CBB752b14 Final Project

(DUE: 27 Apr 2014 11.59pm)

Introduction: High throughput sequencing technology has been used to study a wide variety of biological phenomena, ranging from DNA sequence variation (genomic DNA sequencing/exome sequencing) to RNA expression to protein binding sites on DNA (ChIP-seq) or RNA (CLIP-seq/RIP-seq) to determining RNA secondary structures. Because of the power of sequencing and its decreasing costs, roughly 100 distinct experimental techniques have been designed to harness high throughput sequencing for the study of particular biological problems (http://liorpachter.wordpress.com/seq/).

This assignment challenges you to learn about sequencing technologies in more depth.

- In the CBB assignment, you will program one component of an RNA-seq analysis pipeline.
- In the MBB/MCDB assignment, you will first learn a little bit about RNA-seq computation by implementing a workflow for RNA-seq analysis, using the GALAXY tool (https://usegalaxy.org/). Then, you will select a sequencing-based technology that interests you and investigate some of the challenges associated with analyzing the associated data, as well as considering the potential to leverage the data to make interesting biological predictions.

Details of both assignments are below. If CBB students want to do a different programming project related to their research, they may discuss this possibility with Mark and the TAs.

The workflow is adapted from:

https://main.g2.bx.psu.edu/u/jeremy/p/galaxy-rna-seq-analysis-exercise

Late policy: Barring a valid medical reason (with supporting documentation), or sufficient advance notice of a schedule conflict (at least two weeks before the due date), late projects will not be accepted.

Plagiarism: Following are documents on Yale's policies on academic integrity, and how to avoid plagiarism:
http://www.yale.edu/graduateschool/academics/forms/Avoiding plagiarism.pdf
http://www.yale.edu/graduateschool/academics/forms/integrity_resources.pdf

MBB/MCDB assignment:

Your final project consists of two sections: a semi-computational section (30%) and a literature survey (70%).

Please zip up all the files to be submitted, with filename according to format: netID_firstNameLastName_cbb752b14_finalproj_MBB.zip. The completed assignment should be emailed to cbb752@gersteinlab.org.

Section 1: RNA-seq read mapping in GALAXY

For the semi-computational section, to get a tiny sense of how RNA-seq informatics is done, you will use GALAXY to only do the initial quality control and then the mapping of reads. For this section, you only need to submit the answers to the parts in red.

GALAXY is an open, web-based computational portal, that is specifically designed to be "accessible, reproducible and transparent" to ALL scientists (Giardine et. al., Genome Res., 2005). It was designed in a modular fashion called 'workflows', that are meant to be intuitive and less computationally involved - specially for use by experimental researchers in genomics analyses (Brankenburg et. al., Curr Protoc Mol Biol, 2010). They offer a range of file and data manipulation tools and a myriad of other tools for downstream analyses. No programming experience is required.

1) Importing the data and understanding the data file

a) Unzip the FOUR files in 'cbb752b14_galaxy_rnaseq_files.zip' found on class wiki.
b) Upload them onto your workspace on GALAXY by the "Get Data" tool. These are four FASTQ sample files from the Illumina BodyMap 2.0 project adrenal and brain tissues. They contain data from paired-end reads of 50 base pairs on chromosome 19 physical positions 3000000-3500000.

Q1) Each read is denoted by 4 lines in a typical FASTQ file. In the first entry of adrenal_1.fastq file, what does this line show: '5.544,444344555CC?CAEF@EEFFFFFFFFFFFFFFFFEFFFEFFF'?

2) Quality Control (QC) of the reads.

The "NGS: QC and manipulation" is toolbox for FASTQ and FASTA files manipulation.

(a) Use FASTQC under “NGS: QC and manipulation” to generate a summary of the reads required for QC. There are a number of ways to assess sequence quality using FASTQC (HINT). The easiest way is by per base sequence quality.

3) Trim reads.

Based on your QC in Q2, assess the number of positions to be trimmed for each set of reads. Do the necessary trimming by using **FASTQ trimmer** under **“NGS: QC and manipulation”**. It might be a good idea to name your modules for this step more intuitively.

Q3) How many positions did you trim? Substantiate your assessment.

4) Read mapping using TopHat

The reads, if you recall, are cDNAs, so they do not contain introns. There is a potential issue of reads mapping across splice junctions, so a specialized mapper like TopHat is essential to this endeavor.

a) Go to “NGS: RNA Analysis” to find TopHat. Map the reads to hg19 “Canonical Female build”.
b) Set the mean inner distance between pairs of reads to be 110 for Illumina BodyMap data.
c) You might already have noticed by now that the dataset is made up of paired-end reads.
d) Finally, use default TopHat settings.

