Data integration on inflammatory bowel disease

Lluís Revilla Sancho

2020-10-22

Contents

Preface

The main topic of the thesis is data integration applied in the inflammatory bowel disease (IBD) research. The data I will be integrating are different omics data and the phenotype of the patients. This disease is complex and there are hypothesis pointing that the microbiome is a major factor in the disease. In the precision medicine framework data integration is important to consider all the relevant variables that influence a disease.

The thesis is performed on the IDIBAPS research institute. My colleges are biologist, microbiologists, veterinaries... and we have weekly meetings with the doctors visiting at the nearby hospital.

The thesis program allows one to defend the thesis after 3 years, and up to 5 in total. My timeline is to finish in 2022 after 4 years (I'll add it when it is finished to see how it went) with the guide of my thesis directors Juanjo Lozano and Azucena Sala, who help as bioinformatician and disease expert respectively.

Chapter 1

Introduction

The inflammatory bowel disease (IBD) involves Crohn's disease(CD) and ulcerative colitis(UC). It generally affects the terminal ileum and the colon but it can affect any segment of the gastrointestinal tract. UC is a recurrent, chronic and continuous inflammation of the colon and rectum while the CD is not a continuous inflammation and affects the whole gastrointestinal tract.

IBD etiology is unknown. However, once it has initiated the most prevalent hypothesis of its chronicity suggests an aberrant immunological response to antigens of the commensal microbiome.

Treatments provided for the IBD are palliative. Those treatments include, noninflammatory drugs, suppressors and biologic. The therapeutic options can induce remission in some patients, however they often need continuous treatment to avoid recurrence. Nevertheless, many patients are refractory or intolerant to those therapies and need to undergo surgery.

The disease present unique characteristics that require the usage of integration methods in order to find the specific relationship of the microbiome and the intestine on the disease.

1.1 Integration

Integration is usually defined as:

> "the process of combining two or more things into one" — Cambridge Dictionary

Since the beginning of the integration methods there have been many methods proposed. Some of these methods are specific for one application or data while others are more general. On the recent years with the increase of more

datasets with more variables than previously has seen an increase on the number of methodologies available on several disciplines but mainly on the biological science. This in turn has increased the importance of classification, review and comparison of the tools available, as well as, benchmarking these tools against the same dataset [Wu et al., 2019]. Other times only the strategies used are used to classify the methods [Cavill et al., 2016].

As is common, there are several words that are used integrati(-on, -ve), multi-omics, pluri-omics.

1.1.1 Classification

We can characterize and classify integration methods in several ways. Here I outline some classifications in the bioscience field, with relationships (and references) to concrete methodology and usages.

1.1.1.1 Data type: numeric vs categorical

The most important distinction in integration methods is what kind of data are combined. In general data can be divided between categorical and numeric values, which are usually found in several fields. Sometimes doctors want to understand the relationship between some phenotype they observe as a defined state from the underlying mechanistic point of view. Usually this involves looking how the metabolites, the gene expression, the methylation, the number of variants a gene has, and other numeric variables are related to the observed (categorical) phenotype (like pain).

Depending on what is a method aimed for it handles both data types or just one, often they are used differently. The most common way to handle different type of data is converting the categorical values to a mock variable, where each number represents a category, then the methods use these values to correlate and relate the variables. If the categorical variable has three values (A, B, C) it would be converted to A (1, 0, 0) B (0, 1, 0) and C (0, 0, 1).

1.1.1.2 Aim

The results and methods of data integration depend on the question they seek to answer and on the (biological) question. Most of the times one (or all) of the following results are expected:

- An overview of the role of each 'omics in a biological system

Sometimes several omics are used and the question is which omic method is the best one to describe the disease. Or several methods are used to understand better how two omics interact.

- A better understanding of the relationship (correlation) between the 'omics types

When the relationship between omics is known there are methods to find those relationships and improve upon previous knowledge. For instance, checking that in a particular case or condition there is a given relationship, and that this relationships follows our model or not.

