Computationally analysing bacterial interactions through gene function

Supervisor: Aaron Comeault

## 

## Aim

Making novel genetic comparisons between bacterial genera, with emphasis on differences arising from environment or host, through links between its enironment, its evolutionary history and the functions of the proteins coded for.

(very wordy, but its all there, could prob do with cutting down but i dont know what to cut without messing up the whole thing)

possible cutdown:

Using an understanding of bacterial ecology and evolution to draw novel [conclusions/comparisons] about differences in the genomic function of bacterial genera

or

explore the range of bacterial functions and how they interact using computational genomics

(i still cant pin down this bit)

## Abstract

Bacterial Microbiomes are cryptic environments, in which bacteria cooperate and compete in many different ways. Understanding these relationships could have consequences for many other fields, including health and conservation, due to the universal presence of bacteria. Due to the microscopic nature of these ecosystems, genomics is one of the most reliable ways to study them, and much progress has been made on this in the field’s 30 year history.

Few research projects explore using publicly available data. Using this can be a way of quickly drawing new conclusions about the way different bacteria interact. This project will use computational environments such as R to analyse important functional differences, using the genes of sampled bacteria, and hypothesise how these might affect host organisms and other bacteria in the microbiome. This will lead to a greater understanding of these environments.

(138 words) [i still dont feel confident about this, its better, but is it good?, the plan section for this is confusing]

## Introduction

Bacterial genomics is a very interesting field, it has evolved a lot over the 30 years the field has existed. The first bacterial genome to be sequenced was that of *H. influenzae* in 1995 using shotgun sequencing (Fleischmann et al. 1995). Since then there have been many improvements, allowing a huge breadth of Bacterial Phyla to be discovered and better understood. The knowledge gained through understanding the functional components of bacterial genomes can have implications for many fields, such as: bioforensics; disease treatment; biosurveillance and metabolic modeling (Land et al. 2015). There have been 3 generations of technology, with developments being made into the next every year. This has brought even more improvements, for example through sequencing methods like Oxford Nanopore (Abdi et al. 2024). Many studies have utilised these methods, however few discuss microbe genomics in the context of relationships and interactions between microbiomes and their environment or host, and the factors that effect these relationships, one such paper discussed how the microbiome around a species of potato plant (*Solanum tuberosum*), in many fields including crop yield, how these relationships would evolve in a changeing climate was also discussed. This is a good example of why these analyses are important, potatoes are a very important food globally and ensuring food stability is vital for future development (Petrushin, Filinova, and Gutnik 2024).

The human gut microbiome has been studied frequently, probably due to the relationship between it and public health, for example diseased like Crohn’s disease. One relevant study found that there was a relationship between microbiome size and variation. This analysis phenotyped the samples through barcoding the 16s rRNA gene. Discuss the interaction between host and microbiome(Vandeputte et al. 2017). Another study presented that there can be a huge variation and diversity of skin microbiome of a single host in amphibians. It was also found that members of the same species showed similar microbiomes, whereas there were fewer similarities between different individuals from the same pond (McKenzie et al. 2012). This presents that any member of the same species, regardless of sampling location, can be used as a model for the species and that these microbiomes have spent millenia adapting to their hosts, which will have no doubt caused complex relationships to develop. Another previous study found that the skin microbiome is an important factor in disease survival. There are other important elements to the skin of amphibians, for example respiration, that the present microbiome will effect, making a more whole picture of the relationships and functions therein vital (Harris et al. 2009). The practical uses of bacteria in the context of sustainability are varied. A recent study outlined how bacteria are being used to recycle waste gases or biomass and produce products like ethanol, other researchers have found bacteria that break down certain plastics(Rappuoli et al. 2025).

Relationships between bacteria have been studied previously, in the context of cooperation in biofilm formation. Both antagonism and cooperation were found in the bacterial species studied. The sample was taken from drinking water, six species were identified and 96 agar plates with different combinations of these were created. Previous studies had found that cooperation lead to more stable communities and a few specific beneficial relationships were identified, for example, plasmid conjugation (Simões, Simões, and Vieira 2007). A recent study by (Lunjani et al. 2021) discussed host-microbe interactions in human skin with relevance to how modern factors are changing these relationships and causing inflammatory skin diseases. The study outlines how this knowledge could help in creating personalised probiotic solutions to this problem. There are many ways in which bacteria in a microbiome can interact, as mentioned previously, they can cooperate and produce biofilms and share nutrients, however, there also exists competition wherever there is a niche to exploit and limited resources. These relationships cause selection for genes that cause, for example, the production of antimicrobial compounds to inhibit the competitors.