Q4) Copy and paste the top 10 hits for your ‘splice junctions’, ‘insertions’ and ‘deletions’.

5) Rename your workflow in the format: netID_cbb752b14_finalmywf. Publish your ‘workflow’ and obtain a URL. To publish the workflow, click on the ‘gear’ icon on the right hand side of the ‘Histories’ bar, then choose ‘Share or publish’.

Q5) Copy and paste the URL of your workflow.

Section 2: Analysis of a high throughput sequencing method of your choice

Literature Survey: Limit your writing less than 10 pages (12 pt font, Times New Roman font, double-spaced and 1" border), with citations in the end (citations and Section 1 do not count towards the 10 pages).

Suggested format:

1) Background (~ 1-2 pages)
2) Limitation discussion (~ 1-3 pages)
3) Results: Case Study (~ 3-5 pages)
4) Discussion (~1-2 pages)
5) References NOTE: Some suggested reference managers: Mendeley, Zotero or EndNote. For references, please follow the Nature citation format: http://www.nature.com/nature/authors/gta/ - a5.4.

1. Choose one sequencing method (e.g. from the *Seq page on Lior Pachter's blog (http://liorpachter.wordpress.com/seq/)., although the papers on this page are typically the first demonstration of each technology, and other papers usually have more, and more interesting, data.)

- **Note:** It's fine to choose ChIP-seq/RNA-seq. We just want to give you the flexibility to be creative.

For your chosen technique:

- Describe its basic purpose
- Explore the experimental limitations of the method (via your own observations or from literature survey). How will these limitations affect downstream analyses?
- Chose one of the following:
    - Choose a specific case study that focuses on one of the experimental limitations that you have identified, in which a computational method was used to address the problem. What was the rationale of the computational method? Was it successful? Feel free to add your own ideas about how the analysis to address this experimental problem could be further refined.
    - Could the data from your chosen sequencing method be used to predict another biological phenomenon? Think about the example of predicting RNA expression levels from ChIP-seq signals at the promoters of genes. How would you set up a model to predict your chosen phenotype?

References:

1. Pepke et. al. Computation for ChIP-seq and RNA-seq studies, *Nature Methods* (2009)
2. Wang et. al. RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews* (2009)
3. List of experimental methods that use high throughput sequencing: http://liorpachter.wordpress.com/seq/
4. Cheng, C, Alexander, R, Min, R, Leng, J, Yip, KY, Rozowsky, J, Yan, KK, Dong, X, Djebali, S, Ruan, Y, Davis, CA, Carninci, P, Lassman, T, Gingeras, TR, Guigó, R, Birney, E, Weng, Z, Snyder, M, Gerstein, M**(2012)**. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, 22, 9:1658-67.

CBB/CPSC Assignment

The final project pertains to RNA-seq analysis and consists of **3 sections**. You are required to choose only **ONE** of the sections as your project. **Programming is mandatory in the analysis and the use of portals such as GALAXY will not be accepted.**

For your submission, please provide the following:

1) **Source code** (source code should directly read input files from common directory in bulldogJ (/home1/mbb452/common/), and produce output files in your current directory).
2) **Output file(s)**
3) A short **README file** on description of your files and how to execute your program. Unexplained files will be regarded as missing/unused files.
4) A short **write-up** on the algorithm implemented. Include also your answers to the questions in red and references, maximum 3 pages (excluding references).

Please zip up all your files, with filename according to format: netID_firstNameLastName_cbb752b14_finalproj_sectionNum.zip. The completed assignment should be emailed to cbb752@gersteinlab.org. **For each of your file, please prefix: netID_firstNameLastName_sectionNum_**

NOTE: Some suggested reference managers: Mendeley, Zotero or EndNote. For references, please follow the Nature citation format: http://www.nature.com/nature/authors/gta/#a5.4.

General Introduction: RNA-seq utilizes high-throughput sequencing technology to quantify RNA expression profiles. After extracting all RNAs from cells, these RNAs are converted to complementary DNA (cDNA). cDNAs are then sheared into small fragments, typically 200-300 bp. Each of these short sequences, or "reads", is then determined by next-generation sequencing technology. Computationally, by mapping these short sequences back to the reference genome or transcriptome, gene expression level can be inferred from read abundance at each position. Downstream analysis of RNA-seq includes quantifying differential gene expression, RNA editing, transcriptional profiling, novel gene identification, alternative splicing and SNP discovery.

