Quantitative RNA Sequencing (RNA-seq) and Exome Analysis

Richard A. Radcliffe, Ph.D.
Professor of Pharmacology
School of Pharmacy, Department of Pharmaceutical Sciences
Room V20-3124
(303) 724-3362
richard.radcliffe@ucdenver.edu

Why RNA-seq?

Genetic architecture
Developmental stage
Environmental influences
Tissue type
Disease state

DNA
RNA
RNA-seq
PROTEIN
Phenotype

Why RNA-seq?

“Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease.”

- Catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs
- Determine the transcriptional structure of genes, in terms of their start sites, 5’ and 3’ ends, splicing patterns and other post-transcriptional modifications
- Quantify the changing expression levels of each transcript during development and under different conditions.
- Pathway/network/ontology analysis.

Massively parallel expression analysis

RNA-seq Overview

Adapted from: Pepke et al. (2009) Nat Methods 6:S22-S32
Library Prep

1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
5) 3' ends adenylated and 5' ends repaired
6) DNA sequencing adapters ligated
7) Ligated fragments PCR amplified

Library Prep: Some Considerations

- **RNA fraction**
  - Many different RNA species
  - Poly(A)
  - Size (<200 nt vs. >200 nt)
- Strandedness
- Read length
- Single- vs. pair-end
- Multiplexing
Library Prep: Some Considerations

- **RNA fraction**
  - Many different RNA species
  - Poly(A)
  - Size (<200 nt vs. >200 nt)

- **Strandedness**
  - Overlapping transcripts
  - Annotation of novel transcripts

- Read length
- Single- vs. pair-end
- Multiplexing
The area of the box represents the genome. The area of large green circle is equivalent to the documented extent of transcription, with the darker green area corresponding to that on both strands. CDSs are protein-coding sequences, and UTRs are 5′- and 3′-untranslated sequences in mRNAs. The dots indicate (and in fact overstate) the proportion of the genome occupied by known snoRNAs and miRNAs.

Richard Radcliffe, 1/26/2015
Strandedness

Which strand (gene) did the fragment come from?  
No question about which strand (gene) the fragment came from.
Library Prep: Some Considerations

- **RNA fraction**
  - Many different RNA species
  - Poly(A)
  - Size (<200 nt vs. >200 nt)
- Strandedness
- **Read length**
- Single- vs. pair-end
- Multiplexing

Read Length

- Read length is related to:
  - Sequencing accuracy: quality declines as a function of the length of a read
  - Mapping accuracy: the longer the read, the more accurately it maps
Library Prep: Some Considerations

- RNA fraction
  - Many different RNA species
  - Poly(A)
  - Size (<200 nt vs. >200 nt)
- Strandedness
- Read length
- **Single- vs. pair-end**
- Multiplexing

Single vs. Paired-end

**PE versus SE Illumina**

- Single end (SE): from each cDNA fragment only one end is read.
- Paired end (PE): the cDNA fragment is read from both ends.

**Single-end sequencing**

**Paired-end sequencing**

Zhernakova et al. (2013) *PLoS Genet* e1003594
Library Prep: Some Considerations

- RNA fraction
  - Many different RNA species
  - Poly(A)
  - Size (<200 nt vs. >200 nt)
- Strandedness
- Read length
- Single- vs. pair-end
- **Multiplexing**

Mapping to the Reference Genome

- Bowtie, BWA
- Computational considerations
Mapping to the Genome: Some Considerations

- **Non-unique reads**
  - Gene families
  - Repeat sequences (simple repeats, transposons)
- **Depth**
  - Probability of representation & limits of detection
  - Transcript isoform quantification
  - Variant calling (SNPs, small indels)
- **Reference genome effects**

Non-unique Reads

![Graph showing the relationship between number of multiple alignment reads allowed and fraction of reads suppressed.](https://example.com/graph.png)
Non-unique Reads: Gene Families

Non-unique Reads: Repeats
Mapping to the Genome: Some Considerations

- Non-unique reads
  - Gene families
  - Repeat sequences (simple, SINEs, LINEs, etc.)

