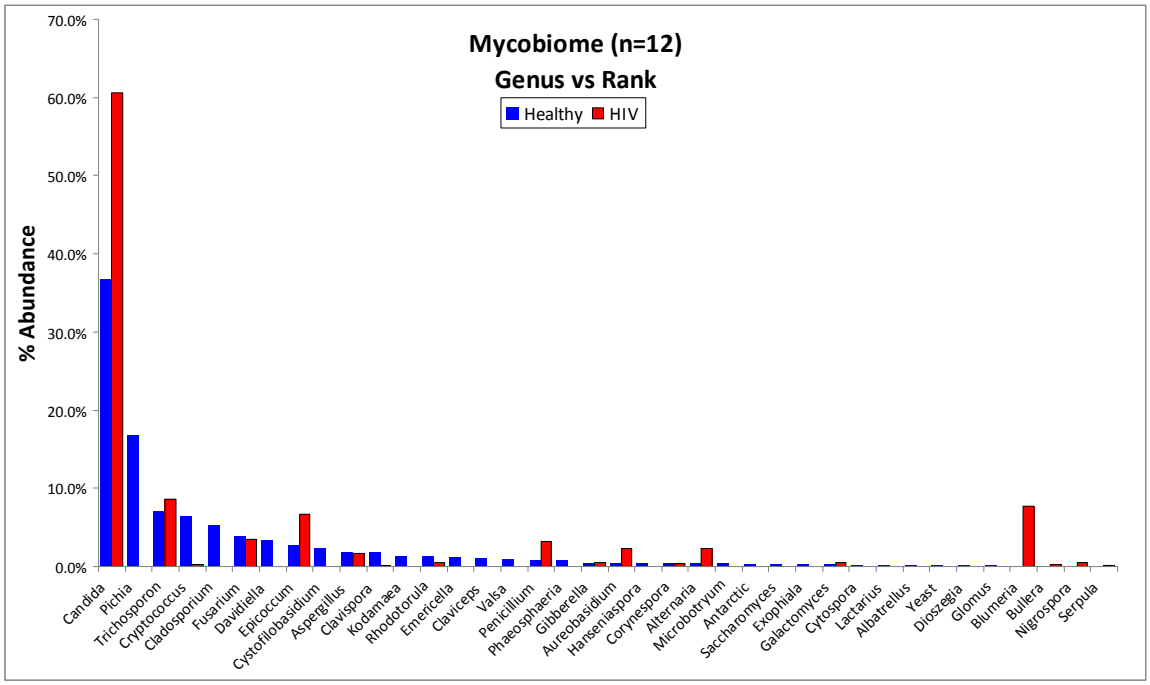



Figure 5.1-1 Oral Rinse Bacterial Genus abundances (Credit M. Retuerto CW)




0-11Figure 5.1-2 Oral Rinse Fungal Genus abundances (Credit M. Retuerto CW)

The analysis for Chapter 5 involves all of the CW Oral Rinse study metabiome – bacterial, fungal, and metabolite sample data was categorized in the same manner as used for metabolites in Chapter 4. To reiterate; the four Classes are – the Control Class is the 12 control samples, Combined HIV Class is the combination of all 12 HIV samples, both HAART and untreated HIV, the HAART Class and untreated. The HAART Class contains the 8 samples comprising the Anti Retroviral Therapy (HAART) oral rinse samples. Lastly, the untreated Class is the remaining 4 HIV samples.

## *5.2 Oral Rinse Bacterial and Fungal Genus Classes and Features*

The combination of all identified metabolites, bacteria at the genus level, and fungus at the genus level across all 24 samples yielded a total of 295 features to be quantified for each sample in each Class. Table 5.1-1 and Table 5.1-2 display the bacterial and fungal abundances by sample, the metabolite quantities are in Table 4.1-1. The analysis specified the inclusion of immeasurable metabolite, bacterial, and fungal values, and required at least 30% measurable values in one of these features of each feature pair to perform a correlation calculation. If this criteria was not met that feature pair was excluded from the correlation calculation, leaving the other N-1 Class feature pairs to be reviewed. The in-class correlations were the basis to determine if there is a significant difference in each metabolite pair correlation *across* Classes. There is a theoretical maximum with 295 features of 43,365 = (295*(295-1))/2 pair-wise feature pair correlations. As stated earlier this total is reduced if the calculation requirements are not met. The Python ‘Spearman Correlation’ routine was executed to create each class correlation output file. The file contains all possible correlations for use as input into the

DCN algorithm later. The correlation maps in this section only include correlations where one of the features is either a fungi or bacteria. Therefore, the correlation pair totals for Section 5.2, are in addition to the metabolite-metabolite feature correlations for the same Class in Section 4.2. The totals are 10 for the Control Class, 5 for the Combined HIV Class, 17 for the HAART Class, and 125 for the untreated HIV Class, as show in the network diagrams. The legend is listed below Figure 5.2-1.

The Chapter 5 DCN diagrams are comprised of nodes and edges. Nodes represent one of three features – bacteria, fungi, or metabolites. The edges appear if there is a significant correlation difference between two features, based on the two classes being compared that comprise the DCN. If a node pair correlation difference has a probability greater than the defined cutoff value, an edge will appear. The cutoff minimum is different per DCN because of the number of significant correlation pairs varies per DCN. To keep the diagram comprehensible the cutoff value for edges to appear is raised to reduce the number for nodes, and correlation difference edges, displayed.

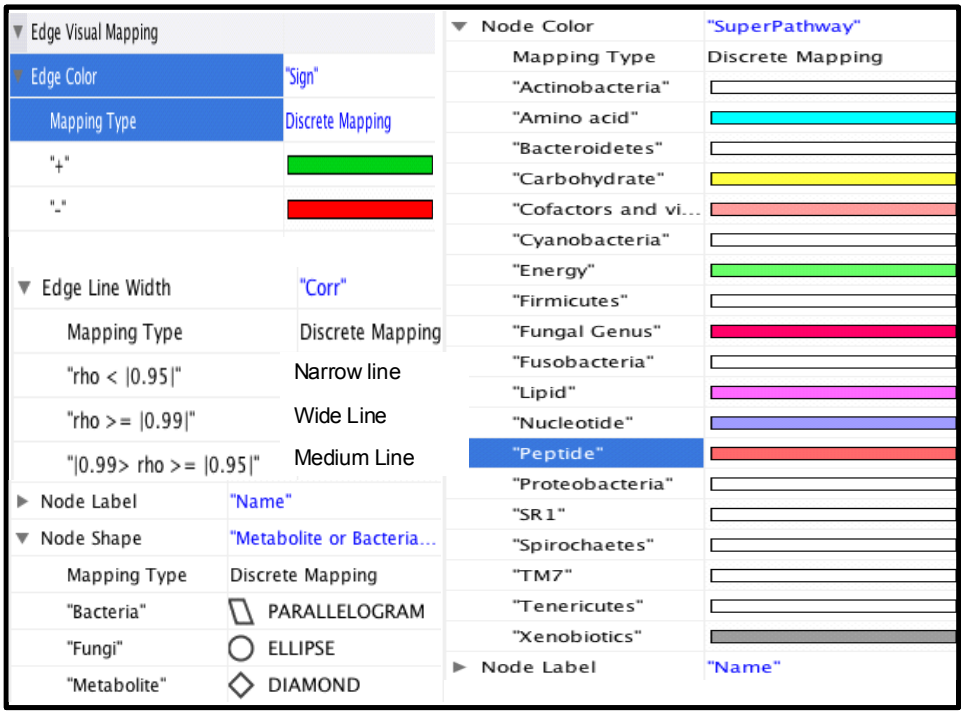



Figure 5.2-1 The Correlation Network Legend for interpreting diagram attributes.

Chapter 5 Correlation Network diagram nodes are bacteria, fungi, or metabolites. The edges are correlation connections between two nodes -- if their correlation value, rho, is greater than the cutoff value. The cutoff applies to |rho| so both large positive and negative correlations are displayed. The cutoff minimum varies per class because of the number of correlation pairs varies. To keep the diagram comprehensible the cutoff display value is raised to reduce, or increase, the number for nodes and connections displayed.

The legend presented in Figure 5.2-1 has 4 main components. First, is the color coding for the edges (correlations). If the correlation, rho, between two features (nodes) is positive, and greater than the network diagram display cutoff, it will display as a green

line. If the feature pair correlation is negative and |rho| >= display cutoff, the edge color is red. Second middle left of legend, is the node pair correlation edge thickness guide, the greater the correlation, the thicker the edge. If the correlation, rho >= 0.99 the edge is thickness, if rho >=0.95 it is the medium thickness, and rho < 0.95 represents the thinness edge. Third legend item, bottom left, is the node shape. It is related to whether the feature (node) is a metabolite, or bacteria, or fungi. Lastly, on the right side of the legend, is the mapping of the KEGG biological superpathway for the metabolites. The metabolite node will display the color indicated in the Figure 5.2-1 based on the superpathway attribute value.

For each of the following 5.2 subsections the data for the figures has been filtered to only display bacterial or fungal nodes and their direct connecting edges. A metabolite node will only appear if it has a direct edge connection with a bacteria or fungus. This is to simplify the network diagrams. The correlation networks for the class for the other non-bacteria and non-fungal nodes are in the corresponding network diagrams in Chapter 4. The direct connection filtered data results are imported into Cytoscape, along with the Metabolite (Node) Metadata.

### 5.2.1 Oral Rinse Control Metabiome Feature Analysis

The result of performing a Spearman Rank Correlation, including all immeasurable values, for the Control Class samples 13 to 24 from Table 4.1-2, for all 295 metabiome features listed generated correlation results that were filtered to only list correlations where rho is greater or equal to |0.84|. There are 9 bacterial/fungal positive

correlations, and 1 bacterial/fungal negative correlation. The overall view of the correlation network created from the results is in Figure 5.2.1-1. The legend for this figure is in Figure 5.2-1.

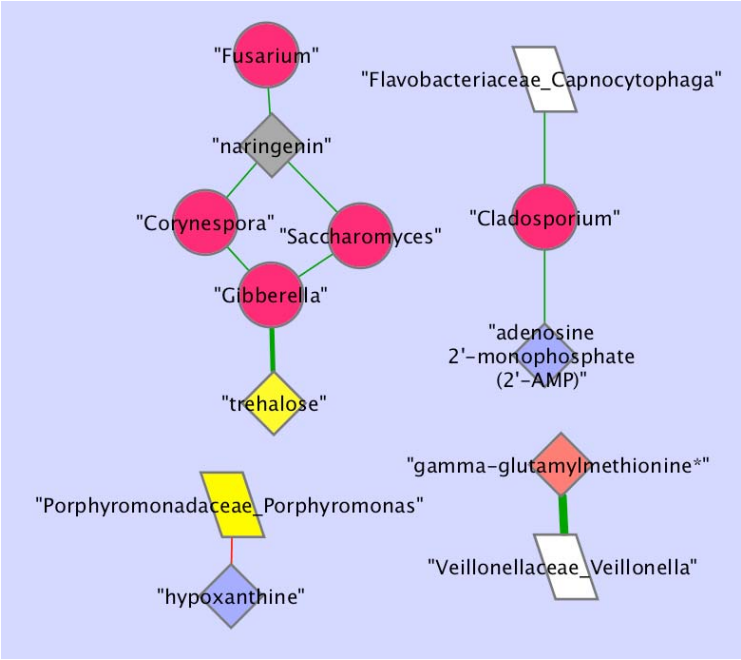



Figure 5.2.1-1 Directly Connected Bacterial or Fungal Control Class Metabolome Correlation Network rho >= |0.84|. The legend is in Figure 5.2-1.

## 5.2.2 Oral Rinse Metabiome Combined HIV Feature Analysis

The result of performing a Spearman Rank Correlation, including all immeasurable values, for the Combined HIV Class samples 1 to 12 from Table 4.1-2, for all 295 metabiome features generated correlation results that were filtered to only list

correlations where rho is greater or equal to |0.84|. There are 5 bacterial/fungal positive correlations and zero bacterial/fungal negative correlations. The overall view of the correlation network created from the results is in Figure 5.2.2-1. The legend for this figure is in Figure 5.2-1.