- A molecular signature[1] leading to more insight into molecular mechanisms

The usage of multifactor analysis might lead to the need to redefine the features that are essential for identifying a phase or a cell line.

- A predictive analytic model towards personalized medicine

If a good model is known it can be used as a prediction if enough information is gathered, which might improve the treatment.

1.1.1.3 Relationship between variables and samples

Depending on the amount of variables and in which samples they have been measures we can classify in two types of integration. Traditionally for each sample few variable has been measured, for instance for a tree only the height and width are measured, however with the new omics techniques (transcriptomics, metabolomics, methylomics, genomics), thousands of variables are measured for the same sample.

- More variables than samples:

For a single sample of RNA around 50k genome identifiers (genes, long non coding RNAs, iRNA, pseudogenes,...) can be measured. High-throughput data analysis typically falls into the category of $p \gg n$ problems, where the number of genes or proteins, p, is considerably larger than the number of samples, n. Which leads to the case where there are (many) more variables than samples, generally "old" statistics don't consider this case, as it has its own complications like co-variance between the variables. When two variables are tightly correlated, discerning which is the lead and which is following is near to impossible.

- More samples than variables:

An example would be when from a cohort of patients the temperature is measured along the stage of a disease: two variables for each sample. If in the cohort there are more than 2 patients, then the number of samples is greater than the number of variables studied. This is described in the literature as $n \gg p$.

[1] A signature is usually a group of features that describe/are representative of a cell line or a process or a stage.

1.1.1.4 Relationship between samples

Depending on the relationship between the samples, the questions answerable and the methods available differ. If each sample has all the expected data we wanted to measure it is a complete case.

Sometimes because the sample is not enough, or there are some technical or organizational problems we might lose a source of data for a sample (which is known as an incomplete case). This results in a new source of variation that has to be dealt with, which complicates the conclusion one can draw from the studies of these kind of data.

Even when all the cases of a sample are complete the samples can come from several sites of the same individual or with different combinations of variables, which makes is relevant to understand.

1.1.1.4.1 Time

Time is one of the factors that sometimes cannot be controlled, despite having programmed visits every two weeks some patients might come early or later due to calendar reasons (holidays), family reasons, or disease state. Sometimes is precisely the object of the study, to see the relationships at different time, or see how the relationships change with time. Simultaneously is very important to consider it because two variables can seem correlated if we don't take time into consideration. Also, to discover causality between two variables the cause must be before the consequence, which highlights the importance of time. Being aware of the time differences and time scales is crucial in most cases.

1.1.1.5 Relationship between variables

Since the lactose operon we know how some genes regulate each other. For other variables we don't know how they are related. For instance, how does the increase in expression of a gene affects the growth of a microorganism? Usually the relationships between variables are mediated by many factors or interactions.

Network approaches relate the variables between them (such as [Koh et al., 2019]). And are fairly new and despite being used in rare occasions they are growing in popularity.

In partial correlations some or all of the other variables and considered on how much do they affect and deduced.

1.1.1.6 Input data

Methods can be classified by the kind of input data required. Some of them need data from the same patients on each data set used to integrate while other

do not.

- Data from the same samples:

These methods do not handle well or at all missing data. They need complete cases/data of the samples in order to be able to integrate the results. These methods include Regularized Generalized Canonical Correlation Analysis (RGCCA) [Tenenhaus and Tenenhaus, 2011, Tenenhaus et al., 2014], Multiple co-inertia analysis (MCIA) [Culhane et al., 2003], Multi-Study Factor Analysis (MSFA) [Vito et al., 2019], Multi-Omics Factor Analysis (MOFA) [Argelaguet et al., 2018], STATegRa [Gomez-Cabrero et al., 2019].

- Data from different samples:

These methods do not need data from the same sample. They draw their conclusions generalizing from the the data available. Some of them handle missing data, while others do use the data at face value. These method includes MetaPhlAn2, HUMAnN, LEfSe [Franzosa et al., 2018, Truong et al., 2015, Segata et al., 2011].

