**Bioinformatics and Biostatistics Core**

**Consultation Checklist**

**JIRA ServiceDesk Request Ticket Code:** S

**Date**: 27 APR 2020

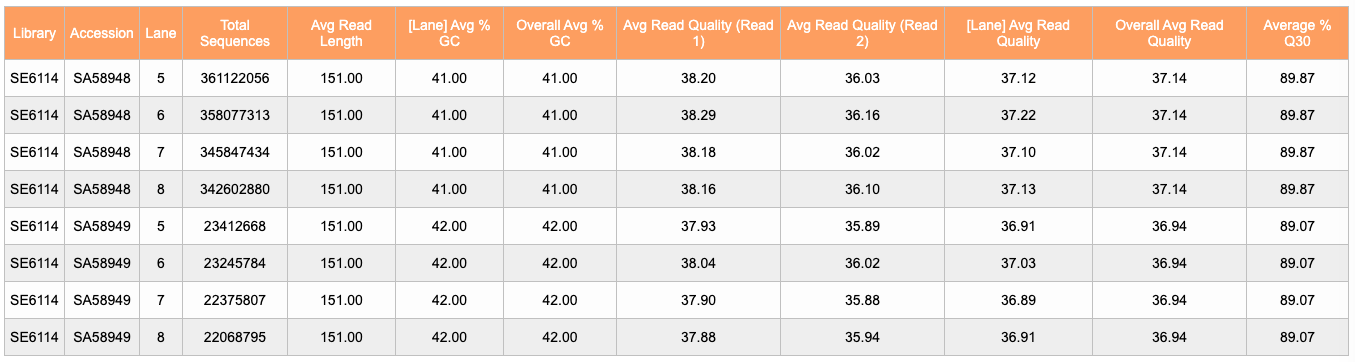
**Client**: Gerd Pfeifer

**Is this related to a previous BBC project:** VBCS-214 (PFEG\_20190808\_CDseq)

**The data** (what are the data? Covariates? number of samples, design, species & preferred reference genome if applicable):

* gDNA libraries
* Species: Mouse
* Reference: mm10
* Samples:
  + UVB irradiated mouse MEF cells (MEF-UVB)
  + Non-treated MEF (MEF-NT).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample ID | External ID | Index 1 | Index 2 | Total reads | File size |
| FT-SA58948 | MEF-UVB | CGCTCATT | GGCTCTGA | 1.4B | 200G |
| FT-SA58949 | MEF-NT | ATTACTCG | GGCTCTGA | 0.1B | 10G |



**Analysis plan** (gross objective, methods, and tools):

**Objective:**

**Methods:**

1. Trim 2x150bp reads to 2x50bp
   1. QC trimming step
2. Align reads with BWA-MEM
   1. All libraries for **MEF-UVB** will be aligned together
   2. All libraries for **MEF-NT** will be aligned together
   3. Allow at least 1 mismatch in alignment (inherent heterozygosity in samples)
3. Remove all duplicate divergent, paired end reads that have a single nucleotide gap (they are probably PCR duplicates), so retain only one read if there are multiple.
4. Analyze base within single nucleotide gap and identify 3’ neighbor.
   1. T/C 3’🗹---->
   2. <----🗹5’A/G
   3. This will create a dinucleotide frequency (we expect 5’TT and 5’TC to be the most frequent, as before).
5. Create a trinucleotide frequency map from the dinucleotides (dipyrimidines): 5’TTN, 5’TCN, 5’CTN, 5’CCN, and others if they are frequent enough.
   1. Frequency histogram of treatment/control
   2. BED file for position information
6. What is the 5’ neighbor of the dinucleotides? This will also create a trinucleotide frequency distribution of the following type: 5’NTT, 5’NTC, 5’NCT, 5’NCC, and others if they are frequent enough.

(Further Steps Delivered later/ in an addendum)

1. Tetranucleotides in the 3’ or 5’ direction, e.g. make 5’NNYY or 5’YYNN.
2. Or we could put the dipyrimidines in the middle and ask what is the frequency of the bases 5’ and 3’: such as 5’NYYN.

**Expected deliverables**:

1. BED file containing genomic loci of dipyrimidines.

**Expected start date**: 28 APR 2020

**Tentative deadline**: 06 MAY 2020 (data needed for May 12 presentation)

**Estimated number of hours to completion (expected billable time)**: 16 hours

BBC Policy Summary

The VARI Bioinformatics and Biostatistics Core has established policies, to facilitate computational and statistical support for the design, planning, conduct, analysis, and reporting of research.

1. Any analyses outside those specifically listed here will require an addendum to this document with a revised timeline and updated number of billable hours. Additionally, any major modifications to the deliverables, defined here as taking >1 hour to generate, will also require an addendum with updated billable hours
2. This timeline is subject to change if projects with higher order in our queue take longer than anticipated.
3. The maximum number of billable hours refers to a conservative estimate of the total hours needed to complete this analysis. If the actual time is less than this estimate, your lab will be billed only the actual hours worked. If the work specifically outlined here takes longer than this estimate, your lab will only be billed the hours quoted here.
4. We will not begin work on your project until you send an affirmative response to this consultation sheet, even if the proposed start date has passed. Please be prompt in your project confirmation.