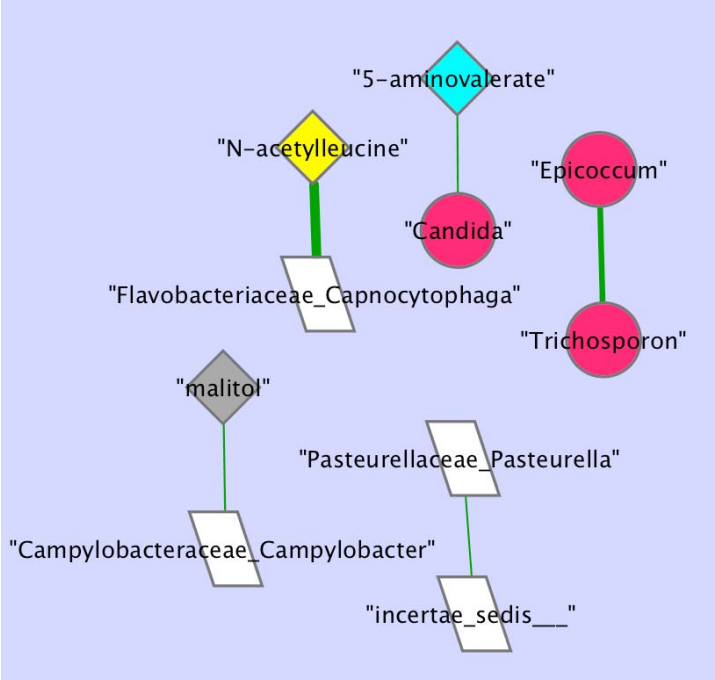



Figure 5.2.2- Directly Connected Bacterial or Fungal Combined HIV Class Metabolome Correlation Network rho >= |0.84|. The legend is in Figure 5.2-1

## 5.2.3 Oral Rinse Metabiome HAART HIV Feature Analysis

The result of performing a Spearman Rank Correlation, including all immeasurable

values, for the HAART HIV Class samples 1 to 8 from Table 4.1-2, for all 295 metabiome features generated correlation results that were filtered to only list correlations where rho is greater or equal to |0.88|. There are 10 bacterial/fungal positive correlations and 7 bacterial/fungal negative correlations. The overall view of the correlation network created from the results is in Figure 5.2.3-1. The legend for this figure is in Figure 5.2-1.

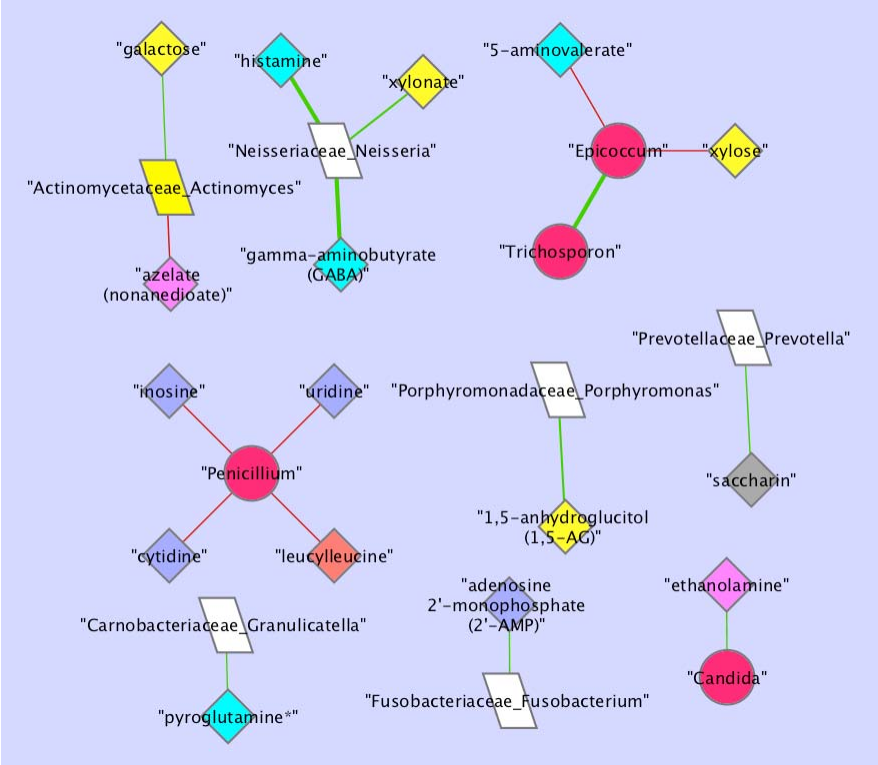



Figure 5.2.3-1 Directly Connected Bacterial or Fungal HARRT Class Metabiome Correlation Network rho >= |0.88|. The legend is in Figure 5.2-1

## 5.2.4 Oral Rinse Metabiome Untreated HIV Feature Analysis

The result of performing a Spearman Rank Correlation, including all immeasurable values, for the untreated HIV Class samples 9 to 12 from Table 4.1-2, for all 295 metabiome features generated correlation results that were filtered to only list correlations where rho is greater or equal to |0.95|. There are 49 positive bacterial/fungal correlations and 76 bacterial/fungal negative correlations. The overall view of the correlation network created from the results is in Figure 5.2.4-1. The legend for this figure is in Figure 5.2-1.

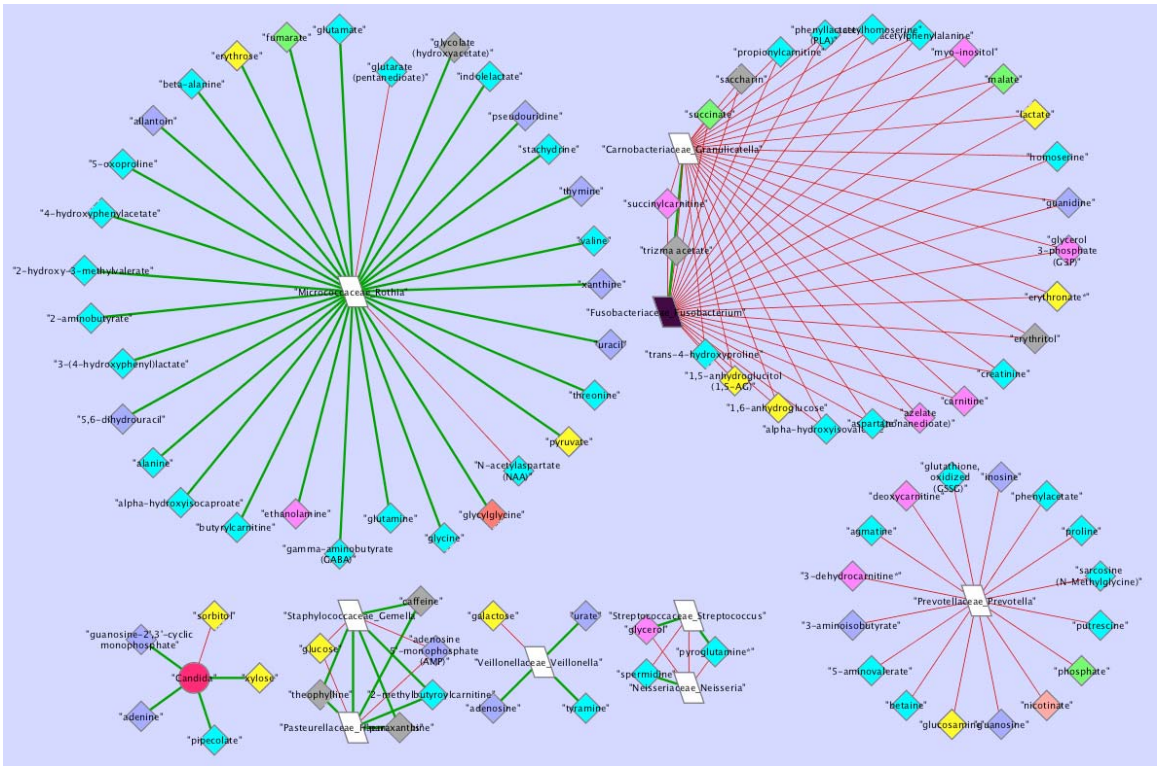



Figure 5.2.4-1 Directly Connected Bacterial or Fungal Untreated HIV Class Metabolome Correlation Network rho >= |.95|. The legend is in Figure 5.2-1

With so few integer rank value samples per feature the ability to achieve a 100% positive or negative correlation is increased, as oppose to using real number data values. Therefore, there are many high probability differential correlations, but their confidence

interval is comparatively weak as discussed in Section 5.4.

## 5.3 Oral Rinse Metabiome Differential Correlation Network Analysis

Section 5.2 covered the analysis performed *within* each class as the basis to determine if there is a significant difference in metabolite pair correlations *across* classes. Here each of the Section 5.2 correlation datasets are paired for Differential Correlation Network (DCN) analysis revealing significantly different correlations from each Feature pair A-B across classes. Using the two class output data files created for Section 5.2, the statistical probability assessment, Eq. 3.5-1, is calculated via the 'Find_Corr_Significant_Diff' routine implemented in the Python language.

The null hypothesis states there will be no significance difference ($p < 5\%$) in the correlations of each pair of features measured in Class 1 as compared to the Class 2. There is a theoretical maximum of 43,365 pair-wise metabiome pair significance comparisons to be reviewed. Four is the minimum required sample pairs from each class being compared. Therefore, if either class's metabiome pair's correlations didn't contain enough sample values, then the correlation difference statistic cannot be computed.

The three pertinent metabiome feature class comparisons presented in the following subsections are – Control vs. Combined HIV, Control vs. HAART HIV, and Control vs. untreated HIV. The DCN diagrams use a common legend that is listed below as Figure 5.3-1. The legend applies to all figures in section 5.3.

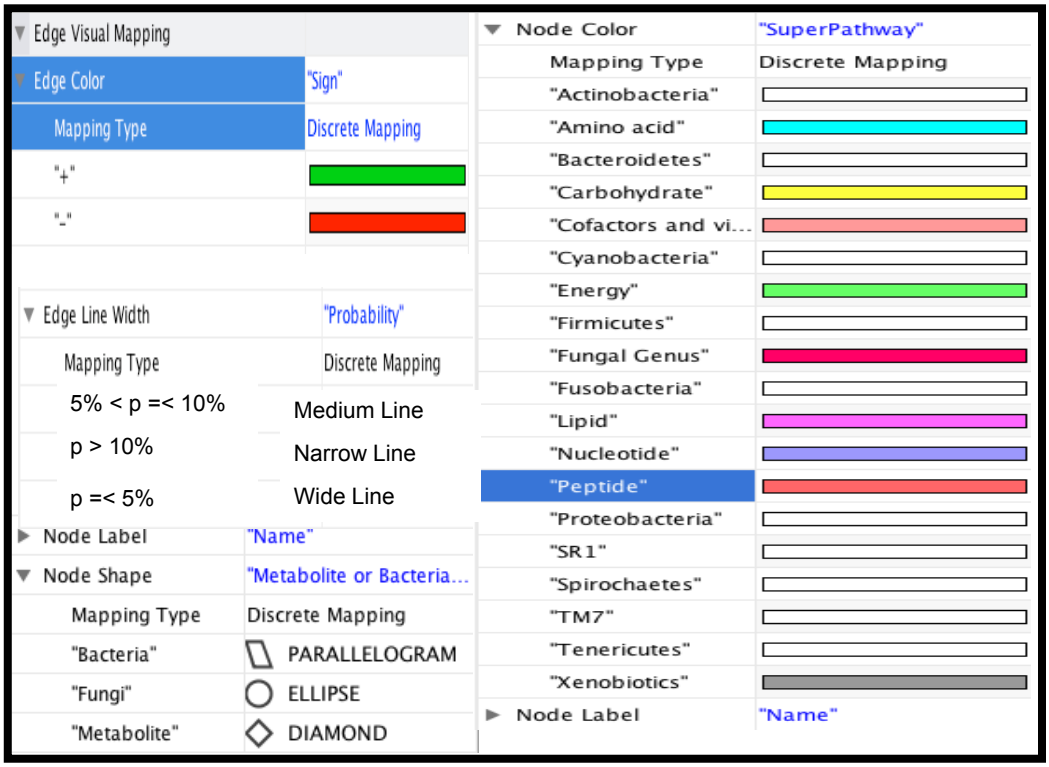



Figure 5.3-1 DCN Map legend for interpreting attributes for Section 5.3.

The first component of the legend figure 5.3-1is the color coding for the edges (correlations *differences*). If the difference between the Control Class metabiomes A-B correlation and the correlation for the same metabiomes, A-B, in the HIV Class being compared exceeds the significant probability cutoff, the edge will appear in the DCN Figure. The color of the edge is green if the correlation value for the Control Class features A-B is lower than the correlation value for the HIV Class features. The edge is red if the Control Class value is greater than the HIV Class correlation value. The middle left portion of the legend is different also. The node pair edge thickness guide now indicates the significance of the probability of the correlation difference, the thicker

the edge the higher the significance. If the correlation difference probability, p <= 5% the edge is thickest, if p <= 10% the edge is medium thickness, and p > 10%, but over the probability significant lower limit, the edge is thinnest. The bottom left of the legend indicates the node shape for all 3 feature types -- bacteria, fungi, or metabolite. Lastly, on the right side of the legend, is the mapping of the KEGG biological superpathway for the metabolome (node). The node will display the color indicated in the Figure 5.3-1 based on the metabolite superpathway attribute.

In each of the following subsections the difference correlation analysis was performed for all 295 metabiome features; bacteria, fungae and metabolites. Metabiome pairs had a probability assigned to their two-class correlation difference and are included in the resulting output data. The probability cutoff of p =<16% was used to allow for additional differential correlations to be reviewed as part of the knowledge discovery. The significance standard was relaxed for the small sample sizes and for inclusion in the knowledge discovery process.

For each of the following 5.3 subsections the data for the figures has been filtered to only display bacterial or fungal nodes and their direct connecting edges. A metabolite node will only appear if it has a direct edge connection with a bacteria or fungus. This is to simplify the network diagrams. The correlation networks for the class for the other non-bacteria and non-fungal nodes are in the corresponding network diagrams in Section 4.3. The direct connection filtered data results are imported into Cytoscape, along with the Metabiome (Node) Metadata.

