STAT 530 Bioinformatics: Homework 5

Due Apr 17, 2017

For problems using R, turn in your answers in the form of a compiled R notebook PDF. **Note: this homework requires downloading a 3GB file.**

Problem 1 (5 points)

Let Y be a random variable that takes values between 0 and 1; for example, suppose Y has a Beta distribution. Suppose you want to study how Y depends on covariates X. Explain why the regression model

$$E(Y \mid X) = X^\top \beta$$

is not appropriate. Then write down a better model for the conditional mean of Y.

Problem 2 (5 points)

When preprocessing RNA-Seq data, it is clear why adapters need to be trimmed. However, why should low-quality bases be trimmed? What can happen if you leave low-quality bases in your reads? (Side note: there are other strategies you can use to deal with low-quality bases, other than trimming them.)

Problem 3

Part (a) (5 points)

Use the file `barcodes.txt` to demultiplex the reads in `hw_rna-seq.txt` using the default number of mismatches in `fastx_barcode_splitter.pl`. What is the smallest and largest number of reads that you see belonging to any subject after demultiplexing? How many reads are unmatched to a subject?

Part (b) (5 points)

Since there are 1000 total reads, each of the 42 subjects should have around $1000/42 \approx 20$ to 30 reads. However, this will not be what you see after you finish the previous problem. Instead you will find that there are very few reads matched to any subject, and that there are a very large number of unmatched reads.

1. You must diagnose the problem. What is causing the large number of unmatched reads?

2. You must also fix the problem. Provide all code you used to fix the problem. What is the smallest and largest number of reads that you see belonging to any subject after demultiplexing? How many reads are unmatched to a subject?

Problem 4 (5 points)

Download RNA-seq data from `ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX201/SRX2011545/SRR4017758/`. These reads are from a single honeybee. This is about a 3 GB file.

1. Install the SRA toolkit (package `sra-toolkit` in Ubuntu).

2. Convert the .sra file you downloaded into .fastq.gz (gzipped) files. This step could take 30 minutes or more.

Note: these are paired end data. You will get two files for this sample.

Report the first 10 bases of the first reads of each output file. Hint: use the `zless` or `zcat` commands to view the file without having to unzip it first. Do not delete these files, you will need them for the next homework.