Modeling the gut microbial ecosystem from shotgun metagenomics data

Laboratory: UMMISCO, UMI 209, SU/IRD.

Thesis supervisor: Edi Prifti (CR IRD, HDR SU)

Co-advisor: Eugeni Belda (IR IRD)

Doctoral School: EDITE ED 130 of Sorbonne University (SU)

Keywords: *Network Annotation Propagation, Multi-layer networks, data integration, microbial ecosystems, annotations, topologies*

Résumé

Contexte. *Les réseaux sont des abstractions très puissantes. Ils peuvent fournir une représentation holistique des relations (arêtes) entre des objets (nœuds), tout en rendant visibles des modèles et des propriétés émergentes, comme des modules ou des nœuds centraux, qui ne seraient pas visibles autrement. Les nœuds correspondent généralement à des variables observées sur un grand nombre d'échantillons (observations). Les relations peuvent être déduites à l'aide de différentes fonctions, basées sur des connaissances externes mais aussi estimées à partir des données, telles que l'information mutuelle, la corrélation ou même les mesures de distance. Une arête peut représenter par exemple un lien entre deux produits achetés ensemble, ou entre deux auteurs qui ont cosigné un article. Il peut également représenter deux gènes fréquemment coexprimés ensemble ou deux espèces bactériennes dépendantes l'une de l'autre. Différents seuils sont généralement choisis pour limiter le nombre d'arêtes aux plus fortes. Une fois la topologie du réseau définie, différentes métriques peuvent être calculées, telles que l'identification des modules, les mesures de centralité, les chemins les plus courts, etc. Le réseau est ensuite visualisé à l'aide de différentes techniques et algorithmes. Les nœuds et les bords peuvent être annotés à l'aide d'informations externes, afin d'illustrer l'enrichissement des modules ou des modèles donnés.*

Cependant, la topologie du réseau dépend fortement des données et des paramètres du pipeline (distance, seuils, ...). Ce problème est d'autant plus important dans les systèmes biologiques que les technologies de quantification à haut débit sont très sensibles. C'est le cas pour l'expression des gènes, l'abondance des métabolites et même l'abondance des espèces du microbiome. Il existe une multitude d'informations biologiques sur ces objets moléculaires, généralement structurées sous forme d'ontologies dans des bases de données de connaissances spécifiques. Dans le contexte des réseaux, ces informations peuvent être utilisées pour annoter les nœuds et les arêtes, mais aussi pour déduire d'autres types de réseaux où les nœuds peuvent être annotés, comme c'est le cas dans l'approche FunNet (Prifti et al 2008). De tels réseaux multicouches peuvent être importants pour comprendre les mécanismes et les propriétés émergentes des phénomènes étudiés.

Objectifs. *Une approche basée sur la propagation dynamique des annotations a été précédemment introduite dans le contexte des réseaux de co-expression (Prifti et al 2008 ; Prifti et al 2010). L'objectif principal de ce projet doctoral est de proposer une nouvelle méthode, qui intègre les annotations dans une topologie de réseau existante, et permet de générer des réseaux multicouches basés sur ces annotations ontologiques. Différentes approches, y compris l'algorithme de propagation de l'état de l'art, seront explorées et évaluées. L'implémentation qui sera faite devra être capable de traiter tous les types de réseaux et d'annotations et différents jeux de données du référentiel SNAP seront explorés (http://snap.stanford.edu). Cependant, un focus particulier sera mis sur l'utilisation des jeux de données du microbiome avec leurs annotations taxonomiques et fonctionnelles. Plus spécifiquement, nous explorerons le contexte des signatures prédictives du microbiome à l'aide de l'approche predomics (Prifti et al., 2020), dans une perspective d'écosystème global.*