### 5.3.1 Metabiome Combined HIV versus Control Class Analysis

This subsection addresses the significant correlation difference calculation and presentation for the Control Class versus the Combined HIV Class. The Control Class data values are from samples 13-24 and the Combined HIV Class data values are from samples 1-12, creating the results set. As in Section 5.2 only the additional correlation differences associated with bacteria and fungi are included. These are in addition to previous corresponding DCNs from Section 4.3. There are 101 correlation differences meeting the above criteria, with 59 showing a higher correlation in the Combined HIV Class as compared to the Control Class, and 42 correlation differences where the reverse is true. To create a network map the direct bacteria or fungal edge filtered results data were imported into Cytoscape, along with the Metabiome (Node) Metadata. The overview of the resulting DCN map is in Figure 5.3.1-1.

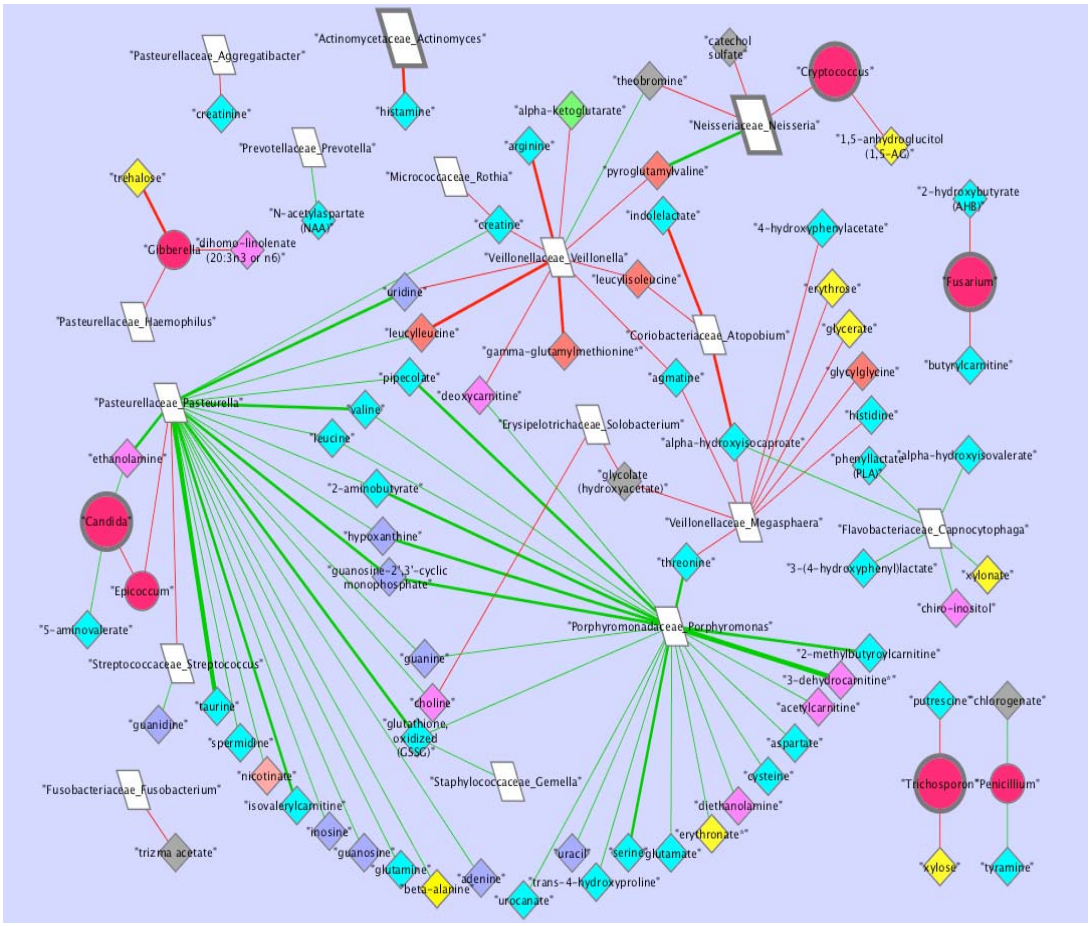



Figure 5.3.1-1 Directly Connected Bacterial or Fungal Only DCN for Control versus Combined HIV Classes p < 17%. The legend is in Figure 5.3-1.

## 5.3.2 Oral Rinse HAART HIV versus Control Class Metabiome Network Analysis

This subsection addresses the significant correlation difference calculation and presentation for the Control Class versus the HAART HIV Class. The Control Class data values are from samples 13-24 in Table 5.2.1-1 as are The HAART HIV Class data values samples 1-8. There are 51 correlation differences meeting the above criteria,

with 37 showing a higher correlation in the HAART HIV Class as compared to the Control Class, and 14 correlation differences are lower. To create a network map the direct bacteria or fungal edge filtered results data was imported into Cytoscape, along with the Metabiome (Node) Metadata. The overview of the resulting DCN map is in Figure 5.3.2-1.

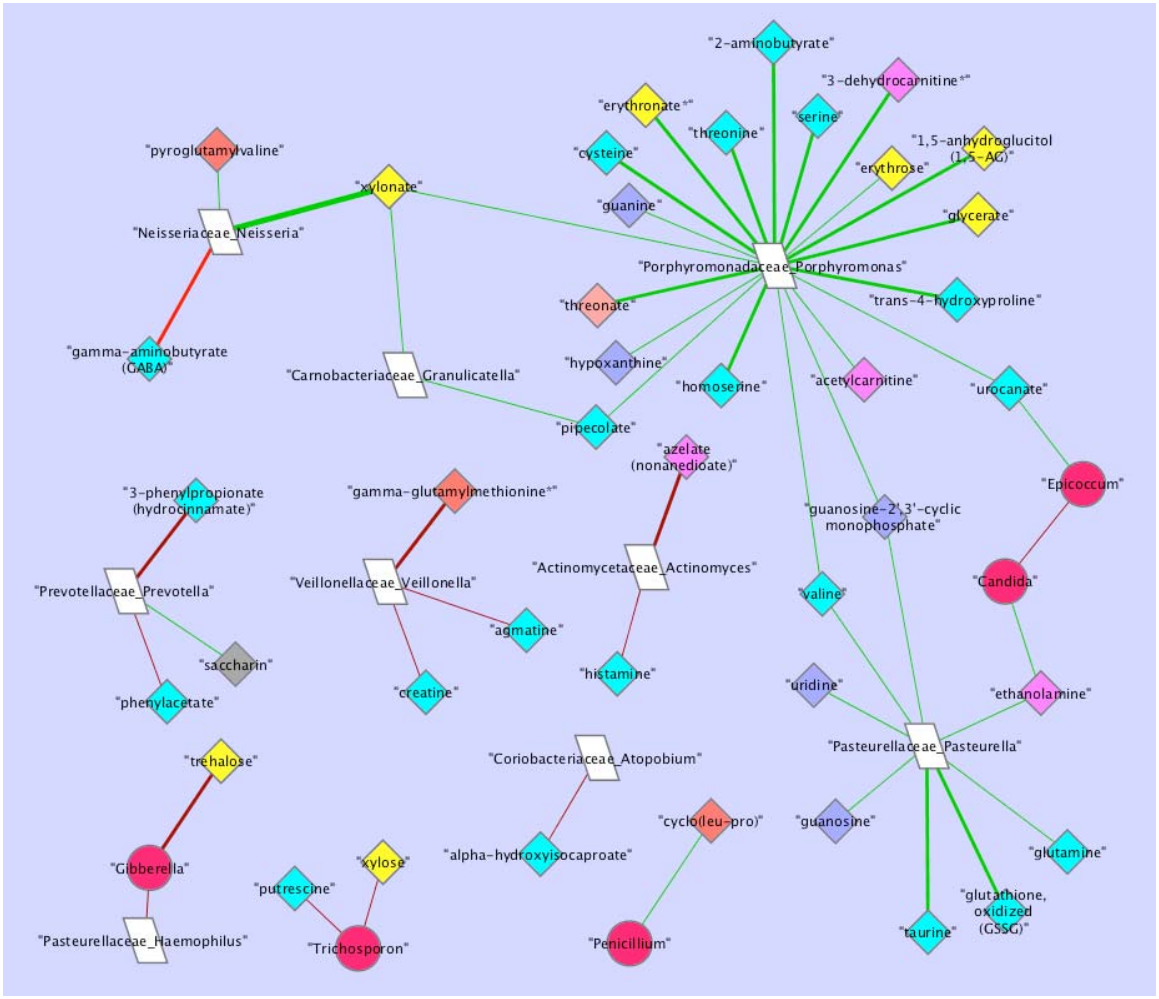



Figure 5.3.2-1 Directly Connected Bacterial or Fungal Only DCN for Control versus HAART HIV Classes p < 17%. The legend is in Figure 5.3-1.

## 5.3.3 Metabiome Untreated HIV versus Control Class Analysis

This subsection addresses the significant correlation difference calculation and presentation for the Control Class versus the untreated HIV Class. The Control Class

data values are from samples 13-24 and the untreated HIV Class data values are from samples 9-12. There are 93 correlation differences meeting the above criteria, with 39 showing a higher correlation in the untreated HIV Class as compared to the Control Class, and 54 correlation differences were the reverse is true. To create a network map the direct bacteria or fungal edge filtered results data was imported into Cytoscape, along with the Metabiome (Node) Metadata. The overview of the resulting DCN map is in Figure 5.3.3-1.

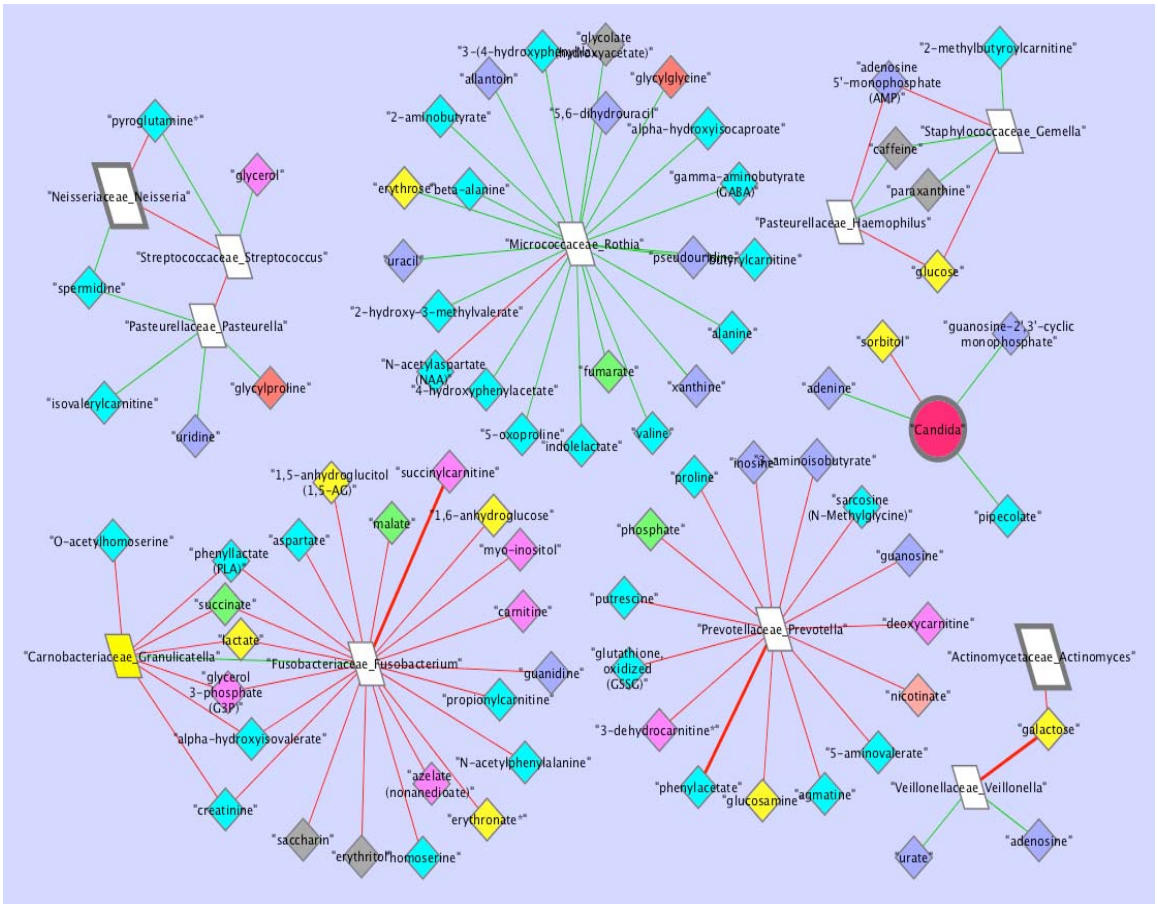
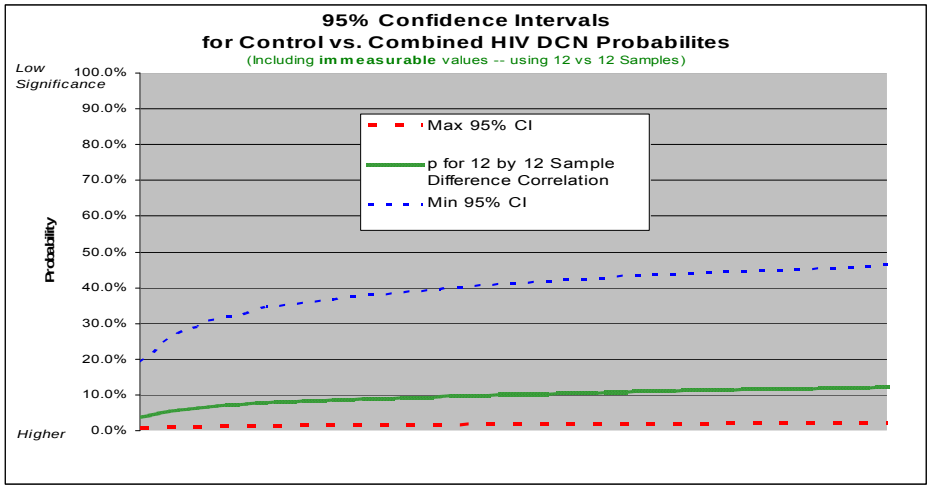