There are many species of bacteria, as such there are almost limitless ways in which they interact with each other in different conditions, gaining better insight into some of the ways in which these relationships occur on a genetic level could have impacts for fields as broad as public health to pollution to zoonotic disease and conservation. This study will aim to investigate some of these relationships in a number of ways using publically available data, and present these in the context of how they might have a beneficial impact on potential host organisms. This work may be cause for future research to be done in the area.

(740)

## Methods

This project is desk-based, and all data analysis will be undertaken computationally. Thus, data must be collected from online sources. There are many data repositories online with free-publicly accessible data that can be utilized. This provides a benefit as the time to pull down this data is much shorter than the time taken to harvest samples from the field, then grow and sequence them in the lab. Relatively few papers specifically focus on this massive wealth of data. The repository used by this project is the American-run National Center for Biotechnology Information (NCBI) [genome database](https://www.ncbi.nlm.nih.gov/datasets/genome/), specifically focusing on the bacterial samples for this project. This will provide a huge amount of data with which I can create novel analyses and draw new conclusions about how the evolutionary history shapes the relationships between bacteria.

Another important database is the Kyoto Encyclopedia of Genes and Genomes (KEGG) [database](https://www.genome.jp/kegg/). This is important as it holds data on the function of genes, these are grouped into “Orthologs” ([KO values](https://www.genome.jp/kegg/ko.html)) of shared function. The database also has data on how these KOs fit into larger biological systems, such as metabolism. These are called “[pathways](https://www.genome.jp/kegg/pathway.html)”. With this information, i can directly correlate the genes expressed in the online bacterial samples to their potential functions in the bacteria and thus, draw conclusions about how those functions could shape relationships with potential hosts and other bacteria in the environment. Specifically, grouped analysis comparing genera will be done to visualise how the evolutionary differences between genera can cause differential gene expression.

The data will need to be transformed between these two databases as first the genes must be isolated from the larger genome and then they must be “[enriched](https://rdrr.io/github/YuLab-SMU/MicrobiomeProfiler/man/enrichKO.html)” to turn those genes into KO pathways. Hence, this will require the use of both local tools, including R and potentially SQL, as well as online tools such as the Hawk Super Computer, ran by [SuperComputing Wales](https://www.supercomputing.wales/). The module used to extract the genes or “annotate” the genomes will be [EggNOG-mapper/2.1.12](https://github.com/eggnogdb/eggnog-mapper/blob/master/README.md). This requires me to use command line code and bash scripting to create scripts that can be uploaded to “slurm” and ran on the supercomputer. For one genome to be annotated takes around 30 minutes, not including time to wait for the job to reach the front of the queue. This number is ideal for small datasets, but once thousands of online genomes need to be annotated this number is not sufficient. Therefore, it is a secondary item of this project to find a way to speed this process up, through refining existing scripts or possibly through finding alternatives, such as [KofamKOALA](https://www.genome.jp/tools/kofamkoala/). As mentioned, R will be used as the tool for data handling and visualisation. This mean that all of the packages on the CRAN repository can be used to assist this work. For example, the package EnrichKO can take the genes and produce the KO values and pathways so that analysis can be done and visualisations can be created. Other useful packages include: Tidyverse for data handling; flextable for tabulation and ggplot2 for visualisation.

Due to the size of the dataset, efficient storage of this data will be vital so that the computational systems do not get overwhelmed. To aid this, a relational database will be created in PostgreSQL. PostgreSQL is good at handling the .JSON data format which is used in many public APIs, including the NCBI. A relational database will allow for the data to be queried in many different ways to answer different questions. For the purposes of data transparency and experiment reproducibility, the scripts used in this project will be uploaded to a github repo for the project. So that any researchers who wish to build on this project can take inspiration from or adapt the scripts used. This repo will also contain a series of notebooks outlining where the data came from, the specifics of the data pipeline and the significance of the outputs.

Finally, the data will be supplemented by over 100 genomes generated in the labs at ECW by Bangor Masters and PhD researchers. These samples come from amphibian skin microbiomes, the known status of these samples mean that more exact analysis can be performed, as well as providing good comparisons, for example, there are not many samples from amphibian skin microbiomes in the database, so a comparison of how the same genera differ when on these hosts compared with the environment could be created.