Résultats attendus. *D'un point de vue méthodologique, le résultat attendu est à la fois un cadre de reconstruction de réseau et de nouvelles méthodes algorithmiques qui permettent d'intégrer la topologie du réseau avec les annotations pour inférer des réseaux multicouches. L'implémentation permettra la visualisation et la manipulation des réseaux. Cette approche sera largement évaluée sur des données simulées pour tester sa robustesse au bruit, mais aussi sur différents types de données réelles provenant du dépôt SNAP de Stanford. Une analyse approfondie des données sur le microbiome provenant de différentes études publiques telles que la cirrhose du foie sera effectuée (Qin et al., 2014). Pour les plus grands réseaux, une attention particulière sera accordée à l'efficacité de calcul et éventuellement au calcul par GPU.*

Abstract

Context. *Networks are very powerful abstractions. They can provide a holistic representation of relations (i.e., edges) between objects (i.e., nodes), while making visible patterns and emerging properties such as modules or special central nodes that wouldn't be seen otherwise. Nodes usually correspond to variables observed on a large number of samples (i.e., observations). Relations can be inferred using a plethora of functions, based on external knowledge but also estimated from the data such as mutual information, correlation or even distance metrics. An edge can represent for instance a link between two products bought together, or between two authors who have co-signed a paper. It can also represent two genes that are co-expressed frequently together or two bacterial species that are dependent on one another. Different thresholds are usually chosen to limit the number of edges to the stronger ones. Once the network topology is defined, different metrics can be computed such as the identification of modules, centrality measures, shortest paths, etc. The network is then visualized using different techniques and algorithms. Nodes and edges can be annotated using external information, to illustrate enrichment in modules or given patterns.*

However, the topology of the network is highly dependent on the data and on the pipeline parameters (distance, thresholds, …). This is even more of an issue in biological systems as the high throughput quantification technologies are very sensitive. Such is the case for gene expression, metabolite abundance and even microbiome species abundance. There is a wealth of biological information on such molecular objects, usually structured as ontologies on specific knowledge databases. In the context of networks, this information can be used to annotate the nodes and edges, and also to infer other types of networks where the nodes are can be annotations as is the case in the FunNet approach (Prifti et al 2008). Such multi-layer networks can be important in understanding the mechanisms and emergent properties of the studied phenomena.

Objectives. *An approach based on the dynamic propagation of annotations was previously introduced in the context of co-expression networks (Prifti et al 2008; Prifti et al 2010). The main objective of this doctoral project, is to propose a novel method, which integrates annotations into an existing network topology, and allows to generate multi-layer networks based on such ontology-based annotations. Different approaches including the state-of-the-art propagation algorithm will be explored and evaluated. The framework should be able to process all types of networks and annotations and different dataset from the SNAP repository will be explored (*[*http://snap.stanford.edu*](http://snap.stanford.edu)*). However, a* ***special focus will be put to the use of microbiome datasets*** *along with their taxonomic and functional annotations. More specifically, we will* ***explore the context of predictive microbiome signatures*** *using the predomics approach (Prifti et al., 2020), from a perspective of the whole ecosystem.*

Expected results. *From a methodological perspective, the expected result is both a framework of network reconstruction as well as novel algorithmic methods which allow integrating network topology along with annotations to infer multi-layer networks. The framework will allow network visualization and manipulation. This approach will be largely benchmarked on simulated data to test for its robustness to noise, but also on different types of real datasets from the SNAP Stanford repository. An in-depth analysis of microbiome data from different public studies such as liver cirrhosis will be performed (Qin et al., 2014). For larger networks, special attention will be given to computational efficiency and eventually GPU computing.*