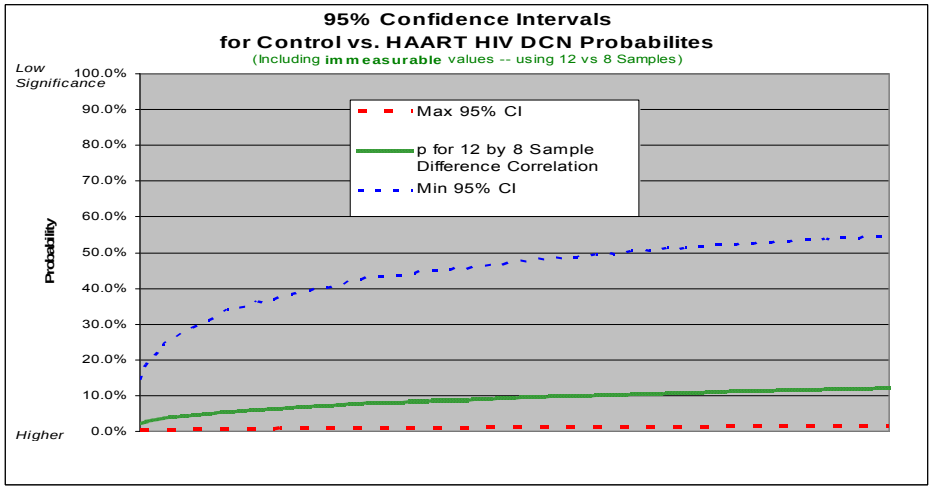

Figure 5.3.3-1 Directly Connected Bacterial or Fungal Only DCN for Control versus untreated HIV Classes with p < 17%. The legend is in Figure 5.3-1.

## *5.4 Metabiome Untreated HIV versus Control Class Analysis*

An issue with sampling is the confidence in the results. The Confidence Interval (CI) was determined for each differential correlation results per the approach outlined in Section 3.5. Since the total number of sample data values was applicable to every same

class comparison, the CI at a give probability, p, was standard for any differential correlation with that probability. The number of samples in the two classes was the other main determinate. The Control Class has 12 samples, Combined HIV, 12, HAART HIV, 8, and untreated HIV had 4 samples. Therefore the total sample values N1 and N2 for the DCN of Control versus either, Combined HIV, HAART HIV, or untreated HIV has 12 by 12, 12 by 8 or 12 by 4 samples respectively. The minimum and maximum 95% CI for each of the highly significant combinations is displayed in Figure 5.4-1. As displayed the most highly significant probabilities have the narrowest 95 %CIs and as expected the greater the sample sizes the narrower the CIs. The figure was created using all 295 Metabiome features spanning the DCNs in both Chapters 4 and 5.



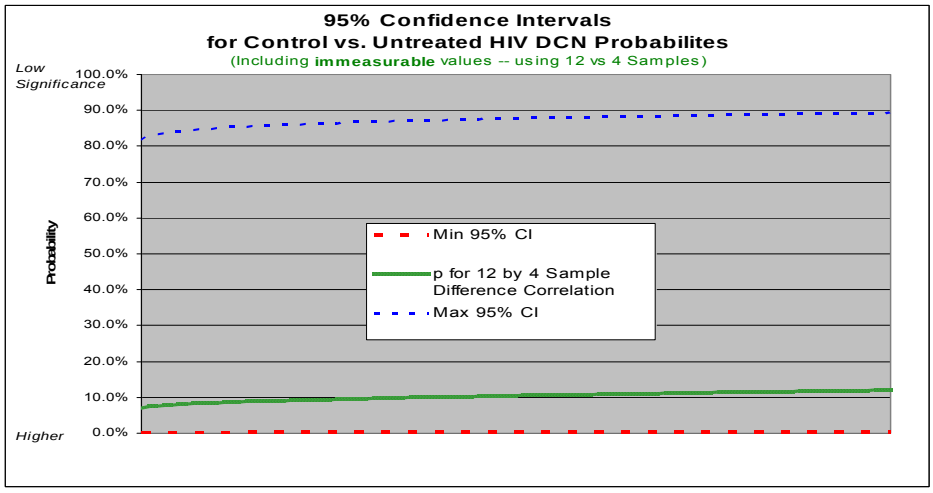




Figure 5.4-1 Comparison of 95% CI by DCN Sample Sizes

## *5.5 Summary*

The metabiome data discussed in this chapter was obtained from the same oral rinse samples presented in Chapter 4. The 24 samples had variable measurable presence or immeasurable, undetectable values, for each of the 295 metabolites. Chapter 5 included the 97 bacterial or fungal genera present across the 24 samples. The abundances was either measurable, or if the total percent abundance across all 24 samples for that genus was under 1% of the total, the genus was considered immeasurable. The analysis performed specified all immeasurable values to be included. First the correlation analysis was performed on all four 295 feature metabiome classes – the Control Class, Combined HIV, HAART HIV, and untreated HIV. The results were explained and network diagrams displayed the pertinent results. Section 5.3 addressed the Differential Correlation Networks (DCN) and their statistically significant correlations. DCNs were produced, the results presented, and DCN maps displayed for – Control vs. combined HIV, Control vs. HAART, and Control vs. untreated HIV Classes. The confidence intervals associated with the various DCNs created was discussed in Section 5.4. Chapter 6 will discuss issues with decisions involving the inclusion, or exclusion, of immeasurable feature abundances or quantities. In Chapter 7 the three oral microbiome difference correlation networks – Control vs. combined HIV, Control vs. untreated HIV, and Control vs. HAART will be reviewed for their findings, and drawing conclusions related to the hypothesis.

# Chapter 6: Data Parameter Impact to Correlation Differences Network Analysis

The purpose of this Chapter is to discuss the issues surrounding data value cutoffs and the treatment of immeasurable data values. The Control Class versus Combined HIV Class Differential Correlation Network is used to demonstrate the effect on the Oral Microbiome Study results. An example will illustrate a hypothetical scenario that shows potential loss of knowledge discovery and that there isn't one correct answer for all scenarios.

## 6.1 Discussion of Treatment of Data Values for DCN Analysis

Critical to the analysis of the data, are the assumptions made in deciding what data is valid for the study. One of the key decisions is how to treat immeasurable values. Does immeasurable indicate for metabiome features; bacteria, fungi, or metabolites, that they are truly not present -- or just below instrument reading limitations? Is the fact they are not detected discount their significance to the analysis? For bacteria, or fungal, abundances, discounting specific values below a researcher defined cutoff indicates they too are not "detected". What is the minimum percentage abundance where the researcher considers readings too prone to error? Additionally, should the taxa be evaluated at the species, or genus, or higher level? Another consideration regarding

correlation calculations is the specification of the minimally acceptable feature pairs required to calculate a correlation. These assumptions impact the Differential Correlation Network findings. The Oral Rinse study research team decided to include immeasurable data values. The Oral Rinse study was based on knowledge discovery and one cannot predict *a priori* what analysis decisions will give the most useful biological interpretation. Table 6.1-1 shows two scenarios demonstrating, with simplistic data, that including, or excluding immeasurable data values, will impact knowledge discovery by hiding potentially useful biological information.

Scenario 1 is excluding immeasurable data values and shows that with sparse datasets it is possible to throw away feature pairs that may be revealing a biological relationship between two features. Correlations do not prove causality but point out where such a relationship may exist.

The second scenario includes immeasurable sample pairs that dilute the significant sample pair information by reducing the overall correlation for the two features. If one feature has all zero values, or even has the same value for all samples, no correlation can be calculated, since the feature sample values variance equals zero. Including immeasurable values increases the number of sample pairs, thereby reducing the Standard Error, and increasing the power of the correlation probability statistic, but may simultaneously reduce significant knowledge discovery results.

Table 6.1-1 Data for two scenarios; showing the impact on feature correlation by including immeasurable values, and excluding them.

| Scenario | Sample # | Feature 1 Single Class Sample Values | Feature 2 Single Class Sample Values | Include Immeasurables ? | Resulting Correlation |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 5 | No | No correlation calculated since |
| 1 | 2 | 0 | 5 | No | all zero values are thrown out, |
| 1 | 3 | 11 | 0 | No | therefore, no pairs to correlate |
| 1 | 4 | 11 | 0 | No | |
| 1 | 5 | 11 | 0 | No | |
| 1 | 6 | 11 | 0 | No | |
| 1 | 7 | 0 | 5 | No | |
| 1 | 8 | 0 | 5 | No | |
| | | | | | |
| | | | | | |
| 2 | 1 | 5 | 0 | Yes | -0.33 |
| 2 | 2 | 5 | 0 | Yes | No significant correlation. |
| 2 | 3 | 0 | 22 | Yes | If feature 1 present => no feature 2, |
| 2 | 4 | 0 | 22 | Yes | and vice versa, MAY be underlying |
| 2 | 5 | 0 | 0 | Yes | biological significance. |
| 2 | 6 | 0 | 0 | Yes | Lost because included immeasurable |
| 2 | 7 | 0 | 0 | Yes | dilution. |
| 2 | 8 | 0 | 0 | Yes | |

The DCN analysis program allows the researcher to specify, via parameter settings, their preferred method of handling the features selection choices. One parameter, to assist in mitigating the effect of too many immeasurable data values, is to set a minimum number of required measurable data values to perform a correlation on two features. Inherently with a Spearman correlation there must be a least 3 feature-pairs to perform a correlation. Additionally, for the DCN, there must be a correlation between each feature-pair A-B in both Classes being compared, since the DCN is the measure of the significance difference between feature-pair correlations. The default setting in the DCN program is 30%. It mandates in this example being set to 30%, for a class with 20 samples that 6 measurable values must be present in one of the features in the pair for

the correlation to be performed. It is useful to remove feature pairs where both feature data values are immeasurable. The parameter does not specify which feature of the pair to apply the restriction, therefore 10% of feature 1's data values could be measurable and a non-overlapping 20% of feature 2's data values being measurable to pass the 30% minimum check. This parameter does not play a factor in the Oral Rinse study since there are only 12 samples in both the Control and Combined HIV Classes; and the significance probability calculation requires 4 or more sample pairs. With 30% of 12 = 3.6 that rounds up to 4, there was no impact to our results. However, for studies with larger sample sizes, it is another setting that can be adjusted to modify the inclusion of subset of immeasurable data values and not completely rule them out.

## 6.2 The Side-by-Side DCN Comparison with and without Immeasurable Values

To demonstrate the impact of immeasurable value inclusion versus exclusion, a comparison of one of the Oral Rinse study's DCNs; the Control Class versus the Combined HIV Class DCN was analyzed. The same samples from the Control versus Combined HIV Class were run two ways. Once excluding immeasurable data values, the other was the normal study approach, including immeasurable values in the analysis. By including the immeasurable values in the correlation determination, the total feature pair data values use to compute the feature pair correlation will equal the number of samples in the class. The affect on the correlations values is variable since the immeasurable values are generally mixed across the sample data values for feature 1 and feature 2. Figure 6.2-2 contains the resulting 'side-by-side' DCN map. The side-by-

side map is a composite of two DCNs, one using only measurable values (solid edge lines), and the other DCN including immeasurable values has dashed lines. The legend is listed separately in Figure 6.2-1, with the key legend modification being the edge line type – dashed or solid. Both DCN maps only show bacteria and fungal features, plus directly linked metabolites as in Section 5.3, where the metabolites must have a direct correlation difference with either a bacteria or fungus.

