**UNIVERSITY OF CAPE TOWN COMPUTATIONAL BIOLOGY DIVISION**

**Bioinformatics Support Request**

Please provide us with more information on your request for support. Complete the form as comprehensively as possible, and please indicate where there is still uncertainty.

**Please note, the earlier we are involved the better – for example, it would be better for us to be involved during the study design and even grant application stage**.

|  |
| --- |
| **CONTACT DETAILS** |
| Date of request | **22-10-2021** |
| Name | **Lauren Martin** |
| Email address | **lcmartin@sun.ac.za** |
| Research Group/Department | **Neuropsychiatric Genetics** |
| Faculty | **Faculty of Medicine and Health Sciences, Stellenbosch University** |
| IF student, name & email of supervisor | **Prof. Sian Hemmings (smjh@sun.ac.za)** |

|  |
| --- |
| **PROJECT DETAILS** |
| 1. What is the scientific question? |
| Does our optimised Illumina iSeq100 full-length 16S rRNA gene sequencing technique and gene reassembly, when evaluated alongside PacBio 16S sequencing and V1-V2 hypervariable sequencing data, improve the species-level taxonomic resolution of the bacterial species present in the maternal vaginal swab samples obtained from women who have given birth to infants with and without Foetal Alcohol Spectrum Disorders (FASD)? |
| 2. Who are the partners on the project? |
| Neuropsychiatric Genetics Research Group, Separations and Illumina. Project has had reagents sponsored by Illumina and Separations conceptualized and proposed the project. |
| 3. What type of collaboration with CBIO is expected? For a project that is done as collaboration or for a fee, we will put the agreement in writing. |
| A project that will be conducted as a collaboration, with CBIO partners added as co-authors on resultant papers. |
| 4. Are there any ethical issues we should be aware of? |
| No |
| 5. How much work is expected from CBIO and when? |
| We would like assistance with the development of a pipeline that makes use of EMIRGE and dada2 to assemble the tagmented amplicon and subsequently assess the bacterial composition and perform statistical analysis of the samples. It would also be great if CBIO could advise on appropriate statistics and scripts to use so that data can be accurately represented and the research question (whether Illumina iSeq100 sequencing technique, when evaluated alongside PacBio 16S sequencing and hypervariable sequencing data, improves bacterial taxonomic resolution). Further advice and advice with the bioinformatics required for potential publications may also be required.We would ideally like to have processed and analysed the data by April/May 2022. This should allow sufficient time for the MSc student to write up her thesis. |
| 6. What type of data will be generated (e.g. sequencing, genotyping, expression, etc.) and what technology platform will be used? |
| Microbial sequencing data of the 16S rRNA gene in the form of Fastq sequencing files will be produced. Sequencing data will be obtained from the Illumina MiSeq, Illumina iSeq100 and PacBio Sequel II instruments. |
| 7. When do you expect the data? Does it need to be transferred from somewhere else? |
| We have V1-V2 hypervariable region data in our possession. We will have sequencing data from the Illumina iSeq100 instrument and the PacBio sequel II instrument by January 2022 the latest. Data transfer to CBIO will be contingent upon a signed Data Transfer Agreement. |
| 8. How large will the data be? How long does it need to stored for, and have you made arrangements for storage?  |
| The data will be roughly 10 -20 GB. Data is and will stored on external hard drives, backed up in the cloud and on the NAS in the possession of the Neuropsychiatric Genetics. |
| 9. What bioinformatics analysis needs to be done? Which tools are required? |
| A pipeline needs to be developed to assemble tagmented 16S rRNA gene reads into full-length 16S rRNA genes. We intend doing this by using an EMIRGE container on HPC singularity platform on the SLURM cluster. Thereafter, microbiome-related bioinformatic and statistical analysis will be performed by making use of dada2, PhyloSeq and Vegan packages on R. Differential abundance analyses could be performed using DeSeq2. Community state types can be inferred from hypervariable sequencing data by using VALENCIA. A method for comparing the different types of sequencing data will also have to be performed. |
| 10. If a collaborative model is being used, what papers are envisaged and who will the authors be? |
| We anticipate that a white paper detailing the sequencing method employed in the study will be published. A possible research article outlining the technique enabling the sequencing of short-read libraries of the full-length 16S rRNA gene amplified from microbial DNA present in vaginal swabs and comparing the differences in species-level resolution achieved between the previously sequenced 16S rRNA V1-V2 region of the microbial DNA extracted from the vaginal swabs and the entire 16S rRNA gene sequencing data (PacBio and Illumina) generated from this project may also be published. The authors would include the MSc student (Lauren Martin), lab advisor, PhD student heading up FASD-microbiome project, and co-supervisor (Natasha Kitchin), CBIO member assisting with the bioinformatics and data analysis (Dr. Katie Lennard), members of the research team that collected vaginal swab samples ( Anna-Susan Marais and Marlene de Vries), principal investigators responsible for initiating the parent FASD study (Prof Philip May, Prof Soraya Seedat (MSc co-supervisor)) and principal investigator of Neuropsychiatric Genetics research group and main supervisor (Prof Sian Hemmings). |
| 11. Can we add a short description and objective of the project to the CBIO website? |
| Yes |

**PLEASE FORWARD THE COMPLETED FORM TO:**

Nicola.mulder@uct.ac.za