Some Issues of Statistical Design & Analysis in RNA-seq Experiment

S. Bashir

University Health Network

November 27, 2012

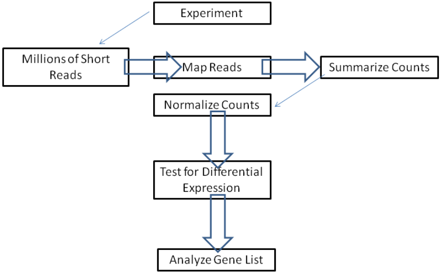
1. Sequencing Experiment
    - Experimental design
    - Biases
    - Read Mapping

2. Analysis
    - Normalization
    - Statistical Analysis

High-Throughput Gene Expression Assays

- Burgeoning field of so-called next-generation sequencing (NGS)/second-generation sequencing/ultra-high-throughput sequencing (UHTS).
- Platforms:
  - Illumina/Solexa's Genome Analyzer, HiSeq systems, MiSeq etc.
  - Applied Biosystems' SOLiD,
  - Roche's 454 Life Sciences.
  - Helicos BioSciences' HeliScope,

RNA Sequencing Pipeline

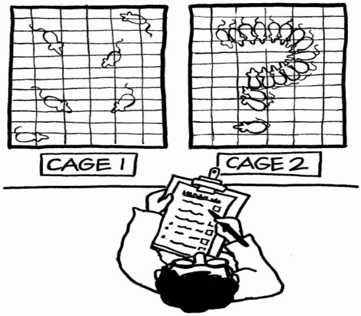


Data Types

- Sequencing assays provide digital measures of sequence abundance, i.e., read counts.
- In contrast, microarrays provide analog measures of sequence abundance, i.e., fluorescence intensities.
- Short sequence reads are aligned against reference sequence, e.g., genome, transcriptome.

Designing your experiment before you start

- How to avoid Confounding sources of variation in the data.
- While it would be nice to be able to partition various sources of technical variation (such as labeling, RNA extraction), it often is too expensive to perform such a design.
- 3 fundamental principles of experimental design, i.e., **replication, randomization & blocking**



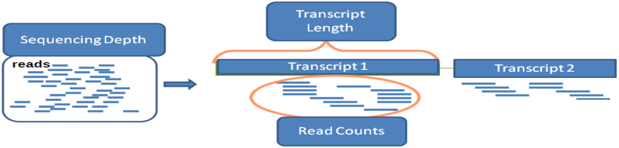
Despite the clear difference between the treated and control groups, something made him question the data.

Design Considerations

- Biological comparison
- Paired end vs single end reads
- Read length & depth
- Replicates
- Pooling

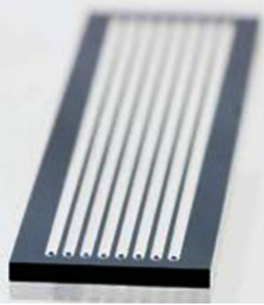
Terminology

- **Sequencing Depth or Coverage**: Total number of reads mapped to the genome/transcriptome, *aka* Library size/Sample size.
- **Transcript/Gene length:** Number of bases.
- **Read counts:** Number of reads mapping to that gene/transcript (expression measurement).



Illumina's Sequencing Technology

- One Flow Cell: Eight lanes
- Up to eight samples are hybridized to an eight-lane flow cell, one lane is often used for the control sample.



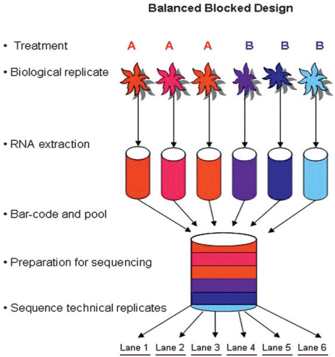
Illumina's Sequencing Technology

- **Multiplexing**
  - A way to save money by sequencing multiple samples on a single unit (an illumina's flow cell)
  - offers the flexibility to construct balanced blocked designs for the purpose of testing differential expression.
- **Barcoding:** To separate inputs, can have many barcodes in a single unit
- 12 different samples can be indexed with unique subsequences and loaded onto each lane. In total, 96 samples can be sequenced per run.
- and the output can be deconvoluted to individual samples.