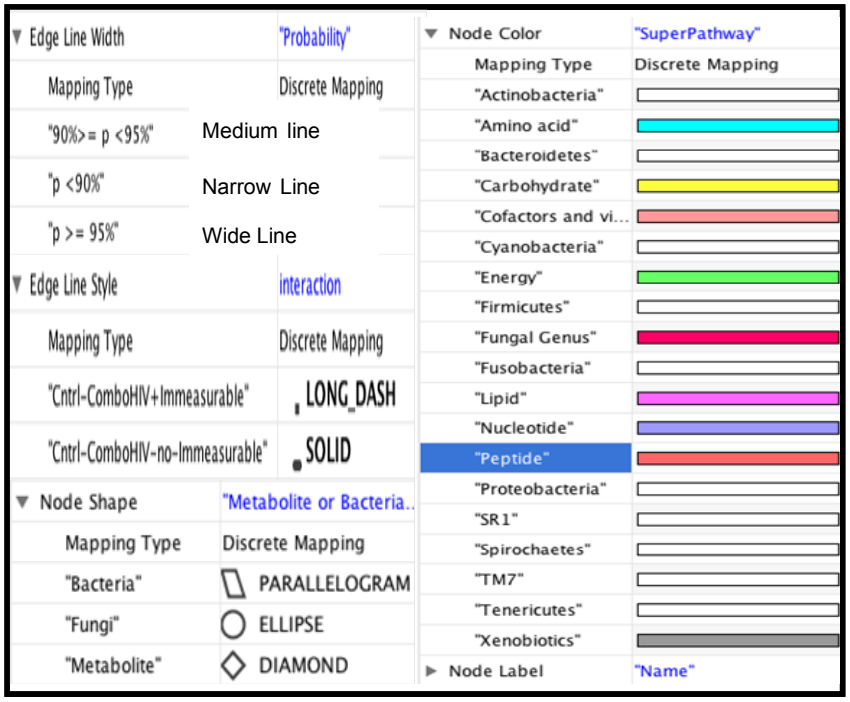



Figure 6.2-1 Legend for side-by-side DCN composite diagram

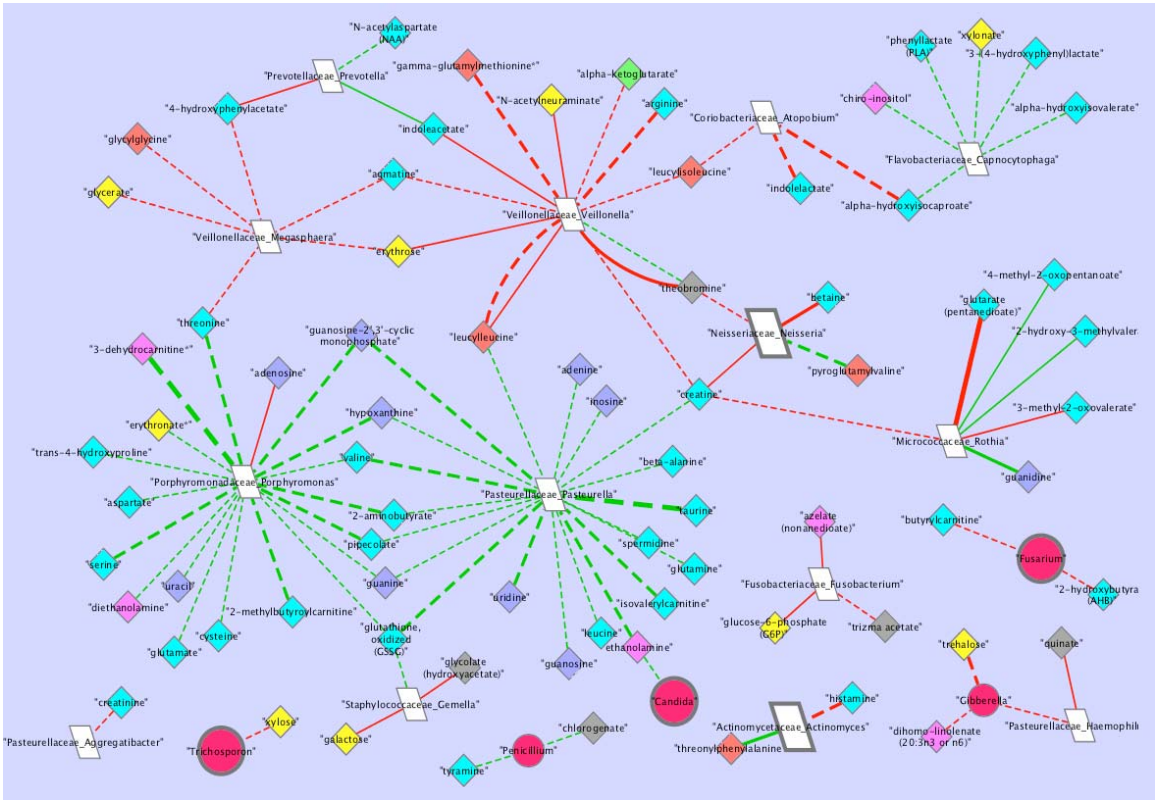



Figure 6.2-2 Side-by-Side Composite Oral Microbiome Control versus Combined HIV DCNs.

Figure 6.2-2 displays the impact of including (dashed line) or excluding immeasurable (solid line) values in the analysis. The Legend is Figure 6.2-1. Again, for clarity the diagram only displays feature pair correlations that have either a bacteria or fungae as one of the features.

## 6.3 DCN Confidence Interval Comparison

The Confidence Interval (CI) is impacted by the number of samples included in the DCN analysis. By excluding immeasurables one may be able to increase the underlying correlations and possibly the differential correlation but at the cost of the 95% CI. As shown in Figure 6.3-1 the 95% CI for the most significant 25 differential correlation results in both DCNs have different 95% CIs associated with the results. In some instances, where the total of 12 samples in each class,only 4 or 5 samples were involved in the calculation. This is another consideration in the parameter decisions involving DCN use for knowledge discovery.

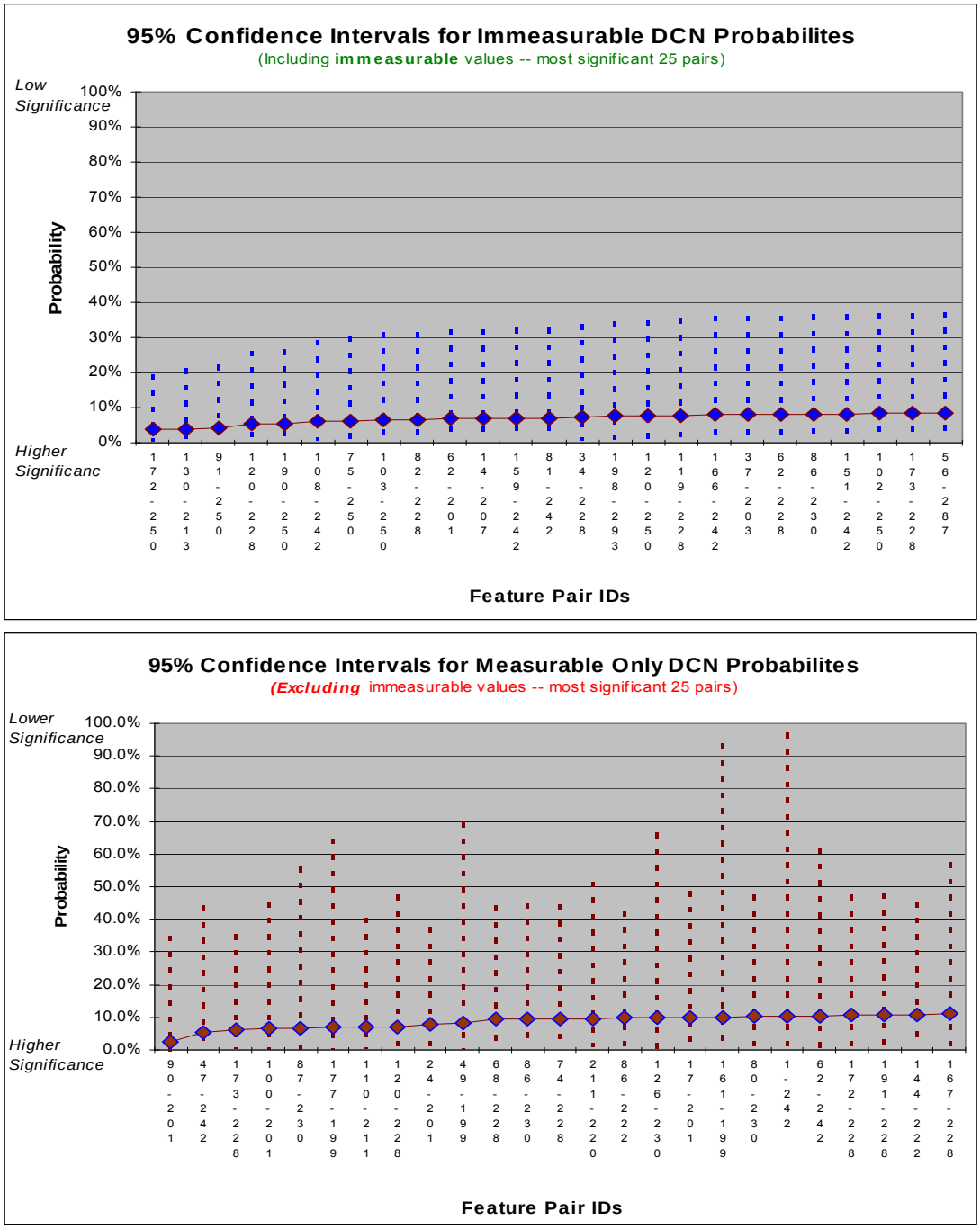



Figure 6.3-1 Comparison of DCN impact on CI by inclusion or excluding immeasurable values

## 6.4 Conclusion

It was shown using mock data that whether immeasurable data (either not present or below measure limit) is included in the DCN analysis or not unintended consequences may occur. Section 6.2 presented a side-by-side view of two DCNs for Control versus the Combination HIV Classes. One DCN included immeasurables and the other excluded them. The key criterion for specifying parameters to control the DCN analysis is for the researcher to follow their experience. Be cognizant that, along with other threshold decisions made during the clinical sample collection, and the wet-lab analysis of bacterial and fungal abundances, key decisions impacting the study findings will be propagated into the data mining phase. Section 6.3 addresses the impact on results 95% CI for the analysis in the previous section. Chapter 7 will distill the Oral Microbiome analysis results from Chapters 4, 5, and 6 and state the findings. Then possible interpretations of these findings, based on the problem statement and hypothesis will be present. Finally conclusions about the ability of DCN to improve knowledge discovery will be highlighted and a roadmap to follow-on opportunities enunciated.

# Chapter 7: Findings, Conclusions, and Implications

Chapter 7 is the culmination of the insights gained from the data and analysis from the previous chapters. First is a quick summary of the study in Section 7.1. Section 7.2 addresses the findings obtained from the analysis of the Oral Metabiome based on12 human healthy and 12 HIV samples. From these findings, section 7.3 will address the Computational Conclusions. The Oral Rinse Metabiome biological conclusions are discussed in Section 7.4. Implications relating to the Differential Correlation Network and beyond Oral Metabiome studies are discussed in Section 7.5. Section 7.6 discusses areas of future research to extend the body of knowledge gleaned from the Oral Metabiome study; and future endeavors for the Differential Correlation Network approach for knowledge discovery.

## *7.1 Summary of Systems Modeling of the Oral Metabolome*

The purpose of this study was to data mine the oral metabiome clinical results in search of knowledge discovery relating to the difference between a healthy metabiome and HIV infected metabiome. The existing set of bioinformatic metabiomic discrimination algorithms have been less than satisfactory in analyzing metabiomes and assigning underlying biological causes for disease.

Restating the hypotheses:

(H1) Examining inter-class differential feature pair correlations will differentiate healthy versus disease classes.

(H2) The application of Differential Correlation Network analysis on experimental data support knowledge discovery related to underlying biological process variation between healthy and disease states.

Within the multitude of metabiomic discrimination algorithms, included are those able to detect significant single feature (metabolite, bacteria, or fungi) differences across classes including; Meta-Stat and the Wilcoxon signed-rank test. Principal Coordinate Analysis and the similar Principal Component Analysis techniques use non-statistical multivariant analysis that cannot apply statistical significance to the results. They have issues as well with sparse data matrices so are evolving to approaches like Sparse PCA. Other multivariant statistical approaches include Libshuff and DOTUR using parsimony tests and OTU distance calculations to determine whole community statistical difference significance. Additionally, approaches similar to Supervised Vector Machines use a trained learning approach to address metabolome statistical analysis.

The analysis of the oral metabiome should play a role in adding insight into the progression HIV disease, considering the visible oral symptoms including lesions, white plaques, and sores associated with HAART. Oral metabolites are produced by both the host, and oral microbes, and they may be an effective indicator of disease status. Identifying, quantifying, and analyzing the oral metabiome gives an improved understanding of the oral cavity's biological functions and the differences that may be indicative of disease processes. Following this line of research could, lay the foundation for simpler, less expensive, and lower risk approaches to identifying HIV in patients or treating complications of HAART.

The data from the Case Western HIV Oral Metabolome clinical samples consisted of 12 Control and 12 HIV oral rinse samples. Oral rinse sample analysis identified 295 features comprised of 198 metabolites, 58 bacterial genera, and 39 fungal genera, across all 24 samples in the study.

The CW Oral Rinse sample data was categorized into four categories; Control Class with12 samples, the Highly Active Anti Retroviral Therapy (HAART) 8 HIV samples, the untreated HIV, with 4 samples, and the Combined HIV Class, merging the HAART HIV and untreated HIV, totaling 12 samples.

The study first analyzed only the198 metabolites for the 4 classes. The pipeline included the generation of intra-class correlation statistics, confidence intervals, and a correlation map. This was followed by comparing each of the 3 HIV Classes individually against the Control Class, via Differential Correlation Network analysis (DCN) and visualizing the results as DCN maps.

The same processes were repeated on the quantitative and abundance measurements for all features in the four classes for the entire Metabiome -- bacteria, fungi, and metabolites.

