UCDAVIS Bioinformatics Core

Experimental Design

Dr. Matthew L. Settles

Genome Center
University of California, Davis
settles@ucdavis.edu

What is Differential Expression

Differential expression analysis means taking *normalized* sequencing fragment count data and performing statistical analysis to discover *quantitative* changes in expression levels between experimental groups.

For example, we use statistical testing to decide whether, for a given gene, an observed difference in fragment counts between group A and group B is significant, that is, whether it is greater than what would be expected just due to natural random variation.

Designing Experiments

Beginning with the question of interest (and work backwards)

- The final step of a DE analysis is the application of a model to each gene in your dataset.

  Traditional statistical considerations and basic principals of statistical design of experiments apply.

  - **Control** for effects of outside variables, avoid/consider possible biases, avoid confounding variables in sample preparation.
  - **Randomization** of samples, plots, etc.
  - **Replication** is essential (triplicates are THE minimum)

- You should know your final (DE) model and comparison contrasts before beginning your experiment.

Three outcomes
Goldilocks and the three bears

- Technical and/or biological variation exceeds that of experimental variation, results in 0 differentially expressed genes
- Experiment induces a significant phenotype with cascading effects and/or little to no biological variation between replicates (ala cell lines), results in 1000s of DE genes. Some of which are directly due to experiment; however, most due to cascading effects.
- Technical artifacts are controlled. Biological variation is induced in the experiment, and cascading effects are controlled, or accounted for, results in 100s of DE genes directly applicable to the question of interest.

General rules for preparing and experiment/ samples

- Prepare more samples then you are going to need, i.e. expect some will be of poor quality, or fail
- Preparation stages should occur across all samples at the same time (or as close as possible) and by the same person
- Spend time practicing a new technique to produce the highest quality product you can, reliably
- Quality should be established using Fragment analysis traces (pseudo-gel images, RNA RIN > 7.0)
- DNA/RNA should not be degraded
  - 260/280 ratios for RNA should be approximately 2.0 and 260/230 should be between 2.0 and 2.2. Values over 1.8 are acceptable
- Quantity should be determined with a Fluorometer, such as a Qubit.

Sample preparation

In high throughput biological work (Microarrays, Sequencing, HT Genotyping, etc.), what may seem like small technical details introduced during sample extraction/preparation can lead to large changes, or technical bias, in the data.

Not to say this doesn't occur with smaller scale analysis such as Sanger sequencing or qRT-PCR, but they do become more apparent (seen on a global scale) and may cause significant issues during analysis.

Be Consistent

BE CONSISTENT ACROSS ALL SAMPLES!!!

