



Universidade do Minho

Taxonomic and functional analysis of metagenomes

Pedro Santos Barbosa

Supervisors:

Professor Miguel Francisco de Almeida Pereira da Rocha

Professor Joel Perdiz Arrais

October, 2013

Acknowledgments/Agradecimentos

Gostaria de iniciar este documento a prestar o meu agradecimento a um conjunto de pessoas que contribuíram para a realização desta tese de mestrado.

Ao Professor Doutor Miguel Rocha pela orientação desta dissertação, por toda a sua disponibilidade, pelos brainstormings, sugestões críticas e eficiente revisão do documento.

Ao Professor Doutor Joel Arrais por ter acreditado em mim desde o primeiro momento em que o contatei, pelas discussões científicas, revisão do documento e pela forma como me acolheu em Coimbra durante a minha estadia na cidade.

Ao Doutor Óscar Dias, homem que desenvolveu o software *Merlin*, pela ajuda inicial na compreensão da arquitectura do programa e pelas discussões acerca do melhor método a implementar na ferramenta.

Aos meus colegas do laboratório pela sabedoria partilhada e boa disposição diária. Em especial, à Sara e ao Filipe Liu pela ajuda e conselhos fornecidos em questões de programação, e à Carla, pela pessoa importante que se tornou ao longo deste caminho e por me aturar nos momentos menos bons.

Aos meus amigos da equipa do Carreira e dos escuteiros, pelos bons momentos proporcionados que me permitiam relaxar e recarregar forças nos dias menos produtivos.

Por último, e não menos importante, gostaria de agradecer aos meus pais e resto da família pelo apoio financeiro, por apostarem na minha formação, pela paciência e confiança demonstrados. Sem vocês não seria quem sou hoje!

Este trabalho foi financiado em parte pelo ERDF - European Regional Development Fund através do programa COMPETE (programa operacional para a competitividade), e Fundos Nacionais através da FCT dentro do projeto COMPETE FCOMP-01-0124-FEDER-015079.

Abstract

Over the years, metagenomics has demonstrated to play an essential role on the study of the microorganisms that live in microbial communities, particularly those who inhabit the human body. Several bioinformatic tools and pipelines have been developed, but usually they only address one question: "Who is there?" or "What are they doing?".

This work aimed to develop a computational framework to answer the two questions simultaneously, that is, perform a taxonomic and functional analysis of microbial communities. *Merlin*, a previously developed software designed for the construction of genome-scale metabolic models for single organisms, was extended to deal with metagenomics data. It has an user-friendly and intuitive interface, not requiring command-line knowledge and further libraries dependencies or installation, as many other tools.

The extended version of *Merlin* can predict the taxonomic composition of an environmental sample based on the results of homology searches, where the proportions of phyla and genera present are discriminated. Regarding the metabolic analysis, it allows to identify which enzymes are present and calculate their abundance, as well as to find out which metabolic pathways are effectively present.

The performance of the tool was evaluated with samples from the Human Microbiome Project, particularly from the saliva. The taxonomic membership predicted in *Merlin* was in agreement with other tools, despite some differences in the proportions. The functional characterization showed a conserved pool of pathways through different samples, although *Merlin* sometimes presented less pathways than expected because the routine is highly dependent on the enzymes annotation. Overall, the results showed the same pattern as reported before: while the pathways needed for microbial life remain relatively stable, the community composition varies extensively among individuals.

In the end, *Merlin* demonstrated to be a reliable standalone alternative to web services for those scientists that have concerns about sharing data.

Resumo

Ao longo dos anos, a metagenómica demonstrou ter um papel essencial no estudo dos microorganismos que vivem em comunidades bacterianas, particularmente aqueles que habitam o corpo humano. Várias ferramentas e pipelines bioinformáticas foram desenvolvidas, mas normalmente estas apenas abordam uma destas questões: "Quem está lá?" ou "O que é que estão a fazer?"

Este trabalho teve como objectivo o desenvolvimento duma ferramenta computacional para responder aos dois problemas em simultâneo, isto é, realizar tanto uma análise taxonómica como funcional de comunidades microbianas. O *Merlin*, um software anteriormente desenvolvido para construir modelos metabólicos à escala genómica para um organismo, foi estendido para tratar dados de metagenómica. O programa possui uma interface intuitiva e amigável do utilizador, não necessitando de conhecimentos de linha de comandos nem de dependências de bibliotecas ou instalação de aplicações adicionais.

Esta versão estendida do *Merlin* prevê a composição taxonómica global dum metagenoma baseado nos resultados de procuras de sequências homólogas, onde as proporções dos géneros e espécies são apresentadas. No que diz respeito à análise metabólica, o *Merlin* permite identificar quais as enzimas presentes e calcular a sua abundância, bem como identificar quais as vias metabólicas que estão efectivamente presentes.

O desempenho da ferramenta foi avaliado com amostras do Projecto do Microbioma Humano, particularmente com amostras da saliva. A composição taxonómica prevista no *Merlin* esteve de acordo com outras ferramentas, apesar de algumas diferenças observadas nas proporções. A caracterização funcional mostrou um conjunto conservado de vias metabólicas nas diferentes amostras, mesmo que o *Merlin* tenha identificado menos enzimas que o esperado, pois o método é bastante dependente do processo de anotação. Globalmente, os resultados revelaram o mesmo padrão reportado anteriormente: enquanto as vias metabólicas necessárias para a vida microbiana se mantêm estáveis, a composição taxonómica varia bastante entre indivíduos.

No final, o *Merlin* demonstrou ser uma alternativa fidedigna a serviços web para aqueles cientistas que têm restrições em divulgar os seus dados não publicados num website.

Contents

Acknowledgements/Agradecimientos	iii
Abstract	v
Resumo	vii
List of figures	xiii
List of tables	xvi
Acronyms	xvii
1 Introduction	1
1.1 Context and Motivation	1
1.2 Objectives	3
1.3 Structure of the thesis	3
2 Metagenomics: concepts and methods	5
2.1 DNA sequencing and assembly	5
2.2 Taxonomic classification of metagenomes	8
2.2.1 <i>Unsupervised</i> methods	8
2.2.2 <i>Supervised</i> methods	9
2.3 Functional analysis of metagenomes	12
2.3.1 Gene Prediction	12
2.3.2 Functional Annotation	15
2.3.3 Pathway inference of communities	20
2.4 Human Microbiome Project	27
2.4.1 How does it work?	27
2.4.2 Bioinformatics for the HMP	29
2.4.3 First achievements and future work	30
2.5 Merlin	31
2.5.1 Identification of genes that encode enzymes	31

2.5.2	Identification of genes that encode transporter proteins and compartments prediction	32
2.5.3	Metabolic reconstruction	33
3	Methodology and algorithms	35
3.1	New complementary features in <i>Merlin</i>	35
3.1.1	Database management	36
3.1.2	Enzymes annotation	38
3.1.3	Uniprot requests	39
3.1.4	Implementation of Local Blast	39
3.2	Metagenomics pipeline	41
3.2.1	Taxonomy inference	41
3.2.2	<i>Merlin</i> 's operation mode for taxonomy	43
3.2.3	Metagenomics functional characterization	46
3.2.4	<i>Merlin</i> 's operation mode for enzymes	46
3.2.5	Metagenomics pathways inference	49
3.2.6	<i>Merlin</i> 's operation mode for pathways	50
3.2.7	Architecture and requirements	53
4	Saliva Microbiome: results and discussion	55
4.1	Annotation of Metagenomes	56
4.2	Taxonomic composition	57
4.2.1	Inference from local BLAST annotations	57
4.2.2	Stringency of the routine parameters	59
4.2.3	Characterization of the Saliva microbiome	60
4.3	Functional capabilities	68
4.3.1	Encoded enzymes	68
4.3.2	Funcional pathways	75
5	Conclusions and further work	85
5.1	Overview	85
5.2	Limitations	87
5.3	Future work	88
	References	90
A	Supplementary material	105

List of Figures

2.1	Flowchart of the main stages and available methods/tools for metagenomics pathway-based functional analysis.	25
2.2	Schematic representation of <i>Merlin</i> architecture (Figure extracted from [154])	34
3.1	<i>Merlin</i> 's view for saving a backup of the project database. . .	37
3.2	<i>Merlin</i> 's view for creating a new database from a backup SQL file.	37
3.3	<i>Merlin</i> 's new ' <i>Merge databases</i> ' view.	37
3.4	<i>Merlin</i> 's new Local Blast operation.	41
3.5	<i>Merlin</i> 's view for taxonomy information of metagenomic datasets.	44
3.6	Detail windows for the ' <i>Taxonomy</i> ' main view (Phylum scores on the left, Genus scores on the right).	45
3.7	Statistics and overall community composition displayed in <i>Merlin</i>	45
3.8	<i>Merlin</i> 's view for metagenomic enzymes information.	47
3.9	<i>Merlin</i> 's detailed information for different genera encoding a selected enzyme.	48
3.10	Detail windows for the ' <i>Enzymes</i> ' main view (Genes on the left, Reactions on the right).	48
3.11	Statistics of the metagenomics enzymes entity.	48
3.12	<i>Merlin</i> 's view for metagenomic pathway information.	51
3.13	<i>Merlin</i> 's detailed information for genera operating a selected pathway.	52
3.14	Detail windows for the ' <i>Pathways</i> ' main view (Enzymes on the left, Reaction on the right).	52
3.15	Statistics of the metagenomics pathways entity.	53
3.16	Schematic representation of <i>Merlin</i> architecture for metagenomic analysis.	53

4.1	Community structure for the saliva samples inferred from local BLAST annotations against SwissProt. Pie charts were generated in R version 2.15.1 using 'plotrix' package.	58
4.2	<i>Merlin</i> predictions of the phyla composition in the samples from saliva. The 'SRS014692' and 'SRS014468' samples are contaminated thus these samples were discarded regarding further analysis.	61
4.3	Phyla distribution of the non contaminated saliva samples stored in KEGG. They can be accessed with the KEGG metagenomes IDs 'T30414', 'T30237' and 'T30194'.	62
4.4	Phyla distribution in the saliva samples taken from MG-RAST. To draw the charts, the data can be downloaded through the following MG-RAST metagenomes IDs: (a) 4478542.3; (b) 4473348.3; (c) 4473411.3;	63
4.5	Overall composition predicted in <i>Merlin</i> for the seven most abundant genera in each sample from saliva. Non classified genes were not included in this chart.	64
4.6	Genus distribution of the saliva samples in different tools. MEGAN was run from the BLAST results of RAPsearch2 against RefSeq database.	65
4.7	Abundance of metabolic enzymes through the three non contaminated samples from saliva across different annotations. Vertical bars represent the samples. Horizontal bars represent the relative abundances of enzymes. Redder colors stand for more abundant patterns, whilst greener cells account for less abundant/absent enzymes. The Heatmap was built using the 'Heatplus' package from Bioconductor [163] with hierarchical clustering using Euclidean distance.	69
4.8	a: Common enzymes found in the non contaminated samples with <i>Merlin</i> (SwissProt). b-d: Comparison of common enzymes found with <i>Merlin</i> (Swissprot) and IMG/M in the non contaminated saliva samples.	70
4.9	Comparison of the BLAST results for a given gene in <i>Merlin</i> using different databases as reference.	72
4.10	BLAST result of a gene with different products and EC numbers within its list of homologues.	72
4.11	Proportion of the genes encoding the enzyme Exonuclease V (EC number: 3.1.11.5) executed by different taxonomic genus in the saliva samples.	74

-
- 4.12 Presence of metabolic pathways in the samples from saliva across different annotations. Vertical bars represent the samples. Horizontal bars represent the binary value for pathway coverage. Red colors stands for present pathways whilst green cells account for the absent ones. 'Heatplus' package from Bioconductor [163] was used with hierarchical clustering algorithm using Euclidean distance. 78
- 4.13 Representation of the present enzymes, marked in red, in the Sulfur metabolism pathway (map00920) for the SRS019120 sample with annotations against SwissProt (local BLAST). . . 80
- A.1 Old database schema for data retrieved from homology searches.105
- A.2 New database schema for data retrieved from homology searches.106

List of Tables

2.1	Specifications of some of the NGS platforms in the market (adapted from [16]).	7
2.2	Current existing pipelines/tools for functional annotation of metagenomic data (adapted from [101])	18
4.1	Description of the saliva samples used in this work. The samples name represent the assigned ID in the HMP data repository.	55
4.2	Remote BLAST against NCBI nr vs Local BLAST against SwissProt for the HMP samples ran in <i>Merlin</i>	56
4.3	Filtering the samples from saliva annotated through local BLAST, across the several steps of the <i>Merlin</i> taxonomic routine, with the default parameters.	58
4.4	Filtering of the samples from saliva annotated through remote BLAST, across the several steps of the <i>Merlin</i> taxonomic routine with different parameters settings.	60
4.5	Average distribution (%) of the five most abundant genus in the three non contaminated samples over the different tools. A top-down list of the genera ordered by their abundances is also presented.	66
4.6	Comparison of the complete enzymes annotated by IMG/M and <i>Merlin</i> for the non contaminated saliva samples.	68
4.7	Statistics regarding the assignment of enzymatic activity to a taxonomic genus in <i>Merlin</i> for the non contaminated samples from saliva using NCBI-nr as the reference database.	74
4.8	Number of pathways assigned with HUMAnN and <i>Merlin</i> for the saliva samples. The unique pathways columns refer to those that were exclusively classified by each method within each sample.	76
4.9	Statistics regarding the assignment of metabolic pathways to a taxonomic genus in <i>Merlin</i> for the non contaminated samples from saliva using NCBI-nr as the reference database.	81

Acronyms

bp	base pairs
CDD	Conserved Domain Database
DACC	Data Analysis and Coordination Center
DNA	Deoxyribonucleic acid
EC	Enzyme Commission
FBA	Flux Balance Analysis
GA IIx	Genome Analyser IIx
GO	Gene Ontology
GUI	Graphical User Interface
HIMI	Human Intestinal Metagenome Initiative
HMMs	Hidden Markov Models
HMP	Human Microbiome Project
HUMAnN	HMP Unified Metabolic Analysis Network
IMMs	interpolated Markov models
KO	Kegg Orthology
LCA	lowest-common ancestor
MGA	MetaGeneAnnotator
NGS	Next Generation Sequencing
NIH	National Institute of Health

nr	non-redundant
ORF	Open Reading Frame
PacBio	Pacific Biosciences
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PE	paired-end
PGM	Personal Genome Machine
PSI-BLAST	Position-Specific Iterated BLAST
RBS	Ribosomal binding site
RPS-BLAST	Reverse Position-Specific BLAST
rRNA	Ribosomal ribonucleic acid
SBML	System Biology Markup Language
SMRT	Single Molecule Real Time
SQL	Structured Query Language
SVM	support vector machine
TB	Terabyte
Tb	Tera base pair
TC	Transport Classification
TIS	Translation initiation site
WGS	Whole Genome Shotgun

Chapter 1

Introduction

1.1 Context and Motivation

For most of the history of life, microorganisms were the only inhabitants on Earth, and they keep dominating the planet in many aspects. Microbial life was essential for the evolution of life and has an important role in human health, agriculture, ecosystem functioning and global geochemical cycles. It is estimated that $4\text{--}6 \times 10^{30}$ prokaryotic cells reside in the planet and they retain 350–550 Petagrams ($1Pg = 10^{15}g$) of carbon, 85–130 Pg of nitrogen and 9–14 Pg of phosphorous, thus making them the largest deposit of those nutrients on Earth [1]. Therefore, sequencing the DNA of these organisms holds great importance, providing a better understanding of our world and enhancing strategies to improve it.

The first sequencing methods appeared in the 1970s, when Sanger and Coulson (1975) proposed a DNA sequencing strategy based on primed synthesis with DNA polymerase [2], which would become known as the Sanger method and enabled the first studies of microbial genomes with the sequencing of the bacteriophages MS2 [3] and ϕ -X174 [4]. The first bacterial genome was sequenced in 1995, the year when the *Haemophilus influenza* genome was published [5]. After this, several genomes were finished, but the real revolution started in 2004 when the so-called Next Generation Sequencing (NGS) technologies appeared and started to overcome the whole-genome Sanger se-

quencing with less expensive and faster methods [6].

Despite the progresses, single organism genomes studies have limitations, since it is necessary to culture an organism in order to sequence its entire genome. Unfortunately, only a small portion of the microorganisms can be cultured, which means that our understanding in the microbial world is highly biased and does not represent the reality in nature [7, 8]. Furthermore, almost all microbes live in multi species communities where they interact and benefit from microbial cooperation. A clonal culture lacks on the true representation of these states, thus making important to obtain genetic information directly from their natural habitats. The early studies on this unknown world focused on the 16S ribosomal RNA gene (rRNA) [9, 10] (which is often conserved within a species and generally different between species), and showed the potential of the field to characterize uncultivated prokaryotes.

With the emergence of the NGS technologies, the first whole genome shotgun (WGS) sequencing projects of bacterial communities were performed [11, 12]. This technology, called metagenomics, enables sequencing the entire community represented in the sample producing large volumes of data, in magnitudes that can reach terabytes (TB) of information in a soil sample [13]. Therefore, new computational challenges have arisen and new methods were needed to analyze these huge amounts of data with the ultimate goal of answering two main questions: "Who is there?" and "What are they doing?".

A great number of tools were released for metagenomics studies since then, either at the taxonomic or functional level, but there is still a lack of choices to perform an integrative analysis of microbial communities at both levels simultaneously. Moreover, if the user is not interested in running a web service, using some of the available standalone programs can be a real hurdle since they are usually command-line based and require further libraries dependencies and installation of additional tools before they can actually be run.

1.2 Objectives

In the context stated above, the main goal of this project focused on developing a user-friendly tool capable of performing a taxonomy description as well as a robust metabolic reconstruction of a microbial community, where enzymes and pathways present in the metagenome are discriminated.

The work was done by adapting the already implemented software *Merlin*. The new features are able to deal with sequences from multiple organisms as input and perform a taxonomic and functional characterization of any metagenome. This process can be performed by users with little bioinformatic skills.

Furthermore, the application of this tool to saliva samples from Human Microbiome Project (HMP) was done, along with an evaluation of its performance. In this work, the following tasks were accomplished:

- Collection of relevant bibliography and use of some existing software to achieve a good comprehension of the *state-of-the-art* methods in the areas involved;
- Development of methods to analyze metagenomes at the taxonomic and functional level and integrate them within *Merlin*;
- Evaluation of the performance of the developed tool, using human saliva microbiome data from HMP;

1.3 Structure of the thesis

Chapter 2

Metagenomics: concepts and methods

The current state of metagenomics and the underlying strategies for taxonomic and functional characterization of microbial communities will be addressed.

Chapter 3

Methodology and algorithms

All the methods developed in the scope of this work will be explained as well as the integration of such methods in the *Merlin* platform.

Chapter 4

Saliva Microbiome: results and discussion

The results obtained by Merlin will be discussed and comparisons with other tools will be performed.

Chapter 5

Conclusions and further work

An overall analysis of this work along with its limitations will be presented.

Chapter 2

Metagenomics: concepts and methods

2.1 DNA sequencing and assembly

The field of metagenomics is highly dependent of the NGS technology availability, and the growing abundance of sequencing platforms has resulted in a constant improvement of its capability. These platforms can be divided into two broad groups: the ones depending on the production of libraries of clonally amplified templates, and more recently, the use of single-molecule sequencing platforms, which determine the sequence of single molecules without amplification [14]. Both systems produce thousands or millions of fragments from random positions of the genome with high coverage, but each platform holds its own specifications.

The first NGS platform was launched in 2005 by the company 454 Life Sciences, which was later acquired by Roche Applied Science. This sequencer relies on pyrosequencing technology, in which instead of using dideoxynucleotides to terminate the chain amplification as Sanger does, it depends on the detection of pyrophosphate released during nucleotide incorporation. Amplification beads are used to capture the fragmented DNA libraries followed by emulsion Polymerase Chain Reaction (PCR). The most famous instrument of Roche 454 (*454 GS FLX Titanium*) was launched in 2008 and

is able to produce long reads of ~ 700 base pairs (bp) with 99.9% accuracy. The great speed of this technology is the most remarkable advantage but the high cost of the reagents and equipment remains a limitation [15].

In recent years, the sequencing industry has been dominated by Illumina, which adopts a sequencing by synthesis approach. The DNA library with the adapters is denaturated into single strands and grafted to a flowcell, followed by bridge amplification to form clusters of clonal DNA fragments. Four different types of nucleotides (ddATP, ddGTP, ddCTP, ddTTP) that contain different fluorescent dye and a removable blocking group will complement the template one base at a time and the signal is captured. The *Illumina Genome Analyser IIx (GA IIx)*, the *HiSeq 2000*, and more recently the *MiSeq* are the most successful sequencing instruments launched by the company. They output paired-end (PE) reads up to 150 bp and have a lower reagent cost comparing to other technologies. One of the known shortcomings of Illumina instruments is its run time (10 days for *GAIIx* and 11 days for *HiSeq 2000*) but *MiSeq* can handle run times of 27 hours, since it is oriented for smaller laboratories and the clinical diagnostic market [16].

More recently, new sequencing platforms were released, such as the *Ion Torrent Personal Genome Machine (PGM)* and the *Pacific Biosciences (PacBio) RS*, both based on revolutionary technologies. Ion Torrent uses semiconductor sequencing technology [17], where it detects the protons released as nucleotides are incorporated by polymerase during synthesis, analyzing the changes in pH for detecting whether the nucleotide was added or not. Ion Torrent is the first commercial machine that does not require fluorescence and camera scanning thus resulting in higher speed, lower cost and smaller instrument size [15]. PacBio introduced the single molecule real time (SMRT) sequencing process [18]. Here, DNA polymerases bounded to the DNA template are loaded into zero-mode waveguides (ZMWs) where replication occurs, producing nucleotide-specific fluorescence. After each run there is a bioinformatics treatment of the single-molecules fragments and consensus sequences are generated, producing longer reads [19].

In Table 2.1 the main features of the described machines are represented.

Table 2.1: Specifications of some of the NGS platforms in the market (adapted from [16]).

	454 GS FLX	GAIIx	HiSeq 2000	MiSeq	Ion Torrent PGM	PacBio RS
Instrument cost	\$500K	\$256K	\$654K	\$128K	\$80K	\$695K
Run time	24 hours	10 days	11 days	27 hours	2 hours	2 hours
Raw Error Rate	1.07%	0.76%	0.26%	0.80%	1.71%	12.86%
Read length	~700bp	up to 150bp	up to 150bp	up to 150bp	~200bp	Average 1500bp
Paired reads	No	Yes	Yes	Yes	Yes	No

The DNA assembly aims to align and merge the reads provided by the sequencing machines in order to reconstruct the original genome (or genomes, in metagenomics). The ideal scenario on a metagenomics assembly would be to have a draft genome assembly for each member of the sample and perform further analysis with high confidence as it is done for single genomes. Although it is possible to assemble individual genomes from low complexity communities, such environments are not representative of the diversity seen in most natural ecosystems [20].

Several genome assemblers designed for metagenomics have been released [21–24], but they all suffer from the same constraints: (i) the species abundance varies within each community and the assemblers tend to build contigs only for those species with high sequence coverage (most dominant ones), discarding the less abundant phyla in the community; (ii) chimerism, in which reads from one species are assembled into contigs from another species; (iii) high level of fragmentation of the contigs, even after tuning the assembler parameters [20, 25, 26].

Recently, a sequence-independent approach was proposed to recover microbial genomes from metagenomics samples based on the differential coverage binning of the reads, allowing separation of the reads into species-level

clusters that can be assembled into single chromosomes [27].

As sequencing platforms and computational methods continue to improve, the task of assembling complex microbial communities will be facilitated. However, a perfect assembling at the species level is still not possible, hampering the downstream analysis for any metagenomics study.

2.2 Taxonomic classification of metagenomes

Since the realization that microbial diversity is much higher than previously observed, the taxonomic characterization of microbial communities has been under attention of the scientific community. The first metagenomics studies focused on 16S rRNA for genetic diversity analysis [28] but the application of this gene has been boosted by the advances in DNA sequencing and barcoded pyrosequencing [29]. NGS technologies can use 16S rRNA amplification primers for targeting hypervariable regions (there are 9 for this gene: V1-V9) enabling the discrimination of bacterial diversity in environmental samples [30].

WGS sequencing is also applied to study microbial composition/diversity of metagenomes and the focus now will be on this topic. There are two types of sequence classification methods: *unsupervised* learning and *supervised* learning. The first approach does not need previous knowledge and classifies the sequences independently (e.g clustering groups of similar sequences together), while the second one classifies them using previously labeled sequences.

2.2.1 *Unsupervised* methods

Some strategies have been developed for this learning approach [31, 32]. These methods are usually performed by a binning process, in which the metagenomics reads are distributed into taxon-specific bins without using assemblies, database searches or alignment with reference genomes. Instead, binning algorithms based on DNA composition (GC content and codon usage) are used for species inference. Given the fact that a significant fraction

of the samples harbors unknown bacteria, these methods have the advantage of performing well on previously unseen data [32].

While LikelyBin uses maximum-likelihood estimations for clustering the data based on their *k-mer* distribution [32], ComposBin applies principal component analysis (PCA) to project the data into an informative lower-dimensional space and then uses the normalized cut clustering algorithm on this filtered dataset to classify sequences into taxon-specific bins [31].

2.2.2 *Supervised methods*

Three main categories have been identified for sequence classification based on *supervised* learning: sequence similarity search, sequence composition methods and phylogenetic methods [33]. A large number of software applications have been released [34–42] and most of them use only one of these approaches, despite some exceptions where two methods are used simultaneously.

Similarity search methods

This approach relies on homology information obtained by database searches and can be further subdivided, whether they are based on hidden Markov models (HMMs) or BLAST. The most basic strategy concerning taxon selection is to search for the best hit in the database, but this type of classification has to be interpreted carefully, since the evolutionary distance between the DNA fragments and the hit is unknown [34]. However, such classification is reliable on higher taxonomic levels (e.g. superkingdom or phylum). CARMA [34], MARTA [35], MetaPhyler [37], MetaPhlAn [43] or MG-RAST [44] are some tools based on similarity searches and each of them has complementary features to improve the classification.

MEGAN [45], a popular software for metagenomics analysis, was pioneer by integrating a version of the lowest-common ancestor (LCA) algorithm for taxonomic labeling. MEGAN maps query sequences to NCBI and for each one that matches the sequence of some gene, the program places the sequence on the lowest LCA node of those species in the taxonomy that are known to

have that gene.

Sequence composition methods

This type of classification depends on the construction of sequence models, which are often built upon interpolated Markov models (IMMs), *naïve* Bayesian classifiers and k-means/k-nearest-neighbor algorithms [33]. Once models are built, this methodology is faster than homology-based methods.

The Phylopythia [40] web server was the first sequence composition-based taxonomic classifier to be released. It is based on a support vector machine (SVM) and outputs great accuracy for long (>1 Kbp) fragments. Another tool, NBC [38, 46], uses a *naïve* Bayes classifier to identify the taxonomy of any sequence. This classifier is trained on unique N -mer frequency profiles of 635 microbial genomes and is claimed to achieve 90% accuracy for highly-represented species.

To improve accuracy, a hybrid method was developed, which uses a combination of IMMs with BLAST. PhymmBL [39] identifies variable-length oligonucleotides specific for each phylogenetic group and the BLAST search is performed to complement and strengthen the results. Despite producing good results on short reads as 100 bp, this tool has the shortcoming of being computationally more expensive than its relatives.

