Experimental design

Goal

- To answer your research question, given logistical constraints.
- You can't do it all!

Types of questions

- What does the transcriptome look like?
- Which genes are on/off?
- What allelic variants are present?
- How much is each transcript expressed?
- How do expression levels vary?
- What are the most differentially expressed genes?
- How much alternative splicing is there?
- … etc.

Types of questions

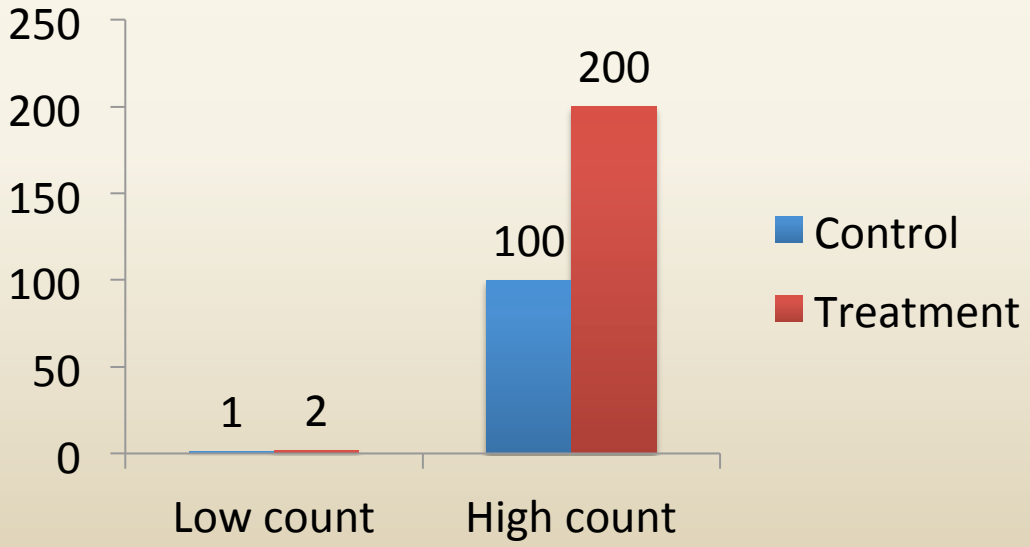
- What does the transcriptome look like?
- Which genes are on/off?
- What allelic variants are present?
- How much is each transcript expressed?
- How do expression levels vary?
- What are the most differentially expressed genes?
- How much alternative splicing is there?
- … etc.

Differential expression analysis

- Statistical power
  - The ability to distinguish differential expression due to treatment effect from background noise

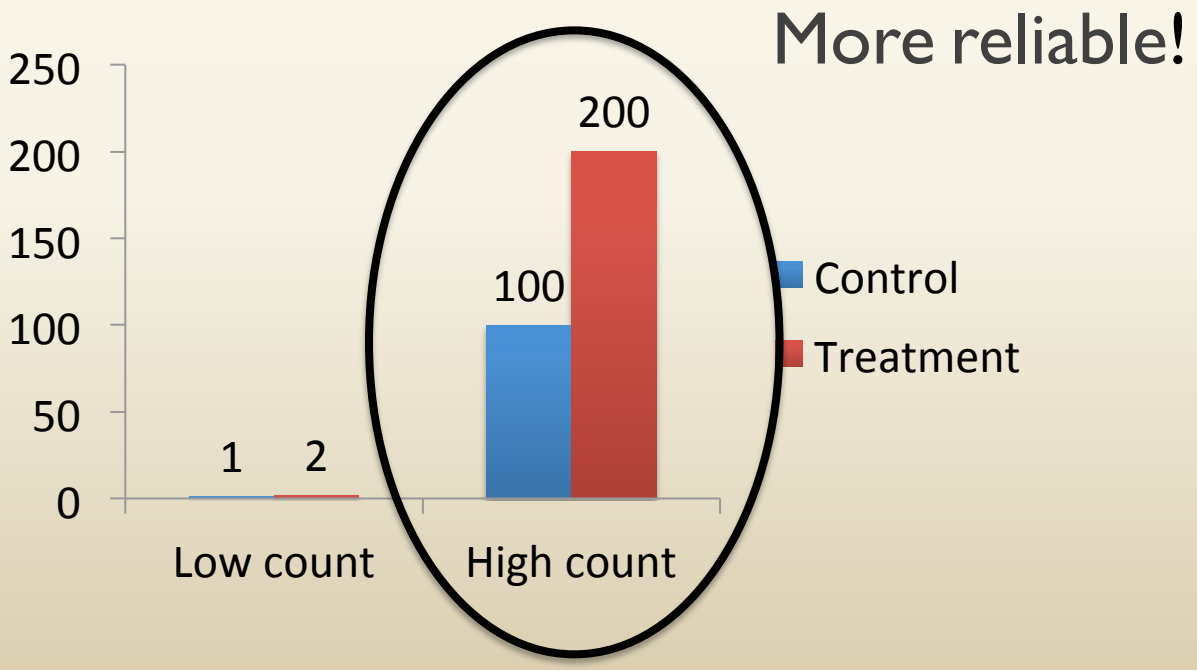
Sources of variation

1) Poisson counting error
   - Uncertainty in count-based measurements
   - Disproportionately large for low-count data



