PhD- FSTC-2019-72
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 12/11/2019 in Esch-sur-Alzette
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Federico BALDINI
Born on 21 November 1991 in Florence (Italy)

DEVELOPING INDIVIDUAL-BASED GUT MICROBIOME METABOLIC MODELS FOR THE INVESTIGATION OF PARKINSON'S DISEASE-ASSOCIATED INTESTINAL MICROBIAL COMMUNITIES

Dissertation defence committee

Dr Ines Thiele, dissertation supervisor
Professor, National University of Ireland Galway

Dr Rejko Krüger, Chairman
Professor, Université du Luxembourg

Dr Ullrich Wüllner
Professor, UKB University of Bonn and DZNE Bonn

Dr Jean-Pierre Trezzi, Vice Chairman
Scientist, Luxembourg Institute of Health

Molecular Systems Physiology

Luxembourg Centre for Systems Biomedicine

Faculty of Life Sciences, Technology and Communication

Doctoral School in Systems and Molecular Biomedicine

Disseration Defence Committee:

Committee members:	Prof. Rejko Krüger
	Dr. Jean-Pierre Trezzi
	Prof. Ullrich Wüllner
Supervisor:	Prof. Ines Thiele

I hereby confirm that the PhD thesis entitled “Developing individual-based gut microbiome metabolic models for the investigation of Parkinson’s disease-associated intestinal microbial communities” has been written independently and without any other sources than cited.

Luxembourg, ______________ ____________________________

Federico Baldini

Acknowledgments

Firstly, I would like to thank all the people that directly and indirectly helped me during these last four years of PhD. I want to thank my PhD supervisor, Prof. Ines Thiele who saw some potential in me when I still was a master student and gave me this incredible scientific opportunity. Ines supported me during all these years, constantly believing in my capabilities, and greatly contributing to my scientific growth and maturity and my career. I also want to acknowledge Prof. Rejko Krüger who hosted me in his lab during the last eight months of PhD, giving me essential support and means for developing our project on Parkinson's disease and allowing me to get closer to the clinical research. I own special thanks to Prof. Andrea Galli and Dr. Marco Fondi for their valuable advice and their time spent listening to me. I would like to thank my friends and former colleagues Dr. Eugen Bauer and Dr. Marouen Ben Guebila, their support and advice were essential for letting me progress and advance during these last four years. Eugen and his way of conducting research were and still remain a great source of inspiration in my scientific career. I admire his incredible capacity to challenge scientific problems in complete autonomy and his interdisciplinarity being able to conduct projects collecting knowledge from different fields. Having scientific discussions with Marouen and Eugen was awesome, and many times when I was stuck, I was able to progress following their suggestions or ideas as an outcome of our discussions. Moreover, I would like to acknowledge my colleagues at the Luxembourg Centre for Systems Biomedicine (LCSB) and at the Parkinson's Research Clinic (PRC) for creating a nice and collaborative environment. At LCSB, I especially want to acknowledge the members of the Molecular Systems Physiology (MSP) and the Clinical and Experimental Neuroscience (CEN) groups of which I was proudly member.

Within my friends, besides Eugen and Marouen, I would like to especially thank Mattia

Graziani and Francesco Bonfiglio for psychologically supporting me and in some moments giving me precious suggestions during these last four years. The long walks and discussions we had with Mattia were instrumental for releasing the tension and "refreshing" my mind, as well as the dinners and Skype calls I had with Francesco. More in general, I would like to thank all my friends (Devrim, Iren, Lorenzo, Benedetta, Alessio, Stefania, Grazia, Sofia). I also want to thank my girlfriend, Raluca Bujoreanu. Raluca greatly supported me during this last and intense year of PhD being always there to listen, suggest, psychologically help me, and her support became essential for me being able to complete this PhD thesis. Last but not least; I would like to thank my parents Stefano Baldini and Francesca Brienza for continuously helping, advising and supporting me during all my studies: the work presented in this thesis is also a result of their involvement in my studies and an indirect result of their efforts as parents.

Contents

1 Introduction 3

1.1 The Human gut microbiota composition and health 3

1.1.1 The human gut microbiota and Parkinson's Disease 5

1.2 Sequencing of microbial communities 7

1.3 Reconstruction of Genome-scale metabolic networks (GEMs) 9

1.4 Flux Balance Analysis 12

1.4.1 Modeling (human gut) microbial communities with constraint-based modeling 15

1.5 Modeling challenges, conclusions, and future perspectives 23

1.6 Scope and aim of the thesis 25

2 The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities 29

2.1 Introduction 30

2.2 Features 30

2.3 Implementation 32

2.4 Discussion 32

2.5 Funding 33

2.6 Documentation 33

3 mgPipe: from microbial community sequencing data to personalized microbiota metabolic models creation and interrogation 41

3.1 Introduction 42

3.2 Features 43

3.3 Implementation 43
3.4 Discussion 45
3.5 Funding 45
3.6 Documentation 45

4 Parkinson's disease-associated alterations of the gut microbiome can invoke disease-relevant metabolic changes 53

4.1 Introduction 54
4.2 Results 55
4.2.1 Species and genus level changes in PD microbiomes 55
4.2.2 PD modifies the effects of basic covariates on the microbiome . . . 57
4.2.3 Microbial abundances, medication intake, and constipation in PD . 58
4.2.4 Genus association with the disease severity 59
4.2.5 Metabolic modelling reveals distinct metabolic secretion capabilities of PD microbiomes 59
4.2.6 PD specific secretion profiles were altered due to changed community structure and species abundances 60
4.3 Discussion 61
4.3.1 Strength and limitation 66
4.4 Methods 67
4.4.1 Description of the Luxembourg Parkinson's study 67
4.4.2 Measurements and neuropsychiatric testing 68
4.4.3 Collection and processing of stool samples 68
4.4.4 Analysis of the microbial composition with 16S rRNA gene sequencing 69
4.4.5 Personalised constraint-based modelling of microbial communities . 69
4.4.6 Analyses of relative abundances 71
4.4.7 Differences between PD and controls in microbial composition and the influence of covariates 72
4.4.8 Analyses of within PD phenotypes in relation to microbial composition 73
4.4.9 Statistical analyses of fluxes 74
4.4.10 Analyses of species contribution to fluxes 74

4.4.11 Material availability . . . 75
4.4.12 Acknowledgment . . . 75

5 Concluding remarks 79

5.1 Sequencing data integration and COBRA modeling . . . 81
5.2 Parkinson's disease and personalized microbiota modeling . . . 82
5.3 Future perspectives . . . 83

A Supplementary material for Chapter 2 103

A.1 Microbiome Modeling Toolbox tutorials . . . 103

B Supplementary material for Chapter 3 113

B.1 mgPipe tutorial . . . 113

C Supplementary material for Chapter 4 123

C.1 Extended information on the simulations . . . 123
C.1.1 Mapping detected species on the gut microbial reconstruction collection 123
C.1.2 Generation of personalized models . . . 123
C.1.3 Supplementary figures . . . 124
C.2 Supplementary tables . . . 125
C.2.1 Supplementary table 1 . . . 125
C.2.2 Supplementary table 2 . . . 125
C.2.3 Supplementary table 3 . . . 125
C.2.4 Supplementary table 4 . . . 125

List of Figures

1.1 Metabolic modeling. 12

1.2 Community modeling techniques. 22

2.1 Overview of the Microbiome Modeling Toolbox. 31

3.1 Overview of the mgPipe features. 44

4.1 Overview of the study approach and the key methods used. 56

4.2 Boxplots of seven significantly changed species in PD versus controls. . . . 58

4.3 Genus alterations in PDs due to interactions with basic covariates. 76

4.4 Result of analysing secretion profiles of microbial communities. 77

4.5 Reported microbial changes at the family level associated with PD in different studies. 78

C.1 Lactobacillus relative abundance positively correlated with the duration of the disease. 124

List of Tables

4.1 Descriptive statistics of the analyses sample from the Luxembourg Parkinson's Disease study and overview over associations. 57

C.1 Number of species present in AGORA in dependency on the number of detected species. 123

C.2 Personalized model characteristics. 124

List of Abbreviations

AGORA	Assembly of gut organisms through reconstruction and analysis
CNS	Central nervous system
COBRA	Constraint-based reconstruction and analysis
CRC	Colorectal cancer
DNA	Deoxyribonucleic acid
ENS	Enteric nervous system
FBA	Flux balance analysis
FDR	False discovery rate
FMT	Fecal microbiota transplant
FVA	Flux variability analysis
GABA	*gamma*-Aminobutyric acid
GEM	Genome-scale metabolic network
HMP	NIH Human Microbiome Project
IBD	Inflammatory bowel disease
IBM	Individual based modeling
OTU	Operational taxonomic unit
PCoA	Principle coordinate analysis
PCR	Polymerase chain reaction
PD	Parkinson's disease
PNS	Peripheral nervous system
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SCFA	Short chain fatty acid
VMH	Virtual metabolic human
WBM	Whole-body metabolism reconstructions
WGS	Whole genome sequencing

Summary

The human phenotype is a result of the interactions of environmental factors with genetic ones. Some environmental factors such as the human gut microbiota composition and the related metabolic functions are known to impact human health and were put in correlation with the development of different diseases. Most importantly, disentangling the metabolic role played by these factors is crucial to understanding the pathogenesis of complex and multifactorial diseases, such as Parkinson's Disease. Microbial community sequencing became the standard investigation technique to highlight emerging microbial patterns associated with different health states. However, even if highly informative, such technique alone is only able to provide limited information on possible functions associated with specific microbial communities composition. The integration of a systems biology computational modeling approach termed constraint-based modeling with sequencing data (whole genome sequencing, and 16S rRNA gene sequencing), together with the deployment of advanced statistical techniques (machine learning), helps to elucidate the metabolic role played by these environmental factors and the underlying mechanisms.

The first goal of this PhD thesis was the development and deployment of specific methods for the integration of microbial abundance data (coming from microbial community sequencing) into constraint-based modeling, and the analysis of the consequent produced data. The result was the implementation of a new automated pipeline, connecting all these different methods, through which the study of the metabolism of different gut microbial communities was enabled. Second, I investigated possible microbial differences between a cohort a Parkinson's disease patients and controls. I discovered microbial and metabolic changes in Parkinson's disease patients and their relative dependence on several physiological covariates, therefore exposing possible mechanisms of pathogenesis of the disease.

Overall, the work presented in this thesis represents method development for the investigation of before unexplored functional metabolic consequences associated with microbial changes of the human gut microbiota with a focus on specific complex diseases such as Parkinson's disease. The consequently formulated hypothesis could be experimentally validated and could represent a starting point to envision possible clinical interventions.

Chapter 1

Introduction

1.1 The Human gut microbiota composition and health

The human gut microbiota is composed by an ensemble of trillions of microorganisms of different species living in the human gut: these microorganisms complement the host with essential functions forming, together with it, a superorganism [189, 65, 165, 60, 169].Thanks to advance in genomic sequencing technologies and several cohort studies carried by different investigators and big consortia (MetaHit and HMP), our knowledge on the human gut microbiota greatly improved in the last years [57, 65]. The human gut microbiota is principally colonized by organisms of the taxa of Bacteroides or Firmicutes [187] and its composition was found to strongly differ between individuals [187]. Nonetheless to this day, the presence of three microbial enterotypes has been identified: one mainly dominated by Prevotellaceae the others by Bacteroides or Ruminococcus [10]. However, diet, ethnic group, type of birth, genetic background, and lifestyle (usage of antibiotics, drugs, ext.) are all factors known to influence the microbiota composition [200].

The establishment of a human gut microbial community is a progressive process [141]. Independently of whether the amniotic liquid would be sterile or not, still an argument of active discussion [98], after a phase of initial colonization by aerobic bacteria, anaerobic ones start their colonization process creating anaerobic dominant microbial communities [141]. The initial colonization process is highly dependent on the delivery mode, with naturally born infants microbiomes dominated mostly by vaginally present microbes in opposition with c-section ones dominated by skin microbes [5, 52]. Overall, the gut microbiota is known to change its composition and richness during the first years of life converging to a more stable composition in adulthood under the effect of the different aforementioned factors [5, 107, 190]. Finally, studies on cohorts of elderly people [42] demonstrated a loss of microbial diversity with changes in the ratio Bacteroidetes Firmicutes, a decreased presence of bifidobacteria, and bigger microbial compositional variations between elderly people compared to the younger controls [42]. Furthermore, the use of antibiotics is known to impact microbial compositions being associated with a decrease in bacterial diversity and with reduced resistance to colonization by other species [48, 97, 175]. The usage of antibiotics is suspected to increase the pool of antibiotic-resistant genes [171].

The crucial role of the human gut microbiota for its host homeostasis is nowadays widely recognized and several studies have associated different diseases (such as obesity, diabetes, irritable bowel disease, colorectal cancer, ext.) with changes of the human gut microbiota composition [43]. Diet also has an important role, being able to impact the human metabolism and the gut microbiota composition [188]. More specifically, loss or gain of the diversity of gut microbiota composition as a consequence of different diets and lifestyles are associated with different diseases such as allergy or atopic diseases [146], enrichment of specific taxa with dysmetabolism and insulin resistance [152], and higher presence of specific families such as fusobacteriaceae with colorectal cancer (CRC) [108, 156]. By consequence, a classifier based on microbial abundance has been proved to be able to identify CRC patients from healthy controls [205]: the accuracy was the same as the fecal occult blood test (FOBT) which is currently used to diagnose CRC [205]. Moreover, corroborating the role of lifestyle in shaping the human gut microbiota, it has been shown that athletes host different microbial compositions then sedentary people [15]. The ratio Bacteroides Firmicutes results altered in obese people with a strong link between human gut microbiota composition and gain of weight also demonstrated by in vivo studies [187]. The existence of an obese enterotype was proved with fecal microbiota transplants (FMTs), transplants from obese mice donors lead to gain in weight in germ-free (GF) mice in contrast to what was happening when the microbiota from lean ones was transplanted in GF [187].

The human gut microbiota has many functions form energy related to helping the development of the host immune system and forms with the human host many so-called axes [130]. Interestingly, one of the main functions of the human gut microbiota is related to metabolic capabilities and the metabolic axis: a bi-directional interaction between several microbial communities and the human host with microbial species being able to modulate and, in some cases, enrich metabolic reactions and metabolic pathways of the human [43]. Several bacterial species, for example, are able to transform indigestible dietary fibers, which would not be metabolized by the host, into beneficial compounds such as butyrate and other short-chain fatty acids (SCFAs) that can be later on uptaken by enterocyte cells [195]. These compounds can benefit the host with different mechanisms such as, improving glucose tolerance, modulating the immune response and appetite (via leptin secretion) [202]. Butyrate, especially, is known for being a beneficial compound given, for example, its anti-inflammatory properties and its excellent usage as an energy source by colonocytes [66, 26, 135, 195]. Many diseases have been correlated with a lack of butyrate such as irritable bowel disease (IBD) [197]. Additionally, the human gut microbiota has been identified as a possible source of B vitamins [111]. Moreover, many compounds produced by the metabolism of microbial species or used as signaling compounds (ex: quorum sensing) by these microbial communities can be uptaken by the host and influence cognition process through the gut-brain axis [123]. Examples of such compounds include neurotransmitters and precursors such as catecholamine, serotonin, and tryptophan [137]. These metabolites can impact the host using different possible ways, locally using host receptors, endocrinaly (portal vein), and in a neurocrine way through the vagus nerve. Gamma-aminobutyrate (GABA) is an excellent example of that as different studies investigated its role in emotion and pathologies such as major depressive disorder [173].

1.1.1 The human gut microbiota and Parkinson's Disease

Parkinson's Disease (PD) is a complex disease where the final result is given by the interaction of genetic factors with environmental ones [99]. While several studies were able to elucidate the role of genetic factors in the pathogenesis of the disease [129, 136, 109] the role and the contribution of different environmental factors are still not fully characterized.

Environmental factors have a high impact on the development of the disease, for example, exposure of humans to certain pesticides has been positively associated with PD [178]. As an example, gut intake of a pesticide, known as rotenone, has been proved to be able to produce neurodegeneration and symptoms similar to the ones present in PD on animal models [143]. The gut is thought to play a crucial role in the pathogenesis of PD: the connection between the human gut and the development of Parkinson's disease is explained by the Braak staging system [28, 29]. The encephalic nerve number X, also known as "nervus vagus," which originates in the mesencephalon and has parasympathetic control over a big number of organs (heart, lungs, gut, ext.), represents a physical link between the human gut, the enteric nervous system (ENS) and the central nervous system (CNS). The current opinion is that the neurodegeneration associated with the development of the disease might start in the olfactory bulb as well as in the enteric nervous system, and from there, passing from the dorsal motor nucleus of the vagus nerve, backpropagate to the CNS [28, 29]. As a matter of fact, alpha-Synuclein was found in the peripheral nervous system (PNS) of the gut and proved able to spread from ENS to CNS [62]. An increased gut permeability (regulated by enteric glial cells) positively correlated with intestinal levels of alpha-synuclein [62]. Notably, confirming the involvement of the gut organ, nonmotor symptoms of Parkinson disease such as constipation [56, 40], can proceed of many years (over decades) the motor symptoms [1, 38], that become evident when great part of the dopaminergic neurons have been through the neurodegenerative phase [7]. Additional studies on autopsies of people with reported constipation in their late years of life showed the presence of Lewy bodies in the substantia nigra even though no evidence of parkinsonism or dementia was reported [2, 148]. Finally, it has been proved that individuals with reduced bowel movements (less than one per day) were at higher risk of developing the disease, with constipation being also the possible consequence of neurodegeneration of ENS [149].

Recent studies [21, 83, 162, 14, 87, 91, 102] (Figure 4.5) have also shown an altered gut composition in the presence of Parkinson's disease. The first study that found alterations in the human gut microbial composition was conducted on a cohort of 144 individuals (72 patients and 72 controls) using 16S rRNA gene sequencing, and identified a reduction of 77.5 % of *Prevotellaceae* in PD patients [162]. Another of these studies [21] was realized conducting whole genome sequencing on stool samples of a population composed of 59 males, in the same range of age, early diagnosticated, and naive from any levodopa treatment. Different abundances of *Verrucomircobiaceae* (*Akkermanisa muciniphila*), Firmicutes, *Prevotellaceae*, and *Eryspelotrychaceae* were found between healthy controls and PDs [21]. Many of these findings were also reported by the other studies before mentioned conducted with 16S rRNA sequencing of stool samples [83, 162, 14, 87, 91, 102]. Among all the reports, increased abundance of *Lacotbaicllaceae* and *Verucomicrobiaceae* were the most common findings.

In conclusion, an altered composition of the human gut microbiota was found in PD patients including early stage and levodopa naive ones. However, if many of the conducted studies (Figure 4.5) seem to partially overlap in their conclusions contributing to the rising of an overall common picture, the average small number of sample size and the variety in the methods used, such as different primers in 16S rRNA gene sequencing, make results difficult to compare. Furthermore, little attention was given, so far, to the functional aspect of the matter. Further studies will need to focus on bigger number of patients, more complex cohorts, and on the functional detail.

1.2 Sequencing of microbial communities

Microbial sequencing became a key tool to understand microbial compositions of samples. Depending on the investigator's needs, targeted or untargeted approaches can be used. Quantitative PCR is a targeted approach and can be used to quantify one or a short a series of selected microorganisms and among its advantages cheapness and simplicity are enumerated [41]. Untargeted techniques are more expensive but they aim at giving more comprehensive information on microbial communities compositions and, depending on the cases, functionalities. In untargeted techniques, the main objective of sequencing is to retrieve the microbial composition of specific samples [199]. To this day, two main untargeted techniques have been used: whole-genome sequencing (WGS), and marker gene analysis, such as 16S ribosomal RNA (rRNA) gene sequencing [41].

Whole-genome sequencing (WGS) marked the birth of metagenomics which is a discipline that studies genomic sequence obtained from any specific environment [199]. In WGS, all the extracted DNA from a sample is sequenced obtaining short reads. These reads can be assembled into contigs which are longer sequences that, according to the case, are long enough to also cover complete operons or gene coding regions. The process of "joining" these reads, extending longer sequences (contigs), is called assembly. Assemblies can be obtained via comparison to a reference genome or *de novo*. Reference assemblers that can construct contigs mapping reads onto a reference genomes and are faster but limited by the available reference genomes and their quality, while *de novo* assemblers, not being limited by a comparison with a reference database, can be used to discover novel organisms [127] but are more computationally expensive [41]. Finally, different contigs can be aligned and merged to form long continuous DNA sequences named scaffolds during a step called scaffolding [41]. The last essential step to obtain taxonomic classification is called binning. During this step, through dedicated software, either on the base of their DNA composition (TETRA, Kraken) [199, 179], either on the base of gene homology (MetaPhlAn2) [185] reds or assembled contigs are grouped together by their likely host genomes, assigning taxonomy.

Marker gene analysis, commonly using 16S rRNA gene amplicons, representing the other commonly used untargeted method, has its rationale in sequencing only a specific gene instead than attempting to sequence all the possible genes [41]. The 16S ribosomal RNA is composed by nine variable regions and in 16S rRNA gene sequencing multiple variable regions of the 16S rRNA gene are sequenced [41]. After preprocessing, reads can be clustered in Operational Taxonomic Units (OTUs): sequences that have a similarity score higher than 97% and used to classify closely related individuals. This clustering can be *de novo* or can happen through the usage of specific reference databases [41]. After the previous steps, different classifiers can be then used to assign taxonomy to the different OTUs or sequences [41]. To obtain taxonomic classification reference amplicon databases are needed such as SILVA [153], Greengenes [47], and RDP [121], and among the most common classifiers, Mothur [163] and RDP-Classifier [196] can provide taxonomic classification to the family and occasionally genus level, while other classifiers such as SPINGO [8] and UTAX [55] can provide species-level taxonomic classification. As sequencing errors are likely to affect OTU clustering alternatives are available [41]: DADA2 [34] can offer a full workflow from quality filtering to taxonomic classification without the usage of OTUs. Besides DADA2, several other pipelines such as MeFit [145], QIIME [35] and Mothour [163] exist and can provide, combining different tools, many bioinformatics operations for marker gene analysis.

In conclusion, microbial sequencing is essential for understanding microbial presence and composition in different samples. Different techniques exist and WGS and marker gene analysis can both provide very useful insights. Both techniques have their pros and cons, for example, marker gene analysis has the benefit of being cheaper than WGS in terms of sequencing costs, easier to process, and less computational expensive allowing the possibility of screening big cohorts. However, this approach is limited in taxonomic resolution and faces other issues related to copy number [4], given that organisms have multiple copies of the marker gene. Moreover, the presence of possible targets of different regions, made, over the years, results from different investigations difficult to compare [41]. On the other hand, WGS techniques provide a better taxonomic resolution and the possibility of conducting functional analysis on the basis of the sequenced genes but are expensive in terms of sequencing and computational effort. Anyhow, the availability of microbial sequencing data and related cohort studies posed the basis for the development of methodologies to functionally analyze and understand enumerative information coming from the sequencing. In this regard, metabolic modeling represents an attractive possibility to add the functional dimension to these microbial sequencing datasets.

1.3 Reconstruction of Genome-scale metabolic networks (GEMs)

A genome-scale metabolic reconstruction consists of the list of all the reactions with related stoichiometry and directionality that are known to be happening in an organism [138]. As a matter of fact, from the genomic annotation for each gene and the corresponding enzyme/s is possible to obtain a list of the reactions coded by different genes (Figure 1.1A). These different reactions can be connected together creating a network representing the metabolism of a specific organism (Figure 1.1B)[138]. Several tools (kbase, ModelSEED) to this day, allow the automated creation of "draft" metabolic reconstructions from specific genomes of eukaryotes or prokaryotes [9, 37, 50]. However, such automated networks are not immune to problems related to genome annotation databases [61] and many times cannot cover organism-specific properties such as directionality of reactions and cofactor utilization [180]. Therefore a phase of manual curation has to be carried to reconstruct high-quality prokaryotic and eukaryotic metabolic networks and a dedicated protocol exists for several years [180]. Recent advances [117] were able to semi-automate part of the process which was summed up in a protocol and consists of five steps [180]: 1) Draft reconstruction 2) Refinement of the reconstruction 3) Conversion of reconstruction into computable format 4) Network evaluation 5) Data assembly and dissemination. The COBRA Toolbox [84] is a MATLAB toolbox providing the tools required for this process. The first four steps are continuously reiterated until the metabolic predictions match the phenotypic characteristics of the organism. During the first step, starting from the genome of an organism and through the usage of specific biochemical databases, a draft reconstruction is obtained (Figure 1.1A). The second phase involves the curation of the retrieved network using the available literature [180]. The retrieved metabolic functions need an individual evaluation. This is needed considering that i) not all the annotations have high confidence scores, and ii) the organism-unspecificity of most biochemical databases that might enumerate enzyme activities proper of different organisms instead of only the ones that are present in the target organism [180]. In fact, using the lowest confidence score, it is possible to use phylogenetically close organisms to fill possible metabolic functions, whenever information of the specific organism is not available [180]. The second phase also includes metabolites formula charging, checks, and formulations on substrate and cofactors usage, reaction stoichiometry and directionality, biomass reactions, and medium growth requirements [180]. The third step is automated through the use of the available functions in the COBRA Toolbox [84] and involves loading the reconstruction into MATLAB, setting an objective function, and transforming the reconstructions into a condition-specific ones setting specific simulation constraints (see chapter on "Flux Balance Analysis" section 1.4) [138, 180]. The metabolic content of the reconstruction is computationally converted into the stoichiometric matrix (S matrix, Figure 1.1C) [138]. In the S matrix rows represent the metabolic content of the reconstruction while each column a different reaction. The S matrix is not a binary matrix and each row entry can have a positive or negative value according to the stoichiometry of each reaction and its directionality. The fourth step "consists of the network verification, evaluation, and validation" [180]. During this phase, an extensive gap-filling process is carried and metabolic dead ends are identified. Gaps are filled through literature knowledge, genome re-annotation and for modeling purposes (e.g.,when a metabolite is essential for the biomass production) assigning to the functionality the lowest confidence score [180]. Moreover, the production of all the biomass precursors is checked and enforced when needed [180]. Finally, predicted physiological properties are compared with observed properties testing the reconstruction accuracy [180].

GEMs can be reconstructed for prokaryotes as well as for eukaryotes. One of the main differences is the number of compartments present. Two main compartments are generally present in reconstructions of prokaryotes an extracellular one and an intracellular one. Compounds uptake rates recapitulating utilization of carbon sources and media needs to be set through setting boundaries of some specific reactions called exchange reactions [138]. Metabolites can be uptaken in the extracellular environment and then, through the addition of some transport reactions, can be transported in the intracellular compartment where most of almost the entirety of the metabolic reactions can happen [138]. Eukaryotic cells maintain a similar structure, but additional compartments are present, representing the different organelles present in the cell. Recon is a global human metabolic model recapitulating human metabolism and the related properties [53, 33]. Using dedicated algorithms to integrate experimental data such as transcriptomic data, it can be used as a template to reconstruct tissue-specific cells [53]. As an example, the algorithm called GIMME has been developed for this purpose and has been initially used to derive the first human genome-scale metabolic models for human skeletal muscle cells [20]. Over the years, different tissue-specific models have been reconstructed and more recently the first two gender-specific whole-body metabolism (WBM) reconstructions have been reconstructed [181]. Notably, such reconstructions can integrate different types of data such as dietary, physiological, and omics [181].

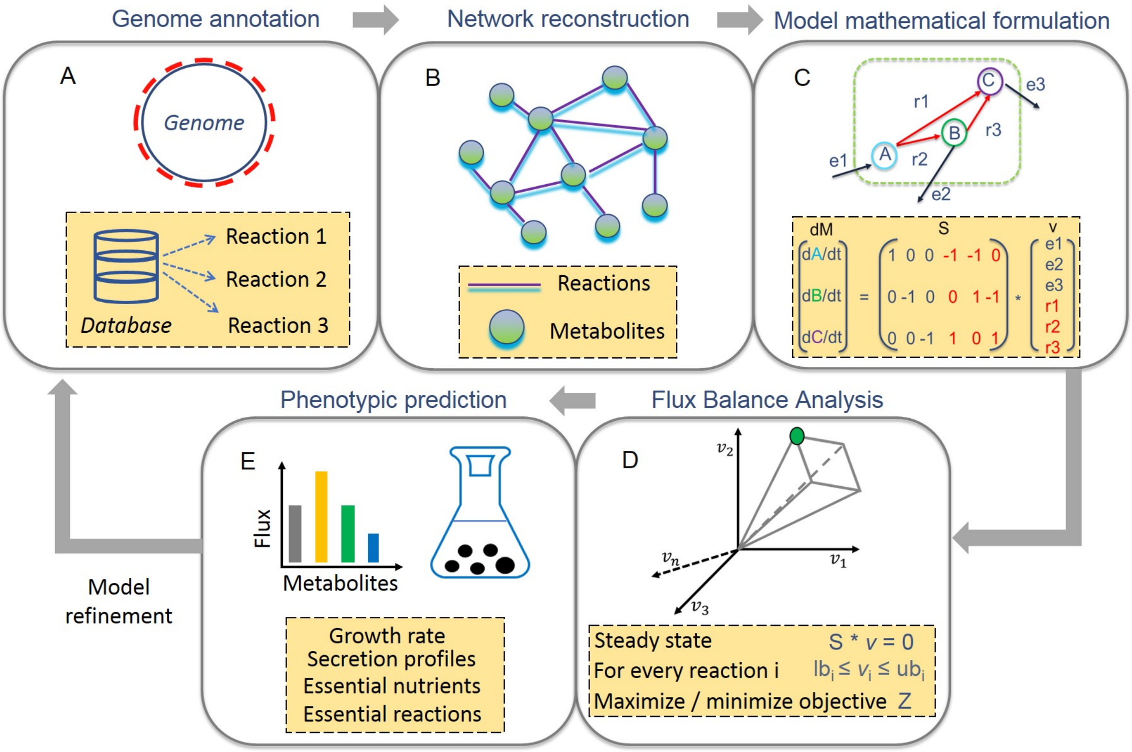


Figure 1.1: **Metabolic modeling. A.** After the genomic annotation phase, reactions coded by different genes are identified. **B.** A network is reconstructed joining metabolites through reactions. **C.** The reconstruction is converted into a computational format. Each compound variation over time is given by the multiplication of S (stoichiometric matrix, representing the stoichiometry of each reaction) for V (vector of fluxes, representing the velocities at which each reaction happens). **D.** Flux balance analysis (FBA) has its basis in the application of three constraints to the reconstruction. **E.** Phenotypic predictions concerning the metabolism of a specific organism can be obtained using FBA.

1.4 Flux Balance Analysis

Flux Balance Analysis (FBA) (Figure 1.1D) is a methodology commonly used to study the metabolism of organisms for which GEMs are available [138]. This method allows the calculation of the flow of metabolites in the network allowing predictions on specific metabolic capabilities [138]. FBA is based on three assumptions i) Steady-state assumption ii) Lower and upper bounds definition for each reaction in the network selecting the minimum and maximum allowable consumption or production rate iii) An objective function for which to optimize (minimize or maximize) the system (Figure 1.1D). As mentioned before, metabolism can be transformed into a mathematical problem and more specifically into a system of equations (Figure 1.1C/D) [138]. Each metabolite can be consumed or produced according to the stoichiometry of the reactions to which it is participating in the network and according to the rate (flux) at which each reaction is happening [138]. The first assumption implies a certain state of disequilibrium, known as steady-state [138], for which no metabolite can accumulate, therefore forcing each produced metabolite to be consumed. The direct consequence of this assumption is that metabolites concentration over time does not change imposing the fact that the multiplication of the stoichiometric matrix S for the vector of fluxes V (representing the flux of metabolites through the network) has to equal to zero [138]. Therefore given the aforementioned constraint, the system of equations describing the metabolism becomes a linear one. The second assumption, defines capacity constraints that together with the existence of the stoichiometric matrix applied to the network creates an allowable solution space [138]. Optimizing for an objective function (third assumption) through the maximization or minimization of the specified objective allows solving the system of linear equations finding a particular flux vector for which the objective is optimized [138]. While the value of the optimized objective function is unique, in many cases, given the lack of different information and parameters, there are several possible different flux distributions allowing that value of objective function [120]. This property is anyhow aligned with the nature of metabolism and its redundancy, for example, reactions redundancy in pathways caused by isoenzymes, contributing to organisms robustness [120]. An approach to studying this redundancy is known as Flux Variability Analysis (FVA) [120]. After selecting a specific flux value for the primary objective function, each reaction in the network is chosen as objective function and the system is optimized to minimize and maximize the flux through that reaction [120]. This operation returns a minimum and maximum allowable value of flux for each reaction in the network under a specific value of primary objective function [120]. The result, besides enabling the study of the variability of the fluxes through the network, can be interesting for many purposes, and FVA can be used to predict reactions directionality or essentiality, as well as blocked reactions in the network (reactions that cannot carry flux under any condition), and essential nutrients under specific conditions. Different objective functions can be chosen according to the modeler's needs and to the need of studying a specific phenotype of interest, but one of the most common studied phenotypes in prokaryotes is growth which is recapitulated by the biomass function. The biomass production is defined as "the rate at which metabolic compounds are converted into biomass constituents such as nucleic acids, proteins, and lipids" [138]. The biomass composition, enumerating all the cellular components (e.g., amino acids, nucleic acids, lipids) needs to be estimated from experimental measurements or, when not fully possible, extrapolated from the genome [180]. The energy consumed for the biosynthesis of the different molecules can also be added in the form of consumed ATP (Growth associated ATP maintenance reaction) [180]. As already mentioned, one of the most intuitive uses of FBA can be to calculate the growth rates of different organisms under specific conditions. Being able to infer the growth of specific organisms under different media and conditions is something extremely interesting that can find immediate usage in many fields such as biotechnology and bioengineering but that can be used to also validate made predictions. This, in the most immediate way, can be achieved evaluating the growth of an organism *in silico* using different carbon sources and comparing it with *in vitro* experiments of phenotypic microarray (PM). However, selecting other objective functions (different from biomass) and using FBA, the production of cofactors and biomass precursors under specific conditions can also be computed [138]. Interestingly, the impact of nutrient and media conditions on the achievement of an objective function value, such as for example, the growth of an organisms, can also be evaluated way changing the conditions of one nutrient or two nutrients at the time using the techniques called robustness analysis and phenotypic phase planes [138]. Briefly, for robustness analysis, over different iterations, the flux through one selected reaction is varied and the objective function value computed as a function of it, while for phenotypic phase plane analysis the fluxes through two reactions are varied at the same time [138]. In addition, FBA can also be used to predict gene knockout impact. Simulating the impact of gene knockouts, as for example constraining the flux of reactions associated to a specific gene to the value of 0 [138], can be interesting for a plethora of applications, from predicting gene essentiality for the growth of an organism, to increase the yield of a specific metabolite. As explained, different phenotypes can be studied using FBA and to this day we collected several examples of different applications: from drug discovery [154] to bioengineering [159] and bioaugmentation. In [154] for example, FBA was used to identify targets for tuberculosis drug design finding essential genes for the production of mycolic acids in *Mycobacterium Tuberculosis* and consequently its pathogenicity. In [159] *in silico* deletions were used to maximize the production of triacylglycerols production by the strain *Acinetobacter baylyi ADP1*. Specifically, four beneficial deletions were selected and studied experimentally through the construction of knock-out strains (MT), and an increase of more than 5.6 fold in the production of triacylglycerols were registered experimentally confirming what identified *in silico* [159].

The human gut microbiota can be considered a complex system composed by different organisms metabolically interacting with each other creating a final state which is a function of these interactions and not only of the single properties of such organisms [18, 17]. To this date, given our little knowledge and comprehension of this system, the difficulty of finding experimental setups to study the properties of the human gut microbiota and its impact on the host, modeling could play a key role in filling the knowledge gap. Starting from well studied and defined in vitro and in vivo minimal communities and expanding towards bigger and less understood human gut microbial communities, metabolic modeling can be used to predict functional properties associated with specific microbial compositions, understanding their structure, functioning and their possible impact on health.

1.4.1 Modeling (human gut) microbial communities with constraint-based modeling

GEMs can be used and, to this date, have been used to study the metabolic properties of the human gut microbiota. In this regard, the year 2017 was a crucial one for the study of the human gut microbiota through metabolic modeling. In fact, before that year, less than 20 manually curated metabolic models of key species composing the human gut microbiota were available [118], while in 2017 AGORA, a collection of over 700 metabolic models of the most commonly found species strains in the human gut was realized . These models were manually curated integrating knowledge on genomics and on metabolism coming from the literature [117]. This was not the only case of large production of microbiota GEMs with other attempts in the latest years. In 2018 Machado et Ali realized CarveMe [116], an automated tool to derive species models from a manually curated “universal” model. CarveMe models performed closely to manually curated models in terms of gene essentiality prediction and substrate utilization and was used to create 74 metabolic models of human gut microbiota members [116].

Consequently, together with advancements and increased availability of GEMs, also frameworks for studying the metabolism of microbial communities and possible related metabolic interactions were developed (Figure 1.2). One of the first attempts of investigating the metabolism of microbial communities can be identified with the usage of the "enzyme soup technique" [106] (Figure 1.2A) where multiple reactions happening in different organisms are added to one single model. This specific framework deals with a situation of limited knowledge and the main objective is the study of the overall community metabolic potential [23]. Individuality is completely lost as there is no attempt to segregate reactions by species/strains, reactions from different species are connected together forming the so-called "meta pathways". Therefore, no information can be provided on the metabolic interactions between the different members of the community, and the main aim of this modeling technique is to predict the biomass and substrate utilization and secretion [23]. This approach was mainly used for natural microbial communities [183, 177]

Another technique used for community metabolic modeling is called compartmentalization (Figure 1.2B) [172, 105]. In this framework, initially deployed to model the presence of different compartments in eukaryotic cells, multiple bacteria models are united together through the creation of compartments allowing the models to share metabolites [172]. A community biomass function is created summing the biomass of the different models. Interestingly there is the possibility to average the different community biomass elements with some coefficients, function of each microbial species presence in experimental active communities. While using this specific setup, two different problems can be solved: the so-called alpha and beta problems [23, 168]. In the alpha problem, organisms abundances are known and uptake and secretion of compounds from the microbial community are predicted on the base of the microbial composition, in the beta problem, secretion and uptake rates are known and fixed and the objective is predicting each species abundance [23]. The compartmentalization approach was initially used for the study of the metabolism of keystone species in pairs predicting possible metabolic relationships between different pairs of species [23]. The first multispecies model realized through a compartmentalization approach was created in the attempt of studying the metabolism of microbial communities composed of methanogens and sulfate-reducing bacteria and was applied to *Desulfovibrio vulgaris* and *Methanococcus maripaludis* metabolic models [172]. The community biomass was an averaged sum (with coefficients coming from experimental values) of the two organisms biomass. Interestingly, the flux of metabolites and the ratio of the two species compared favorably to experimentally measured values [172]. In 2010, the compartmentalization modeling strategy was used to predict pairwise microbial interactions on the base of defined media [105]. The purpose of the study was to define media that would allow the growth of both species but not the growth of each singular species alone, evaluating commensal and mutualistic interactions [105]. Defined possible media that could induce commensalism or mutualism between all pairwise combinations of seven species metabolic models were predicted [105]. Freilich et Ali [63] used the compartmentalization approach to investigate relationships between different microbial pairs using draft automated reconstructed metabolic models. For 6903 microbial pairs (all the combinations of 118 metabolic models) the cooperative and competitive potential was assessed finding cooperation the most common interaction when a moderate number of resources where overlapping, with commensal interactions more diffused [63]. Notably, in 2012 the optCom framework was developed [208]. OptCom builds upon the compartmentalization technique and introduces an additional layer of optimization accounting for species and community fitness. Besides each species optimization, a community-level objective function can be formulated allowing the study of different types of metabolic interactions such as mutualism, commensalism, and parasitism [208]. OptCom was used in a study to understand metabolic interactions between *Bifidobacterium adolescentis L2-32* and *Faecalibacterium prausnitzii A2-165* [58]. The first strain is capable of producing acetate while fermenting and the second can feed on it producing butyrate [58]. FVA was computed for common metabolites finding that butyrate production from *Faecalibacterium prausnitzii* was increased when *Bifidobacterium adolescentis* was present [58]. The compartmentalization approach can be used also to simulate host-microbe metabolic interactions and, more recently, in 2013, Heinken et al., used this approach to derive insights into the host-microbe metabolism joining, for the first time, a mouse metabolic reconstruction affected by inborn error of metabolism, with the model of a Bacteroides strain [79]. The authors proved the ability of *Bacteroides thetaiotaomicron* to rescue the normal phenotype of the mouse metabolic reconstruction [79]. A few years later this compartmentalized host-microbiota model has been expanded joining 11 metabolic models of microbes with the human metabolic reconstruction Recon 2 [182], and, realizing, what represented for many years, and before the work presented in this thesis, the biggest microbial community model created with a compartmentalization technique [81].

The 11 reconstructed GEMs were manually curated and were chosen to represent commensals, probiotics, pathogens, and opportunistic pathogens [81]. The authors created a total of 25 different community models combining the host with each microbe individually, with only two microbes and with five microbes and with the all set of 11 microbes [81]. This effort was done to correlate specific microbial groups under or over-represented in specific disease conditions to possible metabolomic alterations [81]. Metabolic interactions under four different diet regimes were predicted and the microbial community was proved to be able to complement the host with essential metabolites such as precursors of host hormones, glutathione, taurine, and leukotrienes [81]. In addition, it was proved that the synthesis of fundamental neurotransmitters was elevated thanks to the human gut microbiota (increased 34 times in the presence of the 11 microbes) [81]. Using the same community, predicting microbe-microbe and host-microbe interactions, the authors could demonstrate that the host, through the productions of carbohydrates was inducing competition between different microbial pairs as well as the role of oxygen in shaping the behavior of the community, with *Lactobacillus Plantarum* exhibiting mutualistic behavior towards other six community members exclusively under anoxic conditions [80]. In the meantime in 2013, Sohaie and Nielsen created microbiome models for three microbes [168]. Three GEMs chosen as keystone species, *Bacteroides thetaiotamicron*, *Eubacterium rectale*, and *Methanobrevibacter smithii* respectively representing three different phyla (Bacteroidetes, Firmicutes and Euryarchaeota) were reconstructed using the RAVEN toolbox [168, 6]. Substrate uptake and secretion of SCFA was predicted for different combinations of these three microbes: increased secretion of butyrate could be detected when *Bacteroides thetaiotamicron* and *Eubacterium rectale* were co-cultered *in silico*, while methane secretion also increased when combining *M.smithii* and *B. thetaiotaomicron* compared to the one predicted from *M. smithii* alone [168]. Interestingly, predictions of secreted SCFA were in agreement with experimental data [168]. In 2017, using the AGORA GEMs and a pairwise compartmentalization approach, five types of different interactions between microbial pairs were predicted evaluating the impact of different nutrients[117]. Interestingly oxygen presence, simulating an inflammatory status of the human gut, resulted in increasing negative interactions (competition and amensalism) decreasing commensalism and mutualism, while the supplementation with a high fiber diet had an opposite effect [117]. Finally, Kbase developed tools for the creation of microbiota compartmentalized models starting from genomic sequencing: *de novo* assembly is performed using the metagenomic reads, draft strains metabolic models are created from the newly assembled genomes and microbiota models are built using integrating microbial abundance information [9]. In summary, the compartmentalization approach represents a very simple and efficient methodology to simulate the metabolism and related metabolic interactions of different organisms composing microbial communities. As a consequence of this, the majority of the studies were conducted using this approach [23]. This technique can represent a snapshot of a specific scenario describing a microbial community in a specific moment and situation. Compartmentalization is a very powerful technique to investigate, therefore explaining, the metabolic mechanisms behind specific microbial compositions, and to be able to predict properties of microbial communities and their possible impact on the host health. However, the compartmentalization technique has also some limitations. First of all, there is no representation for temporal and spatial components and their relative impact. Secondly, this approach of pre-defining stoichiometric coefficients of the community biomass reaction is not linked to the expectation or condition of growth in an optimal way. Finally, the effects of dietary interventions, in terms of changes in composition or diet components, can also only be evaluated with fixed microbial abundances and consequent changes in microbial abundances cannot be easily predicted.