Phylogenetic methods

The assumption behind these methods lies on the attempt to assign a query sequence on a phylogenetic tree according to a defined model of evolution using maximum likelihood, Bayesian methods or neighbor-joining, for instance [33]. Most of the programs are simply concerned with the placement (and hence classification) of the sequence in the tree and they all require building a multiple alignment for building it (and hence high computational power).

Most of the programs require marker genes, since the initial step in most workflows is to add a query sequence containing a marker gene to a reference alignment. Thus, for the selection of marker genes, these methods are generally combined with similarity searches making this approach a hybrid one.

AMPHORA [47], MLTreeMap [41] and TreePhyler [42] are examples of such approaches.

Software evaluation

A high number of tools has been published for the taxonomic classification of metagenomics samples, and there are even more not reported in this document. Therefore, in the user perspective, the choice of the best software for a specific study might be a challenge. It is crucial to have a reasonable classification accuracy, since it has a direct impact on downstream analysis and further conclusions. Fortunately, Bazinet and Cummings (2012), performed an extensive comparison of the different softwares for each method of *supervised* taxonomic classification described before [33].

They evaluated the performance of the classifiers in two main areas: accuracy and computational requirements. For the homology-based softwares, it became clear that the BLAST step dominates the runtime, with an exception for MetaPhyler that runs pretty quick but only classifies a small portion of the reads. Most of the programs achieve very good and concordant levels of precision and sensitivity.

Concerning the composition methods, NBC displayed the highest average sensitivity and precision, followed by PhymmBL. PhyloPythia took the longest time to train the dataset but the classification step took place $\sim 41x$ faster than PhymmBL. The average precision is lower for these programs in comparison with alignment-based ones, but the fact that classifications were performed at the genus level for composition-based softwares and at the phylum rank for alignment classifications prevented the authors of extracting meaningful conclusions.

Regarding the phylogenetic-based approach the authors only compared two programs: MLTreeMap and TreePhyler. The latter achieved better sensitivity and precision, despite the longer run times.

Overall, composition-based softwares displayed the highest average sensitivity (50.4%) and speed (once they were trained), while homology methods achieved the highest average precision (93.7%). The most precise programs

were CARMA (97.4 %) and MEGAN (98.1%) but they also carry the burden of being the most computationally expensive ones. On the other hand, NBC overcomes all other tools in terms of best combined sensitivity and precision (95.4%).

Concluding, the level of sequence representation in databases, taxonomic diversity, composition of the sample and read lengths influence the performance of each category between data sets, thus not making possible to claim which software is the best [33].

2.3 Functional analysis of metagenomes

2.3.1 Gene Prediction

Gene finding is an essential first step on the genome analysis and correct functional annotation. In a typical bacterial genome, only a small amount of the DNA does not encode protein sequences, being fundamental to distinguish these stretches of DNA from the coding ones. Protein coding sequences have statistical properties that differentiate them from non-coding frequencies, being the sequence composition the most important feature (e.g. the GC content) [48]. These patterns can be extracted using HMMs, models that are usually estimated by maximum likelihood, which maximizes the probability of a gene prediction based on a labeled sequence [49, 50]. Several tools based on HMMs have been produced for gene prediction on single genomes [51–54], in which the model parameters are trained on known annotations to predict unknown genes [55].

However, gene finding on metagenomics datasets is more problematic and this approach cannot be applied, at least directly, with the same confidence due to the assembly limitations. Therefore, the identification of genes directly from metagenomic short reads has been gaining importance.

Homology-based methods

Similarity based methods are applied for gene finding in metagenomics data [56–58], where it is possible to find the genes if their DNA or amino acid

sequences shows strong similarities against reference databases. In this case, annotation success is dependent on the already known genes and their closely related species. Another limiting factor of this method is that the computational cost for this task is high, considering the size of the metagenomics samples [59].

***Ab initio* methods**

On the other hand, the gene prediction can be made based on statistical models [60–64]. These models include features such as codon usage bias and start/stop codon patterns of known genes and have the advantage of predicting known and novel genes with lower computational expenses. A disadvantage of these methods regards that reads may contain sequencing errors that can lead to frame shifts and thus invalid gene predictions [59, 65].

The MetaGeneAnnotator (MGA)[60] integrates statistical models of bacterial, archaea and prophage genes that enables to detect lateral gene transfers and phage infections. It uses a self-training model that takes into account the GC content and the di-codon frequencies of the input sequences as features. In addition, MGA uses a feature that increases the confidence of the translation starts site prediction: a ribosomal binding site (RBS) model based on specific 16S rRNA binding sites.

FragGeneScan [62] builds a model based on HMMs, but also integrates codon usage bias, sequencing error methods and start/stop codon patterns. Actually, this software is the only one that takes into account sequencing errors, which were shown to improve the true positive gene prediction rates [66].

GeneMark [63] was adapted from a previous HMM-driven gene finder [67], by directly estimating the codon and oligonucleotide frequencies from the reads, which enables to apply heuristics that increase the accuracy of the parameter estimation of the HMM model, and thus perform better gene prediction. It also provides separate models for bacteria and archaea.

Another tool, called Orphelia [61], performs predictions in two stages: first, it extracts features such as monocodon usage, dicodon usage and the

translation initiation site (TIS) from sequences. Then, an artificial neural network gathers the sequence features, such as the GC content and the open reading frame (ORF) length, and outputs a probability of an ORF to encode a protein and based in that probability it performs the gene prediction. Orphelia enables gene finding of reads with variable length, while maintaining good performance. For that, the neural network is trained with the specific length of the reads that are being used for gene discrimination.

A widely used gene finder, Glimmer has been recently adapted to deal with metagenomes. The Glimmer-MG [64] uses another approach than GC-content for model parameterization, a phylogenetic classification feature based on the Phymm system [39], which finds evolutionary relatives of the sequences on which to train. Furthermore, it uses an unsupervised method for sequence clustering, SCIMM [68], that groups the reads that might belong to the same organism. Glimmer-MG pipeline integrates these two steps prior to an initial gene prediction, which is performed based on IMMs. The models are retrained within each cluster and features such as insertion/deletion are also added, enabling the final gene predictions. This method has the disadvantage of being substantially more computationally expensive.

Software evaluation

Few comparisons have been made concerning the choice of the best gene finder. Yok and Rosen (2011) [55] studied the performance of GeneMark, MGA and Orphelia separately, along with a combination of the three methods. They evaluated the programs with different read-length datasets and found a trade-off of sensitivity vs. specificity and a decline in these rates for shorter reads. Orphelia and MGA showed high sensitivity, while GeneMark presented the highest specificities values. GeneMark was the best in predicting the start and end of genes for short read lengths, such as the reads produced in the HMP (Illumina \sim 100 bp), while Orphelia has the lowest annotation error for longer read lengths. A combination of the three methods showed the best performance (optimizing prediction and annotation accuracy) for reads between 100-400 bp. For longer reads, a combination of

GeneMark and Orphelia had the best results.

More recently, the developers of Glimmer-MG performed a comparison between their software, FragGeneScan and MetaGeneMark, claiming that their tool outperformed the other ones, both in terms of specificity and sensitivity in real and simulated datasets [64]. It may be important to refer that they only show results in real datasets for 454 reads, excluding from the analysis the short Illumina reads, used in the HMP.

2.3.2 Functional Annotation

Gene prediction is usually followed by functional annotation, which corresponds to the assignment of biological functions to the predicted ORFs. Likewise for gene prediction, the known problems on metagenomics assembly stated earlier are visible here, making this step commonly performed from short sequences [26]. This task is more challenging in metagenomics datasets, because many predicted ORFs are partial, and a large fraction does not have annotated homologues (species with unknown genome sequences) [20]. Furthermore, due to the short sequence size of the metagenomics data, some information (such as gene neighborhood in a genome, gene fusion, coexpression) that is important for function prediction in individual genomes, may not be available in this analysis [26]. Thus, sometimes the gene finding step is skipped from the pipeline and unassembled single reads are used to infer functional information, despite the known fact that the accuracy level is higher as the read-length increases [69]. Below, the existing strategies and methods for functional annotation are presented, feasible from the assembled contigs, predicted genes or unassembled reads.

Read mapping methods

A possible approach for function profiling is a read mapping strategy, in which the reads or predicted genes are simply mapped to reference genomes (MetaHIT, NCBI-nt or IMG/M HMP). The number of matches are counted and the functions scored accordingly [65]. Aligners that rely on the Burrows-Wheeler Indexing system such as BWA [70] and Bowtie [71] are used for this

task.

Parallel versions of BLAST [72] might also be suitable for this purpose, but a better accuracy/time balance is achieved with FR-HIT [73], a tool based on a k-mer hash table for the reference sequences, from which it performs seeding, filtering, and banded alignment to identify the best alignments to the reference sequences. This approach might be hindered due to sequence conservation in functional regions of the proteins across different organisms. A read that maps in one of these regions will probably be assigned to different targets with a similar score [65].

Homology-based methods

Several databases collect multiple sequence alignment of proteins that share a specific function. FIGfams [74] is a collection of protein families that is based on the SEED classification system [75]. It consists in a set of subsystems that were tested and manually curated such that they play a specific function in the cell. SMART [76] is an alternative database that contains protein domains based on HMMs, and owns a sub resource, metaSMART, dedicated to the analysis of domain architectures in various metagenomic data sets. Another databank, and perhaps the most important one, the NCBI Conserved Domain Database (CDD) [77] incorporates proteins from several sources such as Pfam [78] and TIGRFam [79] (profiles generated from HMMs) or COG [80] and Prk [81] (profiles generated from multiple sequence alignments) in order to annotate protein sequences. Alternative databases, such as KEGG [82] or Gene Ontology (GO) [83] provide protein function information at different levels. KEGG infers pathway information for the query sequences, while GO classifies gene products according to three different domains: depending on their cellular location, the overall biological process they are involved and the molecular function of the proteins.

Therefore, search engines were developed to scan proteins against these databases. BLASTx is widely used to search translated sequences against protein databases while BLASTp uses protein sequences as queries. The Reverse Position-Specific BLAST (RPS-BLAST), a variation of the previous

Position-Specific Iterated BLAST (PSI-BLAST) method [72], searches the query sequences against databases of profiles. Another commonly used tool, HMMER [84], looks for homologs of protein sequences using HMMs. The last version of the software is able to detect more remote homologs and be more accurate than BLAST with a similar speed, due to the strength of its underlying mathematical models. For less sensitive, but faster searches, the BLAT alignment tool may be used [85].

Alternatively, it is possible to scan protein databases such as NCBI RefSeq [86], UniProt/UniRef [87] or eggNOG [88] with fast protein search tools designed for next-generation sequencing data, such as RAPsearch2 [89] that uses reduced amino acid alphabet to reduce the overall complexity of the search.

Tools / Workflows for functional annotation

Despite the fact that functional annotation and analysis of metagenomic data sets are problems far from being adequately solved, several tools and pipelines have been produced to perform this task [36, 44, 90–96] (Table 2.2, Figure 2.1).

Almost all of them integrate multiple tools and databases described earlier to improve the analysis. Web-based servers, as is the case of CAMERA [90], MG-RAST [44], and IMG/M [91] host results from published metagenomes, which enable the users to compare their own datasets with those already published. The latter two tools also search for homologs in publicly available metagenomic sequences, increasing the confidence level of the hits. Some pipelines have unique features, such as the IMG/M that has a motif search over the InterPro database [97]. In addition to the Pfam and TIGRFam repositories, Interpro includes protein motifs databases like PROSITE [98] and PRINTS [99]. The MG-RAST web server and Smash community searches for functional interactions between proteins using the STRING database [100], and the web-based METAREP includes prediction of lipoprotein motifs. A detailed list and respective features of each pipeline is described in Table 2.2.

Table 2.2: Current existing pipelines/tools for functional annotation of metagenomic data (adapted from [101])

Tools Annotation	IMG/M	CAMERA	MG-RAST	METAREP	RAMMCAP	Smash commu- nity	MEGAN4	CoMet	WebMGA
Homology search	NCBI RefSeq, SMART, Uniprot	NCBI RefSeq	NCBI RefSeq, SMART, UniProt	NCBI RefSeq, Uniprot	-	SMART, UniProt	NCBI RefSeq	-	NCBI RefSeq
Metagenomic data sets	IMG/M	-	IMG/M	-	-	-	-	-	-
Orthologous groups	COGs	COGs	COGs, eggNOGs	-	COGs	COGs, eggNOGs	-	-	COGs
Protein families	Pfam, TIGRfam	Pfam, TIGRFam	FIGfams	Pfam, TIGRfam	Pfam, TIGRfam	Pfam	-	Pfam	Pfam, TIGRfam
Ontology	GO	GO	GO	GO	GO	-	-	GO	GO
Enzymes, pathways subsystems	KEGG, SEED	KEGG, SEED	KEGG,SEED	PRIAM	-	KEGG	KEGG,SEED	-	KEGG
Protein interactions	-	-	STRING	-	-	STRING	-	-	-
Motif database	InterPro	-	-	-	-	-	-	-	-
Types of prediction	Enzymes, Trans- porter classes	-	-	Enzymes, Transmem- brane helices, lipoprotein motifs	-	Protein networks	-	-	-
Reference	[91]	[90]	[44]	[92]	[96]	[93]	[36]	[95]	[94]

Previous accomplishments and future trends

The perfect scenario for a functional analysis of a community would be to have the individual genomes of every species in the sample and perform further analysis with high reliability. As said before, this is still not possible, and metagenomics studies are usually carried out over a mixture of short contigs and singletons (reads that could not be assembled).

It has been shown that short read lengths have a negative impact on the functional prediction [69, 101] since functional assignments with the same databases and parameters demonstrated discrepant levels of annotations for datasets with different lengths (e.g. ‘Cow Rumen’ metagenome with sequences of length ~ 100 bp [102] vs a Human Gut Japanese with > 1000 bp of mean sequence length [103]). This problem could be attenuated by increasing read-length using the 454 pyrosequencing platform, but the main choice continues to lie on the Illumina technology, due to its higher coverage and lower price.

Another problem is the lack of reference genomes in the databases to provide a more robust functional analysis. Metagenomics datasets harbor many unknown species, with specific functional role in the community context. Therefore, a relevant portion of the sequences will not be assigned to any function, due to the lack of homologue hits on the reference databases. Moreover, due to the low number of sequences from the less abundant species, their functional patterns are usually very difficult to obtain. These facts are evidenced in a comparison of different metagenomes sequenced with the Sanger method (longer reads), in which the annotation of bacterial communities ranged from 50-75 %, meaning that a significant fraction remained unannotated [101]. Single genome sequencing can be used to overcome this problem, as is the case of the HMP microbial reference genomes project [104]

The average genome size in an environmental sample can also affect the functional analysis of the metagenome [105]. It has been shown that differences in relative gene abundance across different metagenomics samples are biased by the average genome size of the environmental samples. Knowing the fact that larger genomes have high levels of novel genes over a small por-

tion of universal and housekeeping genes [106, 107], it is important to take into account their different average genome sizes by normalizing the metagenomics datasets, before inferring biological conclusions from the functional analysis.

Despite all these limitations, if the objective of a study is to analyze the abundance of gene families and perform a functional analysis at the single gene level, the existing methods are, somewhat, accurate enough.

2.3.3 Pathway inference of communities

The gene-pathway-centric view treats the community as a whole and ignores the exact assignment of a gene to a specific organism. This approach is consensual in the metagenomics community and some authors [108, 109] argued that it is possible to say that a metagenome is better characterized by its functional content than by its taxonomic composition, since several different species are able to perform similar biological functions.

If the goal is to analyze the functional content of a metagenome at the pathway level, different strategies are used and the occurrence of genes in pathways is taken into account. In contrast to the gene family abundances, in which many functional categories are found to be statistically different between different metagenomes [103, 108], metabolic pathway comparisons have a much smaller number of differences to distinguish, making data interpretation easier and providing stronger evidences of distinct functional capabilities [109, 110].

Pathway reconstruction lies in finding the most likely set of pathways in a metagenome. Due to the sequencing and annotation limitations, it is very rare to find in a sample all the genes that make up a pathway. Therefore, different approaches can be designed to address this issue. A naïve approach assumes that if a gene that is included in a pathway is present in the dataset, the whole pathway is also present and is scored accordingly. However, this assumption is hampered by the simple fact that genes are commonly present in multiple pathways, and thus the overall list of pathways will be inflated. On the other hand, a more conservative approach considers a pathway if all

its constituents are present in the sample [65].

Another concern to take into account is the different species abundances and coverage. The pathway abundances should be higher for pathways that are present in the most represented organisms of the community. On the other hand, as the diversity in the sample increases, the coverage of the genomes reduces. Therefore, solutions for adjusting pathways abundance must be taken to avoid overestimation based on these factors [65].

Methods for metabolic pathway identification

The KEGG database includes a collection of reference pathways that allows the mapping of annotated proteins for a given organism onto them. Given an annotation, K numbers are created, where each value of K represents an ortholog group of genes that are directly linked to a biochemical step in the KEGG pathway map. Then, it reconstructs pathways based on the assigned K numbers [111]. Similarly, the SEED subsystems can be used. For instance, the MG-RAST server [112] annotates the sequences in FIG families based on the FIGfam database, and then maps these protein families against the SEED subsystems to infer metabolic pathways (subsystems reconstruction, as it is called).

The PathoLogic module of BioCyc Pathway Tools [113] predicts metabolic pathways and operons (co-regulated bacterial genes of a metabolic pathway) based on a machine learning approach that uses MetaCyc [114] (manually curated database of metabolic pathways) as a reference database for the learning process. It takes as input an annotated genome (e.g. set of files in Genbank format) and achieves highly accurate predictions of pathway assignment on single genomes (>91%) [65].

MinPath [115] relies on a more conservative approach, where it finds the minimal set of pathways that can be explained with the supplied protein sequences. It is a parsimony method solved with integer programming which showed a significant reduction in the number of annotated pathways compared with the KEGG and SEED. These two methods may over estimate the number of pathways due to the existing redundancy: different pathways

may share the same biological functions and it is common to find pathways in these databases that are overlapping. Furthermore, some proteins are responsible for multiple functions (different domains, active sites, etc.). Moreover, these approaches may map one protein to multiple homologous proteins in different pathways, with different biological functions (paralogous proteins). As an example in the human genome, the ascorbate (vitamin C) pathway was detected by KEGG due to the presence of a protein that performs the same function in multiple pathways, but it is known that humans cannot synthesize vitamin C. MinPath removes these false assignments, and that is why the number of pathways is reduced with this approach [115]. Furthermore, it does not rely on training as PathoLogic does, so it may be more suitable for metagenomics datasets, since there is yet a long way to have a strong catalog of reference bacterial genomes with respect to the worldwide microbial diversity.

Sharon et al. (2011), [116] proposed two statistical models for pathway analysis that take into account gene length, pathway size and gene overlap: a pathway intersection method and an independent pathways method. Each one relies on two different assumptions about the sharing of genes among pathways. In the independent method, a gene that is shared among several pathways is assumed to have a copy for each pathway in which it appears. This model has shown to strengthen the counting of pathways for highly abundant pathways. The intersection method assumes that each gene present in more than one pathway appears once. This alternative seems better for the pathway abundance prediction on low abundant pathways. However, these models remain theoretical since no software has been distributed.

Tools for analyzing metagenomes in a pathway-based level

An ultimate goal of a gene-pathway-based functional analysis is to find which genes or pathways consistently explain the differences between two or more communities and this is done through statistical methods. ShotgunFunctionalizeR was developed in 2009, being an R package designed for functional comparison of metagenomes. Statistical analyses are performed with classic

binomial and hypergeometric tests, and with generalized linear models with a Poisson canonical logarithmic link [117].

Another comparative metagenomics package, STAMP [118] was developed to provide a stronger statistical analysis for metagenomics communities. It provides a graphical environment system and takes as input the functional and taxonomic profiles generated by MG-RAST and all abundance profiles available at IMG/M. It adds statistical features, such as the effect size (magnitude of the observed difference between samples) and confidence intervals (range of effect size values that have a probability of being compatible with the observed data) making STAMP a valuable tool for comparative metagenomics.

MetaPath [119] is a statistical tool for finding significant metabolic subnetworks from the global metabolic pathway. This global network comes from the network of KEGG reactions of a given sample (obtained from the annotation of the sequences against the KEGG genes database). Afterwards, a scoring step of the metabolic subnetworks is performed using Metastats [120] and a greedy search algorithm that takes into account the topology of the network is used to find the maximum weight subnetworks in the global network.

A very recent promising tool was developed to describe the functional profile of the communities, with special emphasis in human metagenomes. Its name is HUMAnN (HMP Unified Metabolic Analysis Network) [110] and the methodology has the particularity of performing a whole functional pipeline directly from the short unassembled reads. After a first filtering step, in which bad quality reads and human DNA are removed, the sequences are searched against the KEGG Orthology [121] using an accelerated version of the translated BLAST. Gene families' abundances are calculated by simply counting the number of reads associated with a function and for pathway inference it uses the MinPath approach, explained above. HUMAnN distinguishes from the others by some improvements that are added in the analysis: (i) unlikely pathways are removed based on taxonomic profiles from BLAST hits: pathways assigned from taxonomic units that are not identified in the sample; (ii) Gap filling step, to account for rare genes in abundant pathways. The final

outputs for each sample are coverage (presence/absence) and abundance values for KEGG modules and pathways. From these abundance values, further comparative metagenomics studies can be done.

Some of the already described pipelines for functional annotation (Table 2.2), such as the MG-RAST, IMG/M, MEGAN4, METAREP and webMGA, are also able to perform a functional analysis based on pathways (or subsystems). They mainly trust on SEED or KEGG systems to detect pathways from the annotated data and have their own statistical tests to execute comparative metagenomics (Figure 2.1).

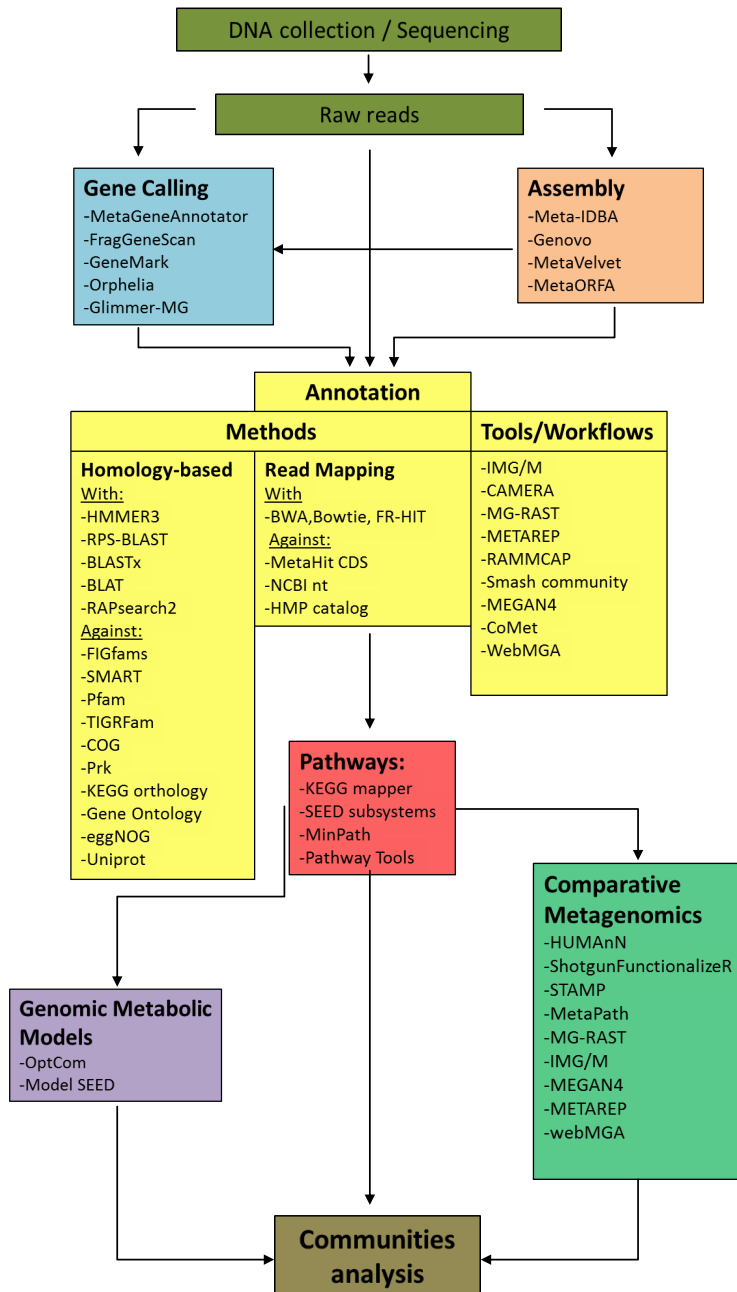


Figure 2.1: Flowchart of the main stages and available methods/tools for metagenomics pathway-based functional analysis.

Metabolic network reconstruction of metagenomes

Multiple metabolic models for single organisms have already been published [122] by integrating known metabolic reactions from databases such as KEGG or MetaCyc with stable annotations (e.g. Uniprot or BRENDA [123] databases). This information can be converted into a mathematical model, that can be analyzed through constraint-based approaches. Tools such as Model SEED [124] and *Merlin* [125] are able to re-construct the models. Then, some procedures can be applied to reduce its complexity and employ physiochemical constraints to find optimal metabolic states via flux balance analysis (FBA). To accomplish these tasks there are software platforms such as OptFlux [126] and the Matlab toolbox COBRA [127].

The use of these stoichiometric genome-scale metabolic models from different organisms has already been proposed [128–130]. Currently, the increasing interest in metabolic simulations of microbial communities is clear, as it is shown by the development of a framework for analyzing metagenomes through FBA, named OptCom [131]. This approach integrates both species and community-level fitness criteria into a multi-objective approach, and allows the assessment of the optimality level of growth for different members of the community (the descriptive mode), and subsequently making predictions regarding metabolic fluxes (the predictive mode).

A novel perspective of pathway and network inference is necessary to span a whole community and its respective interactions. New perspectives are coming from metagenomics and the definition of ‘super-meta-pathways’ has appeared, in which all the functions that make the system are included in the network, irrespectively of the species contributing to specific functions. This approach will reconstruct and model biochemical and regulatory pathways in complex symbiotic interactions, allowing us to have information about the end metabolite of a pathway in a given microorganism (or cell type, in humans) and how the same metabolite enters in a new microorganism (or cell type) to be used in some biological process [65].

Currently, the main limitation for the progress of pathway analysis of metagenomics data remains on assembly and gene function assignments, de-

spite the fast and constant improvements of the analysis tools. The necessary breakthrough for drastic improvements stands on sequencing technologies, by the substantial increase of the read-length. As longer reads become available, all the downstream analysis will be simplified and higher throughput will be achieved.

2.4 Human Microbiome Project

It was in an international meeting held in Paris, 2005, that the first discussion about this topic took place. After this meeting, the National Institute of Health (NIH) discussed the possibility of funding a wider project to study the human microbiome, by analyzing additional body parts not present in previous studies [132].

The HMP was then born, aiming to use high-throughput technologies progresses to fully describe the human microbiome by taking samples from several body sites of at least 250 healthy individuals. Testing different medical conditions, this community expected to use the obtained knowledge to address if there are associations between the changes in the microbiome and the diagnostic of a disease. It was also desired to provide a standardized data resource and develop new tools to enable this type of studies broadly in the scientific community. The ultimate goal of this project focused on demonstrating that it is possible to improve health by performing treatments based on the manipulation of the human microbiome [132].

