Clinical and Translational Science Center

Bulk RNA Sequencing in Clinical and Translational Research

Blythe Durbin-Johnson, Ph.D.
Principal Statistician, Division of Biostatistics

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences (award TR001860).
Outline

- What is RNA sequencing and what can it tell you?
- Designing an RNASeq experiment
- Analyzing your data
- Resources
RNA Sequencing
RNA Sequencing

- RNA sequencing measures gene expression by directly sequencing reverse-transcribed mRNA

- Bulk RNASeq can tell you:
  - Gene expression
  - Transcript expression (with appropriate protocol)

- Limitations
  - Expression measured will be average of that in whole sample
  - Heterogeneity of cell types can add noise
  - Quantification of gene expression is relative
RNA Sequencing

- Steps in RNA Sequencing (short read)
  1. RNA extraction
  2. Enrichment for mRNA
  3. Fragment RNA
  4. Reverse transcribe, add adapters
  5. PCR amplification
  6. Sequence
RNA Sequencing

RNA after rRNA depletion
RNA fragmentation

Illumina RNA ligation method
3’ adapter ligation
5’ adapter ligation
RT
PCR

dUTP method
Random primer RT
dUTP incorporation
Y-shape adapter ligation
dUTP strand degradation
PCR

RT method
Tagged random primer RT
Tagged random primer for 2nd strand synthesis
PCR
Experimental Design
Experimental Design

- Principles of good experimental design in non-omics experiments still apply

- Generally, a minimum of 3 samples/group is required to be able to even conduct any statistical analysis
  - Continuous covariates and very large multifactorial designs are exceptions
  - Adequate power will often require more samples
Experimental Design

- RNA extraction batch tends to be the largest source of technical variability
- If possible, have same person do all RNA extraction in a single session
- Otherwise, **RANDOMIZE** RNA extraction batches so that they are independent of variables of interest, as much as possible
  - Can adjust for RNA extraction batch in statistical model after the fact, if not completely confounded with variables of interest
Experimental Design

▪ NEVER, EVER DO THIS:
  – Day 1: Extract RNA for all treated samples
  – Day 2: Extract RNA for all control samples

▪ Your experiment will be unable to distinguish batch and treatment effects

▪ This is UNFIXABLE with statistics
Experimental Design

- Sample pooling is sometimes necessary to have enough RNA
- In this case, your unit of replication is pool rather than sample
- Still need replicates
  - Don’t pool all of your control samples into one pool and all of your treated samples into another pool
- Each pool should consist of distinct samples
  - Don’t e.g. pool all of your control samples together then split into groups for library prep
Experimental Design

- If you are using more than one lane of sequencing, lane-to-lane variability in sequencing can be mitigated by doing the following:
  - Prepare barcoded libraries—allows samples to be distinguished
  - Pool all libraries
  - Split pool across all lanes being used
- DNA Tech Core at the Genome Center does this
- Make sure your sequencing provider is doing something similar
- If you have more samples than unique barcodes, will need to randomize samples into lanes
Experimental Design

- Power and required sample size depends on:

1. Amount of variability
   - Human >> mouse >> cell line

2. Size of effect to be detected
   - Subtle effects require more samples than large effects

3. Analysis used
Experimental Design

- Resampling of **pilot data** gives best estimates of power
- Other approaches can be unreliable
  - Different methods give very different estimates
  - Too many unverifiable assumptions


- Experience shows, however:
  - 3 replicates/group is typically adequate for cell line studies
  - Human studies require 1 or 2 orders of magnitude above that
Experimental Design

- Matching groups by sex, age, comorbidities, smoking status, etc. reduces bias
- Covariate adjustment in analysis can reduce variability
- However, isolated subjects that are very different from the group can’t be accounted for in analysis:
  - E.g. only male or only smoker in group of women/nonsmokers
  - Try to avoid this situation
  - Throwing a subject from an unreplicated group in “just to see” is a waste of money
- Using subjects as their own control (e.g. pre-post design) can increase power
Analyzing your data
Preprocessing