Context and motivation

Human gut microbiome has been associated with a plethora of human complex diseases, including obesity (Cotillard, Kennedy et al. 2013, Le Chatelier, Nielsen et al. 2013), type-2 diabetes (Qin, Li et al. 2012, Karlsson, Tremaroli et al. 2013), liver cirrhosis (Qin, Yang et al. 2014) and many others. The recent advances in sequencing technologies made possible the exploration of this mostly unknown "world". More specifically, quantitative metagenomics (QM) consists on generating millions of genes from each sample, which are aligned against reference catalogues (*i.e.* genes, genomes, MAGs etc…). The reference gene catalogs can be build using specific bioinformatics pipelines (**Figure 1**) and the most recent ones consists of tens of millions of genes (Qin, Li et al. 2010, Li, Jia et al. 2014). A typical study consists of hundreds of samples, usually comparing several groups of patients. Tools and approaches in dimension reduction have been developed and have allowed translating such tables in much smaller ones composed of metagenomic species (MGS) (Nielsen, Almeida et al. 2014).

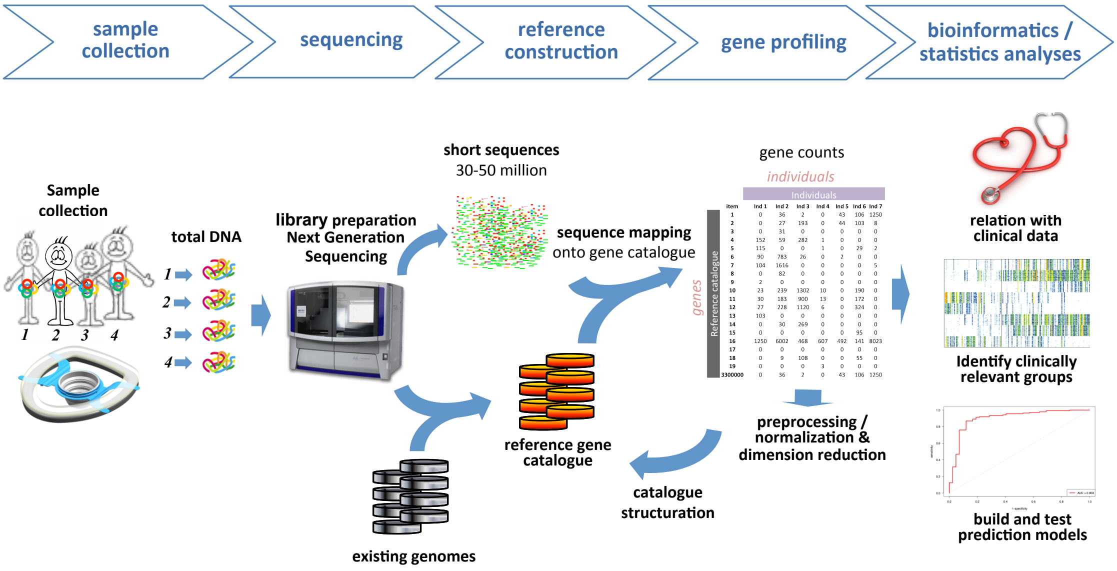


Figure. 1: Overview of a whole-metagenome-sequencing project from sample collection to hypotheses generation.

Several hundreds of different species inhabit the human gut in close relationship with each other where they compete or co-exist in symbiosis in the same niche. Species associated with human disease should be viewed from a broader perspective, that of the whole ecosystem, for a better understanding of the different mechanisms governing such associations. Some preliminary work has been achieved in the lab but also elsewhere using other kind of data (16S). Faust has written a nice review on the subject (Faust, Sathirapongsasuti et al. 2012; Faust and Raes, Nat Rev Microbiol, 2012).

Complex relationships can be explored through the use of comprehensive abstractions such as networks (see Figure 2 for an example). Many tools and methodologies have been developed in the field mostly around biological data such as protein-protein interactions or co-expression data. The use of such analytical frameworks will enable us to explore in depth the gut microbial ecosystem.

