Historically, most biological and medical investigations have examined a few discrete outcomes of interest, and only a few controllable parameters were modified to perturb or improve these outcomes. Investigations of this form went hand-in-hand with the development of inferential statistics, which provide the quantitative tools to detect which perturbations successfully improve outcomes. Biological and clinical research has entered a realm of modifying hundreds or thousands of experimental parameters in high throughput, however, and high-dimensional statistics have been developed to understand which of these modifications in turn significantly improve outcome.
ABSTRACT

Historically, most biological and medical investigations have examined a few discrete outcomes of interest, and only a few controllable parameters were modified to perturb or improve these outcomes. Investigations of this form went hand-in-hand with the development of inferential statistics, which provide the quantitative tools to detect which perturbations successfully improve outcomes. Biological and clinical research has entered a realm of modifying hundreds or thousands of experimental parameters in high throughput, however, and high-dimensional statistics have been developed to understand which of these modifications in turn significantly improve outcome. The field has now reached a point where hundreds or thousands of outcomes can be simultaneously measured as well, but few statistical tools exist to answer the question, "When many experimental or patient outcomes are measured simultaneously, and many experimental parameters or treatments are modified, which modifications are significantly associated with improved outcomes?" This project thus aims to develop novel statistical methods for efficiently associating many controllable predictor variables with many observed response variables with high sensitivity and specificity.
Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
</table>


TOTAL: 5

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper


TOTAL: 4
Number of Papers published in non peer-reviewed journals:

(c) Presentations
31. "From microbes to microbiota and back: using thousands of genomes to understand thousands of metagenomes," Symposium and Workshop on New Methods for Phylogenomics. Austin, TX, 2013
44- "Bug bytes: bioinformatics for metagenomics and microbial community analysis," 8th International Purdue Symposium on Statistics. West Lafayette, IN, 2012

Number of Presentations: 46.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TOTAL:

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TOTAL:

Peer-Reviewed Conference Proceeding publications (other than abstracts):

<table>
<thead>
<tr>
<th>Received</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TOTAL:
<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Manuscript</th>
</tr>
</thead>
<tbody>
<tr>
<td>08/30/2013</td>
<td>6.00</td>
<td>Dan Knights, Mark S. Silverberg, Rinse K. Weersma, Dirk Gevers, Gerard Dijkstra, Hailiang Huang, Andrea D. Tyler, Suzanne van Sommeren, Floris Imhann, Joanne M. Stempak, Caitlin Russell, Jenny Sauk, Jo Knight, Mark J. Daly, Curtis Huttenhower, Ramnik J. Xavier. Complex host genetics drives bacterial dysbiosis in inflammatory bowel disease, IN submission (04 2013)</td>
</tr>
</tbody>
</table>

**TOTAL:** 3

**Number of Manuscripts:**

**Books**

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TOTAL:**
1- ISCB Overton Prize (Harvard School of Public Health, 2015)
2- eLife Sponsored Presentation Series early career award (Harvard School of Public Health, 2014)
3- Presidential Early Career Award for Scientists and Engineers (Harvard School of Public Health, 2012)
**Student Metrics**

This section only applies to graduating undergraduates supported by this agreement in this reporting period.

- The number of undergraduates funded by this agreement who graduated during this period: 1.00
- The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields: 1.00
- The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields: 0.00
- Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale): 1.00
- Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering: 0.00
- The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense: 0.00
- The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

### Names of Personnel receiving masters degrees

<table>
<thead>
<tr>
<th>NAME</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boyu Ren</td>
<td>1</td>
</tr>
</tbody>
</table>

### Names of personnel receiving PHDs

<table>
<thead>
<tr>
<th>NAME</th>
<th>Total Number</th>
</tr>
</thead>
</table>

### Names of other research staff

<table>
<thead>
<tr>
<th>NAME</th>
<th>PERCENT_SUPPORTED</th>
<th>FTE Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yo-Sup Moon</td>
<td>0.85</td>
<td>1.15</td>
</tr>
<tr>
<td>George Weingart</td>
<td>0.30</td>
<td></td>
</tr>
</tbody>
</table>