A typical RNA-seq pipeline consists of a few major components: read alignment and assembly, quantification and downstream analyses. The final projects will focus on major parts in RNA-seq pipeline.

Section 1: Expression quantification

[Grading policy: 80% on programming; 20% on writing. For programming part, grading will be based on:

1) **source code and output (80%)**
2) **programming style (eg. comments) and clear documentation of how your script works (README.txt), in order for us to successfully run your script (20%). Programs that do not compile will get an immediate 0. To receive partial credit, please make sure your program is well-commented]**

File directory on BulldogJ : /home1/mbb452/student19/common/Section1/

Input data 1:

We provide two RNA-seq datasets that contain the aligned reads of chromosome 19 from the liver of an embryonic mouse (GSM850907_heart-E14.5-1.chr19.sam) and adult mouse (GSM723770_RenLab-RNA-Seq-heart-ZY6.chr19.sam), from GSE29278 in the Gene Expression Omnibus (GEO) (Shen et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* 2012). These files are in SAM format.

Input data 2: You are also given a list of genomic elements with their gene names on chromosome 19 (mm9_genes_chr19.gtf) in GTF format.

Sample Test data: We have provided the FPKM values for a subset of 10,000 reads from the adult mouse dataset. The reads are contained in in.sample and the FPKM values are in out.sample.

One of the main challenges in quantification is transcript abundance estimation. There are a number of publicly available software packages that perform such estimates, for instance Cufflinks and ERANGE.

To simplify this problem, we assume that each read maps to a single-isoform gene and that reads in chromosome 18 (this file) represent the entire set of reads. (a) Extract the exons from input data 2. It is optional to include this step in the script below, i.e. not required.

(b) Fragments per kilobase of million fragments mapped, or FPKM, is a metric devised to quantitate relative abundance and also account for bias due to longer exons. Write a script to calculate the FPKM for each GENE.

$$\text{FPKM} = \frac{total\ fragments}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

Q1) The start and end positions in the GTF format are 1-based. How is this different from a 0- based system? Q2) What do values of 0 and 16 in the second column of the SAM file mean? Q3) Are the assumptions made here justified in an actual RNA-seq pipeline? What are other considerations and confounders? Discuss.

References:

1. Shen et al. A map of the cis-regulatory sequences in the mouse genome *Nature* (2012)
2. Roberts et. al., Improving RNA-seq expression estimates by correcting for fragment bias, *Genome Biology* (2011).
3. Trapnell et. al., Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* (2010).

Section 2: Detect and Analyze Differentially Expressed Genes

[Grading policy: 80% on programming; 20% on writing. For the programming part, grading will be based on: 1) source code and output (80%) 2) programming style (eg. comments) and clear documentation of how your script works (README.txt) in order for us to successfully run your script (20%). If any external pre-existing codes or libraries are used, please make sure they are properly placed or installed, so that we can easily rerun your code. Programs that do not compile will get an immediate 0. To receive partial credit, please make sure your program is well commented. In writing your program, you may use libraries and pre-existing code for certain "supporting computations", such as matrix multiplication. You may not use libraries or pre-existing code for any computation that operates directly on the network, including storage and retrieval of network nodes and edges, calculating path lengths, and computation of betweenness centrality. If you are unsure of whether a particular library is permissible, please contact the TFs first.]

File directory on BulldogJ : /home1/mbb452/student19/common/Section2/

Input data 1: Gene expression profiles (chromosome 19) derived from RNA-seq data of two samples, an embryonic mouse heart (**Embryo.fpkm**) and adult mouse heart (**Adult.fpkm**). (Gene expression values are estimated by relative reads abundance mapped to the particular gene. Values are shown in FPKM - Fragments per kilobase of million fragments mapped. FPKM is a well-recognized metric to quantify expression values, taking into account exon length bias.) **Input data 2**: A collected set of mouse protein-protein interaction data (**Mouse.PPI**) from five PPI databases: DIP, BIND, MIPS, MINT and IntAct. Each line represents an interacting gene pair.

Sample test data: We have provided the betweenness centrality and degree values for a random set of genes (selected without regard to differential expression), calculated with the given Mouse.PPI network. These files are betweenness.sample.txt and **betweenness.sample.txt** and **degree.sample.txt**.

