Exploring short read sequences

Martin Morgan
Fred Hutchinson Cancer Research Institute, Seattle, WA

June 27-July 1, 2011

mtmorgan@fhcrc.org
Topics

RNA-seq
  ▶ Experimental design
  ▶ Quality assessment
  ▶ Counting reads
Microbiome
  ▶ Sequence manipulation
RNAseq example work flow – Malone and Oliver (2011)

Sample
- Purify poly(A)+ RNA with oligo(dT) magnetic beads

Microarray
- cDNA synthesis primed with random hexamers
- Dye-swap, hybridization, florescence, analysis

RNA-seq
- Fragment
- cDNA synthesis primed with random hexamers
- Adapter ligation, size select
RNAseq example work flow – Malone and Oliver (2011)

Sample
- Purify poly(A)+ RNA with oligo(dT) magnetic beads

Microarray
- cDNA synthesis primed with random hexamers
- Dye-swap, hybridization, florescence, analysis

RNA-seq
- Fragment
- cDNA synthesis primed with random hexamers
- Adapter ligation, size select
RNAseq example work flow – Malone and Oliver (2011)

Sample
- Purify poly(A)+ RNA with oligo(dT) magnetic beads

Microarray
- cDNA synthesis primed with random hexamers
- Dye-swap, hybridization, florescence, analysis

RNA-seq
- Fragment
- cDNA synthesis primed with random hexamers
- Adapter ligation, size select
Good data: key issues

- **Experimental design** (Auer and Doerge, 2010)
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty (Bravo and Irizarry, 2010)
Good data: key issues

- **Experimental design** (Auer and Doerge, 2010)
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty (Bravo and Irizarry, 2010)

ROC simulation

- Replication (red vs. blue)
- Randomization and blocking (solid vs. dot)
Good data: key issues

- Experimental design (Auer and Doerge, 2010)
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty (Bravo and Irizarry, 2010)

Cumulative proportion of reads occurring 0, 1, … times
Good data: key issues

- Experimental design (Auer and Doerge, 2010)
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty (Bravo and Irizarry, 2010)

Actual (green) versus uniform \( \phi X174 \) coverage

![Cumulative proportion vs. Copies per read (log_{10})](image)
Good data: key issues

- Experimental design (Auer and Doerge, 2010)
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- **Coverage heterogeneity**
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty (Bravo and Irizarry, 2010)

Read count increases with gene length
Good data: key issues

- Experimental design (Auer and Doerge, 2010)
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- **Sequencing uncertainty** (Bravo and Irizarry, 2010)

Reads, stratified by cycle, supporting a spurious SNP call in $\phi X_{174}$
Quality assessment

Subset of Brooks et al. (2011)

- RNAi and mRNA-seq to identify pasilla-regulated alternative splicing
- Purified polyA, random hexamer primed
- Single- and paired end sequences
- Align to reference genome, and to curated splice junctions

```r
> library(ShortRead)
> ## collate statistics
> fqFiles <- list.files(pattern="*.fastq")
> names(fqFiles) <- sub(".fastq", ",", fqFiles)
> qas <- mapply(qa, fqFiles, names(fqFiles),
+ moreArgs=list(type="fastq"))
> qa <- do.call(rbind, qas)
> ## create report
> rpt <- report(qa)
```
Counting hits: `countGenomicOverlaps`

- **Types of overlaps**
- **Decision tree**
- **Performance:** 10’s of seconds to count 10’s of millions of reads against 20,000 regions
Counting hits: `countGenomicOverlaps`

- **Types of overlaps**
- **Decision tree**
- Performance: 10’s of second to count 10’s of millions of reads against 20,000 regions

```plaintext
type
  ▶ "any", "start", "end", "within"
resolution
  ▶ Reads hit 0 genes → discard
  ▶ Reads hit 1 gene → count
  ▶ Reads hit > 1 gene →
    ▶ "none" → discard
    ▶ "divide" → equal division amongst genes
    ▶ "uniqueDisjoint" →
      ▶ Unique disjoint overlap → count
      ▶ Otherwise discard
```
Counting hits: countGenomicOverlaps

- Types of overlaps
- Decision tree
- **Performance**: 10’s of second to count 10’s of millions of reads against 20,000 regions
Sequence manipulation: microbiome

Sampling

1. Sample bacterial communities of 10's of individuals
2. 454 sequencing of 16S RNA
3. Pre-processing
   - Bar codes
   - Primers
4. Phylogenetic placement
5. ‘Ecological’ analysis

Pre-processing tasks

- De-multiplex – simple pattern matching, subset, narrow (remove bar code)
- Primer removal – partial, redundant primer requires full Smith-Waterman matching
Conclusions

- Well-designed experiments include biological replicates, with blocking of potentially confounding variates.
- Biases are likely pervasive in sequence data; the question under investigation may influence whether biases are important.
- *Bioconductor* includes flexible tools for exploring data.
Bioconductor

Who

- FHCRC: Hervé Pagès, Marc Carlson, Nishant Gopalakrishnan, Valerie Obenchain, Dan Tenenbaum, Chao-Jen Wong
- Robert Gentleman (Genentech), Vince Carey (Harvard / Brigham & Women’s), Rafael Irizzary (Johns Hopkins), Wolfgang Huber (EBI, Hiedelberg)
- A large number of contributors, world-wide

Resources

- [http://bioconductor.org](http://bioconductor.org): installation, packages, work flows, courses, events
- Mailing list: friendly prompt help
- Conference: Morning talks, afternoon workshops, evening social. 28-29 July, Seattle, WA. Developer Day July 27