### Sub Contractors (DD882)

### Inventions (DD882)

### Scientific Progress

**Technology Transfer**

HALLA's implementation as a Python package (see http://huttenhower.sph.harvard.edu/halla) has been carried out in collaboration with Weingart Informatics, an independent software development contractor in San Francisco. This has allowed academic development and validation of the algorithm to be carried out efficiently by students and postdoctoral fellows, while Dr. Weingart has provided industry-quality code, unit testing, packaging, documentation, and distribution. He has begun expanding this implementation to a generalizable Python platform for scientific workflow execution, AnADAMA, which may be a target for future industry partnership in the lab.
60119-MA: Scalable biomarker discovery for diverse high-dimensional phenotypes

Associate Professor Curtis Huttenhower, Department of Biostatistics, Harvard School of Public Health

Our final technical report for this project includes completion of all methodological development for HAllA (the Hierarchical All-against All input/output association testing approach), in addition to work previously completed for the MaAsLin (Multivariate Analysis with Linear models) system for high-dimensional biomarker discovery in compositional data. Both are available as open-source software packages with documentation, demonstration data, and tutorials at http://huttenhower.sph.harvard.edu/halla and http://huttenhower.sph.harvard.edu/maaslin, respectively. Our final publication list includes six manuscripts (PMIDs 22699609, 22699610, 24629344, 23949665, 23013615, and 25732063) and two currently in review (HAllA and its application to the oral microbiome).

Problem Statement

Historically, most biological and medical investigations have examined a few discrete outcomes of interest, and only a few controllable parameters were modified to perturb or improve these outcomes. Investigations of this form went hand-in-hand with the development of inferential statistics, which provide the quantitative tools to detect which perturbations successfully improve outcomes. Biological and clinical research has entered a realm of modifying hundreds or thousands of experimental parameters in high throughput, however, and high-dimensional statistics have been developed to understand which of these modifications in turn significantly improve outcome. The field has now reached a point where hundreds or thousands of outcomes can be simultaneously measured as well, but few statistical tools exist to answer the question, "When many experimental or patient outcomes are measured simultaneously, and many experimental parameters or treatments are modified, which modifications are significantly associated with improved outcomes?" This project thus aims to develop novel statistical methods for efficiently associating many controllable predictor variables with many observed response variables with high sensitivity and specificity (Fig. 1).

Figure 1: Overview of HAllA. A) Two or more input datasets are represented in matrix form as features (rows) and samples (columns). B) Continuous data are (optionally) discretized to provide a unified representation of potentially heterogeneous feature types. C) Features within each data set are single linkage hierarchically clustered, using normalized mutual information as the default, fully generalizable similarity metric. D) A hypothesis tree is built by coupling clusters between two datasets at equivalent relative depths. Each hypothesis node has compares two
clusters, with all pairs of children of the two clusters forming the next level of hypothesis testing. E) Hypothesis testing is performed by, first, selecting each cluster's medoid as a representative summary (optionally either multiple correspondence analysis or principle component analysis instead), and a permutation test is then used to determine which pairs are significantly associated between the two datasets. F) Significant associations are reported after false discovery rate controlling for each hypotheses set (family, level, or all).

Results Summary

Generalized multiple input/output association testing: the Hierarchical All-against-All approach

HAllA (Hierarchical All-against-All association testing) is a novel statistical method for well-powered association discovery in high-dimensional heterogeneous datasets, which we have developed, implemented, validated, and applied to diverse datasets. It combines hierarchical hypothesis testing with false discovery rate correction over highly generalizable association measures, yielding high-powered discovery of linear and non-linear patterns in categorical or continuous high-dimensional data. Data and metadata to be associated are hierarchically clustered for dimensionality reduction, and nonparametric permutation testing identifies relationships between the resulting blocks of correlated features.