As we learned, RNA-seq reads can be mapped and assembled to quantify gene expression values. One further application of RNA-seq is to detect differentially expressed genes under different conditions or in different tissues. Genes that are significantly highly or lowly expressed in one sample compared to other samples are termed as differentially expressed genes. Why do some genes tend to have different expression values under different conditions or different tissues? One reason is that they might play important roles that distinguish one tissue from another. To get a glimpse of their functional involvement in particular conditions or samples, we will investigate if they are enriched in certain functional pathways. This pathway information provides us the clue to the functional differences. Here, we provide a simplified procedure aiming to analyze differentially expressed genes in embryo and adult hearts.

1. Identify differentially expressed genes and their functions.

The simplest way to detect differentially expressed genes is to directly compare expression values. For each gene in Input 1, obtain its expression ratio in embryonic and adult mouse hearts. Set **your own ratio cut-off** to define differentially expressed genes.

Q1) Provide the gene expression ratio file indicating which sample is the denominator. What is your cut-off? Justify your choice of this cut-off.

Q2) Based on your cut-off, how many genes have high relative expression in the embryonic heart? How many are more highly expressed in the adult heart?

Upload the gene lists that have high relative expression in embryonic and adult hearts separately to DAVID (a functional annotation tool) to perform function and pathway analysis. (HINT: go to "Functional Annotation", then upload or paste the file with "select identifier - official_gene_symbol")

Q3) Give a snapshot of the DAVID outputs. Present and comment on the results.

2. Network Analysis of differentially expressed genes.

In biological networks, hub proteins tend to be more conserved during evolution and more likely to be essential. Here we try to quantify network positions of differentially expressed genes derived from the previous step.

For **EACH GENE** in the mouse PPI (not just differentially expressed ones), **write your own script** to calculate their degree centrality and betweenness centrality. Degree centrality is defined as the number of edges that a node has. Betweenness centrality is defined as:

$$g(v) = \sum_{s \neq v \neq t} (\sigma_{st}(v)/\sigma_{st})$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of these paths through node v. Your program should produce three-column list where the first column is the gene, the second column is the degree centrality, and the third column is the betweenness centrality.

Q4) List the centralities of the differentially expressed genes, if they exist.

Q5) Use Cytoscape to visualize the network, and highlight the differentially expressed genes.

[**Appendix, additional information on the quantitation of differential gene expression**: The simple cut-off method does not take into account the underlying possible gene expression levels. In other words, gene expression value under

certain conditions possess certain variations. To obtain accurate significant P values, RNA-seq experiments are usually done with replicates. The sound statistical framework is described below.

To test whether an observed difference in a gene's expression is significant, first get the expression ratio in two conditions:

$$Y = \frac{FPKM_a}{FPKM_b}$$

The log of the ratio (T) of expression in two conditions can actually be used as a test statistic, because the quantity:

$$T = \frac{E[\log(Y)]}{Var[\log(Y))]}$$

is approximately normally distributed and can be calculated as

$$T = \frac{E[\log(Y)]}{\mathrm{Var}[\log(Y)]} \approx \frac{\log\left(\frac{FPKM_a}{FPKM_b}\right)}{\sqrt{\frac{\mathrm{Var}[FPKM_a]}{FPKM_a^2} + \frac{\mathrm{Var}[FPKM_b]}{FPKM_b^2}}}$$

With replicated, expression variations under same conditions are easy to get and be applied to this equation. The corresponding P value could be obtained from normal distribution.

In our case, without replicates, cuffdiff (cufflinks software) can tackle this problem based on certain assumptions (for details, go to http://cufflinks.cbcb.umd.edu/howitworks.html#hdif).]

References:

1. Xiao Li, Jiabao Xu, Haoyang Cai, Yizheng Zhang. A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. *Amino Acids* (2010)
2. Trapnell et. al., Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* (2010).

Section 3: SNV calling and functional interpretation

[Grading policy: 60% on programming; 40% on writing. For the programming part, grading will be based on: 1) source code and output (80%) 2) programming style (eg. comments) and clear documentation of how your script works (README.txt), in order for us to successfully run your script (20%). For the writing part, incomplete answers will receive partial credit. Programs that do not compile will get an immediate 0. To receive partial credit, please make sure your program is well-commented.]