To overcome some of these difficulties, during the years, efforts were made for the creation of dynamic modeling setups. In standard FBA, on the base of the given constraints, fluxes are retrieved for a specific instant and metabolites are not allowed to be depleted or accumulated [138]. With dynamic FBA [119], metabolites can accumulate as fluxes are integrated over time and we can assist to the formation of population dynamics as a consequence of environmental changes. For a specific amount of time, in discrete time steps, the biomass of each microbe is optimized and its value retrieved while the cell is secreting and uptaking metabolites on the base of the available nutrients [206]. Dynamic FBA requires kinetic parameters related to compounds uptake [119]. As this methodology represents an expansion of the compartmentalized multispecies FBA, a common compartment, where microbes can share metabolites and allowing microbial metabolic interactions, is also present. The presence of the time component allows the study of these microbial interactions under different conditions: secretion and consumption of metabolites over time as well as changes in the microbial abundances can be predicted. Several studies were conducted performing dFBA on small microbial communities gaining important insights demonstrating the utility of this approach. Useful insights were gained in modeling co-cultures of organisms such as *Escherichia coli* - *Saccharomyces* [73], and *Saccharomyces cerevisiae* - *Scherffersomyces stipitis* [74] as well as a collection of metabolically different cells in the modeling of *Pseudomonas aeruginosa* biofilm [24]. A dynamic version of optCom named d-optCom was also developed [207] and, more recently, further advances made it possible to also model the spatial component of microbial communities [75] creating the so-called population-based community models [17]. The spatial component enables the visualization of consumed and produced metabolites with their relative diffusion and the consequent formation of colonies given by different concentrations of nutrients [54]. One of the first tools developed for population-based modeling is COMETS [75]. COMETS associates dynamic-FBA with diffusion on a lattice [75] and its predictions on a two-species mutualistic community were experimentally validated as well as the ones that were done on a three-member community [75]. This approach was used to study the metabolism of a community composed of six microbes [71]. What was found is that the members of the community were highly metabolically interacting with each other (cross-feeding) and with different interactions according to the different time steps [71]. For many of these cross-feeding interactions, evidence in the literature was found [71]. In 2017 SteadyCom was developed [39]: one of the main purposes of this tool is the prediction of microbial composition at the steady-state. As a matter of fact, an additional constraint "a time-averaged constant growth rate to ensure co-existence and stability" for all the members of the community is imposed to avoid overgrowth of faster-growing organisms [39]. SteadyCom was used to predict population dynamics in a community of nine species on the base of different diets utilized predicting microbial abundances similar to what experimentally observed [39]. Population-based approaches represent a very attractive possibility of modeling complex microbial communities, especially considering the possibility to investigate time and spacial components that are crucial for understanding the way microbial communities might interact and change over time. However, these approaches group different individuals of each species in a colony and rely on the assumption that each colony of species is phenotypically homogeneous in terms of metabolism. To overcome this shortening, individual-based community single and multispecies models (IBM) have been developed (Figure 1.2C) [18, 24, 191]. In individual based modeling each species component is considered a separate individual allowing for the atherogenicity of individuals of the same species forming a colony. In 2017, we developed BacArena [18]. In BacArena a grid of specified dimensions can be created and compounds can be easily added in specific concentrations or in the form of gradients [18]. Compounds are free to diffuse with specified coefficients in the space of the grid. Metabolic models can also be inserted in the grid, randomly or in specified positions and a simulation can be launched for a specific amount of time [18]. For each individual, a set of rules can be specified defining for example, how fast each cell can move in the grid, after what amount of biomass can duplicate (growth rate), and a death rate [18]. During the simulation, each individual can secrete and uptake compounds and grow with the rates predicted with FBA, changing the concentration of the compounds in the grid, therefore creating population and compounds concentration dynamics [18]. Moreover, for each time step, in a discrete manner, compounds concentration can be retrieved, the metabolic phenotype of each cell analyzed and cross-feeding predicted on the base of each individual metabolism [18]. Finally, a spatial component can be observed explaining the possible formation of niches [18]. BacArena was initially applied to simulate microbiota interactions and metabolism in the gut using a small community of seven keystone species species [19] predicting ratios of compounds production and utilization comparable to experimental measurements obtained using gnotobiotic mouse [18] and, more recently, to predict dietary supplementation for Crohn's disease patients [16]. Individual-based modeling techniques and tools such BacArena can be very useful to simulate population dynamics, pro and prebiotics interventions, antibiotics usage, and even fecal microbiota transplant (FMT). The drawback of such tools is the amount of information needed, in general, and for each individual, many information are required, such as diffusion coefficients, physiological information on different organisms. In addition, in IBM modeling, it is very challenging obtaining quick answers form targeted questions. This situation in mainly created by the increase of variables (e.g., individuals proximity), and the increase of computing time, mainly due to calculating diffusion of compounds.

In summary, each of the previously described techniques for modeling the metabolism of microbial communities has its pros and cons, as discussed, and can be useful depending on the biological situation that the modeler wants to simulate and the questions asked. However,

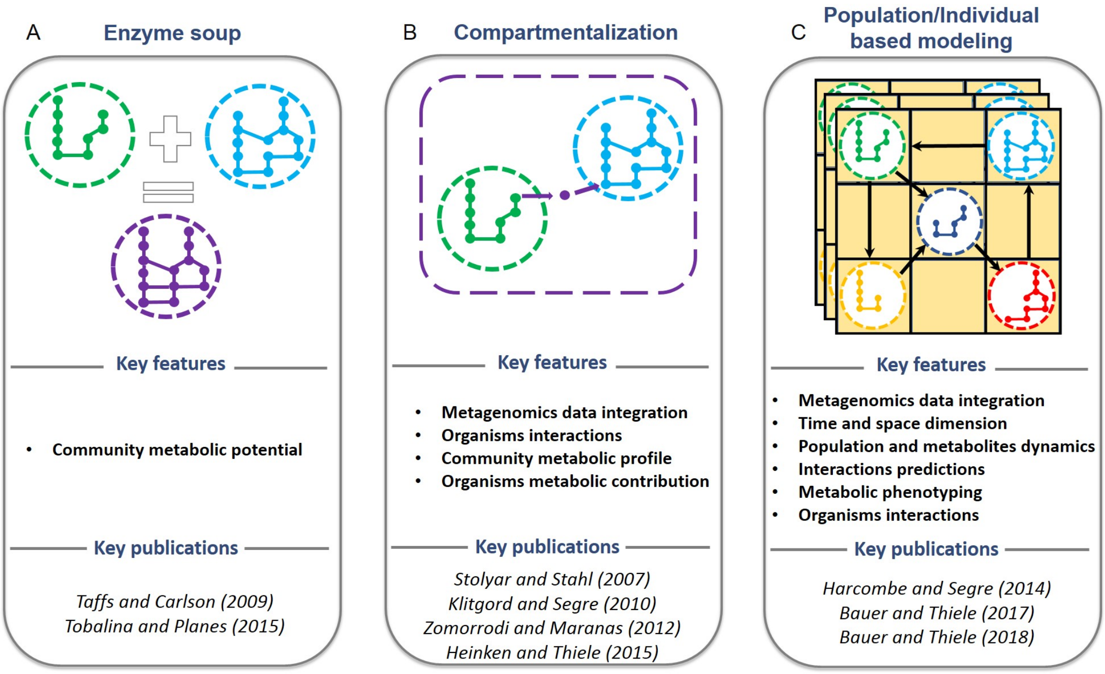

A
Enzyme soup
Key features
• Community metabolic potential
Key publications
Taffs and Carlson (2009)
Tobalina and Planes (2015)
B
Compartmentalization
Key features
• Metagenomics data integration
• Organisms interactions
• Community metabolic profile
• Organisms metabolic contribution
Key publications
Stolyar and Stahl (2007)
Klitgord and Segre (2010)
Zomorrodi and Maranas (2012)
Heinken and Thiele (2015)
C
Population/Individual based modeling
Key features
• Metagenomics data integration
• Time and space dimension
• Population and metabolites dynamics
• Interactions predictions
• Metabolic phenotyping
• Organisms interactions
Key publications
Harcombe and Segre (2014)
Bauer and Thiele (2017)
Bauer and Thiele (2018)

Figure 1.2: **Community modeling techniques. A.** Enzyme soup modeling. **B.** Compartmentalized modeling. **C.** Population and individual based modeling.

even considering these efforts, to this day, powerful and automated tools able to reconstruct community models for hundreds of different manually curated organisms, retrieving and integrating microbial abundance information into community microbial modeling, are still missing. Many of the previously described studies are limited to few organisms, and small microbial communities or miss the integration of experimental data obtained with microbial sequencing. This calls for methods able to implement large-scale production of personalized microbiota models, automatically integrating information retrieved from microbial sequencing data into microbial community modeling to explore different possible metabolic signatures associated with diverse health states. Revisiting existent metagenomics cohort studies and processing future ones adding a causal dimension given by the aforementioned modeling techniques, could reveal essential microbiota mechanisms and signatures in health and disease states.

1.5 Modeling challenges, conclusions, and future perspectives

COBRA modeling approaches represent a powerful tool for analyzing the metabolism of microbial communities such as the ones that we harbor in our gut, and their underlying properties [17]. In the previous chapter, we enumerated different techniques of COBRA modeling and cases in which the related COBRA modeling approach was deployed to model microbial communities. The related studies pioneered the study of microbial communities through metabolic modeling and provided useful insights. However, if many of the introduced approaches can readily be used for the study of the human gut microbiota, many challenges and limitations still need to be tackled and overcome.

First of all, many of the previously conducted studies were carried on small communities composed of two species or at the maximum few species. However, the human gut microbiota is composed of hundreds of different microbial species creating a complex environment [17], trying to reflect this complexity is essential given that the final properties might be changed after the inclusion of more species. In addition, the focus of most of the recent studies is mainly tailored to the modeling of microbial species while we know that other kingdom species such as fungi might play an important role in the homeostasis of the human gut microbiota [95, 90]. On the same line, the impact of virus such as bacteriophages on the microbial populations is something not yet fully unraveled [124] and much is to be understood concerning, especially, the long-term impact of antibiotics. Both agents affect specific organisms composing microbial communities and their presence might be a driver to shape microbial communities composition, and, by consequence, functions. These topics need, for sure, attention and comprehensive and scalable ways to account for them into modeling frameworks still need to be developed. In this regard, the most recent modeling techniques such as IBM based techniques might represent an attractive starting point. Other challenges come, once again, from the microbial nature itself and evolution with mechanisms such as horizontal gene transfer (HGT) and transduction. Both mechanisms are evolutionary ones and can contribute to generating microbial diversity [140] and HGT is known to happen frequently in the human gut [31, 170]. By consequence, the strains that we model in the gut might be genetically different from the ones isolated in other environments containing different genes due to the aforementioned process. Few studies have been carried so far [176, 134], but being able to capture and model this genetic diversity, as well as being able to incorporate evolutionary mechanisms that can happen within the human gut microbiota, is still a challenge.

When modeling the metabolism of the human gut microbiota, one cannot limit oneself to the modeling of the metabolism of the organisms composing it, but also the environmental conditions have to be taken into account. More challenges come from modeling these conditions, conditions such as dietary ones. Even if solutions for this problem have been specifically developed [132] it is still very challenging to be able to track and consequently design media, reflecting, in silico, the composition of each individuals diet. This issue can be addressed in many ways, from developing dedicated apps to track diet [204] with the aim of designing personalized *in silico* diets or applying standard dietary conditions evaluating the response of different microbial compositions between individuals with the assumption that diet already contributed to differently shape microbial communities. Using different modeling setups to run sensibility tests on different nutrients and evaluating their impact on the modeling predictions can also help to tackle the problem, but the difficulty to formulate precise and accurate compounds compositions summing up the diet formulation and nutrients concentration still remains a serious limitation.

Experimental data integration into community modeling represents definitely a double-edged sword. From one side it can help to design better and more reliable modeling setups, while from the other how to integrate this data is not trivial and can contribute to the increase of the number of assumptions. Ideally, different types of data can be integrated into community metabolic modeling from metagenomics [16, 78] to meta-transcriptomic, meta-proteomic and metabolomic data. As for example, metagenomics data can be used to compute relative abundances [101, 16, 78] that can be, later on, integrated in different ways according to the interpretation and the modeling technique used: using a compartmentalization technique relative abundances can be used directly to average each microbial biomass in the community biomass function, while for IBM modeling relative abundances need to be converted in number of individuals of different species composing the total cell count [17]. Alternatively, replication rates can be estimated from metagenomics using tools such as iRep and used as growth rates in community modeling [32, 17].

Finally, more extensive and comprehensive validation of these COBRA modeling techniques is still required and the development of experimental *in vivo* setups [167] might play an essential role in validating the predictions of metabolic community modeling helping, when needed, with further refinement, and directing future research and development efforts.

In conclusion, we introduced different modeling approaches and the related modeling applications to model the metabolism of microbial communities such as the ones composing the human gut microbiota. Each of the introduced methods poses its basis on COBRA modeling, has its advantages and limitations, but can result useful in modeling different scenarios. If many of these methods need still an accurate and extensive validation process, many of the presented findings were validated against literature or using examples of minimal communities, demonstrating the potential of COBRA modeling for microbial communities and the human gut microbiota.

1.6 Scope and aim of the thesis

This thesis work is aimed at investigating the metabolic potential of gut microbial communities in the presence of specific diseases, such as Parkinson's disease, and it is divided into three main objectives. The first objective chapter 2, is the development of methods to create and interrogate personalized community microbiota metabolic models through the integration of samples relative abundances. This objective was accomplished with the development of a toolbox, The Microbiome Modeling Toolbox, which collects methods for microbe-microbe multispecies metabolic modeling with related relative abundance integration and host-microbes multispecies metabolic modeling. The second objective chapter 3 can be seen as an expansion of the first, starting from whole-genome sequencing or 16S rRNA gene sequencing data, a selection of methods are combined to create and interrogate personalized microbiota models prior to the retrieval of sample specific microbial relative abundances. For the third objective chapter 4, we processed 16S rRNA gene sequencing data for a cohort composed of PD patients and controls, constructing personalized microbiota models and comparing the predicted metabolic potential of the related microbial communities. We identified alterations in the microbial composition of PD patients and the related metabolic functional consequences.

Chapter 2: The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities

chapter 2 describes the creation of a Toolbox for microbe-microbe and host-microbe metabolic interactions and for metagenomics data integration into community modeling. Examples of the usage of the toolbox are given (cf. Appendix A). The chapter is a combination of the full reprint of the paper published in Bioinformatics in November 2018 with the related documentation and tutorials (cf. Appendix A).

Contributions

Federico Baldini (FB), Almut Heinken (AH), Ronan M T Fleming (RMTF), and Ines Thiele (IT) designed the study. FB, Stefania Magnusdottir (SM), and AH implemented the code, FB and AH wrote the tutorials and the documentation, FB, AH, and Laurent Heirendt uploaded the code and wrote the testing functions. All authors edited and approved the final manuscript.

Chapter 3: mgPipe: from microbial community sequencing data to personalized microbiota metabolic models creation and interrogation

chapter 3 describes the creation of an automated tool (mgPipe) processing whole genome sequencing and 16S rRNA gene sequencing Illumina data, creating and interrogating, after conducting microbial identification, personalized microbiota metabolic models. The chapter is a combination of the current version of the manuscript in preparation with the related documentation and tutorial (cf. Appendix B).

Contributions

FB and IT designed the study. FB implemented the code, FB wrote the tutorials and the documentation, FB uploaded the code. FB performed the writing of the manuscript.

Chapter 4: Parkinson's disease-associated alterations of the gut microbiome can invoke disease-relevant metabolic changes

chapter 4 describes the creation and interrogation of personalized microbiota models for a cohort composed of PD patients and healthy controls. Relative abundances are obtained from 16S rRNA gene sequencing and used as input for the personalization of the microbiota models. Alterations in microbial composition and consequently microbial community metabolic potential are found in PD patients. The chapter is the full pre-reprint of the submitted manuscript deposited on bioRxiv at http://dx.doi.org/10.1101/691030.

Contributions

FB, Johannes Hertel (JH), Rejko Krüger (RK), and IT designed the study. Lorieza Neuberger-Castillo and Fay Betsou performed the sequencing, FB, Cyrille C. Thinnes, and Estelle Sandt managed the related data, FB performed the marker gene and metabolic modeling analysis, JH performed the statistical analysis, RK and Lukas Pavelka the clinical assessment. All authors edited and approved the final manuscript.

Chapter 5: Concluding remarks

chapter 5 contains the conclusions of the presented work of thesis and the author's personal outlook on the challenges and future perspectives of the field of metabolic modeling of the human gut microbiota.

Contributions

The text was written in full by FB.

Chapter 2

The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities

Baldini, F., Heinken, A., Heirendt, L., Magnusdottir, S., Fleming, R. M., and Thiele, I. (2018). The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. Bioinformatics, 35(13), 2332-2334

Abstract

Motivation

The application of constraint-based modeling to functionally analyze metagenomic data has been limited so far, partially due to the absence of suitable toolboxes.

Results

To address this gap, we created a comprehensive toolbox to model i) microbe-microbe and host-microbe metabolic interactions, and ii) microbial communities using microbial genome-scale metabolic reconstructions and metagenomic data. The Microbiome Modeling Toolbox extends the functionality of the COBRA Toolbox.

Availability and implementation

The Microbiome Modeling Toolbox and the tutorials at https://git.io/microbiomeModelingToolbox.

2.1 Introduction

Microbial community sequencing data are increasingly available for numerous environmental niches [125]. The analysis of this data often relies on investigating which microbes are present in a given sample. However, to further our understanding of the functional contribution of individual microbes in a community as well as the overall functional differences between communities, advanced analysis approaches, such as computational modeling, are required.

One possible approach is the constraint-based reconstruction and analysis (COBRA) approach, which builds genome-scale reconstructions of an organism and enables the prediction of, e.g., phenotypic properties [142]. Through the application of condition-specific constraints, an organism's metabolic reconstruction can be converted into many condition-specific models, which can be analyzed using available toolboxes, such as the Matlab (Mathworks, Inc.) based COBRA Toolbox [84]. Metabolic reconstructions have been assembled for many organisms, including hundreds of gut microbes [117] and human [33]. While the COBRA Toolbox encapsulates many tools developed by the community for biotechnological and biomedical applications, it is currently focused on modeling single organisms or cells. Here, we present the Microbiome Modeling Toolbox, which enables the generation, simulation, and interpretation of 1. pairwise microbe-microbe and host-microbe interactions, and 2. sample-specific microbial community models. By integrating sample-specific metagenomic data, the Microbiome Modeling Toolbox facilitates its analysis in the context of microbial reconstructions.

2.2 Features

The Microbiome Modeling Toolbox (Figure 2.1) enables the generation, simulation, and interpretation of 1. pairwise microbe-microbe and host-microbe interactions, and 2. sample-specific microbial community models.

Pairwise interactions: The pairwise interaction analysis determines metabolic exchange between two metabolic reconstructions. A joint matrix of two individual genome-scale reconstructions is generated, which enables them to freely exchange metabolites (Figure 2.1A). Defined nutrient input, e.g., a particular medium formulation, can be applied via the shared compartment using the corresponding exchange reactions. The pairwise microbial models can be investigated for six possible interaction types (i.e., competition, parasitism, amensalism, neutralism, commensalism, and mutualism) and Pareto optimality frontiers can be calculated. The tutorials *MicrobeMicrobeInteractions* and *HostMicrobeInteractions* illustrate the implemented functionalities.

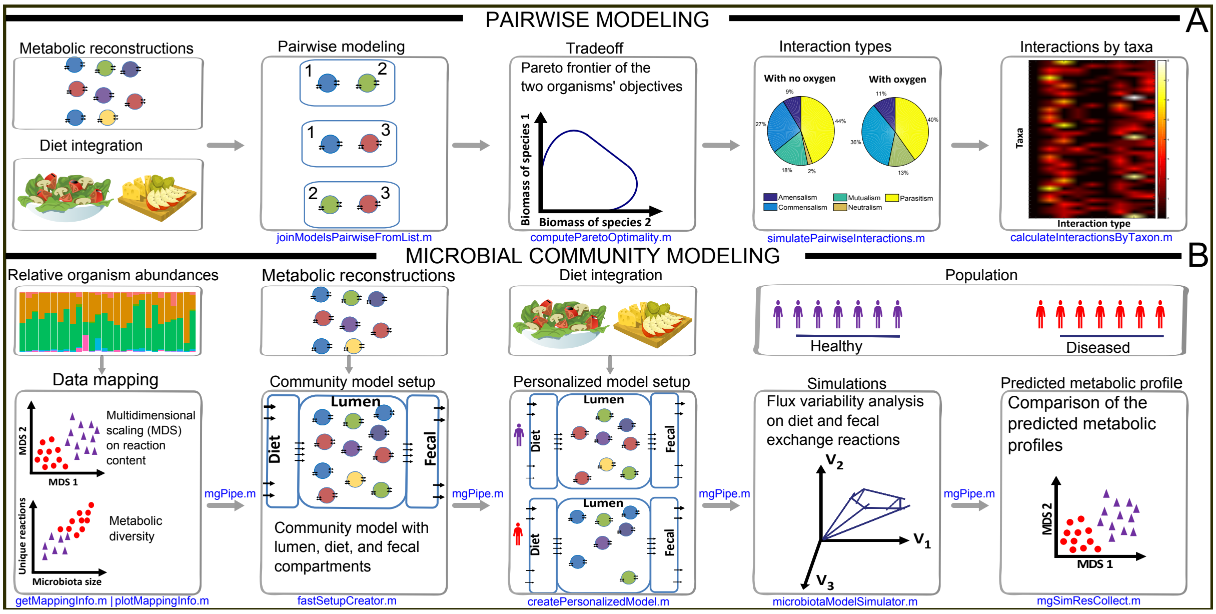

PAIRWISE MODELING
A
Metabolic reconstructions
Diet integration
Pairwise modeling
1
2
1
3
2
3
joinModelsPairwiseFromList.m
Tradeoff
Pareto frontier of the two organisms' objectives
Biomass of species 1
Biomass of species 2
computeParetoOptimality.m
Interaction types
With no oxygen
With oxygen
9%
44%
27%
18%
2%
11%
40%
36%
13%
Amensalism
Commensalism
Mutualism
Neutralism
Parasitism
simulatePairwiseInteractions.m
Interactions by taxa
Taxa
Interaction type
calculateInteractionsByTaxon.m
MICROBIAL COMMUNITY MODELING
B
Relative organism abundances
Metabolic reconstructions
Diet integration
Population
Healthy
Diseased
Data mapping
MDS 2
MDS 1
Multidimensional scaling (MDS) on reaction content
Unique reactions
Microbiota size
Metabolic diversity
getMappingInfo.m | plotMappingInfo.m
mgPipe.m
Community model setup
Lumen
Diet
Fecal
Community model with lumen, diet, and fecal compartments
fastSetupCreator.m
mgPipe.m
Personalized model setup
Lumen
Diet
Fecal
Lumen
Diet
Fecal
createPersonalizedModel.m
mgPipe.m
Simulations
Flux variability analysis on diet and fecal exchange reactions
V2
V1
V3
microbiotaModelSimulator.m
mgPipe.m
Predicted metabolic profile
Comparison of the predicted metabolic profiles
MDS 2
MDS 1
mgSimResCollect.m

Figure 2.1: **Overview of the Microbiome Modeling Toolbox. A.** Pairwise modeling of microbe-microbe and host-microbe interactions. **B.** Microbial community modeling.

Microbial community modeling: Metagenomic data can be analyzed using *mgPipe* (Figure 2.1B), which requires microbe identification and relative abundance data for each sample, obtained with bioinformatic tools, such as QiIME 2[35] and MetaPhlAn [166]. *mgPipe* is divided into three parts: 1. the analysis of individuals' specific microbes abundances, including metabolic diversity and classical multidimensional scaling of the reactions in the identified microbes. 2. Construction of a personalized microbial community model using the identified microbes and their relative abundance data. For each personalized (or sample-specific) model, the corresponding microbial reconstructions are joined by adding reactions to each microbial reconstruction transporting metabolites from the extracellular space to the common lumen compartment. Metabolites present in the lumen compartment are connected to a diet and fecal compartment, enabling the uptake and secretion from/to the environment, respectively. Hundreds of reconstructions can be combined and modeled with using static parallelization. In each microbial community model, the community biomass reaction is personalized using the relative abundance data. Finally, coupling constraints [79] are applied to couple the flux through each microbial reaction to its corresponding biomass reaction flux. And 3. simulation of the personalized microbial community models under different diet regimes, e.g., using flux variability analysis [85]. The differences between maximal uptake and secretion fluxes provide a metabolic profile for each microbial community sample, which can be analyzed using classical multidimensional scaling analyses. Diet-specific constraints (e.g., obtained from https://vmh.life/#nutrition) can be applied to the corresponding diet exchange reactions.

2.3 Implementation

The Microbiome Modeling Toolbox is written in MATLAB (Mathworks, Inc.) and accompagnied with comprehesive documentation and tutorials. The toolbox allows for the integrative analysis of any number of reconstructions, including the human metabolic reonstruction [33]. Metabolic reconstructions can be obtained from, e.g., the VHM (https://vmh.life), BioModels (https://www.ebi.ac.uk/biomodels-main/), and the KBase (https://kbase.us/). A uniform nomenclature of reaction and metabolite abbreviations across the reconstructions is required. The implemented diet constraints require VMH abbreviations. To use higher taxonomical levels create pan-reconstructions (*createPanModels*). For larger datasets and/or bigger microbial community models, we recommend the use of the MATLAB command line or .m files and of a high-performance computing cluster.

2.4 Discussion

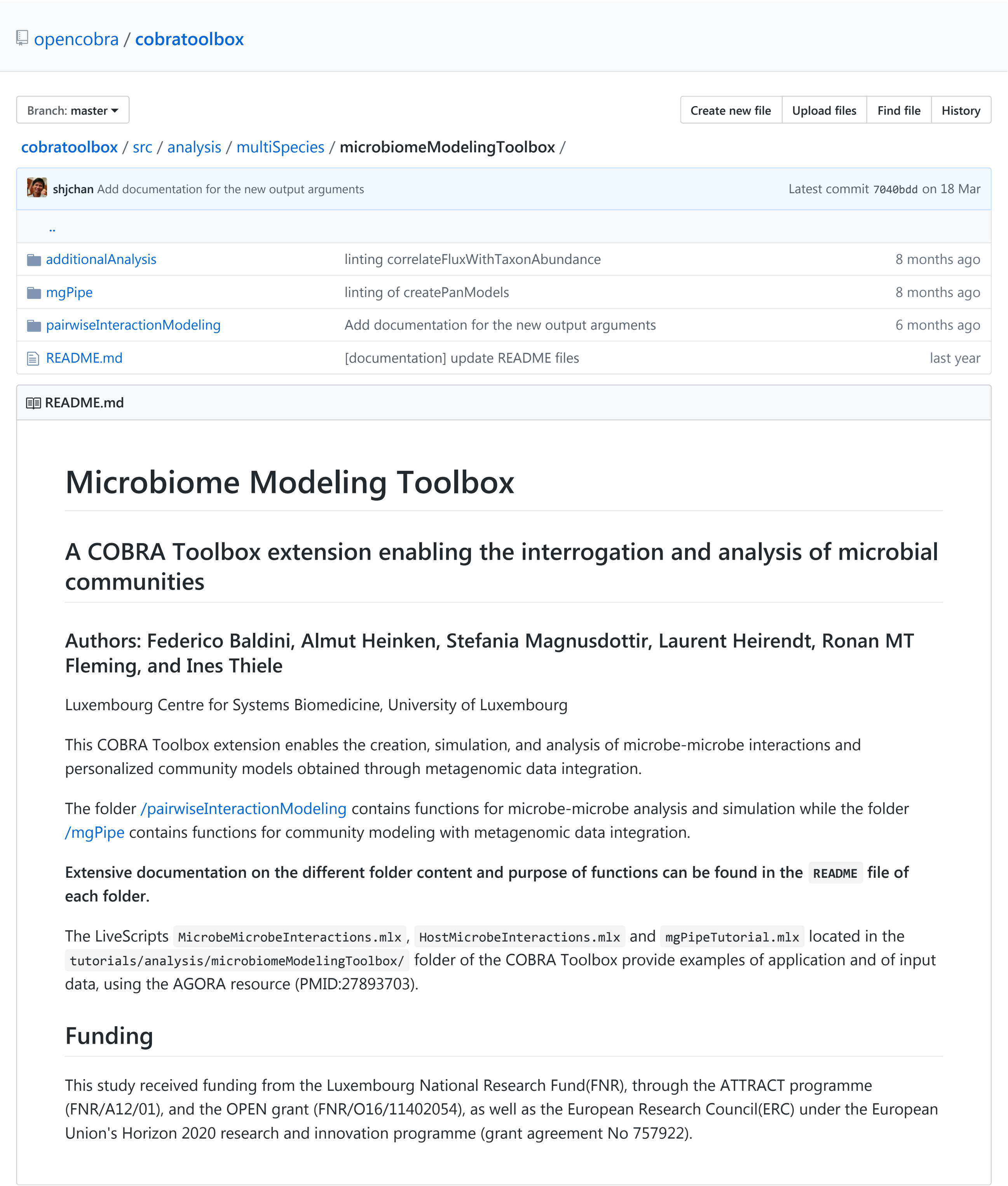
The Microbiome Modeling Toolbox enables the user to investigate microbial interactions at a large scale [79, 117]. Moreover, metagenomically-derived data can be integrated with microbial metabolic reconstructions permitting the prediction of altered functional assessment of different microbial communities, e.g., in health and disease [78, 181].

2.5 Funding

This study received funding from the Luxembourg National Research Fund (FNR), through the ATTRACT programme (FNR/A12/01), and the OPEN grant (FNR/O16/11402054), as well as the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757922).

2.6 Documentation

This section is composed of the software documentation as available online.



cobratoolbox / src / analysis / multiSpecies / microbiomeModelingToolbox /

shjchan Add documentation for the new output arguments — Latest commit 7040bdd on 18 Mar

..		
additionalAnalysis	linting correlateFluxWithTaxonAbundance	8 months ago
mgPipe	linting of createPanModels	8 months ago
pairwiseInteractionModeling	Add documentation for the new output arguments	6 months ago
README.md	[documentation] update README files	last year

README.md

Microbiome Modeling Toolbox

A COBRA Toolbox extension enabling the interrogation and analysis of microbial communities

Authors: Federico Baldini, Almut Heinken, Stefania Magnusdottir, Laurent Heirendt, Ronan MT Fleming, and Ines Thiele

Luxembourg Centre for Systems Biomedicine, University of Luxembourg

This COBRA Toolbox extension enables the creation, simulation, and analysis of microbe-microbe interactions and personalized community models obtained through metagenomic data integration.

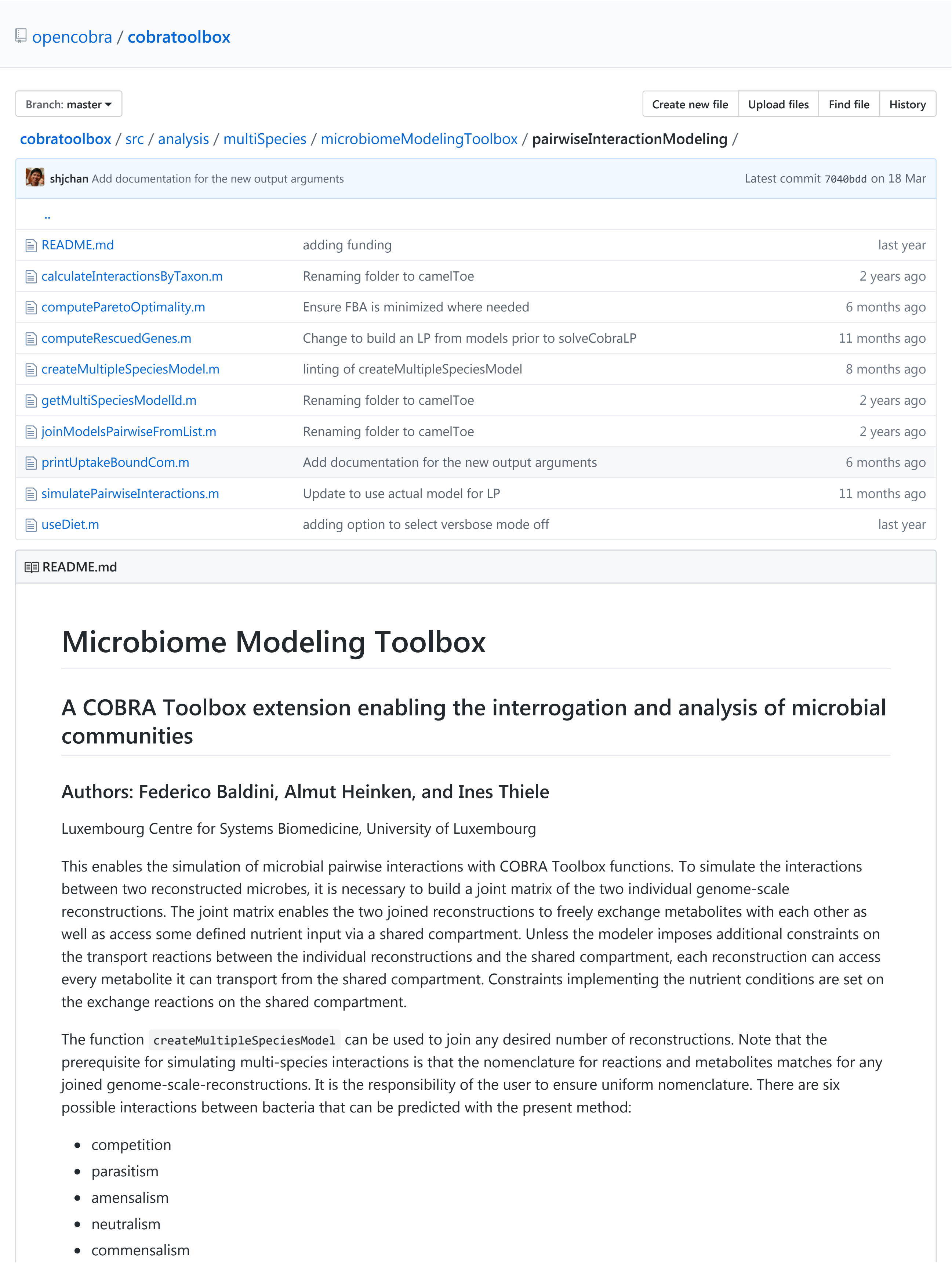
The folder /pairwiseInteractionModeling contains functions for microbe-microbe analysis and simulation while the folder /mgPipe contains functions for community modeling with metagenomic data integration.

Extensive documentation on the different folder content and purpose of functions can be found in the `README` file of each folder.

The LiveScripts `MicrobeMicrobeInteractions.mlx`, `HostMicrobeInteractions.mlx` and `mgPipeTutorial.mlx` located in the `tutorials/analysis/microbiomeModelingToolbox/` folder of the COBRA Toolbox provide examples of application and of input data, using the AGORA resource (PMID:27893703).

Funding

This study received funding from the Luxembourg National Research Fund(FNR), through the ATTRACT programme (FNR/A12/01), and the OPEN grant (FNR/O16/11402054), as well as the European Research Council(ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757922).



cobratoolbox / src / analysis / multiSpecies / microbiomeModelingToolbox / pairwiseInteractionModeling /

shjchan Add documentation for the new output arguments — Latest commit `7040bdd` on 18 Mar

..		
README.md	adding funding	last year
calculateInteractionsByTaxon.m	Renaming folder to camelToe	2 years ago
computeParetoOptimality.m	Ensure FBA is minimized where needed	6 months ago
computeRescuedGenes.m	Change to build an LP from models prior to solveCobraLP	11 months ago
createMultipleSpeciesModel.m	linting of createMultipleSpeciesModel	8 months ago
getMultiSpeciesModelId.m	Renaming folder to camelToe	2 years ago
joinModelsPairwiseFromList.m	Renaming folder to camelToe	2 years ago
printUptakeBoundCom.m	Add documentation for the new output arguments	6 months ago
simulatePairwiseInteractions.m	Update to use actual model for LP	11 months ago
useDiet.m	adding option to select versbose mode off	last year

README.md

Microbiome Modeling Toolbox

A COBRA Toolbox extension enabling the interrogation and analysis of microbial communities

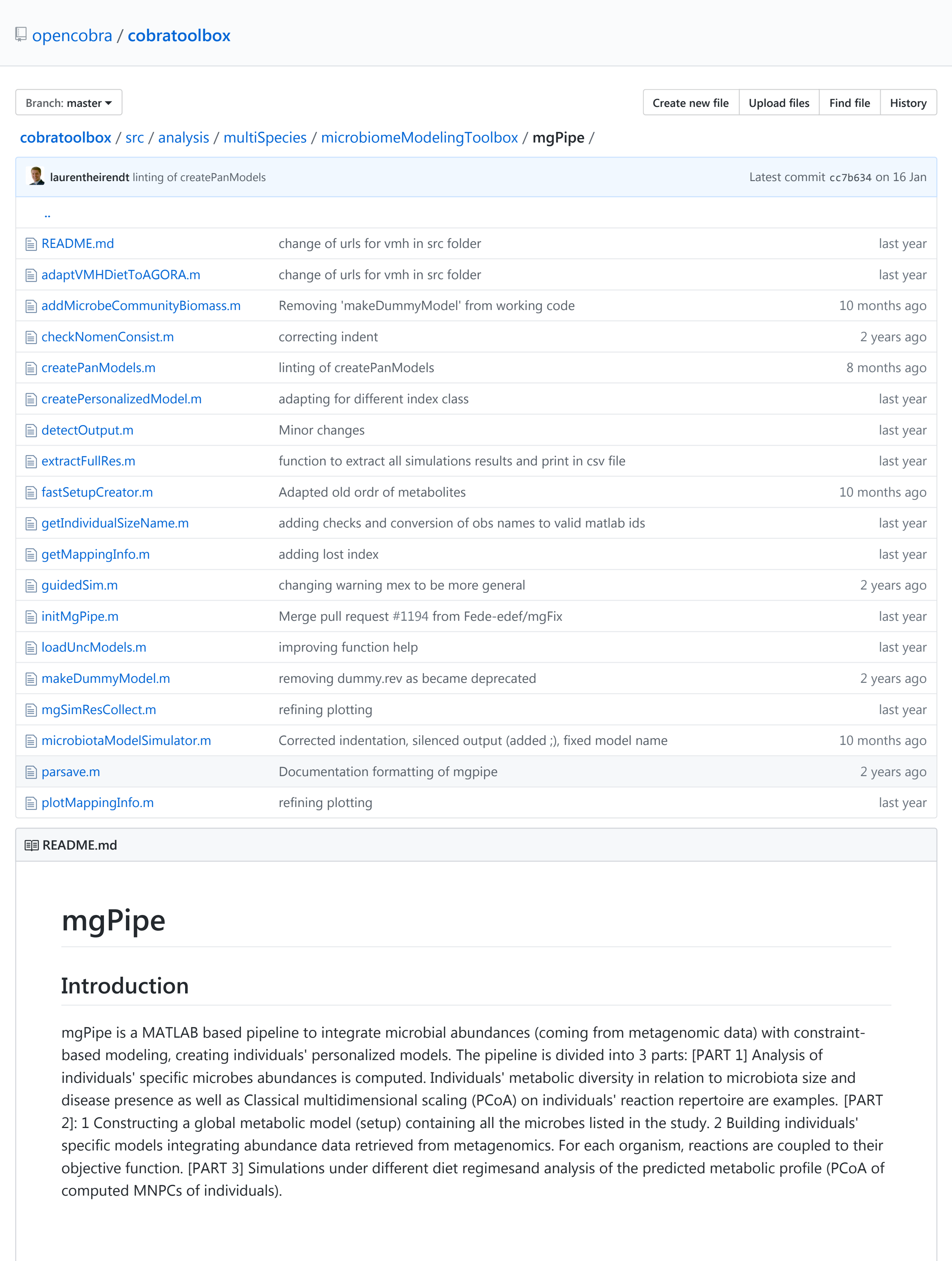
Authors: Federico Baldini, Almut Heinken, and Ines Thiele

Luxembourg Centre for Systems Biomedicine, University of Luxembourg

This enables the simulation of microbial pairwise interactions with COBRA Toolbox functions. To simulate the interactions between two reconstructed microbes, it is necessary to build a joint matrix of the two individual genome-scale reconstructions. The joint matrix enables the two joined reconstructions to freely exchange metabolites with each other as well as access some defined nutrient input via a shared compartment. Unless the modeler imposes additional constraints on the transport reactions between the individual reconstructions and the shared compartment, each reconstruction can access every metabolite it can transport from the shared compartment. Constraints implementing the nutrient conditions are set on the exchange reactions on the shared compartment.

The function `createMultipleSpeciesModel` can be used to join any desired number of reconstructions. Note that the prerequisite for simulating multi-species interactions is that the nomenclature for reactions and metabolites matches for any joined genome-scale-reconstructions. It is the responsibility of the user to ensure uniform nomenclature. There are six possible interactions between bacteria that can be predicted with the present method:

- competition
- parasitism
- amensalism
- neutralism
- commensalism



cobratoolbox / src / analysis / multiSpecies / microbiomeModelingToolbox / mgPipe /

laurentheirendt linting of createPanModels Latest commit cc7b634 on 16 Jan

..

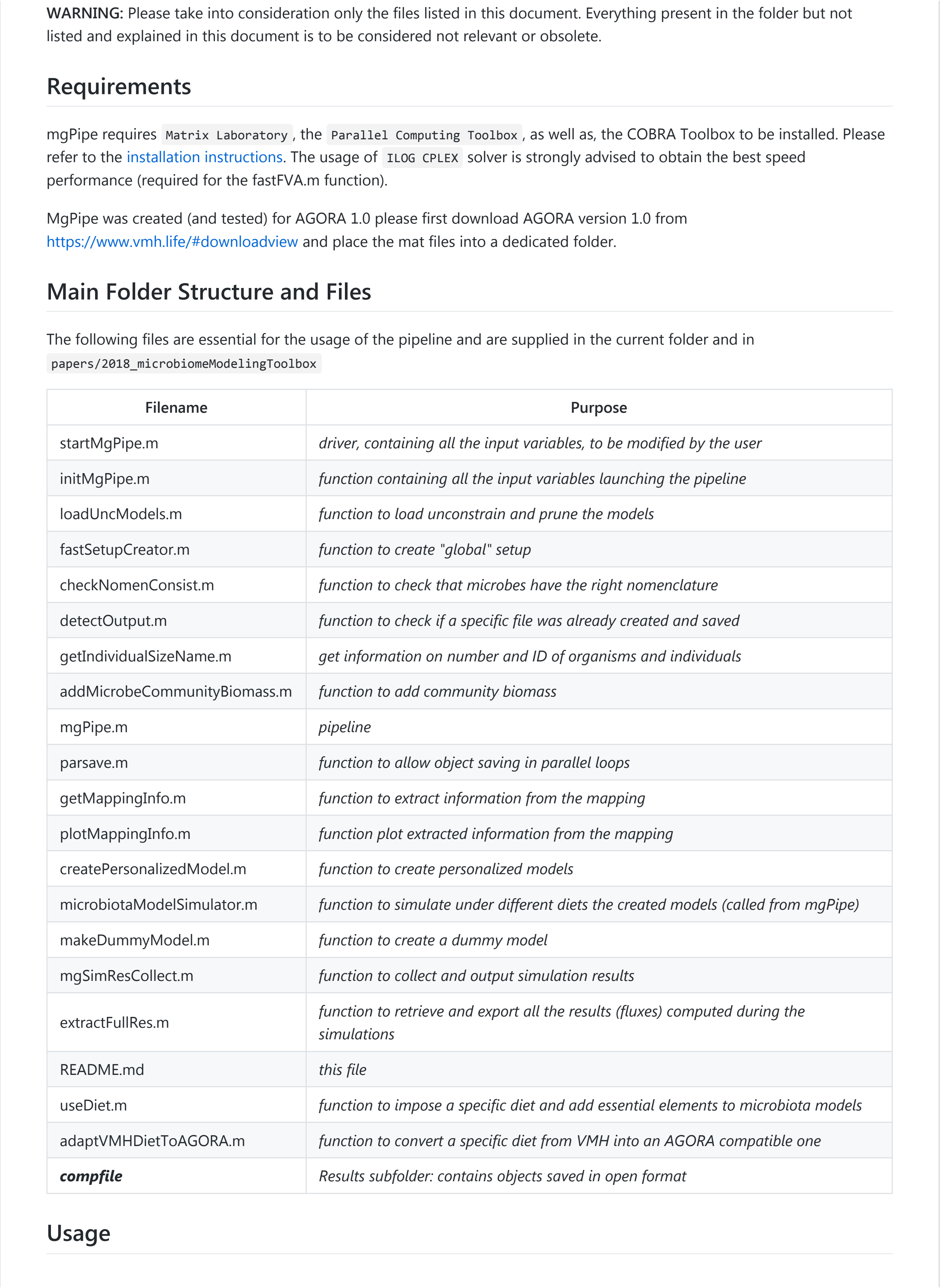
README.md	change of urls for vmh in src folder	last year
adaptVMHDietToAGORA.m	change of urls for vmh in src folder	last year
addMicrobeCommunityBiomass.m	Removing 'makeDummyModel' from working code	10 months ago
checkNomenConsist.m	correcting indent	2 years ago
createPanModels.m	linting of createPanModels	8 months ago
createPersonalizedModel.m	adapting for different index class	last year
detectOutput.m	Minor changes	last year
extractFullRes.m	function to extract all simulations results and print in csv file	last year
fastSetupCreator.m	Adapted old ordr of metabolites	10 months ago
getIndividualSizeName.m	adding checks and conversion of obs names to valid matlab ids	last year
getMappingInfo.m	adding lost index	last year
guidedSim.m	changing warning mex to be more general	2 years ago
initMgPipe.m	Merge pull request #1194 from Fede-edef/mgFix	last year
loadUncModels.m	improving function help	last year
makeDummyModel.m	removing dummy.rev as became deprecated	2 years ago
mgSimResCollect.m	refining plotting	last year
microbiotaModelSimulator.m	Corrected indentation, silenced output (added ;), fixed model name	10 months ago
parsave.m	Documentation formatting of mgpipe	2 years ago
plotMappingInfo.m	refining plotting	last year

README.md

mgPipe

Introduction

mgPipe is a MATLAB based pipeline to integrate microbial abundances (coming from metagenomic data) with constraint-based modeling, creating individuals' personalized models. The pipeline is divided into 3 parts: [PART 1] Analysis of individuals' specific microbes abundances is computed. Individuals' metabolic diversity in relation to microbiota size and disease presence as well as Classical multidimensional scaling (PCoA) on individuals' reaction repertoire are examples. [PART 2]: 1 Constructing a global metabolic model (setup) containing all the microbes listed in the study. 2 Building individuals' specific models integrating abundance data retrieved from metagenomics. For each organism, reactions are coupled to their objective function. [PART 3] Simulations under different diet regimesand analysis of the predicted metabolic profile (PCoA of computed MNPCs of individuals).



