LIDo London Interdisciplinary Doctoral Programme
@ UCL, KCL, QMUL, Birkbeck, LSHTM, RVC

KING'S College LONDON

Creating a gene catalogue for future characterisation of bacterial-fungal community interactions in the oral cavity

Victoria Butt

1 Abstract

The microbial community residing in the human oral cavity is an essential component of oral health. There have been exhaustive studies investigating the composition of bacterial genes (bacteriome) in the oral cavity, yet fungal genes (mycobiome) are less well-characterised due to their lower abundance and to limitations in DNA extraction methods. To overcome this challenge, a concise catalogue containing both bacterial and fungal genes of the oral cavity will be created by processing and combining three large bacteriome-rich databases with literature searches and open-source datasets of the mycobiome. This catalogue will have several uses. Firstly, genes identified from future shotgun metagenomic sequencing will be mapped against this catalogue to detect low abundance fungal genes as well as bacterial genes. Secondly, this catalogue can also be used directly to generate genome-scale metabolic models (GEMs) for each bacterial or fungal species of interest. GEMs are large-scale metabolic networks that can be simulated *in silico* to infer intra- and inter-species phenotypic interactions. Here, I describe how the gene catalogue is created by trimming, filtering, alignment, assembly and quality control measures, using examples of raw genomic read data from oral samples in the Human Microbiome Project (HMP), a bacteriome-rich database.

2 Introduction

The human body has been estimated to be made up of approximately ten times more microbial cells than mammalian cells. The oral cavity is one area of the human body that has a characteristic composition of microbiota that interacts with the human host. In many cases, perturbation in the microbial homeostasis within microbial communities and between microbiota and host, known as dysbiosis, can lead to disease states. Microbes, that are present in a healthy oral cavity but have potential to cause disease, are "pathobionts" that can dominate in this dysbiosis. Pathobiants are microbes that are normally commensal (i.e. obtaining nutrients from the host but not having any harmful nor beneficial effects on the host), but have the potential to become pathogenic from environmental perturbations. *Candida albicans* is an example of a fungal pathobiont which is a dominant pathogen in the common infection, Candidiasis, but is also found as a commensal 50-80% of healthy human oral cavities [1, 2].

To begin to fully understand how entire microbial communities in the oral cavity maintain microbial homeostasis in healthy individuals, and how their dysbiosis can lead to infection and disease, all microbial genes existing in the oral cavity should be identified and quantified, and metabolic interactions across different species, should be characterised. In this study, we will focus on identifying and quantifying bacterial (bacteriome) and fungal (mycobiome) genes in the oral cavity, and investigating the metabolic interactions across taxa, including bacteria-fungi interactions.

2.1 Identifying and quantifying the bacteriome and mycobiome

Bacterial and fungal genes can be identified and quantified using next-generation sequencing (NGS), a sequencing method that can extract genomic information. In NGS, extracted genomic DNA is sheared into short fragments of approximately 100-1000 base pairs. Both ends of these fragments are ligated with adaptors that are required to initiate the sequencing reaction. These adaptors are extended with primers that attach to a surface to stabilise the fragment, such as a flow cell. During the sequencing reaction, single artificial nucleotides are sequentially bound against a fragment's nucleotide to produce a signal, such as an emission of light or change in pH, specific to the nucleotide type i.e. A, C, G or T. These signals are recorded and processed to create sequences of fragment reads. Algorithms are then used to remove adapter and primer reads or other contamination from these fragment reads. After these filtering procedures, the reads are assembled computationally to create large fragments, known as scaffolds. Other algorithms are then used to identify potential protein-coding genes from these scaffolds.