2.4.1 How does it work?

The first phase of the project, named Jumpstart period, began in 2007. At this stage, there were three main goals that the Jumpstart funding supported. Firstly, sequencing 500 new bacterial genomes distributed along the human body to serve as a reference catalog for the subsequent metagenomic annotation and analysis that would be required later in the project. Secondly, the HMP aimed to develop and perform a sampling protocol at five body sites, the gastrointestinal tract, the mouth, the vagina, the skin,

and the nasal cavity (http://hmpdacc.org/micro_analysis/microbiome_analyses.php). Lastly, execute 16S rRNA gene sequencing in the above described body sites using the taken samples and the *Roche-454 FLX Titanium* sequencing platform [132].

The second phase of the project consisted on the improvement of the culture methods to sequence more reference genomes that were not available at that time, achieving a list of 1000 genomes that would be added in a public repository. In addition, all the sequencing centers involved in the HMP (The Baylor College of Medicine, The Broad Institute, The J. Craig Venter Institute and the Washington University School of Medicine) started at this stage to sequence the genomes of viruses and eukaryotic microbes found in the human microbiome and performed the Whole-Genome Shotgun (WGS) sequencing of the 250 individuals sampled in the Jumpstart phase, which produced the metagenomic samples that aimed to characterize the microbiome. Furthermore, one of the main issues of the HMP was addressed at this stage, by the initiation of the HMP Demonstration Projects, which aimed to study the changes in the microbiome that are related to human health and disease, by starting with 15 pilot projects associated with several medical conditions [132].

The Data Analysis and Coordination Center (DACC) (<http://hmpdacc.org/>) was created to store all the generated sequence information from WGS, 16S and reference genome sequencing. Here, it is possible to access all the information related to the project, from the developed software in the course of the HMP to the news, meetings and publications regarding this topic.

In short, the HMP focused in three topics: producing reference genomes, 16S rRNA sequencing and metagenomics sequencing of whole community (WGS). The reference genomes catalog helps on the analysis of WGS data, by enabling the alignment of the metagenomics reads or assembled sequences to the microbial reference genomes. On the other hand, the 16S rRNA sequencing aims to make a taxonomic classification and perform a phylogenetic analysis of the microbiome species. Lastly, the metagenomics sequencing enables, besides calculating organism abundance, to perform a functional annotation of the sequences and infer the metabolic pathways present in the

community, taking a gene-centric view [133] rather than an exact assignment of genes to individual organisms.

2.4.2 Bioinformatics for the HMP

A large amount of data was generated in this project using sequencing technologies. The HMP has released over 100 million 16S rRNA reads and more than 8 Tera base pairs (Tb) of metagenomics sequences [134]. Computational methods were required to deal with this data and extract useful information (Figure 2.1).

Regarding 16s sequencing, a 16S rRNA curation pipeline was developed to reduce the error rates in the individual base calls [135]. Two developed HMP-funded software, *mothur* [136] and *QIIME* [137] use implementations of that pipeline for microbial community taxonomic screening.

The HMP conducted an extensive metagenomics sequencing survey in which 764 samples from 16 body sites were sequenced using the *Illumina GAIIx* platform with 101 bp paired-end reads [134]. Contamination of the samples with human DNA was a concern, thus a human DNA removal step [138] and quality control test was required to speed up and avoid a mislead analysis of the data.

Proceeding with the treatment of metagenomics data, comes up the assembling process comes up. Initial HMP assemblies showed poor results, due to genomic variations between closely related species and the mistake of confusing high abundant organisms with genomic repeats, making the assembling largely fragmented [139]. At the end, no specific tool was developed to perform the assembling of the HMP shotgun data. Instead, an assembling strategy was applied around the *SOAPdenovo* assembler [140] (http://hmpdacc.org/doc/HMP_Assembly_SOP.pdf).

Despite all the efforts of doing an efficient assembly pipeline for metagenomics datasets, the question about the feasibility of assembling hundreds of metagenomes for the HMP was raised, considering the actual limitations of assembling even a single organism alone. Thus, the opportunity of a subsequent analysis pipeline using unassembled reads it was discussed. In spite of

the obvious limitation of the read length in this strategy, it has been shown that tasks such as identification of organisms, community annotation and present pathways on the sample could be addressed using this approach [110, 139] (despite a high level of uncharacterized reads [141]), complementing the 16S rRNA method and gene annotations based on assembled datasets.

A list of all software and online resources associated to the HMP, ranging from the Microbial Reference Genomes methodology to the sampling and analysis of 16S rRNA and WGS can be found at (http://hmpdacc.org/tools_protocols/tools_protocols.php).

2.4.3 First achievements and future work

The first results of this big consortium confirmed the same tendencies as the previous individual studies: each body site owns dominant signature taxa [108, 142, 143]. For instance, *Lactobacillus* is dominant on the vagina, *Bacteroidetes* and *Firmicutes* are abundant in the gut and *Streptococcus* in the oral cavity. Curiously, actively pathogenic species were barely present in the microbial communities of the sampled individuals. On the other hand, the functional pathways derived from metagenomics data show much more stable abundance across the different body habitat than the microbes abundance. [141].

The large amounts of data produced from different body sites and the tools and protocols developed to analyze these data, allowed for the first time a deeper understanding of the human microbiome, both in microorganism composition and in metabolism. Bioinformatics resources need to be continuously improved, so that the analysis of the data represents a closer view of the reality (e.g. metagenomics assembly [26, 139], community pathway inference [110, 141]).

Finally, new microbiome studies will arise, and high-throughput methodologies will appear to address advanced questions such as exchanges between the microbial communities, and between microbes and the host [144]. Moreover, an integration of data from different assays of the human microbiome has already started [145–147], anticipating a bright future on this area, so

that the HMP appears as the first established resource for the human microbiome research and a big step forward on the relation of the Bioinformatics and human health.

2.5 Merlin

Merlin is an in-house-developed software, which performs semi-automatic annotations and constructs draft metabolic models for a single organism given a set of genes [148]. Since this framework is the basis for this project, a detailed description of the *Merlin* methodology will be provided next.

Currently, the software stands on version 2.0 and is available at <http://www.merlin-sysbio.org/>. *Merlin* is an open-source application implemented in *JavaTM* and was built on top of the AIBench (<http://www.aibench.org>) software development framework [149]. It utilizes a relational MySQL database to locally store the data and uses different Java libraries such as BioJava [150], NCBI Entrez Utilities Web Service Java Application Programming Interface (API), UniProtJAPI [151] and KEGG Representational State Transfer (REST) API to access several web services.

Merlin addresses two main objectives: the re-annotation and the genome-scale metabolic reconstruction. The first purpose is based in a similarity-based approach and aims to assign functions to genes that encode enzymes or transporter proteins (skipping regulatory and other genes), the main gene categories involved in metabolism. The second part allows creating a network representation of the metabolic reactions catalyzed by the organism. This reaction set can then be used to simulate *in silico* the phenotype of the organism under several environmental or genetic conditions.

2.5.1 Identification of genes that encode enzymes

The first step of the annotation process is the identification of genes that specifically encode enzymes. Starting from a genome in the FASTA format, it looks for the best homologues using BLASTp, BLASTx or HMMER in databases such as the NCBI non-redundant (nr) database or Uniprot.It

saves the homologues information for every gene: the query sequence, locus identifiers, e-values, scores and organisms. Since it is difficult retrieve enzymatic information from all identified homologue genes of the BLAST output, *Merlin* implements remote similarity alignments to collect information about each of the homologues identified for every gene. The data is retrieved remotely from the *Entrez* Protein database (for each homologue: Taxonomy, Organelle, Locus Tag, Enzyme Commission (EC) number, Product, Molecular weight).

A candidate annotation for each protein is selected based on confidence scores. The scores for each homologue are calculated based on two criteria: (i) the frequency that a given function (EC number) appears in the set of homologues; (ii) the taxonomy, which refers to the level of proximity between the input organism and those in which a function has been found in the homology search. The similarity result with the highest confidence score is selected (gene product, EC number).

After annotating each candidate EC number, the use of a manually curated enzyme database aims to make the annotations more accurate and strengthen the results. BRENDA verifies the function about to be annotated for some genes (e.g. genes with different enzyme assignments in two different similarity searches (NCBI nr and Uniprot), enzymes encoded with partial EC number).

2.5.2 Identification of genes that encode transporter proteins and compartments prediction

Since enzyme transporter proteins cannot be directly classified from homology searches over regular protein databases, they are obtained by performing local alignments using the Smith-Waterman algorithm against the Transport Classification Database (TCDB) [152] to identify the number of genes that encode transporter proteins. This algorithm is very time consuming, so the number of genes to align against TCDB must be reduced. For this purpose, the TMHMM software [153] is incorporated to predict which genes encode transmembrane proteins, and therefore can be related to transport functions.

The ones that have one or more transmembrane helices are considered protein candidates, and those are aligned against TCDB.

For each transporter candidate, Merlin performs the local alignments in the TCDB database and classifies the Transport Classification (TC) family numbers and metabolites associated, in the same way as it is performed for EC numbers. This is done using the taxonomy of each of their TCDB homologue genes and the frequency of the TC family numbers or metabolites within all similar genes.

The prediction of the proteins and metabolites subcellular localization is performed with WoLF PSORT for eukaryotes and PSORTbv3.0 for prokaryotes. For each gene, a main compartment prediction is automatically assigned by these programs along with a secondary compartment if it scores accordingly. To annotate transport systems, besides having transmembrane domains and similarities to TCDB records, the candidate protein must have a localization prediction within a membrane.

2.5.3 Metabolic reconstruction

The construction of the metabolic model starts with the construction of a local MySQL database. Several KEGG data files (with information of reactions, enzymes, organism, etc.) are loaded, which allows to the user, through the *Merlin's Views* operation, to later assemble a genome-scale model, selecting and editing reactions and parameters to be included in the model.

The *Merlin's Integrate* option compares the enzyme information retrieved by similarity with the data already available in the local database. In case of conflict between these data, the user can select which information should be automatically preferred or if the data should be merged.

Lastly, the *Merlin's SBML Builder* operation allows the user to export the model, currently stored in a relational database, to the System Biology Markup Language (SBML) format. This feature allows the user to employ the model for *in silico* simulations in other software applications very easily. A representation of the *Merlin* operation mode is depicted in Figure 2.2.

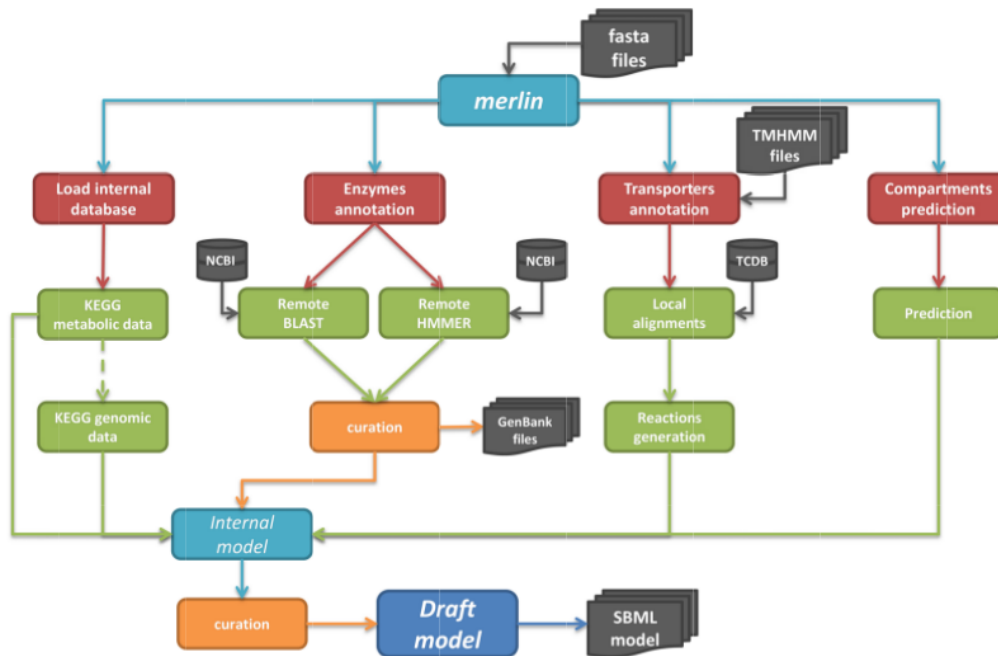


Figure 2.2: Schematic representation of *Merlin* architecture (Figure extracted from [154])

Chapter 3

Methodology and algorithms

Since the underlying methods for annotation incorporated in *Merlin* are based in similarity searches against reference databases and the functional reactions-pathways information is extracted from KEGG, these methodologies were kept for this work. However, there was the need to update *Merlin* to be able to deal with multi-organism data, so a set of changes were performed in the software before the implementation of the algorithms concerning metagenomics itself, that will be described firstly.

3.1 New complementary features in *Merlin*

One of the main issues regarding any metagenomic study is the computational time that the annotation of the reads/genes takes, using as reference databases such as NCBI-nr or UniprotKB. Given the high number of fragments to search in any metagenomic dataset, this task might become a major problem for the user. As an example, *Merlin* v2.0 took more than one month to run an annotation of ≈ 47000 predicted genes from a buccal mucosa sample downloaded from the HMP repository (sample SRS013711), using a desktop computer with an Intel®Core™ 2 Quad CPU Q6600 @ 2.40GHz four processor cores and 4G of RAM. Therefore, some improvements were done towards a more efficient process.

3.1.1 Database management

First, an adjustment in the database schema related to enzymes homology search was performed, to ensure a better data organization and storage (see Figures A.1 and A.2 in Appendix A). These changes in the database structure required the development of methods for keeping old projects still available in this new *Merlin* internal database structure. These methods, written in Java use hash tables to gather all information kept in the old homology schema and load it into a new structure database without losing information. This utility was not added into the *Merlin* graphical user interface (GUI) since this new software release comes already with the new structure incorporated.

Another feature to speed up the annotation was conceived regarding the parallelization of the BLAST/HMMER through several machines. Since the input file in *Merlin* for metagenomics datasets is composed of a high number of predicted genes, it can be useful for the user to split it into several files and run the similarity searches for each of them in different computers with *Merlin* installed. Once this is done, the user may want to gather all the annotations again into one project with its respective database. Next, this process is described in detail:

1. The user saves a backup of the database to export to another computer in the '*Database*' menu and '*Save Database Backup*' option (Figure 3.1) (Note that this option is also helpful to avoid loss of information in case of any problem with the project or the database).
2. The outputted file, that comes in the Structured Query Language (SQL) format, is copied into the computer who hosts the database that will merge the results and the user can create a new project using the newly created database from that file (in the '*Database*' menu and '*New Database from SQL file*' as shown Figure 3.2).
3. Given two projects, the user can merge their databases in the '*Database*' menu in the '*Merge databases*' option (Figure 3.3) being the annotation of all genes of the metagenomics sample together in the same project.

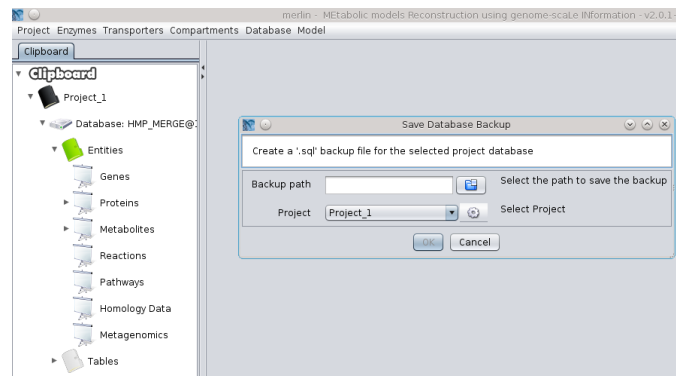


Figure 3.1: *Merlin*'s view for saving a backup of the project database.

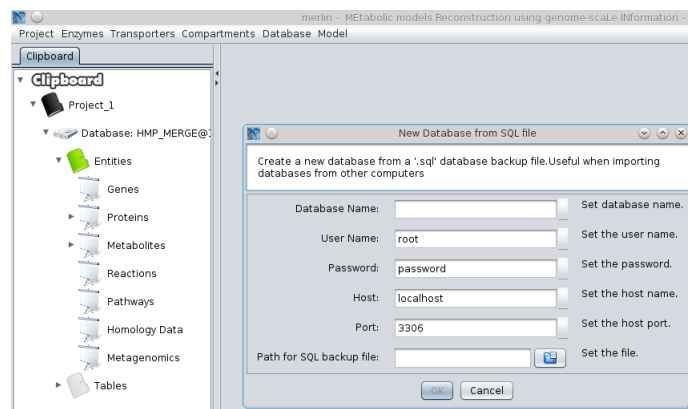


Figure 3.2: *Merlin*'s view for creating a new database from a backup SQL file.

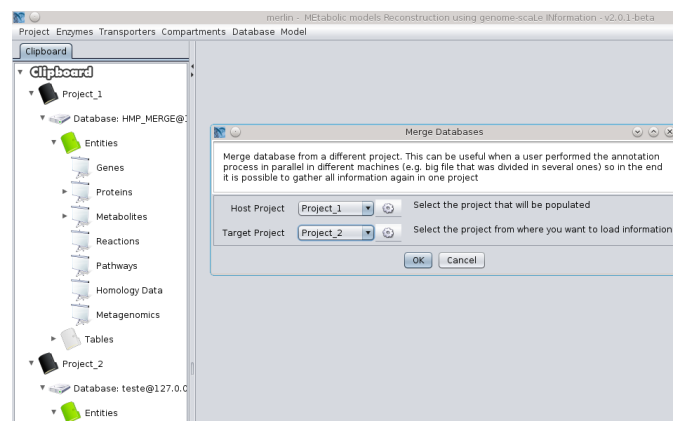


Figure 3.3: *Merlin*'s new 'Merge databases' view.

3.1.2 Enzymes annotation

Merlin uses a routine to assign EC numbers to each gene g . These are done by assigning a weight to the number of times each EC number ec is found within the list of homologues for each gene (frequency), as well as to the taxonomy of the organisms to which such homologues belong to [154]. The equation 3.1 describes how the scoring process is done, where the influence of the frequency score ($Score_f$) and the taxonomy score ($Score_t$) depend on an α parameter that controls the weight to give to each for the overall score:

$$Score_{ec}^g = \alpha \times Score_f + (1 - \alpha) \times Score_t \quad (3.1)$$

For single genome projects the default value of α is 0.5, meaning that the same weight is given to the frequency and taxonomy score. Since the taxonomy score is used to favor homologies with records of closely related taxonomies to the organism being studied, this score does not make sense for metagenomics projects, where the focus stands on the whole community instead of a single organism.

Therefore, for this work the α default value was changed to 1, making the EC number scores calculations only based on the frequency score. This score counts the number of occurrences of an EC number ec within all homologues of that gene, and divides this value by the total number of homologous genes n , as described in equation 3.2.

$$Score_f^{ec} = \sum_{i=1}^n \frac{\nu_{ec_i}}{n} \quad (3.2)$$

where:

$$\nu_{ec_i} = \begin{cases} 1, & \text{if } ec \text{ exists in record } i \\ 0, & \text{otherwise} \end{cases}$$

The score will always have a numeric value between 0 and 1. A minimum score threshold is defined to automatically accept the annotations. For single genome projects this value is set by default to 0.5, which means that scores

smaller than this value will not be annotated, despite the possibility of the user to manually curate it and accept the result.

In metagenomics, given the high amount of genes to be annotated, it is typically not feasible for the user to manually check and curate enzymes annotation, so all the enzymes in the metagenome will normally be automatically processed. Since metagenomics harbors a massive amount of genes that are poorly characterized, it is expected that in some cases, the assigned EC numbers have low confidence scores. Given these facts, the default minimum score threshold for metagenomics projects was set to 0.3 (the user is still able to set the value to a more fitted one for each specific project).

3.1.3 Uniprot requests

In the *Merlin*'s previous version, a query of each EC number candidate to Uniprot is performed using the genus locus identifier (locus tag) to access the existence of a reviewed annotated record for such gene. For single genome projects, this feature is useful since it allows the user to have a degree of confidence for each EC number annotation (if a gene has a reviewed match on Uniprot, the EC number is likely to be well annotated), but for metagenomics projects this is useless.

Since, in metagenomics, genes inputted in *Merlin* come from gene prediction softwares, and therefore putative Open Reading Frames (ORFs) are generated, the gene identifiers will never have a cross-reference to Uniprot, thus making this Uniprot operation worthless. In addition, the Uniprot servers take a high amount of time to answer the requests, slowing down significantly the genome annotation process. Therefore, this step was turned off in metagenomics projects, which led to an evident reduction in the annotation time.

3.1.4 Implementation of Local Blast

Yet concerning the annotation, a local Blast version was implemented to enable some speed improvements in big samples, such as the HMP ones. Since the BLAST output only provides a list of the homologues with their

respective score and e-value for each gene, it is necessary to retrieve more detailed information of each homologue to fill the *Merlin*'s database tables. Regarding this task, a parser of the Uniprot database was developed, which means that for now, it is only possible to execute this operation against that database.

For that, the user must download the reference database (either UniProtKB/ Swiss-Prot, UniprotKB/TrEMBL or both) from the <ftp.uniprot.org> website, as well as the corresponding text file (www.uniprot.org/downloads). Also, a local version of the Blast has to be installed in the machine and the program needs to be added to the environmental variable path. Blast can be downloaded from <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/>.

After the required configuration of the Blast and local databases, *Merlin* is able to perform the operation (Figure 3.4). First, the program blasts the genes against the local database and the results are stored in a temporary directory. Then, the retrieval of homologues information is done by parsing the text file that contains all the information of each record in the database. For each homologue the following fields are saved: UniprotID, Uniprot status, Full name, Ecnnumber, Gene name, Organism, Full taxonomy and Sequence. Afterwards, all this information is loaded into the project database.

Concerning the local Blast operation view, that allows configuring this process (3.4), the first option refers to the Blast program used. In this case, the BLASTp is used, since the predicted ORFs come as amino acid sequences. The user can then choose the local database to perform the annotation, either the SwissProt or the whole UniprotKB (merge of the UniProtKB/Swiss-Prot and the UniProtKB/TrEMBL databases). The user may have to choose a trade-off between sensitivity and time. If SwissProt is selected, which is a small database, the annotation can be very fast despite the loss of information. On the other hand, if the user chooses to use the entire Uniprot (huge database), some problems with the computation times will may occur. However the user will gain on sensibility, i.e more results will be reached in the annotation.

The user has to select the directory where the local database was cre-

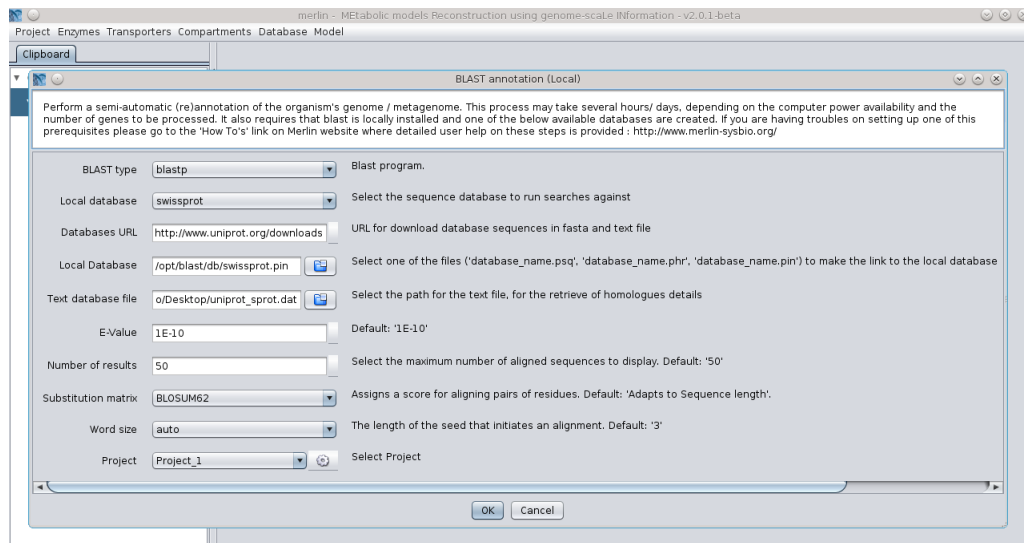


Figure 3.4: *Merlin*'s new Local Blast operation.

ated and the database file for the homologues information retrieval explained above. The normal BLAST parameters, such as minimum e-value, number of results and substitution matrix can also be specified by the user.

3.2 Metagenomics pipeline

3.2.1 Taxonomy inference

The methodology for taxonomic characterization developed in this work relies on a *supervised* approach using similarity search methods, more specifically, the BLAST or HMMER tools. Because *Merlin* has incorporated remote searches using those tools to annotate the genomes, it was easier to develop independent methods for taxonomy classification based on these methods.

Specifically, the purpose of this operation added to *Merlin* is to assign a taxonomic label to each gene, as well as to describe the overall community composition. Thus, it classifies each gene at the phylum and genus level based on the list of homologues obtained from the homology search. Afterwards, given a classification for each gene, it calculates the proportions of each taxon in the whole set of genes.

A routine was developed to assign a phylum and genus to each gene. The assignments are performed giving a weight to the number of times each phylum and genus are found within the gene homologues list. A gene is only classified if it contains a minimum number of homologues and the phylum and genus scores are higher than two defined thresholds.

Therefore, for each gene g , given an ordered set of N homologues (higher than k , the minimum number of homologues allowed), the phylum scores ($Score_{phylum}$) is calculated according the equation 3.3. The ideal value for $Score_{phylum}$ is 1, which would mean that every homologue in the set is of the same phylum.

$$Score_{phylum}^g = \frac{Score_{bestph}}{Score_{max}} \quad (3.3)$$

The best phylum score ($Score_{bestph}$) for each gene represents the phylum candidate from the whole list of homologues that achieved the best score to be tested (equation 3.4). Theoretically, it would be expected that the phylum that occurs most in the list homologues would be the one selected as the candidate. However, it was decided to privilege the first five homology hits, since those are likely to be more taxonomically related to the gene being tested:

$$Score_{bestph}^g = \sum_{i=1}^N f_{bestph_i} \quad (3.4)$$

where:

$$f_{bestph_i} = \begin{cases} 2.0, & \text{if homologue in position } i \text{ belongs to phylum } bestph \text{ and } i \leq 5 \\ 0.5, & \text{if homologue in position } i \text{ belongs to phylum } bestph \text{ and } i > 5 \\ 0, & \text{otherwise} \end{cases}$$

In the cases where two or more candidate phyla have the same score, the routine chooses the taxon that comes first in the homologues list as the best candidate for a given gene.

The maximum score ($Score_{max}$) shown in equation 3.5 represents the highest possible score for gene g given the number of homologues:

$$Score_{max}^g = \sum_{i=1}^N f_i \quad (3.5)$$

where:

$$f_i = \begin{cases} 2.0, & \text{if } i \leq 5 \\ 0.5, & \text{otherwise} \end{cases}$$

This routine explained above describes the methodology developed in this work for phylum scoring. For genus assignment, the procedure is exactly the same, with exception for the minimum score threshold. In the end, a gene will be assigned with a taxonomic label only if it fulfills the following criteria:

- The number of homologues is higher than the minimum number required k (default value is 5), otherwise the routine will not even perform the phylum/genus scores calculations.
- The phylum score $Score_{phylum}^g$ is higher than the phylum threshold (default value is 0.5).
- The genus score $Score_{genus}^g$ is higher than the genus threshold (default value is 0.3).
- The phylum and genus are congruent, that is, the selected genus belongs to the selected phylum.