WARNING: Please take into consideration only the files listed in this document. Everything present in the folder but not listed and explained in this document is to be considered not relevant or obsolete.

Requirements

mgPipe requires `Matrix Laboratory`, the `Parallel Computing Toolbox`, as well as, the COBRA Toolbox to be installed. Please refer to the installation instructions. The usage of `ILOG CPLEX` solver is strongly advised to obtain the best speed performance (required for the fastFVA.m function).

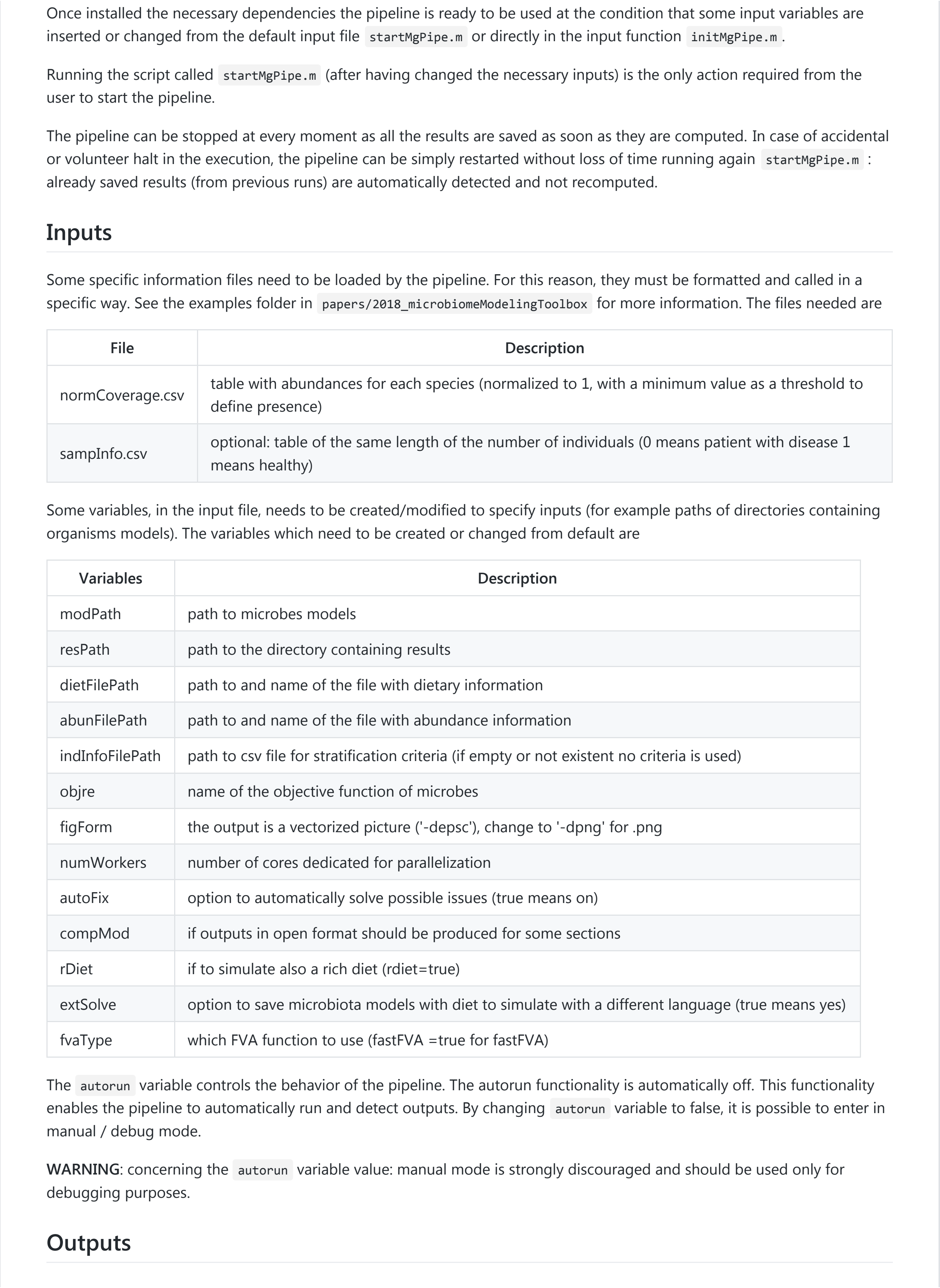
MgPipe was created (and tested) for AGORA 1.0 please first download AGORA version 1.0 from https://www.vmh.life/#downloadview and place the mat files into a dedicated folder.

Main Folder Structure and Files

The following files are essential for the usage of the pipeline and are supplied in the current folder and in `papers/2018_microbiomeModelingToolbox`

Filename	Purpose
startMgPipe.m	*driver, containing all the input variables, to be modified by the user*
initMgPipe.m	*function containing all the input variables launching the pipeline*
loadUncModels.m	*function to load unconstrain and prune the models*
fastSetupCreator.m	*function to create "global" setup*
checkNomenConsist.m	*function to check that microbes have the right nomenclature*
detectOutput.m	*function to check if a specific file was already created and saved*
getIndividualSizeName.m	*get information on number and ID of organisms and individuals*
addMicrobeCommunityBiomass.m	*function to add community biomass*
mgPipe.m	*pipeline*
parsave.m	*function to allow object saving in parallel loops*
getMappingInfo.m	*function to extract information from the mapping*
plotMappingInfo.m	*function plot extracted information from the mapping*
createPersonalizedModel.m	*function to create personalized models*
microbiotaModelSimulator.m	*function to simulate under different diets the created models (called from mgPipe)*
makeDummyModel.m	*function to create a dummy model*
mgSimResCollect.m	*function to collect and output simulation results*
extractFullRes.m	*function to retrieve and export all the results (fluxes) computed during the simulations*
README.md	*this file*
useDiet.m	*function to impose a specific diet and add essential elements to microbiota models*
adaptVMHDietToAGORA.m	*function to convert a specific diet from VMH into an AGORA compatible one*
compfile	*Results subfolder: contains objects saved in open format*

Usage



Once installed the necessary dependencies the pipeline is ready to be used at the condition that some input variables are inserted or changed from the default input file `startMgPipe.m` or directly in the input function `initMgPipe.m`.

Running the script called `startMgPipe.m` (after having changed the necessary inputs) is the only action required from the user to start the pipeline.

The pipeline can be stopped at every moment as all the results are saved as soon as they are computed. In case of accidental or volunteer halt in the execution, the pipeline can be simply restarted without loss of time running again `startMgPipe.m` : already saved results (from previous runs) are automatically detected and not recomputed.

Inputs

Some specific information files need to be loaded by the pipeline. For this reason, they must be formatted and called in a specific way. See the examples folder in `papers/2018_microbiomeModelingToolbox` for more information. The files needed are

File	Description
normCoverage.csv	table with abundances for each species (normalized to 1, with a minimum value as a threshold to define presence)
sampInfo.csv	optional: table of the same length of the number of individuals (0 means patient with disease 1 means healthy)

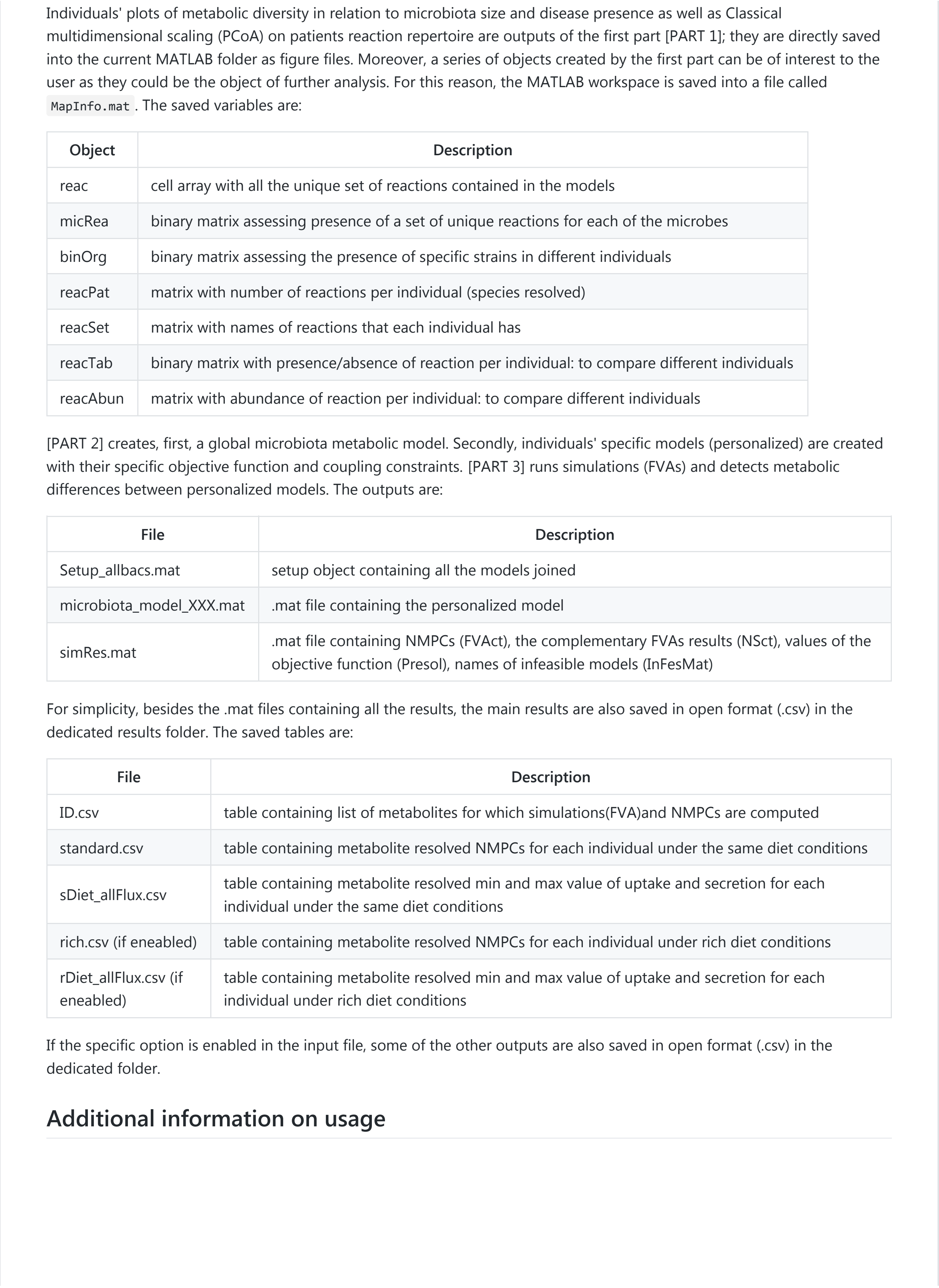
Some variables, in the input file, needs to be created/modified to specify inputs (for example paths of directories containing organisms models). The variables which need to be created or changed from default are

Variables	Description
modPath	path to microbes models
resPath	path to the directory containing results
dietFilePath	path to and name of the file with dietary information
abunFilePath	path to and name of the file with abundance information
indInfoFilePath	path to csv file for stratification criteria (if empty or not existent no criteria is used)
objre	name of the objective function of microbes
figForm	the output is a vectorized picture ('-depsc'), change to '-dpng' for .png
numWorkers	number of cores dedicated for parallelization
autoFix	option to automatically solve possible issues (true means on)
compMod	if outputs in open format should be produced for some sections
rDiet	if to simulate also a rich diet (rdiet=true)
extSolve	option to save microbiota models with diet to simulate with a different language (true means yes)
fvaType	which FVA function to use (fastFVA =true for fastFVA)

The `autorun` variable controls the behavior of the pipeline. The autorun functionality is automatically off. This functionality enables the pipeline to automatically run and detect outputs. By changing `autorun` variable to false, it is possible to enter in manual / debug mode.

WARNING: concerning the `autorun` variable value: manual mode is strongly discouraged and should be used only for debugging purposes.

Outputs



Individuals' plots of metabolic diversity in relation to microbiota size and disease presence as well as Classical multidimensional scaling (PCoA) on patients reaction repertoire are outputs of the first part [PART 1]; they are directly saved into the current MATLAB folder as figure files. Moreover, a series of objects created by the first part can be of interest to the user as they could be the object of further analysis. For this reason, the MATLAB workspace is saved into a file called `MapInfo.mat`. The saved variables are:

Object	Description
reac	cell array with all the unique set of reactions contained in the models
micRea	binary matrix assessing presence of a set of unique reactions for each of the microbes
binOrg	binary matrix assessing the presence of specific strains in different individuals
reacPat	matrix with number of reactions per individual (species resolved)
reacSet	matrix with names of reactions that each individual has
reacTab	binary matrix with presence/absence of reaction per individual: to compare different individuals
reacAbun	matrix with abundance of reaction per individual: to compare different individuals

[PART 2] creates, first, a global microbiota metabolic model. Secondly, individuals' specific models (personalized) are created with their specific objective function and coupling constraints. [PART 3] runs simulations (FVAs) and detects metabolic differences between personalized models. The outputs are:

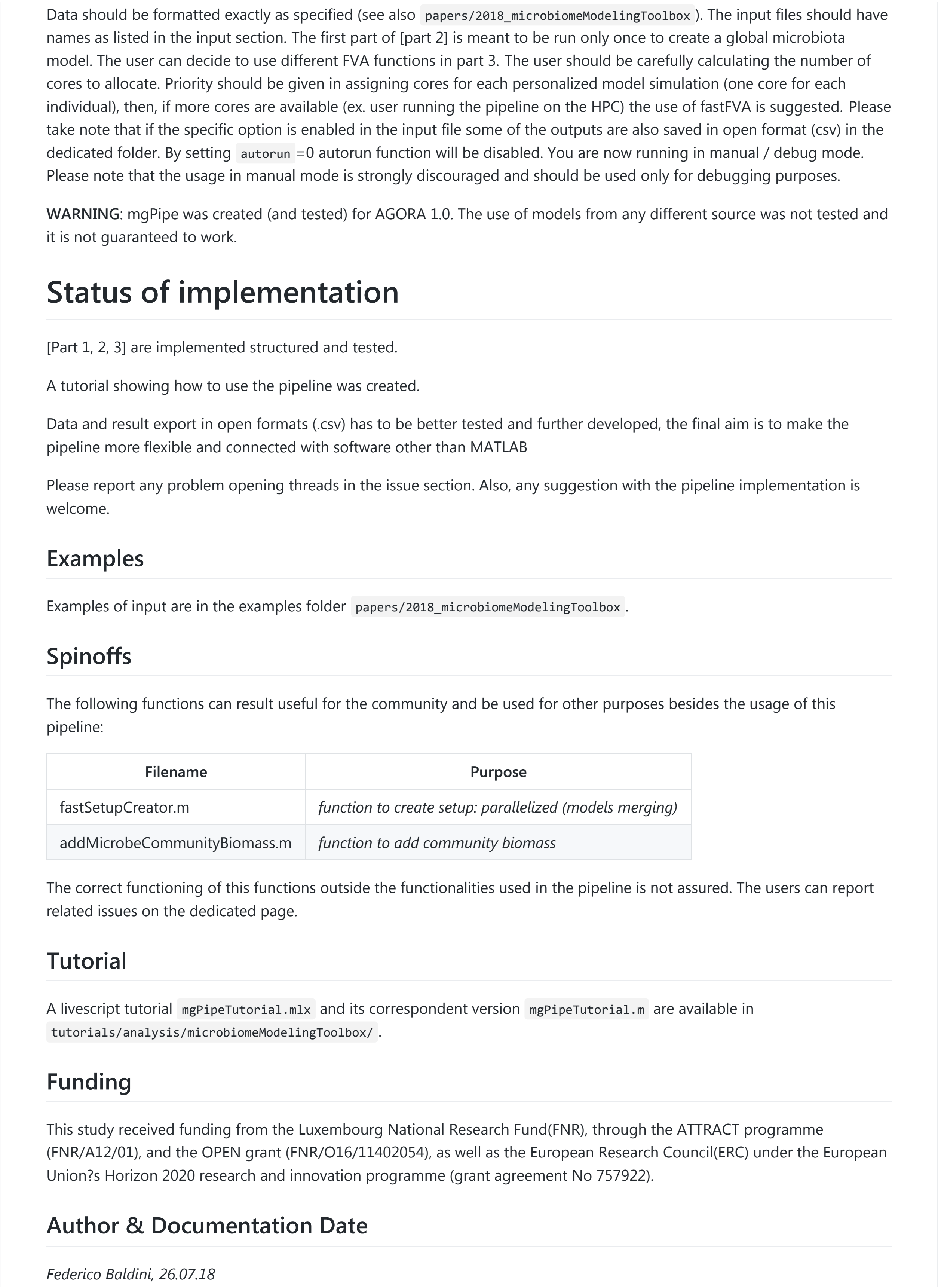
File	Description
Setup_allbacs.mat	setup object containing all the models joined
microbiota_model_XXX.mat	.mat file containing the personalized model
simRes.mat	.mat file containing NMPCs (FVAct), the complementary FVAs results (NSct), values of the objective function (Presol), names of infeasible models (InFesMat)

For simplicity, besides the .mat files containing all the results, the main results are also saved in open format (.csv) in the dedicated results folder. The saved tables are:

File	Description
ID.csv	table containing list of metabolites for which simulations(FVA)and NMPCs are computed
standard.csv	table containing metabolite resolved NMPCs for each individual under the same diet conditions
sDiet_allFlux.csv	table containing metabolite resolved min and max value of uptake and secretion for each individual under the same diet conditions
rich.csv (if eneabled)	table containing metabolite resolved NMPCs for each individual under rich diet conditions
rDiet_allFlux.csv (if eneabled)	table containing metabolite resolved min and max value of uptake and secretion for each individual under rich diet conditions

If the specific option is enabled in the input file, some of the other outputs are also saved in open format (.csv) in the dedicated folder.

Additional information on usage



Data should be formatted exactly as specified (see also `papers/2018_microbiomeModelingToolbox`). The input files should have names as listed in the input section. The first part of [part 2] is meant to be run only once to create a global microbiota model. The user can decide to use different FVA functions in part 3. The user should be carefully calculating the number of cores to allocate. Priority should be given in assigning cores for each personalized model simulation (one core for each individual), then, if more cores are available (ex. user running the pipeline on the HPC) the use of fastFVA is suggested. Please take note that if the specific option is enabled in the input file some of the outputs are also saved in open format (csv) in the dedicated folder. By setting `autorun` =0 autorun function will be disabled. You are now running in manual / debug mode. Please note that the usage in manual mode is strongly discouraged and should be used only for debugging purposes.

WARNING: mgPipe was created (and tested) for AGORA 1.0. The use of models from any different source was not tested and it is not guaranteed to work.

Status of implementation

[Part 1, 2, 3] are implemented structured and tested.

A tutorial showing how to use the pipeline was created.

Data and result export in open formats (.csv) has to be better tested and further developed, the final aim is to make the pipeline more flexible and connected with software other than MATLAB

Please report any problem opening threads in the issue section. Also, any suggestion with the pipeline implementation is welcome.

Examples

Examples of input are in the examples folder `papers/2018_microbiomeModelingToolbox`.

Spinoffs

The following functions can result useful for the community and be used for other purposes besides the usage of this pipeline:

Filename	Purpose
fastSetupCreator.m	*function to create setup: parallelized (models merging)*
addMicrobeCommunityBiomass.m	*function to add community biomass*

The correct functioning of this functions outside the functionalities used in the pipeline is not assured. The users can report related issues on the dedicated page.

Tutorial

A livescript tutorial `mgPipeTutorial.mlx` and its correspondent version `mgPipeTutorial.m` are available in `tutorials/analysis/microbiomeModelingToolbox/`.

Funding

This study received funding from the Luxembourg National Research Fund(FNR), through the ATTRACT programme (FNR/A12/01), and the OPEN grant (FNR/O16/11402054), as well as the European Research Council(ERC) under the European Union?s Horizon 2020 research and innovation programme (grant agreement No 757922).

Author & Documentation Date

Federico Baldini, 26.07.18

Chapter 3

mgPipe: from microbial community sequencing data to personalized microbiota metabolic models creation and interrogation

Manuscript in preparation

Abstract

Motivation

A systematic application of constraint-based modeling to functionally analyze microbial sequencing data is not yet by default performed. Such application, requires deep knowledge in metagenomics and in the metabolic modeling fields, therefore limiting the number of investigators able to perform it.

Results

We created a pipeline (mgPipe) to automatically perform Personalized Microbial Community Metabolic Modeling (PMCM) from whole genome and marker gene sequencing data. mgPipe extends the functionality of The Microbiome Modeling Toolbox which is integrated into the

constraint-based reconstruction and analysis (COBRA) toolbox.

Availability and implementation

mgPipe and the tutorials at https://gitlab.com/ithiele/fpipe.

3.1 Introduction

Microbial community sequencing either performed with whole-genome (WGS) or with marker gene sequencing (MGS), became an established method to study the microbial composition of samples [41].

Methods allowing the study of the metabolic functions of microbial communities were recently developed [12]. Such methods build upon the constraint-based reconstruction and analysis (COBRA) approach. COBRA modeling has its basis in genome-scale reconstruction and predictions of phenotypic properties [84]. The COBRA approach, historically, focused mainly on single organisms modeling, but recently, methods able to create and interrogate multispecies models were developed [17]. The Microbiome Modeling Toolbox, for example, allows the creation of personalized microbiota models integrating relative abundances into a compartmentalized COBRA multispecies framework. These models can be interrogated retrieving the metabolic potential associated with a specific microbial composition [12].

However, although dedicated tools for studying microbial composition and related metabolic functions have been created, such tools are mainly thought for experts in the field of metagenomics or COBRA modeling, limiting their possible uses from the whole microbiota scientific community to few investigators. Here we introduce mgPipe, a "one click button pipeline" able to perform creation and interrogation of personalized gut microbiota models starting from Illumina sequencing data. mgPipe extends the functionalities of the Microbiome Modeling Toolbox creating a facile pipeline that limits to the minimum the user's inputs and actions, enabling non-field experts to perform microbial community analysis.

3.2 Features

mgPipe is divided into three steps.

In the first step, microbial abundances are determined from raw sequencing data. For whole genome sequencing, a referenced mapping is performed using bwa [113] to align reads onto a reference genome composed of the concatenation of all the reference organism genomes as already implemented in [16, 78]. Coverage, relative abundances at a strain taxonomic level are obtained. For 16S rRNA gene sequencing, reads are merged and filtered using the dedicated pipeline MeFit [145] and classified using SPINGO [8] as implemented in [13]. Relative abundances at a genus and species taxonomic level are computed.

In the second step, the Microbiome Modeling Toolbox mgPipe module is run [12]. Personalized microbiota models are created and interrogated obtaining, for each sample, information on metabolite resolved community potential. Briefly, for each sample, a community model is created, through a compartmentalization technique [172], joining all the models of the organisms found in the sample by the first part. A diet and fecal compartments are added allowing the input of a common diet and the secretion of metabolites. A community objective function is obtained averaging the sum of each organisms biomass for the relative abundances coefficients.

In the third and optional step, using particular stratification criteria, relative abundances and secretion profiles are analyzed producing a standard report. The analysis is comprehensive of techniques such as principal coordinate analysis (PCoA), FDR adjusted Wilcoxon rank-sum test (WRST), and feature selection techniques.

3.3 Implementation

mgPipe runs on Bash and is written in Bash, MATLAB (Mathworks, Inc.), R programming, and accompanied by comprehensive documentation and tutorials. mgPipe allows the extrapolation of microbial abundances from sequencing data and the consequent creation and

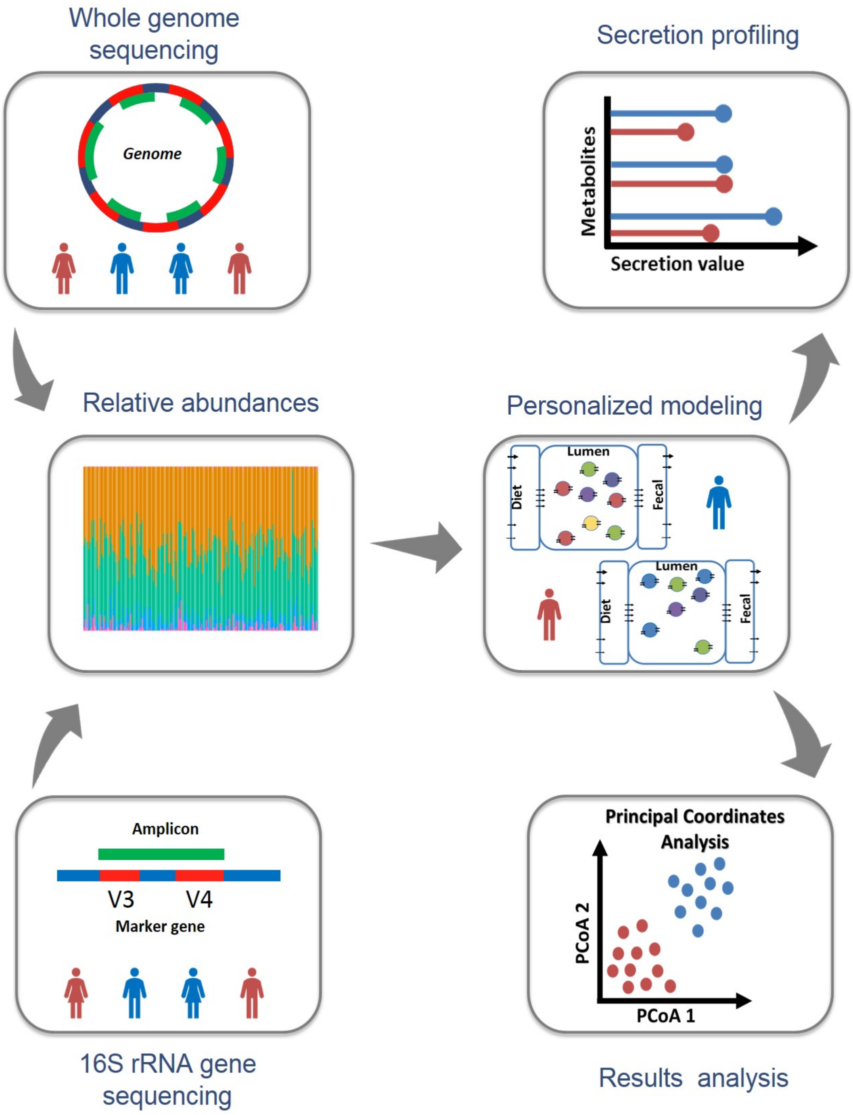


Figure 3.1: **Overview of the mgPipe features.**

interrogation of any number of personalized microbiota models. mgPipe is compatible with the AGORA genome-scale metabolic models. The metabolic reconstructions at a strain taxonomy can be obtained from the VMH website (https://vmh.life) and to use higher taxonomic levels (such as species, for 16S rRNA gene sequencing) the *createPanModels* function of the Microbiome Modeling Toolbox can be used to create pan-models.

3.4 Discussion

mgPipe, via the retrieval of microbial abundances from sequencing data, enables the user to systematically study the metabolic potential of microbial communities. Recent investigations which performed the same or similar steps, highlighted the importance of this approach in cohort studies to identify possible altered metabolic functionalities of microbial communities.

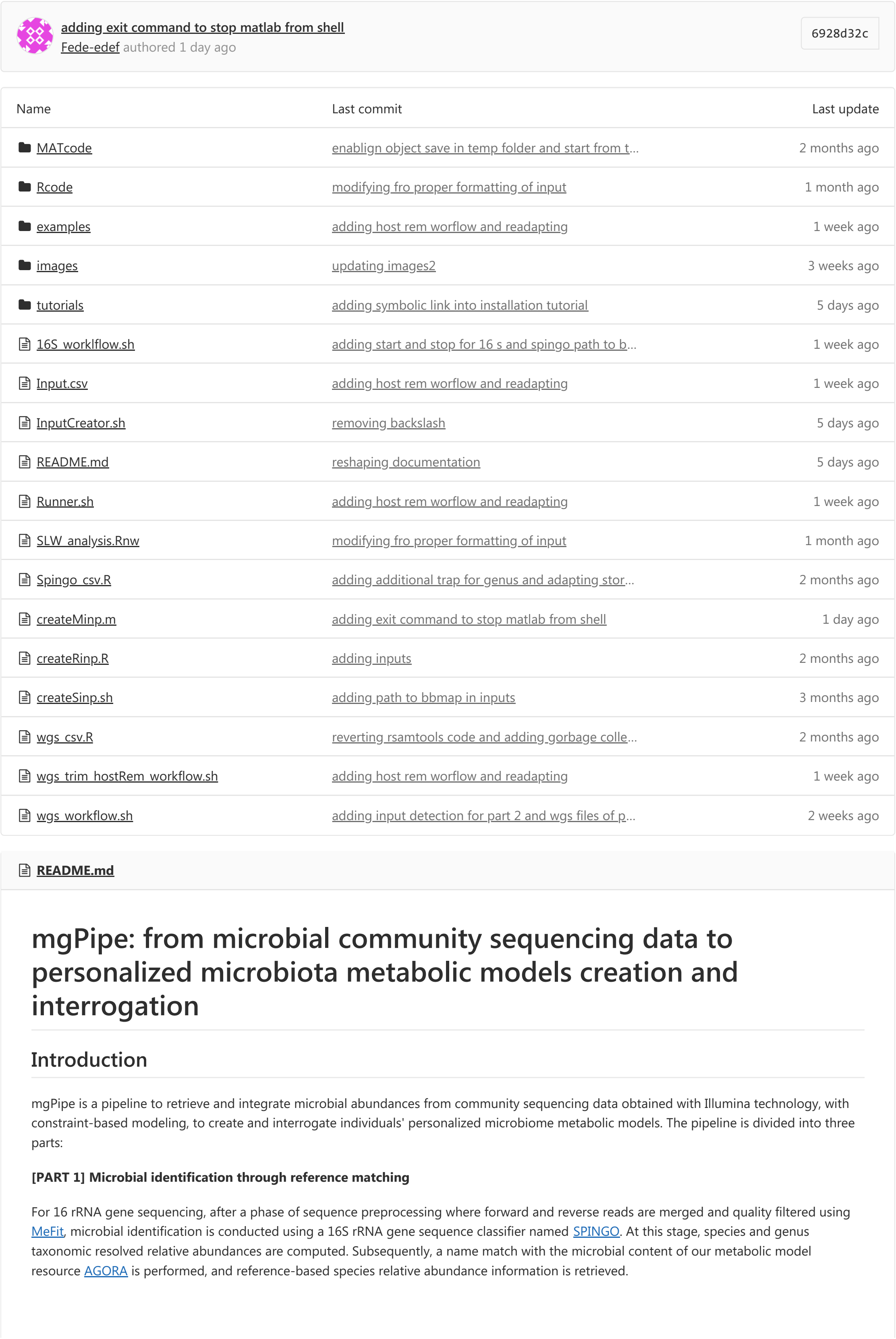
3.5 Funding

This study received funding from the Luxembourg National Research Fund (FNR), through the ATTRACT program [FNR/A12/01], as well as the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [grant agreement No 757922].

Conflict of Interest: none declared.

3.6 Documentation

This section is composed of the software documentation as available to the user.



adding exit command to stop matlab from shell
Fede-edef authored 1 day ago

6928d32c

Name	Last commit	Last update
MATcode	enablign object save in temp folder and start from t...	2 months ago
Rcode	modifying fro proper formatting of input	1 month ago
examples	adding host rem worflow and readapting	1 week ago
images	updating images2	3 weeks ago
tutorials	adding symbolic link into installation tutorial	5 days ago
16S_worklflow.sh	adding start and stop for 16 s and spingo path to b...	1 week ago
Input.csv	adding host rem worflow and readapting	1 week ago
InputCreator.sh	removing backslash	5 days ago
README.md	reshaping documentation	5 days ago
Runner.sh	adding host rem worflow and readapting	1 week ago
SLW_analysis.Rnw	modifying fro proper formatting of input	1 month ago
Spingo_csv.R	adding additional trap for genus and adapting stor...	2 months ago
createMinp.m	adding exit command to stop matlab from shell	1 day ago
createRinp.R	adding inputs	2 months ago
createSinp.sh	adding path to bbmap in inputs	3 months ago
wgs_csv.R	reverting rsamtools code and adding gorbage colle...	2 months ago
wgs_trim_hostRem_workflow.sh	adding host rem worflow and readapting	1 week ago
wgs_workflow.sh	adding input detection for part 2 and wgs files of p...	2 weeks ago

README.md

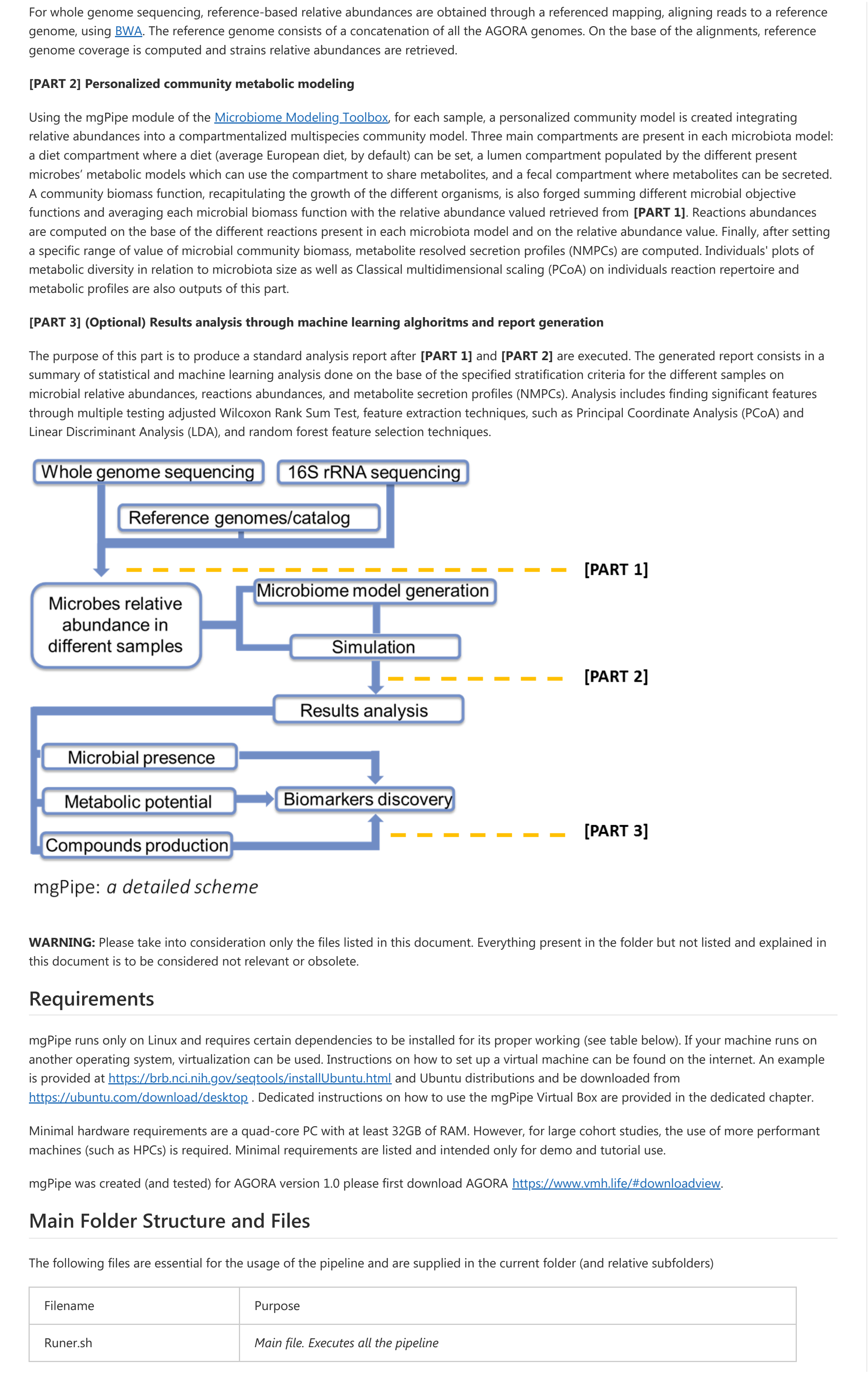
mgPipe: from microbial community sequencing data to personalized microbiota metabolic models creation and interrogation

Introduction

mgPipe is a pipeline to retrieve and integrate microbial abundances from community sequencing data obtained with Illumina technology, with constraint-based modeling, to create and interrogate individuals' personalized microbiome metabolic models. The pipeline is divided into three parts:

[PART 1] Microbial identification through reference matching

For 16 rRNA gene sequencing, after a phase of sequence preprocessing where forward and reverse reads are merged and quality filtered using MeFit, microbial identification is conducted using a 16S rRNA gene sequence classifier named SPINGO. At this stage, species and genus taxonomic resolved relative abundances are computed. Subsequently, a name match with the microbial content of our metabolic model resource AGORA is performed, and reference-based species relative abundance information is retrieved.



For whole genome sequencing, reference-based relative abundances are obtained through a referenced mapping, aligning reads to a reference genome, using BWA. The reference genome consists of a concatenation of all the AGORA genomes. On the base of the alignments, reference genome coverage is computed and strains relative abundances are retrieved.

[PART 2] Personalized community metabolic modeling

Using the mgPipe module of the Microbiome Modeling Toolbox, for each sample, a personalized community model is created integrating relative abundances into a compartmentalized multispecies community model. Three main compartments are present in each microbiota model: a diet compartment where a diet (average European diet, by default) can be set, a lumen compartment populated by the different present microbes' metabolic models which can use the compartment to share metabolites, and a fecal compartment where metabolites can be secreted. A community biomass function, recapitulating the growth of the different organisms, is also forged summing different microbial objective functions and averaging each microbial biomass function with the relative abundance valued retrieved from **[PART 1]**. Reactions abundances are computed on the base of the different reactions present in each microbiota model and on the relative abundance value. Finally, after setting a specific range of value of microbial community biomass, metabolite resolved secretion profiles (NMPCs) are computed. Individuals' plots of metabolic diversity in relation to microbiota size as well as Classical multidimensional scaling (PCoA) on individuals reaction repertoire and metabolic profiles are also outputs of this part.

[PART 3] (Optional) Results analysis through machine learning alghoritms and report generation

The purpose of this part is to produce a standard analysis report after **[PART 1]** and **[PART 2]** are executed. The generated report consists in a summary of statistical and machine learning analysis done on the base of the specified stratification criteria for the different samples on microbial relative abundances, reactions abundances, and metabolite secretion profiles (NMPCs). Analysis includes finding significant features through multiple testing adjusted Wilcoxon Rank Sum Test, feature extraction techniques, such as Principal Coordinate Analysis (PCoA) and Linear Discriminant Analysis (LDA), and random forest feature selection techniques.

mgPipe: *a detailed scheme*

WARNING: Please take into consideration only the files listed in this document. Everything present in the folder but not listed and explained in this document is to be considered not relevant or obsolete.

Requirements

mgPipe runs only on Linux and requires certain dependencies to be installed for its proper working (see table below). If your machine runs on another operating system, virtualization can be used. Instructions on how to set up a virtual machine can be found on the internet. An example is provided at https://brb.nci.nih.gov/seqtools/installUbuntu.html and Ubuntu distributions and be downloaded from https://ubuntu.com/download/desktop . Dedicated instructions on how to use the mgPipe Virtual Box are provided in the dedicated chapter.

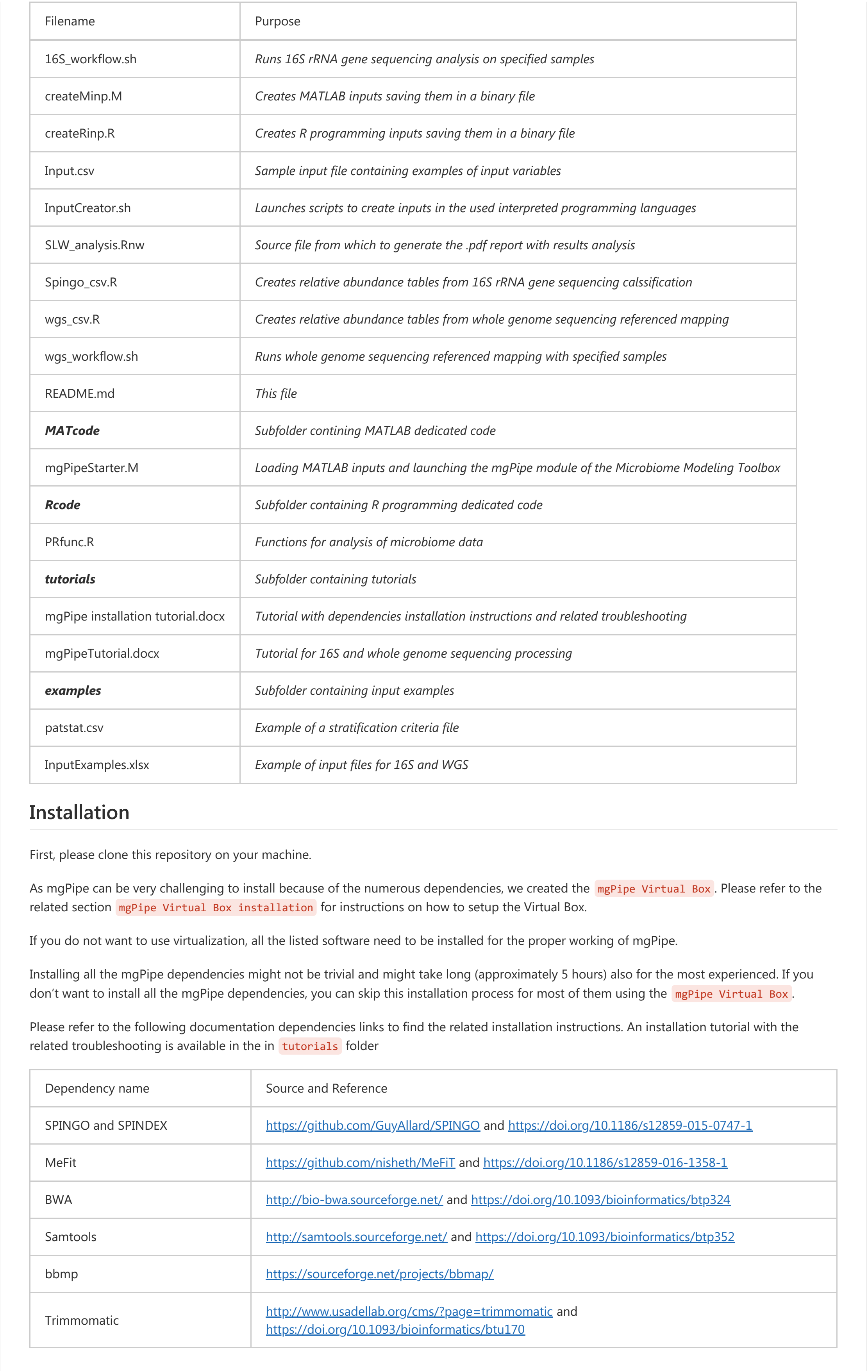
Minimal hardware requirements are a quad-core PC with at least 32GB of RAM. However, for large cohort studies, the use of more performant machines (such as HPCs) is required. Minimal requirements are listed and intended only for demo and tutorial use.

mgPipe was created (and tested) for AGORA version 1.0 please first download AGORA https://www.vmh.life/#downloadview.