1.1.1.7 Output results

According to the output the integration methods can be classified in three groups: Shared factor across the data, specific factors for each data or mixed factors.

- Shared factors:

The integration results in a vector of the samples in a lower dimensional space that is shared by all the data used to integrate. Such methods include iCluster, Multi-Omics Factor Analysis (MOFA) [Argelaguet et al., 2018].

- Specific factors:

The integration results in several vectors of the samples in a lower dimensional space of each data used to integrate. Such methods include Regularized Generalized Canonical Correlation Analysis (RGCCA) [Tenenhaus and Tenenhaus, 2011, Tenenhaus et al., 2014], Multiple co-inertia analysis (MCIA) [Culhane et al., 2003], Multi-Study Factor Analysis (MSFA) [Vito et al., 2019].

- Mixed factors:

The integration results in both previous factors, specific of each data and common to all the data. Such methods include Joint and Individual Variation Explained (JIVE), integrative Non-negative Matrix Factorization (intNMF).

1.1.1.8 Interpretation

Understanding how to interpret the results of the methods is highly tight to understanding the method. If one does a correlation between two variables, the

interpretation of the analysis is clear, if one variable increase, the other one too. However as more complicated methodologies are developed the interpretation becomes less clear, for instance how can one interpret the result of a canonical correlation analysis?

- Individually:

How each variable relates to another, like in the correlation analysis, the relationship between two variables under study. Or by patient: how do interpret that in these patient variable A and B is X and Y?

- Globally:

In a PCA for instance how do we interpret that some variables have the same loading? What happens in a more difficult method like canonical correlation analysis?

There have been some articles about how to interpret those methods on real datasets [Sherry and Henson, 1981]. Others, to benchmark and to learn how to interpret propose analyzing a simulated dataset [Chung and Kang, 2019, Martínez-Mira et al., 2018].

1.1.1.9 Conclusion

The field of integration is large and complex, with high interest in the recent days, specially in the psychology and omics field, which lacks of a large study.

1.1.2 Reviews and benchmarking

The comparison and review of methods independently from original authors have become a crucial step for selecting the right tool for a research [Cantini et al., 2020].

Some of these reviews are focused on a field: metabolomics (ref), genomics, microbiomics... or on a specific characteristic: ?.

1.1.3 Summary

Methods to integrate have many characteristics that allow to classify them which explains the diverse results one can have using one of them.

1.2 Inflammatory bowel disease

Inflammatory bowel disease (IBD) includes the chronic diseases Crohn's disease (CD) and ulcerative colitis (UC). CD is a progressive reincident disease that

can affect all the gastrointestinal tract but shows mostly on both terminal ilium and colon. The UC is a colonic reincident disease characterized by a continuous inflammation of the colon.

Around 4,2 million individuals suffer from IBD in Europe and North America combined. The dysregulation of the inflammatory response observed in IBD requires interplay between host genetic factors and the intestinal microbiome. Several studies support the concept that IBD arise from an exacerbate immune response against commensal gut microorganisms. Nonetheless, the disease could result from an imbalanced microbial composition leading to generalized or localized dysbiosis[2].

The role of the gut microbiome in IBD is an ongoing field of research. Several authors are currently studying the alterations reported in IBD of the intestinal microbiome. However, it is still unclear the cause-effect relation between dysbiosis and IBD. Partly due to the multiple variables that might contribute to the disease progression; for instance, age, diet, usage of antibiotic, tobacco, environment, and eventually socioeconomic status. This could be due to both the genetic predisposition and environmental factors; for instance, bacterial or viral infection, diet, usage of antibiotic, and eventually the socioeconomic status.

(see [Human Microbiome Project Consortium et al., 2012])

The relationship between host and microbiome has been proposed to play a fundamental role to maintain the disease. Little is known of the influence of the gastrointestinal microbiome in the expression of the gastrointestinal tract.

1.2.1 Disease onset

Although so far we do not know what starts the disease there are differences on the disease when appears at different age. There is a rough classification between very early, early or adult on-set disease.

