Analysis of data from high-throughput molecular biology experiments
BB2490 and DD2399 (*KTH*) 7.5 hp – spring 2015

Study questions.

Statistics

Read through the questions from students at the course web, and make sure you can answer them.

- What are multiple-hypothesis corrections, and why are they needed?
- Describe how we normally infer a difference in a particular quantitative train between two different populations.
- You encounter two experimentalists arguing over the definition of p values One claims that the p-value is "the probability of the null hypothesis to be true", the other that it is "probability that replicating the experiment would not yield the same conclusion". Why are they both wrong?
- Why do statisticians normally not advocate Bonferroni corrections, they seem so simple to perform?
- How can it be that we can estimate the number of false positive features in a set of identifications given their p values? How come that the formula is so simple? Why do we not investigate the p-value distribution of false positive features when using well calibrated tests.
- How does the q-value differ from the FDR, and why does that distinction matter?
- Is P(A|B) the same thing as P(B|A)? Why is that question relevant for understanding the difference between a p-value and a PEP (local FDR)?
- What is meant with the local FDR, or PEP. When should we control for PEP (local FDR) rather than FDR?
- What is a type I (one) error?
- Why are we generally focusing more on FPs than FNs for high-throughput experiments?

Genomics and transcriptomics

- What is the main advantage of the "next-generation" sequencing technologies compared to Sanger?
- What read lengths are possible with the Illumina HiSeq sequencing technology?
- Describe the .fastq format.
- How is the base quality score defined? (Definition originating from the Phred program).
- What does a base quality score of 40 mean? (i.e., what probability of a wrong base call does it translate into).
- Explain the "seed-and-extend" procedure for aligning reads to a reference genome.
- Explain why a "seed-and-extend" procedure for alignment is needed.
- What is a "reference genome"?
- Some aligners output a mapping quality. What does a mapping quality of 0 (zero) mean from BWA?
- What is whole-genome shotgun sequencing?
- Explain the following terms:
  - Contig
  - Scaffold
  - N50
- Explain how read-pairs (mate-pairs, or paired-end reads) can be used to obtain the ordering of scaffolds.
- What are the computational challenges of de novo sequence assembly?
- Why are repeat regions so detrimental to de novo sequence assembly?
- What is the OLC approach to de novo sequence assembly?
- How is a de Bruijn graph constructed from a set of (potentially very many) sequence reads?
- What does it mean that *k*=19 in a de Bruijn graph?

- If the read length is *l*, what is the maximum value of *k*?
- Draw a picture of how a repeated region would be represented in a de Bruijn graph (only draw a suitable local graph structure).
- Describe or list a few heuristics used to cope with complex de Bruijn graph structures.
- Describe the typical features of a human gene.
- Define “transcriptome”.
- List at least three phenomena that may contribute to the complexity of a transcriptome.
- Why do you have to be more careful when aligning RNA-seq reads to a reference genome than when aligning reads originating from genomic DNA?
- Why do you need to normalize the RNA-seq counts even when you only want to compare gene expression within a sample?
- What is “RPKM”?
- What is “FPKM”?
- Is RPKM (or FPKM) always the best way of normalizing transcript abundance? Can you think of a situation when RPKM would give values that are not comparable between two samples?
- What does the program TopHat (and TopHat2) do?
- What does the program DEseq do?
- What does the program Macs?
- The endpoint in many RNA-seq experiments is to assess differential expression. The background distribution is often modeled as a Poisson distribution. What is the main criticism of using a Poisson distribution for this? What has been launched as an alternative background distribution? And why is it claimed to be better?
- Suggest a pipeline (work flow) for an RNA-seq experiment, starting with the unmapped reads and ending with an assessment of differential expression. Include examples of programs to use.
- What is the difference between
  - core promoter
  - distal promoter
  - enhancer/silencer
- Is the promoter region always strictly upstreams of the TSS (transcription start site)?
- Explain what is an epigenetic modification (of genomic DNA). Give two examples of such a modification.
- What is a histone?
- What does “open chromatin” mean?
- Give example of 2 repressive and 2 activating histone modifications.
- Describe the ChIP protocol.
- In a ChIP-seq experiment analysis, what is a “peak”?
- Why is there a shift in the position of the peaks originating from reads from the two different strands? How are these peaks merged in the peak calling algorithms (2 different ways of doing this).
- What is “input DNA” and why is it suitable as a control sample?
- How can the result of a ChIP-seq peak calling be validated
  - informatically?
  - experimentally?
- Suggest a way to estimate the FDR (false discovery rate) of a ChIP-seq peak calling.
- Peaks from transcription factor binding sites (TFBS) and from histone modifications (in particular repressive marks) have different expected widths. Which ones are the more narrow?
- In peak calling algorithms, at least three different distributions to model the background have been used. One is negative binomial, as used by DEseq (that can handle both ChIP-seq and RNA-seq data). Name another.
- Why is fold enrichment alone not a good measure of whether a peak is significantly enriched?
- What is the “local lambda” of the MACS peak calling method? (peak calling: the

procedure of going from mapped reads to delineating a set of regions with significantly enriched read density so as to imply that this region was pulled out in the ChIP experiment).

- Can you suggest a way to include the sequence quality values in a ChIP-seq peak calling algorithm?
- What is the "*-seq" paradigm?
- How much of the human genome is believed to be transcribed? (see ENCODE papers).
- What is the difference between exon union and exon intersection approaches when doing read counting for differential expression analysis?
- In differential expression analysis, what is "effect size"?
- When finding out about differential expression, which measure is the most important, effect size or the p-value? (or both?)
- Describe the term "dynamic range" in RNA-seq?
- In "transcriptome reconstruction" (also called transcriptome assembly), what data do you start with? And what is the outcome?
- Name one similarity and one difference between *genome* assembly and *transcriptome* assembly.
- What is the difference between genome *dependent* and genome *independent* transcriptome reconstruction?
- If you have a lot of type I errors in a differential expression analysis, does that mean that you have too many genes predicted to be differentially expressed (i.e., many false positives), or does it mean that you have too few genes predicted to be differentially expressed (i.e., many false negatives)?

Proteomics

- Why do the concentration differences between different proteins represent a challenge for proteomics?
- Why is the large number of types of proteins in a cell represent a problem for proteomics?
- What is the difference between a mono-isotopic mass and an average mass?
- How can one calculate the mass of a peptide?
- Why is electrospray preferred over MALDI as ionization source in shotgun proteomics?
- What is a MS/MS fragmentation mass spectrum, and how does it differ from an MS?
- What is a y-ion?
- Describe the steps of shotgun proteomics.
- Describe some common PTMs.
- Describe how you may extend a spectral search engine to search for PTMs.
- What is the most common null model when assessing the quality of output from spectral search engines for shotgun proteomics?
- What is a PSM?
- In what ways are SRM mass spectrometers different from normal mass spectrometers?
- Why is it hard to use a mass spectrometer for absolute quantification of proteins?
- What are the advantages and disadvantages of SILAC over iTRAQ? What are the pros and cons of labeled versus unlabeled techniques for quantification?