**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 | 18 April 2019 |
| Name | Relebohile Matobole |
| Email address | [Mtbrel001@myuct.ac.za](mailto:Mtbrel001@myuct.ac.za) |
| Research Group/Department | Hair and Skin Research Lab |
| Faculty | Health Sciences |
| IF student, name & email of supervisor | Nonhlanhla Khumalo: [n.khumalo@uct.ac.za](mailto:n.khumalo@uct.ac.za)  Ardeshir Bayat: [ardeshir.bayat@uct.ac.za](mailto:ardeshir.bayat@uct.ac.za) |

|  |
| --- |
| **PROJECT DETAILS** |
| 1. What is the scientific question? |
| The aim is to identify biomarkers associated with keloids and folliculitis keloidalis nuchae and charactering these biomarkers during the wound healing process. The characterization of the biomarkers could give insights into the etiology and treatment target for the disorders. |
| 2. Who are the partners on the project? |
| Pierre Fabre |
| 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. |
| Collaboration – (free-free as agreed between Prof Mulder and Prof Khumalo) |
| 4. Are there any ethical issues we should be aware of? |
| Ethics approval has been granted for the study: HREC REF: 375/2017 |
| 5. How much work is expected from CBIO and when? |
| We expect all the bioinformatics analysis support from CBIO. We would like to have the RNA sequencing analysis and proteomics data by mid-May depending on the feasibility. |
| 6. What type of data will be generated (e.g. sequencing, genotyping, expression, etc.) and what technology platform will be used? |
| 1. **RNA sequencing**: Total RNA transcriptome sequencing by ribo-depletion method and sequenced on a BGISEQ sequencing platform. The data output is 10Gb/sample, 50M reads/sample and it is in FASTQ format. 2. **Proteomics**: a label free quantification mass spectrometry based approach. Samples were prepared on HILIC magnetic bead workflow followed by analysis on the Q-Exactive mass spectrometer. The raw files were processed through the Progenesis QI for proteomics software resulting in a spreadsheet containing expression profiles which will be regulated and validated accordingly. |
| 7. When do you expect the data? Does it need to be transferred from somewhere else? |
| 1. **RNA sequencing**: We have the RNA seq data for the first 80 samples and we are expecting data for 15 more samples in the coming few weeks which are undergoing sequencing pending successful library preparation. The data is in an external hard drive and has been transferred onto the scratch folder on the UCT HPC server. 2. **Proteomics**: We have received the first set of spreadsheet (csv files) and these are in zipped folder on email. The rest of the data will be available in the coming weeks, approximately early May 2019. |
| 8. How large will the data be? How long does it need to stored for, and have you made arrangements for storage? |
| 1. **RNA seq data** will be 2Tb but we currently have 1.6Tb on an external hard drive. The data needs to be stored until the analysis is done. We are hoping to purchase permanent data storage on the Servers for long term data storage 2. **Proteomics data:** The CSV files are 2Mb and we still expect more and in total should not exceed 1G. |
| 9. What bioinformatics analysis needs to be done? Which tools are required? |
| 1. **RNA sequencing**:    1. alignment to the human genome    2. post-alignment QC metrics    3. gene expression values    4. adjusted P-values for differentially expressed genes/proteins for the different time points (day 0, 14 and day 28)    5. expressions visualized using PCA plots, t-SNE, venn diagrams, heatmaps and hierarchical clustering for tissue and cell samples and between and across participants and conditions    6. profiling of gene expression over the 3 time points across the conditions    7. pathway analyses; gene ontology, KEGG, upstream regulators, analysis of downstream effects, diseases and functions    8. identification of possible biomarkers and 2. **Proteomics**:    1. PCA plots and Pearson’s correlation pairwise analysis for quality control    2. Statistical analysis of differential protein expressions    3. Heatmaps, t-SNE and other visualization methods    4. Venn diagrams between tissue samples and cells to find proteins specific to fibroblasts (cells)    5. Pathway analysis and disease functions as well as gene ontology    6. profiling of gene expression over the 3 time points across the conditions    7. Identification of biomarkers |
| 10. If a collaborative model is being used, what papers are envisaged and who will the authors be? |
| As discussed between Prof Mulder and Prof Khumalo. |

**PLEASE FORWARD THE COMPLETED FORM TO:**

[Nicola.mulder@uct.ac.za](mailto:Nicola.mulder@uct.ac.za)