Pupil Bio Bioinformatics Scientist Test

Overview

Pupil Bio is seeking a bioinformatician with expertise in biostatistics, data analysis, and next-generation sequencing (NGS) pipelines to join our team. The ideal candidate will have a strong background in analyzing complex datasets, deriving biological insights, and developing robust pipelines for NGS data.

The present test comprises two challenges designed to evaluate candidates' technical and analytical skills:

1. **Data Analysis and Statistics**: Assessing the ability to analyze and interpret complex data, particularly phased methylation patterns.
2. **NGS Alignment and Mutation Calling**: Evaluating proficiency in processing raw sequencing data, aligning reads to a reference genome, and identifying somatic mutations.

Task 1: Data Handling and Statistical Analysis

Objective: Assess candidates' ability to handle complex data and apply statistical methods effectively.

Background: CpG methylation is an epigenetic marker that varies across tissue types. However, the methylation status of a single CpG site is unreliable as a biomarker due to errors introduced by bisulfite sequencing, sampling techniques, and biological variability.

Definition: Phased Methylation Pattern (PMP) is a unique set of coordinates that includes the DNA strand ('f' for forward (+) or 'r' for reverse (-)), the relative positions of three CpG sites on the same strand (e.g., x:y:z), and their methylation status (e.g., '000' for all unmethylated or '111' for all methylated). It represents a combined epigenetic signature across these CpG sites.

Hypothesis: Phased methylation patterns (PMPs) can act as reliable biomarkers to differentiate tissue types, providing higher specificity compared to individual CpG sites.

Dataset: The dataset (Link to Data) summarizes phased methylation patterns from NGS results across two tissues. Key columns include:

- **Strand**: Indicates the DNA strand ('f' or 'r').

- **CpG Coordinates**: Relative positions of three CpG sites (x:y:z).
- **Methylation Status**: Eight possible patterns ('000' to '111').
- **Sample ID**: Unique identifier for each sample.
- **Replicate**: Indicates technical replicates.
- **Tissue**: Tissue type (Tissue #1 or Tissue #2).

Sub-tasks:

1. **Coverage Analysis** (10 points):
    a. Calculate the median and coefficient of variation (CV) for single CpG coverage in each tissue (5 points).
    b. Generate plots summarizing the coverage statistics (5 points).
2. **Biomarker Identification** (20 points):
    a. Identify PMPs with high specificity for tissue differentiation, minimizing false positives for Tissue #1 while allowing some false negatives. Use statistical or machine learning approaches to assign confidence (e.g., p-values) to each PMP (15 points).
    b. Calculate the mean variant read fraction (VRF) for each PMP in both tissues (5 points).
3. **Address the following questions** (20 points):
    a. How does sequencing depth affect specificity confidence? (5 points).
    b. For the top 10 PMPs, estimate the threshold of reads required to confidently call Tissue #2 at a sequencing depth of 1 million reads. (5 points)
    c. Validate the hypothesis by comparing the specificity of the top 10 PMPs against individual CpG sites.(10 points).

Task 2: NGS Data Analysis

Objective: Evaluate candidates' ability to process and analyze raw sequencing data.

Dataset: The dataset (Link to Data) consists of NGS samples in FASTQ format, including one sample from normal tissue and one from cancer tissue.

Sub-tasks:

1. **Quality Control (10 points):**
    a. Perform quality checks using tools like FastQC and summarize quality metrics (e.g., sequence counts, per-base quality, read duplication levels). (10 points)
2. **Alignment and Mutation Calling** (40 points):

a. Align the samples to the human genome using tools like Bowtie2 or BWA. (10 points).

b. Identify somatic mutations present in the cancer sample but absent in the normal tissue.

   i. **Benchmark Software**: Use established tools such as Mutect2, Strelka2, or VarScan2 for somatic mutation identification and background mutation estimation. (10 points)

   ii. **Custom Code Development**: Write your own scripts, leveraging tools like Samtools, bcftools, or Python/R libraries, to perform mutation detection and calculate the required metrics. (15 points)

c. Use the normal tissue to calculate the median background mutation level. The background mutation level accounts for sequencing errors or biases that can mimic true mutations. Determine how many reads per million are required to confidently call a given mutation. (5 points)

Submission Instructions and deliverables

- **Code Repository:** Upload scripts and code to a GitHub repository. Ensure clear annotations and reproducibility.
- **Report**: Include a PDF report summarizing your findings, supported by visualizations (e.g., plots and tables). Large outputs do not need to be included in the initial submission, but you may be asked to provide such datasets via a custom link if selected for a potential interview.
- **Submission**: Send your GitHub link, report, and any supplemental files to ale.pinto@pupil.bio
- **Deadline**: Submit the completed challenge within two weeks.

Evaluation Criteria

- **Completeness and Accuracy:** Does your analysis fulfill the requirements and provide accurate results?
- **Code Quality and Documentation**: Is your code well-structured, annotated, and reproducible?
- **Visualizations**: Are your reports clear, informative, and visually engaging?

We look forward to seeing your submission. Impress us with your technical expertise and creativity! If you have any questions or need clarification, please don't hesitate to reach out to us at ale.pinto@pupil.bio

Good Luck!