The main differences are [ask Isa?] .

1.2.2 Disease discourse

Intermittent

There is to some extent a disassociated between patient's (clinic) report and the colonoscopy report. When patients report being better but their gut has bigger ulcers or is more inflamed.

[2]A signature is usually a group of features that describe/are representative of a cell line or a process or a stage.

Although Crohn's disease and colon disease are similar the drugs and treatments for each is different. The order of the treatment/drugs changes depending on the disease.

Endoscopic and clinic response.

Mucosal healing

Colectomy

Dysplasia

Fistula

Herpes and citomegalovirus virus.

Scores:

- Mayo
- SESCD

Other measured parameters:

- Weight
- Calprotectin
- PCR (Proteinc C reactive)
- Hemoglobine

1.2.2.1 Treatments

The ones given on the hospital more or less on the order of administration:

- anti-TNFα
- Vedolizumab
- Ustekinumab
- Risankizumab
- Tofacitinib
- surgery
- TRIM/HSCT
- 5-asa
- corticoesteroids
- Azathiprine/Mercaptopurine 6MCP
- Methotrexate MTX
- Tacrolimus FK
- Cyclosporin A CyA
- Infliximab
- Adalimumab
- Antibiotics

All these drugs and procedures are available at the time of writing.

If there is a reduction in doses because the treatment is effective the patients sometimes lose the response to the drug, so after an increase of the dose the do not recover the initial response they had.

1.2.3 Integration on IBD

One of the hypothesis behind the maintenance of the inflammation involves the microbiome. Several studies have been carried out to discover links between microbiome and the inflammation. Some of these studies used metabolites, DNA-seq sequencing of the microbiome content, or targeted 16S sequencing.

Also the technical method used can differ between extracted from stools or from biopsied samples at colonoscopy or from surgical samples.

Some articles use correlation like [Häsler et al., 2016]. There are others that use a combination of methods

Very rarely there is an experimental confirmation. This is due to how complicated it is to test an interaction and the simulate the conditions.

One of the few methods published were the interaction is measured is to expose the ex-vivo sample or cell line with supernatant of a microbiome culture.

Previous methods used include the use of RGCCA, [Tang et al., 2017]

[De Souza et al., 2017]

1.2.4 Summary

The integration of data might help to improve the medicine and reveal links in difficult diseases like IBD. So far it has been applied in IBD with partial success.

Chapter 2

RGCCA

The canonical correlation is a method that using data from the same sample but from different datasets. Characteristics Input Data type: numeric More variables than samples: It is appropriate when the number of variables is higher than the number of samples. Need to be a complete case. Time is not considered as a specially variable Relationship between variables

Output:

Specific factors Interpretation:

Depending on the model and options used.

Over several years of progress [Tenenhaus, 2008, Tenenhaus and Hanafi, 2010, Tenenhaus and Tenenhaus, 2011, 2014, Tenenhaus et al., 2014, 2015] on the field of canonical correlations it provides a robust method and there is an implementation [**?**].

The regularized generalized canonical correlation is a method that combines several datasets, using data from the same sample. It is a good choice when the number of variables is higher than samples. There have been some improvements to generalize the method when $p \gg n$. Current practices include using a pre selected model of relations between blocks. However this model might not be accurate. To help find the fitting model for the data I created an R package. The package, which is named inteRmodel helps finding the right model and how fit it is for your data.

This method applied to an existing dataset of an autologous haematopoietic stem cell transplantation [Corraliza et al.]. From this dataset there is data about the human transcriptome and the 16S DNA present at biopsies from colonoscopy. For several patients we have samples at different time point and at different locations. This allows us to see both location and time differences.

We looked to several models and searched for the model that fit better with the

data. The first model just accounted the transcriptomics and the microbiome data, then in another family of models we added the information we know about those samples. In a further family of models we split the information we know about these samples into three different blocks grouping them according to how are they related between them. We fitted the best model for all the three family of models and we found that the most fitting model was from the family that had the data split on several.