3.2.2 *Merlin's* operation mode for taxonomy

The described methodology was integrated in *Merlin* in a user-friendly interface as depicted in Figure 3.5. This visualization panel is accessed by clicking on the '*Taxonomy*' sub-view under the '*Metagenomics*' entity. It comprises taxonomic information for all genes, including the selected phylum and genus for each gene and their scores. Information for those genes

The screenshot shows the Merlin software interface with the 'Taxonomy' view selected. The main window displays a table of gene taxonomic assignments. Below the table are search and parameter settings.

Info	Genes	Phylum	Phylum Score	Genus	Genus Score	Phylum/Genus are concordant
	gene_id_1006	Bacteroidetes/Chlorobi gr...	1.0	Prevotella	1.0	true
	gene_id_10060	No homologues enough				
	gene_id_10061	Proteobacteria	1.0	Haemophilus	0.6769	true
	gene_id_10062	Proteobacteria	1.0	Haemophilus	0.7344	true
	gene_id_10063	Bacteroidetes/Chlorobi gr...	0.9846	Prevotella	0.8154	true
	gene_id_10064	Bacteroidetes/Chlorobi gr...	1.0	Prevotella	0.8333	true
	gene_id_10065	Bacteroidetes/Chlorobi gr...	1.0	Prevotella	0.8154	true
	gene_id_10066	Bacteroidetes/Chlorobi gr...	1.0	Prevotella	0.6308	true
	gene_id_10067	Proteobacteria	0.9846	Aggregatibacter	0.2154(< 0.3)	no minimum score
	gene_id_10068	Firmicutes	1.0	Veillonella	0.4308	true
	gene_id_10069	Firmicutes	1.0	Veillonella	0.7949	true
	gene_id_1007	Bacteroidetes/Chlorobi gr...	1.0	Prevotella	0.7231	true
	gene_id_10070	Bacteroidetes/Chlorobi gr...	0.975	Prevotella	0.9	true
	gene_id_10071	Firmicutes	1.0	Streptococcus	1.0	true
	gene_id_10072	Proteobacteria	1.0	Haemophilus	0.3692	true
	gene_id_10073	Firmicutes	0.6923	Campylobacter	0.3077	false
	gene_id_10074	Firmicutes	1.0	Lachnoanaerobaculum	0.381	true
	gene_id_10075	No homologues enough				
	gene_id_10076	No homologues enough				
	gene_id_10077	Proteobacteria	1.0	Haemophilus	0.6462	true

Below the table, there is a 'Search' section with a text input field and a 'Name' dropdown. Below that are three sections: 'Set parameters' with input fields for 'Minimum number of homologues' (5), 'Minimum phylum score' (0.5), and 'Minimum genus score' (0.3); 'Recalculate' with a 'Taxonomic composition' button; and 'Export' with a 'Text file' button.

Figure 3.5: Merlin's view for taxonomy information of metagenomic datasets.

that do not present a minimum number of homologues, the scores are lower than the thresholds and the phylum/genus are incongruent is also provided.

As the above figure shows, the user can easily change the values for some parameters of the scoring algorithm: the '*Minimum number of homologues*' text box allows to set a new value for the required number of homologues that each gene must have for the algorithm to perform the calculations; the '*Minimum phylum score*' and '*Minimum genus score*' text boxes can be altered to set the minimum scores for which a gene gets a valid taxonomic assignment. The scores can be re-calculated and the main table updated by clicking on the '*Taxonomic composition*' button.

The '*info*' column provides a button for each gene that allows to access detailed information about the phylum/genus scores for a given gene by showing all taxonomic elements used for that specific classification in a pop-up window (Figure 3.6).

By clicking on the '*Entity view*' tab in the panel, the user is able to see the main statistics of the scoring algorithm (e.g number of genes that did not achieve the minimum scores) as well as the overall community composition, where the percentages of the phylum and genus are discriminated (Figure

Phylum	Scores
Bacteroidetes/Chlorobi group	0.7230769230769231
Firmicutes	0.27692307692307694

Genus	Scores
Prevotella	0.676923076923077
Thermoanaerobacter	0.12307692307692308
Peptostreptococcaceae	0.09230769230769231
Clostridium	0.06153846153846154
Alloprevotella	0.03076923076923077
Barnesiella	0.015384615384615385

Figure 3.6: Detail windows for the 'Taxonomy' main view (Phylum scores on the left, Genus scores on the right).

3.7). This information can also be exported to a text file by clicking on the 'text file' button in the main window.

Taxonomy data	
Total number of genes	81278
Number of genes with no enough homologues(< 5)	30360
Number of genes that were actually included in the taxonomic composition inference	50918
Number of genes with no minimum score achieved (either on phylum or genus level) to be included in the calculations	7506
Number of genes with minimum score achieved but uncharacterized	0
Number of genes with minimum score achieved	43324
Number of genes with minimum score achieved but with no concordant phylum and genus assignments	334
Number of genes with minimum score achieved and with concordant phylum and genus assignments	42990
Number of genes that were used at the end for the taxonomic description of the community	42990
Domain:	
Bacteria	99.80693184461504
Eukaryota	0.0418702023726448
Archaea	0.0023261223540358223
Others:	
Viruses	0.14887183065829263
Phylum:	
Bacteroidetes/Chlorobi group	39.7534310304722
Firmicutes	38.96952779716213
Proteobacteria	14.440567573854384
Fusobacteria	5.531518957897186
Actinobacteria	0.9444056757385438
dsDNA viruses, no RNA stage	0.1302628518260605
unclassified Bacteria	0.07676203768318214
Spirochaetes	0.0721097929751105
Metazoa	0.0209351011863224
ssRNA viruses	0.016282856478250757
Cyanobacteria	0.009304489416143289
Streptophyta	0.009304489416143289
Fungi	0.00453341700071645

Figure 3.7: Statistics and overall community composition displayed in *Merlin*.

3.2.3 Metagenomics functional characterization

This routine aims to characterize the metabolic functions that are present in the metagenome and assign an abundance according to the number of times each of them is encoded in the whole set of genes. Furthermore, it tries to associate the taxonomic genus that encodes each enzyme. This operation is highly dependent on the annotation and the taxonomy, thus it is desirable that these previous steps work well to get better results.

From the set of all enzymes loaded from KEGG, this procedure selects the ones with a complete EC number that have at least one annotated gene in the metagenome, being automatically assumed that those enzymes are present in the community. Regarding the enzyme abundance calculation *Abundance*, given the set of all N annotated metabolic genes (Ω_n) and a set of T genes (Ω_t) encoding for the EC number ec , the routine calculates enzyme abundance according to the equation 3.6:

$$Abundance_{ec} = \frac{|\Omega_{t(ec)}|}{|\Omega_n|} \quad (3.6)$$

Afterwards, this method checks on the genes encoding an EC number if they have a taxonomic genus assignment from the previous taxonomy routine. If so, it is assumed that a specific genus encodes that EC number in the microbial community. On the other hand, if none of the genes encoding an EC number has a genus assignment (e.g. no minimum number of homologues, no minimum score), that enzyme is treated as present but with no taxonomic information regarding it.

3.2.4 *Merlin's* operation mode for enzymes

The visualization panel for metagenomics enzymes was integrated in *Merlin* in the '*Enzymes*' sub-view under the '*Metagenomics*' entity. The main view encompasses information of all encoded enzymes in the dataset. For each enzyme, the number of genes encoding it, the underlying reactions and their abundances are displayed. Furthermore, information is provided concerning the number of genes with taxonomic genera and the number of genes without

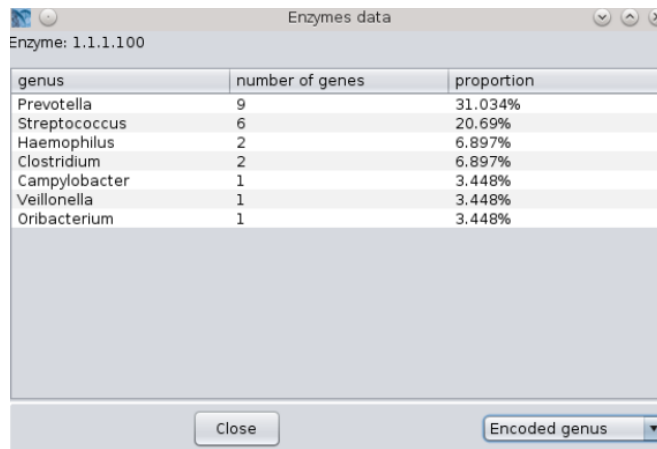
them (Figure 3.8). Since taxonomic information is required to execute this operation, the user must execute the taxonomy routine first by clicking on the '*Taxonomy*' sub-view described earlier, otherwise *Merlin* will throw an error.

Info	Names	ECnumber	Nº of reacti...	Nº of genes	Abundance	Nº of genes encoded by genus	Nº of genes with no genus
	alcohol:NAD+ ...	1.1.1.1	19	3	4.0E-4	No genus minimum score	3 (100.0%)
	(3R)-3-hydroxy...	1.1.1.100	11	29	0.0035	22 (75.862%)	7 (24.138%)
	dUDP-6-deoxy...	1.1.1.133	1	17	0.0021	16 (94.118%)	1 (5.882%)
	(S)-3-hydroxyb...	1.1.1.157	3	4	5.0E-4	2 (50.0%)	2 (50.0%)
	Transferred to...	1.1.1.158	0	30	0.0036	22 (73.333%)	8 (26.667%)
	7alpha-hydrox...	1.1.1.159	1	1	1.0E-4	1 (100.0%)	0 (0.0%)
	(R)-pantoate:...	1.1.1.169	1	3	4.0E-4	2 (66.667%)	1 (33.333%)
	5-amino-6-(5...	1.1.1.193	1	12	0.0015	11 (91.667%)	1 (8.333%)
	IMP:NAD+ oxid...	1.1.1.205	2	42	0.0051	37 (88.095%)	5 (11.905%)
	morphine:NAD...	1.1.1.218	2	1	1.0E-4	1 (100.0%)	0 (0.0%)
	UDP-alpha-D-g...	1.1.1.22	1	5	6.0E-4	4 (80.0%)	1 (20.0%)
	L-histidinol:NA...	1.1.1.23	3	10	0.0012	9 (90.0%)	1 (10.0%)
	shikimate:NAD...	1.1.1.25	1	15	0.0018	12 (80.0%)	3 (20.0%)
	4-phosphono...	1.1.1.262	3	5	6.0E-4	4 (80.0%)	1 (20.0%)
	2-C-methyl-D-e...	1.1.1.267	1	26	0.0032	24 (92.308%)	2 (7.692%)
	(S)-lactate:NA...	1.1.1.27	3	4	5.0E-4	4 (100.0%)	0 (0.0%)
	GDP-beta-L-fu...	1.1.1.271	3	7	8.0E-4	7 (100.0%)	0 (0.0%)
	(R)-2-hydroxyc...	1.1.1.272	5	1	1.0E-4	1 (100.0%)	0 (0.0%)
	(R)-lactate:NA...	1.1.1.28	1	1	1.0E-4	1 (100.0%)	0 (0.0%)
	D-glycerate:NA...	1.1.1.29	2	4	5.0E-4	3 (75.0%)	1 (25.0%)
	L-homoserine:...	1.1.1.3	2	15	0.0018	12 (80.0%)	3 (20.0%)
	(R)-3-hydroxy...	1.1.1.36	2	1	1.0E-4	No genus minimum score	1 (100.0%)
	(S)-malate:NA...	1.1.1.37	1	5	6.0E-4	3 (60.0%)	2 (40.0%)
	(S)-malate:NA...	1.1.1.38	2	14	0.0017	11 (78.571%)	3 (21.429%)
	(R,R)-butane...	1.1.1.4	1	1	1.0E-4	1 (100.0%)	0 (0.0%)
	(S)-malate:NA...	1.1.1.40	2	17	0.0021	17 (100.0%)	0 (0.0%)
	isocitrate:NAD...	1.1.1.41	1	1	1.0E-4	No genus minimum score	1 (100.0%)
	isocitrate:NAD...	1.1.1.42	3	3	4.0E-4	2 (66.667%)	1 (33.333%)

Figure 3.8: *Merlin*'s view for metagenomic enzymes information.

The '*info*' column provides detailed information about the selected enzyme in three different ways: in a table, the most important one, it shows the genera encoding the enzyme (Figure 3.9); another table displays the genes (locus tag) encoding the enzyme as well as their taxonomic assignments; the last one exhibits the reactions assigned to the selected enzyme (Figure 3.10)

The '*Enzymes coverage*' button allows the user to export the enzymes coverage (presence/absence) to a tab-separated text file. In the '*Entity view*' tab in the down side of the panel the main statistics of the enzymes are shown (e.g number of enzymes from each class)(Figure 3.11).

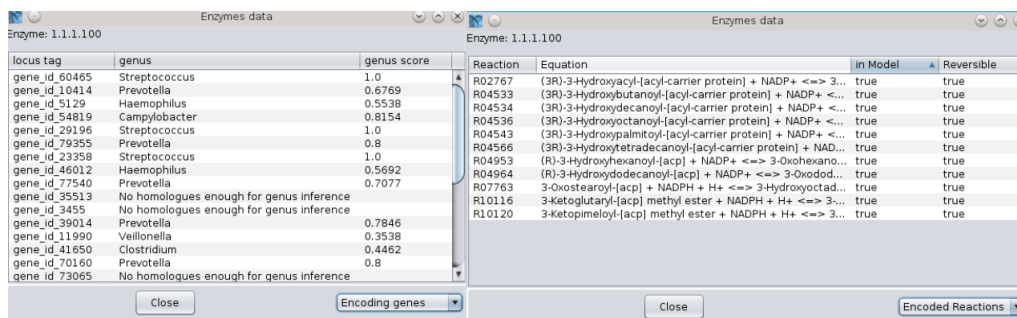


Enzyme: 1.1.1.100

genus	number of genes	proportion
Prevotella	9	31.034%
Streptococcus	6	20.69%
Haemophilus	2	6.897%
Clostridium	2	6.897%
Campylobacter	1	3.448%
Veillonella	1	3.448%
Oribacterium	1	3.448%

Buttons: Close, Encoded genus

Figure 3.9: *Merlin's* detailed information for different genera encoding a selected enzyme.

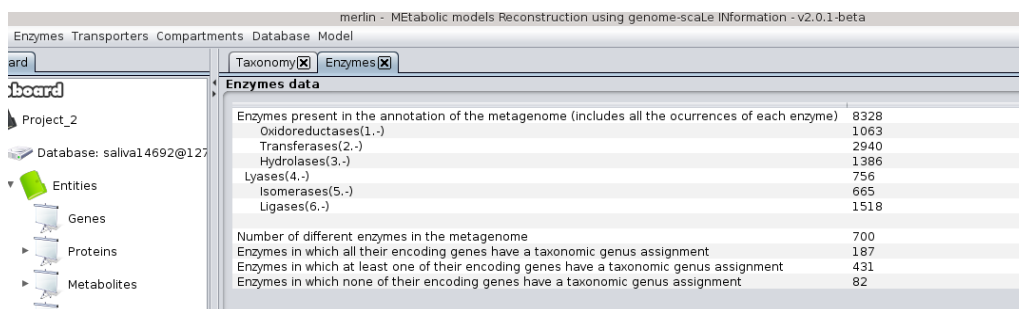


Enzyme: 1.1.1.100

locus tag	genus	genus score	Reaction	Equation	in Model	Reversible
gene_id_60465	Streptococcus	1.0	R02767	(3R)-3-Hydroxyacyl-[acyl-carrier protein] + NADP+ <=> 3...	true	true
gene_id_10414	Prevotella	0.6769	R04533	(3R)-3-Hydroxybutanoyl-[acyl-carrier protein] + NADP+ <...	true	true
gene_id_5129	Haemophilus	0.5538	R04534	(3R)-3-Hydroxydecanoyl-[acyl-carrier protein] + NADP+ <...	true	true
gene_id_54819	Campylobacter	0.8154	R04536	(3R)-3-Hydroxyoctanoyl-[acyl-carrier protein] + NADP+ <...	true	true
gene_id_29196	Streptococcus	1.0	R04543	(3R)-3-Hydroxypalmitoyl-[acyl-carrier protein] + NADP+ <...	true	true
gene_id_79355	Prevotella	0.8	R04566	(3R)-3-Hydroxytetradecanoyl-[acyl-carrier protein] + NAD...	true	true
gene_id_23358	Streptococcus	1.0	R04953	(R)-3-Hydroxyheptanoyl-[acp] + NADP+ <=> 3-Oxoheptano...	true	true
gene_id_46012	Haemophilus	0.5692	R04964	(R)-3-Hydroxydodecanoyl-[acp] + NADP+ <=> 3-Oxododec...	true	true
gene_id_77540	Prevotella	0.7077	R07763	3-Oxostearoyl-[acp] + NADPH + H+ <=> 3-Hydroxyoctadec...	true	true
gene_id_35513	No homologues enough for genus inference		R10116	3-ketoglutaryl-[acp] methyl ester + NADPH + H+ <=> 3-...	true	true
gene_id_3455	No homologues enough for genus inference		R10120	3-ketopimeloyl-[acp] methyl ester + NADPH + H+ <=> 3-...	true	true
gene_id_39014	Prevotella	0.7846				
gene_id_11990	Veillonella	0.3538				
gene_id_41650	Clostridium	0.4462				
gene_id_70160	Prevotella	0.8				
gene_id_73065	No homologues enough for genus inference					

Buttons: Close, Encoding genes, Encoded Reactions

Figure 3.10: Detail windows for the 'Enzymes' main view (Genes on the left, Reactions on the right).



merlin - MEtabolic models Reconstruction using genome-scale INformation -v2.0.1-beta

Enzymes Transporters Compartments Database Model

ard Taxonomy Enzymes

board

Project_2

Database: salv14692@127

Entities

- Genes
- Proteins
- Metabolites

Enzymes data

Enzymes present in the annotation of the metagenome (includes all the occurrences of each enzyme)	8328
Oxidoreductases(1.-)	1063
Transferases(2.-)	2940
Hydrolases(3.-)	1386
Lyases(4.-)	756
Isomerases(5.-)	665
Ligases(6.-)	1518
Number of different enzymes in the metagenome	700
Enzymes in which all their encoding genes have a taxonomic genus assignment	187
Enzymes in which at least one of their encoding genes have a taxonomic genus assignment	431
Enzymes in which none of their encoding genes have a taxonomic genus assignment	82

Figure 3.11: Statistics of the metagenomics enzymes entity.

3.2.5 Metagenomics pathways inference

This section describes the methodology implemented for the classification of functional pathways in complex microbial communities. The main goal here focuses on finding the pathways that are effectively present in the community as a whole, and then find out which organisms may be involved.

When the user performs the loading of KEGG data, pathways information is integrated into the internal database. This information includes the complete enzymes within each pathway, as well as their reactions, amongst others. This is the basis for the routine for metagenomics pathway inference, that is divided in three main stages:

1. Test whether a pathway is effectively present;
2. If so, calculate its abundance;
3. Assign taxonomic information to pathways.

Concerning the first step, this method classifies a pathway as present using hypergeometric tests (equation 3.7). This test calculates the probability that the number of enzymes observed in an enzymes list that compose a pathway occurred by chance. Therefore, given a pathway pt with n enzymes (where n is higher than 3), its probability P is given by:

$$P_{pt} = \frac{\binom{E}{e} \binom{N-E}{n-e}}{\binom{N}{n}} \quad (3.7)$$

where N is the population size of the enzymes that compose all the pathways, E refers to all the enzymes observed in the metagenome and e indicates the number of encoded enzymes (successes) in pathway pt with n enzymes.

To test whether a pathway is statistically significant, *Merlin* compares the value of the p-value P with a threshold t defined by the user (default is 0.1). If P is smaller than t , it means that the probability of the observed situation be explained by chance is so low that the pathway is likely to be present, thus is classified accordingly:

$$f_w = \begin{cases} 1, & \text{if } P_{pt} \leq t \\ 0, & \text{otherwise} \end{cases}$$

where 1 indicates the presence of the pathway pt and 0 its absence.

Pathway abundance abd_p (equation 3.8), is calculated according to a methodology proposed by Abubucker et al [155], representing the average of the upper half enzyme abundances of the pathway, to be robust to low-abundance enzymes:

$$abd_p = \frac{1}{\frac{|p|}{2}} \sum_{i \in [p/2]} w_{i,p} \quad (3.8)$$

where $[p/2]$ stands for the most abundant half of enzymes.

The last stage of the routine tries to describe the genus that is more involved in each pathway. This step is only performed if the pathway is considered present (coverage greater than the threshold), otherwise it would not make sense to assign a genus to a pathway. To make sure that a given genus is operating a pathway, a conservative approach is followed where a genus is assigned to a pathway only if that genus encodes at least 75% of the annotated enzymes within the pathway. This information is pulled out from the enzymes routine, where the genera for each enzyme are discriminated. At the end, this method is able to provide which genera are executing each pathway as well as the opposite, which pathways each genus executes.

3.2.6 *Merlin's* operation mode for pathways

To access and execute the pathways inference routine in *Merlin*, the user must click on the '*Pathways*' view under the '*Metagenomics*' entity. The main view (Figure 3.12) shows the overall information for each pathway, such as the number of complete enzymes that compose it, the number of enzymes from each pathway that were annotated in the sample, the obtained p-value in the hypergeometric test, the pathway abundance and the number of taxonomic genera executing each pathway.

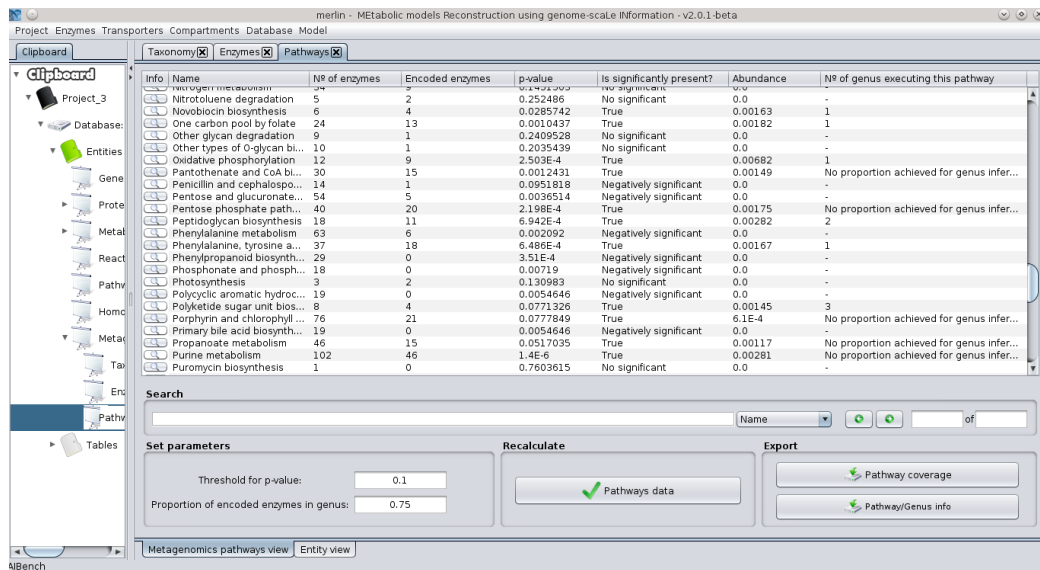
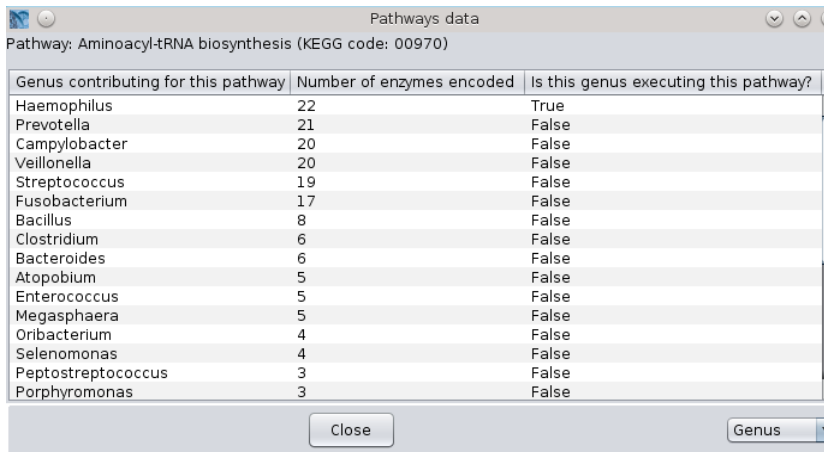


Figure 3.12: Merlin's view for metagenomic pathway information.

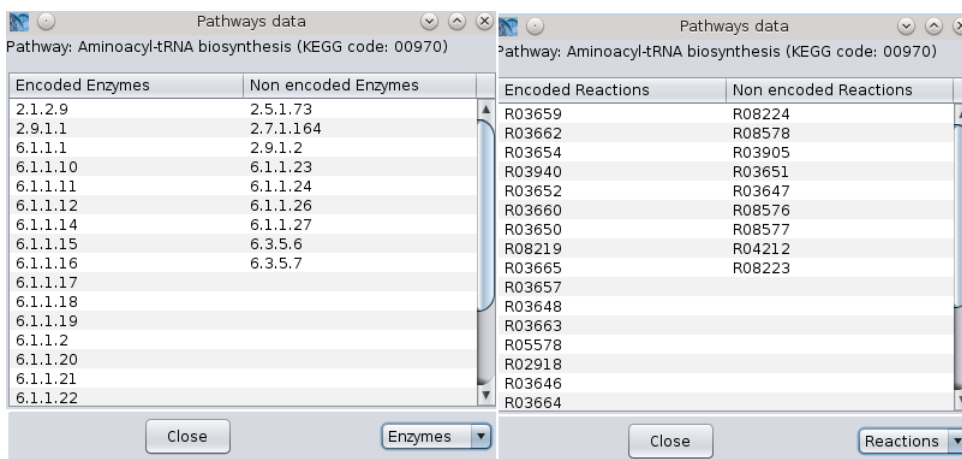
A pathway is considered present if the column '*p-value*' displays a lower value than the threshold defined by the user in the '*Threshold for p-value*' text box. The user can set this parameter to be more or less stringent. Furthermore, he/she can also adopt a less conservative approach for genus assignment by setting the '*Proportion of encoded enzymes in genus*' to a value lower than 0.75. It means that it is possible to choose the proportion of the annotated genes in each pathway that a genus need to encode in order to assign that genus to a specific pathway. For example, if a proportion of 0.5 is selected and the genus *Escherichia* is being tested for pathway *p*, it is only necessary that this genus encodes half of the annotated enzymes to assume that *Escherichia* operates the pathway *p*. The main table can be re-calculated and updated with the new parameters by clicking on the '*Pathways data*' button.

The '*info*' column displays detailed information about the selected pathway, namely at the genus level, where it is shown the genus that executes that pathway (Figure 3.13), and at the enzyme/reactions level where encoded and no encoded enzymes, and encoded and no encoded reactions are shown, respectively (Figure 3.14).