Extracting the microbial DNA for NGS can be achieved in one of two ways: using a targeted or metagenomic approach. Targeted sequencing is a method of extracting and amplifying specific loci of the genome that contains a great depth of taxonomical and phylogenetic information. The small-subunit ribosomal RNA (rRNA) loci (16S) in bacteria, and the internal transcribed space (ITS) loci (between the small-subunit rRNA and the large-subunit rRNA) in fungi are amplified to sequence bacterial and fungal species. In contrast, metagenomics is defined as the direct analysis of all genomic material contained within an environmental niche [3]. Instead of amplifying specific loci, DNA is extracted and immediately sheared into short fragments for NGS. Many studies have aimed to investigate the diversity of human microbial communities using targeted NGS, especially in the gut [4, 5]. Although targeted sequencing is a powerful method of extracting taxonomical information and relative abundance of individual taxa, it has its limitations. Firstly, certain loci have properties where they may be amplified relatively more or less than other loci by PCR, which may create a substantial bias in the relative abundances of the reads, and the subsequent identified taxa. Secondly, targeted sequencing cannot provide detailed information on the true functional diversity in a microbial community. Lastly, this method is limited to taxa for which markers in genetic loci are known and can be amplified. Metagenomic sequencing does not rely on pre-defined loci for amplification, and is free from amplification bias. Crucially, metagenomics can extract a greater diversity of genetic information, and thus resolve functional as well as taxonomical diversity from microbial communities.

Metagenomic analysis has led to the identification of at least 700 bacterial taxa comprising the human oral microbiota [6], but fungal communities still remain less characterised compared to bacterial communities. The composition of the mycobiome has only recently been considered a major component of the human microbiome. It has been estimated 99.1% of microbial genes were bacterial, with the rest being mostly archae, and only 0.1% eukaryotic (including fungal) and viral in origin [7, 8]. The mycobiome's contribution to the entire microbiome may be significantly underestimated for several reasons. Firstly, previous estimates have been reliant upon the available annotated microbial reference sequences of which fungi are hugely under-represented. Secondly, a typical fungal cell is 100-fold larger than a typical bacterial cell, meaning fungi account for larger biomass than number of genomes in microbial communities. Thirdly, fungi are eukaryotes that have metabolic features unique from prokaryotes, which may have important roles in microbial homeostasis. Lastly, it is clear that non-bacterial components of the microbiota respond to changes in diet or to dysbiosis, leading to effects in the immune system [9, 10].

Several studies have used ITS-targeted sequencing to investigate the oral mycobiome, with some agreement yet significant inconsistencies between them. Dupuy et al.'s discovered *Malassezia* as a prominent commensal, but missed core taxa, *Glomus*, *Teratosphaeria*, *Saccharomycetales* and *Dothioraceae*, which was found by Ghannoum and colleagues previously [1, 11]. The potential reasons for these differences between studies are extensive and may be somewhat due to limitations of targeted sequencing methods. To date, there have been no published metagenomic studies of the mycobiome. In the future, we will use metagenomics to quantify the mycobiome as well as the bacteriome.

To be able to quantify the bacteriome and mycobiome using shotgun metagenomic analysis, comprehensive reference catalogues need to be created that contain all possible bacterial and fungal genes that have been found to exist in the oral cavity. Lower abundance genes from metagenomic reads can be detected with higher sensitivity when mapped against a comprehensive reference catalogue, especially for fungal genes. The current catalogue is generated using fragments reads from a total of 682 samples, 382 from the Human Microbiome Project (HMP) [5], 265 from the a rheumatoid arthritis study [12] and 35 samples from King's College London [unpublished]. At this stage, a catalogue containing the mycobiome cannot be generated from these databases due to the type of samples, sample collection and library preparation. Instead, this catalogue contains mostly bacterial genes, but fungal genes will be added from external datasets and literature searches at a later date.

2.2 Inter-species metabolic interactions

Once the catalogue is complete, it will be used to identify metabolic interactions between species in the oral cavity using computational modelling. Models of metabolic networks need to be built for each of the 50 most abundant species before they can be simulated as follows. The genes within the catalogue will be annotated with their known associated biochemical reactions and metabolites of those reactions. These metabolites will be assembled to create a network of reactions called a genome-scale metabolic model (GEM) [13, 14]. A GEM can be represented visually by a graph of nodes, representing metabolites, and arrows representing the directions of the reactions. The GEM can also be represented mathematically by a stoichiometric matrix, S, where each row and column corresponds to a metabolite and reaction, respectively. Matrix values 1 or -1 correspond to a presence of a reactant or substrate, and 0 corresponds to an absence of a metabolite from a reaction [15]. To simulate these GEMs, the reactions fluxes, or rate of production or consumption of metabolites, need to be found for each reaction. This is achieved by setting mathematical constraints on metabolite production and consumption, such as setting a limit in the rate of carbon consumed in a reaction. This is known as constraint-based modelling (CBM). The purpose of CBM is to reduce the search space of the number of possible flux combinations that could occur, avoid non-physiological fluxes and reproduce similar environments seen *in vivo* [16, 17]. Next, a mathematical objective function is defined which aims to maximise (or minimise) the production of a particular metabolite or cellular phenotype. The fluxes are found by simulating the GEM with its constraints and objective function using optimisation algorithms, such as linear programming. To find inter-species interactions, the GEMs are paired and simulated together as described to find the pairwise fluxes, and thus the inter-species metabolic interactions.

