Short read quality assessment

Martin Morgan[1]

June 20-23, 2011

[1] mtmorgan@fhcrc.org

Why sequence?

e.g., RNA-seq

- Expression in novel (un-annotated) regions
- Exon junction / RNA editing insights
- Allele-specific / transcript isoform quantification
- Non-model organisms
- Greater dynamic range and sensitivity?

Lessons from microarrays

- Initially: variability between manufactures, technologies, labs
- MAQC: quality control standards and analysis protocols

Example work flow – [4]

Sample

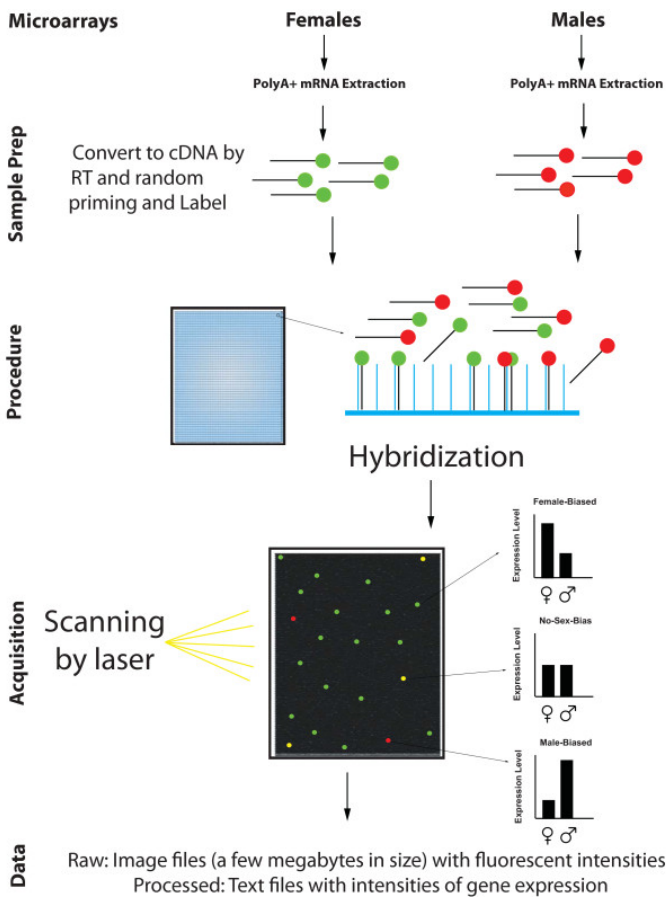
- Purify poly(A)+ RNA with oligo(dT) magnetic beads
- cDNA synthesis primed with random hexamers

Microarray

- Dye-swap, hybridization, florescence, analysis

RNA-seq

- Fragment and size-select
- Illumina adapter ligation



Example work flow – [4]

Sample

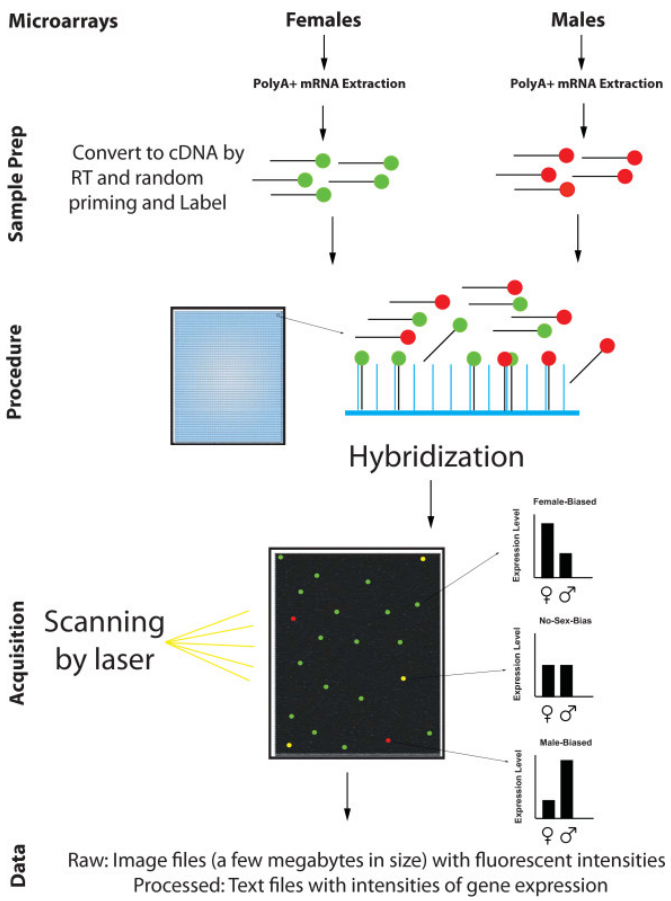
- Purify poly(A)+ RNA with oligo(dT) magnetic beads
- cDNA synthesis primed with random hexamers