Finally, a review of the impact of the input parameter to include or exclude immeasurable data was discussed with oral rinse data run two different ways to demonstrate the impact.

## *7.2 Findings*

The analysis is presented to follow the flow of the earlier detailed analysis in Chapters 4,5, and 6. First each intra-class's correlation findings are presented then the inter-class

DCN. Lastly the side-by-side Control versus Combination HIV Class DCN comparison findings are presented.

### 7.2.1 Metabolite Correlation Findings

An overarching observation was for all four classes, the feature pair correlations where rho > |0.85| are by a large majority positive correlations. A very low percentage of were large negative correlations.

In the review of the Control Class metabolite correlations, the Oral Rinse Metabolome with 198 metabolites (observed in one or more of the 24 samples in this study) resulting in 283 highly correlated pairs, rho > |0.83|, as shown in Figure 4.2.1-1. Interesting, most of the 283 are high positive correlation, leaving only two negative correlations in the entire set. The majority of the correlations include metabolites from the same KEGG super-pathway – 'amino acid'; however, this would be expected since many of the same metabolite superpathway correlations are between other amino acid metabolites. The two main Control Class correlation clusters (a higher cross coupling of edges) in the map are largely amino acid metabolites. In these clusters are the highest rhos, 0.99, value correlations as well.

For the Combined HIV Class Correlation analysis, the Oral Rinse metabolites resulted in a similar number for high correlations as the Control Class resulting in; 257 highly correlated pairs, rho > |0.83|, as shown in Figure 4.2.2-1. Again, most of the 257 high correlations are positively correlated with only 3 negative correlations in the set. The majority of the single metabolite multiple correlations include metabolites from the two

KEGG super-pathways, ‘amino acid’ and ‘nucleotide’. Some of the very highest correlations are the pairings of inosine and tyrosine, lysine and uridine, purtrescine, and cadaverine, xylose and fucose, uracil and hypoxanthine, glutamate and alanine, valine and leucine, valine and serine, caproate and isocaporate, plus throphyline and paraxanthine. Many of these are obvious interconnected biological pathway components that reinforce that the underlying DCN concept appears sound.

The 12 HIV samples initially were run as one combined class, but upon inspection supporting metadata showed they were really two important subclasses -- four untreated HIV and eight Highly Active Anti-Retroviral Therapy (HAART) samples. They were split out and included as two new classes.

Finding from the HAART HIV Class Correlation analysis performed for the Oral Rinse metabolites resulted in several different results than both the Control and Combined HIV correlation results. Of the 302 high correlations, 29 were negatively correlated and 272 positively correlated pairs, with the rho > |0.87|. The cutoff is higher as well to reduce the number of results and artifacts due to the smaller number of samples and using ranks in place of actual data values. As shown in Figure 4.2.3-1, there is only one major cluster of correlations with three lesser clusters and then many extended branches. The mix of KEGG Superpathway metabolites in this major cluster is more diverse compared to the Control and Combined HIV Classes. The correlation intensity (one feature having many correlations) for the positive correlations in the only major cluster, is very high for a majority of the features involved. However, the HAART HIV Class shows a couple metabolites that have a negative correlation with many other features. Stachydrine, Proline, and Glycine are all negatively correlated (with 8 ,7, and 4 correlations respectively) with other disjoint sets of metabolites. Theophyline, is used to prevent and

treat wheezing, shortness of breath, and difficulty breathing caused by asthma, chronic bronchitis, emphysema, and other lung diseases, is both highly positive and negatively correlated in the HAART Class.

The Untreated HIV has an extremely large number 1,157 of very high correlated values were rho = |1.0|, driven by the very small sample size of 4. With the related lessened confidence, increased Confidence Interval length, there are still items of note. First is the six major clusters formed with no cross correlations between them; and there are several KEGG superpathway metabolites in each cluster. Of the 86 negative correlations, Glutarate and theobromine stand out with multiple negative correlation relationships.

### 7.2.2 Metabolite Differential Correlation Network Findings

The four metabolite class correlation evaluations previously allow for the creation of the Differential Correlation Networks (DCN) in Section 4.3. Interestingly, both the untreated HIV versus Control DCN and the Control versus HAART DCN analysis have distinctly different results.

The DCN relationship between the Control Class and the 3 different HIV classes each expresses significant differences that may indicate important biology process variations. However, the correlation and DCN diagrams do not indicate causality, only correlations that need to be matched with biologic repositories for conformation and further insights. The combined information may discern if there is causality, and also if useful biological knowledge discovery insights lie within the network maps.

Interpreting the first DCN analysis of the Control Class versus the Combined HIV Class resulted in the DCN map in Figure 4.3.1-1. It contains a fairly even mix of positive (140) versus negative (103) correlation differences. There are no very highly significant probability differences in the entire DCN map. Azelate, Malitol (a sucrose sugar substitute), and Glycine (amino acid) have several negative correlation differences with other non-overlapping metabolites. However, Azelate and Malitol also negatively correlate with each other. There are both strong negative and positive correlation difference involving alpha-hydroxyisocaproate (oxytocin), while isocaproate and caproate are strongly negative difference correlated. Cotinine, an indicator that someone has been smoking, and indolelactate, involved in Tryptophan metabolism, are both positively difference correlated with many other metabolites.

The DCN between the Control and HAART HIV in Section 4.3.2 shows the most correlation differences with at total of 323 and over a 2:1 ratio of negative correlation difference (216) versus positive correlation differences (107). The negatives are more highly significant as well. The metabolites with multiple significantly different correlations to several other metabolites in the HAART HIV vs. Control DCN are; Glycine (amino acid), Proline (amino acid), Stachydrine (N, N-dimethyl proline), and Agmatine (decarboxylation product of the amino acid arginine), all are Amino Acid superpathway metabolites. Azelate, a cofactor and vitamin superpathway metabolite, has 22 negative correlation differences with other metabolites. Urate, with 15 positive correlation differences, is the lone metabolite on the positive highly differentiated side, but Serine as two high significant correlation differences with phosphate and 2-hydroxybutrate The untreated HIV versus Control DCN findings are comprised of eleven independent cluster “islands”. Generally the difference between Control and Untreated HIV is somewhat

significant positive correlation change. Out of 453, are 368 are positive correlation differences. The map islands each are a mix of superpathway metabolites, including higher cluster intensity than the previously reviewed DCNs. In the largest 31 metabolite cluster, there are only two negatively correlated differences to all of the other cluster metabolites. The first is N-Acetylaspartate (NAA), is a derivative of aspartic acid is the second-most-concentrated molecule in the brain after the amino acid glutamate, plus a reliable diagnostic molecule for treating patients with disease {Premkumar}. The second is glutarate that only have reduced correlation differences in untreated HIV. N-Acetylaspartate with 28 negative significant correlation changes, p<16%, has three that are more significant, p<10%, with glycylglycine, glycolate, and with fumarate. Glutarate with 27 decreased correlation differences also is more significantly decreased against glycolate compared with the Control Class. Theobromine is the center of a small island where it is negative correlation different versus Control with its 11 paired metabolites, with one significant, p<10%. Sorbitol is similar with 2 strong differences out of 5 negative correlation differences.

This research demonstrates a counter-intuitive result regarding the significant differences between classes. Separating untreated HIV from HAART HIV indicates they have a very distinct correlation difference networks. The Control Class versus the untreated HIV DCN has many more less significant positive correlation differences with the Control Class versus HAART HIV DCN. The HAART HIV DCN has many higher significance negative correlations than the Combined HIV DCN or the untreated HIV DCN. Positive difference indicates increase feature pair correlation value in untreated HIV vs. Control and negative difference indicates much lower rho value within the HAART Class versus the Control Class.

### 7.2.3 Metabolome Correlation Findings

An overview of the Oral Rinse bacterial and fungal genus abundances shows, on visual inspection of Figure 5.1-1, that there does not appear to be significant bacterial abundance differences between Control and HIV. At the lower abundances are certain species in the Control Class and others only in HIV Classes.

Both bacterial and fungal genus abundance must greater or equal to 1% of total bacterial or fungal abundance; otherwise they were not included in the study.

The Control versus HIV fungal genus abundance comparison in Figure 5.1-2 displays one very interesting difference of highly populated Control fungae, Candida and Pischia. In the HIV group the Pischia fungi is absent while the Candida abundance is even higher than the Control.

The metabiome correlation analysis and DCN calculations included all 198 metabolites, 58 bacterial genera, and 39 fungal genera. The resulting Correlation diagrams and DCN maps are filtered to only show all Fungi and Bacteria making the rho or probability cutoffs, plus to simplify the maps, only metabolites displayed are those directly correlated to a bacteria or fungae.

Some of the names and traits for frequently referenced fungae in the correlation and DCN results are:

Actinomycetaceae Actinomyces is an agent in chronic periodontal inflammation in man;

Candida albicans, which can cause infections (called candidiasis or thrush) in humans especially in immunocompromised patients;

Epicoccum – causing allergen, irritant, hypersensitivity pneumonitis, and dermatitis in humans;

Fusobacteriaceae Fusobacterium contributes to several human diseases, including periodontal diseases, and should always be treated as a pathogen;

Some of the names and traits for frequently referenced bacteria in the correlation and DCN results are:

Neisseriaceae Neisseria is a non-pathogenic species;

Prevotellaceae Prevotella anaerobic pathogen involved in periodontal infections, including periodontitis and often found in acute necrotizing ulcerative gingivitis;

Porphyromonadaceae Porphyromonas is a family that includes several pathogens of vertebrates, most notably H. influenzae. Other Pasteurellaceae cause gingivitis and chancroid in humans;

Veillonellaceae Veillonella is a human pathogen {Health}

The intra-Control Class correlation findings for Section 5.2.1 shows one fungal bacterial positive correlation between Flavobacteriaceae Capnocytophaga and Cladosporum. Note, all the correlations are in addition to the metabolite-metabolite correlations in Section 4.2. There is one positively correlated group with 3 fungi directly related – Corynespora, Saccharomyces, and Gibberella. Corynespora, Saccharomyce, and Fusarium are all positively linked to the metabolite naringenin (is a flavonoid that is considered to have a bioactive effect on human health as antioxidant, free radical scavenger, anti-inflammatory, carbohydrate metabolism promoter, and immune system modulator. The antiviral effects of naringenin are currently under clinical investigation) {Lyu, 2005}. Veillonellaceae Veillonella is highly correlated to the metabolite gamma-glutamylmethionine. The bacteria Porphyromonadaceae Porphyromonas is negatively correlated with hypoxanthine.

The findings for the Combination HIV Class correlations include a significant correlation,

rho = 0.99 appearing between Flavobacteriaceae Capnocytophaga and the metabolite N-acetylleucine. The fungi Epicoccum and Trichosporon are positively correlated, as is Candida with 5-aminovalerate. Two bacteria Pasteurellaceae Pasteurella and Incertae Sedis are correlated, as is the bacteria Campylobacteraceae Campylobacter with the metabolite malitol.

The correlations in the HAART HIV Class are more numerous than the prior two findings following the earlier trend with the HAART HIV metabolites. The fungi Epicoccum and Trichosporon are again positively correlated as would be expected from viewing the Combined HIV findings. Candida is positively correlated with ethanolamine and this finding differs from the Combined HIV correlation. Additionally, the fungus Penicillium is negatively correlated with 4 metabolites – inosine, uridine, leucylleucine, and cytodine. Pasteurellaceae Pasteurella and Incertae Sedis are still correlated, as they were in the Combined HIV correlation map.

Correlations from the untreated HIV correlation include, Carnobacteriaceae Granulicatella and Fusobacteriaceae Fusobacterium being positively correlated in the untreated HIV correlation analysis, while both are negatively correlated with 24 metabolites. There are two interesting findings, one positive correlation differences including the bacteria Streptococcaceae and Pastuerella, and another a negative correlation between Streptococcaceae and Neisseriaceae. Prevotellaceae Prevotella negatively correlates with 20 metabolites independently. Micrococcaceae_Rothia with rho > 0.95 is positively correlated with 29 metabolites and negatively correlated with two. Candida, rho > 0.95, is positively correlated the metabolites -- xylose, pipecolate, adenine, and guanosine-2',3'-cyclic monophosphate, but negatively correlated with sorbitol. Other bacteria are paired singly with a metabolite in both positively and

negatively correlated combinations.

### 7.2.4 Metabiome Differential Correlation Network Findings

The metabiome difference correlation network analysis involves all possible feature correlations, including pairs that were not highly correlated. If the probability of a feature pair AB changes from low, or no correlation, to a very high correlation between the classes and the sample size is large enough to support significant probability, it will appear in a DCN map even though it may not appear in one of the corresponding input correlation maps used to create the DCN because the feature pair correlation was below the defined cutoff threshold.

The correlation differences between the Control Class and the Combined HIV Class displayed in Figure 5.3.1-1 are extensive. The differences entail many bacteria and fungae. Some notable findings include Pasteurellaceae pasteurella being positively correlated with 22 metabolites and the correlation differential with taurine being very significant. Additionally, Pasteurellaceae Pasteurella is negatively correlated to the fungae Epicoccum, positively differentially correlated with ethanolamine (included in the count above). Additionally, Candida is negatively correlation differentiated with ethanolamine.