Main Folder Structure and Files

The following files are essential for the usage of the pipeline and are supplied in the current folder (and relative subfolders)

Filename	Purpose
Runer.sh	*Main file. Executes all the pipeline*



Filename	Purpose
16S_workflow.sh	*Runs 16S rRNA gene sequencing analysis on specified samples*
createMinp.M	*Creates MATLAB inputs saving them in a binary file*
createRinp.R	*Creates R programming inputs saving them in a binary file*
Input.csv	*Sample input file containing examples of input variables*
InputCreator.sh	*Launches scripts to create inputs in the used interpreted programming languages*
SLW_analysis.Rnw	*Source file from which to generate the .pdf report with results analysis*
Spingo_csv.R	*Creates relative abundance tables from 16S rRNA gene sequencing calssification*
wgs_csv.R	*Creates relative abundance tables from whole genome sequencing referenced mapping*
wgs_workflow.sh	*Runs whole genome sequencing referenced mapping with specified samples*
README.md	*This file*
MATcode	*Subfolder contining MATLAB dedicated code*
mgPipeStarter.M	*Loading MATLAB inputs and launching the mgPipe module of the Microbiome Modeling Toolbox*
Rcode	*Subfolder containing R programming dedicated code*
PRfunc.R	*Functions for analysis of microbiome data*
tutorials	*Subfolder containing tutorials*
mgPipe installation tutorial.docx	*Tutorial with dependencies installation instructions and related troubleshooting*
mgPipeTutorial.docx	*Tutorial for 16S and whole genome sequencing processing*
examples	*Subfolder containing input examples*
patstat.csv	*Example of a stratification criteria file*
InputExamples.xlsx	*Example of input files for 16S and WGS*

Installation

First, please clone this repository on your machine.

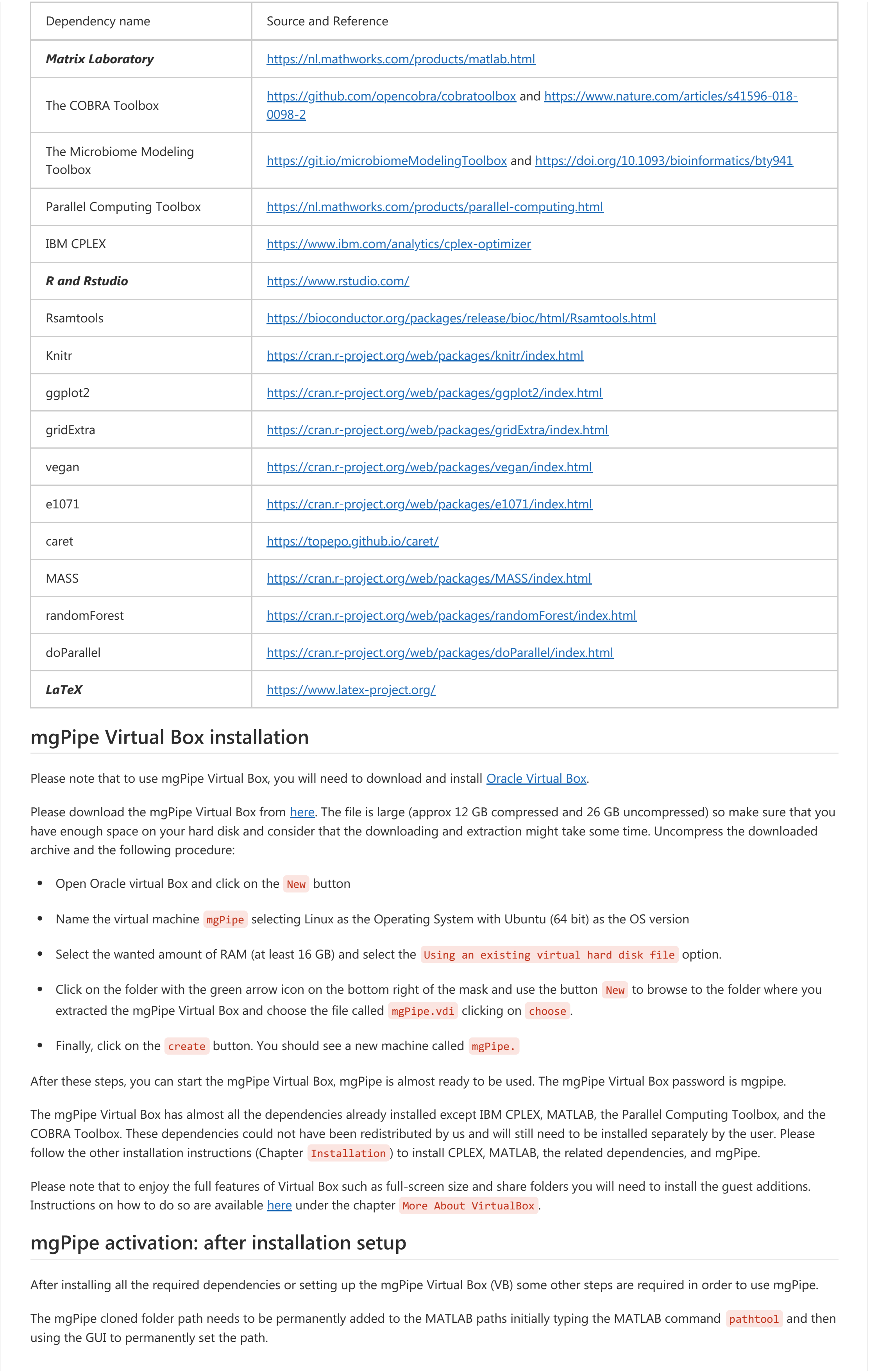
As mgPipe can be very challenging to install because of the numerous dependencies, we created the `mgPipe Virtual Box`. Please refer to the related section `mgPipe Virtual Box installation` for instructions on how to setup the Virtual Box.

If you do not want to use virtualization, all the listed software need to be installed for the proper working of mgPipe.

Installing all the mgPipe dependencies might not be trivial and might take long (approximately 5 hours) also for the most experienced. If you don't want to install all the mgPipe dependencies, you can skip this installation process for most of them using the `mgPipe Virtual Box`.

Please refer to the following documentation dependencies links to find the related installation instructions. An installation tutorial with the related troubleshooting is available in the in `tutorials` folder

Dependency name	Source and Reference
SPINGO and SPINDEX	https://github.com/GuyAllard/SPINGO and https://doi.org/10.1186/s12859-015-0747-1
MeFit	https://github.com/nisheth/MeFiT and https://doi.org/10.1186/s12859-016-1358-1
BWA	http://bio-bwa.sourceforge.net/ and https://doi.org/10.1093/bioinformatics/btp324
Samtools	http://samtools.sourceforge.net/ and https://doi.org/10.1093/bioinformatics/btp352
bbmp	https://sourceforge.net/projects/bbmap/
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic and https://doi.org/10.1093/bioinformatics/btu170



Dependency name	Source and Reference
Matrix Laboratory	https://nl.mathworks.com/products/matlab.html
The COBRA Toolbox	https://github.com/opencobra/cobratoolbox and https://www.nature.com/articles/s41596-018-0098-2
The Microbiome Modeling Toolbox	https://git.io/microbiomeModelingToolbox and https://doi.org/10.1093/bioinformatics/bty941
Parallel Computing Toolbox	https://nl.mathworks.com/products/parallel-computing.html
IBM CPLEX	https://www.ibm.com/analytics/cplex-optimizer
R and Rstudio	https://www.rstudio.com/
Rsamtools	https://bioconductor.org/packages/release/bioc/html/Rsamtools.html
Knitr	https://cran.r-project.org/web/packages/knitr/index.html
ggplot2	https://cran.r-project.org/web/packages/ggplot2/index.html
gridExtra	https://cran.r-project.org/web/packages/gridExtra/index.html
vegan	https://cran.r-project.org/web/packages/vegan/index.html
e1071	https://cran.r-project.org/web/packages/e1071/index.html
caret	https://topepo.github.io/caret/
MASS	https://cran.r-project.org/web/packages/MASS/index.html
randomForest	https://cran.r-project.org/web/packages/randomForest/index.html
doParallel	https://cran.r-project.org/web/packages/doParallel/index.html
LaTeX	https://www.latex-project.org/

mgPipe Virtual Box installation

Please note that to use mgPipe Virtual Box, you will need to download and install Oracle Virtual Box.

Please download the mgPipe Virtual Box from here. The file is large (approx 12 GB compressed and 26 GB uncompressed) so make sure that you have enough space on your hard disk and consider that the downloading and extraction might take some time. Uncompress the downloaded archive and the following procedure:

- Open Oracle virtual Box and click on the `New` button
- Name the virtual machine `mgPipe` selecting Linux as the Operating System with Ubuntu (64 bit) as the OS version
- Select the wanted amount of RAM (at least 16 GB) and select the `Using an existing virtual hard disk file` option.
- Click on the folder with the green arrow icon on the bottom right of the mask and use the button `New` to browse to the folder where you extracted the mgPipe Virtual Box and choose the file called `mgPipe.vdi` clicking on `choose`.
- Finally, click on the `create` button. You should see a new machine called `mgPipe.`

After these steps, you can start the mgPipe Virtual Box, mgPipe is almost ready to be used. The mgPipe Virtual Box password is mgpipe.

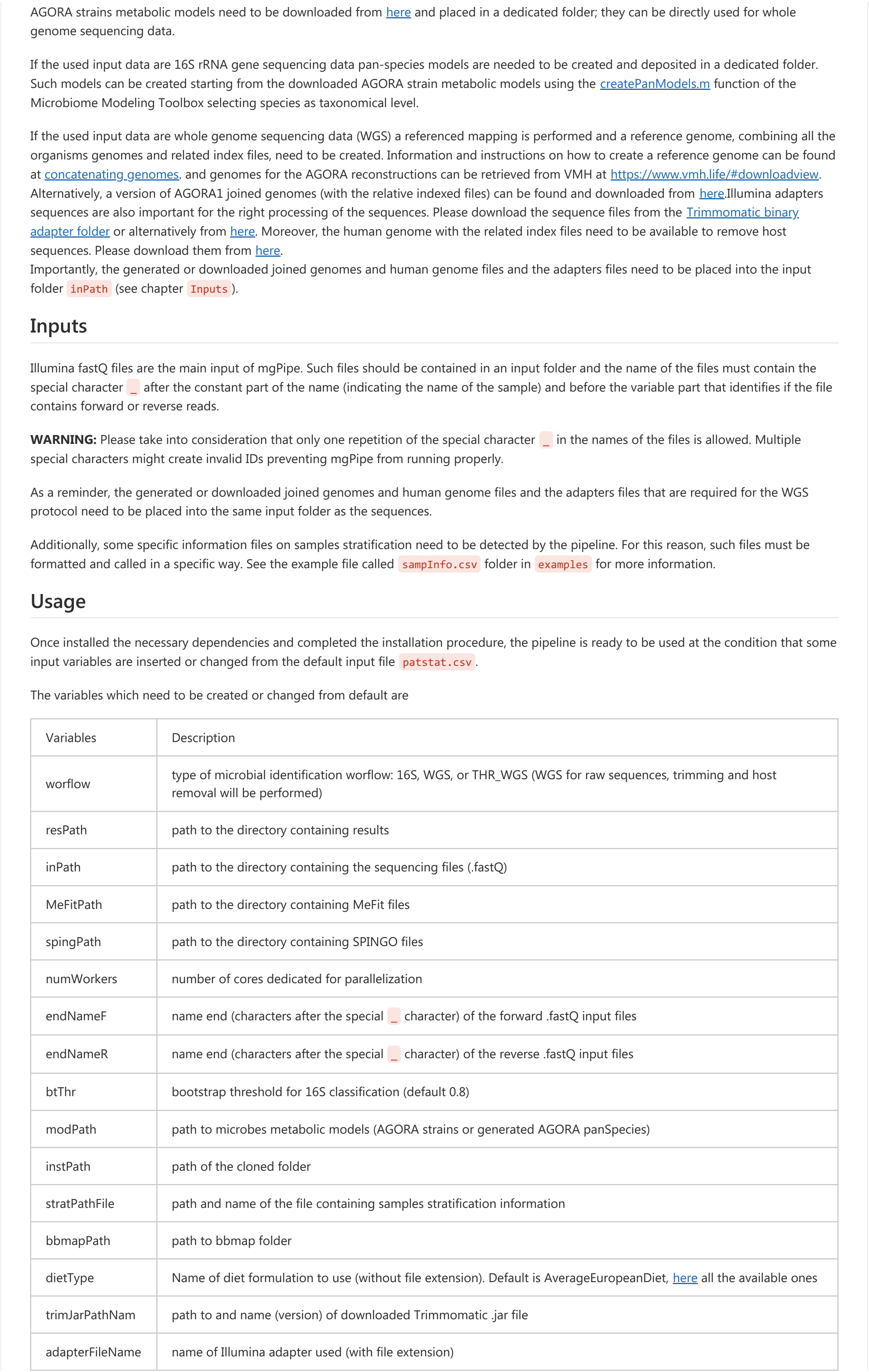
The mgPipe Virtual Box has almost all the dependencies already installed except IBM CPLEX, MATLAB, the Parallel Computing Toolbox, and the COBRA Toolbox. These dependencies could not have been redistributed by us and will still need to be installed separately by the user. Please follow the other installation instructions (Chapter `Installation`) to install CPLEX, MATLAB, the related dependencies, and mgPipe.

Please note that to enjoy the full features of Virtual Box such as full-screen size and share folders you will need to install the guest additions. Instructions on how to do so are available here under the chapter `More About VirtualBox`.

mgPipe activation: after installation setup

After installing all the required dependencies or setting up the mgPipe Virtual Box (VB) some other steps are required in order to use mgPipe.

The mgPipe cloned folder path needs to be permanently added to the MATLAB paths initially typing the MATLAB command `pathtool` and then using the GUI to permanently set the path.



AGORA strains metabolic models need to be downloaded from here and placed in a dedicated folder; they can be directly used for whole genome sequencing data.

If the used input data are 16S rRNA gene sequencing data pan-species models are needed to be created and deposited in a dedicated folder. Such models can be created starting from the downloaded AGORA strain metabolic models using the createPanModels.m function of the Microbiome Modeling Toolbox selecting species as taxonomical level.

If the used input data are whole genome sequencing data (WGS) a referenced mapping is performed and a reference genome, combining all the organisms genomes and related index files, need to be created. Information and instructions on how to create a reference genome can be found at concatenating genomes. and genomes for the AGORA reconstructions can be retrieved from VMH at https://www.vmh.life/#downloadview. Alternatively, a version of AGORA1 joined genomes (with the relative indexed files) can be found and downloaded from here.Illumina adapters sequences are also important for the right processing of the sequences. Please download the sequence files from the Trimmomatic binary adapter folder or alternatively from here. Moreover, the human genome with the related index files need to be available to remove host sequences. Please download them from here.
Importantly, the generated or downloaded joined genomes and human genome files and the adapters files need to be placed into the input folder `inPath` (see chapter `Inputs`).

Inputs

Illumina fastQ files are the main input of mgPipe. Such files should be contained in an input folder and the name of the files must contain the special character `_` after the constant part of the name (indicating the name of the sample) and before the variable part that identifies if the file contains forward or reverse reads.

WARNING: Please take into consideration that only one repetition of the special character `_` in the names of the files is allowed. Multiple special characters might create invalid IDs preventing mgPipe from running properly.

As a reminder, the generated or downloaded joined genomes and human genome files and the adapters files that are required for the WGS protocol need to be placed into the same input folder as the sequences.

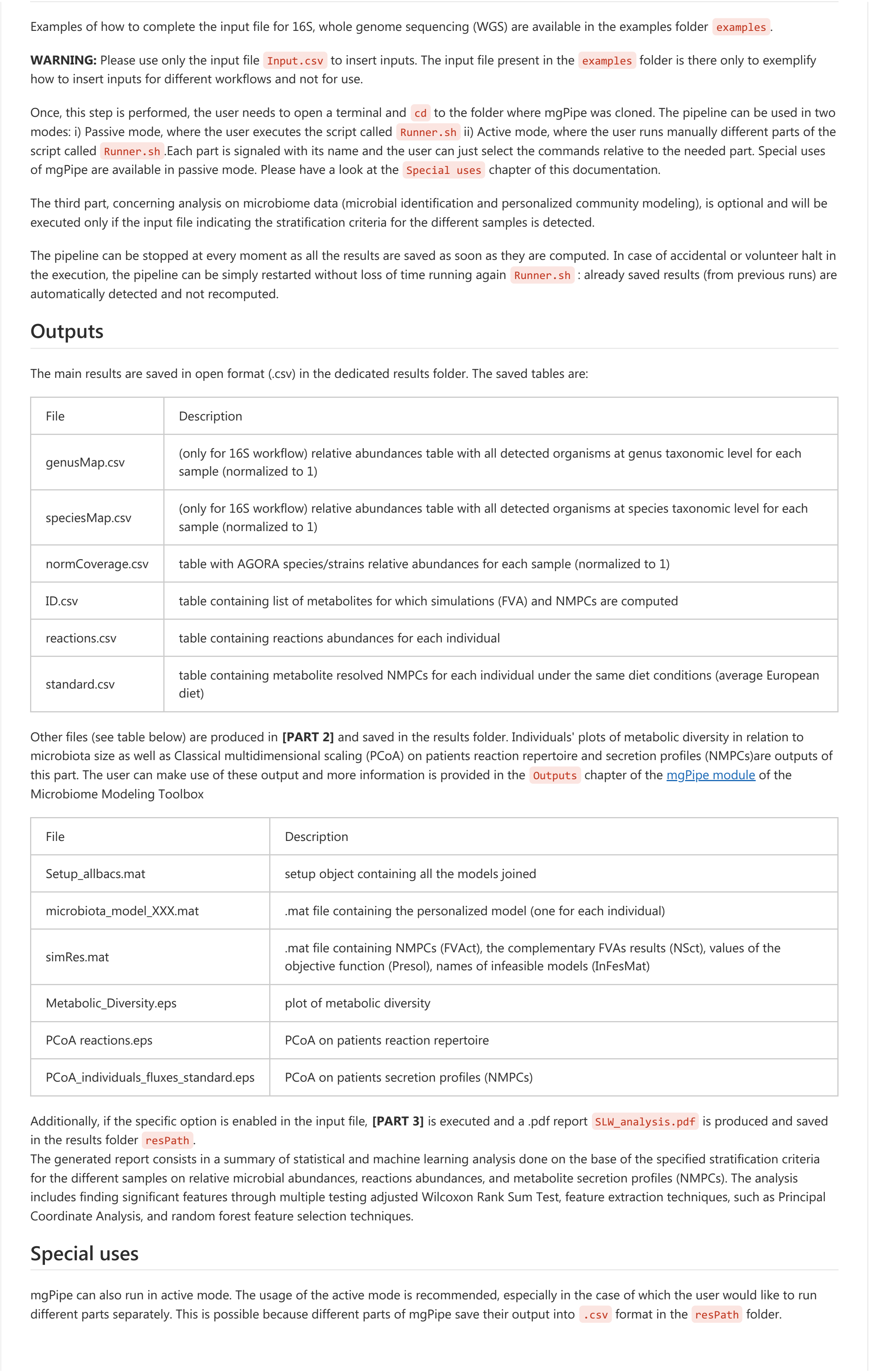
Additionally, some specific information files on samples stratification need to be detected by the pipeline. For this reason, such files must be formatted and called in a specific way. See the example file called `sampInfo.csv` folder in `examples` for more information.

Usage

Once installed the necessary dependencies and completed the installation procedure, the pipeline is ready to be used at the condition that some input variables are inserted or changed from the default input file `patstat.csv`.

The variables which need to be created or changed from default are

Variables	Description
worflow	type of microbial identification worflow: 16S, WGS, or THR_WGS (WGS for raw sequences, trimming and host removal will be performed)
resPath	path to the directory containing results
inPath	path to the directory containing the sequencing files (.fastQ)
MeFitPath	path to the directory containing MeFit files
spingPath	path to the directory containing SPINGO files
numWorkers	number of cores dedicated for parallelization
endNameF	name end (characters after the special `_` character) of the forward .fastQ input files
endNameR	name end (characters after the special `_` character) of the reverse .fastQ input files
btThr	bootstrap threshold for 16S classification (default 0.8)
modPath	path to microbes metabolic models (AGORA strains or generated AGORA panSpecies)
instPath	path of the cloned folder
stratPathFile	path and name of the file containing samples stratification information
bbmapPath	path to bbmap folder
dietType	Name of diet formulation to use (without file extension). Default is AverageEuropeanDiet, here all the available ones
trimJarPathNam	path to and name (version) of downloaded Trimmomatic .jar file
adapterFileName	name of Illumina adapter used (with file extension)



Examples of how to complete the input file for 16S, whole genome sequencing (WGS) are available in the examples folder `examples`.

WARNING: Please use only the input file `Input.csv` to insert inputs. The input file present in the `examples` folder is there only to exemplify how to insert inputs for different workflows and not for use.

Once, this step is performed, the user needs to open a terminal and `cd` to the folder where mgPipe was cloned. The pipeline can be used in two modes: i) Passive mode, where the user executes the script called `Runner.sh` ii) Active mode, where the user runs manually different parts of the script called `Runner.sh`.Each part is signaled with its name and the user can just select the commands relative to the needed part. Special uses of mgPipe are available in passive mode. Please have a look at the `Special uses` chapter of this documentation.

The third part, concerning analysis on microbiome data (microbial identification and personalized community modeling), is optional and will be executed only if the input file indicating the stratification criteria for the different samples is detected.

The pipeline can be stopped at every moment as all the results are saved as soon as they are computed. In case of accidental or volunteer halt in the execution, the pipeline can be simply restarted without loss of time running again `Runner.sh` : already saved results (from previous runs) are automatically detected and not recomputed.

Outputs

The main results are saved in open format (.csv) in the dedicated results folder. The saved tables are:

File	Description
genusMap.csv	(only for 16S workflow) relative abundances table with all detected organisms at genus taxonomic level for each sample (normalized to 1)
speciesMap.csv	(only for 16S workflow) relative abundances table with all detected organisms at species taxonomic level for each sample (normalized to 1)
normCoverage.csv	table with AGORA species/strains relative abundances for each sample (normalized to 1)
ID.csv	table containing list of metabolites for which simulations (FVA) and NMPCs are computed
reactions.csv	table containing reactions abundances for each individual
standard.csv	table containing metabolite resolved NMPCs for each individual under the same diet conditions (average European diet)

Other files (see table below) are produced in **[PART 2]** and saved in the results folder. Individuals' plots of metabolic diversity in relation to microbiota size as well as Classical multidimensional scaling (PCoA) on patients reaction repertoire and secretion profiles (NMPCs)are outputs of this part. The user can make use of these output and more information is provided in the `Outputs` chapter of the mgPipe module of the Microbiome Modeling Toolbox

File	Description
Setup_allbacs.mat	setup object containing all the models joined
microbiota_model_XXX.mat	.mat file containing the personalized model (one for each individual)
simRes.mat	.mat file containing NMPCs (FVAct), the complementary FVAs results (NSct), values of the objective function (Presol), names of infeasible models (InFesMat)
Metabolic_Diversity.eps	plot of metabolic diversity
PCoA reactions.eps	PCoA on patients reaction repertoire
PCoA_individuals_fluxes_standard.eps	PCoA on patients secretion profiles (NMPCs)

Additionally, if the specific option is enabled in the input file, **[PART 3]** is executed and a .pdf report `SLW_analysis.pdf` is produced and saved in the results folder `resPath`.
The generated report consists in a summary of statistical and machine learning analysis done on the base of the specified stratification criteria for the different samples on relative microbial abundances, reactions abundances, and metabolite secretion profiles (NMPCs). The analysis includes finding significant features through multiple testing adjusted Wilcoxon Rank Sum Test, feature extraction techniques, such as Principal Coordinate Analysis, and random forest feature selection techniques.

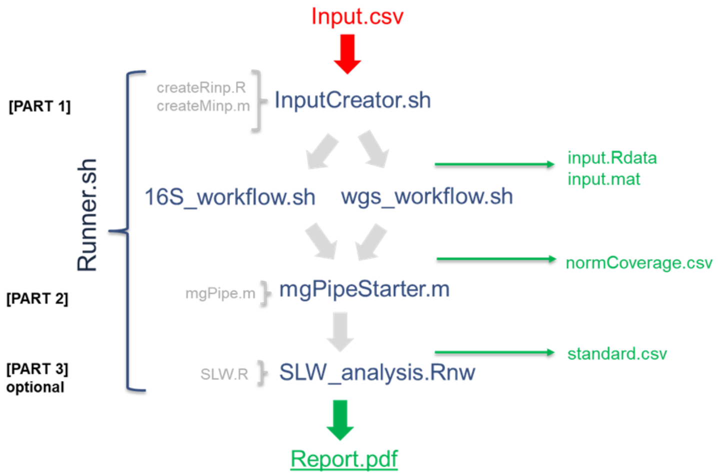
Special uses

mgPipe can also run in active mode. The usage of the active mode is recommended, especially in the case of which the user would like to run different parts separately. This is possible because different parts of mgPipe save their output into `.csv` format in the `resPath` folder.

A typical example is when the user is running the mgPipe virtual Box on a guest OS (maybe even not even UNIX based) where CPLEX MATLAB and the related dependencies (COBRA Toolbox, Microbiome Modeling Toolbox) are already installed. In this case, the user can run **[PART 1]** of `Runner.sh` in active mode on the mgPipe virtual Box, collect the output from **[PART 1]** from the `resPath` folder, run manually the MATLAB mgPipe module of the Microbiome Modeling Toolbox, collect the output, and use it back on the mgPipe virtual Box to execute **[PART 3]**. Headers indicating the beginning of each part are indicated in the file `Runner.sh`

Graphical abstract

The figure below graphically summarizes the pipeline code structure, main input and outputs.



mgPipe: *a detailed scheme indicating the software structure, main code workflows in blue, secondary code in green, the main input in red and the main outputs in green.*

Examples

Examples of input file is in the folder `mgPipe/examples`. The file can be directly modified by the user to run the pipeline.

Status of implementation

Please report any problem opening threads in the issue section.

Tutorial

A tutorial `mgPipeTutorial.docx` is available in `mgPipe/tutorials/`.

An installation tutorial `mgPipe installation tutorial.docx` is available in `mgPipe/tutorials/`.

Funding

This study received funding from the Luxembourg National Research Fund(FNR), through the ATTRACT programme (FNR/A12/01), and the OPEN grant (FNR/O16/11402054), as well as the European Research Council(ERC) under the European Union?s Horizon 2020 research and innovation programme (grant agreement No 757922).

Author & Documentation Date

Federico Baldini, 03.10.19

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, Esch-sur-Alzette, Luxembourg

federico.baldini@uni.lu

Chapter 4

Parkinson's disease-associated alterations of the gut microbiome can invoke disease-relevant metabolic changes

Baldini, F., Hertel, J., Sandt, E., Thinnes, C., Neuberger-Castillo, L., Pavelka, L., Betsou F., Krüger R. and Thiele, I. (2019). Parkinson's disease-associated alterations of the gut microbiome can invoke disease-relevant metabolic changes. bioRxiv, 691030.

Abstract

Parkinson's disease (PD) is a systemic disease clinically defined by the degeneration of dopaminergic neurons in the brain. While alterations in the gut microbiome composition have been reported in PD, their functional consequences remain unclear. Herein, we first analysed the gut microbiome of patients and healthy controls by 16S rRNA gene sequencing of stool samples from the Luxembourg Parkinson's study (n=147 typical PD cases, n=162 controls). All individuals underwent detailed clinical assessment, including neurological examinations and neuropsychological tests followed by self-reporting questionnaires. Second, we predicted the potential secretion for 129 microbial metabolites through personalised metabolic modelling using the microbiome data and genome-scale metabolic reconstructions of human gut microbes. Our key results include: 1. eight genera and nine species changed significantly in their relative abundances between PD patients and healthy controls. 2. PD-associated microbial patterns statistically depended on sex, age, BMI, and constipation. The relative abundances of *Bilophila* and *Paraprevotella* were significantly associated with the Hoehn and Yahr staging after controlling for the disease duration. In contrast, dopaminergic medication had no detectable effect on the PD microbiome composition. 3. Personalised metabolic modelling of the gut microbiomes revealed PD-associated metabolic patterns in secretion potential of nine microbial metabolites in PD, including increased methionine and cysteinylglycine. The microbial pantothenic acid production potential was linked to the presence of specific non motor symptoms and attributed to individual bacteria, such as *Akkermansia muciniphila* and *Bilophila wardswarthia*. Our results suggest that PD-associated alterations of gut microbiome could translate into functional differences affecting host metabolism and disease phenotype.

4.1 Introduction

Parkinson's Disease (PD) is a complex multifactorial disease, with both genetic and environmental factors contributing to the evolution and progression of the disease [99]. While several studies have elucidated the role of genetic factors in the pathogenesis of the disease [104, 25, 139, 51], the role and the contribution of various environmental and lifestyle factors are still not completely understood [64]. Importantly, about 60% of the PD patients suffer from constipation [59], which can start up to 20 years before the diagnosis and is one of the prodromal syndromes [161, 38]. The human being is considered to be a superorganism recognising a complex interplay between the host and microbes [169]. For instance, the human gut microbiome has been shown to complement the host with essential functions (trophic, metabolic, protective) and to influence the host's central nervous system (CNS) via the gut-brain axis through the modulation of neural pathways and GABAergic and serotoninergic signalling systems [36]. Recent studies have reported an altered gut composition in PD [77, 102, 162, 21, 87, 91, 147, 82, 14]. One of these studies has been conducted using samples from recently diagnosed, drug-naive patients [21]. These studies have demonstrated that PD patients have an altered microbiome composition, compared to age-matched controls. However, the functional implications of the altered microbiome remain to be elucidated, e.g., using animal models [158]. A complementary approach is computational modelling, or constraint based reconstruction and analyses (COBRA) [138], of microbiome-level metabolism.

In this approach, metabolic reconstructions for hundreds of gut microbes [117] are combined based on microbiome data [12, 84]. Flux balance analysis (FBA) [138] is then used to compute, e.g., possible metabolite uptake or secretion flux rates of each microbiome model (microbiome metabolic profile) [78] or to study of microbial metabolic interactions (cross feedings) [105, 80]. This approach has been applied to various microbiome data sets to gain functional insights [181, 78, 86], including for PD where we propose that microbial sulphur metabolism could contribute to changes in the blood metabolome of PD patients [86]. In the present study, we aim at investigating microbial changes associated with PD while focusing on possible covariates influencing microbial composition and at proposing functional, i.e., metabolic, consequences arising from the microbiome changes. First, we analysed the faecal microbial composition of PD patients and controls from the Luxembourg Parkinson's study [88] (Figure 4.1). Second, based on the observed significant differences in the composition of microbial communities between PD patients and controls, we created and interrogated personalised computational models representing the metabolism of each individual's microbial community. We demonstrate that the combined microbial composition and functional metabolite analysis provides novel hypotheses on microbial changes associated with PD and disease severity, enabling future mechanism-based experiments

4.2 Results

The Luxembourg Parkinson's Disease study includes patients with typical PD and atypical parkinsonism, as well as matched healthy control subjects from Luxembourg and its neighbouring regions from a broad age-range [88]. For the present study, we focused on typical PD patients and healthy controls over the age of 50 (Table 4.1, Methods). Stool samples were analysed for 147 PD patients and 162 controls using 16S rRNA gene sequences (Methods: Analysis of the microbial composition with 16S rRNA gene sequencing).

4.2.1 Species and genus level changes in PD microbiomes

We investigated disease-associated microbial changes at the species level. We found that the mean species diversity (i.e., the alpha-diversity) did not significantly differ between PD cases and controls (b=-0.04351, 95%-CI:(-.107;0.177), p=0.177), in agreement with earlier

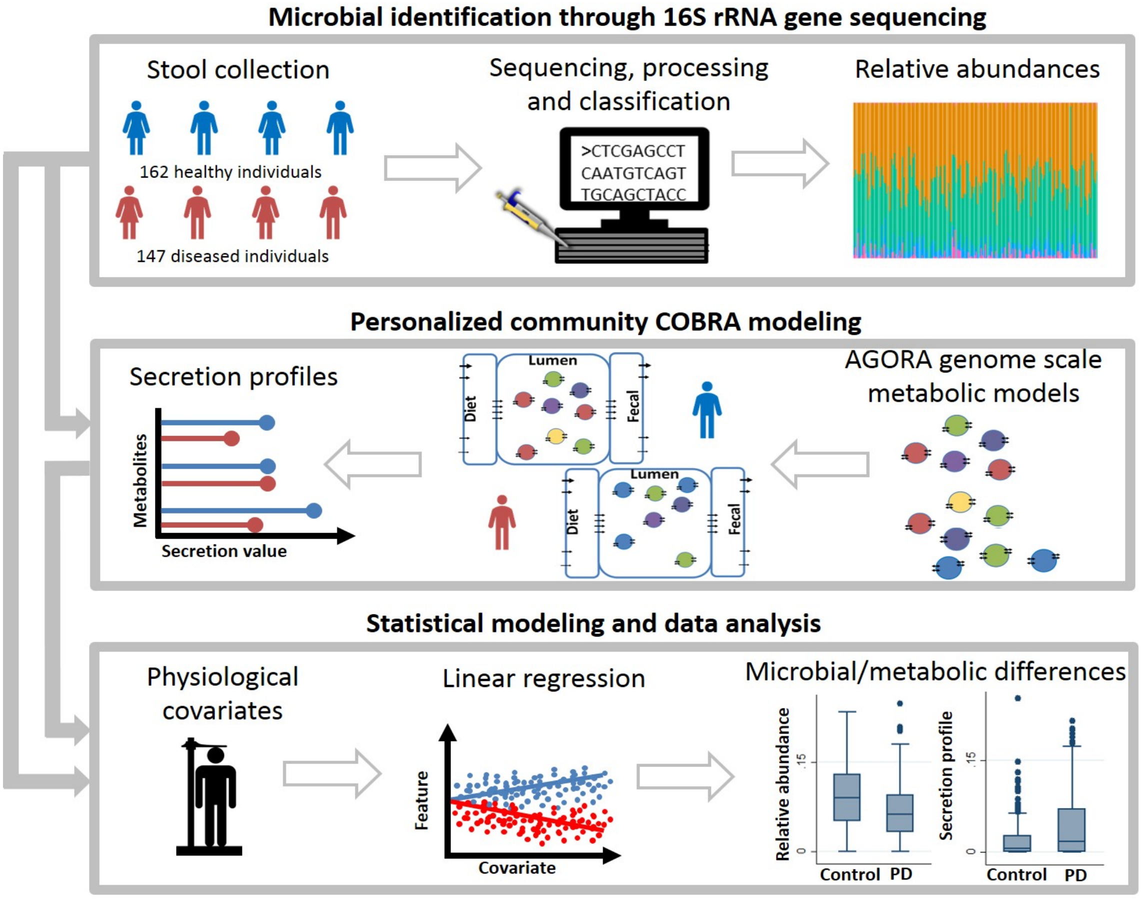
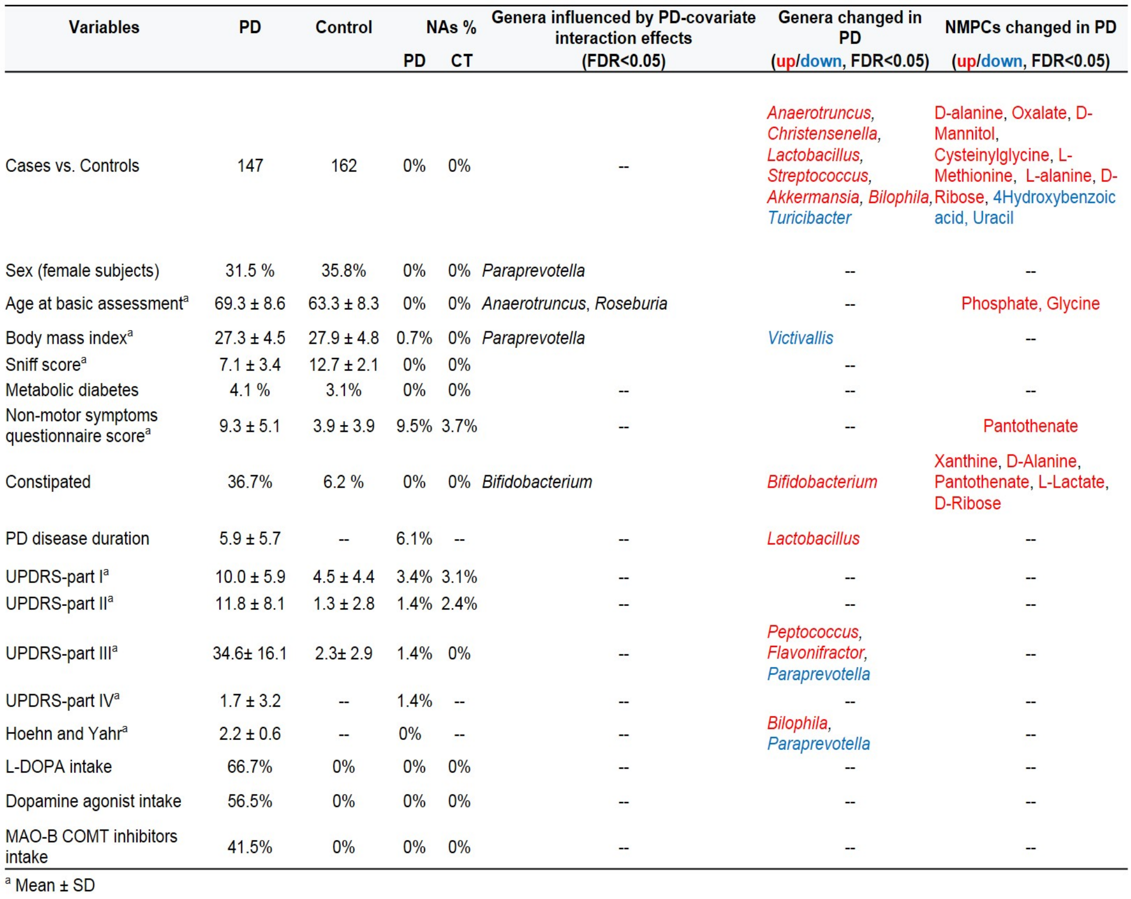


Figure 4.1: **Overview of the study approach and the key methods used.** Relative abundances were derived from 16S rRNA gene sequences (Methods: Analysis of the microbial composition with 16S rRNA gene sequencing) and used as input for the personalised community modelling to simulate metabolites secretion profiles. Relative abundances and secretion profiles were statistically analysed to identify microbial or metabolic differences between PD patients and controls.

studies [162, 21, 91], but in disagreement with two other studies [102, 82]. However, seven species were significantly altered in PD (FDR<0.05, Figure 4.2). Note that when comparing results between different taxonomic levels, changes observed for *Ruminococcus* and *Roseburia* species were not significant on the genus level but only on the species level, highlighting the importance of species-level resolution. The highest effect size was associated with *Akkermansia muciniphila* (Odds ratio (OR)=1.80, 95%-CI=(1.29, 2.51), p=6.02e-04, FDR<0.05; subsection C.2.1) in agreement with the previously reported higher abundance of *A. muciniphila* in PD patients [21, 82]. Subsequently, we examined possible differences at the genus level by performing semiparametric fractional regressions while adjusting for age, sex, the body mass index (BMI), batch, and total read counts. We identified eight genera to be significantly increased in PD (FDR<0.05; Figure 4.3A, Table 4.1), with *Lactobacillus*



Variables	PD	Control	NAs %		Genera influenced by PD-covariate interaction effects	Genera changed in PD	NMPCs changed in PD
			PD	CT	(FDR<0.05)	(up/down, FDR<0.05)	(up/down, FDR<0.05)
Cases vs. Controls	147	162	0%	0%	--	*Anaerotruncus, Christensenella, Lactobacillus, Streptococcus, Akkermansia, Bilophila, Turicibacter*	D-alanine, Oxalate, D-Mannitol, Cysteinylglycine, L-Methionine, L-alanine, D-Ribose, 4Hydroxybenzoic acid, Uracil
Sex (female subjects)	31.5 %	35.8%	0%	0%	*Paraprevotella*	--	--
Age at basic assessment[a]	69.3 ± 8.6	63.3 ± 8.3	0%	0%	*Anaerotruncus, Roseburia*	--	Phosphate, Glycine
Body mass index[a]	27.3 ± 4.5	27.9 ± 4.8	0.7%	0%	*Paraprevotella*	*Victivallis*	--
Sniff score[a]	7.1 ± 3.4	12.7 ± 2.1	0%	0%		--	
Metabolic diabetes	4.1 %	3.1%	0%	0%	--	--	--
Non-motor symptoms questionnaire score[a]	9.3 ± 5.1	3.9 ± 3.9	9.5%	3.7%	--	--	Pantothenate
Constipated	36.7%	6.2 %	0%	0%	*Bifidobacterium*	*Bifidobacterium*	Xanthine, D-Alanine, Pantothenate, L-Lactate, D-Ribose
PD disease duration	5.9 ± 5.7	--	6.1%	--	--	*Lactobacillus*	--
UPDRS-part I[a]	10.0 ± 5.9	4.5 ± 4.4	3.4%	3.1%	--	--	--
UPDRS-part II[a]	11.8 ± 8.1	1.3 ± 2.8	1.4%	2.4%	--	--	--
UPDRS-part III[a]	34.6± 16.1	2.3± 2.9	1.4%	0%	--	*Peptococcus, Flavonifractor, Paraprevotella*	--
UPDRS-part IV[a]	1.7 ± 3.2	--	1.4%	--	--	--	--
Hoehn and Yahr[a]	2.2 ± 0.6	--	0%	--	--	*Bilophila, Paraprevotella*	--
L-DOPA intake	66.7%	0%	0%	0%	--	--	--
Dopamine agonist intake	56.5%	0%	0%	0%	--	--	--
MAO-B COMT inhibitors intake	41.5%	0%	0%	0%	--	--	--

[a] Mean ± SD

Table 4.1: **Descriptive statistics of the analyses sample from the Luxembourg Parkinson's Disease study and overview over associations.** A red label means increased in PD, blue decreased in PD, while – "nothing to report". PD disease duration refers to time since diagnosis at the date of stool sampling. UPDRS=Unified Parkinson Rating Scale, L-DOPA=levodopa, MAO-B=monoaminooxidase B, COMT=Catecholamine-Methyl-Transferase, NMPC=Net maximal production capability.

showing the highest effect size (Odds ratio (OR)=5.75, 95%-CI=(2.29, 14.45), p=1.96e-04, FDR<0.05; subsection C.2.2). In contrast, the genera *Turicibacter* decreased significantly in PD cases (FDR<0.05). To summarise, significant changes could be observed on the species and genus level.

4.2.2 PD modifies the effects of basic covariates on the microbiome

Furthermore, we investigated whether the genus level alterations in PD were affected by basic confounding factors. This interaction analyses uncovered rich effect modifications, revealing that microbiome changes in PD have to be considered in the context of age, BMI, and gender. Our analyses demonstrate that the effects of PD are not homogeneous among important sub-groups of patients. For example, *Paraprevotella* was exclusively reduced in female patients but not in female controls (Figure 4.3B), highlighting gender-dependent alterations of microbial communities in PD. In addition, the effects of BMI and age were modified in PD cases. The PD cases had increased *Anaerotruncus* abundance with age, while non-linear, overall decreasing abundances of *Roseburia* and *Paraprevotella* were observed with age and BMI, respectively (Figure 4.3C). Taken together, these analyses suggest that microbial abundances are shifted in PD cases and that also the effects of important covariates were altered in PD, reflecting the systemic and complex nature of PD.

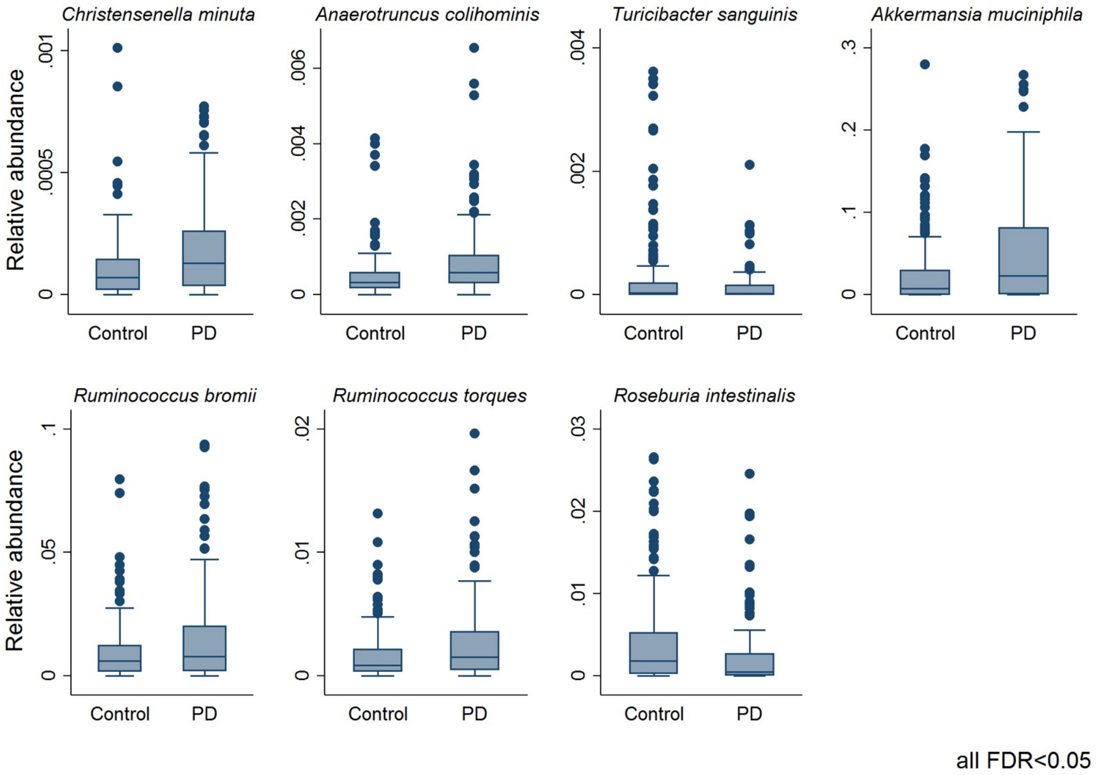


Figure 4.2: **Boxplots of seven significantly changed species in PD versus controls.** (FDR<0.05). Significance levels were determined using multivariable semi-parametrical fractional regressions with the group variable (PD vs. control) as predictor of interest, including age, gender, BMI, and technical variables (i.e., total read-counts and batch effect) as covariates. FDR=false discovery rate.