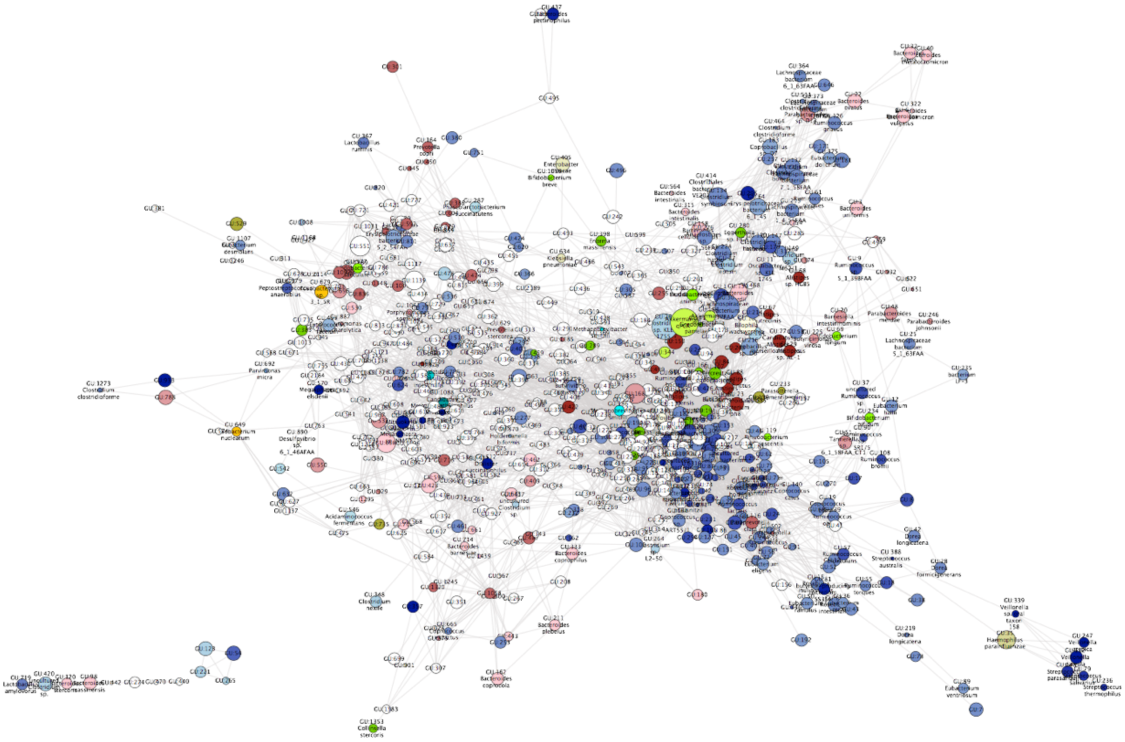
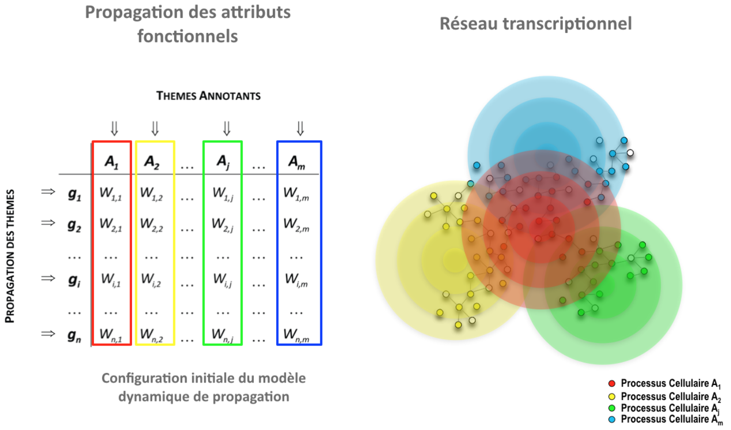


Figure. 2: Example of a co-abundance network of MGS in human gut microbiome data.