Here, I describe the pipeline for creating the catalogue, looking specifically at processing the metagenomic reads of nine samples taken from the bacteriome-rich database, the HMP, and showing how quality control measures can identify and remove poor quality samples.

3 Materials and Methods

The pipeline for creating the gene catalogue from bacteriome-rich databases is described below. Nine samples, taken from the HMP database, are used here as an example of the entire analysis for all 682 samples from HMP [5], the rheumatoid arthritis study [12] and King's College London [unpublished]. The samples, SRR060075, SRR060076, SRR060040, SRR064436, SRR060154, SRR061324, SRR061359, SRR061490 and SRR061492, were downloaded from NCBI: ftp://ftp.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX023/ and ftp://ftp.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX024/ These files contain paired-end reads of sequenced DNA from the oral cavity that were sequenced from an Illumina HiSeq 2000. The following pipeline was run on a high performance cluster computer, "Rosalind" at King's College London. Refer to Supplementary Code (SC) https://github.com/blue-moon22/lido-thesis-2017 for code and parameters used.

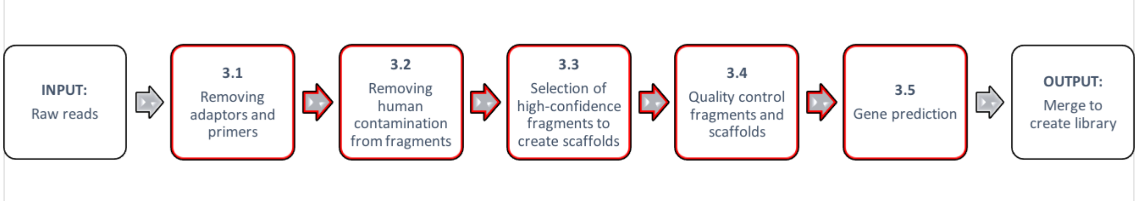
INPUT:
Raw reads
3.1
Removing adaptors and primers
3.2
Removing human contamination from fragments
3.3
Selection of high-confidence fragments to create scaffolds
3.4
Quality control fragments and scaffolds
3.5
Gene prediction
OUTPUT:
Merge to create library

Figure 1: Pipeline for creating the gene catalogue

3.1 Removal of adapters and primers from fragment reads

The raw reads for each sample were trimmed using AlienTrimmer (Version 0.4.0) [18]. AlienTrimmer was used to detect and remove contaminant sequences, adapters and primers, that are found at both ends of the reads and used as part of the sequencing protocol (SC Step 1).

3.2 Removal of human contamination from fragment reads

Human contaminant sequences were removed from the reads, since we are only interested in identifying microbial genes. This was achieved by mapping the reads against a human reference genome using Bowtie2 (Version 2.2.3), and removing those sequences that successfully matched with the human reference genome (SC Step 2). Bowtie2 aligns sequence reads to long reference sequences using memory-efficient algorithms based on the Burrows-Wheeler Transform [19].

3.3 Selection of high-confidence fragments to create scaffolds

To ensure the reads are completely derived from bacterial (and fungal) species and devoid of any other contamination, an additional filtering step was included (SC Step 3). Using Bowtie2, a reference genome was built from assembled scaffolds. These scaffolds were assembled from the trimmed and filtered fragment reads (generated in 3.2) using SPAdes (Version 3.9.0). SPAdes is a genome assembler tool based on resolving de Bruijn graphs, and specialises in single-cell bacterial and fungal read assemblies [20]. As the reads are metagenomic, the metaSPAdes pipeline was used for this assembly [21]. Again, using Bowtie2, the trimmed and filtered reads generated from 3.2 were aligned against the built reference genome. Different scaffolds were generated from fragment reads that aligned, and those reads that did not align were removed. The name, length, and number of mapped reads in each of these scaffolds were extracted using SAMtools (Version 1.3.1), a software tool used to read or write file formats containing scaffold information, and extract scaffold statistics [22]. Scaffolds with a length of less than 500 and containing less than 11 mapped reads were removed.