Additionally I introduced some changes to RGCCA package. The modified package can be found here. This version return the same results as the version on CRAN but provides also some code optimization for a lower computation time. Some of these changes are to to be able to change the design between different dimensions. The idea behind this change is that if the first dimensions correctly fits the data the second dimension might need a different model of relationships. For instance, the first dimensions are dominated by two blocks of data while the second is dominated by another block of data.

2.1 This requires a model of relationships.

More blocks more possible models

Limitations: symmetric Y -> X & X -> Y

2.2 Designing models

What defines a block: Blocks of variables are treated independently

Considering disease

Randomly

2.3 Evaluating models

AVE, inner and outer

Bootstraping/validation on an external cohort.

Biological relevance

Chapter 3

Biological relevance of results

This is a looooooooooooooooooooooong sennnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn-tence wrapped. See you can continue writting right after the dot and it gets written on the next line on the raw file.

References

::: {#refs} :::

Appendix A

Online resources

Some links that I found useful on the thesis and could be useful if you are interested on the multi-omics field.

- Awesome multi-omics: An online repository of references to multi-omics methods.
- Bookdown: The book about how to write this type of books.

Appendix B

Software

Along the years of this thesis several pieces of software have been generated as well as packages. Here they are listed for easier retrieval. They are listed on two ways, one with a brief explanation and another one ordered by what software piece is used on each analysis.

B.1 Listed

An improved/tested version of RGCCA, some modifications on the internal functions to ease the maintenance as well as adding tests and sometimes improving the documentation. Also modified so that it is possible to provide a vector of models so that the model of the first dimension is not the same as the model on the second dimension (not sure if mathematically speaking makes sense but from a biological one I think it might be interesting to have it).

Designed to be used with RGCCA I wrote inteRmodel to ease the bootstrapping and model selection.

A package to design batches to avoid batch effect experDesign and its website on GitHub

A package to analyze sets and fuzzy sets BaseSet. This package was meant to be used with the probabilities that arise from bootstrapping the models. However due to the long times of calculation that it would require it was not used. (A previous iteration of the package called GSEAdv was developed too.)

B.2 By publication

All code of the analysis of the publications is available (in his messed state and complicated history) and a brief description as to why they were used:

Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis:

- TRIM: Mangle with the sample, dataset, explore several methods...
- sgcca_hyperparameters: Explore the effects of the hyperparameters on RGCCA on the provided dataset.
- inteRmodel: Package for easy repeating the methodology developed on TRIM.
- integration: Package with functions that I wrote several times on the TRIM ended up here.

Paper 2:

- Barcelona: Mangle with the sample, dataset ...
- inteRmodel: To repeat the same procedure as in the first analysis.
- integration: To not repeat myself for the same process.

Bibliography

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, jun 2018. ISSN 1744-4292. doi: 10.15252/msb.20178124.

Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. *bioRxiv*, page 2020.01.14.905760, jan 2020. doi: 10.1101/2020.01.14.905760.

Rachel Cavill, Danyel Jennen, Jos Kleinjans, and Jacob Jan Briedé. Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, 17(5): 891–901, sep 2016. ISSN 1467-5463. doi: 10.1093/bib/bbv090.

Ren-Hua Chung and Chen-Yu Kang. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *GigaScience*, 8(5), may 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz045.

Ana M. Corraliza, Elena Ricart, Alicia López-García, Maria Carme Masamunt, Marisol Veny, Miriam Esteller, Aida Mayorgas, Lionel Le Bourhis, Matthieu Allez, Núria Planell, Sudha Visvanathan, Patrick Baum, Carolina España, Raquel Cabezón-Cabello, Daniel Benítez-Ribas, Montserrat Rovira, Julián Panés, and Azucena Salas. Differences in Peripheral and Tissue Immune Cell Populations Following Haematopoietic Stem Cell Transplantation in Crohn's Disease Patients. *Journal of Crohn's and Colitis*. doi: 10.1093/ecco-jcc/jjy203.