HAllA was validated and optimized with synthetic data, outperforming exhaustive all-against-all association testing and alternative similarity measures such as the Maximum Information Coefficient and Spearman correlation in Types I and II error and in runtime (Fig. 2). The recommended HAllA algorithm first groups features by single-linkage agglomerative hierarchical clustering using normalized mutual information (NMI), then compares blocks of features by descending each tree logarithmically. Outliers are removed by comparing the relative dissimilarities within and between datasets, after which cluster medoids are compared (also using NMI) and significant associations determined by permutation testing. False discovery rates are globally controlled while maintaining power by correcting within each tree level. Each choice in the algorithm is modularized and configurable, however, allowing users to select other measures more appropriate for homogeneous data (e.g. continuously valued Spearman correlations) or to use other block summarization techniques (e.g. MCA or PCA). The software implementation and documentation are available at http://huttenhower.sph.harvard.edu/halla.

Figure 2: HAllA efficiently maintains Type I and II error relative to alternative approaches. We simulated 30 datasets with features associated by linear, parabolic, and sinusoidal relationships. HAllA outperformed exhaustive all-against-all (AllA) testing using multiple measures: the maximum information coefficient (MIC), normalized mutual information (NMI) alone, and alternative summarizations including MCA or PCA.

We created the Python package STRUDEL (Stratified Rudimentary Data Exploration) to produce synthetic data with defined correlation structure within datasets and associations between blocks of variables among datasets. Each evaluation dataset contained correlated
features, and each paired dataset contained correlations between blocks of features with both linear and non-linear patterns. Using these simulated data, we evaluated the performance of HAllA and naive all-against-all (AllA) methods using NMI, MIC, distance correlation (dcor), and Spearman correlation as similarity metrics. Simple correlations identify only linear associations, while dcor requires continuous data and performed poorly for discovery of complex nonlinear associations. MIC was very time-consuming even for small datasets often had a high FDR. However, NMI efficiently found all types of associations (linear, parabolic, and sinusoidal) and when applied using hierarchical testing improved both precision and recall. As a negative control, applying HAllA to null datasets containing no associations resulted in the expected baseline false discovery rate (a user configurable parameter).

We applied HAllA to three real-world datasets containing highly diverse data types from several application areas. First, we assessed data published by Martin et al (Hepatology 2007) for 21 liver lipid levels and 120 genes' hepatic transcription levels in 40 wild-type and peroxisome proliferator-activated receptor-α (PPARα)-deficient mice. HAllA recovered all associations previously reported by Gonzalez et al (J. Stat. Software 2008) using canonical correlation analysis, including for example transcriptional activation of xenobiatic metabolism genes Cyp3a11 and CAR1 in conjunction with high fatty acid levels. We further identified novel associations between a cluster of transcripts including CAR1 and ACOTH (fatty acid transport and trafficking) and a cluster of fatty acids including C18.0, C20.3n.6, and C20.4n.6. The associations identified by HAllA were a strict superset of those previously reported, demonstrating the utility of the method as a general-purpose tool that can find general patterns with high sensitivity (Fig. 3).

Figure 3: HAllA identifies validated and novel associations between mouse liver transcripts and fatty acid levels. Associations between hepatic fatty acids and gene expression in data from Martin et al (Hepatology 2007). '*' indicates significant associations identified by HAllA between corresponding genes and fatty acids, a strict superset of those previously reported.

Next, HAllA expanded upon microbe-metabolite associations reported the infant gut microbiome by Kostic et al (CHM 2015). This test of the hygiene hypothesis studied a prospective cohort of 960 Finnish, Russian, and Estonian infants at risk of type 1 diabetes. These subjects were followed for three years, while monthly stool samples and a variety of clinical metadata (e.g. breastfeeding, diet, allergies) were collected. The resulting 104 stool samples were assessed for microbial community composition by 16S rRNA gene sequencing and profiled metabolomically. We applied HAllA to the abundances of 20 genera and 284 metabolites from these data, again identifying a set of relationships that were previously detected (e.g. Veillonella and sphingomyelins 22:0, 24:0, 24:1, C16:0, and C18:0; Ruminococcus and sphingomyelins 24:0 and 24:1). We further recovered an association between Haemophilus and phenylalanine in conjunction with a novel grouping with tryptophan and tyrosine. Blautia spp. were non-linearly associated with long-chain triglycerides, whereas Ruminococcus spp. associated with short-chain triglycerides and Lactobacillus spp. with both. Finally, Enterococcus was positively associated with diacyl and triacyl glycerols, in agreement with independent reports of Enterococcus faecium bioactivity in improving diabetic lipid (Roselino Lipids Health 2012) and triglyceride (Cavallini Lipids Health 2009) levels.
Finally, we applied HAllA to data collected from 204 ileal resection patients in which microbial community profiles and host transcriptomes were assessed from 255 biopsies (Morgan CHM 2015). Host transcription was assayed by Affymetrix microarray and microbiome profiles by 16S rRNA gene sequencing. Previous work correlated antibiotic use, inflammation, biopsy location, and clinical outcome with the transcriptome and microbiome, but extensive dimensionality reduction was required to preserve power while determining microbe-transcript associations. We applied HAllA to a highly-abundant (above 50th percentile) and variant subset (above 95th percentile), comprising 108 microbes and 498 transcripts. We found 576 associations between features and feature clusters; 34 included at least one of the top gene principal component loadings described previously. Of these, 21 corresponded to the genes most influential in the previously reported principal component 9, which represented only 1% of expression variation but was correlated with the most microbes.