3.4 Quality control fragments and scaffolds

The quality of sequencing and the amount of contamination of the samples were considered as different quality control measurements. The quality of sequencing was measured by counting the number of fragment reads before and after trimming in 3.1. A large decrease in the number of reads may suggest problems with the sequencing protocol. The amount of contamination of the sample was measured by counting the number of fragment and scaffold reads before and after removal of potential contamination in 3.2 and 3.3. A large decrease in the number of reads may suggest a large amount of contamination. The measurements were calculated by dividing each count by the total pre-processed fragment/scaffold count for each sample, and multiplying by 1,000,000 to give the normalised read count per million (RPM) (SC Step 4). Samples with outlier counts were excluded from the gene catalogue.

3.5 Gene predictions

The filtered scaffolds for each sample that passed quality control was merged into one fastq file (SC Step 5). Bacterial genes from this file were identified using Prodigal (Version 2.6.3), a protein-coding gene prediction software, mainly used for bacterial and archaeal genomes [23]. Alternative software will be required to identify eukaryotic genes for future construction of the mycobiome content of the catalogue. Gene lengths of less than 60 and genes containing mycoplasm/spiroplasm code or mold, protozoan and coelenterate mitochondrial code were removed. This left potential bacterial (and archaeal) genes only. To create the final catalogue, the filtered genes were clustered using CD-HIT (Version 4.6.6), a greedy incremental clustering algorithm that is used to cluster sequences to remove redundancies [24].

4 Results

4.1 Quality control identifies outliers

A catalogue of bacterial genes was created from metagenomic reads of oral cavity samples taken from three bacteriome-rich databases. In the future, the catalogue will be used for addition of fungal genes to identify and quantify both the bacteriome and mycobiome in the oral cavity, and to investigate inter-species metabolic interactions across bacteria and fungi. To generate a reliable gene catalogue from metagenomic reads, the samples were quality controlled. Out of the nine samples from the HMP, eight were considered high quality and could be included as part of the final catalogue. The quality of sequencing and the amount of contamination of the samples were considered as different quality control measurements, with the quality of sequencing being the most important one. The quality of sequencing was quantified as the change in fragment reads per million (RPM) before and after trimming away adapter and primer reads from the fragments (first and second columns of Fig. 2a). The amount of contamination of the samples was quantified as the change in fragment and scaffold RPM before and after removal of potential (human) contamination (second and third columns of Fig. 2a, and both columns of Fig. 2b). Sample SRR060154 has the highest RPM of pre-processed raw reads and the greatest decrease after trimming compared to other samples in Fig. 2a. This sample's reads contained a high amount of adapters and primers, meaning there may have been problems with the sequencing protocol. Both Fig. 2a and 2b, show a relatively uniform distribution of fragment and scaffold RPMs before and after removal of contamination. Since the quality of the reads are most dependent on the quality of the sequencing, sample SRR060154 was excluded from the catalogue.