- You will get fastq files from your sequencing provider
  - Contains sequences for every read + quality scores
  - Side note: sequencing providers only keep your data for a limited time, plan to download data to your storage ASAP or risk losing it!
Preprocessing

- Raw sequence data need to be preprocessed into a form that can be analyzed readily

- Requires access to a compute cluster + knowledge of linux command line

- Or hire someone to do preprocessing
  - Statisticians, even with ‘omics experience, will generally not do this for you
  - Need to plan specifically for bioinformatics support
Preprocessing

- Data preprocessing includes:
  - Removing bases of unwanted sequence (Ex. vectors, adapter, primer sequence, polyA tails)
  - Merge/join short overlapping paired-end reads
  - Remove low quality bases or N characters
  - Remove reads originating from PCR duplication
  - Remove reads that are not of primary interest (contamination)
  - Remove too short reads

- Cleaned sequence data aligned to genome
- Reads belonging to each gene (exon, transcript) counted/quantified
Statistical Analysis

- Some common analyses of RNASeq data:
  - Visual summaries
  - Differential expression analysis
  - Pathway enrichment analyses
  - Weighted Gene Coexpression Network Analysis (WGCNA)
Visualizations

Multidimensional Scaling Plot
- Shows relative distances between whole transcriptomic profile
- Useful for identifying unusual samples
- More distance between groups often means more DE genes in later analysis
Heatmap

- Shows expression level of individual samples for selected genes
- Most useful when limited to focused set of genes
  - Can’t show gene names for >50 genes
- Genes typically clustered based on hierarchical clustering dendrogram
- Samples often clustered as well

Visualizations

Volcano Plot
- Displays differential expression results
- Plot of $-\log_{10} p$-value by log fold change
- Can quickly show if DE genes are predominantly up- or downregulated
- Top genes are labelled
Differential Expression Analysis

- What genes differ in expression between groups?
- Or, what genes are correlated with a continuous outcome?

Steps in DE:
- Filter low expressed/uninteresting genes
- Normalize data to account for library size differences
- Transform/weight data if required by model
- Fit statistical model to each gene
- Adjust p-values for multiple testing
  
  “Significant” means adjusted P < 0.05
Differential Expression Analysis

- Popular Bioconductor packages for DE include:
  - DESeq, based on negative binomial model fitted to gene counts
  - edgeR, based on negative binomial model fitted to gene counts
  - limma-voom, based on weighted linear models fitted to log-transformed counts per million reads

- All of these can accommodate complicated study designs
- limma allows for random effects
- Comparison papers show limma-voom better controls the false discovery rate at the nominal rate

Differential Expression Analysis

- A table of DE results might look like this (~10K rows not shown):