The HAART HIV vs. Control finding related to Figure 5.3.2-1, presented the some interesting positive correlation differences, many significant, $p<10\%$, especially with the bacteria Porphyromonadaceae porphyromonas and 20 metabolites. Several HAART HIV Class bacteria have a significant statistical negative correlation changes compared to the Control Class. The bacterial metabolite pairs are; Prevotellaceae prevotella and phenylpropionate, Gibberella and trehalose, Veillonellaceae veillonella and gamma-

glutamylmethionine, Actinomycetaceae Actinomyces and azelate, Neisseriaceae neisseria and gamma-aminobutyrate (GABA). However, Neisseriaceae neisseria has a significantly positively correlation difference with xylonite. Pasteurellaceae haemophilus and the fungae Gibberella are the only directly correlation different (positively) in the HAART HIV DCN.

Pasteurellaceae Pasteurella is positively difference correlated with 7 metabolites. Of interest, taurine is known to protect against glutamate excitotoxicity and many other very positive medical effects and glutathione is a tripeptide that contains an unusual peptide linkage between the amine group of cysteine and the carboxyl group of the glutamate side-chain with oxidized (GSSG) playing a very significant metabolic role. As also demonstrated in the Control versus Combined HIV DCN, the HAART DCN, Figure 7.2.4-1, Epicoccum and Candida are directly negatively correlated, corroborating results of Vaz et. al. showing that Epicoccum inhibits Candida (Vaz, Mota et al. 2009). Once the Pasteurella outlier is removed there is no presence of Pasteurella in the HAART sample.

| Control HAART Feature Pairs | Control Correla-tions (rho) | HAART Correla-tions (rho) | Differential Correla-tion Probabil-ities (p) |
|---|---|---|---|
| ethanolamine-Pasteurellaceae_Pasteurella | -0.69 | 0.58 | 12% |
| Pasteurellaceae_Pasteurella-Candida | 0.16 | 0.41 | 42% |
| ethanolamine-Candida | -0.15 | 0.88 | 12% |
| Epicoccum-Candida | 0.32 | -0.79 | 14% |

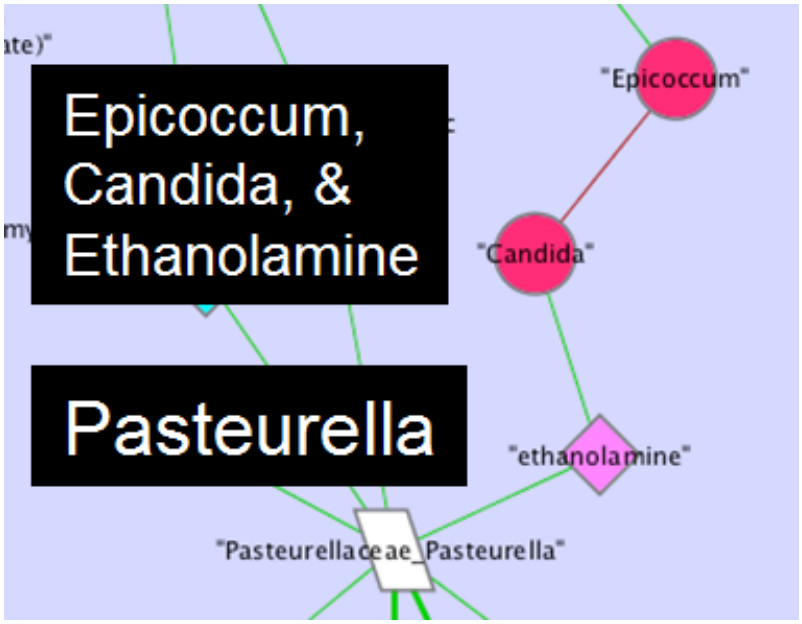



Figure 7.2.4-1 Interesting HAART observation of metabolite, fungus, and bacteria relationship

Additional, we observed that Pasteurella is a major hub that is differentially correlated between the Control and HAART HIV class {Gillevet, 2011}. Candida is differentially correlated with ethanolamine (shown to be involved in the creation of biofilms) is significant in that it may be a place to pursue in the determination of the cause of thrush. Additionally, Pasteurellaceae is differentially correlated to indolelacrate. There is a difference edge between this node and Epicoccum fungi through several metabolites and leading to an indirect linkage to Candida. Another observation is Neisseria is differentially correlated with a number of metabolites but significantly correlated with xylonate.

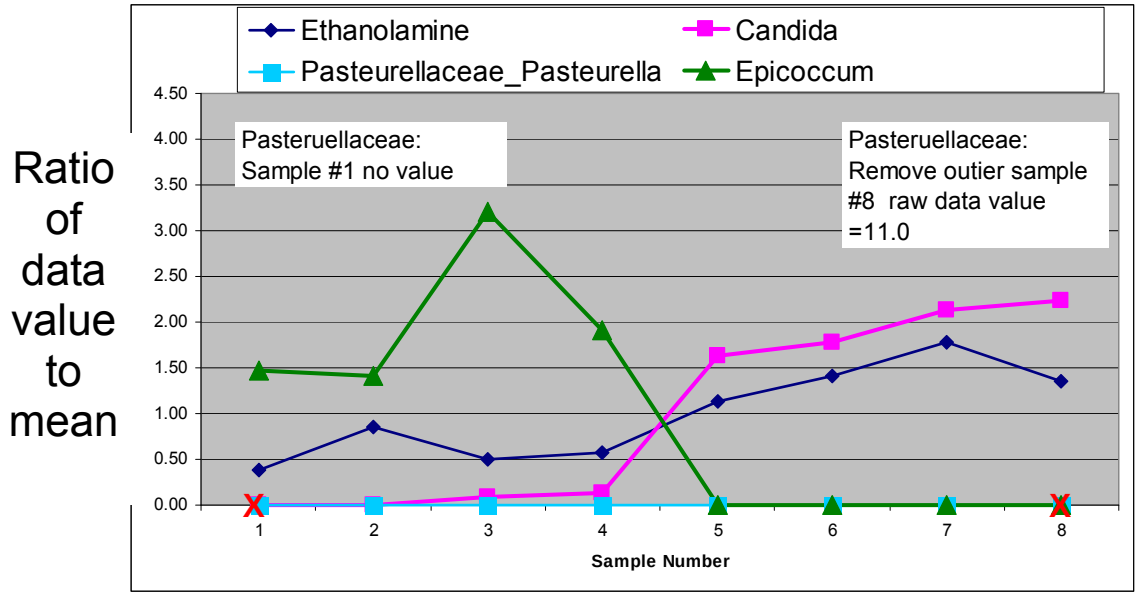



Figure 7.2.4-2 Four HAART metabiome feature data value ratios.

(Compared to their specific HAART mean)

The findings related to untreated HIV and Control DCN reflect the untreated HIV Class Correlation map, with Fusobacteriaceae fusobacterium being negatively correlated in the untreated HIV difference correlation network analysis with 21 metabolites. There is negatively correlation difference between Prevotellaceae prevotella and 15 independent metabolites. Also Candida has a positive difference correlation with adenine, sorbitol, and pipecolate (an enzyme in lysine degradation). There appears to be interesting negative difference correlation islands that include Streptococcaceae, Pastuerella, Streptococcaceae and Neisseriaceae that is also embedded with several Xenobiotic metabolites. Both Fusobacteriaceae fusobacterium and Prevotellaceae prevotella are the center of negative difference correlations with 21 and 15 non-overlapping metabolites respectively.

The oral rinse Metabiome feature correlation profile for the control class versus the three disease classes, demonstrated there were a convincing number of significantly different correlations. These variations between class feature pairs may point to HIV having an underlying biological cause or effect impact on the oral metabiome.

### 7.2.5 Metabiome Differential Correlation Network Immeasurable Findings

The review of the impact of including immeasurable data in the correlation analysis in Chapter 6 indicating significant impact on the resulting networks. The exclusion of immeasurable values reduced the number of DCN feature pairs. There is no correct answer for the DCN knowledge discovery technique and cases were made for both

inclusion and exclusion. If time warrants and the researcher is not clear, running the analysis both ways is a good alternative to ensure all avenues have been pursued. A couple findings of note, one is Porphyromonadaceae porphyromonas and Pasteurella both have many positive correlation differences, 18 and 19, with other metabolites for immeasurable DCN map with nothing corresponding in the other DCN. Secondly, the two Veillonellaceae genera have more negative difference correlations again with immeasurable data included, but the measurable DCN has fewer with higher statistical probability.

## *7.3 Computational Conclusion*

The application of the Differential Correlation Network analysis to the Oral Metabiome has highlighted interesting correlations between the various components – bacteria, fungae, and metabolites. Although the correlations, and their differences between classes, cannot assign causality, one can quickly pinpoint the functional relationships when other biological information is incorporated, either via prior experience, or bioinformatic repositories, to highlight in knowledge discovery. The ability to visualize the significant changes in relationships, linking key attributes, and limiting results to within statistically significant limits, leading to questions about the underlying biological function differences shall assist the researcher in new hypothesis generation.

The DCN methodology brings a different perspective to mining metabiome data in search of underlying biological clues that distinguish two classes. It focuses on every possible metabiome relationship between two classes, and uses a statistical approach to present findings. The usage of network maps with rich attribute visual discriminators

researchers are able to succinctly review hundreds of interactions in one network diagram.

The DCN technique brings to the table a more biological process based approach throughout the metabiome aiding knowledge discovery. It is a straightforward statistically approach to attacking the metabiome and being able to answer the hypotheses present in this paper:

(H1) Examining inter-class differential feature pair correlations will differentiate healthy versus disease classes.

(H2) The application of Differential Correlation Network analysis on experimental data support knowledge discovery related to underlying biological process variation between healthy and disease states.

It cannot be stated with certainty the DCN approach has determined underlying biological function differences, but it has highlighted many areas for further studies. The null hypothesis states; there will be no significance difference in the correlations for a pair of features in healthy class samples compared to a disease class samples. We have demonstrated statistically, with DCN analysis there are differences, but more research is needed to determine if it leads to insightful knowledge discovery.

## *7.4 Biological Conclusions*

The objective of the Oral Rinse study was knowledge discovery regarding possible biomarkers for detecting of HIV in humans and investigate the underlying process involved in HIV disease. One possible set of biomarkers to be determined by this study is the changes involving Phenylalanine and Tryptophan in untreated HIV compared to the control group {M. A. Ghannoum, 2011}. The question to be answered is -- are there

metabolic pathways that are being altered in HIV patients versus healthy individuals? Do these changes reflect the change? Correlations do not indicate causality. Frequently there are visible oral cavity changes in HIV patient's, to include the occurrence of lesions, sores, and Thrush. We found a relationship between ethanolamine and Candida that may indicate some metabolic change in the HAART HIV oral microbiome. Ethanolamine has been studied as a quorum sensing metabolite that may implicate it with the appearance of thrush {Straight, 2009}. This study presented several potential biological functions and processes, based on their corresponding metabiome features to pursue. Are these linked to a metabolic change related to disease? This study was to scrutinize the oral rinse metabiome data in search of knowledge discovery relating healthy to HIV infected individuals. The Oral Rinse study researchers are currently reviewing some of the findings. Does the DCN approach improve researcher's ability to determine the underlying biological causes of disease compared with other bioinformatic algorithms? The findings do lead to some possible explanations and potential markers for the HIV disease. The sample sizes were small, but the findings need to be vetted, and new hypothesis generated to obtain a more in depth understanding, and repeat if necessary analogous to peeling the onion, leading to oral biomarkers for the disease. The metabiome feature relationship correlations within each of the classes are very distinct. The study has presented a multitude of potentially biologically significant oral metabolic differences between healthy and HIV infected samples.

One observations indicates a possible significant finding that ethanolamine is negatively correlated with Candida in the Control but positively correlated in HAART. These are weak correlations but they exist, and may be biologically important. Interestingly, there is a similar relationship with Pasteurella. Is there a causal effect here or just a response to

ethanolamine, or is ethanolamine a secondary effect just being excreted by these organisms?

As stated in the findings, the Control versus Combined HIV DCN and the Control vs. HAART DCN, Figure 7.2.4-1, Epicoccum and Candida are inversely correlated, supporting the findings of Vaz (Vaz, Mota et al. 2009) that the fungi Epicoccum inhibits the Candida fungi. The negative relationship is clearly delineated in Figure 7.2.4-2. This is another area for knowledge discovery regarding the underlying biological cause of Thrush in HAART patients.

The results of the Differential Correlation Network on Oral Rinse Control and HIV samples indicate there are significant differences. To generate follow-on knowledge discovery hypothesis would require significant variations to be discovered in the analysis. Were we able to find noteworthy findings? Do these findings raise biologically important questions that assist the researcher is formulating more focused follow-up hypothesis? These questions have already been partially answered. Follow-up experiments to prepare for new grants are coming from these results. The DCN results were used to support the manuscript Metabolomics Reveals Differential Levels of Oral Metabolites in HIV-Infected Patients submitted for publication to Omics in April 2011 {M. Ghannoum, 2011}.

## *7.5 Differential Correlation Network Implications*

In performing the analysis underlying this study, it became apparent there are areas for improvement and additional features that would make the DCN approach better support the researcher.