Genus contributing for this pathway	Number of enzymes encoded	Is this genus executing this pathway?
Haemophilus	22	True
Prevotella	21	False
Campylobacter	20	False
Veillonella	20	False
Streptococcus	19	False
Fusobacterium	17	False
Bacillus	8	False
Clostridium	6	False
Bacteroides	6	False
Atopobium	5	False
Enterococcus	5	False
Megasphaera	5	False
Oribacterium	4	False
Selenomonas	4	False
Peptostreptococcus	3	False
Porphyromonas	3	False

Figure 3.13: *Merlin*'s detailed information for genera operating a selected pathway.



Encoded Enzymes	Non encoded Enzymes
2.1.2.9	2.5.1.73
2.9.1.1	2.7.1.164
6.1.1.1	2.9.1.2
6.1.1.10	6.1.1.23
6.1.1.11	6.1.1.24
6.1.1.12	6.1.1.26
6.1.1.14	6.1.1.27
6.1.1.15	6.3.5.6
6.1.1.16	6.3.5.7
6.1.1.17	
6.1.1.18	
6.1.1.19	
6.1.1.2	
6.1.1.20	
6.1.1.21	
6.1.1.22	

Encoded Reactions	Non encoded Reactions
R03659	R08224
R03662	R08578
R03654	R03905
R03940	R03651
R03652	R03647
R03660	R08576
R03650	R08577
R08219	R04212
R03665	R08223
R03657	
R03648	
R03663	
R05578	
R02918	
R03646	
R03664	

Figure 3.14: Detail windows for the 'Pathways' main view (Enzymes on the left, Reaction on the right).

Regarding the export options, *Merlin* allows to export the pathways coverage (presence/absence) to a tab-separated text file by clicking on the 'Pathways coverage' button. Selecting the 'Pathway/Genus info' button, a list of pathways operated by each taxonomic genus is exported to a text file. In the 'Entity view' tab in the down side of the panel the main statistics of the pathways routine are shown (Figure 3.15).

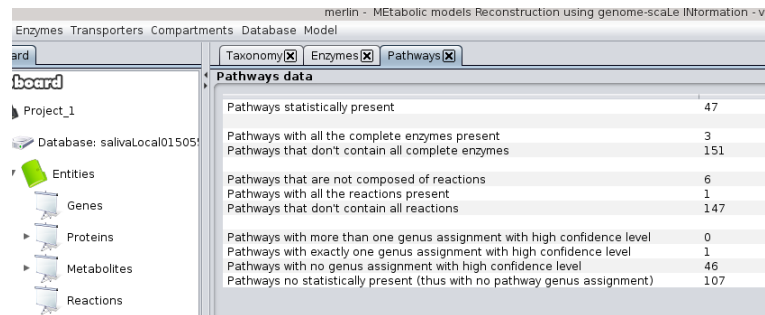


Figure 3.15: Statistics of the metagenomics pathways entity.

3.2.7 Architecture and requirements

The workflow for metagenomics projects in *Merlin* is displayed in Figure 3.16. It is only available on Linux for now, and it worked well on a computer with 3GB of memory, although more memory would be advantageous. The software can be downloaded from http://www.merlin-sysbio.org/meta_merlin.

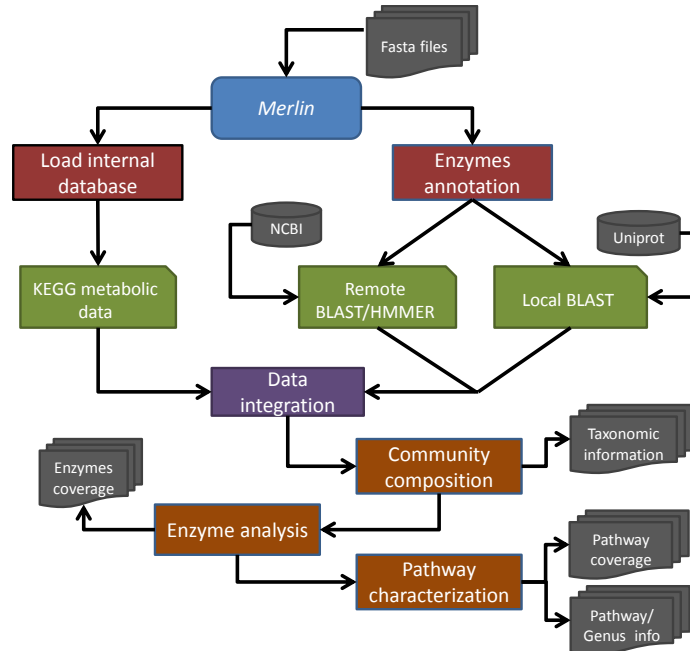


Figure 3.16: Schematic representation of *Merlin* architecture for metagenomic analysis.

Chapter 4

Saliva Microbiome: results and discussion

For this project, the assemblies generated from the saliva samples as part of the HMP were downloaded from the Data Analysis and Coordination Center(DACC), but they were not ready to use in *Merlin*, since it requires a set of genes as input. Each scaffold in HMP contains a high amount of genes, thus a gene prediction software was run to find the putative ORFs for each sample. The MetaGeneMark software was chosen for this task due to its good performance in short reads [55] along with very fast runs. Specifications in the software parameters involved the inclusion of the ribosomal binding site feature '*RBS for gene start prediction*' and the prohibition of the '*Gene overlap*', that is, two genes could not be predicted within the same range of nucleotides. After this step, the files with the predicted genes in FASTA format were ready to be processed in *Merlin*. The dimensions of the five samples used are given in Table 4.1.

Table 4.1: Description of the saliva samples used in this work. The samples name represent the assigned ID in the HMP data repository.

Samples	SRS014692	SRS019120	SRS015055	SRS013942	SRS014468
Number of scaffolds	60611	40761	38301	33912	6251
Number of predicted genes	81279	49665	46189	41906	7883

4.1 Annotation of Metagenomes

Both remote BLAST and local BLAST were run in *Merlin* for the samples from saliva. BLASTp was used with e-value set to 1^{-10} and a maximum number of alignments equal to 50. The local annotation against the whole UniprotKB, which includes Swissprot and TrEMBL was impossible to perform due to the computational power required to execute the operation. With a huge database such as this one, a high performance computing environment is required to perform the local BLAST in a reasonable amount of time and without memory constraints. Therefore, the annotations using a local instance of BLAST were only performed against SwissProt on a desktop computer with an Intel®Core™ i7 CPU 920 @ 2.67GHz with four processor cores with 8 threads and 6Gb of RAM.

Table 4.2: Remote BLAST against NCBI nr vs Local BLAST against SwissProt for the HMP samples ran in *Merlin*.

	Remote Blast			Local Blast		
	Annotated genes	With similarities found	Homologues in the database	Annotated genes	With similarities found	Homologues in the database
SRS014692	81278	62120	986241	81279	32710	183666
SRS019210	49663	45023	994027	49665	20231	155714
SRS015055	46188	41073	1016585	46189	19176	163916
SRS013942	41906	38508	839127	41906	18677	151865
SRS014468	7707	5682	185841	7883	3290	60065
SRS013711 ¹	47412	37272	1302336	47418	18808	154765

¹This Buccal mucosa sample was randomly selected.

Table 4.2 compares the main results from the local and remote BLAST searches. As the table shows, the differences between the two approaches are clear. The remote BLAST against NCBI-nr database provides better results as a big fraction of the genes present similarities. Moreover, the high number of different homologues loaded to *Merlin* demonstrate the greater sensitivity of this database, as well as the high amount of data that the user deals with

metagenomics data. The shortcoming of this type annotation comes with the computational time required. Since it depends on the answers from the NCBI server, all the samples took more than 5 days to run the jobs.

On the other hand, the local BLAST against SwissProt ran much faster (≈ 1 day per sample), but a large number of genes remained unannotated in each dataset ($\approx 60\%$ of the genes). This is explained by the small size of Swissprot, which does not store the diversity and amount of data as the reality harbors. The lower number of homologues also supports this hypothesis (Table 4.2). In the end, the trade-off between sensitivity and time described before was observed. Anyway, the downstream analysis of samples was performed both for local and remote annotations and the results will be discussed later.

4.2 Taxonomic composition

4.2.1 Inference from local BLAST annotations

Merlin predicted the taxonomic composition of the five saliva samples by assigning, if possible, a taxonomic label to every gene. Regarding the samples annotated using SwissProt, and given the low number of homologies found (Table 4.2), it becomes clear that this is not the best approach to taxonomically characterize metagenomes. Since the *Merlin* routine is highly dependent of the BLAST results, poor outputs on this step compromised the performance of the algorithm (Table 4.3).

As the table shows, few of the genes are actually classified with a taxonomic label. An appropriate classification of the sample is limited because a big fraction of the universe of genes is being discarded. Furthermore, using SwissProt as the reference database creates biased results because the known microbial life is not well represented in this database. Instead, few well characterized organisms are highly represented inducing the taxonomic assignments towards these organisms (Figure 4.1).

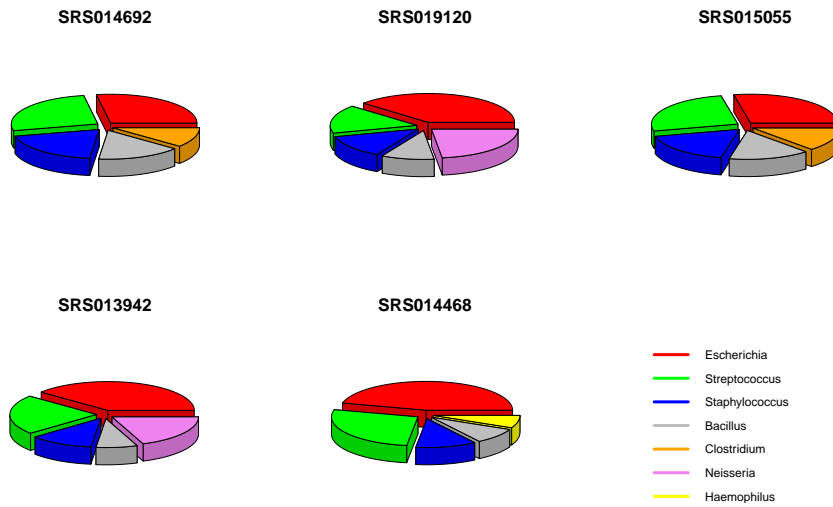


Figure 4.1: Community structure for the saliva samples inferred from local BLAST annotations against SwissProt. Pie charts were generated in R version 2.15.1 using 'plotrix' package.

The pie charts in the figure evidence that the most represented genus in all samples is *Escherichia*. Also, the *Bacillus* and *Staphylococcus* genera are common in the five most represented taxon in saliva. Although these taxonomic genera might be present in the oral cavity, certainly they are not the most representative ones in healthy individuals [141, 156], thus these taxonomic classifications are not correct. Therefore, preventing the analysis to

Table 4.3: Filtering the samples from saliva annotated through local BLAST, across the several steps of the *Merlin* taxonomic routine, with the default parameters.

	Genes	No homologues enough	No minimum score	No concordant taxon	Genes classified
SRS014692	81279	56444	14312	337	9986 (12.29%)
SRS019120	49665	34714	8862	177	5912 (11.90%)
SRS015055	46189	31548	8037	202	6402 (13.86%)
SRS013942	41906	28471	7365	162	5908 (14.10%)
SRS014468	7883	5491	1281	26	1085 (13.76%)

harbor misleading results, the taxonomic classification of metagenomes was performed in *Merlin* only from remote BLAST annotations against NCBI-nr.

4.2.2 Stringency of the routine parameters

Different parameters settings were tested to evaluate the reliability of the scoring algorithms and the main results are provided in Table 4.4. *Merlin* was able to classify more than half of the genes in all samples with the default parameters. Concerning the list of non classified genes, the major fraction refers to those who did not present homologues enough in the homology search, followed by the group of genes who did not achieve the minimum scores (either for phylum or genus classification) (Table 4.4).

When testing the software with a conservative approach (with high values for the thresholds scores) a small proportion of genes were classified, as expected, because most of them did not achieved the minimum scores. On the other hand, when setting *Merlin* with a less conservative approach, it is noted that it does not have a big influence in the proportion of genes classified, despite its increase. In fact, when setting lower values for the thresholds, the percentage of genes with a non concordant genus for a given phylum increases considerably (Table 4.4).

Finally, the change in the parameter defining the minimum number of homologues required for classification to 1 does not show a significant improvement in the results. It is true that the number of genes with no enough homologues was reduced (only genes without similarities still discarded) but this approach increased the number of genes with no minimum scores, which means that genes with few homologues (between 1 and 5) commonly show a great mixture of taxon, avoiding *Merlin* to classify them accordingly.

Overall, the choice of the default parameters seems appropriate. Since in metagenomics a high amount of sequences exist from unknown and poorly described organisms and many genes would not have a list of homologues from the same genus, very strict settings for the scores will cause loss of information. On the other hand, if the user chooses to lower the thresholds down, wrong assignments might happen frequently.

Table 4.4: Filtering of the samples from saliva annotated through remote BLAST, across the several steps of the *Merlin* taxonomic routine with different parameters settings.

	Genes	No homologues enough	No minimum score	No concordant taxon	Genes classified
SRS014692					
Default ¹	81278	30360	7506	334	42990
Conservative ²	81278	30360	29546	0	21372
Generous ³	81278	30360	3047	774	46938
1 homologue required ⁴	81278	19158	14944	334	46743
SRS019120					
Default	49663	9993	5001	202	34429
Conservative	49663	9993	27353	0	12311
Generous	49663	9993	2196	440	36966
1 homologue required	49663	4640	8756	202	35903
SRS015055					
Default	46188	8749	7802	378	29213
Conservative	46188	8749	26689	0	10150
Generous	46188	8749	2470	824	34051
1 homologue required	46188	5115	10019	378	30619
SRS013942					
Default	41906	6564	3762	204	31369
Conservative	41906	6564	24173	0	11168
Generous	41906	6564	1486	398	33446
1 homologue required	41906	3398	5990	204	32296
SRS014468					
Default	7707	2520	937	76	4145
Conservative	7707	2520	4087	0	1100
Generous	7707	2520	435	116	4587
1 homologue required	7707	2025	1306	76	4266

¹Minimum number of homologues = 5, Minimum phylum score = 0.5, Minimum genus score = 0.3;

²Minimum number of homologues = 5, Minimum phylum score = 1, Minimum genus score = 0.75;

³Minimum number of homologues = 5, Minimum phylum score = 0.4, Minimum genus score = 0.2;

⁴Minimum number of homologues = 1, Minimum phylum score = 0.5, Minimum genus score = 0.3;

4.2.3 Characterization of the Saliva microbiome

Having the annotation done, *Merlin* predicted in a very fast and user-friendly way the taxonomic composition of the saliva samples at the phylum and genus level.

Phylum composition

The results demonstrate that at this level of classification, three phyla clearly stand out: *Bacteroidetes*, *Firmicutes* and *Proteobacteria* (Figure 4.2), but none dominates this microbiome. Instead, the *Bacteroidetes* phylum is the most abundant in two samples, the *Proteobacteria* in other two and *Firmicutes* is the most common in one.

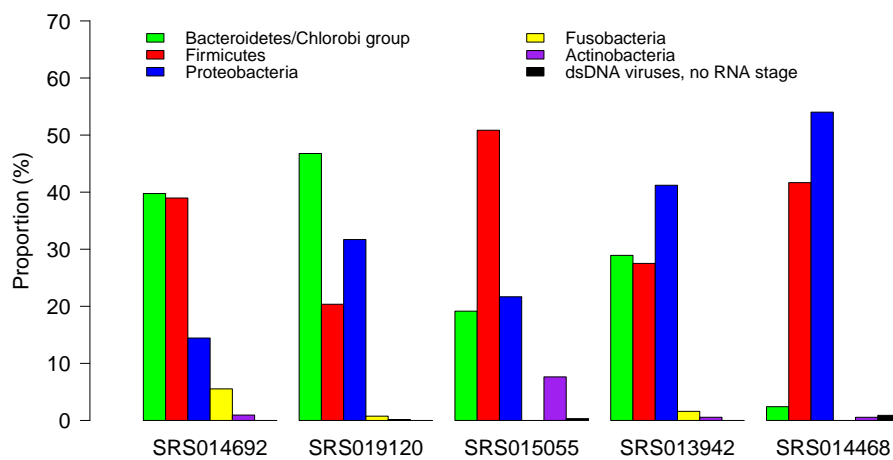


Figure 4.2: *Merlin* predictions of the phyla composition in the samples from saliva. The 'SRS014692' and 'SRS014468' samples are contaminated thus these samples were discarded regarding further analysis.

Having noticed the low abundance of *Bacteroidetes* in the 'SRS014468' sample, and knowing the small number of genes predicted in the first place, only 7883 (Table 4.1), we went to investigate about contamination on this sample. Through a personal communication with Kemi Abolude from HMP, it was mentioned that this sample did not pass the quality control steps in the analysis pipeline along with the 'SRS014692' one, which probably contains human DNA sequences that were not removed. In fact, this very low number of WGS samples from saliva in the HMP (5 sequenced, 3 passed the controls) is justified by the difficulty in sequencing metagenomes from soft tissues, such as saliva, vagina and anterior nares which tend to have a lot of contamination with little to no usable metagenomics sequences [134]. Therefore, only the

samples that passed the quality control tests were used in this work.

A comparison of the results in *Merlin* was performed with other tools. KEGG metagenomes harbors the phylum distribution for the three samples from saliva that passed the quality check controls (Figure 4.3). As the figure shows these distributions are concordant with the *Merlin* predictions: for the 'SRS019120' sample the *Bacteroidetes* phylum is the most abundant, in 'SRS015055' *Firmicutes* appears to be the most common and in the 'SRS013942' sample *Proteobacteria* is dominant (in KEGG Gammaproteobacteria, Betaproteobacteria and Epsilonproteobacteria are treated as phylum when in reality they are classes). A high percentage of the pie charts in KEGG represent undefined organisms. In *Merlin*, this fraction of genes is represented as the ones that were not classified by the routine (Table 4.4).

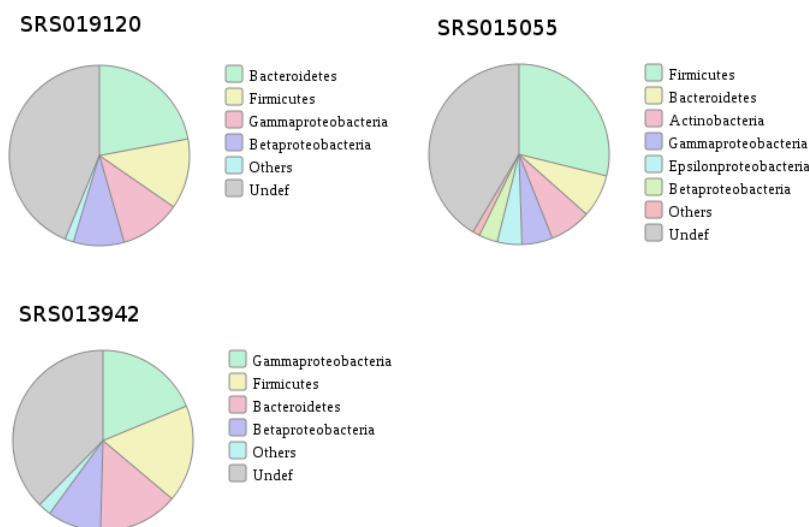


Figure 4.3: Phyla distribution of the non contaminated saliva samples stored in KEGG. They can be accessed with the KEGG metagenomes IDs 'T30414', 'T30237' and 'T30194'.

MG-RAST, a robust system to analyze metagenomes, assigns percentages to the number of reads with predicted proteins and ribosomal RNA genes annotated to a given taxonomic level. The results on this tool for the samples from saliva are described in Figure 4.4.

Predictions with *Merlin* and MG-RAST are concordant, where the three described phyla dominate. Particularly, details such as the significant amount

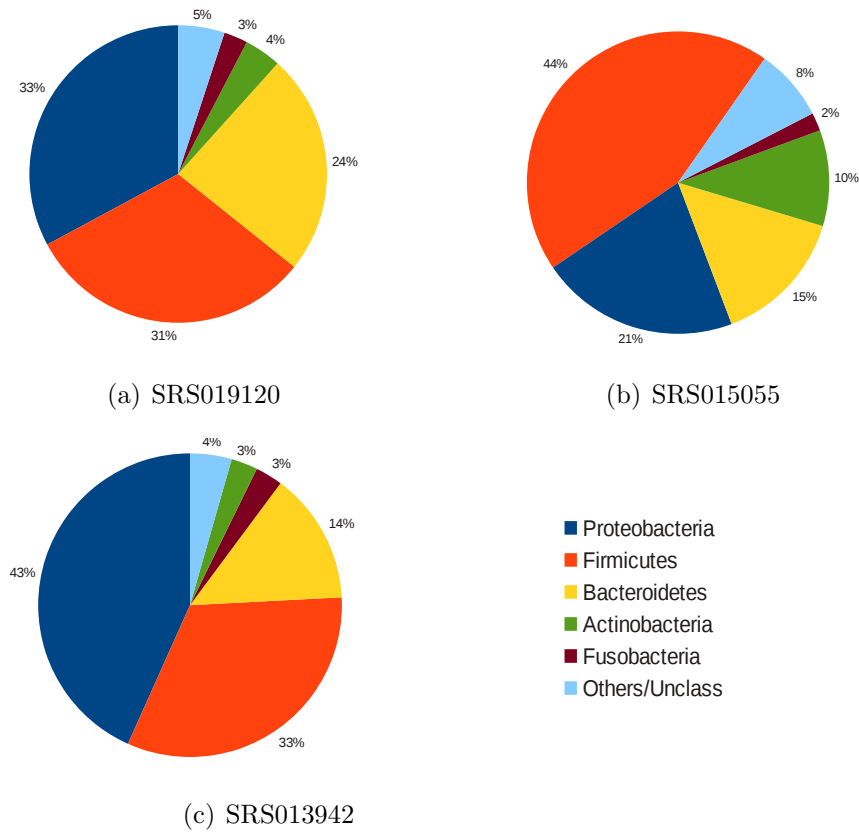


Figure 4.4: Phyla distribution in the saliva samples taken from MG-RAST. To draw the charts, the data can be downloaded through the following MG-RAST metagenomes IDs: (a) 4478542.3; (b) 4473348.3; (c) 4473411.3;

of *Actinobacteria* ($\sim 10\%$) in the 'SRS015055' sample alone and the dominance of *Proteobacteria* in the 'SRS013942' are evidences of the good performance of the tool (Figures 4.2 and 4.4). However, the dominance of *Bacteroidetes* in the 'SRS019120' sample is not in accordance with MG-RAST predictions. These small differences can be explained by the underlying methods used in MG-RAST for annotation (SEED subsystems and FIGfams database), which are different from the ones employed in *Merlin*.

The results obtained in *Merlin* for phylum classification are accurate enough for a taxonomic analysis of microbial communities. Despite the low number of samples with good quality, the results are concordant with previ-

ous studies of the saliva microbiome [157, 158]. *Firmicutes*, *Proteobacteria* and *Bacteroidetes* dominate the phylum distribution, but individuals from *Actinobacteria* and *Fusobacterium* might be found frequently. It is not possible to infer a core microbiome, even at phylum level, because its composition is influenced by several factors such as the host physiology, the diet and the local environment [159]. Since most of the previous studies focused on 16S rRNA amplicon pyrosequencing, a higher number of WGS samples from saliva would be very useful to achieve a more robust analysis of this microbiome. However, the problems in performing metagenomics sequencing on saliva prevented the HMP consortium to achieve better results [134].

Genus composition

Concerning the genus distribution, *Merlin* shows a strong diversity on composition and proportions over different samples (Figure 4.5). The *Prevotella*, *Streptococcus*, *Veillonella*, *Neisseria* and *Haemophilus* genera seem common in all non contaminated microbiomes, although their abundances vary significantly.

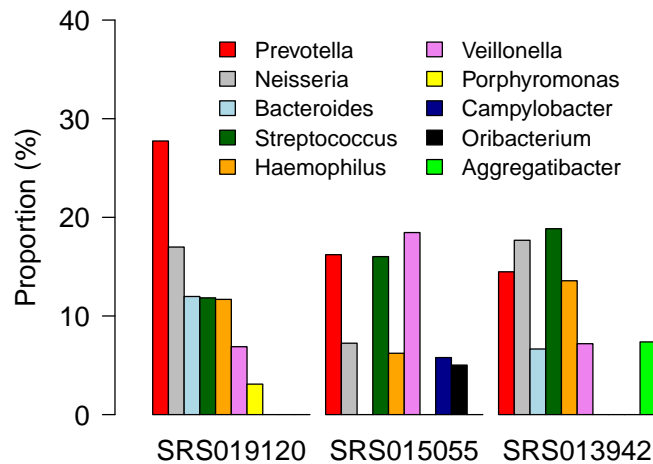
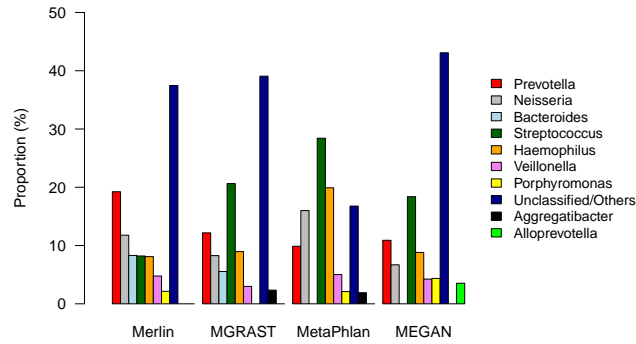
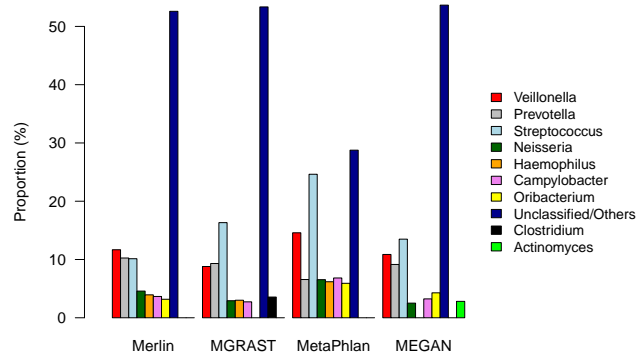


Figure 4.5: Overall composition predicted in *Merlin* for the seven most abundant genera in each sample from saliva. Non classified genes were not included in this chart.

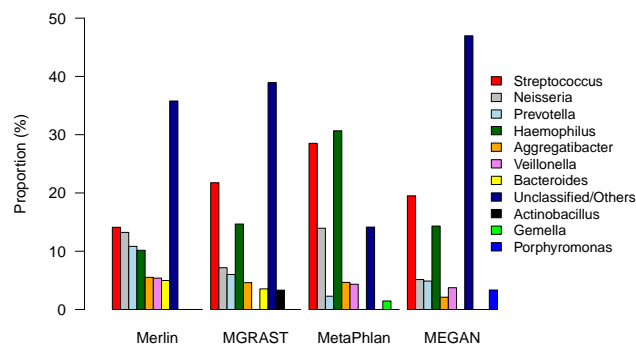
A comparison of *Merlin* results with other tools was performed and it is displayed in Figure 4.6. In these charts, the unclassified and very low abundant genus assigned in *Merlin* were also included for an easier comparison.