Aedín C. Culhane, Guy Perrière, and Desmond G. Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1):59, dec 2003. ISSN 1471-2105. doi: 10.1186/1471-2105-4-59.

Heitor S.P. De Souza, Claudio Fiocchi, and Dimitrios Iliopoulos. The IBD interactome: An integrated view of aetiology, pathogenesis and therapy. 14, aug 2017. doi: 10.1038/nrgastro.2017.110.

Eric A. Franzosa, Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, Rob Knight, J. Gregory Caporaso, Nicola Segata, and et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11):962, Nov 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0176-y. URL https://www.nature.com/articles/s41592-018-0176-y.

David Gomez-Cabrero, Sonia Tarazona, Isabel Ferreirós-Vidal, Ricardo N. Ramirez, Carlos Company, Andreas Schmidt, Theo Reijmers, Veronica von Saint Paul, Francesco Marabita, Javier Rodríguez-Ubreva, Antonio Garcia-Gomez, Thomas Carroll, Lee Cooper, Ziwei Liang, Gopuraja Dharmalingam, Frans van der Kloet, Amy C. Harms, Leandro Balzano-Nogueira, Vincenzo Lagani, Ioannis Tsamardinos, Michael Lappe, Dieter Maier, Johan A. Westerhuis, Thomas Hankemeier, Axel Imhof, Esteban Ballestar, Ali Mortazavi, Matthias Merkenschlager, Jesper Tegner, and Ana Conesa. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Scientific Data*, 6(1): 1–15, oct 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0202-7.

Robert Häsler, Raheleh Sheibani-Tezerji, Anupam Sinha, Matthias Barann, Ateequr Rehman, Daniela Esser, Konrad Aden, Carolin Knecht, Berenice Brandt, Susanna Nikolaus, Sascha Schäuble, Christoph Kaleta, Andre Franke, Christoph Fretter, Werner Müller, Marc-Thorsten Thorsten Hütt, Michael Krawczak, Stefan Schreiber, and Philip Rosenstiel. Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. *Gut*, pages gutjnl–2016–311651, 2016. ISSN 0017-5749. doi: 10.1136/gutjnl-2016-311651.

Barbara A. Human Microbiome Project Consortium, Karen E. Nelson, Mihai Pop, Heather H. Creasy, Michelle G. Giglio, Curtis Huttenhower, Dirk Gevers, Joseph F. Petrosino, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Ashlee M. Earl, Michael G. FitzGerald, Robert S. Fulton, Kymberlie Hallsworth-Pepin, Elizabeth A. Lobos, Ramana Madupu, Vincent Magrini, John C. Martin, Makedonka Mitreva, Donna M. Muzny, Erica J. Sodergren, James Versalovic, Aye M. Wollam, Kim C. Worley, Jennifer R. Wortman, Sarah K. Young, Qiandong Zeng, Kjersti M. Aagaard, Olukemi O. Abolude, Emma Allen-Vercoe, Eric J. Alm, Lucia Alvarado, Gary L. Andersen, Scott Anderson, Elizabeth Appelbaum, Harindra M. Arachchi, Gary Armitage, Cesar A. Arze, Tulin Ayvaz, Carl C. Baker, Lisa Begg, Tsegahiwot Belachew, Veena Bhonagiri, Monika Bihan, Martin J. Blaser, Toby Bloom, J. Vivien Bonazzi, Paul Brooks, Gregory A. Buck, Christian J. Buhay, Dana A. Busam, Joseph L. Campbell, Shane R. Canon, Brandi L. Cantarel, Patrick S. Chain, I-Min A. Chen, Lei Chen, Shaila Chhibba, Ken Chu, Dawn M. Ciulla, Jose C. Clemente, Sandra W. Clifton, Sean Conlan, Jonathan Crabtree, Mary A. Cutting, Noam J. Davidovics, Catherine C. Davis, Todd Z. DeSantis, Carolyn Deal, Kimberley D. Delehaunty, Floyd E. Dewhirst, Elena Deych, Yan Ding, David J. Dooling, Shannon P. Dugan, Wm. Michael Dunne, Jr., A. Scott Durkin, Robert C. Edgar, Rachel L. Erlich, Candace N. Farmer, Ruth M.