Networks are very powerful abstractions. They can provide a holistic representation of relations (i.e., edges) between objects (i.e., nodes), while making visible patterns and emerging properties such as modules or special central nodes that wouldn't be seen otherwise. Nodes usually correspond to variables observed on a large number of samples (i.e., observations). Relations can be inferred using a plethora of functions, based on external knowledge but also estimated from the data such as mutual information, correlation or even distance metrics. An edge can represent for instance a link between two products bought together, or between two authors who have co-signed a paper. It can also represent two genes that are co-expressed frequently or two bacterial species that are dependent on one another on a given ecosystem. Different thresholds are usually chosen to limit the number of edges to the stronger ones. Once the network topology is defined, different metrics can be computed such as the identification of modules, centrality measures, shortest paths, etc. The network is then visualized using different techniques and algorithms. Nodes and edges can be annotated using external information, to illustrate enrichment in modules or given patterns.

However, the topology of the network is highly dependent on the data as well as on the pipeline and the different parameters (distance metrics, filtering thresholds, etc.) used to infer it. The topology can eventually influence the interpretation of the results, including the identification of key nodes. This is even more of an issue in biological systems as the high throughput technologies allowing to quantify the abundance of molecular objects are very sensitive. Such is the case for gene expression, metabolite abundance and even microbiome species abundance to name a few. There is a wealth of biological information on such molecular objects, usually structured as ontologies on specific knowledge databases. In the context of networks, this information can be used to simply annotate the nodes and edges, but also to infer other types of networks where the nodes are can be annotations as is the case in the FunNet approach (Prifti et al 2008). Such multi-layer networks can be very important in understanding the mechanisms and emergent properties of the studied phenomena. An approach based on the dynamic propagation of annotations was previously introduced in the context of co-expression networks (Prifti et al 2008; Prifti et al 2010). Figure 1 illustrates the concept behind the propagation approach (Figure 1).


Propagation des attributs fonctionnels
Réseau transcriptionnel
Themes Annotants
Propagation des themes
Configuration initiale du modèle dynamique de propagation
● Processus Cellulaire A_1
● Processus Cellulaire A_2
● Processus Cellulaire A_j
● Processus Cellulaire A_m

		A_1	A_2	…	A_j	…	A_m
⇒	g_1	$W_{1,1}$	$W_{1,2}$	…	$W_{1,j}$	…	$W_{1,m}$
⇒	g_2	$W_{2,1}$	$W_{2,2}$	…	$W_{2,j}$	…	$W_{2,m}$
	…	…	…	…	…	…	…
⇒	g_i	$W_{i,1}$	$W_{i,2}$	…	$W_{i,j}$	…	$W_{i,m}$
	…	…	…	…	…	…	…
⇒	g_n	$W_{n,1}$	$W_{n,2}$	…	$W_{n,j}$	…	$W_{n,m}$

Figure. 1: Schematic of the annotation propagation process in network abstractions

The human gut microbiome has been recently associated with a large and growing number of human complex diseases as for instance obesity (Cotillard, Kennedy et al. 2013, Le Chatelier, Nielsen et al. 2013), type-2 diabetes (Qin, Li et al. 2012, Karlsson, Tremaroli et al. 2013), liver cirrhosis (Qin, et al. 2014) and many others. The recent advances in sequencing technologies made possible the exploration of this mostly unknown “world” through the new and hot field of metagenomics. Quantitative metagenomics (QM) consists of measuring the abundance of each gene from a reference metagenome and gene groups named MGS (metagenomic species) (Nielsen, Almeida et al. 2014) representing the ecosystem of interest. **Several hundreds of different species inhabit the human gut in close relationship with each other where they compete or coexist in symbiosis in the same niche** (Faust, Sathirapongsasuti et al. 2012; Faust and Raes, Nat Rev Microbiol, 2012)**. These relations can be quantified and modeled with network abstractions** (see Figure 2).