4.2.3 Microbial abundances, medication intake, and constipation in PD

The Luxembourg Parkinson's study enrols patients of all stages of PD. Therefore, the patients have considerable inter-individual variance in PD-related features, such as constipation and intake of medication (Table 4.1). We analysed whether these features had an impact the microbiome composition in PD. In our data, we could not find any evidence for an effect of the three medication types on the microbiome, i.e., levodopa, COMT inhibitors, or MAO-B inhibitors, when correcting for multiple testing (subsection C.2.2). In contrast, constipation, a prevalent non-motor symptom in PD patients (Lesser 2002), was associated with an increased abundance of *Bifidobacterium*, with a clear effect in constipated PD cases (Figure 4.3D). However, since there were only ten constipated controls (Table 4.1), these results must be confirmed in larger cohorts.

4.2.4 Genus association with the disease severity

We next investigated whether the stage of the disease, i.e., defined by Hoehn and Yahr staging, NMS, and UPDRS (Unified Parkinson Rating Scale) scores, and its subscales, was associated with altered genus abundance. For the Hoehn and Yahr staging, *Paraprevotella* showed a negative association and *Bilophila* showed a positive association, both of which were significant after multiple testing (Figure 4.3E). For the UPDRS III subscale score (i.e., motor symptoms, (Table 4.1), three genera, being *Peptococcus*, *Flavonifractor*, and *Paraprevotella*, survived correction for multiple testing (Figure 4.3F). In contrast, the other UPDRS subscales and the NMS were not significantly associated with microbial changes, after correction for multiple testing. Note that these analyses were performed while adjusting for disease duration. When analysing the association pattern of disease duration, we found *Lactobacillus* positively correlated with the disease duration (FDR<0.05, C.1). In conclusion, our data suggest that the microbial composition may be utilised as a correlate of disease severity.

4.2.5 Metabolic modelling reveals distinct metabolic secretion capabilities of PD microbiomes

To obtain insight into the possible functional consequence of observed microbiome changes in PD, we used metabolic modelling (cf. Methods). Briefly, we mapped each of the 309 microbiome samples on the generic microbial community model consisting of 819 gut microbial reconstructions [117, 78] (cf. Appendix C) to derived personalised microbiome models [12]. We then computed a net maximal production capability (NMPC) for 129 different metabolites that could be secreted by each microbial community model (cf. Methods), providing thereby a characterisation of the differential microbial metabolic capabilities in PDs and controls. The secretion of nine metabolites had differential NMPCs in PD (Figure 4.4A, all FDR<0.05) as determined by multivariable regressions adjusting for age, sex, BMI, and technical covariates. Moreover, although less dominant in comparison to the abundance data, PD-covariate interactions were also prevalent, with the uracil secretion potential showing a sex-specific effect and cysteine-glycine showing a BMI-dependent PD-effect (Figure 4.4B, 4C). In subsequent analyses, we tested for associations of the NMPCs with constipation, medication, disease duration, Hoehn-Yahr staging, NMS, and UPDRS III scores, complementing thereby the analyses on the abundance level. Notably, we found xanthine, D-alanine, L-lactic acid, D221 ribose, and pantothenic acid positively associated with constipation (Figure 4.4B), while no NMPC was associated with medication or with disease duration. However, the pantothenic acid secretion potential was positively associated with higher NMS scores, interestingly both in PD and in controls (Figure 4.4D), while no NMPC survived correction for multiple testing regarding associations with the UPDRS III score and Hoehn-Yahr staging. To conclude, these results suggest that the altered microbial composition in PD could result in broad changes in metabolic capabilities, which manifested themselves additionally in non-motor symptoms and constipation.

4.2.6 PD specific secretion profiles were altered due to changed community structure and species abundances

Next, we analysed which microbes contributed to the differential secretion profiles by correlating the NMPCs to the abundance data (Figure 4.4E/F, subsection C.2.3). Six metabolite NMPCs had strong contribution or where even dominated by single genera (Figure 4D), while for the other four NMPCs no single dominant genus could be identified. We then computed the contribution value of each genus to the production of each secreted metabolite (NMPC). From the aforementioned genera, which were associated on genus or species level with PD, only *Akkermansia*, *Acidaminococcus*, and *Roseburia* had substantial metabolic contributions (over 25%). *Acidaminococcus* was responsible for 64% of the variance in cysteine-glycine production and *Roseburia* for 30% of the variance in uracil production potential. *Akkerman-*

sia impacted the secretion profiles the most and substantially contributed to the metabolism of nine metabolites (Figure 4.4F), including the neurotransmitter gamma243 aminobutyric acid (GABA) and two sulfur species, being hydrogen sulfide and methionine. GABA was also significantly altered between PD and controls on a nominal level missing FDR corrected significance narrowly (b=0.18, 95%-CI:(0.06;0.30), p=0.003, FDR=0.0501). These analyses demonstrate the added value of metabolic modeling to investigate altered metabolic functions from the whole microbial composition.

4.3 Discussion

In this study, we aimed to elucidate compositional and functional changes in the fecal microbiome of PD patients. Therefore, we analysed 16S rRNA data from a cohort of typical PD patients (n=147) and controls (n=162), and performed personalized microbial computational modeling. We identified i) eight genera and seven species that changed significantly in their relative abundances between PD patients and healthy controls. ii) PD associated microbial patterns that were dependent on sex, age, BMI, constipation, and iii) in PD patients altered secretion potentials, particularly in sulfur metabolism, using metabolic modeling of microbial communities. Overall, our work demonstrated compositional and functional differences in the gut microbial communities of Parkinson's disease patients providing novel experimentally testable hypothesis related to PD pathogenesis.

The microbial compositional analyses of our cohort identified significantly different microbial abundance distributions between PD patients and healthy controls (Table 4.1). Up to date, 13 studies have described altered colonic microbial compositions associated with PD and an overall picture starts to arise (Figure 4.5). For instance, the microbial the families of *Verrucomicrobiaceae* and *Lactobacillaceae* have been consistently found to have an increased abundance in PD (Figure 4.5). In accordance, our study also reports increased abundance in PD of *Akkermansia*, *Christensenella*, and *Lactobacillus*. Similarly, *Bifidobacteria* has also been repeatedly associated with PD (Figure 4.5) but in our study, we could show that the *Bifidobacteria* association dependent on constipation (Figure 4.3) highlighting the need for incorporating disease-specific phenotypes as covariates into the statistical design.

At the same time, inconsistencies between the studies remain and they may be due to differences in study design, inclusion criteria, faecal sampling, RNA extraction protocols, and metagenomic and statistical methods. For instance, we used a relatively large, PD cohort while Bedarf and colleagues [21] studied a small cohort of drug-naïve, male PD patients and male controls (Figure 4.5). Three studies included individuals of Chinese descent [114, 115, 150] while the other studies focused on Caucasian individuals. It has been shown that microbial composition is associated with ethnic background, geography, and dietary habits [186, 46, 203], which may explain some of the discrepancies. The differences between the studies hence highlight the importance of performing meta-analysis to identify global microbial signatures, as it has been done for, e.g., colorectal cancer [198]. Such meta-analysis may also permit to investigate subgroups of PD, as the number of cases and controls would be substantially increased and thus provide higher statistical power. For instance, we observed various effect modulators that were not reported before in humans (Table 4.1), such as *Paraprevotella* abundance reduction being specific to women. This result is apparently in contradiction with findings from Bedarf and colleagues [21] who reported decreased levels of *Prevotellaceae* in a cohort of only male PD patients. However, once again, differences might be explained by different inclusion criteria, methodologies, and related possible sex-specific effects. Interestingly, a recent study reported a higher abundance of *Paraprevotella* in male mice compared to female mice [92]. Despite the lack of extensive studies on gender-specific differences in microbiome composition, we suggest that machine learning procedures on microbiome data should be performed in a sex-stratified manner. Larger cohorts, e.g., through meta-analysis of published cohorts would allow the identification of generalizable microbial differences in PD patients and also, specific microbial changes associated with certain traits and physiological characteristics, as suggested by our data.

We could not detect an effect of the dopaminergic, PD specific medication on the microbiome composition, after correction for multiple testing. Also the fact that key findings from the study of Bedarf and colleagues [21] were reproduced in other cohorts of PD patients under medication, including ours, support that notion. Nonetheless, in previous studies, *Dorea* and *Phascolarctobacterium* genera have been negatively associated with levodopa equivalent doses [150] and members of the family of *Bacillaceae* have been correlated with levodopa treatment [82]. Consequently, it cannot be excluded that medication is associated with microbial changes, albeit the association may be weaker than the effects of other covariates. As PD drugs are often taken in combinations, it would require a larger sample size than used in our study to permit the investigation of all possible drug combinations. The lack of clear association is somewhat expected as levodopa is absorbed in the upper part of the small intestine [174] and thus small intestinal rather than large intestinal microbes may play a more prominent role in levodopa bioavailability. Consistently, a recent study showed that bacterial tyrosine decarboxylases restrict the bioavailability of levodopa [192]. Interestingly, 193/818 reconstructed microbes [117], commonly found in the human gut, carry genes encoding for proteins that convert levodopa into dopamine [132]. Levodopa is always given with decarboxylase inhibitors, such as carbidopa or benserazide, targeting the human decarboxylases, but it cannot be excluded that they also act on the microbial counterpart. However, Van Kessel et al. have shown that carbidopa as well as benserazide is only a weak inhibitor of the microbial tyrosine decarboxylase [192].

We identified a positive association of *Bilophila* abundance with the Hoehn and Yahr staging, which captures motor impairment and disability independent of disease duration. Indeed, the abundance of *Bilophila* was not associate with disease duration, indicating mainly dependency on the progression of symptoms. This finding is consistent with experimental mice studies demonstrating the pro-inflammatory effect of *Bilophila* overgrowth [49, 128]. Notably, the Hoehn and Yahr staging was also positively associated on a nominal level with the predicted pyruvate secretion profile (subsection C.2.4), which was accordingly significantly increased in PD patients on a nominal level alongside with L- and D-alanine. *Bilophila* has the rare capability to use taurine, an inhibitory neurotransmitter with neuroprotective effects [160, 201], as an energy source [110]. This pathway is initiated by the taurine: pyruvate aminotransferase [110], converting pyruvate and taurine into L-alanine and sulfoacetaldehyde. The only microbe of the 818 species in our AGORA collection encoding the corresponding gene was *Bilophila*, which was significantly increased (FDR<0.05) and hence, the corresponding reaction (VMH ID: TAURPYRAT) was increased in abundance in PD microbiomes as well. In a previous study [86], we have shown that blood taurine conjugated bile acids were positively associated with motor symptoms. *Bilophila* may be a marker of disease progression in PD, and it could modulate human sulphur metabolism through its taurine degradation capabilities. Alterations in sulphur metabolism have been already described when using computational modelling of microbiomes from a cohort of early diagnosed and levodopa naive PD patients [21, 86] as well as an increased concentration of methionine and derived metabolites in blood samples [86]. Furthermore, we and others have reported alterations in bile acids and taurine-conjugated bile acids in PD patients [69, 86]. Our present study suggests again a key role of *Bilophila* in host-microbiome sulphur co-metabolism, which may link with bile acid metabolism.

Interestingly, an increased abundance of *B. wadsworthia* has been linked to constipation [194]. *B. wadsworthia* is the only microbe in the AGORA collection capable of the metabolic reaction converting pyruvate and taurine to L-alanine and sulphoacetaldehyde (VMH ID: TAURPYRAT). Therefore, an increased production of L-alanine might be due to the increased *B. wadsworthia* abundance. This resulting higher production rate of L-alanine could then lead to an increased conversion into D-alanine via the alanine racemase (VMH ID: ALAR), which was present in 808/818 gut microbes in the AGORA collection. Accordingly, D-alanine was one of the three metabolite secretion profiles increased in constipated PD patients (Figure 4.4E). This hypothesis of *B. wadsworthia* playing a role in constipation of PD patients would need to be experimentally validated, especially since we could not find statistically significant changes in the association between the abundance of *B. wadsworthia* and constipated individuals. In contrast, we found an increase in *Bifidobacteria* abundance in constipated individuals and particularly in constipated PD patients. This result disagreed with an earlier study on individuals with chronic constipation, which reported a decrease in *Bifidobacteria* abundance [103]. Overall, the available data suggest that complex alterations in microbial composition are associated with constipation but may differ between diseases.

The mucin degrading microbe, *A. muciniphila*, represents about 1-4% of the faecal microbiome in humans [126]. Numerous diseases have been associated with a decrease in A. muciniphila abundance [164, 70], while an increase has been consistently reported in PD patients (Figure 4.5). The *A. muciniphila* abundance had the largest contribution to the significantly altered metabolite secretion profiles (Figure 4.4E), including the neurotransmitter gamma-aminobutyric acid (GABA). While its predicted secretion potential was only nominally increased in PD patients the present study, higher GABA secretions rates have also been predicted based on microbiome data from early stage levodopa naive PD patients [86]. Importantly, GABA receptors have been found in the enteric nervous system, gut muscle, gut epithelial layers, and endocrine-like cells [93] and its gut receptors are thought to be related to gastric motility (peristalsis), gastric emptying, and acid secretion [93]. Experiments with the GABAb agonist baclofen have shown that GABAb receptors can reduce gastric mobility in the colon of rabbits (via cholinergic modulation) [184]. Interestingly, *A. muciniphila* has been shown to be positively associated with gastrointestinal transit time [67, 193]. GABA could reach the CNS via blood stream as a lipophilic compound, being able to pass the blood brain barrier. Additionally, microbial GABA could affect the brain-gut axis by contributing the human GABA pools, especially as it has been shown that the microbiome can affect GABA receptor density in the CNS via the vagus nerve [30]. To establish whether and which role *A. muciniphila* and GABA may play a role in prodomal PD, further experimental studies will be required.

In order to move beyond mere cataloguing of microbial changes associated with diseases, pathway-based tools [3] have been developed, in which microbial sequences (or reads) are mapped, e.g., onto KEGG ontologies present in the KEGG database [100]. Using such tools, Bedarf et al reported decreased glucuronate degradation and an increase in tryptophan degradation and formate conversion [21]. Similarly, Heinz-Buschart et al. reported 26 KEGG pathways to be altered in PD microbiomes [82]. In our study, we complemented the compositional analysis with computational modelling to gain insight into potential functional, i.e., metabolic, consequences of changed microbe abundances in PD. The advantage of our approach is that the functional assignments may be more comprehensive than more canonical methods, such as KEGG ontologies because (1) the underlying genome-scale metabolic reconstructions have been assembled based on refined genome annotations and have been manually curated to ensure that the reaction and gene content is consistent with current knowledge about the microbe's physiology, and (2) each of these reconstructions, alone or in combinations, are amenable to metabolic modelling and thus functional and metabolic consequences of a changed environment (e.g., nutrients or other microbes in the models) can be computed. These simulations are thus allowing to predict functional consequences and not only pathway or reaction enrichment, as typically done.

4.3.1 Strength and limitation

Here, we present microbiome analyses in a large population-based, monocentric case control study on PD from a defined area (Figure 4.5). Capitalising on the overall clinical spectrum of PD of the LuxPark cohort, which reflects a representative sample of PD patients of different disease stages from a defined geographical area, we demonstrated that microbial composition is not only altered in PD but also that the observed associations of PD with changes in the composition of the microbiome should be interpreted in the context of age, sex, BMI, and constipation. This information is of importance for clinical translation, highlighting the need for both, (i) a personalised and (ii) a holistic approach, to understand the role of microbial communities in PD pathogenesis. In a second step targeting the potential functional changes related to PD-associated microbiomes, we performed metabolic modelling based on the AGORA collection [117] of genome-scale metabolic reconstructions, allowing for the predictions of metabolite secretion profiles. Thus, our analyses facilitated a detailed investigation of the altered metabolism of PD-related microbial communities in the gut, pointing towards a role of the known pro-inflammatory species *B. wadsworthia* interacting with the host on sulphur metabolism. Hence, metabolic modelling provides a valuable tool for deciphering the metabolic activity of microbial communities in PD.

However, despite the partial confirmation of previous results by our study (Table 5), several limitations should be kept in mind. First, certain covariates were not investigated, such as diet, exercise, and smoking. Whether these covariates alter the PD-specific signature is yet to be analysed. Although our study belongs to the three largest studies performed yet on PD, our sample size was still too small to deliver insights on combinations of drugs. Furthermore, 16S RNA sequencing, as applied in our study, is not allowing analyses on the strain level and may lead to misclassifications [96], and follow-up studies based on shotgun sequencing are needed to further corroborate our results. However, our results are notably well aligned with a previous shotgun sequencing study [21], which would further support a role of 16S RNA sequencing as a cost-efficient screening method. Being cross-sectional in nature, causal inference is not possible. Consequently, although metabolic modelling has been numerous times been shown to correctly predict attributes of living systems [133, 11, 131], our hypothesis on the role of *B. wadsworthia* in PD interlinking sulphur metabolism with disease severity requires experimental validation. To conclude, by combining metabolic modelling with comprehensive statistical analyses, we identified a promising research target in PD and refined the understanding of PD-related microbial changes.

4.4 Methods

4.4.1 Description of the Luxembourg Parkinson's study

For this study, data and biospecimen of the LuxPark cohort were utilised [88]. The Luxembourg Parkinson's study includes a variegated group of patients with typical PD and atypical parkinsonism, and controls from Luxembourg and its neighbouring regions [88]. Controls were partly sampled among relatives of patients. The corresponding information on the family relation between controls and cases was not available. Cancer diagnosis with ongoing treatment, pregnancy, and secondary parkinsonism (drug induced parkinsonism and parkinsonism in the frame of normotensive hydrocephalus) were exclusion criteria for enrolling in the patient or healthy control group. For 454 individuals (controls: n=248, PD: n=206) from the LuxPark cohort, stool samples were available and used for 16S RNA gene sequencing data (see below). Within LuxPark, controls were selected among spouses of chosen patients and volunteers and individuals from other independent Luxembourgish studies [45, 157]. As we aimed to target specifically typical PD (IPD), we excluded all individuals with age below 50 (controls: n=47, PD: n=9) and all individuals with an unclear status of PD diagnosis or an atypical PD diagnosis (PD: n=47). PD patients were defined as typical PD, according to the inclusion criteria by the United Kingdom Parkinson's Disease Society Brain Bank Clinical Diagnostic Criteria (Hughes et al. 1992). Furthermore, we excluded control patients with a United Parkinson's Disease Rating Scale (UPDRS) III score above ten, except for one control where the high UPDRS III score was caused by an arm injury. Furthermore, we excluded control persons who took dopaminergic medications (n=5), and individuals who reported to have taken antibiotics in the last six months (controls: n=20, PD: n=13). Note that excluded observations behave sub-additive, because of overlap between the exclusion criteria (i.e. individuals below age 50 and taking antibiotics). Finally, 309 individuals (controls: n=162, cases: n=147) were included in the statistical analyses.

All study participants gave written informed consents, and the study was performed in accordance with the Declaration of Helsinki. The LuxPark study [88] was approved by the National Ethics Board (CNER Ref: 201407/13) and Data Protection Committee (CNPD Ref: 446/2017).

4.4.2 Measurements and neuropsychiatric testing

All patients and healthy controls were assessed by a neurologist, neuropsychologist or trained study nurse during the comprehensive battery of clinical assessment. Olfaction testing was conducted using the Sniffin' Sticks 16-item version (SS) within the LuxPark cohort [88]. Antibiotics usage was defined as intake of antibiotic within the previous six months to stool collection. For assessing PD-related motor and non-motor symptoms, the UPDRS rating scales I-IV were used [68]. The severity of the disease was reflected by the Hoehn and Yahr staging [89]. Non-motor symptoms were measured via the NMS questionnaire [155]. The use of medication was recorded, and PD498 specific medication was classified into three classes, 1) levodopa, 2) dopamine receptor agonist, and 3) MAO-B/COMT inhibitors.

4.4.3 Collection and processing of stool samples

All samples were processed following standard operating procedures [112, 122]: stool samples were collected at home by patients using the OMNIgene.GUT stool tubes (DNA Genotek) and sent to the Integrated Biobank Luxembourg (IBBL) where one aliquot of 1 ml was used for DNA extraction. For the DNA extraction, a modified Chemagic DNA blood protocol was used with the MSM I instrument (PerkinElmer), the Chemagic Blood kit special 4 ml (Ref. CMG-1074) with a lysis buffer for faecal samples, and MSM I software. Samples were lysed using the SEB lysis buffer (included in the kit) and vortexed to obtain a homogenous suspension that was incubated for 10min at 70°C, then 5min at 95°C. Lysates (1.5mL) were centrifuged for five minutes at 10,000 g at RT. Supernatants were transferred to a 24XL deep-well plate. Plates were processed using the MSM I automated protocol.

4.4.4 Analysis of the microbial composition with 16S rRNA gene sequencing

The V3-V4 regions of the 16S rRNA were sequenced at IBBL using an Illumina 516 Platform (Illumina MiSeq) using 2x300bp paired-end reads [88]. The gene517 specific primers targeted the V3 - V4 regions of the 16S rRNA gene. These primers were designed with Illumina overhang adapters and used to amplify templates from genomic DNA. Amplicons were generated, cleaned, indexed, and sequenced according to the Illumina- demonstrated 16S Metagenomic Sequencing Library Preparation Protocol with certain modifications. In brief, an initial PCR reaction contained at least 12.5 ng of DNA. A subsequent limited-cycle amplification step was performed to add multiplexing indices and Illumina sequencing adapters. Libraries were normalised, pooled, and sequenced on the Illumina MiSeq system using 2x300 bp paired-end reads.

The demultiplexed samples were processed merging forward and reverse reads and quality filtered using the dedicated pipeline "Merging and Filtering tool (MeFit)" [145] with default parameters. To obtain a reliable microbial identification, identification to both genus and species taxonomic level was obtained using the SPINGO (SPecies level IdentificatioN of metaGenOmic amplicons) classifier [8] with default parameters. Relative abundances were computed, for each sample, using an R (R Foundation for Statistical Computing, Vienna, Austria) [94] custom script. Briefly, for each sample, the counts of each genera/species were retrieved, and then the sum of the counts of all the genera/species was used to normalise to a total value of 1 each genera/species count.

4.4.5 Personalised constraint-based modelling of microbial communities

AGORA consists of a set of 819 strains of microbes commonly found in the human gut [117, 132]. To match species taxonomic resolution, we combined strain models of the same species in one species model (‘panSpeciesModel.m’) using the function ‘createPanModels.m’ of the microbiome modelling toolbox [12]. Briefly, reactions of multiple strains are combined into one pan-reconstruction. The pan542 biomass reaction is built from the average of all strain-specific biomass reactions. Microbial abundances were mapped onto a set of 646 species performing an automatic name matching between SPINGO species taxonomic assignment and panSpecies names. A threshold for assessing the bacterial presence of a relative abundance value of 0.0001 was used to reduce the time of computations while limiting the order of magnitude simulations results of stoichiometric coefficients to ten. A total of 259 species overlapped between our set of species models and SPINGO species assignment when considering species identified at least in 10 % of samples (cf. Appendix C). The retrieved microbial abundance information for each sample was integrated into a community modelling setup obtaining personalised microbiome models using the automated module of the microbiome modelling toolbox [12] called mgPipe within the COBRA toolbox [84] (commit: b097185b641fc783fa6fea4900bdd303643a6a7e). Briefly, the metabolic models of the community members are connected by a common compartment, where each model can secrete/uptake metabolites. An average European diet was set as input for each microbiome model [132]. A community objective function was formulated based on the sum of each microbial model objective function and constrained to a lower bound of 0.4 per day and upper bound of one per day. A set of exchange reactions connects the shared compartment to the environment enabling to predict metabolite uptake and secretion flux rates (metabolic profiles/NMPCs) consistent with the applied constraints. The personalisation of each microbiome model was achieved by adjusting stoichiometric coefficients in the community biomass reactions to each sample’s relative microbial abundance and removing species undetected from the community models.

Relative reactions abundances were calculated by summing the number of species having the reaction in a microbiome model and scaling the sum by the respective species relative abundance. Community metabolic profiles of these microbial communities were assessed using flux variability analysis on the exchange reactions [72]. AGORA microbial metabolic reconstructions used for the construction of the community models were downloaded from the VMH (www.vmh.life, [132]). All computations were performed in MATLAB version

2018a (Mathworks, Inc.), using the IBM CPLEX (IBM, Inc.) solver through the Tomlab (Tomlab, Inc.) interface.

4.4.6 Analyses of relative abundances

For descriptive statistics, metric variables were described by means and standard deviations, while nominal variables were described by proportions. Missing values were not imputed, and the pattern of missing values was not assessable via the ADA platform [88]. The read counts for each metagenomic feature (e.g., genera and species) were divided by total read counts such that relative abundances were retrieved. Relative abundances were checked for outliers. Observations with more than four standard deviations from the mean were excluded from analyses. Only genera and species detected in more than 50% of all samples were included in the analyses, resulting in 62 genera and 127 species.

The metagenomic data was analysed using fractional regressions as developed by [144]. Fractional regressions, first applied to econometric problems, are semiparametric methods designed to model fractional data without the need of specifying the distribution of the response variable. Fractional regressions are further inherently robust against heteroscedasticity and can be parametrised in odds ratios, delivering convenient interpretations of the regression coefficients. All statistical models included technical covariates, batch, total read counts, and unclassified read counts (reads for which a taxonomic assignment was not possible independently from any threshold of confidence estimate value used). The read count variables were included into the statistical model, as it has been shown that normalisation by division can introduce bias if certain statistical assumptions implied by the application of division are not fulfilled [86]. In the case of metagenomic data, the effect of read counts would be removed by division if the observations would be sampled from a multinomial distribution. However, this is not a given as species and genera correlate amongst each other, violating the assumptions needed to construct multinomial distributions. In consequence, read count normalisation by division is prone to introduce a bias into metagenomic data; a potential bias, we corrected for by including the read counts as covariates into the model.

Before fitting the final statistical models, we explored the associations of basic covariates (age, sex, and BMI) with metagenomic features using fractional regressions as described above to avoid misspecifications of the statistical models. Since the data showed a high range in age and BMI, we checked for potential non-linear associations by including these variables into the models as restricted cubic splines [76] using three knots defined by the 5%-percentile, the median, and the 95%-percentile. As in the case for age, we found species with indications of non-linear age-associations with p<0.01, age was modelled in all analyses via restricted cubic splines.

All p-values are reported two-tailed. Statistical analyses were performed in STATA 14/MP (College Station, Texas, USA). Summary statistics of the performed analyses are given in the Supplementary files (section C.2).

4.4.7 Differences between PD and controls in microbial composition and the influence of covariates

To analyse difference between genus abundances between PD and controls, fractional regressions were carried out with the relative abundance of the genus as the response variable, while including technical covariates, age (restricted cubic splines), sex, and BMI into the statistical modelling. The predictor of interest was the study group indicator variable. We corrected for multiple testing using the Benjamini-Hochberg procedure [22] by setting the false discovery rate (FDR) to 0.05. Consequently, we corrected for 62 tests when reporting genera results. These analyses were repeated analogously for the taxonomic level of species, while correcting for multiple testing via the FDR.

Next, we explored the possibility of statistical interactions between basic covariates (age, sex, and BMI) and the group indicator. For these analyses, we once again modelled age and BMI via restricted cubic splines allowing for non-linear interaction terms. We only tested two-way interaction terms. All interaction terms were introduced simultaneously into the statistical model and tested on significance via a Wald test [76], correcting for multiple testing via the FDR. For the globally significant test, the single interaction terms were investigated to explore which covariate-group interaction contributed to the overall significance. For interpretation, the interaction terms were visually inspected by plotting the predictions conditional on technical covariates. These analyses were then rerun with species abundances as response variable instead of genus abundances.

We assessed the influence of constipation on the microbial composition. We introduced the binary predictor constipation (yes/no) as additional predictor into the model and the corresponding group-constipation interaction term. Both terms were tested simultaneously on zero with a Wald test. The analyses were once again adjusted for technical covariates, age (restricted cubic splines), sex, and BMI, and we corrected for multiple testing via the FDR.

4.4.8 Analyses of within PD phenotypes in relation to microbial composition

We investigated the association pattern of medication and clinical features regarding the microbial composition. These analyses were only performed on the IPD cases, while controls were excluded from the analyses. First, we analysed the disease duration as measured in years between the date of the stool sampling and the year of the diagnosis. The analyses were conducted as before via fractional regressions with the genus abundances as the response variable, while adjusting for technical covariates, age (restricted cubic splines), sex, and BMI. Then, we assessed in separate analyses the UPDRS III score as an indicator for motor symptoms, the non-motor symptoms as measured by the NMS, the Hoehn-Yahr staging of the disease as a global measure of disease progression, and the sniff-score. All these analyses were performed adjusted for technical covariates, age (restricted cubic splines), sex, BMI, and disease duration. Each of these series of regression represents 62 test, which was accounted for using the FDR. The impact of medication was analysed by examining three classes of medication, a) levodopa, b) mono-amino oxidase/catechol-O-methyltransferase inhibitors, and c) dopamine receptor agonists. We generated three corresponding binary phenotypes (intake/no intake) and added these three variables simultaneously to the statistical model determining the significance of this add-on via a Wald test. We then tested each medication-class in separate analyses, strictly correcting for multiple testing via the FDR (186 tests in total). The analyses were performed adjusted for technical covariates, age (restricted cubic splines), sex, BMI, and disease duration.

4.4.9 Statistical analyses of fluxes

The NMPCs were log transformed such that the skewness of the distribution was minimised [27]). This type of transformation was applied because of the very differently skewed distributions of the single NMPCs. Then, outliers were excluded using the 4-SD outlier rule as before. Only fluxes with more than 50% non-zero values were retained in analyses. Furthermore, NMPCs with distributions not suitable for statistical analyses (e.g., distributions with a high number of observations with exact the same numerical value) were excluded resulting in 129 NMPCs included into analyses.

The NMPCs were analysed with mixed linear regressions including the batch as random effects. Including the batch variable as a random effect has a higher statistical power in comparison to the fixed effect approach, but relies on more restrictive assumptions. We tested the corresponding random effect assumption by Hausman specification tests and found no indications of violations of the Hausman specification test. Note that this possibility to account for batch effects via random effects is not available with fractional regressions where batch effects were corrected via fixed effects.

We performed the same analyses as with the metagenomic data, with the sole exception of replacing the fractional regression model with the linear mixed model. In all other aspects, the analyses followed the same scheme.

4.4.10 Analyses of species contribution to fluxes

To investigate the contribution of species and genera, we calculated for all included genera and all analysed fluxes the pairwise correlation and the corresponding variance contribution (the squared correlation). We classified every correlation above 0.5 (equal to 25% of variance contribution) as a strong correlation in accordance with classical classifications of effect size [44].

4.4.11 Material availability

All 16S rRNA sequences can be requested from I.T. (ines.thiele@nuigalway.ie). The mgPipe pipeline is available within the COBRA toolbox (https://github.com/opencobra/cobratoolbox), and the custom scripts with related documentation are available at the GitHub repository: https://github.com/ThieleLab/CodeBase/tree/master/ND_collect.

4.4.12 Acknowledgment

This study was funded by grants from the Luxembourg National Research Fund (FNR) within the National Centre of Excellence in Research (NCER) on Parkinson's disease (FNR/ NCER13/BM/11264123) and the PEARL programme (FNR/P13/6682797 to RK), by European Union's Horizon 2020 research and innovation programme grant agreement No 692320 to RK, and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757922) to IT.

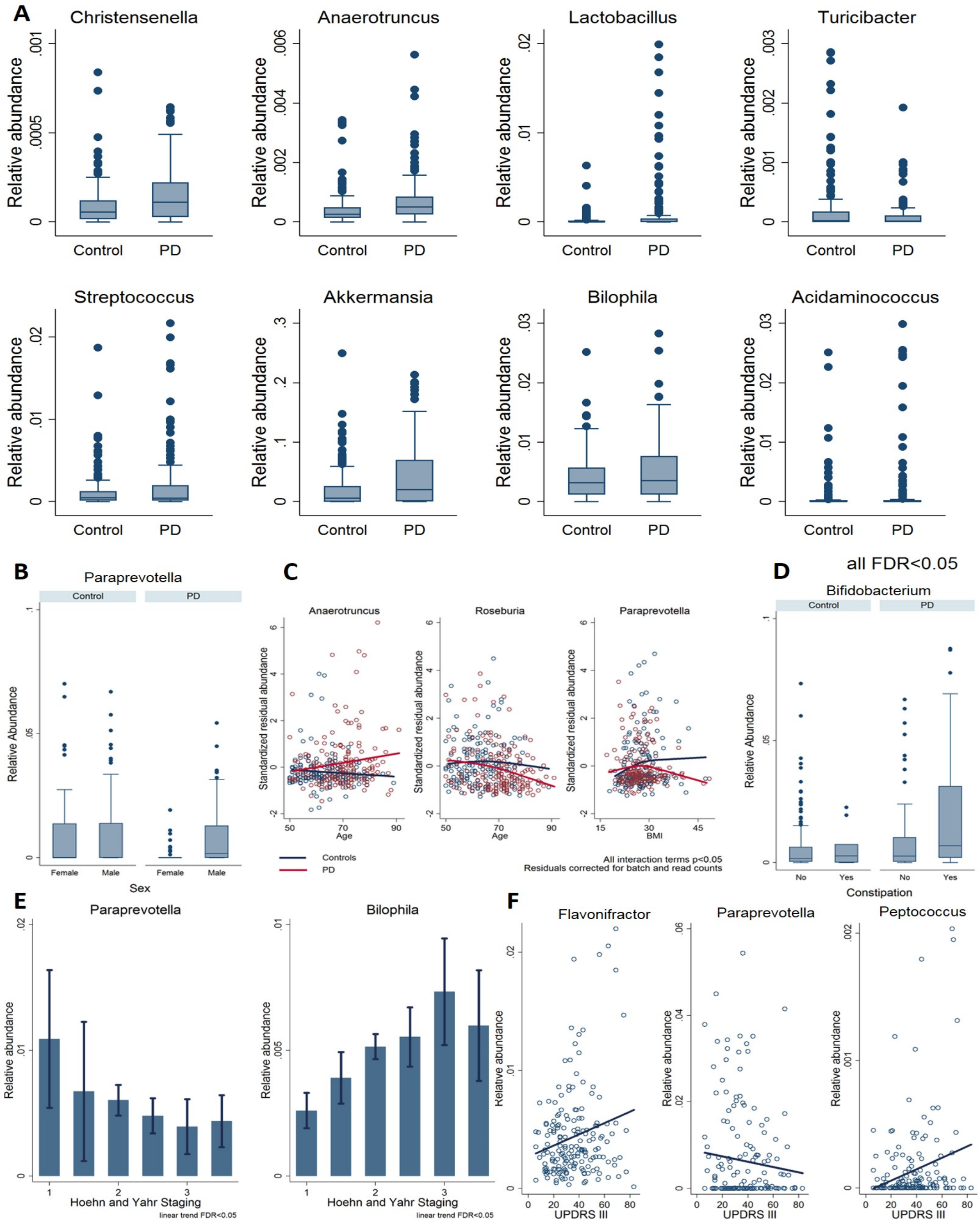
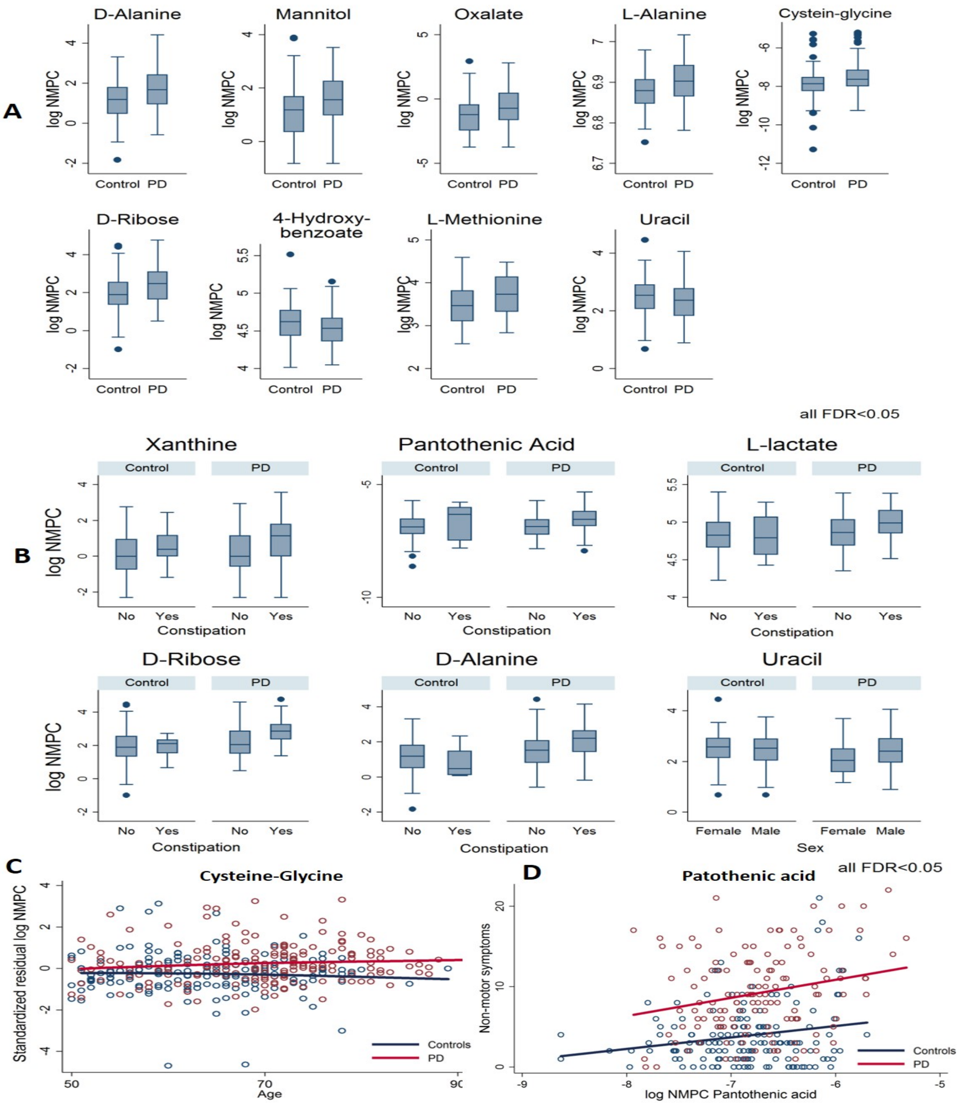


Figure 4.3: **Genus alterations in PDs due to interactions with basic covariates.** **A.** Boxplots of the seven significant species (FDR<0.05). **B.** Female PD patients have a reduced abundance of *Paraprevotella* (FDR<0.05). **C.** Genus abundance age and BMI dependencies of Anaerotruncus, Roseburia, and Paraprevotella (global test on all interaction terms, FDR<0.05). For graphical assessment of the interaction terms the z-transformed residual abundances are displayed after correction for technical covariates (batch and read counts). **D.** The genus relative abundance of Bifidobacterium was increased in patients reporting to be constipated (FDR<0.05). **E.** Genus association with disease staging showed a decrease of relative abundance of Paraprevotella and an increase of Bilophila genus over increasing Hoehn and Yahr scale values (FDR<0.05). **F.** An increased score in motor symptoms (UPDRS III) was associated with an increased trend in abundances of Flavonifractor and Peptococcus and a decreased trend in Paraprevotella abundance (FDR<0.05). UPDRS=Unified Parkinson Rating Scale, BMI=body mass index,FDR=false discovery rate.



E

Genera contributing above 25% Variance to NMPC s different between cases and controls

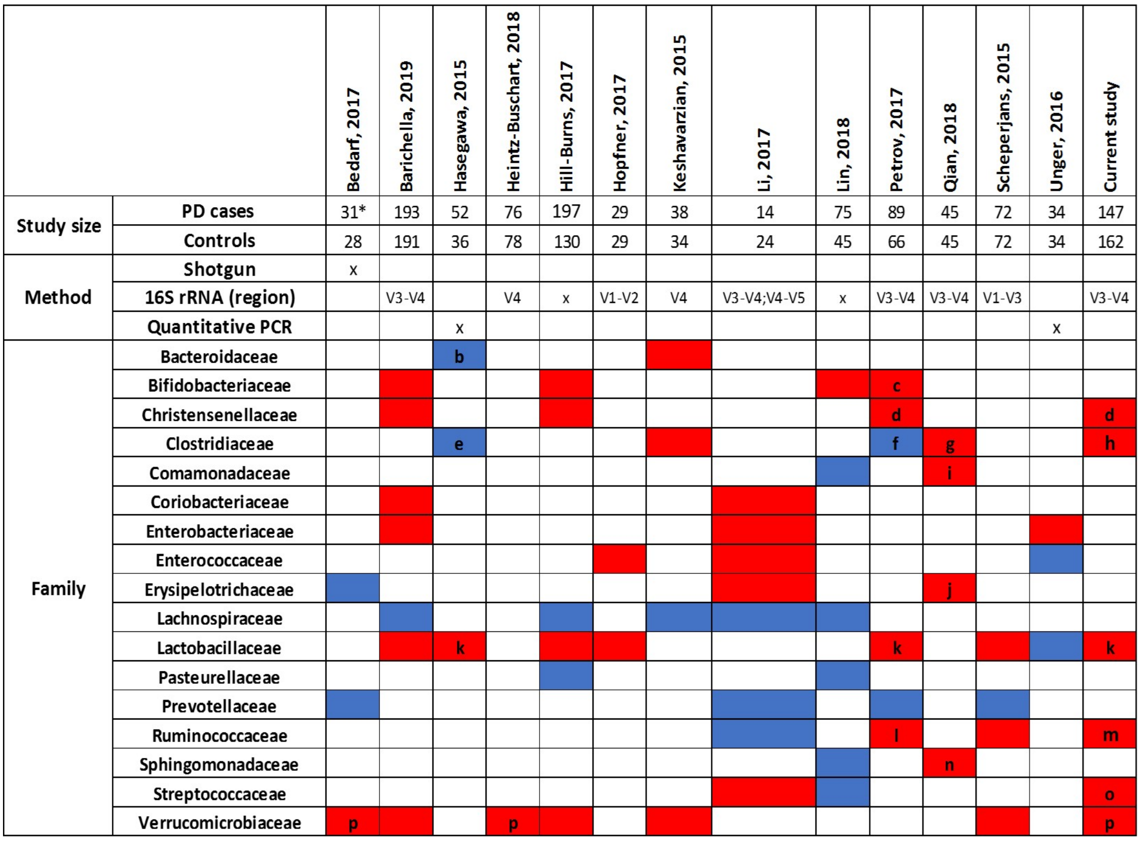
Metabolite	Genus	Explained Variance	Effect sign
D-Alanine	*Veillonella*	27.47%	+
L-Alanine	*Akkermansia, Prevotella*	27.97%,31.02%	+,-
Cysteine-Glycine	*Acidaminococcus*	66.18%	+
4-Hydroxybenzoate	*Faecalibacterium*	85.16%	+
L-Methionine	*Akkermansia*	65.89%	+
Uracil	*Roseburia*	29.76%	+

F

NMPCs for which *Akkermansia* abundance explained more than 25% of variance

Metabolite	Explained Variance	Effect sign
GABA	66.51%	+
Alanine	27.97%	+
L-Asparagine	26.29%	-
Ethanol	30.06%	+
Glycerol	32.71%	+
Hydrogen sulphide	53.16%	+
L-Methionine	65.89%	+
Phosphate	34.30%	+

Figure 4.4: **Result of analysing secretion profiles of microbial communities. A.** Box plots for NMPCs differential between cases and controls with FDR<0.05. **B.** NMPCs with sex-specific PD signature or constipation effects (all FDR<0.05). **C.** Differential age trajectory between cases and controls for cysteine-glycine (p<0.05). **D.** Association of pantothenic acid with non254 motor symptoms. **E.** Genera contributing more than 25% to NMPCs different between cases and controls. **F.** *Akkermansia* contribution to community production of 12 metabolites expressed as a percentage of total production for each compound. Metabolites highlighted in red were significantly increased in PD (FDR<0.05). NMPC=net maximal production capacity, GABA=gamma-aminobutyrate, H2S=hydrogen sulphide, FDR=false discovery rate. Effect sign "–": negative correlation. Effect sign "+": positive correlation.