(732)

[i really switched it up, not sure if i went too far with this one, so including the older methods section below the references for reference, and in case i need to roll back a bit, also, should i have references in this bit?, should i put in a flowchart or something visual like that?]

## References

Abdi, Ghloamareza, Maryam Abbasi Tarighat, Mukul Jain, Reshma Tendulkar, Mugdha Tendulkar, and Mukul Barwant. 2024. “Revolutionizing Genomics: Exploring the Potential of Next-Generation Sequencing.” In, edited by Vijai Singh and Ajay Kumar, 1–33. Singapore: Springer Nature. <https://doi.org/10.1007/978-981-99-8401-5_1>.

Asaf, Sajjad, Numan ,Muhammad, Khan ,Abdul Latif, and Ahmed and Al-Harrasi. 2020. “Sphingomonas: From Diversity and Genomics to Functional Role in Environmental Remediation and Plant Growth.” *Critical Reviews in Biotechnology* 40 (2): 138–52. <https://doi.org/10.1080/07388551.2019.1709793>.

Becker, Matthew H., Robert M. Brucker, Christian R. Schwantes, Reid N. Harris, and Kevin P. C. Minbiole. 2009. “The bacterially produced metabolite violacein is associated with survival of amphibians infected with a lethal fungus.” *Applied and Environmental Microbiology* 75 (21): 6635–38. <https://doi.org/10.1128/AEM.01294-09>.

Edwards, David J., and Kathryn E. Holt. 2013. “Beginner’s Guide to Comparative Bacterial Genome Analysis Using Next-Generation Sequence Data.” *Microbial Informatics and Experimentation* 3 (1): 2. <https://doi.org/10.1186/2042-5783-3-2>.

Fleischmann, Robert D., Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, et al. 1995. “Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd.” *Science* 269 (5223): 496–512. <https://doi.org/10.1126/science.7542800>.

Harris, Reid N, Robert M Brucker, Jenifer B Walke, Matthew H Becker, Christian R Schwantes, Devon C Flaherty, Brianna A Lam, et al. 2009. “Skin Microbes on Frogs Prevent Morbidity and Mortality Caused by a Lethal Skin Fungus.” *The ISME Journal* 3 (7): 818–24. <https://doi.org/10.1038/ismej.2009.27>.

Land, Miriam, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R. Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, et al. 2015. “Insights from 20 Years of Bacterial Genome Sequencing.” *Functional & Integrative Genomics* 15 (2): 141–61. <https://doi.org/10.1007/s10142-015-0433-4>.

Lunjani, Nonhlanhla, Sinead Ahearn-Ford, Felix S. Dube, Carol Hlela, and Liam O’Mahony. 2021. “Mechanisms of Microbe-Immune System Dialogue Within the Skin.” *Genes & Immunity* 22 (5-6): 276–88. <https://doi.org/10.1038/s41435-021-00133-9>.

McKenzie, Valerie J, Robert M Bowers, Noah Fierer, Rob Knight, and Christian L Lauber. 2012. “Co-Habiting Amphibian Species Harbor Unique Skin Bacterial Communities in Wild Populations.” *The ISME Journal* 6 (3): 588–96. <https://doi.org/10.1038/ismej.2011.129>.

McKnight, Donald T., Roger Huerlimann, Deborah S. Bower, Lin Schwarzkopf, Ross A. Alford, and Kyall R. Zenger. 2022. “The Interplay of Fungal and Bacterial Microbiomes on Rainforest Frogs Following a Disease Outbreak.” *Ecosphere* 13 (7): e4037. <https://doi.org/10.1002/ecs2.4037>.

Petrushin, Ivan S., Nadezhda V. Filinova, and Daria I. Gutnik. 2024. “Potato Microbiome: Relationship with Environmental Factors and Approaches for Microbiome Modulation.” *International Journal of Molecular Sciences* 25 (2): 750. <https://doi.org/10.3390/ijms25020750>.

Rappuoli, Rino, Nguyen K. Nguyen, David E. Bloom, Charles Brooke, Rachel M. Burckhardt, Alan D. Dangour, Dilfuza Egamberdieva, et al. 2025. “Microbes Can Capture Carbon and Degrade Plastic Why Aren’t We Using Them More?” *Nature* 639 (8056): 864–66. <https://doi.org/10.1038/d41586-025-00875-w>.