Microarray

- Dye-swap, hybridization, florescence, analysis

RNA-seq

- Fragment and size-select
- Illumina adapter ligation



Example work flow – [4]

Sample

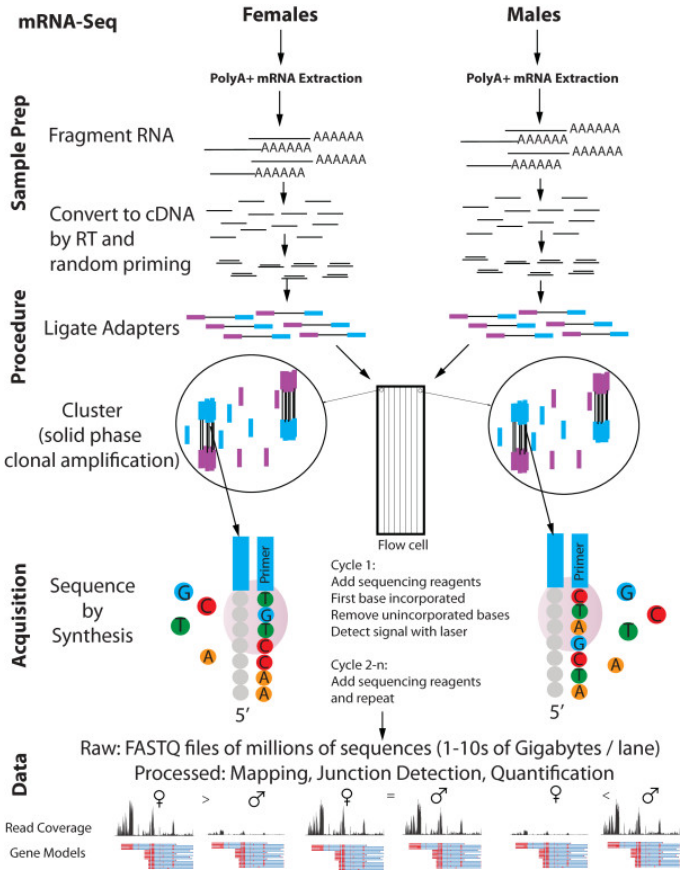
- Purify poly(A)+ RNA with oligo(dT) magnetic beads
- cDNA synthesis primed with random hexamers