File directory in BulldogJ: /home/cpsc752/student19/common/Section3/

Input data

To simplify this process**, spleen_chr18.pileup** is provided (pileup format file facilitating SNP/indel calling). For detailed information, see Pileup format. To obtain this format from RNA-seq aligned data (from the same study as the data from section 1) [1], SAMtools (pileup function) is used to convert Bam file to Pileup format. This file contains reads in mouse chr18: 3,000,000 - 6,000,000. **Input data 2**: You are also given a list of genomic elements with their gene names on chromosome 19 (**mm9_genes_chr18.gtf**) in GTF format.

Ten column pileup format

The ten-column (consensus) pileup incorporates additional consensus information:

```
  1   2 3 4  5  6  7  8     9      10
-----------------------------------------
chrM 412 A A 75  0 25 2     .,      II
chrM 413 G G 72  0 25 4    ..t,    IIIH
chrM 414 C C 75  0 25 4    ...a    III2
chrM 415 C T 75 75 25 4    TTTt    III7
```

where:

Column Definition

---

1. Chromosome
2. Position (1-based)
3. Reference base at that position
4. Consensus bases
5. Consensus quality (Phred scaled consensus quality)
6. SNP quality (Phred scaled probability of difference from reference bases)

7. Maximum mapping quality
8. Coverage (# reads aligning over that position)
9. Bases within r
10. Quality values (ASCII: phred+33 scale)

---

1. SNV (single nucleotide variant) calling.

Similar to direct genomic DNA sequencing, RNA-seq technology can also be used to call variants. Instead of calling genome-wide variants, RNA-seq variant calling identifies **expressed** variants. Compared to variants that occur in silent genes, expressed variants are much more likely to possess functional impact on proteins, thus affecting phenotypes. There are a number of software that perform variant calling, such as VarScan, Bcftool, GATK and SOAPsnp. These methods apply sophisticated statistical models to infer variants.

Here, we will perform a simple variant filtering process to find potential SNVs from the raw pileup file. Please write **your own code** following the filtering criteria listed below and report identified SNVs in the following format: Chromosome, Position, Reference base at that position, Number of A reads, Number of C reads, Number of G reads, Number of T reads, Quality adjusted read coverage, also genes (if any) the SNV resides in).

Filtering:

1) minimum SNP quality is 20.
2) minimum Maximum mapping quality is 25.
3) Filter out read mapping bases with quality (phred quality score) lower than 20.
4) After filtering, the read depth should between 3 and 100.
5) After filtering, at least one mapped reads contain mutation.

Q1) What is Phred Quality Score? How is it defined? Q2) The simple filtering process will certainly reduce false positive rate. What are the potential drawbacks of this approach? Q3) Provide the SNV calling results. Among these identified SNVs, do you observe variants with partially mutated bases (bases in aligned reads)? If yes, give some explanations. For positions with 100% mutated bases, what are the situations that can result in such cases?

2. Variant Functional Interpretation.

Variants occurring in different genomic locations may have different functional impacts. For example, non-synonymous variants in translated genes are more likely to have an impact than synonymous variants. Many methods have been developed to assess SNV effect, such as VEP (Variant Effect Predictor) and VAT (Variant Annotation Tool). Here, we will use VEP to annotate our identified SNVs. Go to http://may2012.archive.ensembl.org/Mus_musculus/Info/Index, click “Manage Data”, then “Variant Effect Predictor”. (Convert your SNV file to proper

format, use “+” as strand information).

Q4) Paste your result from VEP. What do synonymous and nonsynonymous SNVs mean? Even in nonsynonymous SNVs, you could further quantify their potential damage effect. Please provide some other ideas that could be used to further interpret or validate SNV effects. Q5) RNA-seq, in theory, should capture mRNA sequences. However, some SNVs occur in intronic regions which are not transcribed. What could possibly explain this phenomenon that we are observing these ‘intronic’ SNVs in RNA-seq?

Go to the UCSC Genome Browser, and navigate to those SNV positions in the mm9 mouse genome assembly. Look at the vertebrate multiple alignment for these positions. If you are unfamiliar with the UCSC genome browser, please check out this link: http://www.openhelix.com/ucsc.

Q6) Show a print-screen of the multiple alignment of one of the SNVs found in UCSC. How can one tell if these SNVs are evolutionarily important? Based on this, can you infer if ANY positions are possibly more important than others?

References:

1. Shen et al. A map of the cis-regulatory sequences in the mouse genome *Nature* (2012)
2. Heng Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics* (2011)
3. Li H, Handsaker B, et al., 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (2009)