Exploring short read sequences

Martin Morgan[1]
Fred Hutchinson Cancer Research Institute, Seattle, WA

June 27-July 1, 2011

[1] mtmorgan@fhcrc.org

Topics

RNA-seq

- Experimental design
- Quality assessment
- Counting reads

Microbiome

- Sequence manipulation

RNAseq example work flow – Malone and Oliver (2011)

Sample

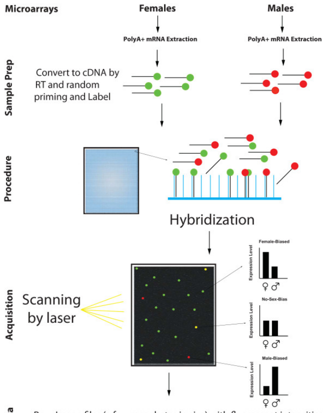
- Purify poly(A)+ RNA with oligo(dT) magnetic beads

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



Raw: Image files (a few megabytes in size) with fluorescent intensities
Processed: Text files with intensities of gene expression

RNAseq example work flow – Malone and Oliver (2011)

Sample

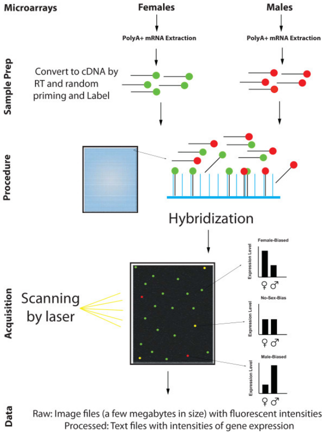
- Purify poly(A)+ RNA with oligo(dT) magnetic beads

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

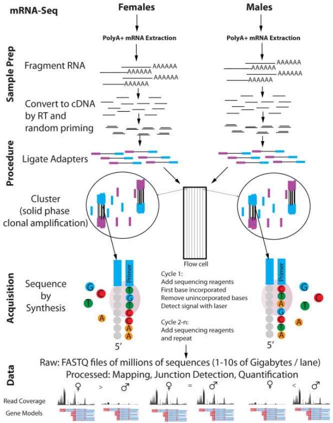
- Purify poly(A)+ RNA with oligo(dT) magnetic beads

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



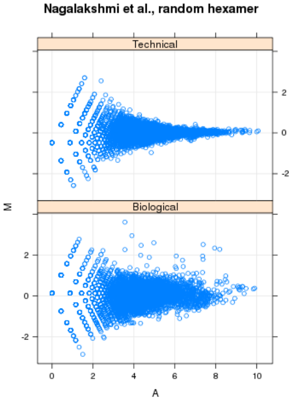
Cycle 1:
Add sequencing reagents
First base incorporated
Remove unincorporated bases
Detect signal with laser

Cycle 2-n:
Add sequencing reagents
and repeat

Raw: FASTQ files of millions of sequences (1-10s of Gigabytes / lane)
Processed: Mapping, Junction Detection, Quantification

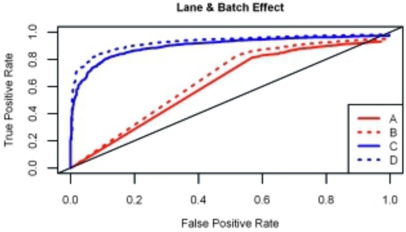
Good data: key issues

- **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)

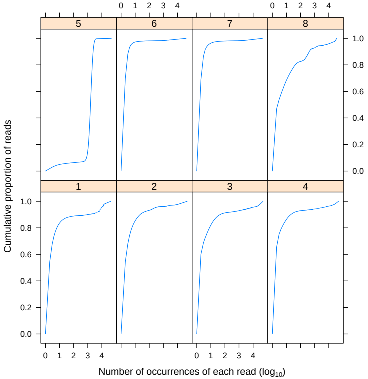


ROC simulation

- Replication (red vs. blue)
- Randomization and blocking (solid vs. dot)