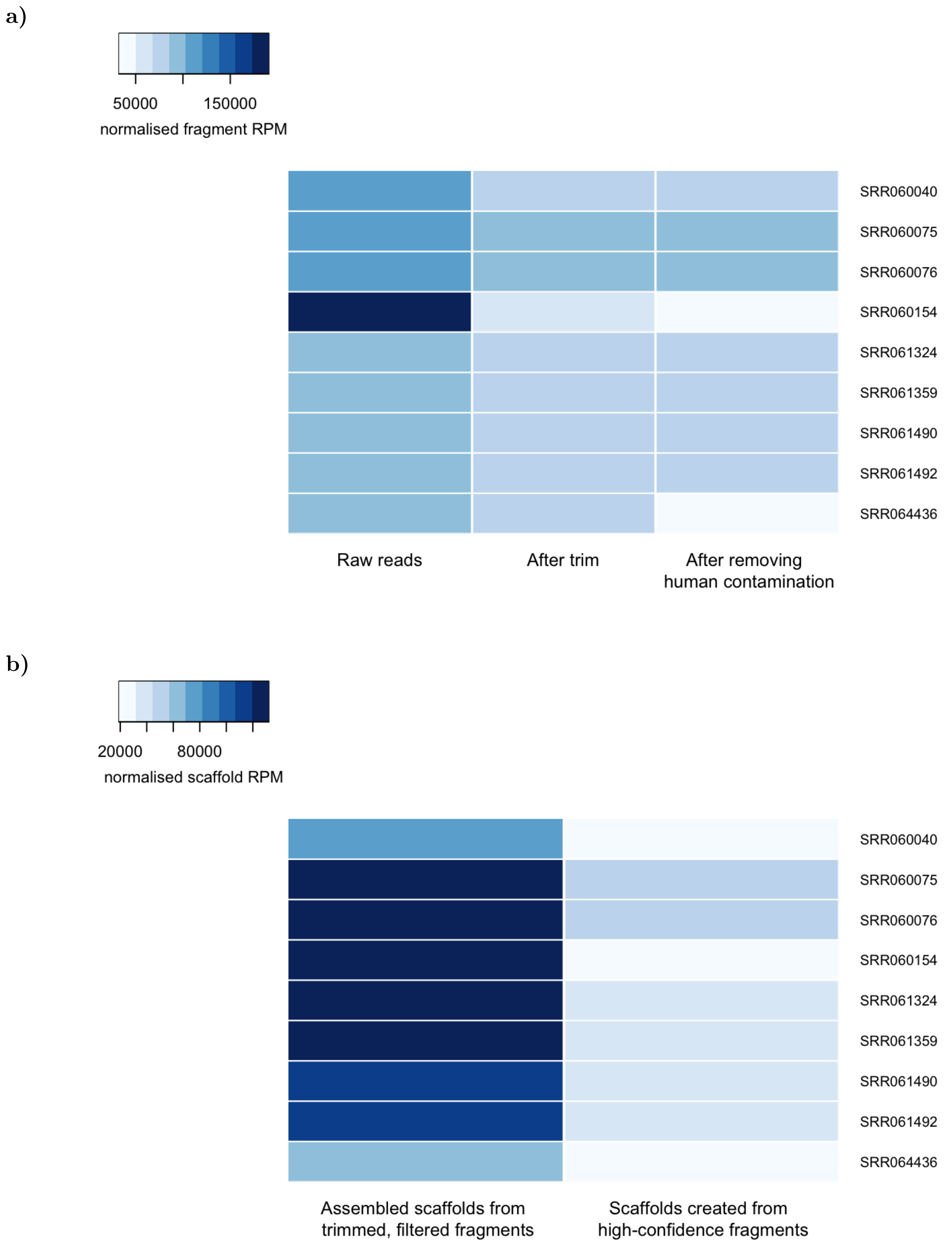

Figure 2: For each of the nine samples, the normalised reads per million (RPM) of a) fragment reads before and after each filtering step: before trimming (pre-processed raw reads), after trimming by AlienTrimmer, and after removal of human contamination; b) scaffolds assembled from trimmed and filtered fragments, and scaffolds created from high-confidence fragments aligned to a reference genome built from the assembled scaffolds

5 Discussion

In metagenomic studies of the human microbiome, there has been a no coverage to date of the fungal genome for several reasons. Firstly, there are very few comprehensively annotated catalogues of fungal genes in the human microbiome. Secondly, many common DNA extraction protocols are limited in extracting fungal DNA due to the difficulty in breaking melanin in some fungal cell walls. Several studies have successfully extracted fungal DNA using several different protocols, but the species predictions have not been completely consistent [25].

We will overcome the first issue by creating a comprehensive catalogue of both bacterial and fungal genes that exist in the oral cavity. At present, the catalogue has been successfully generated from the pipeline to include mostly bacterial genes, and fungal genes will need to be added at a later stage from external, open-source datasets and literature searches. To address the second issue, we will be investigating various DNA extraction protocols that are optimised for extracting fungal DNA from saliva for metagenomic analysis. Thus far, we have attempted to extract fungal DNA using an adapted extraction method described previously [26] and the MoBio PowerSoil DNA Isolation Kit, with limited success. We will trial other extraction protocols, including MolYsis as described previously [27] and other protocols from studies that have extracted fungal DNA for ITS-targeted sequencing. It is hoped the final optimised protocol will be used to extract bacterial and fungal DNA from human saliva samples, and mapped against the catalogue to give a personalised profile of the bacteriome and mycobiome for an individual.