(a) SRS019120



(b) SRS015055



(c) SRS013942

Figure 4.6: Genus distribution of the saliva samples in different tools. MEGAN was run from the BLAST results of RAPsearch2 against RefSeq database.

In a first look, the high percentage of unclassified sequences in all charts is evident. It is important to refer that these values are inflated because the organisms with very low abundance (treated as 'others'), which are many, also contribute for these results. Anyway, these large dark blue bars demonstrate the potential of metagenomics on unveiling new forms of life as a great amount of organisms remain unknown.

The non contaminated samples ('SRS019120', 'SRS015055', 'SRS013942) show a great mixture of taxon making impossible to describe a common pattern in the microbiome: Despite the *Prevotella*, *Streptococcus*, *Veillonella*, *Neisseria* and *Haemophilus* are the overall most abundant genera in all samples, no consistency was found between the samples and the tools (Table 4.5). *Streptococcus* appears to be the dominant genus in almost all software (except in *Merlin*), but the proportion within these tools vary a lot (17.1% in MEGAN, 27.2% in MetaPhlAn). Furthermore, *Prevotella* stands for the most abundant genus in *Merlin* and the second one in MG-RAST, but it is the less represented in MetaPhlAn. In addition, *Haemophilus* appears to occupy between $\sim 7.5\%$ to 9% in *Merlin*, MG-RAST and MEGAN but MetaPhlAn indicates that this genus is the second most abundant in saliva reaching $\sim 19\%$ of the overall composition.

Table 4.5: Average distribution (%) of the five most abundant genus in the three non contaminated samples over the different tools. A top-down list of the genera ordered by their abundances is also presented.

	<i>Merlin</i>	MG-RAST	MetaPhlAn	MEGAN
<i>Streptococcus</i>	10.816	19.570	27.186	17.126
<i>Prevotella</i>	13.443	9.165	6.240	8.310
<i>Neisseria</i>	9.863	6.122	12.161	4.781
<i>Haemophilus</i>	7.397	8.879	18.919	8.513
<i>Veillonella</i>	7.276	3.930	7.980	6.281
Ordered abundance	Prevotella Streptococcus Neisseria Haemophilus Veillonella	Streptococcus Prevotella Haemophilus Neisseria Veillonella	Streptococcus Haemophilus Neisseria Veillonella Prevotella	Streptococcus Haemophilus Prevotella Veillonella Neisseria

The different proportions of each taxon on the different tools can be explained considering the way each method works. MetaPhlAn is by far the

fastest one because it uses a reduced reference database of marker genes, representative of 3000 prokaryotic genomes. This can be a disadvantage, because a genus might be poorly represented in the database, thus making the software miss some assignments. The low average abundances of *Prevotella* may be explained by this issue. Furthermore, only reads that match a clade-specific gene are maintained for classification, the others are ignored. That is why the average values of genus distribution are higher within this tool (Table 4.5).

MG-RAST displays higher concordance with *Merlin*, but it also trusts on marker genes for classification making the method less sensitive [33]. MEGAN requires a BLAST (or similar) file as input, so its sensitivity depends both on the software used to perform the similarity searches and the reference database. *Merlin* searches against the whole universe of existing genes (NCBI-nr) and holds a conservative approach in which a gene is only classified if it passes some thresholds. Therefore, the classifications preserve a high degree of confidence and the results displayed in Table 4.5 are reliable. Nonetheless, two constraints are holding back the *Merlin* acclamation: the computational time required to perform the BLAST searches and the fact that many information is lost through the assembly process (*Merlin* does not accept reads as input, requires assembly and gene prediction steps before).

Previous studies of the oral flora at the genus level reveal a diverse microbiome composition. *Streptococcus* is clearly the most common genus in all oral sites, with the exception of saliva, where *Veilonella* and *Prevotella* individuals dominate [160]. Keijser et al. (2008) [161] claim that *Prevotella* and *Streptococcus* account for almost 40% of the saliva abundance, whilst the Yang et al. (2012) study [158] classifies *Neisseria* as the most abundant genus followed for *Prevotella*. These results support *Merlin* in the sense that *Prevotella* individuals are highly abundant in saliva, unlike MetaPhlan displays(4.5). However, the proportions should be interpreted carefully.

4.3 Functional capabilities

4.3.1 Encoded enzymes

The enzymes encoded by genes in the microbes that compose microbial communities are a suitable way to address their functional capabilities. A comparison of the enzymes annotated by *Merlin* and IMG/M-HMP for the three non contaminated samples from saliva is displayed in Table 4.6. IMG/M-HMP web server, designed for the HMP project, uses diverse data sources for annotations, but regarding enzyme assignments, this tool uses KEGG orthology. On the other hand, *Merlin* uses the *Entrez* protein database (done by remote similarity searches) and the Uniprot text file (done by local similarity searches) to retrieve EC numbers from the homologues of each gene.

Table 4.6: Comparison of the complete enzymes annotated by IMG/M and *Merlin* for the non contaminated saliva samples.

	IMG/M-HMP		<i>Merlin</i> (SwissProt)		<i>Merlin</i> (NCBI-nr)	
	Encoded enzymes	Unique enzymes	Encoded enzymes	Unique enzymes	Encoded enzymes	Unique enzymes
SRS019210	12143 (24.34%)	957	10058 (20.25%)	977	4871 (9.81%)	605
SRS015055	11988 (25.81%)	997	9776 (21.17%)	977	2287 (4.95%)	506
SRS013942	10642 (25.46%)	954	8922 (21.29%)	957	2739 (6.54%)	507

The results displayed in Table 4.6 show a large discrepancy between the assignments based on SwissProt and NCBI-nr, where the annotations based on SwissProt seem to agree in cardinality with those stored in IMG/M-HMP. The percentage of the metagenome that encodes enzymes is slightly different ($\sim 25\%$ for IMG/M-HMP, $\sim 21\%$ for *Merlin* with SwissProt), but the number of different enzymes annotated is very similar. These proportions of enzymes are a bit smaller than the usually found in a bacterial genome (33%, regarding essential genes) [162], but *Merlin* does not account for incomplete EC numbers as enzymes (in IMG/M-HMP only complete enzymes

Figure 4.7 demonstrates that the enzyme assignments are more dependent on the chosen method for their annotations rather than the sample itself, as the heatmap grouped together the samples from the same method. The more similar results between IMG/M-HMP and *Merlin* using local BLAST against SwissProt were also evident. The few enzymes identified in remote annotations based on NCBI-nr obtained were highlighted by the predominance of green colors over the heatmap.

The reason for these results can be explained by the stable functional capability of the human microbiome described before ([141],[134]). In fact, no big differences were observed between samples when using the same method (Figure 4.7). Figure 4.8(a) proves the previous statement, where a comparison of enzymes assignments in *Merlin* with annotations against SwissProt shows that the majority of enzymes were common to the three saliva samples.

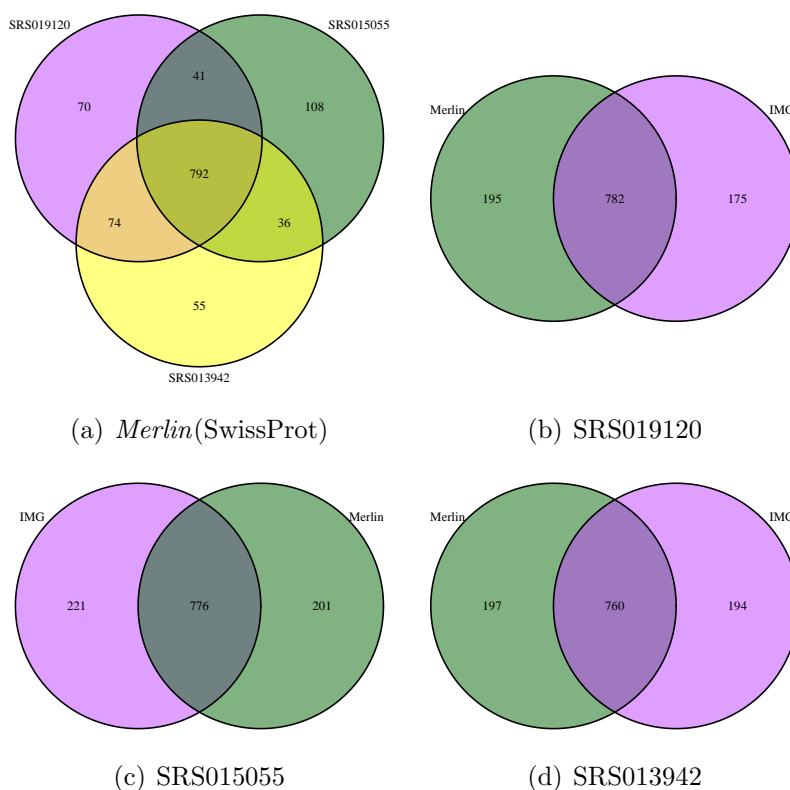


Figure 4.8: **a:** Common enzymes found in the non contaminated samples with *Merlin* (SwissProt). **b-d:** Comparison of common enzymes found with *Merlin* (Swissprot) and IMG/M in the non contaminated saliva samples.

Moreover, the total number of only 1433 different enzymes found across the different annotations, from the large universe of 6130 described in BRENDA, demonstrates the uniformity of the microbiome. Even though the samples were clustered according the underlying methods, figures 4.8(b), 4.8(c), 4.8(d) support the reliability of *Merlin* with SwissProt as the reference database as a significant portion of identified enzymes are in agreement with IMG/M.

The apparent bad results in *Merlin* for annotations against a big database such as the NCBI-nr can have several interpretations. Many genes present an EC number assignment with SwissProt, but the same does not occur using NCBI-nr. An example of such a gene is reported in Figure 4.9, where it is clear that using SwissProt, all the homologues for that gene harbor the same three EC numbers (multi functional protein). On the other hand, using NCBI-nr, only two homologues are described with those EC numbers, thus leading *Merlin* discard these EC numbers because no minimum score was achieved. This is probably wrong because the same product "Fatty acid oxidation complex subunit alpha" was obtained in all homologues and they are likely to perform the same functions. Furthermore, a low e-value for SwissProt annotations was used (1^{-10}), so those enzyme assignments present a high degree of confidence.

To address these poor annotated reports of EC numbers in NCBI-nr, one strategy could be to change the scoring routine for the annotation. Currently, it is based on the portion of EC numbers that appear within the list of all homologues, but this approach does not seem efficient to deal with situations such as the one described above. Instead, an EC number could be automatically assigned to a gene if it appeared concordantly even if a relatively reduced number of times.

Other explanation for this issue might be related to the huge amount of data stored in the NCBI-nr. Due to its huge diversity of possible products, the list of homologues for a given gene may exhibit different functions, thus no EC number will be dominant and the gene will not be assigned to anything (Figure 4.10).

Reference ID	Locus ID	Status	Organism	E Value	Score (bits)	Product	EC Number
WP_005706863	WP_005706863		Haemophilus p...	0	261.0	fatty-acid oxidat...	
WP_005708429	WP_005708429		Haemophilus p...	0	254.0	fatty-acid oxidat...	
WP_005824377	WP_005824377		Actinobacillus ...	0	172.0	fatty-acid oxidat...	
WP_005819179	WP_005819179		Actinobacillus ...	0	168.0	fatty-acid oxidat...	
WP_006817427	ASU2_05250		Actinobacillus s...	0	167.0	fatty acid oxidat...	
WP_005622491	WP_005622491		Actinobacillus d...	1.4013e-45	166.0	fatty-acid oxidat...	4.2.1.17, 1.1.1.35, 5.1.2.3
WP_005612217	WP_005612217		Actinobacillus p...	1e-37	144.0	fatty-acid oxidat...	
YP_001651902	APJL_0900		Actinobacillus p...	1e-37	144.0	fatty oxidation c...	
WP_005601219	WP_005601219		Actinobacillus p...	1e-37	144.0	fatty-acid oxidat...	
WP_005608038	WP_005608038		Actinobacillus p...	1e-37	144.0	fatty-acid oxidat...	
WP_005597407	WP_005597407		Actinobacillus p...	1e-37	144.0	fatty acid oxidat...	
WP_001968741	APF7_0947		Actinobacillus p...	1e-37	144.0	fatty acid oxidat...	
YP_001053589	APL_0988		Actinobacillus p...	1e-37	144.0	fatty acid oxidat...	4.2.1.17, 1.1.1.35, 5.1.2.3
YP_007548020	YP_007548020		Bibersteinia tre...	3e-35	137.0	Fatty oxidation ...	
WP_005712339	WP_005712339		Haemophilus p...	9e-35	136.0	fatty acid oxidat...	
YP_002475272	HAPS_0676		Haemophilus p...	2e-34	135.0	fatty oxidation c...	
YP_008123929	K756_05955		Haemophilus p...	2e-34	135.0	fatty oxidation c...	
WP_010129919	WP_010129919		Haemophilus sp...	3e-33	132.0	fatty-acid oxidat...	
WP_007524632	WP_007524632		Haemophilus sp...	3e-32	129.0	fatty acid oxidat...	

(a) NCBI-nr

Reference ID	Locus ID	Status	Organism	E Value	Score (bits)	Product	EC Number
Z2E6	Q0T2E6		Shigella flexneri	0.000000000000...	70.9	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
ZL0	A8A2L0		Escherichia coli	0.000000000000...	68.9	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
WV4	B2TWV4		Shigella boydii	0.000000000000...	68.9	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
ZM2	Q3YZM2		Shigella sonnei	0.000000000000...	68.9	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
IQ0	Q83QQ0		Shigella flexneri	0.000000000000...	68.9	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
AS	B1IXAS		Escherichia coli	0.000000000001	68.6	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
YB7	Q31YB7		Shigella boydii	0.000000000001	68.2	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
FG4	Q8FFG4		Escherichia coli	0.000000000002	67.8	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
VS2	B7NSV2		Escherichia coli	0.000000000002	67.4	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
IM2	B7M6M2		Escherichia coli	0.000000000003	67.4	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
IP24	B7NP24		Escherichia coli	0.000000000004	67.0	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
YX4	B5YX4		Escherichia coli	0.000000000004	66.6	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
IQ4	B6IQ4		Escherichia coli	0.000000000004	66.6	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
CP2	Q8XCP2		Escherichia coli	0.000000000004	66.6	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
YFB8	A7ZFB8		Escherichia coli	0.000000000004	66.6	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35
IGV7	B7MGV7		Escherichia coli	0.000000000005	66.6	Fatty acid oxida...	4.2.1.17, 5.1.2.3, 1.1.1.35

(b) SwissProt

Figure 4.9: Comparison of the BLAST results for a given gene in *Merlin* using different databases as reference.

Reference ID	Locus ID	Status	Organism	E Value	Score (bits)	Product	EC Number
WP_009293951	WP_009293951		Campylobacter ...	0	242.0	ATPase	
WP_002942236	WP_002942236		Campylobacter ...	0	237.0	ATPase	
YP_001408094	CV52592_1377		Campylobacter ...	0	223.0	heavy metal tra...	3.6.3.-
WP_018136617	WP_018136617		Campylobacter ...	0	223.0	ATPase	
WP_009651533	WP_009651533		Campylobacter ...	0	224.0	ATPase	
WP_002945357	WP_002945357		Campylobacter ...	0	197.0	ATPase	
WP_002952114	WP_002952114		Campylobacter ...	0	199.0	heavy metal tra...	
WP_002946912	WP_002946912		Campylobacter ...	0	199.0	ATPase	
WP_009497046	WP_009497046		Campylobacter ...	0	197.0	Lead, cadmium...	3.6.3.4, 3.6.3.3...
YP_892006	CFR240_0831		Campylobacter ...	0	176.0	metal transport...	
WP_018712333	WP_018712333		Campylobacter ...	0	168.0	hypothetical pr...	
WP_016645928	WP_016645928		Campylobacter ...	0	167.0	heavy metal tra...	
YP_005553033	ABLL_0841		Arcobacter sp. L	2.00386e-43	159.0	heavy metal tra...	
YP_003656835	Arnit_2680		Arcobacter nitr...	3.9937e-43	159.0	heavy metal tra...	
YP_006404461	YP_006404461		Sulfurospirillum ...	0	168.0	ATPase	
WP_005871640	WP_005871640		Campylobacter ...	1.99965e-42	157.0	ATPase	
NP_907711	WS1571		Wolonia succin...	9.99967e-42	154.0	ATPase	
YP_005538189	ABED_0662		Arcobacter butz...	2.94273e-44	162.0	heavy metal tra...	
YP_008330643	A7H1H_0699		Arcobacter butz...	5.04467e-44	161.0	heavy metal tra...	
YP_001489645	Abu_0711		Arcobacter butz...	5.04467e-44	161.0	heavy metal tra...	3.6.3.4
WP_006802845	WP_006802845		Helicobacter w...	7e-40	150.0	ATPase	
YP_001405955	CHAB381_0353		Campylobacter ...	7e-40	149.0	metal transport...	
WP_005021857	WP_005021857		Helicobacter pu...	5.04467e-44	161.0	ATPase	
WP_005673100	WP_005673100		Leitropia mirab...	3e-37	143.0	ATPase	
YP_962155	YP_962155		Shewanella sp...	8e-36	139.0	ATPase	
YP_006011182	YP_006011182		Shewanella put...	8e-36	139.0	ATPase	
YP_001184704	YP_001184704		Shewanella put...	7e-36	139.0	ATPase	
YP_157340	ebA609		Aromatoleum ar...	3e-35	137.0	cation transpor...	

Figure 4.10: BLAST result of a gene with different products and EC numbers within its list of homologues.

The opposite situation can also occur, even if in much lower frequencies: *Merlin* assigns an enzymatic function to a given gene when using the NCBI-nr database, but using SwissProt it does not. This is explained by the high number of genes that do not get any similarities when searching against a small database such as SwissProt.

Enzymes abundance and taxonomic relationships

Regarding the most common enzymes found in saliva, no surprises were found, since the ones involved in the basics of microbial life were assigned a significant number of times. Enzymes such as the DNA polymerase (EC: 2.7.7.7) for DNA replication, RNA polymerase (EC: 2.7.7.6) for transcription, DNA topoisomerases and helicases (EC: 5.99.1.2, 5.99.1.3 and 3.6.4.12) involved in DNA unpacking and ATP synthase (EC:3.6.3.14) for ATP synthesis and hydrolysis are examples of such proteins.

Merlin has a feature to associate which microorganisms are encoding each enzyme. This can be useful when looking for unique enzymes encoded by a single group of bacteria. Since this feature depends on the taxonomy routine discussed before, and knowing beforehand that taxonomic inferences from annotations against small databases do not perform well, the analysis of the enzymes and taxonomy simultaneously was not performed for the local BLAST annotations against SwissProt .

Therefore, the relationship between enzymes and taxonomy was only analyzed with the annotations based on NCBI-nr database. Table 4.7 demonstrates, from the set of unique enzymes in each sample, the efficiency of *Merlin* in assigning a taxonomic genus to each enzyme.

This feature is naturally highly dependent of the taxonomy routine accuracy, but the results are elucidative of the performance of the software. It was possible to assign a functional role to a genus in all genes encoding an EC number in more than half of the enzymes present in all samples (Table 4.7). Only a very small fraction of the enzymes did not get any taxonomic association, thus making *Merlin* suitable to study small particularities in the functional capabilities of microbial communities. Most of the encoded

Table 4.7: Statistics regarding the assignment of enzymatic activity to a taxonomic genus in *Merlin* for the non contaminated samples from saliva using NCBI-nr as the reference database.

	Unique enzymes	All the genes ¹	At least one ²	0 genes ³
SRS019210	605	328	245	32
SRS015055	506	321	155	30
SRS013942	507	427	67	13

¹Number of enzymes in which all their encoding genes have a taxonomic assignment.

²Number of enzymes in which at least one of their encoding genes have a taxonomic assignment.

³Number of enzymes in which none of their encoding genes have a taxonomic assignment.

enzymes were uniformly distributed across the most abundant genera described before, but exceptions such as the one displayed in Figure 4.11 might happen frequently.

Enzymes data		
Enzyme: 3.1.11.5		
genus	number of genes	proportion
Neisseria	17	80.952%
Streptococcus	2	9.524%
Haemophilus	1	4.762%

(a) SRS019210

Enzymes data		
Enzyme: 3.1.11.5		
genus	number of genes	proportion
Neisseria	7	87.5%

(b) SRS015055

Enzymes data		
Enzyme: 3.1.11.5		
genus	number of genes	proportion
Neisseria	24	85.714%
Streptococcus	2	7.143%
Actinobacillus	1	3.571%

(c) SRS013942

Figure 4.11: Proportion of the genes encoding the enzyme Exonuclease V (EC number: 3.1.11.5) executed by different taxonomic genus in the saliva samples.

As the figure shows, the genus *Neisseria* is almost the only one that encodes the Exonuclease V enzyme. This enzyme is a helicase-nuclease that is also responsible for repairing double-strand breaks in DNA by homologous recombinations. These double-strand breaks can be caused by UV light, chemical mutagens or by errors in DNA replications, therefore their repair is essential for cell viability [164, 165]. To the best of our knowledge, no biological correlation exists to the fact that almost only *Neisseria* encodes this enzyme, but this is the type of information that *Merlin* can provide to enhance further studies to understand the biological meaning for that.

This *Merlin* operation is interesting but the results are still limited in the sense that few enzymes were classified in the first place. If we want to achieve a reasonable number of enzymes representing the reality better, annotations against SwissProt would be more valuable (some false positives would also arise). However, taxonomic information would be lost with this methodology. On the other hand, using a big database such as the NCBI-nr enables a first hint about which genus encode for a given type of enzyme, but the universe of all identified enzymes is small compared to the reality.

4.3.2 Functional pathways

Merlin predicted the pathways present in each metagenomic sample using hypergeometric tests based on the number of enzymes encoded in each. The results obtained by HUMAnN, the pipeline developed by HMP to infer community function, were used to compare with those produced by *Merlin* (Table 4.8).

A few considerations need to be made before the interpretation of the results. *Merlin* only analyzes pathways where at least one complete EC number exists. Therefore, only a list of 154 pathways are tested for significance every time the pathways routine is run. HUMAnN calls for coverage in every pathway in KEGG associated to KO numbers, instead of EC numbers. Hence, the universe of pathways is larger because there are many pathways with no associated enzymes (EC numbers), but all of them have genes (KO numbers) involved in other functions. However, in this work, the pathways with no en-

Table 4.8: Number of pathways assigned with HUMAnN and *Merlin* for the saliva samples. The unique pathways columns refer to those that were exclusively classified by each method within each sample.

	HUMAnN		<i>Merlin</i> (SwissProt)		<i>Merlin</i> (NCBI-nr)	
	Present pathways	Unique pathways	Present pathways	Unique pathways	Present pathways	Unique pathways
SRS019210	50	10	51	3	40	1
SRS015055	56	20	47	3	37	1
SRS013942	45	8	47	3	37	2

zymes in their constituents, were filtered out from HUMAnN results for an easier comparison.

There are also differences in the pathway coverage methods. The hypergeometric tests employed in *Merlin* try to find significantly enriched pathways, so pathways with p-values lower than the threshold set by the user are automatically treated as present (binary value 1), otherwise they are absent (binary value 0). HUMAnN calculates coverage as the likelihood that all genes needed to operate a pathway are encoded, by estimating the fraction of KOs in the pathway that are confidently present, that is, with abundance greater than the overall sample median [155]. The coverage values provided range from 0 and 1. Thus, for a comparative analysis, values higher than 0.5 were treated as present (binary value 1) and those with values lower than 0.5 were handled as absent (binary value 0).

Table 4.8 compares pathway classifications by the different methods. *Merlin* assignments based on annotations against SwissProt presented a similar number of pathways to HUMAnN, but with a big difference on unique pathways within samples. As expected, the samples annotated against NCBI-nr harbored a lower number of pathways, since the number of encoded enzymes was smaller too (Table 4.6). However, the differences observed between the two strategies for enzymes were not that evident for pathways. This is explained by the nature of the hypergeometric test, which is independent of the sample size (encoded enzymes in each metagenome) regarding a finite population size (sum of all enzymes in all pathways). Even though the sample size is very small for annotations against NCBI-nr, the assignment of pathways is independent of this, which enabled *Merlin* to classify a reasonable

amount of pathways. Nonetheless, pathways containing few enzymes (e.g. Biosynthesis of ansa-mycins (map01051, 3 enzymes), beta-Lactam resistance (map00312, 1 enzyme)) are prone to be false negatives in *Merlin*. Due to this low number of enzymes, the hypergeometric test will display high p-values as the result does not show statistic significance to discard the random chance as the main reason for the observed situation. That is why *Merlin* only tests for significance in pathways composed by more than 3 enzymes. As a result, the Biosynthesis of vancomycin group of antibiotics pathway (map01055, 1 enzyme) was classified as present in HUMAnN, but in *Merlin* it did not, even though the enzyme was encoded. Anyway, the number of pathways in such conditions is small and usually they are not metabolic. Therefore, these conditions have little influence in the overall performance of the tool.

To inspect if the inferred pathways were concordant across different methods, an heatmap was constructed (Figure 4.12). This data representation enabled to cluster the samples according to their similarity in terms of coverage.

Figure 4.12 demonstrates again that the method used for pathways assignment influences more the clustering of the samples than the sample itself. The three samples were stable between them, while the applied methodology seems responsible for the differences in the results. Particularly, HUMAnN shows singular patterns, while the two approaches using *Merlin* are mostly concordant. This is confirmed by the significant number of unique pathways assigned in HUMAnN comparing to *Merlin*, which presents few unique pathways (Table 4.8).

This uniqueness may be explained mainly by the limitations in the enzymes annotation reported before: fewer enzymes were classified, thus less pathways were identified. As explained before, *Merlin* requires assembling the metagenome before the input of genes. HUMAnN infers community function directly from the short reads which prevents the loss of information that is inherent to *Merlin*. In addition, pathways with many incomplete EC numbers, such the case of the Nitrotoluene degradation (map00633) are treated as absent on *Merlin*. This happens because *Merlin*, a software firstly designed to build metabolic models for single species, discards incomplete EC numbers from the pathway constituents because no reactions are associated

to them, and thus they are not important to the model. Such pathways will have few complete enzymes left, so *Merlin* is likely to classify them as absent.

