Broken String Biosciences Bioinformatician pre-interview coding test

The purpose of this exercise is for you to process some of the outputs of an INDUCE-seq sequencing run in order to describe and interpret the number of breaks that occur at specific sites in the genome

Introduction

Our patented INDUCE-seq technology reveals the precise position and frequency of DNA double strand breaks (DSBs) throughout the genome. This approach is unique as it accurately represents DNA break events in true proportions and without experimentally introduced bias. Each read represents a single DSB that has occurred in situ in the cells that have been processed. We can locate where the break has occurred by mapping the read to the human genome. The sequence of the read is not important apart for a mechanism to locate the break. Instead the 5' end of the read is the exact position where the DSB was labelled.

As an internal control we often use the well-characterised DIvA (DSB Inducible via AsiSI) cell line, in which treatment with 4-hydroxytamoxifen (4OHT) triggers nuclear localisation of the AsiSI enzyme. This results in DSBs occurring at the AsiSI restriction enzyme sites within the genome. However only a proportion of the approximately 1400 possible AsiSI cut sites predicted by presence of the GCGAT/CGC sequence are actually cut, primarily due to the chromatin state at each possible site.

We provide you with single-end fastqs files from an INDUCE-seq experiment with 16 samples where a proportion of the samples are controls derived from DIvA cells without treatment, and the remainder are DIvA cells that have been treated with 4OHT. Given that INDUCE-seq determines breaks in an unbiased way, the reads representing breaks will be derived from naturally occurring endogenous breaks, in addition to those that are induced by the AsiSI enzyme active only in treated samples. The hypothesis is that treated samples can be distinguished from controls based on the number of breaks occurring at predicted AsiSI sites.

Starting Data

The data is provided at this URL: https://gitlab.com/brokenstringbio/analysis/coding-test. In order to keep the size of the fastqs small and the computation tractable, the reads have been both down-sampled and filtered to only include those that belong to chr21 and another small chromosome.

Instructions

1. Write a series of steps or a pipeline using **bwa mem** and **bedtools** to
   a. **Map the reads to chromosome 21**
      The sequence of chromosome 21 is provided within the same data folder as the fastqs for indexing with bwa prior to mapping. After mapping you will have a bam file per sample.
   b. **Convert the position of the reads contained in the bam files to genomic intervals in a bed file**
      Convert the bam file to the commonly-utilised and standardised bed file format, for downstream processing.
   c. **Process the bed file so that the coordinates are adjusted to just include the break site**

The bed files from step 1b contains the positions of the mapped reads. However as stated in the introduction, we are only interested in the 5' end where the break occurred. For each interval in the bed file, adjust the positions as follows to produce a new bed file.

   i. If the read is on the + strand adjust the **end** position to be **start** + 1
   ii. If the read is on the – strand adjust the **start** position to be **end** -1

d. **Intersect the breaks encoded in the bed file with predicted AsiSI sites**
   We provide a bed file of AsiSI cut sites in the human genome (T2T v2.0 release) named chr21_AsiSI_sites.t2t.bed. Intersect this with the break sites in the sample bed from step 1c

At this stage you will have a final bed file for each sample containing the number of breaks that occur at each AsiSI site within chromosome 21.

In case you have been unable to complete step 1 in its entirety, please upload what you have been able to do and proceed with pre-made outputs found in the data/pre-made_results directory

2. Using python and/or python libraries. analyse the data by
   a. **Reading in the data**
      From your step 1 there will be a tab-separated bed file for each sample containing AsiSI breaks
   b. **Sum the number of AsiSI breaks**
      Each sample will contain zero or more breaks at each of the sites on chr21. Find the sum of the AsiSI breaks per sample.
   c. **Normalize the number of AsiSI breaks**
      The original bed file for each sample (step 1c) will contain the total number of breaks per sample (also provided as data/pre-made_results/sample_total_breaks.tsv). In order to account for different amounts of starting material, divide the sum of AsiSI breaks (step 2b) by **total** breaks/1000, so that the data consists of the normalised sum AsiSI breaks for each sample.
   d. **Plot the data**
      Plot the data so that it is possible to determine if there are cluster of samples representing control and treated subsets.

Questions

1. Which of the samples are likely to be controls or treated?
2. Are there any you are uncertain of?
3. Can you explain the samples in the uncertain group?
4. What is the maximum percentage of possible AsiSI cut sites on chromosome 21 (as described in the chr21_AsiSI_sites.t2t.bed file) that is observed in a single sample?

Result submission

Please submit your answer and code to a publicly available git repository