Illumina's Sequencing Technology : Balanced Block Designs

- All the samples of RNA are barcoded & pooled into the same batch and then sequenced in one lane of a flow cell.
- Any batch effects are the same for all the samples, and all effects due to lane will be the same for all samples.
- This can be achieved barcoding the RNA immediately after fragmentation
- Sequencing lanes can also serve as blocks when bar-coding during library preparation

Illumina's Sequencing Technology: Bar Coding



©Auer and Doerge 2010

Biological Replicates

- If you have limited resources, it is generally far better to have biological replication (independent biological samples for a given treatment) than technical replication
- Biological replicates essential for differential expression analysis
- Technical replicates useful for trouble shooting only, not needed as low technical variation in the technology
- For a sufficient number of biological replicates certain designs can accommodate lane and/or flow cell as blocking factor.

Pooling & Balanced Designs

- Limited RNA obtainable
- **Complete Pooling:** All samples from one treatment group are pooled
  - No replication for one treatment. This approach does not provide an estimate of variability and therefore can not be used for statistical analysis.
- **Sub-Pooling:** Subsets of samples are randomly selected and pooled but there are still **multiple replicates** within each group.
- Multiple pools per group required
- Better power than complete pooling
- Equal pooling with replicates in the same pool contributing equally has better power
- Balanced designs have better statistical power

Modes of Sequencing

- **Single-end Read:** One read sequenced from one end of each cDNA insert
- **Paired-end Read:** two reads sequenced from each cDNA sample insert (one from each end)
- The reads are typically $30 - 400$ *bp*, depending on the DNA-sequencing technology used.
- The costs of paired end sequencing are higher than single end sequencing

Technical Effects

- Sequence eight samples simultaneously in the eight lanes in one flow cell in illumina, variation from one flow cell to another resulting in flow cell effect.
- In addition, there exists variation between the individual lanes within a flow cell due to systematic variation in sequencing cycling and/or base-calling.
- The flow cell and lane effects are relatively small.
- Blocking design can be used to eliminate the flow cell and lane effects
- Among these sources of variation, the library preparation effect is the largest.

Technical Effects

- RNA-seq Biases
    - Sequencing depth
    - RNA composition effect,
    - Differences in the counts distribution among samples.

Read Mapping

- Following sequencing, the resulting reads are either aligned to a reference genome (fasta file) or reference transcripts, or assembled de novo without the genomic sequence to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene
- An important summary statistic is the number of reads in a class; for RNA-Seq, this read count has been found to be (to good approximation) linearly related to the abundance of the target transcript.
- Open source Tuxedo suite comprising Bowtie, TopHat, & Cufflinks
- SAM/BAM files

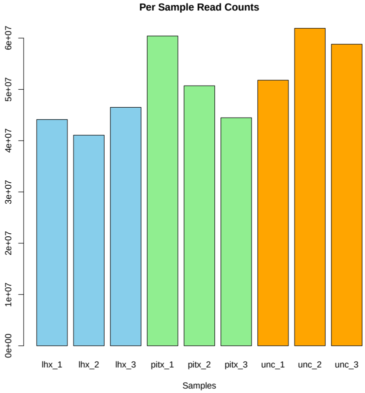
Data

- The count data are presented as a table which reports, for each sample, the number of reads that have been assigned to a gene.
- Analogous analysis also arise for other assay types, such as comparative ChIP-Seq.
- Interest lies in comparing read counts between different biological conditions, i.e., differential expression analysis.
- A gene is declared differentially expressed if an observed difference or change in read counts between two experimental conditions is statistically significant,

Data

Transc	lhx1	lhx2	lhx3	pitx1	pitx2	pitx3	unc1	unc2	unc3
1	0	0	0	0	0	0	1	1	0
2	0	0	0	0	0	0	1	2	0
3	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	1	0	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Library Size



Some Normalization Methods

Minimize the technical biases by normalizing counts data.