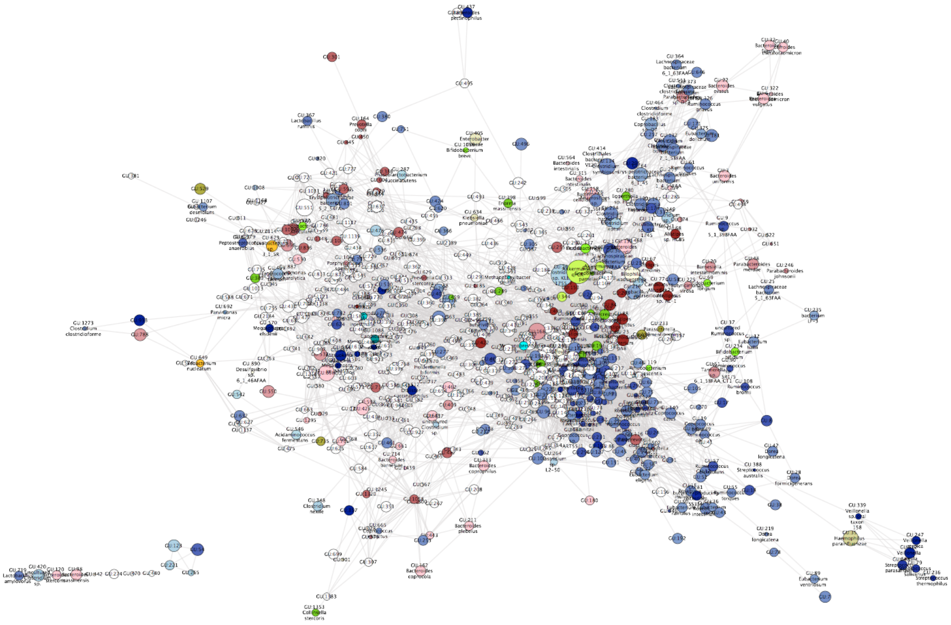


Figure. 2: Example of a co-abundance network of MGS in human gut microbiome data. Nodes are annotated (colored) by taxonomic information and size of the MGS (node size).

Understanding the microbial ecosystem is a complex research task since very little is known on the relations between bacterial species, the functions they participate to and the microbial species where the genes are physically located and expressed. For a better exploration and understanding of this ecosystem and its impact on the human host we have developed novel network abstractions, such as Scalenet (Affeldt et al 2018), that allow building a more robust topology of the ecosystem network.

Objectives

The main objective of this doctoral project, is to **propose a novel end-to-end method which reconstructs interaction networks from different types of OMICs data, integrates annotations into an existing network topology, and allows to generate multi-layer networks based on such ontology-based annotations**. For network reconstruction, different feature x feature interaction metrics (pairwise correlations, (di)similarity-based metrics, mutual information criteria or beta-coefficients product of pairwise linear regression) will be combined to estimate robust measures of interaction (between genes, species, functions, etc.) that will be the basis for network reconstruction with algorithms like Scalenet mentioned before. With the reconstructed networks, different approaches including the state-of-the-art propagation algorithms and novel deep-learning based approaches like Graph Neural Networks (GNN) will be explored and evaluated for predicting functional associations between taxonomic groups, and exploring the use of annotation propagation techniques (message passing, attention mechanisms, label propagation) based on different GNN architectures to infer functional annotations for uncharacterized taxa. Finally, the modularity of the multi-layer networks will be explored by different network decomposition algorithms (MCL, edge-betweenness, fast greedy, etc) in order to identify densely connected areas of the network reflecting potential signatures of complex multi-feature interactions and co-occurrence patterns. The framework **should be able to process all types of networks and annotations** and different dataset from the SNAP repository will be explored (http://snap.stanford.edu). However, a **special focus will be put to the use of microbiome datasets** along with their taxonomic and functional annotations to gain a better understanding of microbial ecosystem functioning by integrating taxonomic and functional data in a multi-layer network context. More specifically, we will capitalize on available large-scale cross-sectional studies like Metacardis to reconstruct these multi-layer networks in different clinical groups representing different degrees of severity in cardiometabolic and cardiovascular diseases and study the main topological differences between the disease-specific networks. Also, by capitalizing on longitudinal metagenomic datasets (Microbaria and Microobes cohorts) we will explore the topological changes in these multi-layer networks derived from nutritional interventions (Microobes study) or bariatric surgery interventions (Microbaria study) to gain insights in the main drivers of ecosystem change. Finally, we will **explore the context of predictive microbiome signatures** using the predomics approach (Prifti et al., 2020) and what functional knowledge could be gained from the multi-layer network topologies vs. an approach based on the metabolic modelling of microbial communities to interpret the mechanistic relevance of the retrived signatures from a perspective of the whole ecosystem.