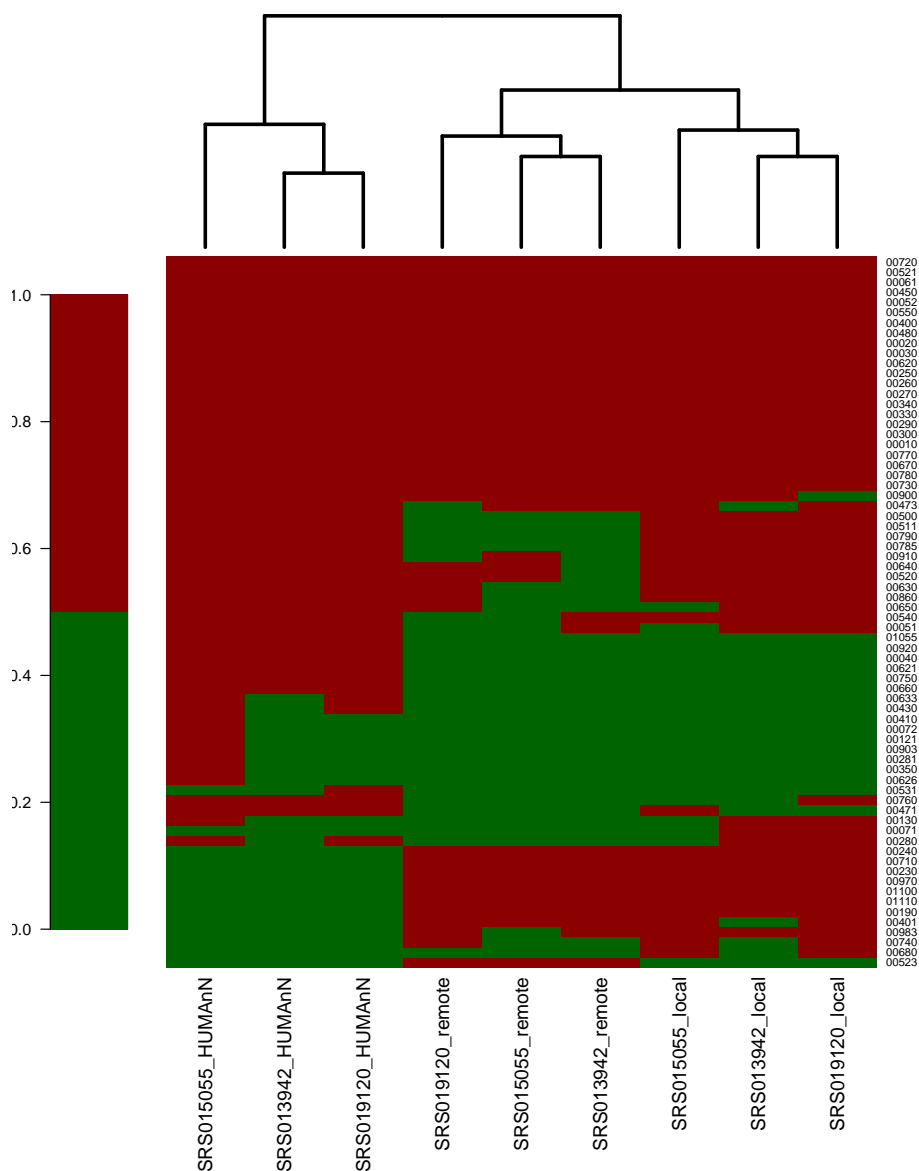


Figure 4.12: Presence of metabolic pathways in the samples from saliva across different annotations. Vertical bars represent the samples. Horizontal bars represent the binary value for pathway coverage. Red colors stands for present pathways whilst green cells account for the absent ones. 'Heatplus' package from Bioconductor [163] was used with hierarchical clustering algorithm using Euclidean distance.

On the other hand, some pathways were identified in *Merlin* but not in HUMAnN (Figure 4.12). Essential pathways such as the Purine and Pyrimidine metabolism (map00230 and map00240, respectively), the Aminoacyl-tRNA biosynthesis (map00970) and oxidative phosphorylation (map00190) are in this group. In HUMAnN, these pathways achieve coverage scores below 0.5 (from 0.45 to 0.49) and that is the reason why they were treated as absent for this comparison, but in fact they might be present given their importance. The fact that this tool depends on gene/enzyme abundances for pathway coverage calculation might be the reason for their low scores.

Despite all this, 23 pathways were significantly found in all samples across the different methods (Figure 4.12), which supports the stability of functional pathways in the human microbiome described before [134, 155]. Moreover, only 70 different pathways out of 154 possible (with complete EC numbers) were found over all methods and samples, demonstrating that this small functional variation in saliva is in agreement with other body sites [141]. This common metabolic content involves the basics of microbial life and metabolism and evidences the good performance of *Merlin* when compared with HUMAnN, a powerful tool developed in the HMP project by several research groups.

Nevertheless, when it comes to identifying more specialized processes within each body habitat, this approach might exhibit some shortcomings. It is true that the functional signature of a microbial community is more stable than its taxonomic composition, but some functional processes are body habitat specific, otherwise some functions related to it could not be performed. Abubucker et al. (2012) discussed this issue by claiming that using these large pathways as functional objects lacks on specificity and only a small portion of most KEGG pathways are treated as present in the human microbiome because the underlying methods require that a significant fraction of the pathway constituents are present [155]. Frequently, in real life a microorganism does not need to operate a whole pathway, instead it only executes smaller paths to achieve the production of a given metabolite. Therefore, some pathways might be described as absent in *Merlin* and HUMAnN, when in fact they occur in the natural environment.

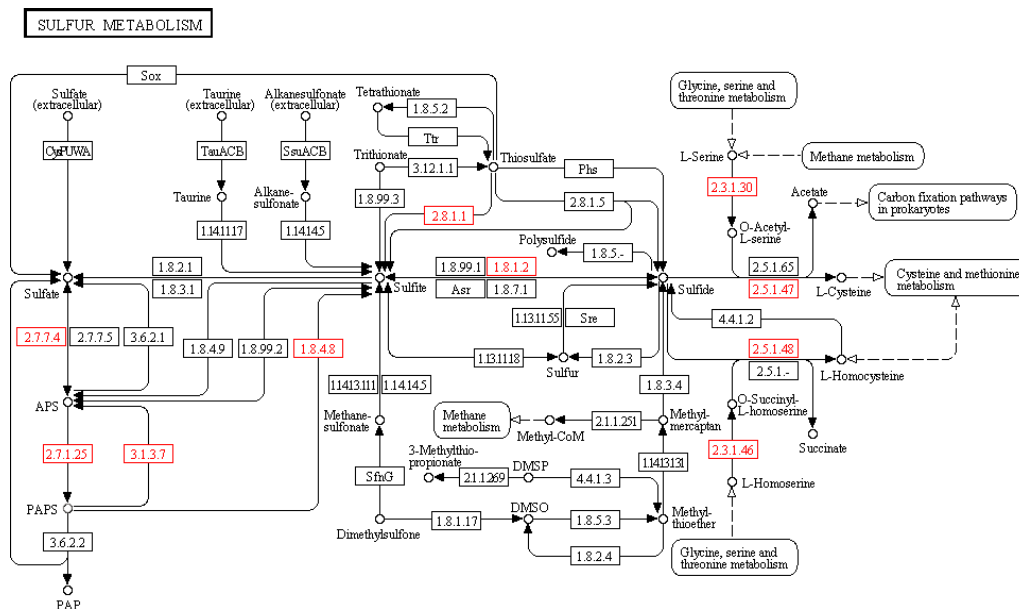


Figure 4.13: Representation of the present enzymes, marked in red, in the Sulfur metabolism pathway (map00920) for the SRS019120 sample with annotations against SwissProt (local BLAST).

An example of such pathway is displayed in Figure 4.13. The Sulfur metabolism pathway (map00920) is an essential element of life and the sulfate reduction is responsible for the biosynthesis of S-containing amino acids (methionine and cysteine). In the specific case of the figure, *Merlin* only assigned 10 enzymes out of the 45 that characterize this pathway, thus it did not achieve statistical significance. However, Figure 4.13 demonstrates operability, since the necessary reactions to reduce sulfate and produce cysteine are there. Despite the fact that the majority of enzymes are not present, this subpathway might be activated to keep the system viable.

Similarly to this example, many others might happen in nature, therefore these missing pathway assignments influence, somehow, the final result. An alternative to address these problems was proposed by Abubucker et al.(2012), that suggests instead of analyzing large pathways, to focus on smaller modules within them. They found little site-specific abundance variations in big pathways, but KEGG modules showed greater inter subject variability, thus more appropriate for comparative studies. Findings such as

the enrichment of arginine transport (M00229) and methionine biosynthesis (M00017) in the oral habitats would not be possible using larger pathways [155].

Combining membership with functional reconstruction

As for enzymes, *Merlin* has a feature to associate which microorganisms are executing each pathway. Once again, the analysis of the pathways and taxonomy simultaneously was not performed for the local BLAST annotations against SwissProt, given their bad results in the taxonomy routine. Table 4.9 provides information about how many genera are operating the identified pathways. It is possible to see that the majority of identified pathways in the saliva samples did not get taxonomic matches, that is, few pathways obtained a genus encoding at least 75% of their identified enzymes.

Table 4.9: Statistics regarding the assignment of metabolic pathways to a taxonomic genus in *Merlin* for the non contaminated samples from saliva using NCBI-nr as the reference database.

	Present pathways	No genus inference	> than 1 genus ¹	1 genus ²
SRS019210	40	26	4	10
SRS015055	37	31	2	4
SRS013942	37	21	2	14

¹Number of pathways that are confidently operated by more than 1 genus.

²Number of pathways that are confidently operated exactly by 1 genus.

This operation is very promising, since it is supposed to identify pathways that are associated with a specific group of organisms. However, the hypothesis to describe the functional capability of a microbial community refers that the basics for microbial metabolism remain present in the set of organisms that comprise an environment, with exception for few specific modules. This was exactly what did not happen in this implemented feature in *Merlin*. As Table 4.9 shows, the number of pathways in which only one genus was confidently linked was higher than those pathways where several genera were associated. This lack of concordance contradicts the hypothesis of a core

functional pool across the sample members and given such conditions in the real life, most bacteria would not even survive.

The weakness of the results is mainly explained by the loss of information inherent to a metagenomic analysis through assembling of the reads. The complete assembly of a given species in a metagenome rarely happens, therefore the contigs/scaffolds representing, for instance, a *Streptococcus* specie may fail into assemble the part of the genome that encodes for a set of enzymes present in a given pathway [109, 166]. Therefore, even if those enzymes might be encoded by any other species and they are identified in the pathway, the test for *Streptococcus* operating the pathway in *Merlin* will fail due to these artifacts. Another reason regards the low number of enzymes encoded with annotations against NCBI-nr, thus missing several pathway constituents frequently.

Nevertheless, some results were achieved with high degree of confidence with this feature: abundant microorganisms in saliva belonging to *Neisseria*, *Haemophilus*, *Streptococcus* and *Prevotella* genera operate important pathways such as the Aminoacyl-tRNA biosynthesis (map00970), Valine, leucine and isoleucine biosynthesis (map00290), Biotin metabolism (map00780), D-Alanine metabolism (map00473), Peptidoglycan biosynthesis (map00550), Purine metabolism (map00230) and Citrate cycle (TCA cycle, map00020). This number significantly increased when the threshold for assigning a genus to a pathway was reduced to 0.5. However, false negative assignments were introduced since only 50% of the enzymes from the same genus were required to assign a genus, even in those pathways that are composed already by few enzymes.

It the end, *Merlin* was able to properly characterize the functional pool of the saliva microbiome, that is somehow similar as other body habitats. If the objective lies on getting an overall idea of the microbial metabolic pathways where the boundaries between species are ignored, *Merlin* is perfectly capable of performing this task in a user-friendly. However, if the user wants to look up for specific variations among environments and adaptations to nutrient changes and metabolite availabilities, he may want to look at a deeper level for presence/absence of specific metabolic processes represented in KEGG

modules or SEED subsystems, for instance. Additionally, further functional potential remain unknown as a substantial amount of gene families is still uncharacterized [167], thus much work still needs to be done.

Chapter 5

Conclusions and further work

5.1 Overview

The emergence of metagenomics in recent years as a discipline with potential to advance knowledge in a wide variety of fields such as medicine, engineering and agriculture has led to the development of a great number of tools to analyze this type of data at different levels [33, 65]. However, a significant portion of the tools are web-based services, which sometimes do not correspond to the user preferences. If the user wants to run his own data on a local computer and tune several parameters towards his goals, he might be able to run some standalone programs available, but such tools frequently require some computational knowledge since they are command-line based and require installation of other tools to work properly.

This work presented an upgrade to *Merlin*, a software firstly designed to construct metabolic models, which is now able to perform a reliable analysis of microbial communities. It enables the study of metagenomes in a user-friendly way without further dependencies and installations, allowing a microbiologist, ecologist or geneticist to use the tool easily without many informatics concerns.

Merlin incorporates two common approaches to study metagenomes: taxonomic composition and functional capability. Since the software was originally developed in Java, these new features were implemented in the same

programming language as well. It requires as input a file with gene encoding sequences, therefore a metagenomics prediction software must be run before over the scaffolds generated by the assembling software. Since the execution of the features depends on the results from an homology search, the BLAST tool has to be run first to perform the annotation of the sample and afterwards loaded into the *Merlin* internal database. Finally, the results from BLAST are used to feed the taxonomic and functional routines, respectively.

The overall taxonomic composition of the community can be easily obtained, where the proportions of phyla and genera are discriminated. Regarding the metabolic analysis, *Merlin* allows to identify which enzymes are present and calculate their abundance, as well as to find out which metabolic pathways are effectively present. A first attempt to correlate the functional capability with the taxonomic members of the community was also done.

The performance of the tool in the saliva microbiome showed the same pattern as observed before: while the pathways needed for microbial life remain relatively stable, the community composition varies extensively among individuals. The taxonomic membership is influenced by several factors such as the environment [159], age [168] or diet [169], thus it was not possible to infer a core structure for the microbiome. Furthermore, a larger number of samples from saliva would have been valuable, but the difficulty in sequencing such microbiome in the HMP hampered this work to achieve better results.

Nowadays, the main goal of studying the human microbiome can be addressed: improve the human health based on the manipulation of the microbes that live in the human body. Several diseases have been associated to shifts in the microbiome [170–172], and the current possibilities to explain the mechanisms behind such conditions enhances the emergence of new treatments to fight those diseases. Furthermore, the inclusion of metatranscriptomics and metaproteomics studies will be of great importance to fully comprehend how and why metabolic processes and microbial composition are altered in diseased conditions [173, 174].

5.2 Limitations

As all scientific work has its limitations, no exceptions in this case were observed. The main limitations of this work will be described next:

Merlin architecture

- The software was firstly designed to construct genome-scale metabolic models for single organisms, thus all the *Merlin* structure is projected towards that goal.
- The fact that *Merlin* requires a file with gene encoding sequences as input making the assembly of the reads necessary. Loss of information occurs as low abundant species are frequently discarded due to their low sequencing coverage along with the fact that closely related species might be assembled together. Moreover, the highly fragmented contigs and scaffolds might compromise the performance of the gene prediction software.

Annotation

- High computational time to perform remote BLAST searches against NCBI-nr.
- Implementation of local BLAST is currently only feasible against SwissProt. Due to the huge size of TrEMBL, it is not advisable to use such database in a regular computer.
- Large number of genes without similarities (inherent to metagenomics).

Taxonomic routine

- No usable results for projects with annotations against SwissProt.
- High dependence of the database used for annotation.

Functional capability routines

- Few enzymes are classified using remote BLAST against NCBI-nr, which compromises the performance of the enzymes and pathways routines.
- Use of the KEGG pathways based on EC numbers for pathway identification. Incomplete EC numbers within pathways are discarded because no reaction is associated to them.
- Association between taxonomic composition and functional pathways shows poor results.

5.3 Future work

Although the main goals proposed for this project have been accomplished, some features could be added to improve the tool:

- Implement annotations against KEGG Orthology (KO), or any other catalog of orthologs (COG,NOG) that can be mapped to KOs. This feature would increase the speed of the process maintaining high sensitivity. Moreover, concerning the functional pathways routine, this implementation would allow to take into account non enzymatic pathways, as well as the inclusion of the incomplete EC numbers that compose some pathways.
- Use of smaller functional modules to characterize the metabolic potential of the metagenomic samples. At this point, *Merlin* uses large pathways with up to several hundred genes, but this strategy lacks on specificity. The use of KEGG modules (each contain an average of ~ 10 genes) would be beneficial.
- Include a systems biology approach to understand the dynamics of a microbial ecosystem. It has been shown that there is a high level of molecular and metabolic interactions between microbes of a certain

species, as well as ecological interactions between the numerous species comprising the microbiome [175, 176]. Therefore, the future lies on the *in silico* construction of metabolic models for microbial communities and the time to bridge this gap is closer than ever.

Bibliography

- [1] W. B. Whitman. “Prokaryotes: The unseen majority”. *Proceedings of the National Academy of Sciences* 95 (12): 6578–6583, 1998.
- [2] F. Sanger and A. Coulson. “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. *Journal of Molecular Biology* 94 (3): 441–448, 1975.
- [3] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene”. *Nature* 260 (5551): 500–507, 1976.
- [4] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. “Nucleotide sequence of bacteriophage ϕ X174 DNA”. *Nature* 265 (5596): 687–695, 1977.
- [5] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. “Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.” *Science* 269 (5223): 496–512, 1995.
- [6] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church. “Advanced sequencing technologies: methods and goals.” *Nature reviews Genetics* 5 (5): 335–44, 2004.
- [7] R. I. Amann, W. Ludwig, and K. H. Schleifer. “Phylogenetic identification and in situ detection of individual microbial cells without cultivation.” *Microbiological reviews* 59 (1): 143–69, 1995.
- [8] N. R. Pace. “A Molecular View of Microbial Diversity and the Biosphere”. *Science* 276 (5313): 734–740, 1997.
- [9] J. L. Stein, T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong. “Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon.” *Journal of bacteriology* 178 (3): 591–599, 1996.
- [10] P. Hugenholtz, B. M. Goebel, and N. R. Pace. “Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity.” *Journal of bacteriology* 180 (18): 4765–4774, 1998.

- [11] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Neelson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. "Environmental genome shotgun sequencing of the Sargasso Sea." *Science* 304 (5667): 66–74, 2004.
- [12] S. G. Tringe, C. von Mering, A. Kobayashi, A. a. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. "Comparative metagenomics of microbial communities." *Science* 308 (5721): 554–7, 2005.
- [13] J. Gans, M. Wolinsky, and J. Dunbar. "Computational improvements reveal great bacterial diversity and high metal toxicity in soil." *Science* 309 (5739): 1387–90, 2005.
- [14] N. J. Loman, C. Constantinidou, J. Z. M. Chan, M. Halachev, M. Sergeant, C. W. Penn, E. R. Robinson, and M. J. Pallen. "High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity." *Nature reviews. Microbiology* 10 (9): 599–606, 2012.
- [15] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. "Comparison of next-generation sequencing systems." *Journal of biomedicine & biotechnology* 2012 (1): 251364, 2012.
- [16] M. a. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." *BMC genomics* 13 (1): 341, 2012.
- [17] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth, and J. Bustillo. "An integrated semiconductor device enabling non-optical genome sequencing." *Nature* 475 (7356): 348–52, 2011.
- [18] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. "Real-time DNA sequencing from single polymerase molecules." *Science* 323 (5910): 133–8, 2009.
- [19] E. B. Fichot and R. S. Norman. "Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform". en. *Microbiome* 1 (1): 10, 2013.
- [20] J. C. Wooley, A. Godzik, and I. Friedberg. "A primer on metagenomics." *PLoS computational biology* 6 (2): e1000667, 2010.

- [21] Y. Ye and H. Tang. “An ORFome assembly approach to metagenomics sequences analysis.” *Journal of bioinformatics and computational biology* 7 (3): 455–71, 2009.
- [22] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. “Meta-IDBA: a de Novo assembler for metagenomic data”. *Bioinformatics* 27 (13): i94–i101, 2011.
- [23] J. Laserson, V. Jojic, and D. Koller. “Genovo: De Novo Assembly for Metagenomes”. *Journal of Computational Biology* 18 (3): 429–443, 2011.
- [24] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. “MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads.” *Nucleic acids research* 40 (20): e155, 2012.
- [25] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. “Use of simulated data sets to evaluate the fidelity of metagenomic processing methods.” *Nature methods* 4 (6): 495–500, 2007.
- [26] J. Wooley and Y. Ye. “Metagenomics: facts and artifacts, and computational challenges”. *Journal of computer science and technology* 25 (1): 71–81, 2009.
- [27] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. r. L. Nielsen, G. W. Tyson, and P. H. Nielsen. “Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.” *Nature biotechnology* 31 (6): 533–8, 2013.
- [28] G. Muyzer, E. C. de Waal, and A. G. Uitterlinden. “Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA.” *Applied and environmental microbiology* 59 (3): 695–700, 1993.
- [29] M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight. “Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex.” *Nature methods* 5 (3): 235–7, 2008.
- [30] N. Shah, H. Tang, T. G. Doak, and Y. Ye. “Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics.” *Pacific Symposium on Biocomputing* 16: 165–76, 2011.
- [31] S. Chatterji, I. Yamazaki, Z. Bai, and J. Eisen. “CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads”. *Research in Computational Molecular Biology*. 2007, 1–19. arXiv: 0708.3098.
- [32] A. Kislyuk, S. Bhatnagar, J. Dushoff, and J. S. Weitz. “Unsupervised statistical clustering of environmental shotgun sequences.” *BMC bioinformatics* 10 (1): 316, 2009.
- [33] A. L. Bazinet and M. P. Cummings. “A comparative evaluation of sequence classification programs.” *BMC bioinformatics* 13 (1): 92, 2012.
- [34] W. Gerlach and J. Stoye. “Taxonomic classification of metagenomic shotgun sequences with CARMA3.” *Nucleic acids research* 39 (14): e91, 2011.
- [35] M. Horton, N. Bodenhausen, and J. Bergelson. “MARTA: a suite of Java-based tools for assigning taxonomic status to DNA sequences.” *Bioinformatics (Oxford, England)* 26 (4): 568–9, 2010.

- [36] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. "Integrative analysis of environmental sequences using MEGAN4." *Genome research* 21 (9): 1552–60, 2011.
- [37] B. Liu, T. Gibbons, M. Ghodsi, T. Treangen, and M. Pop. "Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences." *BMC genomics* 12 (Suppl 2): S4, 2011.
- [38] G. L. Rosen, E. R. Reichenberger, and A. M. Rosenfeld. "NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads." *Bioinformatics* 27 (1): 127–9, 2011.
- [39] A. Brady and S. L. Salzberg. "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models." *Nature methods* 6 (9): 673–6, 2009.
- [40] A. C. McHardy, H. G. Martín, A. Tsigos, P. Hugenholtz, and I. Rigoutsos. "Accurate phylogenetic classification of variable-length DNA fragments." *Nature methods* 4 (1): 63–72, 2007.
- [41] M. Stark, S. A. Berger, A. Stamatakis, and C. von Mering. "MLTreeMap—accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies." *BMC genomics* 11 (1): 461, 2010.
- [42] F. Schreiber, P. Gumrich, R. Daniel, and P. Meinicke. "TreePhyler: fast taxonomic profiling of metagenomes." *Bioinformatics (Oxford, England)* 26 (7): 960–1, 2010.
- [43] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. "Metagenomic microbial community profiling using unique clade-specific marker genes." *Nature methods* 9 (8): 811–4, 2012.
- [44] E. M. Glass, J. Wilkening, A. Wilke, D. Antonopoulos, and F. Meyer. "Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes." *Cold Spring Harbor protocols* 2010 (1): pdb.prot5368, 2010.
- [45] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. "MEGAN analysis of metagenomic data." *Genome research* 17 (3): 377–86, 2007.
- [46] G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj. "Metagenome fragment classification using N-mer frequency profiles." *Advances in bioinformatics* 2008 (1): 205969, 2008.
- [47] M. Wu and J. A. Eisen. "A simple, fast, and accurate method of phylogenomic inference." *Genome biology* 9 (10): R151, 2008.
- [48] J. W. Fickett and C. S. Tung. "Assessment of protein coding measures." *Nucleic acids research* 20 (24): 6441–50, 1992.
- [49] D Kulp, D Haussler, M. G. Reese, and F. H. Eeckman. "A generalized hidden Markov model for the recognition of human genes in DNA." *Proceedings International Conference on Intelligent Systems for Molecular Biology*. Vol. 4. 1996, 134–42.
- [50] A Krogh. "Two methods for improving performance of an HMM and their application for gene finding." *Proceedings International Conference on Intelligent Systems for Molecular Biology* 5: 179–86, 1997.

- [51] V Solovyev and A Salamov. “The Gene-Finder computer tools for analysis of human and model organisms genome sequences.” *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* 5: 294–302, 1997.
- [52] A. L. Delcher, D Harmon, S Kasif, O White, and S. L. Salzberg. “Improved microbial gene identification with GLIMMER.” *Nucleic acids research* 27 (23): 4636–41, 1999.
- [53] S. E. Cawley, A. I. Wirth, and T. P. Speed. “Phat—a gene finding program for Plasmodium falciparum.” *Molecular and biochemical parasitology* 118 (2): 167–74, 2001.
- [54] I. Korf. “Gene finding in novel genomes.” *BMC bioinformatics* 5 (1): 59, 2004.
- [55] N. G. Yok and G. L. Rosen. “Combining gene prediction methods to improve metagenomic gene annotation.” *BMC bioinformatics* 12 (1): 20, 2011.
- [56] L. D. Altschul S, Gish W, Miller W, Myers E. “Basic Local Alignment Search Tool”. *J Mol Biol* 215 (3): 403–410, 1990.
- [57] D Frishman, A Mironov, H. W. Mewes, and M Gelfand. “Combining diverse evidence for gene recognition in completely sequenced bacterial genomes.” *Nucleic acids research* 26 (12): 2941–7, 1998.
- [58] J. H. Badger and G. J. Olsen. “CRITICA: coding region identification tool invoking comparative analysis.” *Molecular biology and evolution* 16 (4): 512–24, 1999.
- [59] K. J. Hoff. “The effect of sequencing errors on metagenomic gene prediction.” *BMC genomics* 10 (1): 520, 2009.
- [60] H. Noguchi, T. Taniguchi, and T. Itoh. “MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes.” *DNA research : an international journal for rapid publication of reports on genes and genomes* 15 (6): 387–96, 2008.
- [61] K. J. Hoff, T. Lingner, P. Meinicke, and M. Tech. “Orphelia: predicting genes in metagenomic sequencing reads.” *Nucleic acids research* 37 (Web Server issue): W101–5, 2009.
- [62] M. Rho, H. Tang, and Y. Ye. “FragGeneScan: predicting genes in short and error-prone reads.” *Nucleic acids research* 38 (20): e191, 2010.
- [63] W. Zhu, A. Lomsadze, and M. Borodovsky. “Ab initio gene identification in metagenomic sequences.” *Nucleic acids research* 38 (12): e132, 2010.
- [64] D. R. Kelley, B. Liu, A. L. Delcher, M. Pop, and S. L. Salzberg. “Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering.” *Nucleic acids research* 40 (1): e9, 2012.
- [65] C. De Filippo, M. Ramazzotti, P. Fontana, and D. Cavalieri. “Bioinformatic approaches for functional annotation and pathway inference in metagenomics data”. *Briefings in Bioinformatics* 13 (6): 696–710, 2012.
- [66] T. Thomas, J. Gilbert, and F. Meyer. “Metagenomics - a guide from sampling to data analysis.” en. *Microbial informatics and experimentation* 2 (1): 3, 2012.
- [67] A. V. Lukashin and M Borodovsky. “GeneMark.hmm: new solutions for gene finding.” *Nucleic acids research* 26 (4): 1107–15, 1998.