Microarray

- Dye-swap, hybridization, florescence, analysis

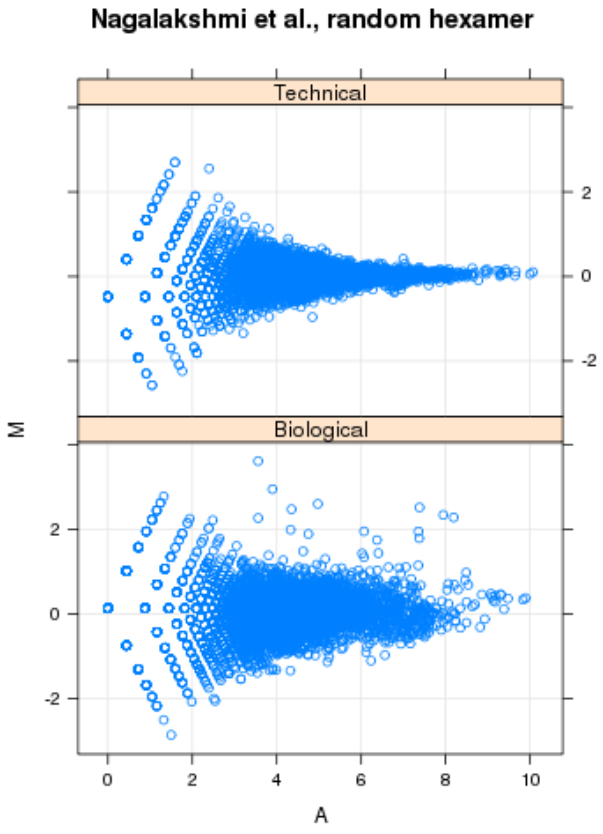
RNA-seq

- Fragment and size-select
- Illumina adapter ligation



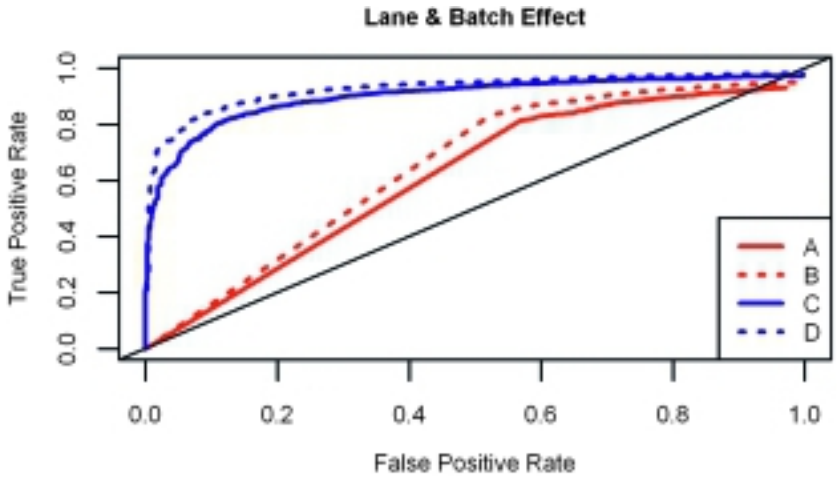
Key issues

- **Experimental design** [1]
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty [2]



Key issues

- **Experimental design** [1]
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty [2]

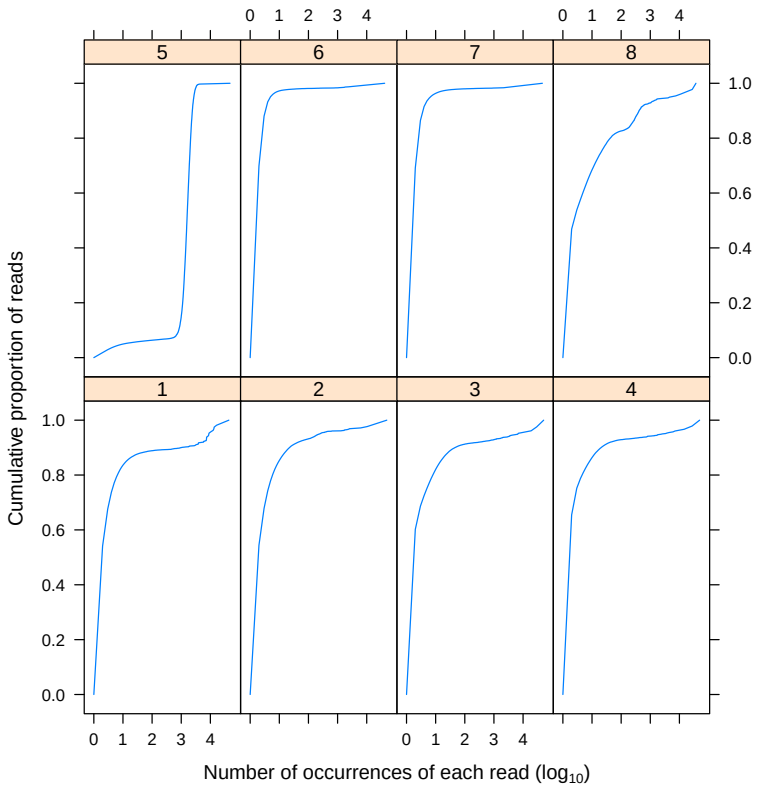


ROC simulation

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

Key issues

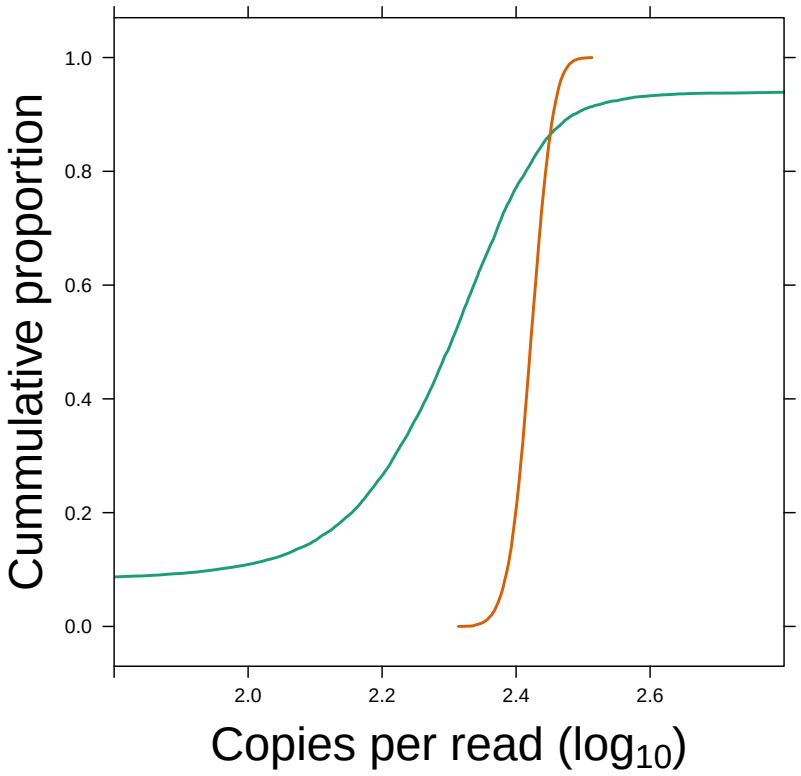
- Experimental design [1]
  - Replication
  - Randomization and blocking, e.g., batch effects
- **Depth of coverage**
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty [2]