Expected results

From a methodological perspective, the expected result is both **a framework of network reconstruction as well as novel algorithmic methods which allow integrating network topology along with annotations to infer multi-layer networks**. The framework will allow network visualization and manipulation. This approach will be largely benchmarked on simulated data to test for its robustness to noise, but also on different types of real datasets from the SNAP Stanford repository. An in-depth analysis of microbiome data from different public studies such as Metacardis, Microbaria Microobes. For larger networks, special attention will be given to computational efficiency and eventually GPU computing.

References

Prifti, Edi, Jean-Daniel Zucker, Karine Clement, and Corneliu Henegar. “FunNet: An Integrative Tool for Exploring Transcriptional Interactions.” *Bioinformatics* 24, no. 22 (November 15, 2008): 2636–38. https://doi.org/10.1093/bioinformatics/btn492.

Prifti, Edi, Jean-Daniel Zucker, Karine Clément, and Corneliu Henegar. “Interactional and Functional Centrality in Transcriptional Co-Expression Networks.” *Bioinformatics* 26, no. 24 (December 15, 2010): 3083–89. https://doi.org/10.1093/bioinformatics/btq591.

Cotillard, Aurélie, Sean P. Kennedy, Ling Chun Kong, Edi Prifti, Nicolas Pons, Emmanuelle Le Chatelier, Mathieu Almeida, et al. “Dietary Intervention Impact on Gut Microbial Gene Richness.” *Nature* 500, no. 7464 (August 29, 2013): 585–88. https://doi.org/10.1038/nature12480.

MetaHIT consortium, Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, et al. “Richness of Human Gut Microbiome Correlates with Metabolic Markers.” *Nature* 500, no. 7464 (August 29, 2013): 541–46. https://doi.org/10.1038/nature12506.

Karlsson, F. H., V. Tremaroli, I. Nookaew, G. Bergstrom, C. J. Behre, B. Fagerberg, J. Nielsen and F. Backhed (2013). "Gut metagenome in European women with normal, impaired and diabetic glucose control." Nature 498(7452): 99-103.

Qin, Nan, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, et al. “Alterations of the Human Gut Microbiome in Liver Cirrhosis.” *Nature* 513, no. 7516 (September 4, 2014): 59–64. https://doi.org/10.1038/nature13568.

MetaHIT Consortium, H Bjørn Nielsen, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, et al. “Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes.” *Nature Biotechnology* 32, no. 8 (August 2014): 822–28. https://doi.org/10.1038/nbt.2939.

Faust, K., J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes and C. Huttenhower (2012). "Microbial co-occurrence relationships in the human microbiome." PLoS Comput Biol 8(7): e1002606.

Faust, Karoline, and Jeroen Raes. “Microbial interactions: from networks to models.” Nature Reviews Microbiology 10, no. 8 (August 1, 2012): 538–550.

Prifti, Edi, Yann Chevaleyre, Blaise Hanczar, Eugeni Belda, Antoine Danchin, Karine Clément, and Jean-Daniel Zucker. “Interpretable and Accurate Prediction Models for Metagenomics Data.” *GigaScience* 9, no. 3 (March 1, 2020): giaa010. https://doi.org/10.1093/gigascience/giaa010.