Thanks for applying to Asimov! As part of our interview process, we would like to evaluate your technical skills in computational biology. We kindly ask you to complete the following task within a few weeks. We're aiming for this task to approximately a half-day of work. Please reach out to joe@asimov.io if you have any questions!

Task

1. Download *E. coli* RNA-seq data from this link. This data contains multiple fastq files of *E. coli* cDNA, generated during RNA sequencing. The files are described in more detail below.
2. You may find it useful to download the *E. coli* DH10B reference genome from NCBI as well as an annotated reference for *E. coli* DH10B and a plasmid which was transformed into DH10B.
3. Write software to answer the following question: which genes are differentially expressed between state 1 and state 2?

And please:

- Feel free to write your solution in any programming language and use any publically available third party libraries.
- Have an easy to run script to reproduce any reported statistics, graphics, and conclusions from (3).
- Write a README describing how your software works and how to run your scripts.

RNA-seq File Details

There are 8 files in the downloaded RNA-seq tarball: 2 states; 2 replicates per state; 2 parts per replicate per state.

- The two states represent different experimental conditions. Task (3c) is to identify genes that are differentially expressed between these states.
- The two replicates (per state) are for measuring the experimental noise for each state.
- There are two parts (per replicate per state) to ensure file sizes stay below 1Gb. Files such as *ecoli_state1_rep1.fastq* and *ecoli_state1_rep1_2.fastq* are from the same state and replicate.

In addition there are 2 files in the downloaded references tarball:

- An annotated reference for DH10-beta
- An annotated plasmid that was transformed into DH10-beta

Task Details

We would like to see your creativity and technical skills when solving this challenge. We kindly ask that you push yourself in (just) one of the following dimensions:

1. Software. Have strong unit tests, a scalable implementation and clean integrations with third party libraries
2. Statistics. Take some time to talk through more rigorous statistical analysis of transcriptional activity and experiment repeatability.
3. Genomics. What is unique about these two *E. coli* populations compared with wild type? Does the RNA profile contain sufficient information to infer details about the experimental conditions?
4. Visualization. Show us some beautiful, interactive visualizations of the alignments, transcriptional activity across states or between replicates

Please email all code, instructions to build the software and any supplemental material to joe@asimov.io.