Sepulveda, Jorge L. 2020. “Using r and Bioconductor in Clinical Genomics and Transcriptomics.” *The Journal of Molecular Diagnostics* 22 (1): 3–20. <https://doi.org/10.1016/j.jmoldx.2019.08.006>.

Simões, Lúcia Chaves, Manuel Simões, and Maria João Vieira. 2007. “Biofilm Interactions Between Distinct Bacterial Genera Isolated from Drinking Water.” *Applied and Environmental Microbiology* 73 (19): 6192–6200. <https://doi.org/10.1128/AEM.00837-07>.

Vandeputte, Doris, Gunter Kathagen, Kevin D’hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, et al. 2017. “Quantitative Microbiome Profiling Links Gut Community Variation to Microbial Load.” *Nature* 551 (7681): 507–11. <https://doi.org/10.1038/nature24460>.

Walterson, Alyssa M., and John Stavrinides. 2015. “Pantoea: Insights into a Highly Versatile and Diverse Genus Within the Enterobacteriaceae.” *FEMS Microbiology Reviews* 39 (6): 968–84. <https://doi.org/10.1093/femsre/fuv027>.

### Plan (just for the review, NOT in final document)

*OUTLINE FOR DISSERTATION PLANS:*  
*Name:*  
*Title of the project:*  
*Name of the Supervisor: So we know who to send the plan to for grading!*  
*Hypothesis/ Aim: clearly state what you want to test.*  
*Abstract/ Summary: max 150 words (remember your abstract in the essay last semester). This is different*  
*from the lay summary. Include a problem statement, your intended approach, your outcomes (potential*  
*findings of your study or the outcomes of discussions with your supervisor over the approach you will take)*  
*and the implications of your study.*

*\*Introduction/ Background: What has been published so far? Why is it important to test this hypothesis/do*  
*this work?*

*Finish the introduction by stating your objective. For some more exploratory projects that use*  
*published literature as the source material & the data, specific objectives might be difficult to define at this stage. Speak to your supervisor if you are unsure. 750 words.*

*†Materials and Methods: What kind of equipment will you use? What are the reagents? What methods/*  
*experimental design will you exploit? How will you collect your data? What software? Which statistical*  
*analyses? 750 words.*

*Figures and flowcharts fit in nicely into Materials and Methods. Figure titles & legends do not count towards*  
*the word limit. Students undertaking literature-based meta-analyses might present some preliminary data.*  
*Methodologies vary between different subjects and different groups; get your supervisor’s help for this part.*  
*References (obviously).*

*†Commitment (optional; 100 words): Reflect on why you chose this project, what you expect from the project*  
*and how you would define the success or failure of the project.*  
*Risk Assessment form; Ethics Checklist: A key part of assessing the viability of a project is assessing the*  
*hazards & risks involved. You will be carrying out the project, you need to understand these risks. Similarly,*  
*as researchers, we have a responsibility to the animals and people we work with. Because of this, failure to*  
*include either form may result in being awarded 0%. These forms do not contribute to the word count but*  
*starting the research without approved RA or ethics clearance is breaking the law and is a disciplinary offence*  
*under the university regulations.*

#### old abstract

*Bacterial genomics is a modern field with a storied history. Much research has been done in the field in recent years, however very little taps into the wealth of free genomic data held on websites like the NCBI database. On sites like these there are millions of public bacterial genomes that can be used to better understand the genetic differences and genomic makeup. This information, combined with metadata about the environment or host, can generate novel comparisons about the functional relationships between bacteria and how their evolutionary histories have developed. Previous studies have found that species can have similar microbiomes despite geographic seperation. There are many computer-based tools that can assist with this, including environments like R, the hawk supercomputer and SQL servers. There are many R packages and other online tools that can assist with this. Using all of these, novel visualisations can be made to present these findings.*

*(150 words) [maybe talk more specifically about microbiomes, “genomics is the tool, microbiomes are the things being researched”]*

### *old methods section*

*All analysis will be undertaken computationally by me using data either collected from online repositories, such as the NCBI database, or given to me by the researchers at the ECW labs in Bangor as a by-product of their research. This will involve the use of many R packages and tools on the Hawk supercomputer, operated by SCW, for example the modules gtdb-tk/2.1.1 and eggnog-mapper/2.1.12. This will require me to develop my skills with command-line code as well as slurm and bash scripting. There are many tools in R packages that facilitate high-quality bioinformatics, making use of these is critical for data manipulation and clear data presentation (Sepulveda 2020).*

