**Microbes, Microbiomes and Bioinformatics DTP**

**SECTION 2 – Information about the project**

|  |  |
| --- | --- |
| **9** | **Project title** |
|  | **Aggressive prostate cancer in African men: are microbes involved? Investigating decontamination, microbiomes and human genomics** |
| **10** | **Strategic/Enabling Themes (select all that apply)** |
|  | * *Microbes (bacteria, viruses, fungi, parasites)* * *Microbiomes* * *Microbial bioinformatics* * *The one-health approach* |
| **11** | **Primary organisation where student will be based** |
|  | * *University of East Anglia* |

**SECTION 4 – Quality of the Science**

|  |  |
| --- | --- |
| **20** | **Aims and objectives of the proposed research (up to 100 words)** |
|  | This PhD project will investigate whether microbes are associated with adverse clinical outcomes in men of African ancestry with prostate cancer. The project will include wet-lab and *in silico* approaches to optimise the detection of microbes.  1. Benchmark *in silico* approaches for human depletion and decontamination of cancer whole genome sequence metagenomics.  2. Investigate experimental techniques to perform human depletion techniques prior to metagenomic sequencing.  3. Apply pipelines for short read metagenomic classification to African-centric datasets.  4. Investigate identified bacteria in objective 3 in urine and tissue samples.  5. Search for associations with clinical, genomic and anthropological features. |
| **21** | **Background to the proposed research (up to 500 words, including references)** |
|  | Prostate cancer is the most common non-skin cancer in men1. One in eight men will obtain a diagnosis at some point in their lives, which doubles to 1 in 4 men with African ancestry, which also tends to be more aggressive.  We recently identified that certain bacteria are associated with aggressive prostate cancer2. What is not yet clear is if microbial differences are contributing to different prostate cancer outcomes between ethnicities. Most large-scale sequencing initiatives have focused on populations of European ancestry, giving rise to inequalities between previous research efforts and the populations that need it most.  Our laboratory has a research focus on classifying microbes in human tissue whole genome sequencing data3. We recently uncovered contamination and human reads misclassified as microbial taxa, critically impacting cancer metagenomic studies4–6. Ensuring that human reads are sufficiently handled is therefore a paramount task in accurately identifying microbial sequences associated with human cancers. This can be done by either depleting human reads using laboratory-based approaches, or by computationally identifying the human reads and removing them from analysis.  Recently, software has emerged with the purpose of removing human sequencing reads and preserving microbial reads (such as sra-human-srubber and Hostile7). To date, no benchmark has been done to identify how software compare. One aim of this project is to benchmark human read removal, with a particular focus on how well these perform with regards to African sequences which may not be as well represented in human reference genomes8. Similarly, the issue of contamination is critical for high performance of taxonomic classification pipelines, particularly in low-biomass or identifying microbes in cancer whole genome sequencing studies9. This project will test synthetic and real, controlled communities to benchmark approaches to microbial decontamination including software such as SCRuB10 and Deepurify11.  This project will then combine the optimal human depletion and decontamination approaches into a pipeline for metagenomic classification. This project will benefit from unique African-centric sequencing datasets (HEROIC 1K project) to investigate whether or not microbes are associated with anthropological features and adverse clinical outcomes in African prostate cancer. Presence of key bacteria will also be investigated and validated on tissue and urine samples. Associations will also be investigated alongside anthropological findings. This project will provide evidence on whether microbes are associated with aggressive African prostate cancer and will suggest avenues for future translational research.  **References**  **1** Lozano, R. *et al.* *Lancet (London, England)* **380**, 2095–2128 (2012)  **2** Jaratlerdsiri, W. *et al.* *Nature* 1–8 (2022)  **3** Hurst, R. *et al.* *European Urology Oncology* **5**, 412–419 (2022)  **4** Gihawi, A. *et al.* *Genome Biology* **20**, 208 (2019)  **5** Gihawi, A. *et al.* *Microbial Genomics* **9**, 001088 (2023)  **6** Gihawi, A. *et al.* *mBio* **0**, e01607-23 (2023)  **7** Constantinides, B. *et al.* *Bioinformatics* **39**, btad728 (2023)  **8** Sherman, R. M. *et al.* *Nat Genet* **51**, 30–35 (2019)  **9** Salter, S. J. *et al.* 1–12 (2014)  **10** Austin, G. I. *et al.* *Nat Biotechnol* **41**, 1820–1828 (2023)  **11** Zou, B. *et al.* 2023.09.27.559668 (2023) |