Good data: key issues

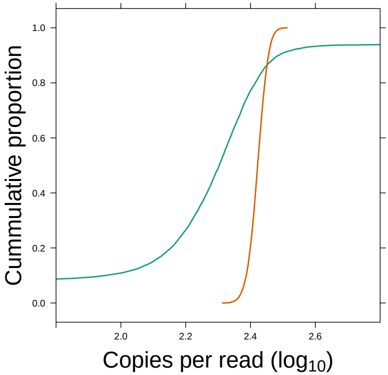
- Experimental design (Auer and Doerge, 2010)
  - Replication
  - Randomization and blocking, e.g., batch effects
- **Depth of coverage**
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty (Bravo and Irizarry,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



Cumulative proportion of reads occuring 0, 1, . . . times

Good data: key issues

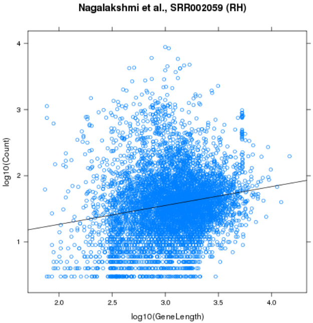
- Experimental design (Auer and Doerge, 2010)
    - Replication
    - Randomization and blocking, e.g., batch effects
- Depth of coverage
    - Statistical power
    - Library complexity
- **Coverage heterogeneity**
    - Estimation biases
    - Legitimate comparison
- Sequencing uncertainty (Bravo and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



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

Good data: key issues

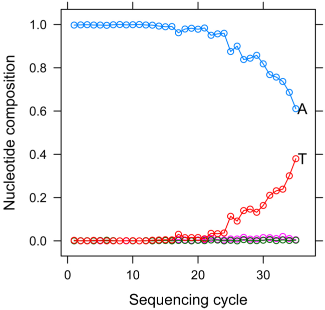
- Experimental design (Auer and Doerge, 2010)
    - Replication
    - Randomization and blocking, e.g., batch effects
- Depth of coverage
    - Statistical power
    - Library complexity
- **Coverage heterogeneity**
    - Estimation biases
    - Legitimate comparison
- Sequencing uncertainty (Bravo and 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 $\phi X174$

Quality assessment

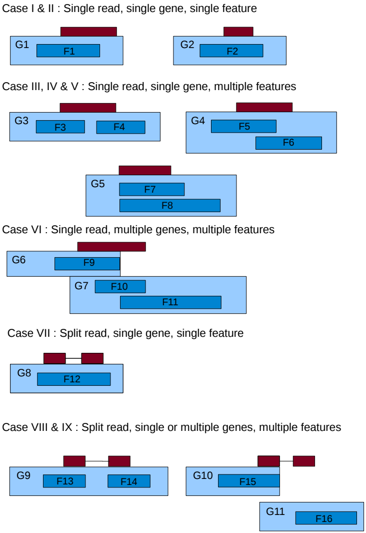
Subset of Brooks et al. (2011)

- 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

```
> 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 second to count 10's of millions of reads against 20,000 regions



Case I & II : Single read, single gene, single feature
G1
F1
G2
F2
Case III, IV & V : Single read, single gene, multiple features
G3
F3
F4
G4
F5
F6
G5
F7
F8
Case VI : Single read, multiple genes, multiple features
G6
F9
G7
F10
F11
Case VII : Split read, single gene, single feature
G8
F12
Case VIII & IX : Split read, single or multiple genes, multiple features
G9
F13
F14
G10
F15
G11
F16

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

`type`

- `"any", "start", "end", "within"`

`resolution`

- Reads hit 0 genes → discard
- Reads hit 1 gene → count
- Reads hit $>$ 1 gene →
  - `"none"` → discard
  - `"divide"` → equal divsion 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 indivdiuals
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: 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

P. L. Auer and R. W. Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185:405–416, Jun 2010.

H. C. Bravo and R. A. Irizarry. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, 66:665–674, Sep 2010.

A. N. Brooks, L. Yang, M. O. Duff, K. D. Hansen, J. W. Park, S. Dudoit, S. E. Brenner, and B. R. Graveley. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.*, 21:193–202, Feb 2011.

J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.*, 9:34, 2011.