UCDAVIS Bioinformatics Core

Generating RNA-seq libraries

Considerations

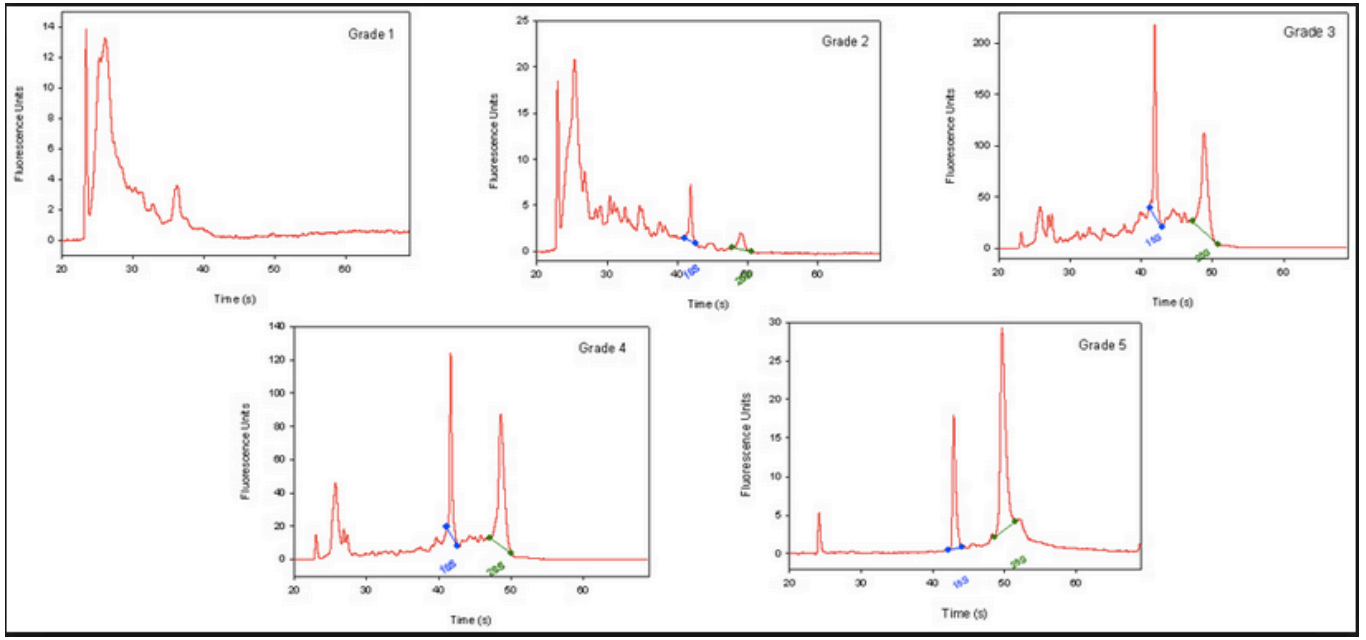
- QA/QC of RNA samples
- What is the RNA of interest
- Library Preparation
  - Stranded Vs. Unstranded
- Size Selection/Cleanup
- Final QA

UCDAVIS Bioinformatics Core

QA/QC of RNA samples

RNA Quality and RIN (RQN on AATI Fragment Analyzer)

- RNA sequencing begins with high-quality total RNA, only an Agilant BioAnalyzer (or equivalent) can adequately determine the quality of total RNA samples. RIN values between 7 and 10 are desirable.



BE CONSISTANT!!!

RNA of interest

- From “total RNA” we extract “RNA of interest”. Primary goal is to NOT sequence 90% (or more) ribosomal RNAs, which are the most abundant RNAs in the typical sample. there are two main strategies for enriching your sample for “RNA of interest”.
  - polyA selection. Enrich mRNA (those with polyA tails) from the sample by oligo dT affinity.
  - rRNA depletion. rRNA knockdown using RiboZero (or Ribominus) is mainly used when your experiment calls for sequencing non-polyA RNA transcripts and non-coding RNA (ncRNA) populations. This method is also usually more costly.

rRNA depletion will result in a much larger proportion of reads which align to intergenic and intronic regions of the genome.

Library Preparation

- Some library prep methods first require you to generate cDNA, in order to ligate on the Illumina barcodes and adapters.
  - cDNA generation using oligo dT (3’ biased transcripts)
  - cDNA generation using random hexomers (less biased)
  - full-length cDNAs using SMART cDNA synthesis method
- Also, can generate strand specific libraries, which means you only sequence the strand that was transcribed.
  - This is most commonly performed using dUDP rather than dNTPs in cDNA generation and digesting the “rna” strand.
  - Can also use a RNA ligase to attach adapters and then PCR the second strand and remainder of adapters.

Size Selection/Cleanup/qA

Final insert size optimal for DE are ~ 150bp (or kit suggestion)

- Very important to be consistent across all samples in an experiment on how you size select your final libraries. You can size select by:
  - Fragmenting your RNA, prior to cDNA generation.
    - Chemically heat w/magnesium
    - Mechanically (ex. ultra-sonicator)
- Cleanup/Size select after library generation using SPRI beads or (gel cut)
- QA the samples using an electrophoretic method (Bioanalyzer) and quantify with qPCR.

Most important thing is to be consistent!!!

UCDAVIS Bioinformatics Core

[SUMMARY] Generating RNA-seq libraries

Considerations

- QA/QC of RNA samples [Consistency across samples is most important.]
- What is the RNA of interest [polyA extraction is recommended.]
- Library Preparation
  - Stranded Vs. Unstranded [Standard stranded library kits]
- Size Selection/Cleanup [Target mean 150bp or kit recommendation]
  - Final QA [Consistency across samples is most important.]