- [68] D. R. Kelley and S. L. Salzberg. “Clustering metagenomic sequences with interpolated Markov models.” *BMC bioinformatics* 11 (1): 544, 2010.
- [69] K. E. Wommack, J. Bhavsar, and J. Ravel. “Metagenomics: read length matters.” *Applied and environmental microbiology* 74 (5): 1453–63, 2008.
- [70] H. Li and R. Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform.” *Bioinformatics (Oxford, England)* 25 (14): 1754–60, 2009.
- [71] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.” *Genome biology* 10 (3): R25, 2009.
- [72] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” *Nucleic acids research* 25 (17): 3389–402, 1997.
- [73] B. Niu, Z. Zhu, L. Fu, S. Wu, and W. Li. “FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes.” *Bioinformatics (Oxford, England)* 27 (12): 1704–5, 2011.
- [74] F. Meyer, R. Overbeek, and A. Rodriguez. “FIGfams: yet another set of protein families.” *Nucleic acids research* 37 (20): 6643–54, 2009.
- [75] R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H.-y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. “The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.” *Nucleic acids research* 33 (17): 5691–702, 2005.
- [76] I. Letunic, T. Doerks, and P. Bork. “SMART 7: recent updates to the protein domain annotation resource.” *Nucleic acids research* 40 (Database issue): D302–5, 2012.
- [77] A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant. “CDD: a Conserved Domain Database for the functional annotation of proteins.” *Nucleic acids research* 39 (Database issue): D225–9, 2011.
- [78] R. D. Finn, J. Mistry, J. Tate, P. Coghill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman. “The Pfam protein families database.” *Nucleic acids research* 38 (Database issue): D211–22, 2010.
- [79] D. H. Haft. “TIGRFAMs: a protein family resource for the functional identification of proteins”. *Nucleic Acids Research* 29 (1): 41–43, 2001.
- [80] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. “The COG database: a tool for genome-scale analysis of protein functions and evolution.” *Nucleic acids research* 28 (1): 33–6, 2000.

- [81] W. Klimke, R. Agarwala, A. Badretdin, S. Chetvernin, S. Ciufo, B. Fedorov, B. Kiryutin, K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova. "The National Center for Biotechnology Information's Protein Clusters Database." *Nucleic acids research* 37 (Database issue): D216–23, 2009.
- [82] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. "KEGG for integration and interpretation of large-scale molecular data sets." *Nucleic acids research* 40 (Database issue): D109–14, 2012.
- [83] T. Gene and O. Consortium. "Gene Ontology : tool for the unification of biology". *Nature genetics* 25 (1): 25–29, 2000.
- [84] R. D. Finn, J. Clements, and S. R. Eddy. "HMMER web server: interactive sequence similarity searching." *Nucleic acids research* 39 (Web Server issue): W29–37, 2011.
- [85] W. J. Kent. "BLAT — The BLAST-Like Alignment Tool". *Genome research* 12 (4): 656–664, 2002.
- [86] K. D. Pruitt, T. Tatusova, and D. R. Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic acids research* 35 (Database issue): D61–5, 2007.
- [87] T. U. Consortium. "Reorganizing the protein space at the Universal Protein Resource (UniProt)." *Nucleic acids research* 40 (Database issue): D71–5, 2012.
- [88] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork. "eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges." *Nucleic acids research* 40 (Database issue): D284–9, 2012.
- [89] Y. Zhao, H. Tang, and Y. Ye. "RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data." *Bioinformatics (Oxford, England)* 28 (1): 125–6, 2012.
- [90] S. Sun, J. Chen, W. Li, I. Altintas, A. Lin, S. Peltier, K. Stocks, E. E. Allen, M. Ellisman, J. Grethe, and J. Wooley. "Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource." *Nucleic acids research* 39 (Database issue): D546–51, 2011.
- [91] V. M. Markowitz, I.-m. A. Chen, K. Chu, E. Szeto, K. Palaniappan, Y. Grechkin, A. Ratner, B. Jacob, A. Pati, M. Huntemann, K. Liolios, I. Pagani, I. Anderson, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides. "IMG/M: the integrated metagenome data management and comparative analysis system." *Nucleic acids research* 40 (Database issue): D123–9, 2012.
- [92] J. Goll, D. B. Rusch, D. M. Tanenbaum, M. Thiagarajan, K. Li, B. a. Methé, and S. Yooseph. "METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics." *Bioinformatics (Oxford, England)* 26 (20): 2631–2, 2010.
- [93] M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes, and P. Bork. "Smash-Community: a metagenomic annotation and analysis tool." *Bioinformatics (Oxford, England)* 26 (23): 2977–8, 2010.
- [94] S. Wu, Z. Zhu, L. Fu, B. Niu, and W. Li. "WebMGA: a customizable web server for fast metagenomic sequence analysis." *BMC genomics* 12 (1): 444, 2011.

- [95] T. Lingner, K. P. Asshauer, F. Schreiber, and P. Meinicke. “CoMet—a web server for comparative functional profiling of metagenomes.” *Nucleic acids research* 39 (Web Server issue): W518–23, 2011.
- [96] W. Li. “Analysis and comparison of very large metagenomes with fast clustering and functional annotation.” *BMC bioinformatics* 10 (1): 359, 2009.
- [97] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coghill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajananathan, P. D. Thomas, C. H. Wu, C. Yeats, and S.-Y. Yong. “InterPro in 2011: new developments in the family and domain prediction database.” *Nucleic acids research* 40 (Database issue): D306–12, 2012.
- [98] C. J. A. Sigrist, L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo. “PROSITE, a protein domain database for functional characterization and annotation.” *Nucleic acids research* 38 (Database issue): D161–6, 2010.
- [99] T. K. Attwood, P Bradley, D. R. Flower, A Gaulton, and N Maudling. “PRINTS and its automatic supplement , prePRINTS”. *Nucleic acids research* 31 (1): 400–402, 2003.
- [100] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguéz, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. “The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.” *Nucleic acids research* 39 (Database issue): D561–8, 2011.
- [101] T. Prakash and T. D. Taylor. “Functional assignment of metagenomic data: challenges and applications.” *Briefings in bioinformatics* 13 (6): 711–27, 2012.
- [102] J. M. Brulc, D. A. Antonopoulos, M. E. B. Miller, M. K. Wilson, A. C. Yannarell, E. A. Dinsdale, R. E. Edwards, E. D. Frank, J. B. Emerson, P. Wacklin, P. M. Coutinho, B. Henrissat, K. E. Nelson, and B. A. White. “Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (6): 1948–53, 2009.
- [103] K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, T. D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, M. Hattori, K. K. Urokawa, T. I. Toh, T. K. Uwahara, K. O. Shima, H. T. Oh, A. T. Oyoda, H. M. Ori, Y. O. Gura, D. S. E. Hrlich, K. I. Toh, T. T. Akagi, and Y. S. Akaki. “Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.” en. *DNA research : an international journal for rapid publication of reports on genes and genomes* 14 (4): 169–81, 2007.
- [104] NIH. “A Catalog of Reference Genomes from the Human Microbiome”. *Science* 328 (5981): 994–999, 2010.

- [105] B. Beszteri, B. Temperton, S. Frickenhaus, and S. J. Giovannoni. "Average genome size: a potential source of bias in comparative metagenomics." *The ISME journal* 4 (8): 1075–7, 2010.
- [106] E. van Nimwegen. "Scaling laws in the functional content of genomes." *Trends in genetics : TIG* 19 (9): 479–84, 2003.
- [107] J. Raes, J. O. Korb, M. J. Lercher, C. von Mering, and P. Bork. "Prediction of effective genome size in metagenomic samples." *Genome biology* 8 (1): R10, 2007.
- [108] P. J. Turnbaugh, M. Hamady, T. Yatsunencko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon. "A core gut microbiome in obese and lean twins." *Nature* 457 (7228): 480–4, 2009.
- [109] B. Liu. "Computational Metagenomics : Network , Classification and Assembly". PhD thesis. University of Maryland: USA, 2012.
- [110] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methé, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower. "Metabolic reconstruction for metagenomic data and its application to the human microbiome." *PLoS computational biology* 8 (6): e1002358, 2012.
- [111] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa. "KAAS: an automatic genome annotation and pathway reconstruction server." *Nucleic acids research* 35 (Web Server issue): W182–5, 2007.
- [112] F Meyer, D Paarmann, M D'Souza, R Olson, E. M. Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, J Wilkening, and R. a. Edwards. "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes." *BMC bioinformatics* 9 (1): 386, 2008.
- [113] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi. "Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology." *Briefings in bioinformatics* 11 (1): 40–79, 2010.
- [114] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." *Nucleic acids research* 40 (Database issue): D742–53, 2012.
- [115] Y. Ye and T. G. Doak. "A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes." *PLoS computational biology* 5 (8): e1000465, 2009.
- [116] I. Sharon, S. Bercovici, R. Y. Pinter, and T. Shlomi. "Pathway-based functional analysis of metagenomes." *Journal of computational biology : a journal of computational molecular cell biology* 18 (3): 495–505, 2011.
- [117] E. Kristiansson, P. Hugenholtz, and D. Dalevi. "ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes." *Bioinformatics (Oxford, England)* 25 (20): 2737–8, 2009.

- [118] D. H. Parks and R. G. Beiko. "Identifying biologically relevant differences between metagenomic communities." *Bioinformatics* 26 (6): 715–21, 2010.
- [119] B. Liu and M. Pop. "MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets." *BMC proceedings* 5 (Suppl 2): S9, 2011.
- [120] J. R. White, N. Nagarajan, and M. Pop. "Statistical methods for detecting differentially abundant features in clinical metagenomic samples." *PLoS computational biology* 5 (4), 2009.
- [121] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. "KEGG for representation and analysis of molecular networks involving diseases and drugs." *Nucleic acids research* 38 (Database issue): D355–60, 2010.
- [122] M. A. Oberhardt, B. O. Palsson, and J. A. Papin. "Applications of genome-scale metabolic reconstructions." *Molecular systems biology* 5 (1): 320, 2009.
- [123] C. Söhngen, A. Chang, and D. Schomburg. "Development of a classification scheme for disease-related enzyme information." *BMC bioinformatics* 12 (1): 329, 2011.
- [124] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. "High-throughput generation, optimization and analysis of genome-scale metabolic models." *Nature biotechnology* 28 (9): 977–82, 2010.
- [125] O. Dias, A. K. Gombert, E. C. Ferreira, and I. Rocha. "Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*." *BMC genomics* 13 (1): 517, 2012.
- [126] I. Rocha, P. Maia, P. Evangelista, P. Vilaça, S. a. Soares, J. P. Pinto, J. Nielsen, K. R. Patil, E. C. Ferreira, and M. Rocha. "OptFlux: an open-source software platform for in silico metabolic engineering." *BMC systems biology* 4 (1): 45, 2010.
- [127] N. E. Lewis, H. Nagarajan, and B. O. Palsson. "Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods." *Nature reviews. Microbiology* 10 (4): 291–305, 2012.
- [128] S. Stolýar, S. Van Dien, K. L. Hillesland, N. Pinel, T. J. Lie, J. a. Leigh, and D. a. Stahl. "Metabolic modeling of a mutualistic microbial community." *Molecular systems biology* 3 (92): 92, 2007.
- [129] R. Taffs, J. E. Aston, K. Brileya, Z. Jay, C. G. Klatt, S. McGlynn, N. Mallette, S. Montross, R. Gerlach, W. P. Inskip, D. M. Ward, and R. P. Carlson. "In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study". *BMC systems biology* 3 (114): 114, 2009.
- [130] N. Klitgord and D. Segrè. "Ecosystems biology of microbial metabolism." *Current opinion in biotechnology* 22 (4): 541–6, 2011.
- [131] A. R. Zomorodi and C. D. Maranas. "OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities". *PLoS Computational Biology* 8 (2): e1002363, 2012.

- [132] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. a. Schloss, V. Bonazzi, J. E. McEwen, K. a. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, a. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer. “The NIH Human Microbiome Project.” *Genome research* 19 (12): 2317–23, 2009.
- [133] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. “A bioinformatician’s guide to metagenomics.” *Microbiology and molecular biology reviews* : *MMBR* 72 (4): 557–78, Table of Contents, 2008.
- [134] T. Human and M. Project. “A framework for human microbiome research.” *Nature* 486 (7402): 215–21, 2012.
- [135] J. Consortium, H. Microbiome, P. Data, and G. Working. “Evaluation of 16S rDNA-based community profiling for human microbiome research.” *PloS one* 7 (6): e39315, 2012.
- [136] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. a. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. “Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.” *Applied and environmental microbiology* 75 (23): 7537–41, 2009.
- [137] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunencko, J. Zaneveld, and R. Knight. “QIIME allows analysis of high-throughput community sequencing data.” *Nature methods* 7 (5): 335–6, 2010.
- [138] K. Rotmistrovsky and R. Agarwala. “BMTagger : Best Match Tagger for removing human reads from metagenomics datasets BMTagger screening”: 2–7, 2011.
- [139] D. Gevers, M. Pop, P. D. Schloss, and C. Huttenhower. “Bioinformatics for the Human Microbiome Project”. *PLoS Computational Biology* 8 (11), 2012.
- [140] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. “De novo assembly of human genomes with massively parallel short read sequencing.” *Genome research* 20 (2): 265–72, 2010.
- [141] The Human Microbiome, P. Consortium, T. Human, and M. Project. “Structure, function and diversity of the healthy human microbiome.” *Nature* 486 (7402): 207–14, 2012.
- [142] E. A. Grice, H. H. Kong, G. Renaud, A. C. Young, C. S. Program, G. G. Bouffard, R. W. Blakesley, T. G. Wolfsberg, M. L. Turner, and J. A. Segre. “A diversity profile of the human skin microbiota”. *Genome research* 18 (1): 1043–1050, 2008.

- [143] J. I. Gordon, R. Knight, E. K. Costello, C. L. Lauber, M. Hamady, and N. Fierer. “Bacterial community variation in human body habitats across space and time.” *Science (New York, N.Y.)* 326 (5960): 1694–7, 2009.
- [144] L. M. Proctor. “The Human Microbiome Project in 2011 and beyond.” *Cell host & microbe* 10 (4): 287–91, 2011.
- [145] P. J. Turnbaugh and J. I. Gordon. “An invitation to the marriage of metagenomics and metabolomics.” *Cell* 134 (5): 708–13, 2008.
- [146] P. J. Turnbaugh, C. Quince, J. J. Faith, A. C. McHardy, T. Yatsunencko, F. Niazi, J. Affourtit, M. Egholm, B. Henrissat, R. Knight, and J. I. Gordon. “Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (16): 7503–8, 2010.
- [147] B. L. Cantarel, A. R. Erickson, N. C. VerBerkmoes, B. K. Erickson, P. a. Carey, C. Pan, M. Shah, E. F. Mongodin, J. K. Jansson, C. M. Fraser-Liggett, and R. L. Hettich. “Strategies for metagenomic-guided whole-community proteomics of complex microbial environments.” *PLoS one* 6 (11): e27173, 2011.
- [148] O. Dias and M. Rocha. “Merlin : Metabolic Models Reconstruction using Genome-Scale Information”. *Computer Applications in Biotechnology* 11 (1): 120–125, 2010.
- [149] D Glez-Peña, M Reboiro-Jato, P Maia, M Rocha, F Díaz, and F Fdez-Riverola. “AIBench: a rapid application development framework for translational research in biomedicine.” *Computer methods and programs in biomedicine* 98 (2): 191–203, 2010.
- [150] R. C. G. Holland, T. A. Down, M Pocock, A Prlić, D Huen, K James, S Foisy, A Dräger, A Yates, M Heuer, and M. J. Schreiber. “BioJava: an open-source framework for bioinformatics.” *Bioinformatics (Oxford, England)* 24 (18): 2096–7, 2008.
- [151] S. Patient, D. Wieser, M. Kleen, E. Kretschmann, M. Jesus Martin, and R. Apweiler. “UniProtJAPI: a remote API for accessing UniProt data.” *Bioinformatics (Oxford, England)* 24 (10): 1321–2, 2008.
- [152] M. H. Saier, C. V. Tran, and R. D. Barabote. “TCDB: the Transporter Classification Database for membrane transport protein analyses and information.” *Nucleic acids research* 34 (Database issue): D181–6, 2006.
- [153] A Krogh, B Larsson, G von Heijne, and E. L. Sonnhammer. “Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.” *Journal of molecular biology* 305 (3): 567–80, 2001.
- [154] O. Dias. “Reconstruction of the Genome-scale Metabolic Network of *Kluyveromyces Lactis*”. PhD thesis. University of Minho, 2013.
- [155] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. IZARD, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methé, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower. “Metabolic reconstruction for metagenomic data and its application to the human microbiome.” *PLoS computational biology* 8 (6): e1002358, 2012.

- [156] J. r. A. Aas, B. J. Paster, L. N. Stokes, I. Olsen, and F. E. Dewhirst. “Defining the normal bacterial flora of the oral cavity.” *Journal of clinical microbiology* 43 (11): 5721–32, 2005.
- [157] E. Zaura, B. J. F. Keijsers, S. M. Huse, and W. Crielaard. “Defining the healthy “core microbiome” of oral microbial communities.” *BMC microbiology* 9 (1): 259, 2009.
- [158] F. Yang, X. Zeng, K. Ning, K.-L. Liu, C.-C. Lo, W. Wang, J. Chen, D. Wang, R. Huang, X. Chang, P. S. Chain, G. Xie, J. Ling, and J. Xu. “Saliva microbiomes distinguish caries-active from healthy human populations.” *The ISME journal* 6 (1): 1–10, 2012.
- [159] J. Li, I. Nasidze, D. Quinque, M. Li, H.-P. Horz, C. André, R. M. Garriga, M. Halbwax, A. Fischer, and M. Stoneking. “The saliva microbiome of Pan and Homo.” *BMC microbiology* 13 (1): 204, 2013.
- [160] D. L. Mager, L. A. Ximenez-Fyvie, A. D. Haffajee, and S. S. Socransky. “Distribution of selected bacterial species on intraoral surfaces.” *Journal of clinical periodontology* 30 (7): 644–54, 2003.
- [161] B. Keijsers, E. Zaura, S. Huse, J. van der Vossen, F. Schuren, R. Montijn, J. ten Cate, and W. Crielaard. “Pyrosequencing analysis of the Oral Microflora of healthy adults”. *Journal of Dental Research* 87 (11): 1016–1020, 2008.
- [162] F. Gao and R. R. Zhang. “Enzymes are enriched in bacterial essential genes.” *PloS one* 6 (6): e21683, 2011.
- [163] A. Ploner. *Heatplus: Heatmaps with row and/or column covariates and colored clusters*. 2012.
- [164] M. S. Dillingham, M. Spies, and S. C. Kowalczykowski. “RecBCD enzyme is a bipolar DNA helicase.” *Nature* 423 (6942): 893–7, 2003.
- [165] G. R. Smith. “How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist’s view.” *Microbiology and molecular biology reviews : MMBR* 76 (2): 217–28, 2012.
- [166] M. Pignatelli and A. Moya. “Evaluating the fidelity of de novo short read metagenomic assembly using simulated data.” *PloS one* 6 (5): e19984, 2011.
- [167] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. r. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. r. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang. “A human gut microbial gene catalogue established by metagenomic sequencing.” *Nature* 464 (7285): 59–65, 2010.
- [168] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, A. Bousvaros, J. Korzenik, B. E. Sands, R. J. Xavier, and C. Huttenhower. “Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.” *Genome biology* 13 (9): R79, 2012.

- [169] M. J. Claesson, I. B. Jeffery, S. Conde, S. E. Power, E. M. O'Connor, S. Cusack, H. M. B. Harris, M. Coakley, B. Lakshminarayanan, O. O'Sullivan, G. F. Fitzgerald, J. Deane, M. O'Connor, N. Harnedy, K. O'Connor, D. O'Mahony, D. van Sinderen, M. Wallace, L. Brennan, C. Stanton, J. R. Marchesi, A. P. Fitzgerald, F. Shanahan, C. Hill, R. P. Ross, and P. W. O'Toole. "Gut microbiota composition correlates with diet and health in the elderly." *Nature* 488 (7410): 178–84, 2012.
- [170] B. Liu, L. L. Faller, N. Klitgord, V. Mazumdar, M. Ghodsi, D. D. Sommer, T. R. Gibbons, T. J. Treangen, Y.-C. Chang, S. Li, O. C. Stine, H. Hasturk, S. Kasif, D. Segrè, M. Pop, and S. Amar. "Deep sequencing of the oral microbiome reveals signatures of periodontal disease." *PloS one* 7 (6): e37919, 2012.
- [171] H. H. Kong, J. Oh, C. Deming, S. Conlan, E. A. Grice, M. A. Beatson, E. Nomicos, E. C. Polley, H. D. Komarow, P. R. Murray, M. L. Turner, and J. A. Segre. "Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis." *Genome research* 22 (5): 850–9, 2012.
- [172] D. Börnigen, X. C. Morgan, E. A. Franzosa, B. Ren, R. J. Xavier, W. S. Garrett, and C. Huttenhower. "Functional profiling of the gut microbiome in disease-associated inflammation". *Genome medicine* 5 (7): 1–13, 2013.
- [173] P. Wilmes and P. L. Bond. "Metaproteomics: studying functional gene expression in microbial ecosystems." *Trends in microbiology* 14 (2): 92–7, 2006.
- [174] S. Mitra, P. Rupek, D. C. Richter, T. Urich, J. a. Gilbert, F. Meyer, A. Wilke, and D. H. Huson. "Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG." *BMC bioinformatics* 12 (Suppl 1): S21, 2011.
- [175] P. I. Trosvik, K. Rudi, K. O. Straetkvern, K. S. Jakobsen, T. Naes, and N. C. Stenseth. "Web of ecological interactions in an experimental gut microbiota." *Environmental microbiology* 12 (10): 2677–87, 2010.
- [176] S. Shoaie, F. Karlsson, A. Mardinoglu, I. Nookaew, S. Bordel, and J. Nielsen. "Understanding the interactions between bacteria in the human gut through metabolic modeling". *Scientific Reports* 3: 1–10, 2013.

Appendix A

Supplementary material

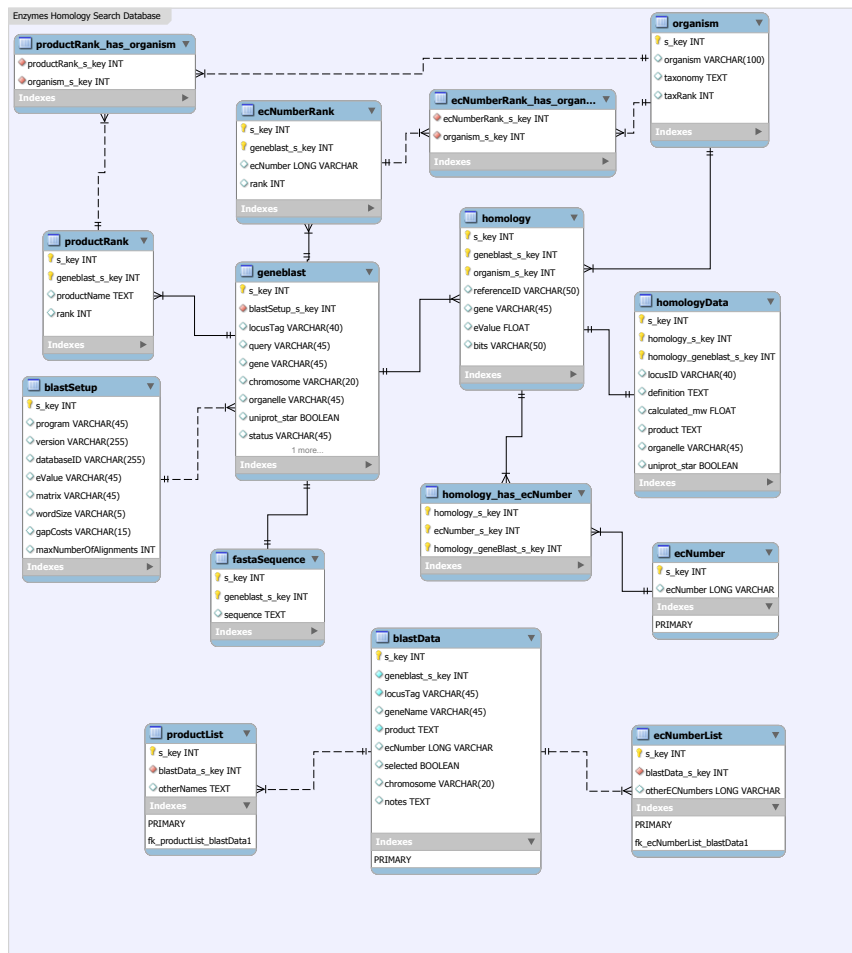


Figure A.1: Old database schema for data retrieved from homology searches.

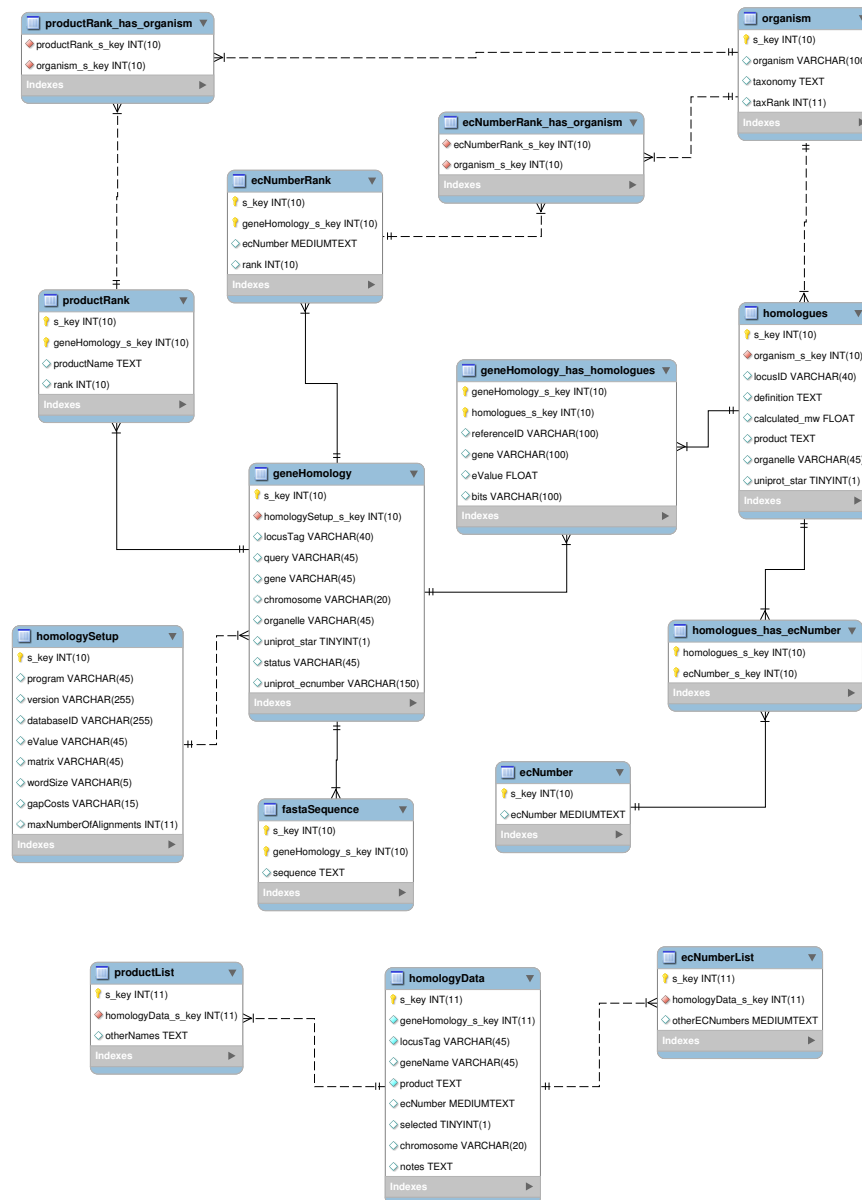


Figure A.2: New database schema for data retrieved from homology searches.