		Bedarf, 2017	Barichella, 2019	Hasegawa, 2015	Heintz-Buschart, 2018	Hill-Burns, 2017	Hopfner, 2017	Keshavarzian, 2015	Li, 2017	Lin, 2018	Petrov, 2017	Qian, 2018	Scheperjans, 2015	Unger, 2016	Current study
Study size	PD cases	31*	193	52	76	197	29	38	14	75	89	45	72	34	147
	Controls	28	191	36	78	130	29	34	24	45	66	45	72	34	162
Method	Shotgun	x													
	16S rRNA (region)		V3-V4		V4	x	V1-V2	V4	V3-V4;V4-V5	x	V3-V4	V3-V4	V1-V3		V3-V4
	Quantitative PCR			x										x	
Family	Bacteroidaceae			🟦 b				🟥							
	Bifidobacteriaceae		🟥			🟥				🟥	🟥 c				
	Christensenellaceae		🟥			🟥					🟥 d				🟥 d
	Clostridiaceae			🟦 e				🟥			🟦 f	🟥 g			🟥 h
	Comamonadaceae									🟦		🟥 i			
	Coriobacteriaceae		🟥						🟥						
	Enterobacteriaceae		🟥						🟥					🟥	
	Enterococcaceae						🟥		🟥					🟦	
	Erysipelotrichaceae	🟦							🟥			🟥 j			
	Lachnospiraceae		🟦			🟦		🟦	🟦	🟦					
	Lactobacillaceae		🟥	🟥 k		🟥	🟥				🟥 k		🟥	🟦	🟥 k
	Pasteurellaceae					🟦				🟦					
	Prevotellaceae	🟦							🟦		🟦		🟦		
	Ruminococcaceae								🟦		🟥 l		🟥		🟥 m
	Sphingomonadaceae									🟦		🟥 n			
	Streptococcaceae								🟥	🟦					🟥 o
	Verrucomicrobiaceae	🟥 p	🟥		🟥 p	🟥		🟥					🟥		🟥 p

Figure 4.5: **Reported microbial changes at the family level associated with PD in different studies.** Only those bacterial families are shown, for which significant associations with species or genera have been reported in at least two studies comparing stool samples from patients and controls. Red - increased in PD, Blue - decreased in PD. a: *Actinomycetales*, b: *Bacteroides fragilis*, c: *Bifidobacterium*, d: *Christensenella*, e: *Clostridium coccoides/ leptum*, f: *Faecalibacterium*, *Dorea*, g: *Clostridium* IV/XVIII, *Butyricicoccus*, *Anaerotruncus*, h: *Anaerotruncus*, i: *Aquabacterium*, j: *Holdemania*, k: *Lactobacillus*, l: *Oscillospira*, m: *Ruminococcus romii*, *Ruminococcus torques*, n: *Sphingomonas*, o: *Streptococcus*, p: *Akkermansia*. * Drug-naive, de novo PD patients only. Based on (Barichella et al. 2019) [14].

Chapter 5

Concluding remarks

The increasing availability of microbial sequencing data advanced rapidly our knowledge of the human gut microbiota. Thanks to pioneer work on these data, we now understand more of the role of the human gut microbiota and we are able to correlate microbial composition with several diseases, different lifestyles and diets. The related studies allowed us to understand all we currently know about the human gut microbiota and provided answers to questions such as “How does the human gut microbiota composition change over the years?”, “Is microbial composition changing with different lifestyles?”, “Are there common trends between similar populations?” and “Can microbial composition be correlated to specific diseases?”. Anyhow, microbial sequencing data represent a gold mine to be dug to understand the role of the human gut microbiota, and the full potential of these data has not been yet exploited. As mentioned before, nowadays, most of the studies in the human gut microbiota field, focus on statistical correlation rather than trying to explain the related functional mechanisms. Thereby, great part of the investigators’ efforts was dedicated to performing fine statistics to find, for each dataset, correlations with a specific predictor and relative covariates, more than trying to understand the underlying mechanisms. The time to add another dimension to these microbiota studies, the functional one, has come. Besides few cases, where other type of omics data such as transcriptomic and proteomic where correlated to metagenomics in the attempt of disclosing functional aspects of gut microbial communities [82], very marginal attention was dedicated to developing methods for understanding these aspects. These multi-omics approaches can provide powerful insides but are extremely expensive and their application resulted mainly limited in size and number by funds availability. The good news is that during the last 10 years COBRA modeling, that was mainly used to simulate the metabolism of single cells, evolved and nowadays, also thanks to the work presented in this thesis, reached the maturity to be applied to simulate multi-species metabolism, investigating functional consequences of microbial abundances. The perfect example of what just discussed is the case of Parkinson's disease. Multiple investigators hypothesized the origin of PD and the related neurodegeneration in the gut [28, 29]. By consequence, in the last five years, several studies were conducted analyzing the composition of the human gut microbiota in PD patients [21, 83, 162, 14, 87, 91, 102]. Even if with some discrepancies and differences between studies, a common picture was outlined, with microbial composition found different in PD patients, especially in few families such as *Verrucomicrobiaceae*. However, no consequent mechanism of this differential microbial enrichment was advanced.

The described case on PD, as well as many other microbiome cases, calls for powerful methods to study the mechanisms behind emerging correlations. The work presented in this thesis tries to answer this call, delivering methods allowing researchers to investigate functional aspects of the matter. Firstly, we developed a toolbox (The Microbiome Modeling Toolbox) to model multi-species (host-microbe and microbe-microbe) scenarios [12]. Within this toolbox, a facile pipeline (mgPipe) is implemented, allowing the researchers to produce and interrogate -on a large scale- personalized microbiota models through the integration with metagenomic data. Secondly, we expanded this pipeline adding steps for relative abundance retrieval from 16S rRNA gene sequencing and whole-genome sequencing. Our aim was to develop a fast and automated tool combining multiple steps from different fields, allowing study replication and non-filed experts to be able to conduct such analysis. In this way, microbial sequencing data can be functionally explored. The advantage of this approach is that it poses its basis on analysis of already available data, surely mgPipe can be used for future studies, but can also help investigators to revisit old studies without the necessity of producing new data. We demonstrated this in two recent studies on IBD and PD [86, 78] where we applied this approach to already finalized investigations. We were able to give new conclusions from already studied data, and in one case we were able to partially experimentally validate our new predictions [86]. Finally, in the third part of this thesis, we used this approach on new and non-studied before data on PD patients being able to infer, and partially validate against literature, changes in the metabolite secretion of gut microbial communities of PD patients in dependency of differential microbial abundances [13]. Taken together, this work of thesis demonstrates the importance and future central role that COBRA modeling can play in filling the knowledge gaps, advancing our knowledge of the functional metabolic aspect of the human gut microbiota.

5.1 Sequencing data integration and COBRA modeling

The human gut microbiota consists of several hundreds of microbial species [151] interacting together and therefore, with the host. All the knowledge on the human gut microbiota that we currently dispose of comes from the sequencing of microbial communities. However, even if the availability of these data advanced rapidly our knowledge, we are still limited in understanding key functionalities of these species such has their metabolic impact on the host. Understanding the metabolic potential and the metabolic interactions of microbial communities, such as the ones that we harbor in our gut, is definitely not an easy task, and methods to do so, in vitro and in vivo, are expensive and limited. In this regard, computational modeling can represent an attractive target. Thanks to recent efforts, COBRA modeling approaches enabling the study of the metabolism of microbial communities were developed [17]. chapter 2 and chapter 3 describe two connected methodologies to process and integrate microbial sequencing data into COBRA modeling. Thanks to a pipeline named mgPipe, for each sample, starting from sequencing data, personalized microbiota metabolic models can be created and the relative secretion profiles can be computed. For each sample, mgPipe uses a compartmentalization approach to combine several hundreds of organisms GEMs in newly assembled microbiota models integrating at the same time relative abundance information. In each microbiota model, organisms can interact with each other sharing metabolites through a common compartment. Each microbiota model disposes of a community biomass objective function where the sum of each organisms biomass is averaged for their abundance coefficients driving the personalization process and make each microbiota model unique. On the base of a defined media and carbon sources availability, and a selected objective function, the metabolism of each microbial community can be inferred from the secreted metabolites. The action of screening secreted metabolites is called secretion profiling, and secretion profiles are also important to understand how the microbial community could impact the host health. Furthermore, once the secretion profiles are computed microbiota models can be used to trace back each secreted metabolite identifying organisms responsible for their secretion. All these possibilities can greatly contribute to disclose the functional metabolic mechanisms behind gut microbial communities structure and their related impact on human homeostasis. mgPipe was created to be able, from microbial sequencing data, to automatically and efficiently perform the creation and interrogation of personalized microbiota models. Such models can include hundreds of organisms and can be generated for numerous cohorts of individuals. In the first study where mgPipe was used [78], we were able to reconstruct personalized microbiota models for a cohort composed of controls and IBD patients inferring functional metabolic properties related to the metabolism of bile acids and tracing back and comparing among individuals, each microbial contribution to the overall bile acids production.

5.2 Parkinson's disease and personalized microbiota modeling

Recent studies have highlighted a possible role of the human gut microbiota in the development of PD [21, 83, 162, 14, 87, 91, 102]. Such studies were mainly based on microbial sequencing of cohorts composed of PD patients and found alterations in microbial compositions in the presence of the disease. Even though the previous studies signed a hallmark in the comprehension of the human gut microbiota composition of Parkinson's Disease patients, the related results are difficulty generalizable due to the small number of samples and to the low degree of heterogeneity of the described cohorts. In fact, on average the number of patients in each of these studies was quite low, and sometimes uniform patients groups were selected and covariates were eliminated through matching. Moreover, very few details were provided on the functional mechanisms that these reported changes would have triggered. The integration of metagenomics data with a modeling technique for microbial communities based on constraint-based modeling already proved to be helpful to elucidate metabolic functional insights of a similar type of multidimensional datasets [86, 78]. In chapter 4, we analyzed a cohort composed of typical PD patients and controls. Our enrolment criteria wanted to reflect the complexity of the disease, and we selected a numerous number of cases (in the top three of the studies, so far, ever conducted on PD microbiota) under different medications, age, stage of the disease, of different sex, and different degree of motor and non motor symptoms. In agreement with previous investigations, we could find a different enrichment of several bacteria genus and species in PDs also in dependence of specific covariates, and using mgPipe (chapter 2 and chapter 3) we constructed personalized microbiota models. Under a common European diet, a differential abundance of microbial species in the gut of PD patients translated into a different secretion of certain metabolites. Such predictions were largely overlapping with previous ones that we computed in another cohort study [86], differences can be explained by different inclusions criteria and degree of cohort diversity. These predictions can help to shed light on the role of the human gut microbiota in PD, and with what related mechanisms different organisms could contribute to the pathogenesis of the disease.

5.3 Future perspectives

Personalized metabolic human gut microbiota models can be used to infer new mechanistic insights related to specific microbial compositions. The computed predictions can be useful in many ways from evaluating the impact of different diets on the metabolism of gut microbial communities to evaluate the metabolic role of the human gut microbiota in diseases. Understanding the role of the human gut microbiota in specific diseases is crucial to be able to intervene and personalized gut microbiota modeling could be used to evaluate the efficacy of possible interventions such as pro or prebiotics supplementations and fecal microbiota transplant. To this day probiotics interventions are mainly based on supplementation of few so-called “beneficial” strains. However, the human gut microbiota is a complex system [17], and we have no idea of the role and behavior of these strains which could be dependent on the microbial population already present. Furthermore, even if a more stringent and comprehensive experimental validation of the predictions is required, at the moment, being able to formulate testable hypothesis on gut microbial metabolic mechanisms is already something of great value that could speed the process of discovery delivering possible testable targets. Such predictions could be expanded through integration with the whole-body metabolism reconstruction [181]. The personalized microbiota models created by mgPipe are already compatible with the whole body metabolism reconstruction [181] and their integration could be extremely helpful to understand how the microbiome could metabolically impact organs that are anatomically far from the gut, such as the brain.

Potentially, personalized gut microbial metabolic modeling (PGM) could find application also in other fields. The related predictions can be used to reveal mechanistic insights for different diseases as we demonstrated in chapter 4, but could also potentially revolutionize the world of nutrition. To this day there is not any diet that takes into account the ensemble of microbial composition and functions. A diet that would reflect microbiota composition and function could be more effective and potentially, for specific diseases, even therapeutic.

Finally, the work presented in this thesis could contribute to assemble and refine diagnosis systems for several diseases. If bigger datasets would be needed for such an application, in any case, the work presented in this thesis can be seen as the first milestone towards personalized gut microbiota analysis.

Bibliography

[1] Abbott, R. D., Petrovitch, H., White, L., Masaki, K., Tanner, C., Curb, J., Grandinetti, A., Blanchette, P., Popper, J., and Ross, G. (2001). Frequency of bowel movements and the future risk of parkinson's disease. *Neurology*, 57(3):456–462.

[2] Abbott, R. D., Ross, G. W., Petrovitch, H., Tanner, C. M., Davis, D. G., Masaki, K. H., Launer, L. J., Curb, J. D., and White, L. R. (2007). Bowel movement frequency in late-life and incidental lewy bodies. *Movement disorders: official journal of the Movement Disorder Society*, 22(11):1581–1586.

[3] Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology*, 8(6):e1002358.

[4] Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., and Polz, M. F. (2004). Divergence and redundancy of 16s rrna sequences in genomes with multiple rrn operons. *Journal of bacteriology*, 186(9):2629–2635.

[5] Adlerberth, I. and Wold, A. (2009). Establishment of the gut microbiota in western infants. *Acta paediatrica*, 98(2):229–238.

[6] Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS computational biology*, 9(3):e1002980.

[7] Albin, R. L., Young, A. B., and Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in neurosciences*, 12(10):366–375.

[8] Allard, G., Ryan, F. J., Jeffery, I. B., and Claesson, M. J. (2015). Spingo: a rapid species-classifier for microbial amplicon sequences. *BMC bioinformatics*, 16(1):324.

[9] Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., et al. (2018). Kbase: the united states department of energy systems biology knowledgebase. *Nature Biotechnology*, 36(7).

[10] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174.

[11] Aurich, M. K. and Thiele, I. (2016). Computational modeling of human metabolism and its application to systems biomedicine. In *Systems medicine*, pages 253–281. Springer.

[12] Baldini, F., Heinken, A., Heirendt, L., Magnusdottir, S., Fleming, R. M., and Thiele, I. (2018). The microbiome modeling toolbox: from microbial interactions to personalized microbial communities. *Bioinformatics*, 35(13):2332–2334.

[13] Baldini, F., Hertel, J., Sandt, E., Thinnes, C., Neuberger-Castillo, L., Pavelka, L., Betsou, F., Krueger, R., and Thiele, I. (2019). Parkinson's disease-associated alterations of the gut microbiome can invoke disease-relevant metabolic changes. *bioRxiv*, page 691030.

[14] Barichella, M., Severgnini, M., Cilia, R., Cassani, E., Bolliri, C., Caronni, S., Ferri, V., Cancello, R., Ceccarani, C., Faierman, S., et al. (2019). Unraveling gut microbiota in parkinson's disease and atypical parkinsonism. *Movement Disorders*, 34(3):396–405.

[15] Barton, W., Penney, N. C., Cronin, O., Garcia-Perez, I., Molloy, M. G., Holmes, E., Shanahan, F., Cotter, P. D., and O'Sullivan, O. (2018). The microbiome of professional athletes differs from that of more sedentary subjects in composition and particularly at the functional metabolic level. *Gut*, 67(4):625–633.

[16] Bauer, E. and Thiele, I. (2018a). From metagenomic data to personalized in silico microbiotas: predicting dietary supplements for crohn's disease. *NPJ systems biology and applications*, 4(1):27.

[17] Bauer, E. and Thiele, I. (2018b). From network analysis to functional metabolic modeling of the human gut microbiota. *MSystems*, 3(3):e00209–17.

[18] Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., and Kaleta, C. (2017). Bacarena: individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLoS computational biology*, 13(5):e1005544.

[19] Becker, N., Kunath, J., Loh, G., and Blaut, M. (2011). Human intestinal microbiota: characterization of a simplified and stable gnotobiotic rat model. *Gut microbes*, 2(1):25–33.

[20] Becker, S. A. and Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082.

[21] Bedarf, J. R., Hildebrand, F., Coelho, L. P., Sunagawa, S., Bahram, M., Goeser, F., Bork, P., and Wüllner, U. (2017). Functional implications of microbial and viral gut metagenome changes in early stage l-dopa-naïve parkinson's disease patients. *Genome medicine*, 9(1):39.

[22] Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 72(4):405–416.

[23] Biggs, M. B., Medlock, G. L., Kolling, G. L., and Papin, J. A. (2015). Metabolic network modeling of microbial communities. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(5):317–334.

[24] Biggs, M. B. and Papin, J. A. (2013). Novel multiscale modeling tool applied to pseudomonas aeruginosa biofilm formation. *PLoS One*, 8(10):e78011.

[25] Bonifati, V., Rizzu, P., Squitieri, F., Krieger, E., Vanacore, N. a., Van Swieten, J., Brice, A., Van Duijn, C., Oostra, B., Meco, G., et al. (2003). Dj-1 (park7), a novel gene for autosomal recessive, early onset parkinsonism. *Neurological Sciences*, 24(3):159–160.

[26] Boren, J., Lee, W.-N. P., Bassilian, S., Centelles, J. J., Lim, S., Ahmed, S., Boros, L. G., and Cascante, M. (2003). The stable isotope-based dynamic metabolic profile of butyrate-induced ht29 cell differentiation. *Journal of Biological Chemistry*, 278(31):28395–28402.

[27] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

[28] Braak, H., Del Tredici, K., Rüb, U., De Vos, R. A., Steur, E. N. J., and Braak, E. (2003). Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of aging*, 24(2):197–211.

[29] Braak, H., Ghebremedhin, E., Rüb, U., Bratzke, H., and Del Tredici, K. (2004). Stages in the development of parkinson's disease-related pathology. *Cell and tissue research*, 318(1):121–134.

[30] Bravo, J. A., Forsythe, P., Chew, M. V., Escaravage, E., Savignac, H. M., Dinan, T. G., Bienenstock, J., and Cryan, J. F. (2011). Ingestion of lactobacillus strain regulates emotional behavior and central gaba receptor expression in a mouse via the vagus nerve. *Proceedings of the National Academy of Sciences*, 108(38):16050–16055.

[31] Brito, I., Yilmaz, S., Huang, K., Xu, L., Jupiter, S., Jenkins, A., Naisilisili, W., Tamminen, M., Smillie, C., Wortman, J., et al. (2017). Corrigendum: Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 544(7648):124.

[32] Brown, C. T., Olm, M. R., Thomas, B. C., and Banfield, J. F. (2016). Measurement of bacterial replication rates in microbial communities. *Nature biotechnology*, 34(12):1256.

[33] Brunk, E., Sahoo, S., et al. (2018). Recon3d enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36(3):272.

[34] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581.

[35] Caporaso, J. G., Kuczynski, J., Stombaugh, J., et al. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335.

[36] Carabotti, M., Scirocco, A., Maselli, M. A., and Severi, C. (2015). The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology*, 28(2):203.

[37] Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., et al. (2013). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 42(D1):D459–D471.

[38] Cersosimo, M. G., Raina, G. B., Pecci, C., Pellene, A., Calandra, C. R., Gutiérrez, C., Micheli, F. E., and Benarroch, E. E. (2013). Gastrointestinal manifestations in parkinson's disease: prevalence and occurrence before motor symptoms. *Journal of neurology*, 260(5):1332–1338.

[39] Chan, S. H. J., Simons, M. N., and Maranas, C. D. (2017). Steadycom: Predicting microbial abundances while ensuring community stability. *PLoS computational biology*, 13(5):e1005539.

[40] Chaudhuri, K. R., Martinez-Martin, P., Schapira, A. H., Stocchi, F., Sethi, K., Odin, P., Brown, R. G., Koller, W., Barone, P., MacPhee, G., et al. (2006). International multicenter pilot study of the first comprehensive self-completed nonmotor symptoms questionnaire for parkinson's disease: the nmsquest study. *Movement disorders: official journal of the Movement Disorder Society*, 21(7):916–923.

[41] Claesson, M. J., Clooney, A. G., and O'toole, P. W. (2017). A clinician's guide to microbiome analysis. *Nature Reviews Gastroenterology & Hepatology*, 14(10):585.

[42] Claesson, M. J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J. R., Falush, D., Dinan, T., Fitzgerald, G., et al. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4586–4591.

[43] Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–1270.

[44] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. abingdon.

[45] Crichton, G. E. and Alkerwi, A. (2014). Association of sedentary behavior time with ideal cardiovascular health: the oriscav-lux study. *PLoS One*, 9(6):e99829.

[46] De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from europe and rural africa. *Proceedings of the National Academy of Sciences*, 107(33):14691–14696.

[47] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Appl. Environ. Microbiol.*, 72(7):5069–5072.

[48] Dethlefsen, L., Huse, S., Sogin, M. L., and Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rrna sequencing. *PLoS biology*, 6(11):e280.

[49] Devkota, S., Wang, Y., Musch, M. W., Leone, V., Fehlner-Peach, H., Nadimpalli, A., Antonopoulos, D. A., Jabri, B., and Chang, E. B. (2012). Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in il10-/- mice. *Nature*, 487(7405):104.

[50] Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A. A., and Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the seed and model seed. In *Systems Metabolic Engineering*, pages 17–45. Springer.

[51] Di Fonzo, A., Dekker, M., Montagna, P., Baruzzi, A., Yonova, E., Guedes, L. C., Szczerbinska, A., Zhao, T., Dubbel-Hulsman, L., Wouters, C., et al. (2009). Fbxo7 mutations cause autosomal recessive, early-onset parkinsonian-pyramidal syndrome. *Neurology*, 72(3):240–245.

[52] Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107(26):11971–11975.

[53] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782.

[54] Earle, K. A., Billings, G., Sigal, M., Lichtman, J. S., Hansson, G. C., Elias, J. E., Amieva, M. R., Huang, K. C., and Sonnenburg, J. L. (2015). Quantitative imaging of gut microbiota spatial organization. *Cell host & microbe*, 18(4):478–488.

[55] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461.

[56] Edwards, L., Pfeiffer, R., Quigley, E., Hofman, R., and Balluff, M. (1991). Gastrointestinal symptoms in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 6(2):151–156.

[57] Ehrlich, S. D., Consortium, M., et al. (2011). Metahit: The european union project on metagenomics of the human intestinal tract. In *Metagenomics of the human body*, pages 307–316. Springer.

[58] El-Semman, I. E., Karlsson, F. H., Shoaie, S., Nookaew, I., Soliman, T. H., and Nielsen, J. (2014). Genome-scale metabolic reconstructions of bifidobacterium adolescentis l2-32 and faecalibacterium prausnitzii a2-165 and their interaction. *BMC systems biology*, 8(1):41.

[59] Fasano, A., Visanji, N. P., Liu, L. W., Lang, A. E., and Pfeiffer, R. F. (2015). Gastrointestinal dysfunction in parkinson's disease. *The Lancet Neurology*, 14(6):625–639.

[60] Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology*, 8(7):e1002606.

[61] Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129.

[62] Forsyth, C. B., Shannon, K. M., Kordower, J. H., Voigt, R. M., Shaikh, M., Jaglin, J. A., Estes, J. D., Dodiya, H. B., and Keshavarzian, A. (2011). Increased intestinal permeability correlates with sigmoid mucosa alpha-synuclein staining and endotoxin exposure markers in early parkinson's disease. *PloS one*, 6(12):e28032.

[63] Freilich, S., Zarecki, R., Eilam, O., Segal, E. S., Henry, C. S., Kupiec, M., Gophna, U., Sharan, R., and Ruppin, E. (2011). Competitive and cooperative metabolic interactions in bacterial communities. *Nature communications*, 2:589.

[64] Gatto, N. M., Rhodes, S. L., Manthripragada, A. D., Bronstein, J., Cockburn, M., Farrer, M., and Ritz, B. (2010). α-synuclein gene may interact with environmental factors in increasing risk of parkinson's disease. *Neuroepidemiology*, 35(3):191–195.

[65] Gevers, D., Knight, R., Petrosino, J. F., Huang, K., McGuire, A. L., Birren, B. W., Nelson, K. E., White, O., Methé, B. A., and Huttenhower, C. (2012). The human microbiome project: a community resource for the healthy human microbiome. *PLoS biology*, 10(8):e1001377.

[66] Gibson, G. R., Rastall, R. A., and Fuller, R. (2003). The health benefits of probiotics and prebiotics. *Gut flora, nutrition, immunity and health*, pages 52–76.

[67] Gobert, A. P., Sagrestani, G., Delmas, E., Wilson, K. T., Verriere, T. G., Dapoigny, M., Del'homme, C., and Bernalier-Donadille, A. (2016). The human intestinal microbiota of constipated-predominant irritable bowel syndrome patients exhibits anti-inflammatory properties. *Scientific reports*, 6:39399.

[68] Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., et al. (2008). Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170.

[69] Graham, S., Rey, N., Ugur, Z., Yilmaz, A., Sherman, E., Maddens, M., Bahado-Singh, R., Becker, K., Schulz, E., Meyerdirk, L., et al. (2018). Metabolomic profiling of bile acids in an experimental model of prodromal parkinson's disease. *Metabolites*, 8(4):71.

[70] Grander, C., Adolph, T. E., Wieser, V., Lowe, P., Wrzosek, L., Gyongyosi, B., Ward, D. V., Grabherr, F., Gerner, R. R., Pfister, A., et al. (2018). Recovery of ethanol-induced akkermansia muciniphila depletion ameliorates alcoholic liver disease. *Gut*, 67(5):891–901.

[71] Granger, B. R., Chang, Y.-C., Wang, Y., DeLisi, C., Segre, D., and Hu, Z. (2016). Visualization of metabolic interaction networks in microbial communities using visant 5.0. *PLoS computational biology*, 12(4):e1004875.

[72] Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC bioinformatics*, 11(1):489.

[73] Hanly, T. J. and Henson, M. A. (2011). Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures. *Biotechnology and bioengineering*, 108(2):376–385.

[74] Hanly, T. J. and Henson, M. A. (2014). Dynamic model-based analysis of furfural and hmf detoxification by pure and mixed batch cultures of s. cerevisiae and s. stipitis. *Biotechnology and bioengineering*, 111(2):272–284.

[75] Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., Bonilla, G., Kar, A., Leiby, N., Mehta, P., et al. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell reports*, 7(4):1104–1115.

[76] Harrell, F. E. (2001). Resampling, validating, describing, and simplifying the model. In *Regression modeling strategies*, pages 87–103. Springer.

[77] Hasegawa, S., Goto, S., Tsuji, H., Okuno, T., Asahara, T., Nomoto, K., Shibata, A., Fujisawa, Y., Minato, T., Okamoto, A., et al. (2015). Intestinal dysbiosis and lowered serum lipopolysaccharide-binding protein in parkinson's disease. *PloS one*, 10(11):e0142164.

[78] Heinken, A., Ravcheev, D. A., Baldini, F., Heirendt, L., Fleming, R. M., and Thiele, I. (2019). Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome*, 7(1):75.

[79] Heinken, A., Sahoo, S., Fleming, R. M., and Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut microbes*, 4(1):28–40.

[80] Heinken, A. and Thiele, I. (2015a). Anoxic conditions promote species-specific mutualism between gut microbes in silico. *Applied and environmental microbiology*, 81(12):4049–4061.

[81] Heinken, A. and Thiele, I. (2015b). Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes*, 6(2):120–130.

[82] Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., Wampach, L., Schneider, J. G., Hogan, A., De Beaufort, C., et al. (2017). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology*, 2(1):16180.

[83] Heintz-Buschart, A., Pandey, U., Wicke, T., Sixel-Döring, F., Janzen, A., Sittig-Wiegand, E., Trenkwalder, C., Oertel, W. H., Mollenhauer, B., and Wilmes, P. (2018). The nasal and gut microbiome in parkinson's disease and idiopathic rapid eye movement sleep behavior disorder. *Movement Disorders*, 33(1):88–98.

[84] Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdottir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639.

[85] Heirendt, L., Thiele, I., and Fleming, R. M. (2017). Distributedfba. jl: high-level, high-performance flux balance analysis in julia. *Bioinformatics*, 33(9):1421–1423.

[86] Hertel, J., Harms, A. C., Heinken, A., Baldini, F., Thinnes, C. C., Glaab, E., Vasco, D. A., Pietzner, M., Stewart, I. D., Wareham, N. J., et al. (2019). Integrated analyses of microbiome and longitudinal metabolome data reveal microbial-host interactions on sulfur metabolism in parkinson's disease. *Cell reports*, 29(7):1767–1777.

[87] Hill-Burns, E. M., Debelius, J. W., Morton, J. T., Wissemann, W. T., Lewis, M. R., Wallen, Z. D., Peddada, S. D., Factor, S. A., Molho, E., Zabetian, C. P., et al. (2017). Parkinson's disease and parkinson's disease medications have distinct signatures of the gut microbiome. *Movement Disorders*, 32(5):739–749.

[88] Hipp, G., Vaillant, M., Diederich, N. J., Roomp, K., Satagopam, V. P., Banda, P., Sandt, E., Mommaerts, K., Mosch, S., Longhino, L., et al. (2018). The luxembourg parkinson's study: A comprehensive approach for stratification and early diagnosis. *Frontiers in aging neuroscience*, 10:326.

[89] Hoehn, M. M. and Yahr, M. D. (1967). Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427.

[90] Hoffmann, C., Dollive, S., Grunberg, S., Chen, J., Li, H., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2013). Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. *PloS one*, 8(6):e66019.

[91] Hopfner, F., Künstner, A., Müller, S. H., Künzel, S., Zeuner, K. E., Margraf, N. G., Deuschl, G., Baines, J. F., and Kuhlenbäumer, G. (2017). Gut microbiota in parkinson disease in a northern german cohort. *Brain research*, 1667:41–45.

[92] Huang, R., Li, T., Ni, J., Bai, X., Gao, Y., Li, Y., Zhang, P., and Gong, Y. (2018). Different sex-based responses of gut microbiota during the development of hepatocellular carcinoma in liver-specific tsc1-knockout mice. *Frontiers in microbiology*, 9:1008.

[93] Hyland, N. P. and Cryan, J. F. (2010). A gut feeling about gaba: focus on gabab receptors. *Frontiers in pharmacology*, 1:124.

[94] Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

[95] Iliev, I. D., Funari, V. A., Taylor, K. D., Nguyen, Q., Reyes, C. N., Strom, S. P., Brown, J., Becker, C. A., Fleshner, P. R., Dubinsky, M., et al. (2012). Interactions between commensal fungi and the c-type lectin receptor dectin-1 influence colitis. *Science*, 336(6086):1314–1317.

[96] Janda, J. M. and Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9):2761–2764.

[97] Jernberg, C., Löfmark, S., Edlund, C., and Jansson, J. K. (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME journal*, 1(1):56.

[98] Jiménez, E., Marín, M. L., Martín, R., Odriozola, J. M., Olivares, M., Xaus, J., Fernández, L., and Rodríguez, J. M. (2008). Is meconium from healthy newborns actually sterile? *Research in microbiology*, 159(3):187–193.

[99] Kalia, L. V., Lang, A. E., Hazrati, L.-N., Fujioka, S., Wszolek, Z. K., Dickson, D. W., Ross, O. A., Van Deerlin, V. M., Trojanowski, J. Q., Hurtig, H. I., et al. (2015). Clinical correlations with lewy body pathology in lrrk2-related parkinson disease. *JAMA neurology*, 72(1):100–105.

[100] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361.

[101] Karlsson, F. H., Nookaew, I., and Nielsen, J. (2014). Metagenomic data utilization and analysis (medusa) and construction of a global gut microbial gene catalogue. *PLoS computational biology*, 10(7):e1003706.

[102] Keshavarzian, A., Green, S. J., Engen, P. A., Voigt, R. M., Naqib, A., Forsyth, C. B., Mutlu, E., and Shannon, K. M. (2015). Colonic bacterial composition in parkinson's disease. *Movement Disorders*, 30(10):1351–1360.

[103] Khalif, I., Quigley, E., Konovitch, E., and Maximova, I. (2005). Alterations in the colonic flora and intestinal permeability and evidence of immune activation in chronic constipation. *Digestive and Liver Disease*, 37(11):838–849.

[104] Kitada, T., Asakawa, S., Hattori, N., Matsumine, H., Yamamura, Y., Minoshima, S., Yokochi, M., Mizuno, Y., and Shimizu, N. (1998). Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*, 392(6676):605.

[105] Klitgord, N. and Segrè, D. (2010). Environments that induce synthetic microbial ecosystems. *PLoS computational biology*, 6(11):e1001002.

[106] Klitgord, N. and Segrè, D. (2011). Ecosystems biology of microbial metabolism. *Current opinion in biotechnology*, 22(4):541–546.

[107] Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585.

[108] Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., Clancy, T. E., Chung, D. C., Lochhead, P., Hold, G. L., et al. (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell host & microbe*, 14(2):207–215.

[109] Larsen, S., Hanss, Z., and Krüger, R. (2018). The genetic architecture of mitochondrial dysfunction in parkinson's disease. *Cell and tissue research*, 373(1):21–37.

[110] Laue, H. and Cook, A. M. (2000). Biochemical and molecular characterization of taurine: pyruvate aminotransferase from the anaerobe bilophila wadsworthia. *European journal of biochemistry*, 267(23):6841–6848.

[111] LeBlanc, J. G., Milani, C., de Giori, G. S., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Current opinion in biotechnology*, 24(2):160–168.

[112] Lehmann, S., Guadagni, F., Moore, H., Ashton, G., Barnes, M., Benson, E., Clements, J., Koppandi, I., Coppola, D., Demiroglu, S. Y., et al. (2012). Standard preanalytical coding for biospecimens: Review and implementation of the sample preanalytical code (sprec). *Biopreservation and biobanking*, 10(4):366–374.

[113] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[114] Li, W., Wu, X., Hu, X., Wang, T., Liang, S., Duan, Y., Jin, F., and Qin, B. (2017). Structural changes of gut microbiota in parkinson's disease and its correlation with clinical features. *Science China Life Sciences*, 60(11):1223–1233.

[115] Lin, A., Zheng, W., He, Y., Tang, W., Wei, X., He, R., Huang, W., Su, Y., Huang, Y., Zhou, H., et al. (2018). Gut microbiota in patients with parkinson's disease in southern china. *Parkinsonism & related disorders*, 53:82–88.

[116] Machado, D., Andrejev, S., Tramontano, M., and Patil, K. R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research*, 46(15):7542–7553.

[117] Magnúsdóttir, S., Heinken, A., Kutt, L., , et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature biotechnology*, 35(1):81.

[118] Magnúsdóttir, S. and Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current opinion in biotechnology*, 51:90–96.

[119] Mahadevan, R., Edwards, J. S., and Doyle III, F. J. (2002). Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical journal*, 83(3):1331–1340.

[120] Mahadevan, R. and Schilling, C. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–276.

[121] Maidak, B., Cole, J., and Lilburn Jr, T. (2001). Ctp the rdp-ii (ribosomal database project). *Nucleic acids research*, 29:173–174.

[122] Mathay, C., Hamot, G., Henry, E., Georges, L., Bellora, C., Lebrun, L., de Witt, B., Ammerlaan, W., Buschart, A., Wilmes, P., et al. (2015). Method optimization for fecal sample collection and fecal dna extraction. *Biopreservation and biobanking*, 13(2):79–93.

[123] Mayer, E. A., Tillisch, K., and Gupta, A. (2015). Gut/brain axis and the microbiota. *The Journal of clinical investigation*, 125(3):926–938.

[124] Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome research*, 21(10):1616–1625.

[125] Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., Salazar, G. A., Pesseat, S., Boland, M. A., Hunter, F. M. I., et al. (2017). Ebi metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic acids research*, 46(D1):D726–D735.

[126] Naito, Y., Uchiyama, K., and Takagi, T. (2018). A next-generation beneficial microbe: Akkermansia muciniphila. *Journal of clinical biochemistry and nutrition*, 63(1):33–35.

[127] Narayanasamy, S., Jarosz, Y., Muller, E. E., Heintz-Buschart, A., Herold, M., Kaysen, A., Laczny, C. C., Pinel, N., May, P., and Wilmes, P. (2016). Imp: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome biology*, 17(1):260.

[128] Natividad, J. M., Lamas, B., Pham, H. P., Michel, M.-L., Rainteau, D., Bridonneau, C., Da Costa, G., van Hylckama Vlieg, J., Sovran, B., Chamignon, C., et al. (2018). Bilophila wadsworthia aggravates high fat diet induced metabolic dysfunctions in mice. *Nature communications*, 9(1):2802.

[129] Nichols, W., Pankratz, N., Marek, D., Pauciulo, M., Elsaesser, V., Halter, C., Rudolph, A., Wojcieszek, J., Pfeiffer, R., Foroud, T., et al. (2009). Mutations in gba are associated with familial parkinson disease susceptibility and age at onset. *Neurology*, 72(4):310–316.

[130] Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., and Pettersson, S. (2012). Host-gut microbiota metabolic interactions. *Science*, 336(6086):1262–1267.

[131] Nielsen, J. (2017). Systems biology of metabolism: a driver for developing personalized and precision medicine. *Cell metabolism*, 25(3):572–579.

[132] Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., Daníelsdóttir, A. D., Krecke, M., Merten, D., Haraldsdóttir, H. S., et al. (2018). The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic acids research*, 47(D1):D614–D624.

[133] Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular systems biology*, 5(1).

[134] Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299.

[135] Ogawa, H., Rafiee, P., Fisher, P. J., Johnson, N. A., Otterson, M. F., and Binion, D. G. (2003). Butyrate modulates gene and protein expression in human intestinal endothelial cells. *Biochemical and biophysical research communications*, 309(3):512–519.

[136] Olzmann, J. A., Brown, K., Wilkinson, K. D., Rees, H. D., Huai, Q., Ke, H., Levey, A. I., Li, L., and Chin, L.-S. (2004). Familial parkinson's disease-associated l166p mutation disrupts dj-1 protein folding and function. *Journal of Biological Chemistry*, 279(9):8506–8515.

[137] O'Mahony, S. M., Clarke, G., Borre, Y., Dinan, T., and Cryan, J. (2015). Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. *Behavioural brain research*, 277:32–48.

[138] Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3):245.

[139] Paisán-Ruız, C., Jain, S., Evans, E. W., Gilks, W. P., Simón, J., Van Der Brug, M., De Munain, A. L., Aparicio, S., Gil, A. M., Khan, N., et al. (2004). Cloning of the gene containing mutations that cause park8-linked parkinson's disease. *Neuron*, 44(4):595–600.

[140] Pál, C., Papp, B., and Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature genetics*, 37(12):1372.

[141] Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS biology*, 5(7):e177.

[142] Palsson, B. (2006). Systems biology: properties of reconstructed networks. *Cambridge: Cambridge Univ Pr.*

[143] Pan-Montojo, F., Anichtchik, O., Dening, Y., Knels, L., Pursche, S., Jung, R., Jackson, S., Gille, G., Spillantini, M. G., Reichmann, H., et al. (2010). Progression of parkinson's disease pathology is reproduced by intragastric administration of rotenone in mice. *PloS one*, 5(1):e8762.

[144] Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of applied econometrics*, 11(6):619–632.

[145] Parikh, H. I., Koparde, V. N., Bradley, S. P., Buck, G. A., and Sheth, N. U. (2016). Mefit: merging and filtering tool for illumina paired-end reads for 16s rrna amplicon sequencing. *BMC bioinformatics*, 17(1):491.

[146] Penders, J., Thijs, C., van den Brandt, P. A., Kummeling, I., Snijders, B., Stelma, F., Adams, H., van Ree, R., and Stobberingh, E. E. (2007). Gut microbiota composition and development of atopic manifestations in infancy: the koala birth cohort study. *Gut*, 56(5):661–667.

[147] Petrov, V., Saltykova, I., Zhukova, I., Alifirova, V., Zhukova, N., Dorofeeva, Y. B., Tyakht, A., Kovarsky, B., Alekseev, D., Kostryukova, E., et al. (2017). Analysis of gut microbiota in patients with parkinson's disease. *Bulletin of experimental biology and medicine*, 162(6):734–737.

[148] Petrovitch, H., Abbott, R. D., Ross, G. W., Nelson, J., Masaki, K. H., Tanner, C. M., Launer, L. J., and White, L. R. (2009). Bowel movement frequency in late-life and substantia nigra neuron density at death. *Movement disorders*, 24(3):371–376.

[149] Pfeiffer, R. F. (2003). Gastrointestinal dysfunction in parkinson's disease. *The lancet neurology*, 2(2):107–116.

[150] Qian, Y., Yang, X., Xu, S., Wu, C., Song, Y., Qin, N., Chen, S.-D., and Xiao, Q. (2018). Alteration of the fecal microbiota in chinese patients with parkinson's disease. *Brain, behavior, and immunity*, 70:194–202.

[151] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59.

[152] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55.

[153] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596.

[154] Raman, K., Rajagopalan, P., and Chandra, N. (2005). Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS computational biology*, 1(5):e46.

[155] Romenets, S. R., Wolfson, C., Galatas, C., Pelletier, A., Altman, R., Wadup, L., and Postuma, R. (2012). Validation of the non-motor symptoms questionnaire (nms-quest). *Parkinsonism & related disorders*, 18(1):54–58.

[156] Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., and Han, Y. W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating e-cadherin/β-catenin signaling via its fada adhesin. *Cell host & microbe*, 14(2):195–206.

[157] Ruiz-Castell, M., Kandala, N.-B., Kuemmerle, A., Schritz, A., Barré, J., Delagardelle, C., Krippler, S., Schmit, J.-C., and Stranges, S. (2016). Hypertension burden in luxembourg: individual risk factors and geographic variations, 2013 to 2015 european health examination survey. *Medicine*, 95(36).

[158] Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schretter, C. E., Rocha, S., Gradinaru, V., et al. (2016). Gut microbiota regulate motor deficits and neuroinflammation in a model of parkinson's disease. *Cell*, 167(6):1469–1480.

[159] Santala, S., Efimova, E., Kivinen, V., Larjo, A., Aho, T., Karp, M., and Santala, V. (2011). Improved triacylglycerol production in acinetobacter baylyi adp1 by metabolic engineering. *Microbial cell factories*, 10(1):36.

[160] Saransaari, P. and Oja, S. (2007). Taurine release in mouse brain stem slices under cell-damaging conditions. *Amino acids*, 32(3):439–446.

[161] Savica, R., Carlin, J., Grossardt, B., Bower, J. H., Ahlskog, J., Maraganore, D., Bharucha, A. E., and Rocca, W. A. (2009). Medical records documentation of constipation preceding parkinson disease: A case-control study. *Neurology*, 73(21):1752–1758.

[162] Scheperjans, F., Aho, V., Pereira, P. A., Koskinen, K., Paulin, L., Pekkonen, E., Haapaniemi, E., Kaakkola, S., Eerola-Rautio, J., Pohja, M., et al. (2015). Gut microbiota are related to parkinson's disease and clinical phenotype. *Movement disorders*, 30(3):350–358.

[163] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541.

[164] Schneeberger, M., Everard, A., Gómez-Valadés, A. G., Matamoros, S., Ramírez, S., Delzenne, N. M., Gomis, R., Claret, M., and Cani, P. D. (2015). Akkermansia muciniphila inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice. *Scientific reports*, 5:16643.

[165] Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., Huttenhower, C., and Izard, J. (2012a). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome biology*, 13(6):R42.

[166] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012b). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811.

[167] Shah, P., Fritz, J. V., Glaab, E., Desai, M. S., Greenhalgh, K., Frachet, A., Niegowska, M., Estes, M., Jäger, C., Seguin-Devaux, C., et al. (2016). A microfluidics-based in vitro model of the gastrointestinal human–microbe interface. *Nature communications*, 7:11535.

[168] Shoaie, S., Karlsson, F., Mardinoglu, A., Nookaew, I., Bordel, S., and Nielsen, J. (2013). Understanding the interactions between bacteria in the human gut through metabolic modeling. *Scientific reports*, 3:2532.

[169] Sleator, R. D. (2010). The human superorganism–of microbes and men. *Medical hypotheses*, 74(2):214–215.

[170] Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., and Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241.

[171] Sommer, M. O., Dantas, G., and Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*, 325(5944):1128–1131.

[172] Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., and Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular systems biology*, 3(1).

[173] Strandwitz, P., Kim, K. H., Terekhova, D., Liu, J. K., Sharma, A., Levering, J., McDonald, D., Dietrich, D., Ramadhar, T. R., Lekbua, A., et al. (2019). Gaba-modulating bacteria of the human gut microbiota. *Nature microbiology*, 4(3):396.