Additionally, we found that the abundance of *Escherichia* was correlated with host expression of MUC1 and nitric oxide synthase. Expression of the sodium / bile acid transporter SLC10A2, the primary bile transporter in the distal ileum, was correlated with the genera *Coprococcus*, *Blautia*, and *Clostridium* from the family Erysipelotrichaceae. Notably, antibiotic use was associated with increased SLC10A2 expression and total bile acids, in agreement with Miyata et al (J. Pharma. Exp. Therapy 2011), and both Crohn’s disease and ulcerative colitis have previously been associated with decreased metagenomic abundance of bile salts among the Lachnospiraceae and Erysipelotrichaceae (Labbe PLoS ONE 2014).

The manuscript describing HAllA is currently in review, and it has previously been presented at the 2015 Broad Institute retreat, the 2015 Statistical and Applied Mathematical Sciences Institute workshop on Discovering Patterns in Human Microbiome Data, the 2014 Intelligent Systems for Molecular Biology (ISMB) conference, and the 2014 Dana-Farber Cancer Insitute Biostatistics and Computational Biology seminar series. Its software and evaluation package are complete, documented, and supported by online tutorials and an active user group (see http://huttenhower.sph.harvard.edu/halla and http://bitbucket.org/biobakery/halla). It is currently in use by ongoing collaborations with Dr. Jessica Green at the University of Oregon, Dr. Frank Nestle (King’s College for the MAARS study), and Dr. Mihai Netea (Radboud University Nijmegen). Postdoctoral fellow Dr. Gholamli Rahnavard is the current lead, in collaboration with research associate Dr. Eric Franzosa and completing work previously executed by former research assistant Yo Sup Moon and postdoctoral fellows Timothy Tickle (currently at the Broad Institute) and Levi Waldron (currently at Hunter College).

**Multiple input/output microbial community methods: Multivariate Analysis with Linear models**

The Multivariate Analysis with Linear models (MaAsLin) method performs efficient high-dimensional association testing specifically for microbial communities. Microbial taxonomic and functional profiles possess unique statistical characteristics, in particular a combination of sparsity (many zero values), compositionality (fixed per-sample sum) or count data, and longitudinal variability (i.e. repeated measures), including typical epidemiological characteristics of heterogeneous high-dimensional metadata. MaAsLin was developed as a well-powered statistical association methodology and software appropriate for microbial data in conjunction with any experimental design and associated phenotypic, clinical, environmental, or ’omic metadata.
The MaAsLin algorithm combines four main steps to associate relative abundance profiles with heterogeneous metadata (Fig. 4). First, data are preprocessed to remove outliers, handle missing values, and ensure quality control and consistency. Next, one of three link functions appropriate for microbial profiles is applied: a variance-stabilizing arcsin square-root transformation by default, or a log-linear Gaussian link, or a negative Binomial count link. The first (arcsin-sqrt) is an approximation that is less theoretically appropriate than the latter two, but performs comparably well in evaluations and is substantially more computationally efficient and numerically stable. Third, dimensionality is reduced if necessary using a feature selection step (boosting by default, LASSO and univariate selection optionally available). Finally, significant associations are identified using a mixed effects linear model (zero-inflated by default).