| **22** | **Methodology and work packages within the programme of proposed research (up to 400 words)** |
|  | **1. Benchmark and develop pipelines for the metagenomic classification of tumour whole genome sequence data.**  1.1 Investigate *in silico* human read depletion.  a. Generate simulated sequencing datasets combining bacteria and human.  b. Investigate differences between human references (i.e. GRCh37/38, CHM13, Pangenome, African specific sequences). Unique kmers in each genome will be identified and investigated for potential misclassification.  c. Benchmarking existing methods to remove human reads e.g. BBDuk and SRA-human-scrubber and the effect on microbial misclassification.  1.2 Investigate resolving contamination.  a. Identify and test algorithmic approaches to remove contamination i.e. SCRuB.  b. Identify the best approach for resolving contamination using lists of contaminants e.g. kitome.  1.3 Combine the best approaches into a pipeline for metagenomic classification.  **2. Investigate experimental techniques to perform human depletion techniques prior to metagenomic sequencing.**  Overabundance of human DNA compared to bacterial DNA limit the efficacy of metagenomics investigations. Methods in the literature have been applied to several sample  types (e.g., stool, saliva, tissue) but not with prostate secretions. The student will continue the investigation of these approaches using mock communities plus cancer cell lines and clinical samples.  **3. Apply pipeline for short read metagenomic classification to African-centric datasets.**  The pipeline developed in O1 will be applied to the following prostate cancer WGS datasets:   * Ethnically diverse population in the HEROIC Prostate Cancer Precision Health (PCaPH) project in collaboration with the University of Sydney (n=400. This includes ethnically European (n=102), Asian (n=23) and African samples (n=120). The African cohort is currently being expanded to n=1000. * Pan-Prostate Cancer Group (n=2,176). * Genomics England (n=580).   **4. Investigate identified bacteria in objective 3 in urine and tissue samples.**  4.1 Quantitative PCR assays will be designed and tested for key bacteria targets and analysed in urine samples from the HEROIC study (n=250) and other studies (n=100).  4.2 Develop and update a custom NanoString assay to detect up to 100 key bacteria taxa.  4.3 Apply custom Nanostring assay to the same samples.  **5. Search for associations with clinical, genomic and anthropological features.**  The student will use the results from O3 and O4 to investigate associations between the microbes detected and various features. Unsupervised machine learning will illuminate differences between ethnicities and markers for aggressive disease such as Gleason Score. Survival analyses will be performed. Significant associations will be determined between microbe presence and tumour molecular characterisations. This will give insight into the effect the bacteria are having on the tumour. |
| **23** | **What is the potential for publication outputs? (up to 100 words)** |
|  | This PhD research project includes a combination of bioinformatics approaches and lab-based experiments to determine bacteria effects on prostate cancer and other mechanisms, which may translate to diagnostic biomarker development and/or potential treatment option.  There is scope for at least two solid publications from the proposed work. Firstly, the benchmark of human depletion approaches and decontamination would provide significant benefit to the microbial bioinformatics community working on human and low-biomass samples. Secondly, there is the potential to publish the investigation of bacteria associated with aggressive prostate cancer and adverse clinical outcomes in men of African ancestry, which would provide an opportunity for patient benefit. |