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

Key issues

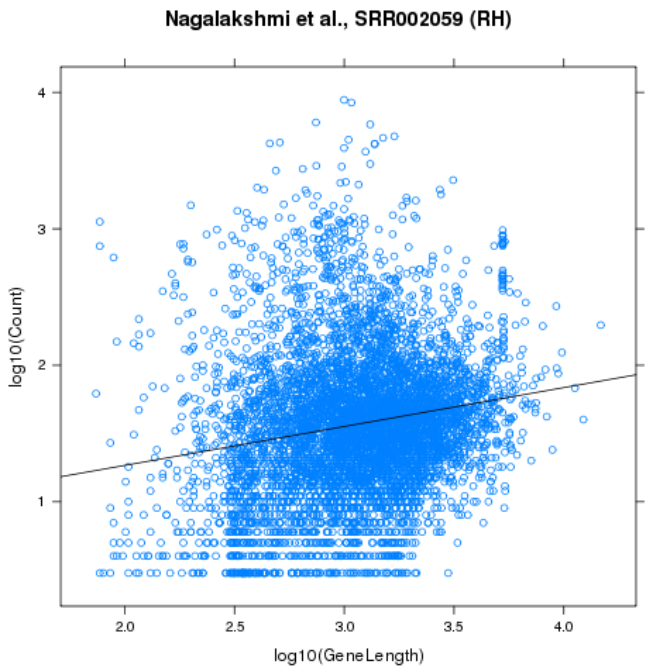
- Experimental design [1]
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- **Coverage heterogeneity**
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty [2]



Actual versus uniform $\phi X174$ coverage

Key issues

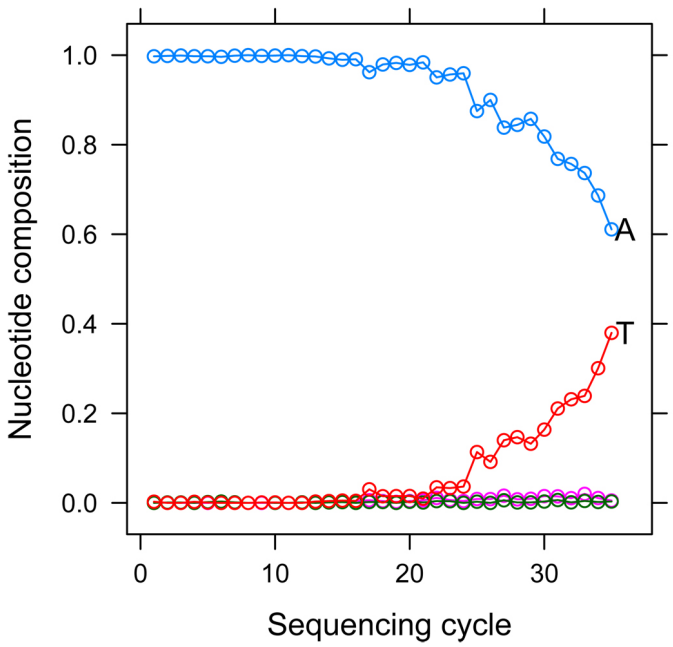
- Experimental design [1]
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- **Coverage heterogeneity**
  - Estimation biases
  - Legitimate comparison
- Sequencing uncertainty [2]



Read count increases with gene length

Key issues

- Experimental design [1]
  - Replication
  - Randomization and blocking, e.g., batch effects
- Depth of coverage
  - Statistical power
  - Library complexity
- Coverage heterogeneity
  - Estimation biases
  - Legitimate comparison
- **Sequencing uncertainty** [2]



Reads, stratified by cycle, supporting a spurious SNP call in $\phi X174$

Case study

Subset of Brooks et al. [3]

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

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.