- **Depth**
  - Probability of representation & limits of detection
  - Transcript isoform quantification
  - Variant calling (SNPs, small indels)

- Reference genome effects

Depth: Transcript Quantification
Depth: Variant Calling

Mapping to the Genome: Some Considerations

- Non-unique reads
  - Gene families
  - Repeat sequences (simple, SINEs, LINEs)
- Depth
  - Probability of representation & limits of detection
  - Variant calling (SNPs, small indels)
  - Transcript isoform quantification
- **Reference genome effects**
Reference Genome Effects

Analysis

- QC
- Assembly/Quantification
  - Reads Per Kilobase Exon per Million Mapped Reads (RPKM)
- Differential expression
- Pathway/network functional analysis
- Annotation
  - Novel exons – novel splice junctions – novel genes
Quality Control

- Pre-library construction:
  - RNA quality
- Pre-alignment:
  - Per base quality
  - Per read quality
  - Nucleotide distribution per position
  - GC content
  - Sequence over-representation
- Post-alignment:
  - Mean coverage, 5'-3' and 3'-5'
  - Ribosomal RNA contamination
  - Percent mapped reads

Quality Control: RNA Degradation

![RNA Degradation Graphs]

Quality Control

- Quality per position
- Quality per read
- Nucleotide distribution

Analysis

- QC
- **Assembly/ Quantification**
  - Reads Per Kilobase Exon per Million Mapped Reads (RPKM)
- Differential expression
- Pathway/network functional analysis
- Annotation
  - Novel exons – novel splice junctions – novel genes
Assembly/Quantification: RPKM

\[ RPKM = \frac{C}{LN} \]

Analysis

- QC
- Assembly/Quantification
  - Reads Per Kilobase Exon per Million Mapped Reads (RPKM)
- **Differential expression**
- Pathway/network functional analysis
- Annotation
  - Novel exons – novel splice junctions – novel genes
Differential Expression

Analysis

• QC
• Assembly/Quantification
  – Reads Per Kilobase Exon per Million Mapped Reads (RPKM)
• Differential expression
• **Pathway/network functional analysis**
• Annotation
  – Novel exons – novel splice junctions – novel genes
Pathway/Network Functional Analysis

Weighted Gene Co-expression Network Analysis (WGCNA)

Gene Ontology (GO) Cluster Analysis

HMGP 7620: Advanced Genome Analysis

Darlington et al. (2013) Genes Brain Behav 12:263-274

Analysis

• QC
• Assembly/Quantification
  – Reads Per Kilobase Exon per Million Mapped Reads (RPKM)
• Differential expression
• Pathway/network functional analysis
• Annotation
  – Novel exons – novel splice junctions – novel genes
Annotation

Exome Sequencing

- Why
  - Identification of variants (SNPs, CNVs, small InDels)
  - Linkage/association/pedigree studies
  - Clinical diagnostics
- How
  - Isolate, fragment DNA
  - Build library
  - Exome enrichment
  - Sequence
  - Align to reference genome
  - Variant calling
  - Higher order genetic analysis
Exome Enrichment

Variant Calling

Altman et al. (2012) Hum Genetics 131:1541-1554
Examples of intragenic deletion and duplication detected by WES and confirmed by exome aCGH. Each bar in the graphs (a)–(c) and (e)–(g) represents an exon. (a–c) WES data from a family trio in which the (a) proband has inherited a whole-gene duplication of KRT34 from the (b) father, whereas the (c) mother shows normal copy number at that gene. (e–g) WES data from a family trio in which the (e) proband has inherited a partial-gene heterozygous deletion in the SYCP2L gene from the (g) mother, whereas the (f) father shows normal copy number at those exons. Each dot in panels d and h represents an oligonucleotide probe in the gene of interest on the exome array, with a duplication shown by probes deviating to a positive log2 ratio (marked in red) and a deletion shown by probes deviating to a negative log2 ratio (marked in green). Panels d and h show confirmation of the KRT34 duplication and the SYCP2L deletion, respectively, by exome aCGH. aCGH, array comparative genomic hybridization; WES, whole-exome sequencing.
Variant Calling: CNVs/Indels

Genetic Analysis: Mendelian Inheritance

Assumptions:
- Only consider small indels and SNPs
- Causal variants are coding
- Causal variants alter protein sequence
- Near complete penetrance
Examples of intragenic deletion and duplication detected by WES and confirmed by exome aCGH. Each bar in the graphs (a)–(c) and (e)–(g) represents an exon. (a–c) WES data from a family trio in which the (a) proband has inherited a whole-gene duplication of KRT34 from the (b) father, whereas the (c) mother shows normal copy number at that gene. (e–g) WES data from a family trio in which the (e) proband has inherited a partial-gene heterozygous deletion in the SYCP2L gene from the (g) mother, whereas the (f) father shows normal copy number at those exons. Each dot in panels d and h represents an oligonucleotide probe in the gene of interest on the exome array, with a duplication shown by probes deviating to a positive log2 ratio (marked in red) and a deletion shown by probes deviating to a negative log2 ratio (marked in green). Panels d and h show confirmation of the KRT34 duplication and the SYCP2L deletion, respectively, by exome aCGH. aCGH, array comparative genomic hybridization; WES, whole-exome sequencing.

Radcliffe, Richard, 2/1/2015
A Few References

RNA-seq:

Exome sequencing: