

## Analysis of High Throughput Sequence Data

### Homework Set

- 1. Prepare RNA-seq analysis environment. (Due February 27)** Prepare bioinformatic software environment for analysis of RNA-seq data, including tools for mapping, spliced alignment (optional), calculation of RPKM, and gene expression ratios for pairs of RNA-seq datasets. The student is expected to identify, select, install, and test installation of bioinformatic software to accomplish RNA-seq analysis tasks, and to become familiar with command line options in each tool to constrain analysis, including parameters for number of map sites per read, mismatches per read, and seed length.

*Example protocol to accomplish this homework module: Install Bowtie, TopHat (optional), SAMtools, Cufflinks. Test all tools and familiarize with Bowtie's -m, -n and seed length options.*

Evaluation: Oral confirmation and discussion in class that infrastructure is in place.

- 2. Establish experimental design (due February 29)** Based on class discussion of an RNA-seq time-series experiment measuring effects of "Factor X" on a selected human cell line, define one or more questions to ask based on the data and create an experimental design to answer the questions. Describe motivation, question, and design in one page or less, preferably with a table or diagram for the design.

*Example questions:*

*What genes are upregulated or downregulated under Factor X treatment compared to "buffer" treatment?*

*Do some genes continue to increase while others decrease during the treatment time period?*

*Are any of the genes whose mRNA levels increase transcription factors?*

*Are any of these genes biomarkers for signaling pathways, and thus indicators of what Factor X may be?*

*Example design: Pairwise comparison of Factor X treated vs buffer treated at each timepoint (12h, 24h, 48h).*

Evaluation: Participate in class discussion of trade-offs and motivations in experimental design. Turn in one page description of motivating question(s) and experimental design.

- 3. Execute gene expression comparisons (due March 5)** Execute pairwise comparisons of datasets based on experimental design established in #2. It is preferred that each student perform their own mapping of RNA-seq data to the hg19 reference genome via Bowtie or comparable tool, but mapping should not be a bottleneck. So in addition to fastq files, BAM files are also provided at <http://genomes.ucsd.edu/~tg/wi12/cse282-a/>. Reference chromosome sequences from hg19 and gene annotation (exon) coordinates are provided as well. Reads in the BAM files were mapped with bowtie allowing one match site maximum and three mismatches maximum per read. The student is expected to gain familiarity with the trade-offs in choice of mapping parameters and facility with handling

RNA-seq data to obtain relative gene expression values and normalized RPKM (reads per thousand transcript bases and million reads mapped) values for genes in Genes.gtf.

*Example: Use the “cuffdiff” command from Cufflinks to generate “genex” output files that contain RPKM values for each input dataset, log2 ratio values of dataset 1 vs dataset 2 for each comparison, and significance scores.*

Evaluation: Participate in class discussion of data generated, including content, implications, and computer demands. Oral confirmation that pairwise comparisons were accomplished.

- 4. Analyze results in context of experimental design (due March 7)** Describe results from pairwise comparisons (above) in 1-2 pages. As needed, provide one or several pages of supplementary data as tables or lists.

Address the following:

- a. Were the questions that motivated your experimental design answered? Discuss the answers you obtained and describe how the data supported these answers.
- b. Were there surprises in the data?
- c. How many genes were significantly up- or down-regulated in each pairwise comparison and overall (total number of significantly differentially regulated genes)?
- d. Describe how you can use your data to infer the identity of Factor X.
- e. What is Factor X? Why?
- f. What type of cell was used in the experiment?
- g. Which cellular processes seem to be invoked by Factor X?  
(for e, f, g, points will be given for well-supported answers based on data; bonus points for correct answer)

Evaluation: Turn in written description of results and answers to questions **4.a-g**.

- 5. SNV identification and evaluation (due March 9)** To become familiar with single nucleotide polymorphism/variant (SNP/SNV) inference, use SAMtools or a comparable tool to generate variant tables for each of the three Factor X treated datasets (one variant file per dataset/BAM file). Merge the BAM files and generate a variant table from the overall merged dataset (i.e., three datasets merged into one large dataset). Submit the four variant files to SeattleSeq 134 to annotate SNPs at <http://snp.gs.washington.edu/SeattleSeqAnnotation134/>.

*(Hints: If you call variants with SAMtools, you will want to use “custom” input format with reference allele in column 3 and both first and second alleles in column 4. Use option “alleles submitted”. For gene locations, use NCBI and CCDS 2011.)*

The steps above will produce four variant files and four SeattleSeq output files. Use this data to answer the following questions:

- a. How many SNVs were generated in each of the four datasets?

- b.** If different, is there a correlation between numbers of reads mapped and SNVs called?
- c.** For each of the four output files, how many of each of the following types of SNVs did you find?
  - i.** Splicesite changing
  - ii.** Nonsense (introduced early stop codon)
  - iii.** “Probably damaging”
- d.** How many variants are coding (exon or splicesite) versus non-coding (inc. intron, near gene, UTR...)?
- e.** How many coding variants are “known” versus “novel” for each of the four variant files?
- f.** If different, does the number of novel variants correlate with number of reads mapped per dataset?
- g.** Focus on splicesite and nonsense SNVs in the merged dataset only: What genes are impacted? Speculate, what might be cellular processes that would be altered or impacted by total knock-out of these genes?
- h.** Select one or several genes from **g.** that look interesting to you. Query Pubmed with the gene name and “knock-out” – are there any clues to diseases that might result? Query OMIM with gene name. Are there known diseases associated with mutations?
- i.** Select a splicesite change (if none, indicate “none were found”). Are there one or more alternative acceptor/donor splicesites nearby? How far away? Upstream or downstream? In-frame or frame-shifting? If frame-shifting, how many supurious codons will be introduced?  
*(Hint: Use the UCSC Genome Browser to enter the location of the variant and extract DNA ~100 bases upstream and downstream.)*

**Evaluation:** Turn in written answers to questions **5.a-i** plus one paragraph describing the method you used to generate variant tables including command line commands.