<table>
<thead>
<tr>
<th>Gene.stable.ID</th>
<th>Gene.name</th>
<th>logFC</th>
<th>AveExpr</th>
<th>P.Value</th>
<th>adj.P.Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSMUSG000000103477</td>
<td>5930409G06Rik</td>
<td>1.664192137</td>
<td>1.837293153</td>
<td>0.001950398</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000020721</td>
<td>Helz</td>
<td>-0.363050365</td>
<td>7.501167612</td>
<td>0.003238648</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000026051</td>
<td>Ecrg4</td>
<td>-1.255273365</td>
<td>2.499099269</td>
<td>0.003247058</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000029798</td>
<td>Herc6</td>
<td>-1.445749184</td>
<td>2.050030151</td>
<td>0.003562989</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000038872</td>
<td>Zfhx3</td>
<td>-0.378363542</td>
<td>8.239061179</td>
<td>0.003879801</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000052675</td>
<td>Zfp112</td>
<td>1.11791828</td>
<td>3.113930496</td>
<td>0.004248763</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000037108</td>
<td>Zcwpw1</td>
<td>0.591054065</td>
<td>4.725995765</td>
<td>0.004378957</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000038010</td>
<td>Ccdc138</td>
<td>-0.840359112</td>
<td>4.178710837</td>
<td>0.004715516</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000014905</td>
<td>Dnajb9</td>
<td>-0.501504798</td>
<td>5.243296442</td>
<td>0.0050183</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000022311</td>
<td>Csmd3</td>
<td>-0.877654845</td>
<td>3.42558905</td>
<td>0.005062491</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000090272</td>
<td>Mndal</td>
<td>1.371770654</td>
<td>2.262543233</td>
<td>0.005264431</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000044968</td>
<td>Napepld</td>
<td>-1.666212304</td>
<td>2.06202902</td>
<td>0.005863092</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000025507</td>
<td>Pidd1</td>
<td>1.268521656</td>
<td>2.077240363</td>
<td>0.006642031</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000034912</td>
<td>Mdga2</td>
<td>-0.53890529</td>
<td>5.107036276</td>
<td>0.006768523</td>
<td>0.999161216</td>
</tr>
<tr>
<td>ENSMUSG00000067336</td>
<td>Bmpr2</td>
<td>-0.303627268</td>
<td>7.83187622</td>
<td>0.006771913</td>
<td>0.999161216</td>
</tr>
</tbody>
</table>
Pathway Analysis

- DE analyses can be difficult to interpret

- Pathway or gene ontology enrichment analyses can summarize DE results into a more manageable form

- What pathways/gene sets are overrepresented among significant genes, or at the top of the DE results?

- Common databases:
  - KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways
  - GO (Gene Ontology), a controlled vocabulary for describing gene products
  - Reactome pathways
  - MSigDB (Molecular Signatures Database), used by GSEA
  - Ingenuity Pathway Analysis (requires license for use)
Pathway Analysis

- Enrichment analyses often take one of two approaches:
  1. Is a given gene set overrepresented in my gene list (e.g. significant genes)?
     - Uses Fisher’s Exact Test or hypergeometric test
     - Approach taken by DAVID ([https://david.ncifcrf.gov/helps/tutorial.pdf](https://david.ncifcrf.gov/helps/tutorial.pdf))
  2. Is a given gene set ranked higher in my DE analysis results (or other ranked list) than would be expected by chance
     - Kolmogorov-Smirnov test, GSEA’s leading edge analysis

- Gene ontology is complicated by directed acyclic graph structure of GO terms
  - R package topGO applies either of the above approaches in a way that preferentially tests more specific terms (e.g. “positive regulation of granzyme B production”) over less specific ones (“immune system process”).
Pathway Analysis

- Enrichment analysis results are a descriptive tool, not a smoking gun
- Top KEGG enrichment results using data from 2018 paper:

WGCNA

- Weighted Gene Coexpression Network Analysis (Langfelder and Horvath, 2008) identifies modules of coexpressed genes:

WGCNA

- “Eigengenes” of modules (first principal components of gene expression) provide useful summary
  - ~10 eigengenes vs. 10K genes
  - Easier to calculate/interpret correlations given large quantities of metadata
  - Great for integrating matched data from multiple omics methods

Resources

- Available to anyone on fee-for-service basis:
  - UCD Genome Center DNA Technologies Core
    Library preparation, sequencing, and other services
    https://dnatech.genomecenter.ucdavis.edu/
  - UCD Genome Center Bioinformatics Core
    Wide range of analysis services including everything mentioned in this talk
    https://bioinformatics.ucdavis.edu/

- For IDDRC projects:
  - IDDRC BBRD Core
    Statistical analysis including gene expression data

- Cancer Center Genomics Shared Resource:
  https://health.ucdavis.edu/cancer/research/sharedresources/ger.html
Acknowledgements

- UC Davis CTSC
- MIND Institute IDDRC
  P50 HD103526