- **Proportion of reads:** number of reads (n) mapping to an exon (gene) divided by the total number of reads (N), n/N.
- **RPKM:** Reads Per Kilobase of exon (gene) per Million mapped sequence reads, $10^9 n/(NL)$, where L is the length of the transcriptional unit in bp (Mortazavi et al., Nat. Meth., 2008)
- **FPKM (Trapnell et al., 2010):** Instead of counts, Cufflinks software generates FPKM values (Fragments Per Kilobase of exon per Million fragments mapped) to estimate gene expression, which are analogous to RPKM.

Normalization Methods Cont'd

- **TMM (Robinson and Oshlack, 2010):** Trimmed Mean of M values.
  - For seq data, gene-wise log-fold change are : $M_g = \log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N'_k}$
  - Absolute Expression levels as $A_g = \frac{1}{2}\log_2\{Y_{gk}/N_k . Y_{gk'}/N'_k\}$
  - Trim both observed M values (30%) & A values (5%) (defaults)
  - Take the weighted average of M_g values, using inverse of the approximate variance as weights
  - Normalization factors across several samples calculated taking one sample as reference & calculating TMM for the others
  - The RNA seq data themselves do not need to be modified, but normalization factors incorporated into statistical methods for differential expression

$$\log_2(TMM_k) = \frac{\sum\limits_{g \in G^*} w_{gk} M_{gk}}{\sum\limits_{g \in G^*} w_{gk'}}$$

Normalization Methods Cont'd

- **Upper-quartile (Bullard et al., 2010):** Counts are divided by upper-quartile of counts for transcripts with at least one read.
- **Conditional Quantile Normalization,** as in microarray normalization (Hansen et al., 2012). This combines the robust generalized regression to correct for GC-bias & quantile normalization

Statistical Distributions

- For microarray normal distribution based methods are most common
- Sequencing data is counts, transformation of count data is not well approximated by continuous distributions, especially in the lower count range and for small samples
- Statistical distributions for discrete data are used
- Relevant distributions are
    - Binomial distribution
    - Poisson distribution
    - Negative binomial distribution

Poisson Distribution

- The read counts were first modelled using a Poisson distribution.
  - Poisson distribution is used for count data
  - It has only one parameter, i.e., mean
  - It assumes that mean and variance are the same ▸ Go to.
- RNA seq data represent overdispersion (i.e., variance of counts larger than mean).
- Biological variability of RNA-seq count data cannot be captured using the Poisson distribution
- Negative Binomial (NB) distribution takes into account overdispersion; hence, it has been used to model RNA-seq data

Statistical Model

- The expression quantification problem can be framed in terms of generalized linear models (GLM),

$$\ln(E[y_{gi}|N_i]) = \log N_i + x_i^T \beta_g,$$

  where
  - y_{gi} : read count for the gth gene in the ith sample
  - x_i: is the vector of covariates
  - ($log\ N_i$): offset, e.g., the total number of mapped reads
  - β_g is the vector of regression coefficients
  - and possibly other technical effects
- Information sharing among genes (Bayesian gene-wise dispersion estimation)

Negative Binomial Distribution

- The negative binomial distribution is common when count data has variance significantly greater than its mean (overdispersed)
- The NB distribution has mean μ and variance $\mu + \alpha\mu^2$; as α goes to 0, the distribution goes to a Poisson
- It is used to model biological replicates
- The number of replicates in data sets of interest is often too small to estimate both parameters, mean and variance, reliably for each gene.
- For edgeR, Robinson and Smyth assumed that mean and variance are related by $\sigma^2 = \mu + \alpha\mu^2$, with a single proportionality constant α that is the same throughout the experiment and that can be estimated from the data.