**SECTION 5 – Suitability and Feasibility**

|  |  |
| --- | --- |
| **24** | **Explain how the proposed project fits the remit and priorities of the Medical Research Council (up to 100 words).** |
|  | This research fits well into the remit and priorities of the MRC. Firstly, this constitutes global health research and will benefit from existing international collaborations, particularly those in Africa (HEROIC and PCaPH projects) and Australia (Professor Vanessa Hayes, Sydney). Secondly, this fits in with the infection and immunity theme. Thirdly, a significant proportion of this project will fit into the MRC theme of better methods and better research. Fourthly, there is considerable scope for translational benefit for patients across the world, which is also a focus of the MRC. |
| **25** | **Track record of the supervisory team with this kind of research (up to 100 words)** |
|  | We have been working in microbiomes and cancer (with a focus on prostate cancer) for over 10 years. We have received multiple grants for this work from Prostate Cancer UK, The Bob Champion Cancer Trust, and Big C. We were founding members of the Prostate Cancer Foundation bacteria working group. We have had papers on this topic in European Urology Oncology, Genome Biology and Nature Genetics. More broadly, we have decades of experience in cancer molecular biology and bioinformatics, that has led to high impact results. |
| **26** | **Up to two recent relevant publications from the supervisory team** |
|  | 1. Hurst R, Meader E, Gihawi A, Rallapalli G, Clark J, Kay GL, et al. Microbiomes of Urine and the Prostate Are Linked to Human Prostate Cancer Risk Groups. European Urology Oncology [Internet]. 2022 [cited 2022 Sep 20];5:412–9. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2588931122000566>  2. Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C. S., Brewer, D. S., Pertea, M., & Salzberg, S. L. (2023). Major data analysis errors invalidate cancer microbiome findings. In I. B. Zhulin (Ed.), mBio (Vol. 14, Issue 5). American Society for Microbiology. <https://doi.org/10.1128/mbio.01607-23> |
| **27** | **Provide evidence that the project is achievable within the resource limitations and the given timeframe (up to 100 words)** |
|  | Most of the infrastructure and resources required for this project are in place: the NanoString equipment required for the bacteria detection custom assay is in the UEA Norwich Medical School Diagnostic laboratory. We have access to other NanoString bacteria assay datasets to demonstrate that the technology is applicable and feasible to use for this study. Additional costs include a laptop (£1,500), 5TB of research storage (£50 per TB per year). Funding will be required for laboratory reagents, qPCR assays, sequencing runs and NanoString custom assay. |
| **28** | **How is the project best suited to a PhD student rather than a post-doc or technician? What is the intellectual challenge? (up to 100 words)** |
|  | This project is not suitable for a lab technician as this is not a data generating project (a good proportion of data is already generated) and requires considerable scientific insight to interpret the results. The intellectual challenge is in this and conceiving of future experiments. The student will conduct original investigations, shown an understanding of how results relate to a wider field of knowledge, and make a significant contribution to the field. This project is not suitable for a postdoc as topics will be explored in greater depth and more hypotheses will be generated than is normal in a project grant. |
| **29** | **How will the student be able to show originality and independence? (up to 100 words)** |
|  | The project is novel and so the student will show originality throughout. The student will be expected to perform the experiments and data analyses independently, especially after the first year. Although numerous decisions will be taken by the student the main areas when originality and independence will be displayed are in:  - The interpretation of results and the proposal of optimal approaches to the removal of host DNA sequences and decontamination  - Further development of lab based qPCR assays, NanoString assays to detect specific bacteria  - The interpretation of associations between bacteria and clinical features |
| **30** | **What aspects of the project are risky and how will risks be mitigated? (up to 100 words)** |
|  | Funding and infrastructure is already in place at the University of Sydney to sequence these samples. Sequencing has already taken place and data is already available for N=190 samples.  Sample Sizes – We have N=2,176 samples from the PPCG (ethnically European) stored locally which we will apply the updated pipeline to.  Supervision – Dr Gihawi is relocating to Sydney for a year during the time frame of this project (April 2026). We will apply to the Turing award to allow for the student to visit. The student will be well supported in-person by the rest of the supervisory team and will have remote meetings with Dr Gihawi. The student will be firmly into their PhD project at this time. |