Sequencing Depth

- Coverage is determined differently for ”Counting” based experiments (RNAseq, amplicons, etc.) where an expected number of reads per sample is typically more suitable.
- The first and most basic question is how many reads per sample will I get
  Factors to consider are (per lane):
  1. Number of reads being sequenced
  2. Number of samples being sequenced
  3. Expected percentage of usable data
  4. Number of lanes being sequenced

$$\frac{reads}{sample} = \frac{reads.sequenced * 0.8}{samples.pooled} \text{ X num.lanes}$$

- Read length, or SE vs PE, does not factor into sequencing depth.

Sequencing

Characterization of transcripts, or differential gene expression

Factors to consider are:

- Read length needed depends on likelihood of mapping uniqueness, but generally longer is better and paired-end is better than single-end. (2 x >75bp is best)
- Interest in measuring genes expressed at low levels (<< level, the >> the depth and necessary complexity of library)
- The fold change you want to be able to detect (< fold change more replicates, more depth)
- Detection of novel transcripts, or quantification of isoforms requires >> sequencing depth

The amount of sequencing needed for a given sample/experiment is determined by the goals of the experiment and the nature of the RNA sample.

Barcodes and Pooling samples for sequencing

- Best to have as many barcodes as there are samples
  - Can purchase barcodes from vendor, generate them yourself and purchase from IDTdna (example), or consult with the DNA technologies core.
- Best to pool all samples into one large pool, then sequence multiple lanes
- IF you cannot generate enough barcodes, or pool into one large pool, RANDOMIZE samples into pools.
  - Bioinformatics core can produce a randomization scheme for you.
  - This must be considered/determined PRIOR to library preparation

Cost Estimation

- RNA extraction and QA/QC (Per sample)
- Enrichment of RNA of interest + library preparation (Per sample)
  - Library QA/QC (Bioanalyzer and Qubit)
  - Pooling ($10/library) [If you do your own libraries]
- Sequencing (Number of lanes)
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

http://dnatech.genomecenter.ucdavis.edu/prices/

Example: 12 samples, ribo-depletion libraries, target 30M reads per sample, Hiseq 3000 (2x100).

Illumina sequencing

- http://www.illumina.com/systems/hiseq-3000-4000/specifications.html

2500

MiSeq

	HISEQ 3000 SYSTEM	HISEQ 4000 SYSTEM
No. of Flow Cells per Run →	1	1 or 2
Data Yield: ← 2 × 150 bp 2 × 75 bp 1 × 50 bp	 650-750 Gb 325-375 Gb 105-125 Gb	 1300-1500 Gb 650-750 Gb 210-250 Gb
Clusters Passing Filter (Single Reads) (8 lanes per flow cell) →	2.1-2.5 billion	4.3-5 billion
Quality Scores: ↑ 2 × 50 bp 2 × 75 bp 2 × 150 bp	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1-3.5 days	< 1-3.5 days
Human Genomes per Run*	up to 6	up to 12
Exomes per Run**	up to 48	up to 96
Transcriptomes per Run***	up to 50	up to 100

Cost Estimation

- 12 Samples
  - QA Bioanalyzer = $98 for all 12 samples
  - Library Preparation (ribo-depletion) = $383/sample = $4,596
- Sequencing = $2,346 per lane PE100
  - 2.1 - 2.5 Billion reads per run / 8 lanes = Approximately 300M reads per lane
  - Multiplied by a 0.8 buffer equals 240M expected good reads
  - Divided by 12 samples in the lane = 20M reads per sample per lane.
  - Target 30M reads means 2 lanes of sequencing $2346 x 2 = $4692
- Bioinformatics, simple comparison design, DE only $2000
  - This is the most basic analysis, for in depth collaborative analysis double sequencing budget.

Total = $98 + $4596 + $4692 + $2000 = $11,386

Approximately $950 per sample @ 40M reads per sample

UCDAVIS Bioinformatics Core

RNA-seq pipeline overview