GEMs can be created from the current bacterial gene-rich catalogue for each bacterial species present, and can be computed to infer large-scale bacterial community interactions in the oral cavity. To aid the precision of the GEMs, transcriptomic data will also be integrated into the catalogue to determine the expression status of a gene, identify potential gene isoforms that confer different metabolic functions and to consider any metabolic feedback loops. Once the fungal genes are added to the catalogue, the GEMs can be iterated to include fungal community interactions as well as bacterial. As well as being able to model metabolic interactions on a large scale, GEMs have the potential to generate a profile of anti-microbial resistance genes (ARGs) in the microbial community, also known as the "resistome" [28]. ARG profiles would allow identification of genes involved in the resistance of antimicrobial drugs, especially genes where their involvement is not obvious from targeted studies. Novel antibiotics can be developed based on these anti-microbial genes products, which have potential to tackle the current and future world's antibiotic-resistance crisis. At a later stage, optimised DNA extraction methods and metagenomic shotgun sequencing of human saliva samples could generate personalised GEM profiles for an individual. Using personalised GEMs that are modelled to an individual's microbial community could reveal how an individual's microbiota will affect their response to certain antimicrobial drugs.

In the future, a comprehensive gene catalogue of the bacteriome and mycobiome, and its uses in accurate fungal and bacterial species identification and creation of GEMs, has the potential to identify therapeutic targets, such as influential genes that could be targeted by novel antimicrobial drugs, and diagnostic biomarkers to give precise disease status.

6 Acknowledgements

I would like to thank my supervisors, Dr David Moyes and Dr Peter Mullany for their support, Dr David Gomez-Cabrero Lopez for sharing the code, Dr Elizabeth Witherden for help with fungal DNA extraction, and Dr Saeed Shoaie for information on GEMs.

References

[1] M. A. Ghannoum, R. J. Jurevic, P. K. Mukherjee, F. Cui, M. Sikaroodi, A. Naqvi, and P. M. Gillevet, "Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals," *PLoS Pathogens*, vol. 6, p. e1000713, Jan. 2010.

[2] D. Williams and M. Lewis, "Pathogenesis and treatment of oral candidosis," *Journal of Oral Microbiology*, vol. 3, Jan. 2011.

[3] D. D. Roumpeka, R. J. Wallace, F. Escalettes, I. Fotheringham, and M. Watson, "A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data," *Frontiers in Genetics*, vol. 8, p. 23, 2017.

[4] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon, "Human gut microbiome viewed across age and geography," *Nature*, vol. 486, pp. 222–227, May 2012.

[5] Human Microbiome Project Consortium, "A framework for human microbiome research," *Nature*, vol. 486, pp. 215–221, June 2012.

[6] H. F. Jenkinson, "Beyond the oral microbiome," *Environmental Microbiology*, vol. 13, pp. 3077–3087, Dec. 2011.

[7] 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. 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. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, MetaHIT Consortium, P. Bork, S. D. Ehrlich, and J. Wang, "A human gut microbial gene catalogue established by metagenomic sequencing," *Nature*, vol. 464, pp. 59–65, Mar. 2010.

[8] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, MetaHIT Consortium, P. Bork, J. Wang, and MetaHIT Consortium, "An integrated catalog of reference genes in the human gut microbiome," *Nature Biotechnology*, vol. 32, pp. 834–841, Aug. 2014.

[9] S. Devkota, Y. Wang, M. W. Musch, V. Leone, H. Fehlner-Peach, A. Nadimpalli, D. A. Antonopoulos, B. Jabri, and E. B. Chang, "Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in Il10-/- mice," *Nature*, vol. 487, pp. 104–108, July 2012.

[10] L. Bull-Otterson, W. Feng, I. Kirpich, Y. Wang, X. Qin, Y. Liu, L. Gobejishvili, S. Joshi-Barve, T. Ayvaz, J. Petrosino, M. Kong, D. Barker, C. McClain, and S. Barve, "Metagenomic analyses of alcohol induced pathogenic alterations in the intestinal microbiome and the effect of Lactobacillus rhamnosus GG treatment," *PloS One*, vol. 8, no. 1, p. e53028, 2013.

