Short read quality assessment

Martin Morgan\textsuperscript{1}

June 20-23, 2011
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 occurring 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