*As mentioned, this project relates to a larger work conducted by Bangor Masters and Phd researchers, as such, I will be receiving the genomes of up to 100 more bacteria sampled in Bangor labs, with the associated metadata. I will then add these to my existing data-pool and rerun past research in light of this data-influx. This includes; CheckM analysis, GTDBTK analysis and eggNOG-mapper analysis. This will update the phylogenetic trees and heatmaps that already exist, and allow for a database of these samples to be made, into which I can put the associated metadata.*

*Part of my work will be to discover and use more advanced computational applications. For example, the use of PostgreSQL for data storage, as opposed to the file system on my machine. This is to develop my skills as a bioinformatician, especially in reference to creating novel diagrams in R.*

*For example, a specific genera of interest, identified in 1 sample of the 10 local samples is the genus Pantoea. This is an interesting genus with a large variety of functions and relationships. Isolates from this genus have been found to have functions ranging from herbicide degridation to nitrogen fixation to antibiotic production. Samples have been found in hospitals, suggesting a possible link to disease in some capacity, this could make this genera interesting from a Chytridiomycosis point of view. An understanding of this genus could be beneficial for future research in the area (Walterson and Stavrinides 2015). Another genera of interest is Sphingomonas, samples have been found to be beneficial to host plant health, with little known about the genomics and internal mechanisms of how they produce these benefits (Asaf et al. 2020).*

*The discovery of novel avenues of discovery will be important to this project. Requiring me to study the literature surrounding bacterial genomics thoroughly to find more methods. For example, the chemical violacein is important in chytridiomycosis defence (Becker et al. 2009), thus, I can look to identify it in the online samples, in order to create a distribution or possible evolutionary history of the gene that codes for this chemical.*

*There is a wealth of data surrounding the public samples, for example, location data can feed the creation of maps that identify the source location of the samples, and figures can be created surrounding the host they were sampled from, these fields can then also feed secondary analysis, such as how samples from e.g. Europe and Africa differ.*

*Further public analysis will include intra-specific comparisons of genera inside families where more than 1 Bangor sample has been identified, i.e. Sphingomonas and Microbacterium, these will help illuminate how the genera inside these families differ from each other. This will be a simple, if time consuming, process, as the analysis pipeline has already been created, the new public samples just need to be processed.*

*The data is downloaded directly off of the NCBI and KEGG websites, this then feeds into a data-analysis pipeline that processes the raw genomes in .fasta format through eggNOG-mapper, these are then run through the enrichKO package in R to find the KEGG map-ids, which, after further processing in R, can then be turned into heatmaps to identify important pathways in these groups, grouping by the genera. This is taken to mean any pathway where there is more than 80% enrichment in one group and less than 50% in the others. The use of the R package GTDB-TK can also be used to create phylogenetic trees in R. Metadata is also downloaded off of both sites to help with clarity in the visuals, as well as providing further avenues of study. Once the important genes are identified, then the implications of those pathways in reference to the bacterial community and host can be illuminated.*

*(723)*

*Poison dart frogs were chosen for this because other research was already being conducted on them by Bangor researchers, so the sequenced bacterial MAGs were available for me to conduct secondary analysis on. They are also convenient due to the importance of microbiomes to life, the skin of amphibians is especially important, due to its use as a site for respiration, thus the microbiome therein should have important interactions with the host. This research could have implications for developing methods for combating chytridiomycosis infection in these animals. For example, the relationship between microbiome and fungal infection is not simple, and past research into probiotic solutions has had issues, so a greater understanding could help develop novel solutions or reveal more about existing processes (McKnight et al. 2022).*

*This study shares methods with a previous analysis involving E coli, which similarly used the wealth of publically available samples to do analysis with bioinformatics (Edwards and Holt 2013). An understanding of their relationships with their skin microbiomes, and the interspecific functions, variation and competition therein, through funcitonal genomics might give us a better idea of how these often cryptic animals fit into their niches in their environment. This could help us develop plans to protect them from dangers such as chytridiomycosis through a deeper understanding of the underlying mechanisms at play, as well as a potential host of novel information generated surrounding these mysterious microbiomes and possibly inspiring more scientists to follow on from this study. This will include analyses using public data, a lesser-used methodology in the area, this has the benefit that analysis can be done faster than with traditional methods, as the repositories contain years worth of data that is almost immediately accessible. This project could have real-world impacts on the future study of microbiome genomics.*

*(feel like this should be methods, but where?)*