One area of improvement is the simplification of the analysis pipeline. Creating

automated pipelines for common analysis steps to streamline the effort as well as reduce human error in processing, where several workbenches such as Galaxy {Giardine, 2005} or QIIME {Caporaso} to supply a framework for pipelines.

An area to address is the treatment of all immeasurable values for a feature in one class when it appears in the second class. With no measurable values the feature will not be correlated to other features in the class and therefore will not be part of a DCN between the classes. This presents the dilemma observed with Pichia and Candida {M. Ghannoum, 2011}, where Pichia is not present in the HIV abundance tables, but is present in the Control data. There may be biological significance if Pichia were correlated with other features even when absent in one class.

Adding a repository for results and pertinent experimental data would enhance the DCN capability, such as with a tool like Drupal. This information could be combined with results from other studies to either improve the overall finding significance by increasing the sample counts, or extending the analysis beyond the initial metabiome to cross metabiomes. Results captured from continuing, or extending, the System Modeling of the Oral Metabolome initiative, can be combined to generate an organ specific, temporal, extracellular function-metabolite healthy and disease biological pathway mappings. Linking metabiomes to see the impact of one on the other could also present interesting possibilities for experimentation. Transforming the individual experimental results into a repository that many experiments can contribute results would also improve researcher communication, by sharing their information in a common way.

These significant Differential Correlation Network maps only show that two features have a different relationship between the two classes, but do not demonstrate cause and effect, or even imply a biological significance. The technique shifts the data and shows

areas that could be of interest and deserve a more detailed review. To improve attribute availability, automatically, or semi-automatically link significant findings to other repositories that would confer more knowledge about the underlying biological feature model. One approach would be to link the Cytoscape diagrams, via plug-ins, to other biological repositories would enhance the ease of analysis from the resulting diagrams. One example is, linking the metabolite data to the KEGG inter-cellular processes to validate the resulting correlation perturbation significance. These are some approaches that would enhance the DCN core that is being used for knowledge discovery.

## *7.6 Future Research for Oral Metabiome*

Results from this study provided critical information that will form the basis of follow-on Oral Metabolome hypothesis-driven experiments. Studies being considered for follow-on DCN analysis include Oral Metabolome studies factoring in individual variability, such as, the effect of drugs, diet, or locality.

One area of continued research is to apply the DCN approach beyond the oral microbiome to other environments, involving humans or other host species. The oral microbiome has been implicated in downstream intestinal inflammatory bowel disease (Singhal, Dian et al.). Studies have attempted to understand the interactions within the oral microbiome and its impact on Candida (Thein, Samaranayake et al. 2006). The entire human metabiome plays a significant role in healthy individuals (Ghannoum, Jurevic et al.) (Dewhirst, Chen et al.).

Another is to apply DCN to longitudinal studies where the timeframes could be longer, e.g. days, to look for variation in the Metabiome, or based on very short timeframes, e.g. minutes, to investigate whether there is any kind of oscillations or other phenomenon

occurring in the oral metabolome. Specifically intended is a longitudinal study of the Oral Metabolome control and HIV classes used in this research. This follow-on longitudinal study should be based on short time durations, measured in hours or days. This would give a better perspective on possible temporal relationships of the oral metabiome features for each class that a single snapshot may be missing.

Let the DCN technique delve deeper into various environmental metabiomes and enlighten us to the biological perturbations that define disease.

# References

# References


(HMP), H. M. P. (2011). "Human Microbiome Project (HMP)." http://nihroadmap.nih.gov/hmp/.

Baati, H., S. Guermazi, et al. "Microbial community of communities based on 16S rRNA analysis." Can J Microbiol **56**(1): 44-51.

Bik, E. M., C. D. Long, et al. "Bacterial diversity in the oral cavity of 10 healthy individuals." Isme J **4**(8): 962-74.

Camilolab.slu (2011). "Ecological Ordination."

Caporaso, J. G., J. Kuczynski, et al. "QIIME allows analysis of high-throughput community sequencing data." Nat Methods **7**(5): 335-6.

Chan, E. K., H. C. Rowe, et al. "The complex genetic architecture of the metabolome." PLoS Genet **6**(11): e1001198.

Costanzo, M., A. Baryshnikova, et al. "The genetic landscape of a cell." Science **327**(5964): 425-31.

de Tayrac, M., S. Le, et al. (2009). "Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach." BMC Genomics **10**: 32.

Dewhirst, F. E., T. Chen, et al. "The human oral microbiome." J Bacteriol **192**(19): 5002-17.

Dhanoa, A. and Q. K. Fatt (2009). "Non-typhoidal Salmonella bacteraemia: epidemiology, clinical characteristics and its' association with severe immunosuppression." Ann Clin Microbiol Antimicrob **8**: 15.

Diggle, M. A. and S. C. Clarke (2004). "Pyrosequencing: sequence typing at the speed of light." Mol Biotechnol **28**(2): 129-37.

Fakhrai-Rad, H., N. Pourmand, et al. (2002). "Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms." Hum Mutat **19**(5): 479-85.

Fischbach, M. A. and N. J. Krogan "The next frontier of systems biology: higher-order and interspecies interactions." Genome Biol **11**(5): 208.

Franca, L. T., E. Carrilho, et al. (2002). "A review of DNA sequencing techniques." Q Rev Biophys **35**(2): 169-200.

Ghannoum, M. A., R. J. Jurevic, et al. (2010). "Characterization of the oral fungal microbiome (mycobiome) in healthy individuals." PLoS Pathog **6**(1): e1000713.

Giardine, B., C. Riemer, et al. (2005). "Galaxy: a platform for interactive large-scale genome analysis." Genome Res **15**(10): 1451-5.

Gillevet, P. (2011). "Internal Discussions Bioinformatics GMU."

Gillevet, P., M. Sikaroodi, et al. "Quantitative assessment of the human gut microbiome using multitag pyrosequencing." Chem Biodivers **7**(5): 1065-75.
Hamady, M., C. Lozupone, et al. "Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data." Isme J **4**(1): 17-27.
Hsiao, W. W. and C. M. Fraser-Liggett (2009). "Human Microbiome Project--paving the way to a better understanding of ourselves and our microbes." Drug Discov Today **14**(7-8): 331-3.
Hui Zou!, T. H., Robert Tibshirani‡ (2004). "Sparse Principal Component Analysis."
Huson, D. H., D. C. Richter, et al. (2009). "Methods for comparative metagenomics." BMC Bioinformatics **10**(Suppl 1): S12.
Kell, D. B. (2005). "Metabolomics, machine learning and modelling: towards an understanding of the language of cells." Biochem Soc Trans **33**(Pt 3): 520-4.
Kohl, M., S. Wiese, et al. "Cytoscape: software for visualization and analysis of biological networks." Methods Mol Biol **696**: 291-303.
Kurokawa, K., T. Itoh, et al. (2007). "Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes." DNA Res **14**(4): 169-81.
Lau, S. C. and W. T. Liu (2007). "Recent advances in molecular techniques for the detection of phylogenetic markers and functional genes in microbial communities." FEMS Microbiol Lett **275**(2): 183-90.
Lazarevic, V., K. Whiteson, et al. "Study of inter- and intra-individual variations in the salivary microbiota." BMC Genomics **11**: 523.
Liu, W. T., T. L. Marsh, et al. (1997). "Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA." Appl Environ Microbiol **63**(11): 4516-22.
Loscalzo, J., I. Kohane, et al. (2007). "Human disease classification in the postgenomic era: a complex systems approach to human pathobiology." Mol Syst Biol **3**: 124.
M. Ghannoum, P. G., M. Retuerto, P, Mukherjee, R. Jurevic; R. Brown (2011). "Manuscript in progress."
Mahmoud A. Ghannoum1*, P. K. M., Richard R Jurevic2, Mauricio Retuerto1, Robert E. Brown3, Masoumeh Sikaroodi3, Jennifer Webster-Cyriaque4, and Patrick M. Gillevet3 (2011). "Metabolomics Reveals Differential Levels of Oral Metabolites in HIV-Infected Patients." Manuscript in preparation.
Maidak, B. L., J. R. Cole, et al. (2001). "The RDP-II (Ribosomal Database Project)." Nucleic Acids Res **29**(1): 173-4.
Major, H. J., R. Williams, et al. (2006). "A metabonomic analysis of plasma from Zucker rat strains using gas chromatography/mass spectrometry and pattern recognition." Rapid Commun Mass Spectrom **20**(22): 3295-302.
Martin, F. P., M. E. Dumas, et al. (2007). "A top-down systems biology view of microbiome-mammalian metabolic interactions in a mouse model." Mol Syst Biol **3**: 112.
Morgenthal, K., W. Weckwerth, et al. (2006). "Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation." Biosystems

**83**(2-3): 108-17.
Naqvi, A., H. Rangwala, et al. "Analysis of multitag pyrosequence data from human cervical lavage samples." Chem Biodivers **7**(5): 1076-85.
Nasidze, I., J. Li, et al. (2009). "Global diversity in the human salivary microbiome." Genome Res **19**(4): 636-43.
Nossa, C. W., W. E. Oberdorf, et al. "Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome." World J Gastroenterol **16**(33): 4135-44.
Nyamundanda, G., L. Brennan, et al. "Probabilistic Principal Component Analysis for Metabolomic Data." BMC Bioinformatics **11**(1): 571.
Oberg, A. L. and O. Vitek (2009). "Statistical design of quantitative mass spectrometry-based proteomic experiments." J Proteome Res **8**(5): 2144-56.
Peterson, J., S. Garges, et al. (2009). "The NIH Human Microbiome Project." Genome Res **19**(12): 2317-23.
Rattray, R. M., S. Perumbakkam, et al. "Microbiomic comparison of the intestine of the earthworm Eisenia fetida fed ergovaline." Curr Microbiol **60**(3): 229-35.
Roessner, U., C. Wagner, et al. (2000). "Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry." Plant J **23**(1): 131-42.
Ronaghi, M. (2001). "Pyrosequencing sheds light on DNA sequencing." Genome Res **11**(1): 3-11.
Rosner, B. (2006). "Fundamentals of Biostatistics Sixth Edition." Thompson Publisher: 498-488 Fisher Z.
Sahota, G. and G. D. Stormo "Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes." Bioinformatics **26**(21): 2672-7.
Schloss, P. D. and J. Handelsman (2005). "Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness." Appl Environ Microbiol **71**(3): 1501-6.
Schloss, P. D. and J. Handelsman (2006). "Introducing TreeClimber, a test to compare microbial community structures." Appl Environ Microbiol **72**(4): 2379-84.
Schloss, P. D., B. R. Larget, et al. (2004). "Integration of microbial ecology and statistics: a test to compare gene libraries." Appl Environ Microbiol **70**(9): 5485-92.
Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-504.
Shiboski, C. H. (2002). "HIV-related oral disease epidemiology among women: year 2000 update." Oral Dis **8 Suppl 2**: 44-8.
Singhal, S., D. Dian, et al. "The Role of Oral Hygiene in Inflammatory Bowel Disease." Dig Dis Sci.
Singleton, D. R., M. A. Furlong, et al. (2001). "Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples." Appl Environ Microbiol **67**(9): 4374-6.
Sreekumar, A., L. M. Poisson, et al. (2009). "Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression." Nature **457**(7231): 910-4.

Thein, Z. M., Y. H. Samaranayake, et al. (2006). "Effect of oral bacteria on growth and survival of Candida albicans biofilms." Arch Oral Biol **51**(8): 672-80.
Turnbaugh, P. J., R. E. Ley, et al. (2007). "The human microbiome project." Nature **449**(7164): 804-10.
Vaz, A. B., R. C. Mota, et al. (2009). "Antimicrobial activity of endophytic fungi associated with Orchidaceae in Brazil." Can J Microbiol **55**(12): 1381-91.
Watson, J. D. (1990). "The human genome project: past, present, and future." Science **248**(4951): 44-9.
White, J. R., N. Nagarajan, et al. (2009). "Statistical methods for detecting differentially abundant features in clinical metagenomic samples." PLoS Comput Biol **5**(4): e1000352.
Wilcoxon, F. (1945). "Individual comparisons by ranking methods." Biometrics **1**: 80-83.
Wishart, D. S. (2007). "Current progress in computational metabolomics." Brief Bioinform **8**(5): 279-93.
Wu, G. D., J. D. Lewis, et al. "Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags." BMC Microbiol **10**: 206.
Zaura, E., B. J. Keijser, et al. (2009). "Defining the healthy "core microbiome" of oral microbial communities." BMC Microbiol **9**: 259.

# Curriculum Vitae

Mr. Robert Brown was born in Dearborn Michigan; he received his B.G..S., with majors in Chemistry and Biology, from the University of Michigan in 1975. He work in biology labs and in writing applications for computer based design before obtaining his Master in Computer Science from the University of Southern California in 1983. He started working in the Top Secret realm as a contractor and eventually joined the U.S. Government as an employee where he then decided to pursue a PhD in Bioinformatics. He is listed as an author in one submission to PLoS ONE titled “Metabolomics Reveals Differential Levels of Oral Metabolites in HIV-Infected Patients”. Previously he was published in Federal Computer Week Sept 11, 2006 titled “Service Oriented Architecture”. Mr. Brown is a National Merit Semifinalist and currently a Member of Tau Beta Pi National Engineering Honor Society plus the American Association for the Advancement of Science (AAAS).