[11] A. K. Dupuy, M. S. David, L. Li, T. N. Heider, J. D. Peterson, E. A. Montano, A. Dongari-Bagtzoglou, P. I. Diaz, and L. D. Strausbaugh, "Redefining the Human Oral Mycobiome with Improved Practices in Amplicon-based Taxonomy: Discovery of Malassezia as a Prominent Commensal," *PLoS ONE*, vol. 9, p. e90899, Mar. 2014.

[12] X. Zhang, D. Zhang, H. Jia, Q. Feng, D. Wang, D. Liang, X. Wu, J. Li, L. Tang, Y. Li, Z. Lan, B. Chen, Y. Li, H. Zhong, H. Xie, Z. Jie, W. Chen, S. Tang, X. Xu, X. Wang, X. Cai, S. Liu, Y. Xia, J. Li, X. Qiao, J. Y. Al-Aama, H. Chen, L. Wang, Q.-j. Wu, F. Zhang, W. Zheng, Y. Li, M. Zhang, G. Luo, W. Xue, L. Xiao, J. Li, W. Chen, X. Xu, Y. Yin, H. Yang, J. Wang, K. Kristiansen, L. Liu, T. Li, Q. Huang, Y. Li, and J. Wang, "The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment," *Nature Medicine*, vol. 21, pp. 895–905, July 2015.

[13] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen, "The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum," *PLoS computational biology*, vol. 9, no. 3, p. e1002980, 2013.

[14] G. J. E. Baart and D. E. Martens, "Genome-scale metabolic models: reconstruction and analysis," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 799, pp. 107–126, 2012.

[15] E. J. O'Brien, J. M. Monk, and B. O. Palsson, "Using Genome-scale Models to Predict Biological Capabilities," *Cell*, vol. 161, pp. 971–987, May 2015.

[16] S. Magnúsdóttir, A. Heinken, L. Kutt, D. A. Ravcheev, E. Bauer, A. Noronha, K. Greenhalgh, C. Jäger, J. Baginska, P. Wilmes, R. M. T. Fleming, and I. Thiele, "Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota," *Nature Biotechnology*, vol. 35, pp. 81–89, Jan. 2017.

[17] A. Bordbar, J. M. Monk, Z. A. King, and B. O. Palsson, "Constraint-based models predict metabolic and associated cellular functions," *Nature Reviews Genetics*, vol. 15, pp. 107–120, Feb. 2014.

[18] A. Criscuolo and S. Brisse, "AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads," *Genomics*, vol. 102, pp. 500–506, Nov. 2013.

[19] B. Langmead, "Aligning short sequencing reads with Bowtie," *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. CHAPTER, pp. Unit–11.7, Dec. 2010.

[20] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," *Journal of Computational Biology*, vol. 19, pp. 455–477, Apr. 2012.

[21] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaSPAdes: a new versatile metagenomic assembler," *Genome Research*, vol. 27, pp. 824–834, May 2017.

[22] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2078–2079, Aug. 2009.

[23] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, p. 119, 2010.

[24] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics (Oxford, England)*, vol. 22, pp. 1658–1659, July 2006.

[25] A. Vesty, K. Biswas, M. W. Taylor, K. Gear, and R. G. Douglas, "Evaluating the Impact of DNA Extraction Method on the Representation of Human Oral Bacterial and Fungal Communities," *PloS One*, vol. 12, no. 1, p. e0169877, 2017.

[26] J.-P. Furet, O. Firmesse, M. Gourmelon, C. Bridonneau, J. Tap, S. Mondot, J. Doré, and G. Corthier, "Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR," *FEMS Microbiology Ecology*, vol. 68, pp. 351–362, June 2009.

[27] C. D. McCann and J. A. Jordan, "Evaluation of MolYsis$^{\text{TM}}$ Complete5 DNA Extraction Method for Detecting Staphylococcus aureus DNA from Whole Blood in a Sepsis Model Using PCR/Pyrosequencing," *Journal of microbiological methods*, vol. 99, pp. 1–7, Apr. 2014.

[28] G. D. Wright, "The antibiotic resistome: the nexus of chemical and genetic diversity," *Nature Reviews. Microbiology*, vol. 5, pp. 175–186, Mar. 2007.