Farrell, Karoline Faust, Michael Feldgarden, Victor M. Felix, Sheila Fisher, Anthony A. Fodor, Larry Forney, Leslie Foster, Valentina Di Francesco, Jonathan Friedman, Dennis C. Friedrich, Catrina C. Fronick, Lucinda L. Fulton, Hongyu Gao, Nathalia Garcia, Georgia Giannoukos, Christina Giblin, Maria Y. Giovanni, Jonathan M. Goldberg, Johannes Goll, Antonio Gonzalez, Allison Griggs, Sharvari Gujja, Brian J. Haas, Holli A. Hamilton, Emily L. Harris, Theresa A. Hepburn, Brandi Herter, Diane E. Hoffmann, Michael E. Holder, Clinton Howarth, Katherine H. Huang, Susan M. Huse, Jacques Izard, Janet K. Jansson, Huaiyang Jiang, Catherine Jordan, Vandita Joshi, James A. Katancik, Wendy A. Keitel, Scott T. Kelley, Cristyn Kells, Susan Kinder-Haake, Nicholas B. King, Rob Knight, Dan Knights, Heidi H. Kong, Omry Koren, Sergey Koren, Karthik C. Kota, Christie L. Kovar, Nikos C. Kyrpides, Patricio S. La Rosa, Sandra L. Lee, Katherine P. Lemon, Niall Lennon, Cecil M. Lewis, Lora Lewis, Ruth E. Ley, Kelvin Li, Konstantinos Liolios, Bo Liu, Yue Liu, Chien-Chi Lo, Catherine A. Lozupone, R. Dwayne Lunsford, Tessa Madden, Anup A. Mahurkar, Peter J. Mannon, Elaine R. Mardis, Victor M. Markowitz, Konstantinos Mavrommatis, Jamison M. McCorrison, Daniel McDonald, Jean McEwen, Amy L. McGuire, Pamela McInnes, Teena Mehta, Kathie A. Mihindukulasuriya, Jason R. Miller, Patrick J. Minx, Irene Newsham, Chad Nusbaum, Michelle O'Laughlin, Joshua Orvis, Ioanna Pagani, Krishna Palaniappan, Shital M. Patel, Matthew Pearson, Jane Peterson, Mircea Podar, Craig Pohl, Katherine S. Pollard, Margaret E. Priest, Lita M. Proctor, Xiang Qin, Jeroen Raes, Jacques Ravel, Jeffrey G. Reid, Mina Rho, Rosamond Rhodes, Kevin P. Riehle, Maria C. Rivera, Beltran Rodriguez-Mueller, Yu-Hui Rogers, Matthew C. Ross, Carsten Russ, Ravi K. Sanka, J. Pamela Sankar, Fah Sathirapongsasuti, Jeffery A. Schloss, Patrick D. Schloss, Thomas M. Schmidt, Matthew Scholz, Lynn Schriml, Alyxandria M. Schubert, Nicola Segata, Julia A. Segre, William D. Shannon, Richard R. Sharp, Thomas J. Sharpton, Narmada Shenoy, Nihar U. Sheth, Gina A. Simone, Indresh Singh, Chris S. Smillie, Jack D. Sobel, Daniel D. Sommer, Paul Spicer, Granger G. Sutton, Sean M. Sykes, Diana G. Tabbaa, Mathangi Thiagarajan, Chad M. Tomlinson, Manolito Torralba, Todd J. Treangen, Rebecca M. Truty, Tatiana A. Vishnivetskaya, Jason Walker, Lu Wang, Zhengyuan Wang, Doyle V. Ward, Wesley Warren, Mark A. Watson, Christopher Wellington, Kris A. Wetterstrand, James R. White, Katarzyna Wilczek-Boney, Yuan Qing Wu, Kristine M. Wylie, Todd Wylie, Chandri Yandava, Liang Ye, Yuzhen Ye, Shibu Yooseph, Bonnie P. Youmans, Lan Zhang, Yanjiao Zhou, Yiming Zhu, Laurie Zoloth, Jeremy D. Zucker, Bruce W. Birren, Richard A. Gibbs, Sarah K. Highlander, George M. Weinstock, Richard K. Wilson, and Owen White. A framework for human microbiome research. *Nature*, 486(7402):215–21, jun 2012. ISSN 1476-4687. doi: 10.1038/nature11209.