[174] Streubel, A., Siepmann, J., and Bodmeier, R. (2006). Drug delivery to the upper small intestine window using gastroretentive technologies. *Current opinion in pharmacology*, 6(5):501–508.

[175] Sullivan, Å., Edlund, C., and Nord, C. E. (2001). Effect of antimicrobial agents on the ecological balance of human microflora. *The Lancet infectious diseases*, 1(2):101–114.

[176] Szappanos, B., Fritzemeier, J., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R. A., et al. (2016). Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nature communications*, 7:11607.

[177] Taffs, R., Aston, J. E., Brileya, K., Jay, Z., Klatt, C. G., McGlynn, S., Mallette, N., Montross, S., Gerlach, R., Inskeep, W. P., et al. (2009). In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC systems biology*, 3(1):114.

[178] Tanner, C. M., Kamel, F., Ross, G. W., Hoppin, J. A., Goldman, S. M., Korell, M., Marras, C., Bhudhikanok, G. S., Kasten, M., Chade, A. R., et al. (2011). Rotenone, paraquat, and parkinson's disease. *Environmental health perspectives*, 119(6):866–872.

[179] Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. O. (2004). Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC bioinformatics*, 5(1):163.

[180] Thiele, I. and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93.

[181] Thiele, I., Sahoo, S., Heinken, A., Heirendt, L., Aurich, M. K., Noronha, A., and Fleming, R. M. (2018). When metabolism meets physiology: Harvey and harvetta. *bioRxiv*, page 255885.

[182] Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., et al. (2013). A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419.

[183] Tobalina, L., Bargiela, R., Pey, J., Herbst, F.-A., Lores, I., Rojo, D., Barbas, C., Peláez, A. I., Sánchez, J., von Bergen, M., et al. (2015). Context-specific metabolic network reconstruction of a naphthalene-degrading bacterial community guided by metaproteomic data. *Bioinformatics*, 31(11):1771–1779.

[184] Tonini, M., Crema, A., Frigo, G., Rizzi, C., Manzo, L., Candura, S., and Onori, L. (1989). An in vitro study of the relationship between gaba receptor function and propulsive motility in the distal colon of the rabbit. *British journal of pharmacology*, 98(4):1109–1118.

[185] Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, 12(10):902.

[186] Turnbaugh, P. J., Bäckhed, F., Fulton, L., and Gordon, J. I. (2008). Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell host & microbe*, 3(4):213–223.

[187] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027.

[188] Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science translational medicine*, 1(6):6ra14–6ra14.

[189] Ursell, L. K., Clemente, J. C., Rideout, J. R., Gevers, D., Caporaso, J. G., and Knight, R. (2012). The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites. *Journal of allergy and clinical immunology*, 129(5):1204–1208.

[190] Vaishampayan, P. A., Kuehl, J. V., Froula, J. L., Morgan, J. L., Ochman, H., and Francino, M. P. (2010). Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome biology and evolution*, 2:53–66.

[191] Van Hoek, M. J. and Merks, R. M. (2017). Emergence of microbial diversity due to cross-feeding interactions in a spatial model of gut microbial metabolism. *BMC systems biology*, 11(1):56.

[192] van Kessel, S. P., Frye, A. K., El-Gendy, A. O., Castejon, M., Keshavarzian, A., van Dijk, G., and El Aidy, S. (2019). Gut bacterial tyrosine decarboxylases restrict levels of levodopa in the treatment of parkinson's disease. *Nature communications*, 10(1):310.

[193] Vandeputte, D., Falony, G., Vieira-Silva, S., Tito, R. Y., Joossens, M., and Raes, J. (2016). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut*, 65(1):57–62.

[194] Vandeputte, D., Falony, G., Vieira-Silva, S., Wang, J., Sailer, M., Theis, S., Verbeke, K., and Raes, J. (2017). Prebiotic inulin-type fructans induce specific changes in the human gut microbiota. *Gut*, 66(11):1968–1974.

[195] Vanhoutvin, S., Troost, F., Kilkens, T., Lindsey, P., Hamer, H., Jonkers, D., Venema, K., and Brummer, R.-j. M. (2009). The effects of butyrate enemas on visceral perception in healthy volunteers. *Neurogastroenterology & motility*, 21(9):952–e76.

[196] Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267.

[197] Wang, W., Chen, L., Zhou, R., Wang, X., Song, L., Huang, S., Wang, G., and Xia, B. (2014). Increased proportions of bifidobacterium and the lactobacillus group and loss of butyrate-producing bacteria in inflammatory bowel disease. *Journal of clinical microbiology*, 52(2):398–406.

[198] Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature medicine*, 25(4):679.

[199] Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46.

[200] Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108.

[201] Wu, J.-Y., Wu, H., Jin, Y., Wei, J., Sha, D., Prentice, H., Lee, H.-H., Lin, C.-H., Lee, Y.-H., and Yang, L.-L. (2009). Mechanism of neuroprotective function of taurine. In *Taurine 7*, pages 169–179. Springer.

[202] Yadav, H., Lee, J.-H., Lloyd, J., Walter, P., and Rane, S. G. (2013). Beneficial metabolic effects of a probiotic via butyrate-induced glp-1 hormone secretion. *Journal of biological chemistry*, 288(35):25088–25097.

[203] Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., et al. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222.

[204] Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094.

[205] Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology*, 10(11).

[206] Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., and Lovley, D. R. (2011). Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *The ISME journal*, 5(2):305.

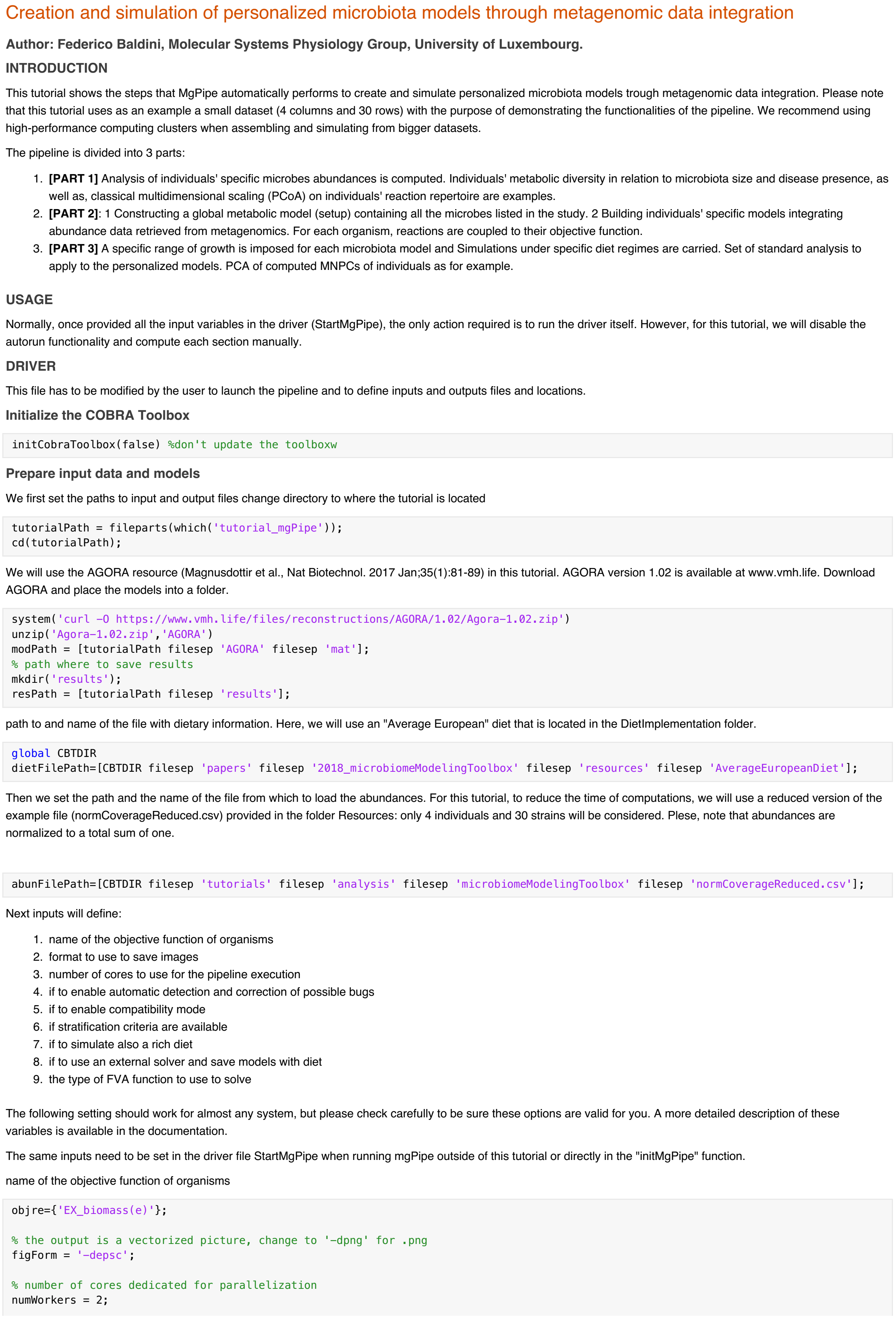
[207] Zomorrodi, A. R., Islam, M. M., and Maranas, C. D. (2014). d-optcom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS synthetic biology*, 3(4):247–257.

[208] Zomorrodi, A. R. and Maranas, C. D. (2012). Optcom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS computational biology*, 8(2):e1002363.

Appendix A

Supplementary material for Chapter 2

A.1 Microbiome Modeling Toolbox tutorials



Creation and simulation of personalized microbiota models through metagenomic data integration

Author: Federico Baldini, Molecular Systems Physiology Group, University of Luxembourg.

INTRODUCTION

This tutorial shows the steps that MgPipe automatically performs to create and simulate personalized microbiota models trough metagenomic data integration. Please note that this tutorial uses as an example a small dataset (4 columns and 30 rows) with the purpose of demonstrating the functionalities of the pipeline. We recommend using high-performance computing clusters when assembling and simulating from bigger datasets.

The pipeline is divided into 3 parts:

1. **[PART 1]** Analysis of individuals' specific microbes abundances is computed. Individuals' metabolic diversity in relation to microbiota size and disease presence, as well as, classical multidimensional scaling (PCoA) on individuals' reaction repertoire are examples.
2. **[PART 2]**: 1 Constructing a global metabolic model (setup) containing all the microbes listed in the study. 2 Building individuals' specific models integrating abundance data retrieved from metagenomics. For each organism, reactions are coupled to their objective function.
3. **[PART 3]** A specific range of growth is imposed for each microbiota model and Simulations under specific diet regimes are carried. Set of standard analysis to apply to the personalized models. PCA of computed MNPCs of individuals as for example.

USAGE

Normally, once provided all the input variables in the driver (StartMgPipe), the only action required is to run the driver itself. However, for this tutorial, we will disable the autorun functionality and compute each section manually.

DRIVER

This file has to be modified by the user to launch the pipeline and to define inputs and outputs files and locations.

Initialize the COBRA Toolbox

```
initCobraToolbox(false) %don't update the toolboxw
```

Prepare input data and models

We first set the paths to input and output files change directory to where the tutorial is located

```
tutorialPath = fileparts(which('tutorial_mgPipe'));
cd(tutorialPath);
```

We will use the AGORA resource (Magnusdottir et al., Nat Biotechnol. 2017 Jan;35(1):81-89) in this tutorial. AGORA version 1.02 is available at www.vmh.life. Download AGORA and place the models into a folder.

```
system('curl -O https://www.vmh.life/files/reconstructions/AGORA/1.02/Agora-1.02.zip')
unzip('Agora-1.02.zip','AGORA')
modPath = [tutorialPath filesep 'AGORA' filesep 'mat'];
% path where to save results
mkdir('results');
resPath = [tutorialPath filesep 'results'];
```

path to and name of the file with dietary information. Here, we will use an "Average European" diet that is located in the DietImplementation folder.

```
global CBTDIR
dietFilePath=[CBTDIR filesep 'papers' filesep '2018_microbiomeModelingToolbox' filesep 'resources' filesep 'AverageEuropeanDiet'];
```

Then we set the path and the name of the file from which to load the abundances. For this tutorial, to reduce the time of computations, we will use a reduced version of the example file (normCoverageReduced.csv) provided in the folder Resources: only 4 individuals and 30 strains will be considered. Plese, note that abundances are normalized to a total sum of one.

```
abunFilePath=[CBTDIR filesep 'tutorials' filesep 'analysis' filesep 'microbiomeModelingToolbox' filesep 'normCoverageReduced.csv'];
```

Next inputs will define:

1. name of the objective function of organisms
2. format to use to save images
3. number of cores to use for the pipeline execution
4. if to enable automatic detection and correction of possible bugs
5. if to enable compatibility mode
6. if stratification criteria are available
7. if to simulate also a rich diet
8. if to use an external solver and save models with diet
9. the type of FVA function to use to solve

The following setting should work for almost any system, but please check carefully to be sure these options are valid for you. A more detailed description of these variables is available in the documentation.

The same inputs need to be set in the driver file StartMgPipe when running mgPipe outside of this tutorial or directly in the "initMgPipe" function.

name of the objective function of organisms

```
objre={'EX_biomass(e)'};

% the output is a vectorized picture, change to '-dpng' for .png
figForm = '-depsc';

% number of cores dedicated for parallelization
numWorkers = 2;
```

```
% autofix for names mismatch
autoFix = true;

% if outputs in open formats should be produced for each section
compMod = false;

% if documentation (.csv) on stratification criteria is available
indInfoFilePath='none';

% to enable also rich diet simulations
rDiet = false;

% if to use an external solver and save models with diet
extSolve = false;

% the type of FVA function to use to solve
fvaType = true;

% to turn off the autorun to be able to manually execute each part of the pipeline
autorun = false;

[init,modPath,~,resPath,dietFilePath,abunFilePath,indInfoFilePath,objre,figForm,numWorkers,autoFix,compMod,rDiet,extSolve,fvaType,autorun]=
```

PIPELINE: [PART 1]

The number of organisms, their names, the number of samples and their identifiers are automatically detected from the input file.

```
[patNumb, sampName, strains] = getIndividualSizeName(abunFilePath)
```

Now we detect from the content of the results folder if PART1 was already computed: if the associated file is already present in the results folder its execution is skipped else its execution starts

```
[mapP] = detectOutput(resPath, 'mapInfo.mat');

if ~isempty(mapP)
    s= 'mapping file found: loading from resPath and skipping [PART1] analysis';
    disp(s)
    load(strcat(resPath,'mapInfo.mat'))
end
```

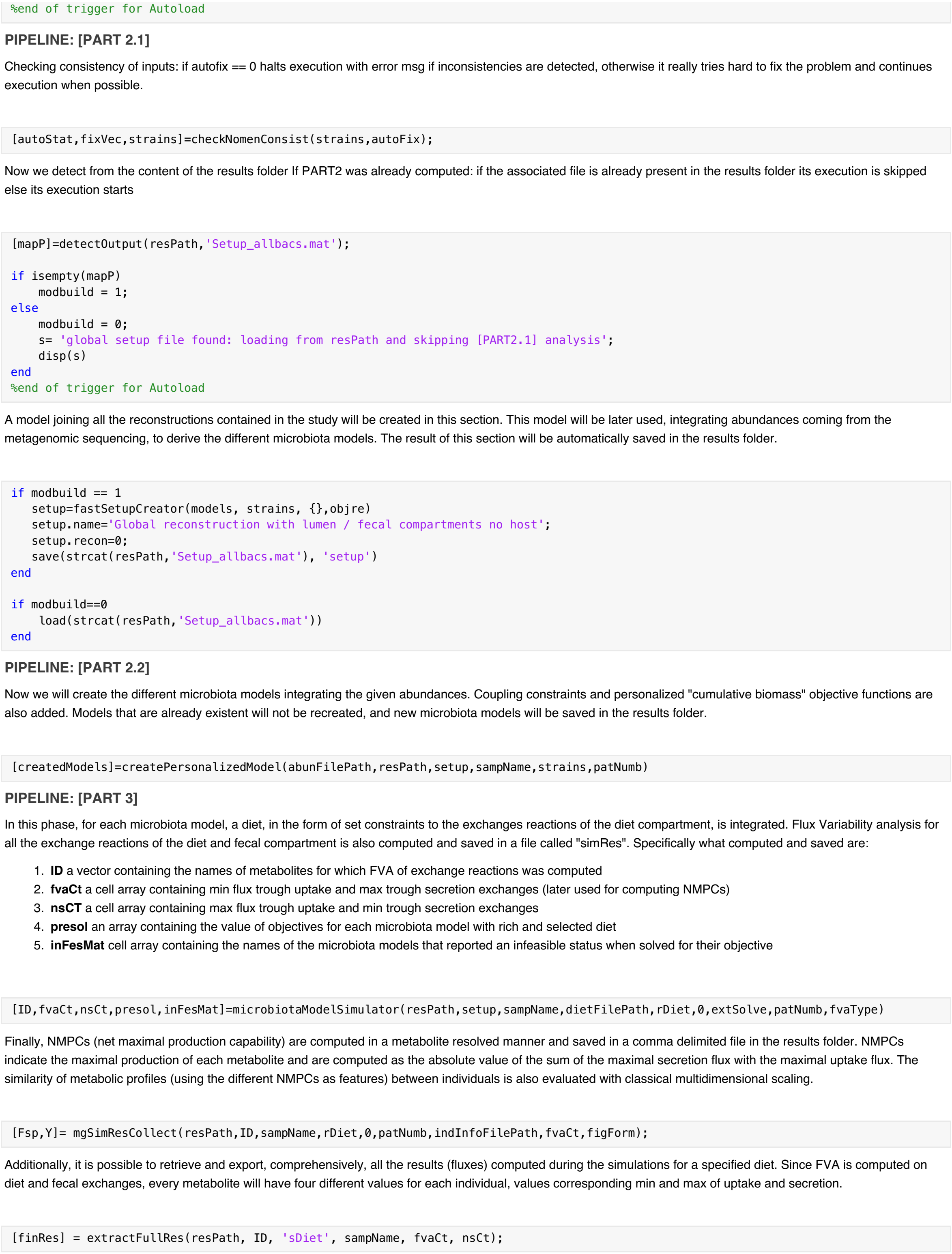
In case PART 1 was not computed we will compute it now. We will first load the models and create a cell array containing them. This cell array will be used as input by many functions in the pipeline. Any possible constraint from each model reactions will be removed. Moreover we will run and subsequentially plot the results of some analysis that are computed. The main outputs are:

1. **Metabolic diversity** The number of mapped organisms for each individual compared to the total number of unique reactions (extrapolated by the number of reactions of each organism).Please, note that bigger circles with a number inside represent overlapping individuals for metabolic diversity.
2. **Classical multidimensional scaling of each individual reactions repertoire**

Other outputs computed during this phase are saved together with the previous ones into the **.mat** file called **mapInfo.mat**. If the **compMod** option is enabled (disabled here and by default in the **mgPipe** pipeline) these results are outputted as different **.csv** files. For simplicity reasons we will not discuss these additional outputs in this tutorial: for a description of them, please refer to the documentation.

```
[mapP] = detectOutput(resPath,'mapInfo.mat')
```

```
if isempty(mapP)
    % Loading models
    models = loadUncModels(modPath,strains,objre);

    % Computing genetic information
    [reac,micRea,binOrg,patOrg,reacPat,reacNumb,reacSet,reacTab,reacAbun,reacNumber] = getMappingInfo(models,abunFilePath,patNumb);
    writetable(cell2table(reacAbun,'VariableNames',['Reactions';sampName]'),strcat(resPath,'reactions.csv'));

    % Plotting genetic information
    [PCoA] = plotMappingInfo(resPath,patOrg,reacPat,reacTab,reacNumber,indInfoFilePath,figForm,sampName,strains);

    if compMod == 1
       mkdir(strcat(resPath,'compfile'))
       writetable([array2table(reac),array2table(reacTab,'VariableNames',sampName')],[resPath 'compfile' filesep 'ReacTab.csv'])
       writetable(cell2table(reacSet,'VariableNames',sampName'),[resPath 'compfile' filesep 'reacSet.csv'])
       writetable([array2table(strains),array2table(reacPat,'VariableNames',sampName')],[resPath 'compfile' filesep 'ReacPat.csv'])
       csvwrite(strcat(resPath,'compfile/PCoA_tab.csv'),PCoA)
    end

    %Save all the created variables

    %Create tables and save all the created variables
    reacTab=[array2table(reac),array2table(reacTab,'VariableNames',sampName')],[resPath 'compfile' filesep 'ReacTab.csv'];
    reacSet=cell2table(reacSet,'VariableNames',sampName');
    reacPat=[array2table(strains),array2table(reacPat,'VariableNames',sampName')];

    save(strcat(resPath,'mapInfo.mat'))
end
```



```
%end of trigger for Autoload
```

PIPELINE: [PART 2.1]

Checking consistency of inputs: if autofix == 0 halts execution with error msg if inconsistencies are detected, otherwise it really tries hard to fix the problem and continues execution when possible.

```
[autoStat,fixVec,strains]=checkNomenConsist(strains,autoFix);
```

Now we detect from the content of the results folder If PART2 was already computed: if the associated file is already present in the results folder its execution is skipped else its execution starts

```
[mapP]=detectOutput(resPath,'Setup_allbacs.mat');

if isempty(mapP)
    modbuild = 1;
else
    modbuild = 0;
    s= 'global setup file found: loading from resPath and skipping [PART2.1] analysis';
    disp(s)
end
%end of trigger for Autoload
```

A model joining all the reconstructions contained in the study will be created in this section. This model will be later used, integrating abundances coming from the metagenomic sequencing, to derive the different microbiota models. The result of this section will be automatically saved in the results folder.

```
if modbuild == 1
   setup=fastSetupCreator(models, strains, {},objre)
   setup.name='Global reconstruction with lumen / fecal compartments no host';
   setup.recon=0;
   save(strcat(resPath,'Setup_allbacs.mat'), 'setup')
end

if modbuild==0
    load(strcat(resPath,'Setup_allbacs.mat'))
end
```

PIPELINE: [PART 2.2]

Now we will create the different microbiota models integrating the given abundances. Coupling constraints and personalized "cumulative biomass" objective functions are also added. Models that are already existent will not be recreated, and new microbiota models will be saved in the results folder.

```
[createdModels]=createPersonalizedModel(abunFilePath,resPath,setup,sampName,strains,patNumb)
```

PIPELINE: [PART 3]

In this phase, for each microbiota model, a diet, in the form of set constraints to the exchanges reactions of the diet compartment, is integrated. Flux Variability analysis for all the exchange reactions of the diet and fecal compartment is also computed and saved in a file called "simRes". Specifically what computed and saved are:

1. **ID** a vector containing the names of metabolites for which FVA of exchange reactions was computed
2. **fvaCt** a cell array containing min flux trough uptake and max trough secretion exchanges (later used for computing NMPCs)
3. **nsCT** a cell array containing max flux trough uptake and min trough secretion exchanges
4. **presol** an array containing the value of objectives for each microbiota model with rich and selected diet
5. **inFesMat** cell array containing the names of the microbiota models that reported an infeasible status when solved for their objective

```
[ID,fvaCt,nsCt,presol,inFesMat]=microbiotaModelSimulator(resPath,setup,sampName,dietFilePath,rDiet,0,extSolve,patNumb,fvaType)
```

Finally, NMPCs (net maximal production capability) are computed in a metabolite resolved manner and saved in a comma delimited file in the results folder. NMPCs indicate the maximal production of each metabolite and are computed as the absolute value of the sum of the maximal secretion flux with the maximal uptake flux. The similarity of metabolic profiles (using the different NMPCs as features) between individuals is also evaluated with classical multidimensional scaling.

```
[Fsp,Y]= mgSimResCollect(resPath,ID,sampName,rDiet,0,patNumb,indInfoFilePath,fvaCt,figForm);
```

Additionally, it is possible to retrieve and export, comprehensively, all the results (fluxes) computed during the simulations for a specified diet. Since FVA is computed on diet and fecal exchanges, every metabolite will have four different values for each individual, values corresponding min and max of uptake and secretion.

```
[finRes] = extractFullRes(resPath, ID, 'sDiet', sampName, fvaCt, nsCt);
```

Computation and analysis of microbe-microbe metabolic interactions

Note: This tutorial is a draft and needs completion. Contributions welcome!

Author: Almut Heinken, Molecular Systems Physiology Group, University of Luxembourg.

This tutorial demonstrates how to join a given list of microbial COBRA models in all possible combinations and compute the metabolic

interactions between the microbes depending on the implemented diet. Moreover, the tradeoff between the growth of different joined microbes is computed. The tutorial can be adapted to any number of AGORA models and dietary conditions analyzed.

We will use the AGORA resource (Magnusdottir et al., Nat Biotechnol. 2017 Jan;35(1):81-89) in this tutorial. Please download AGORA version 1.02 from https://vmh.life and place the models into a folder.

Define the path to the folder where you stored the AGORA models.

```
modelPath='YOUR_PATH_TO_AGORA/';
```

Import a file with information on the AGORA organisms including reconstruction names and taxonomy.

```
[~,infoFile,~]=xlsread('AGORA_infoFile.xlsx');
```

Initialize the COBRA Toolbox.

```
initCobraToolbox
```

Creation of pairwise models

For the sake of this tutorial, we will use ten random AGORA reconstructions from the info file.

```
modelList = infoFile(randi([2 length(infoFile)],1,10),1);
```

Uncomment the following line to join all AGORA reconstructions in all combinations. NOTE: this is very time-consuming due to the large number of model combinations analyzed.

```
modelList=infoFile(2:end,1);
```

You may also enter a custom selection of AGORA reconstructions as a cell array named modelList.

Load the AGORA reconstructions to be joined.

```
for i=1:size(modelList,1)
    load(strcat(modelPath,modelList{i,1},'.mat'));
    % make sure the fields in the reconstruction structure are in the correct
    % format, incorrect format causes errors when joining the models
    model = convertOldStyleModel(model);
    inputModels{i,1}=model;
end
```

Let us define some parameters for joining the models. Set the coupling factor c, which defined how the flux through all reactions in a model is coupled to the flux through its biomass reaction. Allowed flux span through each reaction= -(c * flux(biomass)) to +(c * flux(biomass)).

```
c = 400;
```

Set the threshold u, which defines the flux through each reaction that is allowed if flux through the biomass reaction is zero.

```
u = 0;
```

Define whether or not genes from the models are merged and kept in the joined models. If set to true the joining is more time-consuming.

```
mergeGenes = false;
```

Define the number workers for parallel pool to allow parallel computing.Recommended if a large number of microbe models is computed. Set to zero if parallel computing is not available.

```
numWorkers = 4;
```

Join the models in all possible combinations.

```
[pairedModels,pairedModelInfo]=joinModelsPairwiseFromList(modelList,inputModels,'c',c,'u',u,'mergeGenesFlag',mergeGenes,'numWorkers',numWork
```

Computation of pairwise interactions

The interactions between all microbes joined in the first step will be simulated on given dietary conditions. Here, we will use four dietary conditions used in Magnusdottir et al., Nat Biotechnol. 2017.: Western Diet without oxygen, Western Diet with oxygen, High fiber diet without oxygen, and High fiber diet with oxygen. Let us define the input parameters for the simulation of pairwise interactions.

```
% Name the four dietary conditions that will be simulated.
conditions = {'WesternDiet_NoOxygen','WesternDiet_WithOxygen','HighFiberDiet_NoOxygen','HighFiberDiet_WithOxygen'};
```

Define the corresponding constraints to implement for each diet. The input file needs to be a string array.

```
dietConstraints{1}={'EX_fru[u]','-0.14899','1000';'EX_glc_D[u]','-0.14899','1000';'EX_gal[u]','-0.14899','1000';'EX_man[u]','-0.14899','100
dietConstraints{2}={'EX_fru[u]','-0.14899','1000';'EX_glc_D[u]','-0.14899','1000';'EX_gal[u]','-0.14899','1000';'EX_man[u]','-0.14899','100
dietConstraints{3}={'EX_fru[u]','-0.03947','1000';'EX_glc_D[u]','-0.03947','1000';'EX_gal[u]','-0.03947','1000';'EX_man[u]','-0.03947','100
dietConstraints{4}={'EX_fru[u]','-0.03947','1000';'EX_glc_D[u]','-0.03947','1000';'EX_gal[u]','-0.03947','1000';'EX_man[u]','-0.03947','100
```

NOTE: if you design your own diet, make sure that exchange reaction abbreviations correspond to the lumen exchanges in the joint models ('EX_compound[u]').

Define what counts as significant difference between single growth of the microbes and growth when joined with another microbe-here we choose 10%.

```
sigD = 0.1;
```

Simulate the pairwise interactions on the four dietary conditions.

```
for i = 1:length(conditions)
    % assign dietary constraints
    [pairwiseInteractions]=simulatePairwiseInteractions(pairedModels,pairedModelInfo,'inputDiet',dietConstraints{i},'sigD',sigD,'saveSoluti
Interactions.(conditions{i})=pairwiseInteractions;
end
```

Analysis of computed pairwise interactions

The computed microbe-microbe interactions will be plotted by type and analyzed in the context of the taxonomy of the joined strains. There are six possible types of interactions total that can result in increased growth (+), no change in growth (=) or decreased growth (-) compared with the single condition for each joined microbe.

- Competition (-/-)
- Parasitism   (+/-)
- Amensalism   (=/-)
- Neutralism   (=/=)
- Commensalism (+/=)
- Mutualism   (+/+)

This results in nine different outcomes total from the perspective of each joined microbe.

Plot the percentage of interactions computed.

```
figure('rend','painters','pos',[10 10 900 600])
typesIA=unique(pairwiseInteractions(2:end,10));
for i = 1:length(conditions)
    pairwiseInteractions=Interactions.(conditions{i});
    listIA=pairwiseInteractions(2:end,10);
    for j=1:length(typesIA)
        dat(j)=sum(strcmp(listIA(:),typesIA{j}));
    end
    subplot(2,2,i)
    pie(dat)
    set(gca,'FontSize',10)
    h=title(conditions{i});
    set(h,'interpreter','none')
    title(conditions{i})
end
legend1=legend(typesIA);
set(legend1,'Position',[0.42 0.45 0.2 0.2],'FontSize',12)
suptitle('Percentage of computed pairwise interactions')
```

Next, the percentage of interactions will be calculated on different taxon levels (genus, family, order, class, phylum) using the taxon information contained in AGORA_infoFile.xlsx. Here, the interactions will be considered from the perspective of each joined microbe resulting in nine possible interactions total.

Calculate the percentage of interactions predicted for each taxon included in the list of microbes analyzed.

```
for i = 1:length(conditions)
    pairwiseInteractions=Interactions.(conditions{i});
    [InteractionsByTaxon]=calculateInteractionsByTaxon(pairwiseInteractions,infoFile);
    TaxonSummaries.(conditions{i})=InteractionsByTaxon;
end
```

Combine the four conditions into one structure.

```
InteractionsByTaxonCombined=struct;
for i = 1:length(conditions)
    InteractionsByTaxon=TaxonSummaries.(conditions{i});
    taxLevels=fieldnames(InteractionsByTaxon);
    if i==1
        for j=1:length(taxLevels)
            InteractionsByTaxonCombined.(taxLevels{j})=InteractionsByTaxon.(taxLevels{j});
            InteractionsByTaxonCombined.(taxLevels{j})(2:end,1)=strcat(InteractionsByTaxonCombined.(taxLevels{j})(2:end,1),'_',conditions{i}
        end
    else
        for j=1:length(taxLevels)
            rowLength=size(InteractionsByTaxonCombined.(taxLevels{j}),1);
            InteractionsByTaxonCombined.(taxLevels{j})=[InteractionsByTaxonCombined.(taxLevels{j});InteractionsByTaxon.(taxLevels{j})(2:end,
            InteractionsByTaxonCombined.(taxLevels{j})(rowLength+1:end,1)=strcat(InteractionsByTaxonCombined.(taxLevels{j})(rowLength+1:end,
        end
    end
end
```

Let us plot the distributions of interactions for all dietary conditions combined on the level of genera as an example. Note: The xticklabels/yticklabels function is only available in MATLAB R2016b or newer. Older versions of MATLAB will be unable to display the labels.

```
for i=5
    xlabels=InteractionsByTaxonCombined.(taxLevels{i})(1,2:end);
    ylabels=InteractionsByTaxonCombined.(taxLevels{i})(2:end,1);
    data=string(InteractionsByTaxonCombined.(taxLevels{i})(2:end,2:end));
    data=str2double(data);
    figure;
    imagesc(data)
    colormap('hot')
    colorbar
    set(gca,'xtick',1:length(xlabels));
    xticklabels(xlabels);
    set(gca,'ytick',1:length(ylabels));
    yticklabels(ylabels);
    xtickangle(90)
    set(gca,'TickLabelInterpreter', 'none');
    title(taxLevels{i})
end
```

Pareto optimality analysis

Another way to analyze the metabolic interactions between two microbes in Pareto optimality analysis. In this method, the tradeoff between two competing objectives (e.g., the biomasses of two joined microbes) is calculated. The resulting Pareto frontier depicts all possible outcomes of co-growth between the two microbes under the given constraints.

Let us compute the Pareto frontier for five randomly chosen pairs from the list of AGORA models.

```
modelInd = randi([2 length(infoFile)],2,5);
```

The Pareto frontier will be computed on the Western diet without oxygen.

```
dietConstraints{1}={'EX_fru[u]','-0.14899','1000';'EX_glc_D[u]','-0.14899','1000';'EX_gal[u]','-0.14899','1000';'EX_man[u]','-0.14899','100
```

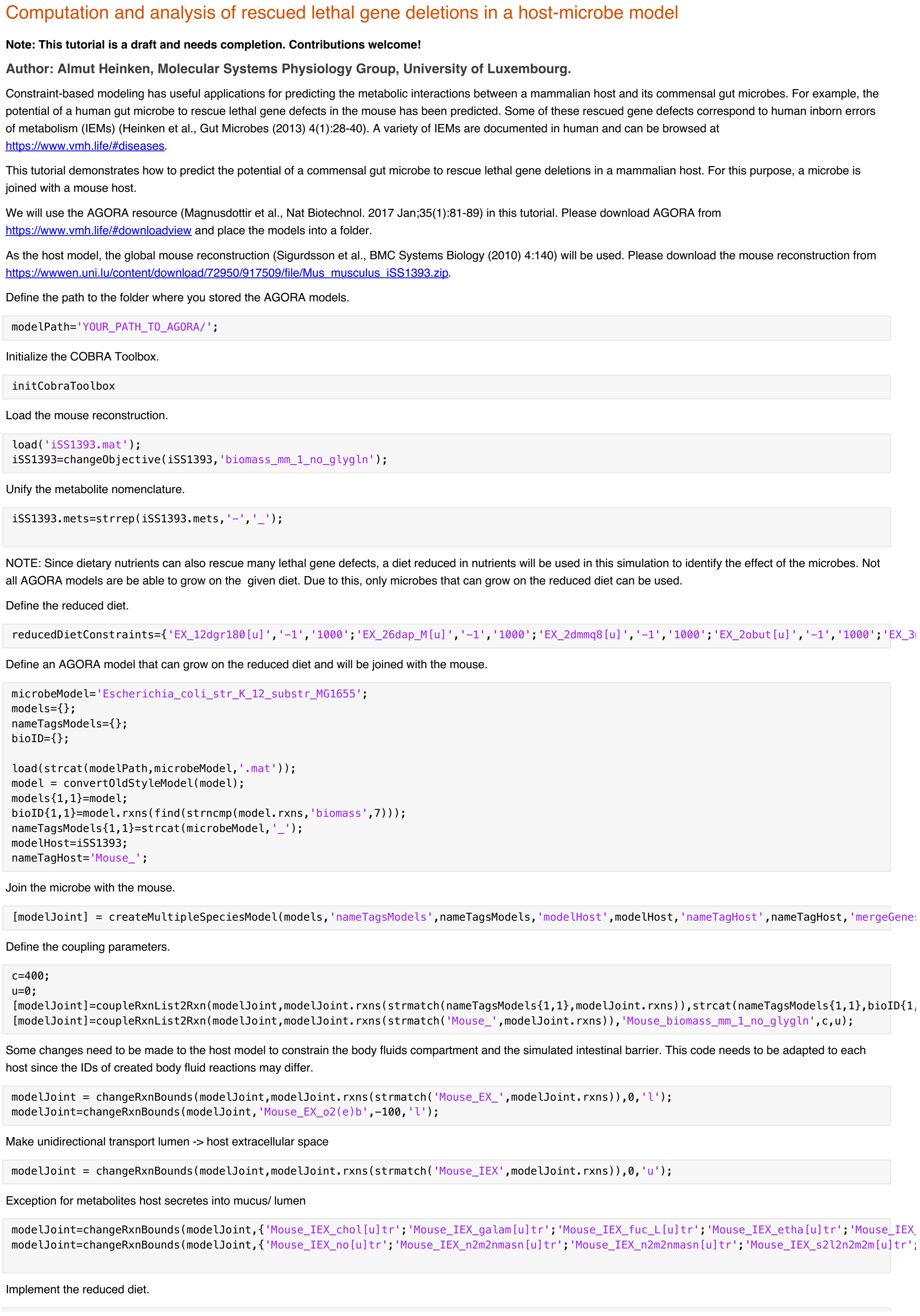
By default, the points of the frontier will be generated at steps of 0.001.

```
dinc=0.001;
```

Perform the Pareto optimality analysis for the five pairs. The shape of the computed Pareto frontier, which represents all possible optimal solutions of simultaneously optimized growth, depends on the metabolic networks of the two joined microbes.

```
for i=1:size(modelInd,2)
    load(strcat(modelPath,infoFile{modelInd(1,i),1},'.mat'));
    model = convertOldStyleModel(model);
    models{1,1}=model;
    bioID{1,1}=model.rxns(find(strncmp(model.rxns,'biomass',7)));
    nameTagsModels{1,1}=strcat(infoFile{modelInd(1,i),1},'_');
    load(strcat(modelPath,infoFile{modelInd(2,i),1},'.mat'));
    model = convertOldStyleModel(model);
    models{2,1}=model;
    nameTagsModels{2,1}=strcat(infoFile{modelInd(2,i),1},'_');
    bioID{2,1}=model.rxns(find(strncmp(model.rxns,'biomass',7)));
    [pairedModel] = createMultipleSpeciesModel(models,'nameTagsModels',nameTagsModels);
    [pairedModel]=coupleRxnList2Rxn(pairedModel,pairedModel.rxns(strmatch(nameTagsModels{1,1},pairedModel.rxns)),strcat(infoFile{modelInd(1,
    [pairedModel]=coupleRxnList2Rxn(pairedModel,pairedModel.rxns(strmatch(nameTagsModels{2,1},pairedModel.rxns)),strcat(infoFile{modelInd(2,
    pairedModel=useDiet(pairedModel,dietConstraints{1});
    [ParetoFrontier] = computeParetoOptimality(pairedModel,strcat(infoFile{modelInd(1,i),1},'_',bioID{1,1}),strcat(infoFile{modelInd(2,i),1}
end
```

Can you interpret the shapes of the five Pareto frontiers that were computed? Are there microbe pairs that are always competing with each other? Are there pairs in which one microbe can benefit the other at certain points in the curve and vice versa?



Computation and analysis of rescued lethal gene deletions in a host-microbe model

Note: This tutorial is a draft and needs completion. Contributions welcome!

Author: Almut Heinken, Molecular Systems Physiology Group, University of Luxembourg.

Constraint-based modeling has useful applications for predicting the metabolic interactions between a mammalian host and its commensal gut microbes. For example, the potential of a human gut microbe to rescue lethal gene defects in the mouse has been predicted. Some of these rescued gene defects correspond to human inborn errors of metabolism (IEMs) (Heinken et al., Gut Microbes (2013) 4(1):28-40). A variety of IEMs are documented in human and can be browsed at https://www.vmh.life/#diseases.

This tutorial demonstrates how to predict the potential of a commensal gut microbe to rescue lethal gene deletions in a mammalian host. For this purpose, a microbe is joined with a mouse host.

We will use the AGORA resource (Magnusdottir et al., Nat Biotechnol. 2017 Jan;35(1):81-89) in this tutorial. Please download AGORA from https://www.vmh.life/#downloadview and place the models into a folder.

As the host model, the global mouse reconstruction (Sigurdsson et al., BMC Systems Biology (2010) 4:140) will be used. Please download the mouse reconstruction from https://wwwen.uni.lu/content/download/72950/917509/file/Mus_musculus_iSS1393.zip.

Define the path to the folder where you stored the AGORA models.

```
modelPath='YOUR_PATH_TO_AGORA/';
```

Initialize the COBRA Toolbox.

```
initCobraToolbox
```

Load the mouse reconstruction.

```
load('iSS1393.mat');
iSS1393=changeObjective(iSS1393,'biomass_mm_1_no_glygln');
```

Unify the metabolite nomenclature.

```
iSS1393.mets=strrep(iSS1393.mets,'-','_');
```

NOTE: Since dietary nutrients can also rescue many lethal gene defects, a diet reduced in nutrients will be used in this simulation to identify the effect of the microbes. Not all AGORA models are be able to grow on the given diet. Due to this, only microbes that can grow on the reduced diet can be used.

Define the reduced diet.

```
reducedDietConstraints={'EX_12dgr180[u]','-1','1000';'EX_26dap_M[u]','-1','1000';'EX_2dmmq8[u]','-1','1000';'EX_2obut[u]','-1','1000';'EX_3
```

Define an AGORA model that can grow on the reduced diet and will be joined with the mouse.

```
microbeModel='Escherichia_coli_str_K_12_substr_MG1655';
models={};
nameTagsModels={};
bioID={};

load(strcat(modelPath,microbeModel,'.mat'));
model = convertOldStyleModel(model);
models{1,1}=model;
bioID{1,1}=model.rxns(find(strncmp(model.rxns,'biomass',7)));
nameTagsModels{1,1}=strcat(microbeModel,'_');
modelHost=iSS1393;
nameTagHost='Mouse_';
```

Join the microbe with the mouse.

```
[modelJoint] = createMultipleSpeciesModel(models,'nameTagsModels',nameTagsModels,'modelHost',modelHost,'nameTagHost',nameTagHost,'mergeGene
```

Define the coupling parameters.

```
c=400;
u=0;
[modelJoint]=coupleRxnList2Rxn(modelJoint,modelJoint.rxns(strmatch(nameTagsModels{1,1},modelJoint.rxns)),strcat(nameTagsModels{1,1},bioID{1,
[modelJoint]=coupleRxnList2Rxn(modelJoint,modelJoint.rxns(strmatch('Mouse_',modelJoint.rxns)),'Mouse_biomass_mm_1_no_glygln',c,u);
```

Some changes need to be made to the host model to constrain the body fluids compartment and the simulated intestinal barrier. This code needs to be adapted to each host since the IDs of created body fluid reactions may differ.

```
modelJoint = changeRxnBounds(modelJoint,modelJoint.rxns(strmatch('Mouse_EX_',modelJoint.rxns)),0,'l');
modelJoint=changeRxnBounds(modelJoint,'Mouse_EX_o2(e)b',-100,'l');
```

Make unidirectional transport lumen -> host extracellular space

```
modelJoint = changeRxnBounds(modelJoint,modelJoint.rxns(strmatch('Mouse_IEX',modelJoint.rxns)),0,'u');
```

Exception for metabolites host secretes into mucus/ lumen

```
modelJoint=changeRxnBounds(modelJoint,{'Mouse_IEX_chol[u]tr';'Mouse_IEX_galam[u]tr';'Mouse_IEX_fuc_L[u]tr';'Mouse_IEX_etha[u]tr';'Mouse_IEX_
modelJoint=changeRxnBounds(modelJoint,{'Mouse_IEX_no[u]tr';'Mouse_IEX_n2m2nmasn[u]tr';'Mouse_IEX_n2m2nmasn[u]tr';'Mouse_IEX_s2l2n2m2m[u]tr';
```

Implement the reduced diet.

```
modelJoint=useDiet(modelJoint,reducedDietConstraints);
```

Run the prediction of rescued genes. This will take some time.

```
[OptSolKO,OptSolWT,OptSolRatio,RescuedGenes,fluxesKO]=computeRescuedGenes('modelJoint',modelJoint,'Rxn1','Mouse_biomass_mm_1_no_glygln','Rx
```

Show the mouse genes that caused a lethal phenotype when deleted in germfree mouse but not in presence of the microbe:

```
RescuedGenes.Mouse_biomass_mm_1_no_glygln.RescuedLethalGenes
```

The gene identifiers are NCBI Gene IDs and can be looked up to find the corresponding human genes and associated inborn errors of metabolism (IEMs). The reactions associated with the IEMs can subsequently be browsed at https://www.vmh.life/#diseases. For example, the gene 22247.1 encodes UMP synthase and its deletion can be resuced by the presence of E. coli. The corresponding IEM in human is orotic aciduria (https://www.vmh.life/#disease/OROA).

