Homework 4
BIOS-7659/CPBS-7659
Due 10/16 in class

1. RNA-seq Data and QC

   Obtaining Data

   - We will be reanalyzing data from Risso et al., 2011 `https://www.ncbi.nlm.nih.gov/pubmed/22177264`. Please read the Methods section to learn about their yeast data set.
   - The Sequence Read Archive (SRA) at NCBI is a next generation sequencing data repository. At SRA `http://www.ncbi.nlm.nih.gov/sra`, you can obtain the yeast data from Risso et al., 2011 using the identifier SRP009873. Once you search for this entry, there will be 11 results. Click on "RNA-Seq library made from YPD replicate 1" for the first two entries in Table 1 from the paper. You can click on the "SRR390924" link to get information on one of the runs, which has identifier "SRX112044".
   - To download the .fastq file for "RNA-Seq library made from YPD replicate 1", go to `http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=search_seq_name` and input SRX112044 (from above). Select the first sample SRR390924 and the FASTQ format on the right to download (this will take a few minutes, be patient).

   Online Software

   - Go to `https://usegalaxy.org/` and set up a Galaxy account using "Register" from the "User" tab at the top. For each tool you use in Galaxy, please read the description to understand what the tool returns.
   - Once you are on Galaxy, go to the "Analyze Data" tab at the top, select "Get Data" at the left menu, and then "Upload File." Select your file for upload with the following options: "Type" = "fastq" and "Genome" = "S. cerevisiae Apr. 2011". Depending on server activity, this step may take up to an hour. Once the upload is completed, you will get notice on the right under "History" and if you click on the eye icon, some of the data will be displayed in the browser.

   (a) Read information: Describe the .fastq format and include the first entry in your solution. Is the first entry a high quality or low quality read? What are the lengths of the reads? Find the number of reads in the file, either by examining the file directly or from the SRA entry.

   (b) Summary Statistics: Under "NGS: QC and manipulation" on the left menu, select "FASTQ Groomer". This is a conversion step that is necessary for running other analyses on Galaxy. Select your uploaded data for "File to Groom," "Sanger

& Illumina" for the quality score type and then "Execute". Once this step is completed, you will have a "FASTQ Groomer ..." entry under "History". Under that same left menu select "FASTQ Summary Statistics" and run it with the new "FASTQ Groomer ..." entry you created.

- Once the summary statistics step is completed, describe the columns returned by "FASTQ Summary Statistics".

- Download the table of summary statistics (click on disc icon) and read them into R to create plots (include in your solution) showing the nucleotide content across positions. What trends to do you see?

- Under "NGS: QC and manipulation", to plot the quality scores by position in the read, select "Compute quality statistics" and after that is completed, plot those results using "Draw quality score boxplot". Include the boxplot in your solution. What trend do you see across read position for quality scores?

2. RNA-seq Mapping using Bowtie

- Select "NGS: Mapping" on the left menu and then "Map with Bowtie for Illumina" under options for Illumina. Select the same reference genome as above (sacCer3), your "FASTQ Groomer ..." entry, "single-end" library, and keep the other default options.

(a) Bowtie Mapping: Running Bowtie creates a .sam file. Look at the completed Bowtie mapping entry in "History" (click on eye icon). The format of .sam files are described in `http://samtools.sourceforge.net/SAM1.pdf`. Based on the description of the "FLAG" field in Section 1.4, was the first read listed in the .sam file mapped? If not, which read going down the list is the first one mapped?

- To filter out unmapped reads, select "NGS: SAM Tools" and then "Filter SAM or BAM, output SAM or BAM". Create filters by selecting "yes" for "Filter on bitwise flag" button and for "Skip alignments with any of these flag bits set," select 'The read is unmapped" and "The read fails platform/vendor quality checks." Output "sam" format. Once this is completed, approximately what percentage of reads were filtered out (check the display in "History")?

(b) Visualization of Mapping: To view the mapping results at UCSC Genome Browser, select "NGS: SAM Tools" and then "SAM-to-BAM". Keep the default option and select your filtered mapping data. Once this is completed, in the entry listed in "History" you will find an option for "display at UCSC", click on that. You will see a display of the genome in a linear format. Each horizontal entry in the browser is a track and displays information for that part of the genome. The amount of information viewed can be controlled by the pull-down menus for each track. The track for your mapping results will be under "Custom Tracks", select the "squish view."

- What chromosome are you viewing? How many genes do you see on this part of the chromosome? (you may need to zoom in or out using the browser control buttons). Which genes have higher coverage? Include a figure in your solution of the browser image.

(c) Quantitation for Genes: Download the "Saccharomyces_cerevisiae.R64-1-1.85-v2.gtf" file from Canvas. This file contains sequence features of the Saccharomyces cerevisiae genome. Use "Get Data", "Upload file" to upload into galaxy (using .gtf file format and genome version as above). Use "NGS: RNA Analysis", "htseq-count", to count the number of reads for each of those sequence features (use the "SAM-to-BAM" alignment in your history, use"gene" for the"Feature type" option and "gene_id" for the "ID Attribute" option).

- Once this step is completed, click on the eye icon. What does HT-seq return? Include a few lines of your output.

3. Extra Credit - Statistical Theory Used in Microarray Pre-Processing - No due-date (you can turn in anytime in the semester)

   (a) In class (9/13), the paper Huber et al., (2002) was discussed ("Variance stabilization applied to microarray data calibration and to the quantification of differential expression." *Bioinformatics* 1:S96-104).

   `http://bioinformatics.oxfordjournals.org/content/18/suppl_1/S96.long`

   The authors assume a quadratic relationship between the variance and mean (Eq. 2 in the paper). Show that the variance stabilizing transformation based on this assumption is the `arsinh` function (Eq. 4 in the paper).

   (b) In class (9/13), background correction methods were discussed. Typically a background intensity value is subtracted from the foreground intensity, but there are also 'convolution' methods that model the signal intensity as

   $$observed\ signal\ (O) = true\ signal\ (S) + background\ noise\ (B)$$

   In the Affy package documentation `http://www.bioconductor.org/packages/release/bioc/vignettes/affy/inst/doc/builtinMethods.pdf` page 2 (section 2.2), they give the expected value of s given $O = o$ assuming that $S \sim \exp(\alpha)$ and $B \sim N^+(\mu, \sigma^2)$ (truncated non-negative Gaussian), with densities

   $$f(s) = \alpha e^{-\alpha x}, \quad f_N^+(b) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(b-\mu)^2/2\sigma^2} 1_{[0,\infty]}(b),$$

   where 1_A is an indicator function of the interval A.

   i. Define b as $o - s$ and find the joint density of o and s, $f(o, s)$.
   ii. Derive the marginal density of o.
   iii. Find the conditional distribution of s given $O = o$.
   iv. Derive the expectation given on page 2 (bottom of Section 2.2) in the Affy package documentation.

   The mode of the MM (mismatch) values are used to estimate μ, the average squared distance in the left tail of the distribution of MM (left of the MM mode) is used to estimate σ^2 and an exponential is fit to the right tail of the distribution of PM (right of the mode) to estimate α.