Hiromi W. L. Koh, Damian Fermin, Christine Vogel, Kwok Pui Choi, Rob M. Ewing, and Hyungwon Choi. iOmicsPASS: Network-based integration of multiomics data for predictive subnetwork discovery. *npj Systems Biology and Applications*, 5(1):1–10, jul 2019. ISSN 2056-7189. doi: 10.1038/s41540-019-

0099-y.

Carlos Martínez-Mira, Ana Conesa, and Sonia Tarazona. MOSim: Multi-Omics Simulation in R. Preprint, Bioinformatics, sep 2018.

Nicola Segata, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S. Garrett, and Curtis Huttenhower. Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6):R60, Jun 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-6-r60. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-6-r60.

Alissa Sherry and Robin K Henson. Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. 1981.

Mei San Tang, Rowann Bowcutt, Jacqueline M. Leung, Martin J. Wolff, Uma M. Gundra, David Hudesman, Lisa B. Malter, Michael A. Poles, Lea Ann Chen, Zhiheng Pei, Antonio G. Neto, Wasif M. Abidi, Thomas Ullman, Lloyd Mayer, Richard A. Bonneau, Ilseung Cho, and P'ng Loke. Integrated Analysis of Biopsies from Inflammatory Bowel Disease Patients Identifies SAA1 as a Link Between Mucosal Microbes with TH17 and TH22 Cells. *Inflammatory Bowel Diseases*, 23(9):1544–1554, sep 2017. ISSN 1078-0998. doi: 10.1097/MIB.0000000000001208.

Arthur Tenenhaus and Michel Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2):257–284, apr 2011. ISSN 00333123. doi: 10.1007/s11336-011-9206-8.

Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*, 238(2):391–403, oct 2014. ISSN 0377-2217. doi: 10.1016/j.ejor.2014.01.008.

Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583, jul 2014. ISSN 1465-4644. doi: 10.1093/biostatistics/kxu001.

Arthur Tenenhaus, Cathy Philippe, and Vincent Frouin. Kernel Generalized Canonical Correlation Analysis. *Computational Statistics & Data Analysis*, 90:114–131, oct 2015. ISSN 01679473. doi: 10.1016/j.csda.2015.04.004.

Michel Tenenhaus. Component-based Structural Equation Modelling. *Total Quality Management & Business Excellence*, 19(7-8):871–886, aug 2008. ISSN 1478-3363, 1478-3371. doi: 10.1080/14783360802159543.

Michel Tenenhaus and Mohamed Hanafi. A Bridge Between PLS Path Modeling and Multi-Block Data Analysis. In Vincenzo Esposito Vinzi, Wynne W. Chin, Jörg Henseler, and Huiwen Wang, editors, *Handbook of Partial Least Squares*, pages 99–123. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-540-32825-4 978-3-540-32827-8. doi: 10.1007/978-3-540-32827-8_5.

Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, Oct 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3589. URL http://www.nature.com/articles/nmeth.3589.

Roberta De Vito, Ruggero Bellio, Lorenzo Trippa, and Giovanni Parmigiani. Multi-study factor analysis. *Biometrics*, 75(1):337–346, 2019. ISSN 1541-0420. doi: 10.1111/biom.12974.

Cen Wu, Fei Zhou, Jie Ren, Xiaoxi Li, Yu Jiang, and Shuangge Ma. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High-Throughput*, 8(1):4, mar 2019. doi: 10.3390/ht8010004.