We will now identify the mechanisms of rescued KO phenotypes. Metabolites secreted by each species into the lumen may be taken up by the joined species and provide the metabolites that are essential due to the gene defect. To find the lumen exchange reactions of E. coli:

```
microbeExchanges=find(strncmp(modelJoint.rxns,strcat(microbeModel,'_IEX'),length(strcat(microbeModel,'_IEX'))));
```

Now, let us find out which of the metabolites secreted by E.coli was essential for rescuing the defect in mouse UMP synthase.

```
[model,hasEffect,constrRxnNames,deletedGenes] = deleteModelGenes(modelHost,'22247.1');
constrRxnNames = strcat(nameTagHost,constrRxnNames);
modelJoint=changeRxnBounds(modelJoint,constrRxnNames,0,'b');
modelJoint=changeObjective(modelJoint,'Mouse_biomass_mm_1_no_glygln');
modelJoint=changeRxnBounds(modelJoint,strcat(nameTagHost,'ATPM'),0,'l');
modelJoint=changeRxnBounds(modelJoint,strcat(nameTagsModels{1},'DM_atp_c_'),0,'l');
```

To print out the E. coli exchange that had to carry flux to rescue the orotic aciduria-like mouse phenotype, use the following code:

```
for i=1:length(microbeExchanges)
    if isempty(strfind(modelJoint.rxns{microbeExchanges(i)},'biomass'))
        % prevent secretion flux through the exchanges one by one while predicting mouse biomass
        modelJointDel=changeRxnBounds(modelJoint,modelJoint.rxns{microbeExchanges(i)},0,'u');
        solution=solveCobraLP(modelJointDel);
        if solution.obj<0.0000000001
            fprintf('%s \n',modelJoint.rxns{microbeExchanges(i)},' is essential for rescuing orotic aciduria.')
        end
    end
end
```

Can you explain what you observe? You can look up the metabolite ID of the respective exchange at https://www.vmh.life. Hint: Check the description of orotic aciduria at https://www.vmh.life/#disease/OROA. Follow the external link to OMIM (Online Mendelian Inheritance in Man) to find more information.

Appendix B

Supplementary material for Chapter 3

B.1 mgPipe tutorial

mgPipe: from microbial community sequencing data to personalized microbiota metabolic models creation and interrogation

Author: Federico Baldini, Molecular Systems Physiology Group, University of Luxembourg

INTRODUCTION

This tutorial shows the steps performed by mgPipe to retrieve reference-based microbial abundances from Illumina sequencing data of microbial communities and to create and interrogate personalized metabolic models of microbiota communities. mgPipe can be used for 16S rRNA gene sequencing data as well as whole-genome sequencing data. mgPipe is divided into three parts:

1. **[PART1]** Relative abundances with different taxonomic resolution (genus and species for 16S rRNA gene sequencing and strain for whole-genome sequencing) are retrieved from .fastQ files.
2. **[PART 2]** Creation and interrogation of personalized microbiota metabolic models integrating relative abundances retrieved in the first step
3. **[PART 3] (Optional)** Analysis of microbial abundances and modeling outputs.

For the sake of this tutorial, we will download some data from two different studies one of which on Crohn's Disease, namely PRJEB22832 and SRP057027 available at EBI metagenomics (https://www.ebi.ac.uk/metagenomics/) and respectively produced using 16S rRNA gene sequencing and whole-genome sequencing techniques. We will download only a limited number of observations (four samples) with the only purpose of demonstrating the functionalities of the pipeline. Please note that the purpose of this tutorial in solely to demonstrate the functionalities of mgPipe and no biological interpretation should be extrapolated by the processed samples and the relative results. Furthermore, we recommend using high-performance computing clusters when assembling and simulating from bigger datasets. Please, before proceeding to the next sessions, make sure that you completed the mgPipe installation installing all the required dependencies and creating the necessary files as explained in the "Installation" and "mgPipe activation" chapters of the documentation.

USAGE

Normally, once completed the installation procedure and provided all the input variables in the input file (Input.csv), the only action required is to run the script Runner.sh using the so-called "passive mode" of execution of the pipeline. However, for this tutorial, we will not run mgPipe in passive mode, and we will use the active mode running each section of Runner.sh manually. This tutorial will be divided into two main sections, one for 16S rRNA gene sequencing and one for whole-genome sequencing. However, the majority of the steps will be common so we advise you to read the full document as we will initially describe the protocol for 16S rRNA gene sequencing and report only the different steps for whole-genome sequencing.

16S rRNA GENE SEQUENCING

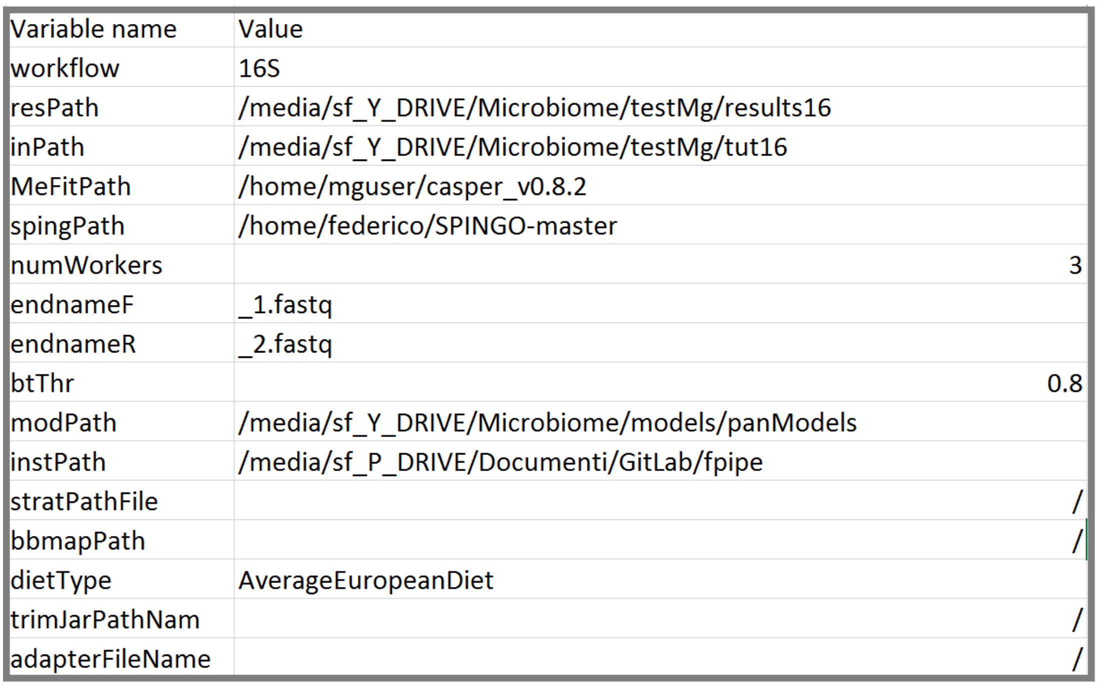
For this tutorial, we will download samples from a small dataset available at EBI metagenomics. In this case, the study PRJEB22832 under the accession ERP104539 (https://www.ebi.ac.uk/ena/data/view/PRJEB22832). As first step, please download and unzip all the sequences from the study (2 x 4 samples) and place them in a dedicated folder called, for example, "samples.

Now we are ready to proceed with **[PART 1]** of the pipeline. Go to the mgPipe installation folder (where you cloned the repository) and open the csv file called Input.csv. As explained in the "Inputs" chapter of the documentation, this file will contain several inputs aiming to define:

- type of microbial identification workflow: 16S or WGS/THR_WGS (whole genome sequencing)
- path to the directory containing results

- path to the directory containing the sequencing files (.fastQ)
- path to the directory containing MeFit files
- path to the directory containing SPINGO files
- number of cores dedicated for parallelization
- name end (characters after the special _ character) of the forward .fastQ input files
- name end (characters after the special _ character) of the reverse .fastQ input files
- bootstrap threshold for 16S classification (default 0.8)
- path to microbes metabolic models (AGORA strains or generated AGORA panSpecies)
- path of the cloned folder
- path and name of the file containing samples stratification information (optional)
- path to bbmap folder (for whole-genome sequencing, WGS)
- name of diet formulation to use (without file extension). Default is "AverageEuropeanDiet"
- path to and name (version) of downloaded Trimmomatic .jar file (for THR_WGS)
- name of Illumina adapter used (with file extension) (for THR_WGS)

Please, edit the file Input.csv to enter your specific inputs, save and close it. In our case, considering the folders we prepared, or desktop power and the names of the sequences, the file will look like the following

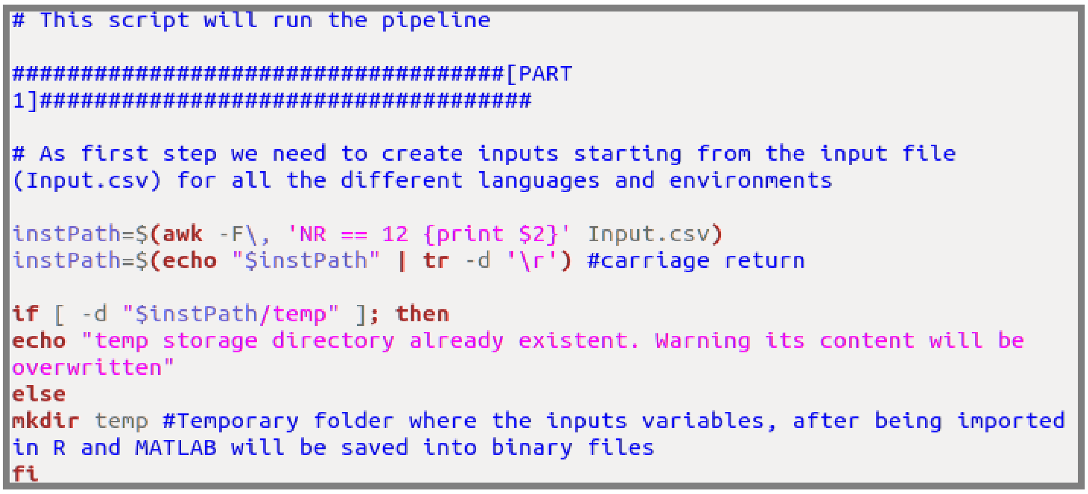


Variable name	Value
workflow	16S
resPath	/media/sf_Y_DRIVE/Microbiome/testMg/results16
inPath	/media/sf_Y_DRIVE/Microbiome/testMg/tut16
MeFitPath	/home/mguser/casper_v0.8.2
spingPath	/home/federico/SPINGO-master
numWorkers	3
endnameF	_1.fastq
endnameR	_2.fastq
btThr	0.8
modPath	/media/sf_Y_DRIVE/Microbiome/models/panModels
instPath	/media/sf_P_DRIVE/Documenti/GitLab/fpipe
stratPathFile	/
bbmapPath	/
dietType	AverageEuropeanDiet
trimJarPathNam	/
adapterFileName	/

Now we are all set to start executing mgPipe and we will do this opening and pasting the commands of the script called Runner.sh. Please, press Ctrl - Alt + T or double click on the icon to open the Terminal and browse to the folder where mgPipe was installed using the "cd" command.

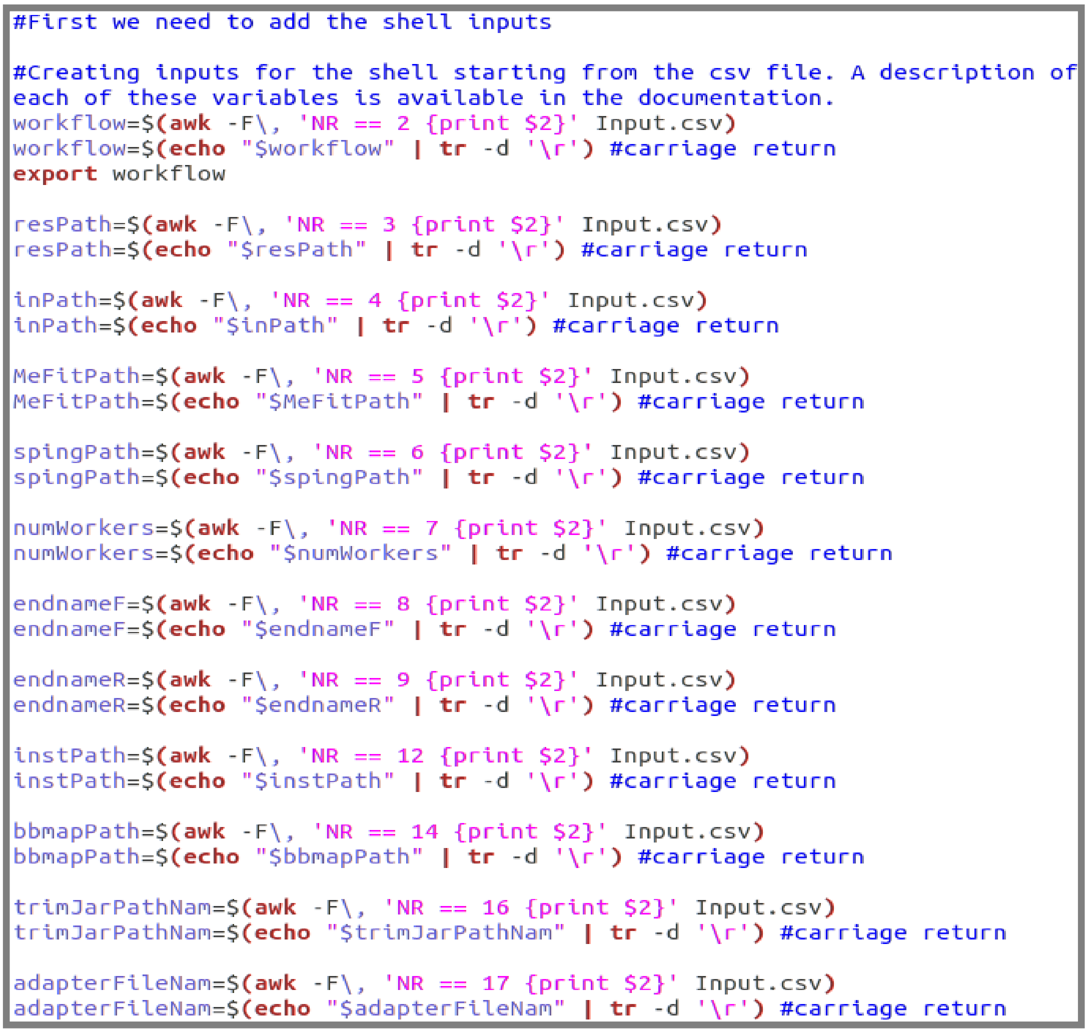
```
> cd /yourfileSystem/fpipe
```

The first command of the script Runner.sh will save the path of where the pipeline is (instPath) and will create a folder called "temp" where later on, the temporary folder where the inputs variables, after being imported in R and MATLAB will be saved into binary files. Execute:



```
# This script will run the pipeline

#####################################[PART
1]#####################################

# As first step we need to create inputs starting from the input file
(Input.csv) for all the different languages and environments

instPath=$(awk -F\, 'NR == 12 {print $2}' Input.csv)
instPath=$(echo "$instPath" | tr -d '\r') #carriage return

if [ -d "$instPath/temp" ]; then
echo "temp storage directory already existent. Warning its content will be
overwritten"
else
mkdir temp #Temporary folder where the inputs variables, after being imported
in R and MATLAB will be saved into binary files
fi
```

The next lines of code will be importing inputs for the terminal parsing them from Input.csv. A description of each of these variables is available in the documentation. Execute:



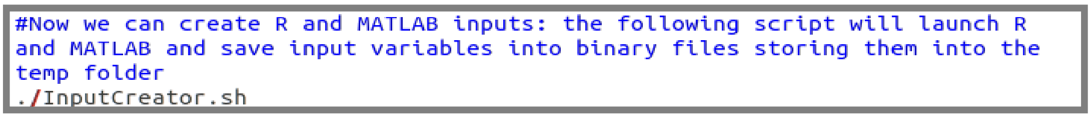
```
#First we need to add the shell inputs

#Creating inputs for the shell starting from the csv file. A description of
each of these variables is available in the documentation.
workflow=$(awk -F\, 'NR == 2 {print $2}' Input.csv)
workflow=$(echo "$workflow" | tr -d '\r') #carriage return
export workflow

resPath=$(awk -F\, 'NR == 3 {print $2}' Input.csv)
resPath=$(echo "$resPath" | tr -d '\r') #carriage return

inPath=$(awk -F\, 'NR == 4 {print $2}' Input.csv)
inPath=$(echo "$inPath" | tr -d '\r') #carriage return

MeFitPath=$(awk -F\, 'NR == 5 {print $2}' Input.csv)
MeFitPath=$(echo "$MeFitPath" | tr -d '\r') #carriage return

spingPath=$(awk -F\, 'NR == 6 {print $2}' Input.csv)
spingPath=$(echo "$spingPath" | tr -d '\r') #carriage return

numWorkers=$(awk -F\, 'NR == 7 {print $2}' Input.csv)
numWorkers=$(echo "$numWorkers" | tr -d '\r') #carriage return

endnameF=$(awk -F\, 'NR == 8 {print $2}' Input.csv)
endnameF=$(echo "$endnameF" | tr -d '\r') #carriage return

endnameR=$(awk -F\, 'NR == 9 {print $2}' Input.csv)
endnameR=$(echo "$endnameR" | tr -d '\r') #carriage return

instPath=$(awk -F\, 'NR == 12 {print $2}' Input.csv)
instPath=$(echo "$instPath" | tr -d '\r') #carriage return

bbmapPath=$(awk -F\, 'NR == 14 {print $2}' Input.csv)
bbmapPath=$(echo "$bbmapPath" | tr -d '\r') #carriage return

trimJarPathNam=$(awk -F\, 'NR == 16 {print $2}' Input.csv)
trimJarPathNam=$(echo "$trimJarPathNam" | tr -d '\r') #carriage return

adapterFileNam=$(awk -F\, 'NR == 17 {print $2}' Input.csv)
adapterFileNam=$(echo "$adapterFileNam" | tr -d '\r') #carriage return
```

As a demonstration of the results of the mentioned chunk of code, one could execute the command

> echo $inPath

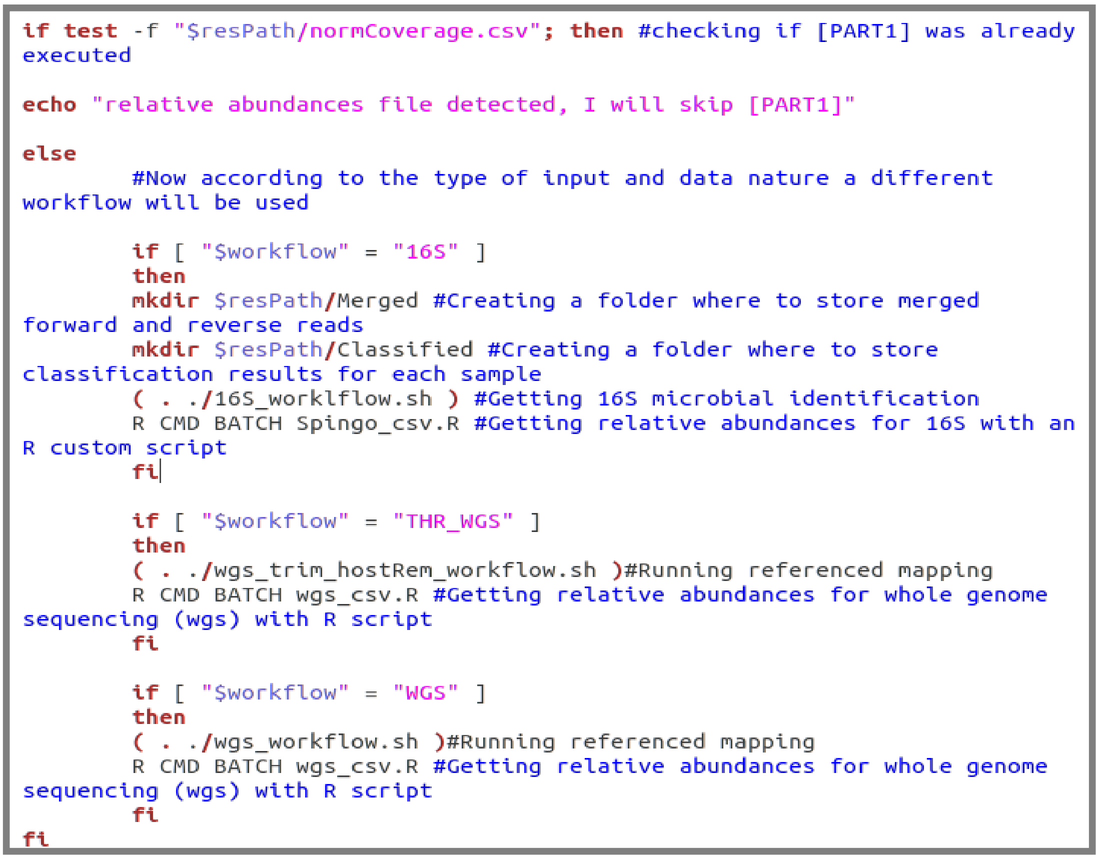
Retrieving the path to our directory where we placed the downloaded fastQ files. This indicates us, that the inputs were properly loaded in the terminal environment.

Now we will create R and MATLAB inputs: the following command will execute a script that will launch R and MATLAB and save input variables into binary files storing them into the temp folder.



```
#Now we can create R and MATLAB inputs: the following script will launch R
and MATLAB and save input variables into binary files storing them into the
temp folder
./InputCreator.sh
```

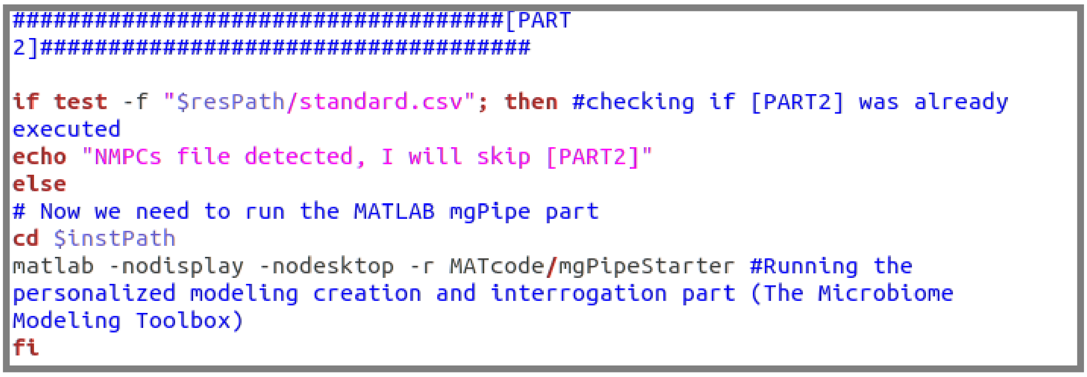
The next chunk of code, according to the type of input and data nature, will run a different workflow. In our case, the 16S protocol will be executed. As explained by the comments next to the code, initially two folders will be created: a folder “Merged” where to store merged forward and reverse reads and a folder where to store classification results for each sample “Classified”. After that, the 16S rRNA gene sequencing workflow will be executed. Briefly, for each sample forward and reverse reads are merged and quality filtered, and the merged sequences taxonomically classified. Finally, an R script parses the classification results for each sample and creates relative abundances tables for different taxonomic resolution and a relative abundance file based on the species present in our metabolic resource (AGORA).



```
if test -f "$resPath/normCoverage.csv"; then #checking if [PART1] was already
executed

echo "relative abundances file detected, I will skip [PART1]"

else
        #Now according to the type of input and data nature a different
workflow will be used

        if [ "$workflow" = "16S" ]
        then
        mkdir $resPath/Merged #Creating a folder where to store merged
forward and reverse reads
        mkdir $resPath/Classified #Creating a folder where to store
classification results for each sample
        ( . ./16S_worklflow.sh ) #Getting 16S microbial identification
        R CMD BATCH Spingo_csv.R #Getting relative abundances for 16S with an
R custom script
        fi

        if [ "$workflow" = "THR_WGS" ]
        then
        ( . ./wgs_trim_hostRem_workflow.sh )#Running referenced mapping
        R CMD BATCH wgs_csv.R #Getting relative abundances for whole genome
sequencing (wgs) with R script
        fi

        if [ "$workflow" = "WGS" ]
        then
        ( . ./wgs_workflow.sh )#Running referenced mapping
        R CMD BATCH wgs_csv.R #Getting relative abundances for whole genome
sequencing (wgs) with R script
        fi
fi
```

After this part is executed, the result folder should contain a file called genusMap.csv with relative abundances at genus taxonomic resolution, a file called speciesMap.csv with relative abundances at species taxonomic resolution, and a file called normCoverage.csv with relative abundances including only the species contained in the AGORA resource.

The next step, using the AGORA specific relative abundances, will be running the personalized modeling creation and interrogation part from The Microbiome Modeling Toolbox **[PART 2]**.



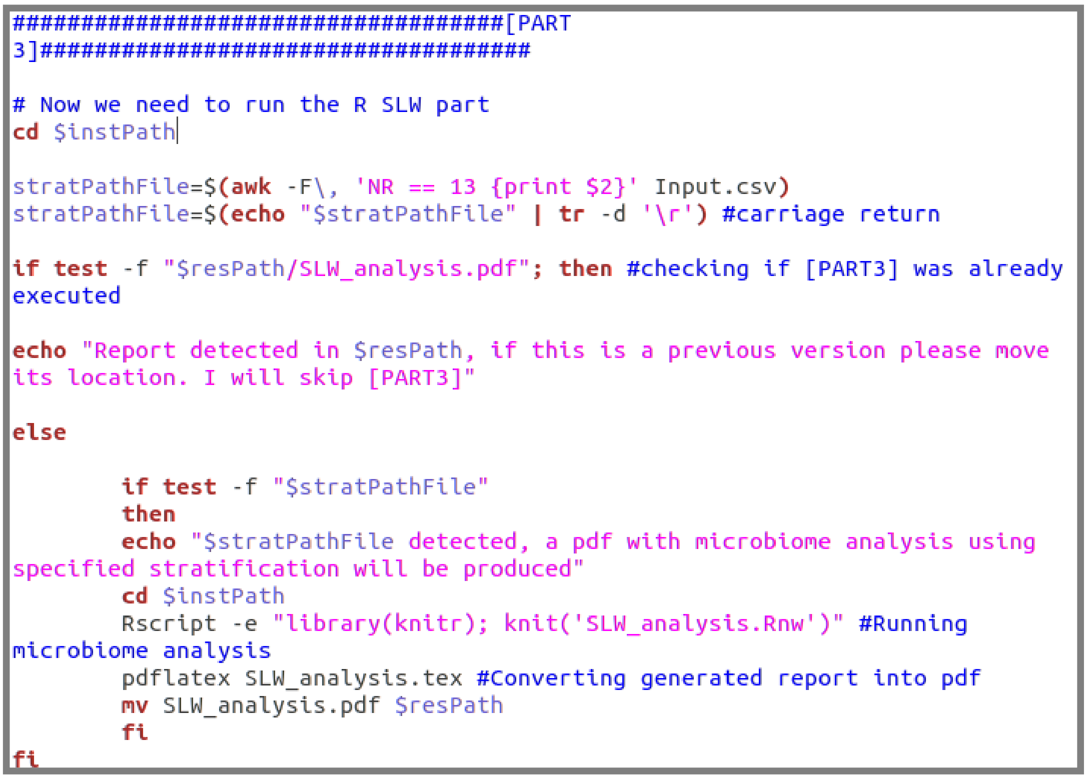
```
####################################[PART
2]######################################

if test -f "$resPath/standard.csv"; then #checking if [PART2] was already
executed
echo "NMPCs file detected, I will skip [PART2]"
else
# Now we need to run the MATLAB mgPipe part
cd $instPath
matlab -nodisplay -nodesktop -r MATcode/mgPipeStarter #Running the
personalized modeling creation and interrogation part (The Microbiome
Modeling Toolbox)
fi
```

This command will run a specific MATLAB script that loads the inputs and launche the automated mgPipe.m module of the Microbiome Modeling Toolbox. Initially, reactions abundance will be computed accounting, in a continuous way, for microbial reaction content in each sample. After that, a global community model will be created (Setup.mat), then, integrating the relative abundances information, personalized modes for each sample will be created (microbiota_model_xxxx.mat) and interrogated for their secretion profiles (NMPCs). The metabolite resolved metabolic profiles will be saved in the file called standard.csv.

To recapitulate, the main outputs produced and saved in the results folder are:

- genusMap.csv -> relative abundances at genus taxonomic resolution
- speciesMap.csv -> relative abundances at species taxonomic resolution
- normCoverage.csv -> AGORA matched species relative abundances
- reactions.csv -> reaction abundances
- standard.csv -> reaction abundances

One can decide to take these files separately and run any statistical analysis using his/her favorite set of analysis. By default, individuals plots of metabolic diversity in relation to microbiota size as well as Classical multidimensional scaling (PCoA) on individuals reaction repertoire and metabolic profiles (NMPCs) are also computed and saved in this part (see documentation). Optionally, the user can decide to run **[PART 3]** providing a file containing information on sample stratification. The code to run **[PART 3]** is embedded at the end of Runner.sh and can be executed pasting the following chunk of code:



```
####################################[PART
3]####################################

# Now we need to run the R SLW part
cd $instPath

stratPathFile=$(awk -F\, 'NR == 13 {print $2}' Input.csv)
stratPathFile=$(echo "$stratPathFile" | tr -d '\r') #carriage return

if test -f "$resPath/SLW_analysis.pdf"; then #checking if [PART3] was already
executed

echo "Report detected in $resPath, if this is a previous version please move
its location. I will skip [PART3]"

else

        if test -f "$stratPathFile"
        then
        echo "$stratPathFile detected, a pdf with microbiome analysis using
specified stratification will be produced"
        cd $instPath
        Rscript -e "library(knitr); knit('SLW_analysis.Rnw')" #Running
microbiome analysis
        pdflatex SLW_analysis.tex #Converting generated report into pdf
        mv SLW_analysis.pdf $resPath
        fi
fi
```

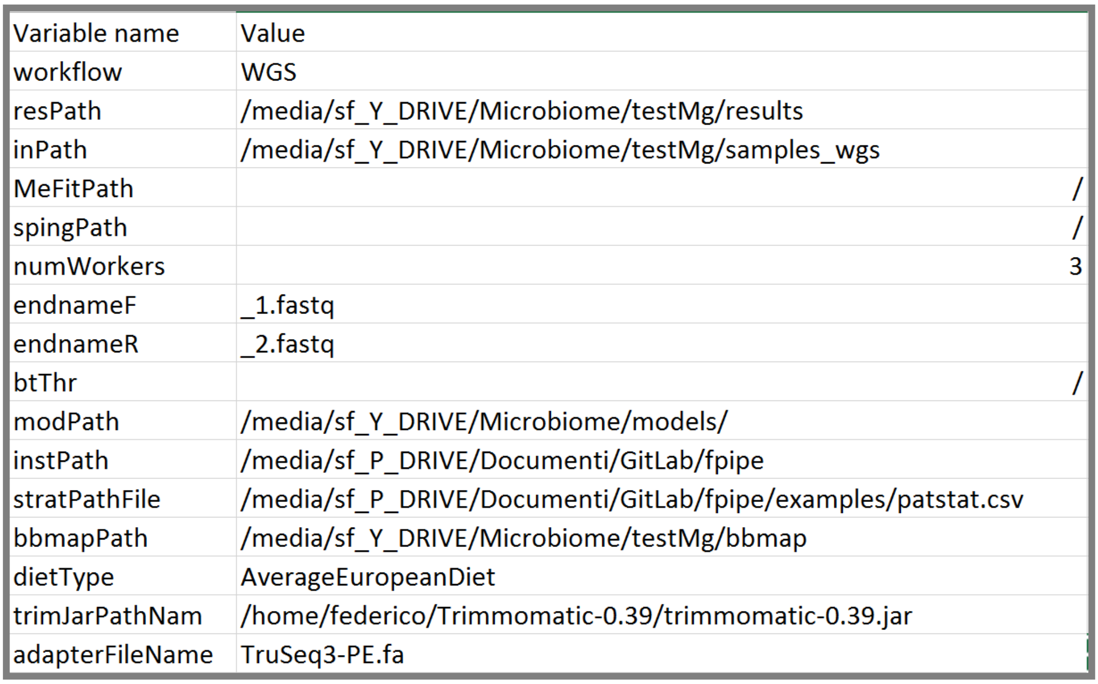
In this case, since we have no information on sample stratification, statistical analysis will not run, and no report will be produced. We will see in the next session an example of a generated report with **[PART 3]** stratification analysis.

WHOLE GENOME SEQUENCING (WGS)

For this tutorial, we will download four samples from a dataset available at EBI metagenomics. In this case, the study SRP057027 on pediatric Crohn's disease patients. The selected samples are SRR2145316 and SRR2145317 for the patients, while SRR2145359 and SRR2145371 were selected for the controls. (https://www.ebi.ac.uk/ena/data/view/SRRXXX)

As first step, please download and unzip all the sequences from the study (2 x 4 samples) and place them in a dedicated folder. In our case, the dedicated folder is called "samples_wgs". Then place into the same folder the indexed reference genome files "joined_genomes.fasta", the human genome files, (https://drive.google.com/drive/folders/1-iZt0teOg2YHyOPVuv3KXuuT9nZeeyDa?usp=sharing), and the adapters files (https://drive.google.com/drive/folders/19Tso01988X7032VaToQQZtt5O2-Vtk2V?usp=sharing) . As a final operation, since a stratification criteria is possible (Healthy vs. Crohn's Disease) we need to specify labels for each of the four observations. For this example, a file containing an example of a stratification file is contained in the folder "examples".

Now we are ready to proceed with **[PART 1]** of the pipeline. Go to the mgPipe installation folder (where you cloned the repository) and open the csv file called Input.csv. We already explained in the previous paragraph of this tutorial (16S rRNA GENE SEQUENCING) what the different variables mean. Please, edit the file Input.csv to enter your specific inputs, save and close it. In our case, considering the folders we prepared, or desktop power and the names of the sequences, the file will look like the following:



Variable name	Value
workflow	WGS
resPath	/media/sf_Y_DRIVE/Microbiome/testMg/results
inPath	/media/sf_Y_DRIVE/Microbiome/testMg/samples_wgs
MeFitPath	/
spingPath	/
numWorkers	3
endnameF	_1.fastq
endnameR	_2.fastq
btThr	/
modPath	/media/sf_Y_DRIVE/Microbiome/models/
instPath	/media/sf_P_DRIVE/Documenti/GitLab/fpipe
stratPathFile	/media/sf_P_DRIVE/Documenti/GitLab/fpipe/examples/patstat.csv
bbmapPath	/media/sf_Y_DRIVE/Microbiome/testMg/bbmap
dietType	AverageEuropeanDiet
trimJarPathNam	/home/federico/Trimmomatic-0.39/trimmomatic-0.39.jar
adapterFileName	TruSeq3-PE.fa

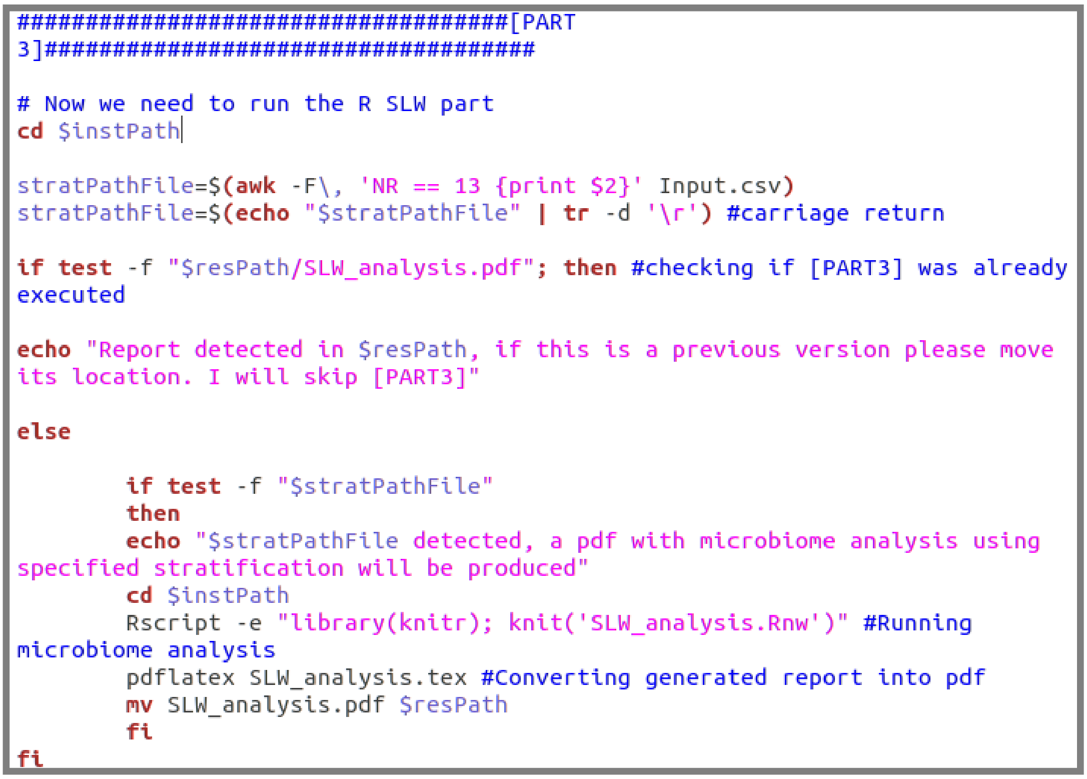
Please note that the first input (“Workflow”) that defines the protocol uses for whole genome sequencing can be “WGS” or “THR_WGS” depending on the type of input sequences. The difference is that using the “THR_WGS” input additional preprocessing steps such as sequences trimming (with adapters removal) and host sequences removal are performed. As these steps are commonly performed by sequencing centers, ask your sequencing center if they already performed them. For the purpose of this tutorial, considering exclusively the time of processing we will use the “WGS” input.

We can now process **[PART1]** as written in Runner.sh. The steps to follow are the ones that we already described these steps in the previous chapter of this tutorial, the only difference is the workflow that will be executed. Given the input “WGS” the 16S rRNA gene sequencing workflow will not run; the whole genome sequencing workflow will be launched by the last lines of the **[PART1]** code contained in Runner.sh

After this part is executed, the lines of code corresponding to **[PART2]** can be executed as showed in the previous chapter.
Once again, the main outputs of the first two parts will be the following:

- normCoverage.csv -> AGORA matched species relative abundances
- reactions.csv -> reaction abundances
- standard.csv -> reaction abundances

Now we are ready to execute **[PART 3]**.



```
###################################[PART
3]###################################

# Now we need to run the R SLW part
cd $instPath

stratPathFile=$(awk -F\, 'NR == 13 {print $2}' Input.csv)
stratPathFile=$(echo "$stratPathFile" | tr -d '\r') #carriage return

if test -f "$resPath/SLW_analysis.pdf"; then #checking if [PART3] was already
executed

echo "Report detected in $resPath, if this is a previous version please move
its location. I will skip [PART3]"

else

        if test -f "$stratPathFile"
        then
        echo "$stratPathFile detected, a pdf with microbiome analysis using
specified stratification will be produced"
        cd $instPath
        Rscript -e "library(knitr); knit('SLW_analysis.Rnw')" #Running
microbiome analysis
        pdflatex SLW_analysis.tex #Converting generated report into pdf
        mv SLW_analysis.pdf $resPath
        fi
fi
```

This time a stratification file is present and analysis based on each observation labels will run. An automatically generated report SLW_analysis.pdf will be stored in the results folder (resPath). The report consists of the concatenation of different analysis explained by text and related figures. Additionally, a folder containing the different figures of the report called “figure” will be created in the pipeline folder.

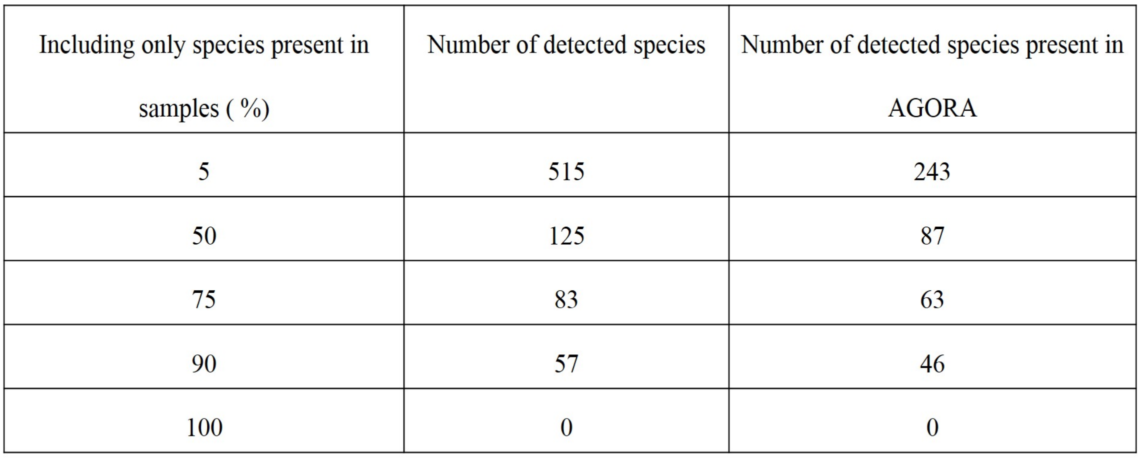
Appendix C

Supplementary material for Chapter 4

C.1 Extended information on the simulations

C.1.1 Mapping detected species on the gut microbial reconstruction collection

Currently, strain-specific metabolic reconstructions have been published for 819 gut microbes, named the AGORA collection [1-3], corresponding to 646 species. In the analyses data-set of the current study (n=309), 515 species were detected at least in 5% of the stools samples, 243 overlapped with AGORA. A total of 125 species were detected in at least 50% of the samples with an overlap of 87 AGORA species. Thus, 70% of the identified species were covered by the AGORA selection C.1. We conclude that our AGORA collection covers most of the frequently found species in our data set.



Including only species present in samples (%)	Number of detected species	Number of detected species present in AGORA
5	515	243
50	125	87
75	83	63
90	57	46
100	0	0

Table C.1: **Number of species present in AGORA in dependency on the number of detected species in at least 5,50%,75%, and 100% of the samples.**

C.1.2 Generation of personalized models

As a next step, we generated a generic microbiome metabolic reconstruction consisting of 257 microbial metabolic reconstructions, after removing all species that had a relative abundance below 1e-4. This generic microbiome reconstruction was then personalized to each sample by eliminating all species in a sample below this threshold and by adjusting the community biomass reaction coefficients to the normalized relative abundance data, as obtained with Spingo [4]. In absence of personal nutrition information, an average European diet was used to constrain each microbiome model[3, 5]. In average, the personalized microbiome models contained 67 species, 77,390 (non-unique) reactions, and 69,265 (non-unique) metabolites C.2. Furthermore, on average, the microbiome models covered 2,727 unique reactions and 67 species (Table B). The number of unique reactions, total reactions and total metabolites was slightly higher in PD in comparison with controls.

	Overall	PD	Control
Species	67 ± 11	69 ± 11	66 ± 10
Unique Reactions	2727 ± 96	2747 ± 101	2708 ± 86
Total Reactions	77390 ± 11853	79059 ± 12639	75875 ± 10933
Total Metabolites	69265 ± 10679	70757 ± 11353	67910 ± 9868

Mean ± Standard Deviation

Table C.2: **Personalized model characteristics.**

C.1.3 Supplementary figures

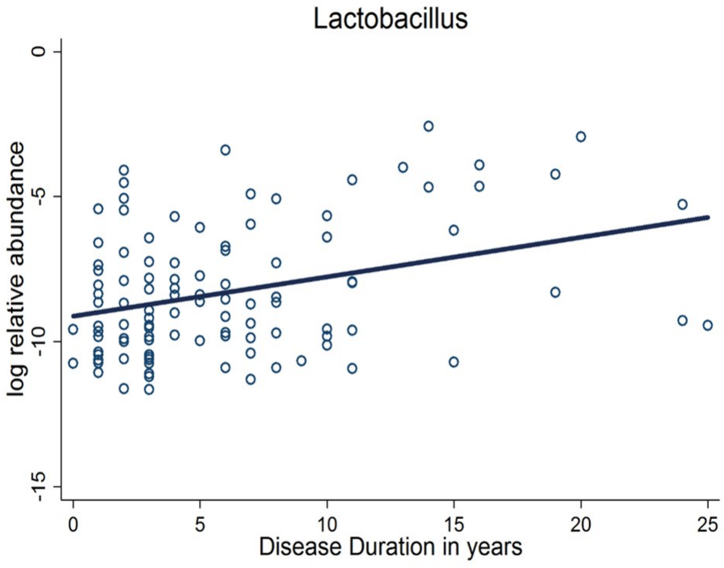


Figure C.1: **Lactobacillus relative abundance positively correlated with the duration of the disease (FDR<0.05).**

C.2 Supplementary tables

C.2.1 Supplementary table 1

Supplementary table 1 available on request

C.2.2 Supplementary table 2

Supplementary table 2 available on request

C.2.3 Supplementary table 3

Supplementary table 3 available on request

C.2.4 Supplementary table 4

Supplementary table 4 available on request