MaAsLin has been applied in a variety of published and ongoing studies during the total project period:


• Tiffany Hsu, Regina Joice, Jose Vallarino, Caleb Abu-Ali, Erica M. Hartmann, Afrah Shafquat, Casey DuLong, Catherine Baranowski, Dirk Gevers, Jessica L. Green, Xochitl C. Morgan, John D. Spengler, Curtis Huttenhower. "Urban transit system microbial communities differ by surface type and interaction with humans and environment." in review

• Daniela Börnigen, Boyu Ren, Robert Pickard, Jingfeng Li, Erica M. Hartmann Wei Hong Xiao, Timothy Tickle, Jennifer Rider, Dirk Gevers, Mary Ellen Davey, Curtis Huttenhower, Maura Gillison. "Alterations in the oral microbiome associated with oral cancer risk factors and their contributions to pathogenesis." in review

MaAsLin has also been presented in a wide range of venues, including:

• "High-precision functional profiling of the gut microbiome for characterization during inflammatory disease." University of Pittsburgh Immunology seminar. Pittsburgh, PA, 2015


• "High-precision Functional Profiling of Microbial Communities and the Human Microbiome." 41st Annual Northeast Bioengineering Conference. Albany, NY, 2015

• "A Tour of the bioBakery: Computational Tools for Microbial Community Analysis." Broad Institute Medical and Population Genetics seminar. Cambridge, MA, 2015 (presented by Eric Franzosa)


• "The microbiome in IBD and analysis methods for microbial communities." International Inflammatory Bowel Disease Genetics Consortium meeting. Barcelona, Spain, 2015


• "Metagenomics, metatranscriptomics, and multi’omic integration." Massachusetts General Hospital Center for the Study of Inflammatory Bowel Disease research symposium. Boston, MA, 2014

• "High-precision methods for metagenomic and metatranscriptomic profiling." New York University Medical School seminar. New York, NY, 2014

• "Gut microbial epidemiology and biogeography." University of Washington Genome Sciences seminar. Seattle, WA, 2014


• "Computational Approaches for the Human Microbiome in Health and Disease," 12th Biennial Congress of the Anaerobe Society of the Americas. Chicago, IL, 2014


• "Host-microbiome transcriptional crosstalk and clinical outcome in a large ileal pouch-anal anastomosis (IPAA) cohort," 109th International Titisee Conference. Titisee, Germany, 2014


• "Bug bytes: bioinformatics for the human microbiome in health and disease," University of Michigan Molecular and Clinical Epidemiology of Infectious Diseases (MAC-EPID) symposium. Ann Arbor, MI, 2013
• "Computational methods for meta'omic characterization of the human microbiome," Tufts Computer Science Department Colloquium Series. Medford, MA, 2013
• "Adding depth to human microbiome studies with multi’omic data integration," International Human Microbiome Congress. Hangzhou, China, 2013
• "From microbes to microbiota and back: using thousands of genomes to understand thousands of metagenomes," Harvard School of Public Health Bioinformatics Core Forum. Boston, MA, 2013
• "From microbes to microbiota and back: using thousands of genomes to understand thousands of metagenomes," Symposium and Workshop on New Methods for Phylogenomics. Austin, TX, 2013
• "From microbial surveys to mechanisms of interaction in the human microbiome," University of Colorado at Boulder BioFrontiers Institute seminar. Boulder, CO, 2013
• "Meta'omic characterization of microbial community function in health and disease," American Society for Microbiology Conference on Beneficial Microbes. San Antonio, TX, 2012
• "Bug bytes: bioinformatics for metagenomics and microbial community analysis," 8th International Purdue Symposium on Statistics. West Lafayette, IN, 2012

The software is currently available with documentation and demonstration data at http://huttenhower.sph.harvard.edu/maaslin, with an additional Galaxy interface online at http://huttenhower.sph.harvard.edu/galaxy. The project is currently led by postdoctoral fellow Dr. Ayshwarya Subramanian, with previous contributions by former postdoctoral fellows Dr. Soumya Bannerjee (currently at Children's Hospital) and Dr. Timothy Tickle (currently at the Broad Institute), Ph.D. student Emma Schwager, and undergraduate research assistant Yiren Lu.