Negative Binomial Distribution:Hypothesis

- The count for a given gene in sample i come from negative binomial distributions with the mean μ_{gi} and variance $\sigma^2 = \mu_{gi} + \alpha\mu_{gi}^2$, with a single proportionality constant α
- The experimental condition r has no influence on the expression of the gene under consideration:

$$H_o : \mu_{g1} = \mu_{g2}$$

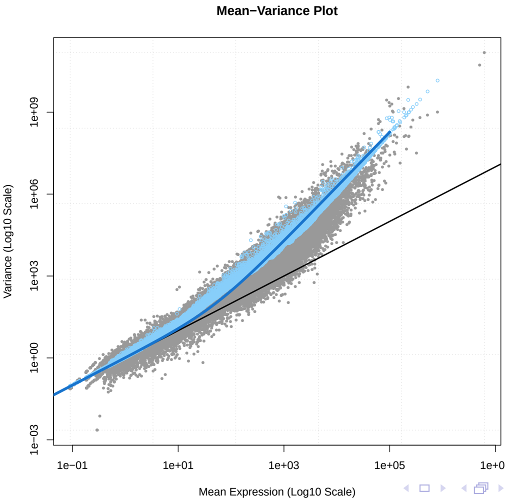
Test Statistic

Dependent on the software

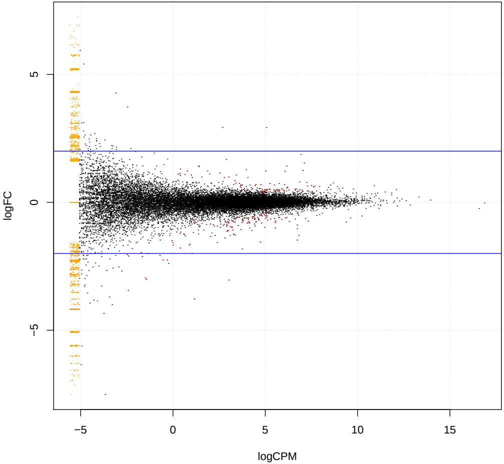
- Exact test statistic (Robinson and Smyth., 2008)
- Log-likelihood ratio (LLR) statistics based on log-linear regression models

Model Fitting

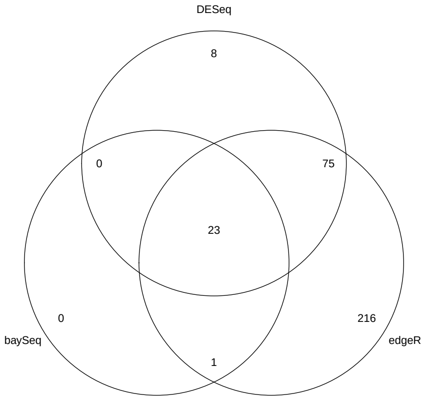
Go to



Smear Plot



Venn Diagram



Some Bioconductor and R packages

- library(DEGseq) (Wang et al., 2010): MA-plots based methods (MATR and MARS), assuming Normal distribution for *M&A*.
- library(edgeR) (Robinson et al., 2010): Exact test based on Negative Binomial distribution or GLM for multi-factor designs followed by Empirical Bayes method to evaluate the differences across transcripts.
- library(DESeq) (Anders and Huber, 2010): Exact test based on Negative Binomial distribution and a shrinkage estimator for the distribution's variance
- library(baySeq) (Hardcastle et al., 2010): Estimation of the posterior likelihood of differential expression (or more complex hypotheses) via empirical Bayesian methods using Negative Binomial distributions.

Message

- Design the experiment properly, i.e., try to reduce all technical sources of variability
- The best way to ensure reproducibility and accuracy of results is to include **independent biological replicates**
- Proper use of multiplexing
- **Balanced Block Designs** are better than, their unblocked counterparts in term of power and type I error and are far better when batch and/or lane effects are present

Thank You