Sequence Analysis: Differential Representation

Martin Morgan, Hervé Pagès, Nishant Gopalakrishnan
Fred Hutchinson Cancer Research Center

9-10 December, 2010

Sequence Analysis

- Work Flows
- Example data
- Differential Representation

Subsequent Analysis

- Lab activity
- References

Work Flows: Differential Representation

Prior to analysis

- Biological experimental design – treatments, replication, etc.
- Sequencing preparation – library preparation, manufacturer protocol, etc.

Analysis

1. Pre-processing (sequencing, alignment, quality assessment)
2. Count, e.g., reads per transcript – ChIP-seq; RNA-seq; novel transcript identification; microbiome; . . .
3. Differential representation
4. Annotation
5. . . .

`http://bioconductor.org/workflows` for common analyses.

Third-party tools

- Primary data generation.
- Aligners
  - Differing in alignment flexibility (e.g., mismatches vs. indels); error models (e.g., SOLiD homopolymers); performance
  - *Bowtie*, *BWA*, *SSAHA2*, ...
- Domain-specific
  - ChIP-seq: *MACS*; ...
  - RNA-seq: *GSNAP*, *TopHat* (alignment); *Cufflinks* (isoform assembly), ...
  - Variants: *samtools*, ...
  - Microbiome: ?
- Comprehensive: *GATK*; *BioPerl*, *Biopython*, *HTSeq*
- SeqAnswers

Bioconductor entry points

- Quality assessment.
- Preliminary read processing, e.g., demultiplexing, remediation
- Specialized alignment, e.g., `matchPDict` in *Biostrings*.
- 'Upstream' domain-specific work flows, e.g., ChIP-seq peak calling (*chipseq*), RNA-seq reads per transcript (*GenomicRanges* / *IRanges* / . . .)
- Statistical analysis of designed experiments, e.g., *edgeR*, *DESeq*
- Specialized analysis, e.g., microbiome sequence processing and ecological analysis (*vegan*, *ape*, . . .)

Example Data

Nagalakshmi et al., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science* 320: 1344–1349.

- Original ‘RNA-seq’ experiment
- Two different primers to generate DNA from poly(A) RNA:

RH	Random hexamer
dT	oligo(dT)

- Biological and technical replicates
- Illumina GAI – relatively small number ($<$5 million / lane) of short (33bp) reads; poor trailing base quality.

Counting Reads

- Retrieve results from SRA, reference sequence from UCSC.
- Align to reference using BWA
- Use *GenomicFeatures* to identify exons
- `IRanges::countOverlaps` to count reads
- See `browseVignette("SeattleIntro2010")`

Bioconductor Solutions

Data

- Matrix (transcript × samples) of counts (caution: no special treatment of overlapping transcripts!)
- Designed experiment – random hexamer vs. oligo(dT)

edgeR [3]

- Negative binomial error model (originally:(over-dispersed Poisson).
- Empirical Bayes to moderate over-dispersion.
- Recently: much more flexible experimental design – negative binomial GLM – `glmFit`

DESeq [1]

- Negative binomial error model
- Variance and mean estimation using local regression.

Issues in Analysis

Normalization

- Between-sample differences in total count
- Within-sample trade-offs in reads per transcript
- Approaches: robust estimates via trimmed or geometric mean counts or quantiles (e.g., 75th) per sample

Dispersion: overcoming poor estimates

- *edgeR*: empirical Bayes, common dispersion.
- *DESeq*: estimate per-gene mean and variance, then robust fit across genes to model mean / variance relationship

Significance

- Exact test (single factor; analogous to Fisher exact test)
- GLM likelihood ratio – comparison of fitted to reduced model

DESeq in a Nutshell

```
> ## counts: matrix of counts
> ## conditions: vector of treatments, corresponding
> ##   to each column of counts
> cds <- newCountDataSet(counts, conditions)
> cds <- estimateSizeFactors(cds)
> cds <- estimateVarianceFunctions(cds)
> ## 'top table' of differentially expressed regions
> res <- nbinomTest(cds, "Condition_1", "Condition_2")
```

Subsequent analysis

- Annotation work flows
- Novel domain-specific approaches, e.g., ChIP-seq motif discovery
- Standard analyses tailored to sequence data, e.g., *goseq*.
- Application of microarray-style analyses.

Lab Activity

- Exploratory assessment of ‘hits per transcript’
- *DESeq* work flow
- Evaluation of results

References

S. Anders and W. Huber.
Differential expression analysis for sequence count data.
Genome Biol, 11:R106, Oct 2010.

U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder.
The transcriptional landscape of the yeast genome defined by RNA sequencing.
Science, 320:1344–1349, Jun 2008.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth.
edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.
Bioinformatics, 26:139–140, Jan 2010.