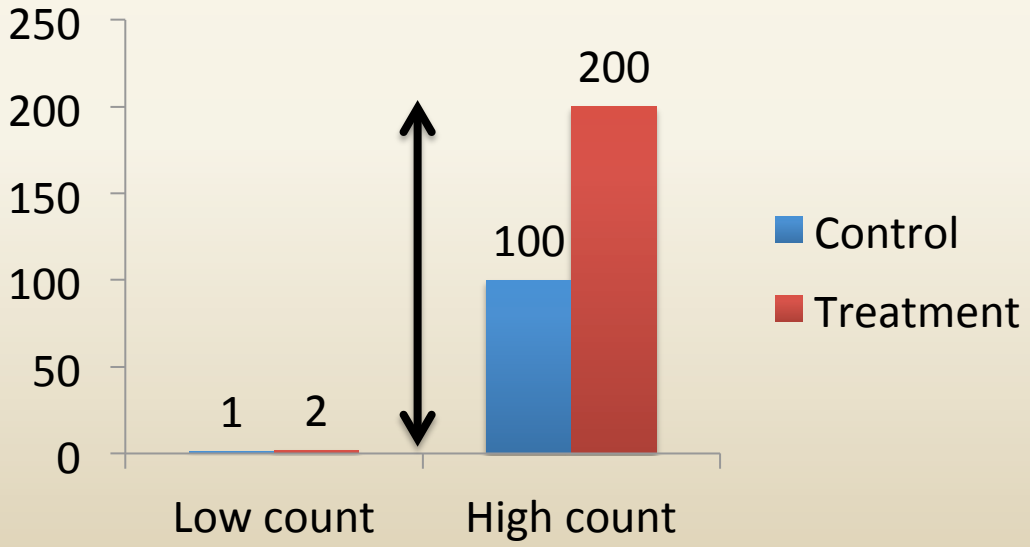
Sources of variation

1) Poisson counting error
   - Uncertainty in count-based measurements
   - Disproportionately large for low-count data



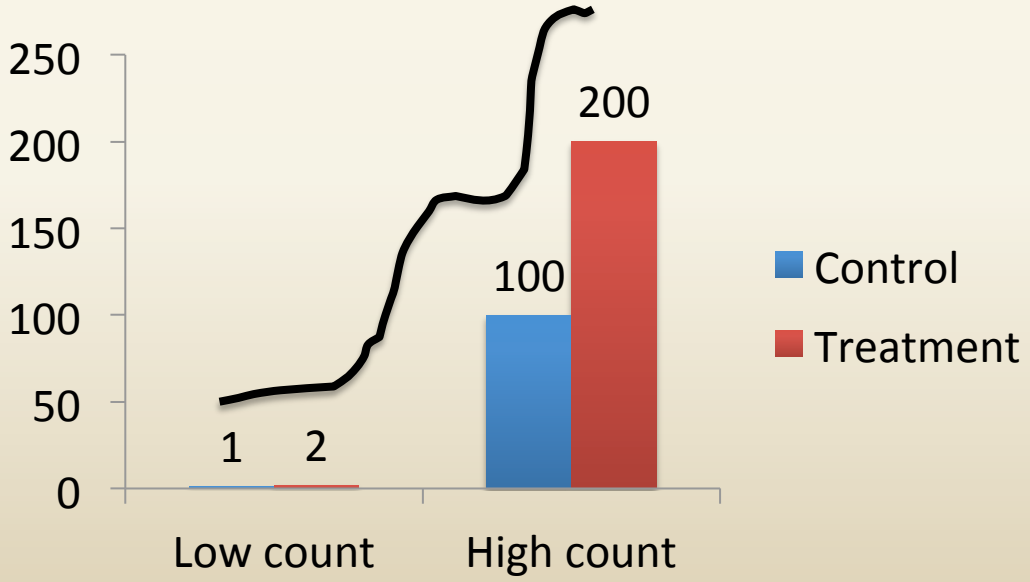
Sources of variation

- Dynamic range = spread between highest and lowest counts in a dataset



Sources of variation

- Expression landscape = magnitude and proportion of expression differences between samples



Sources of variation

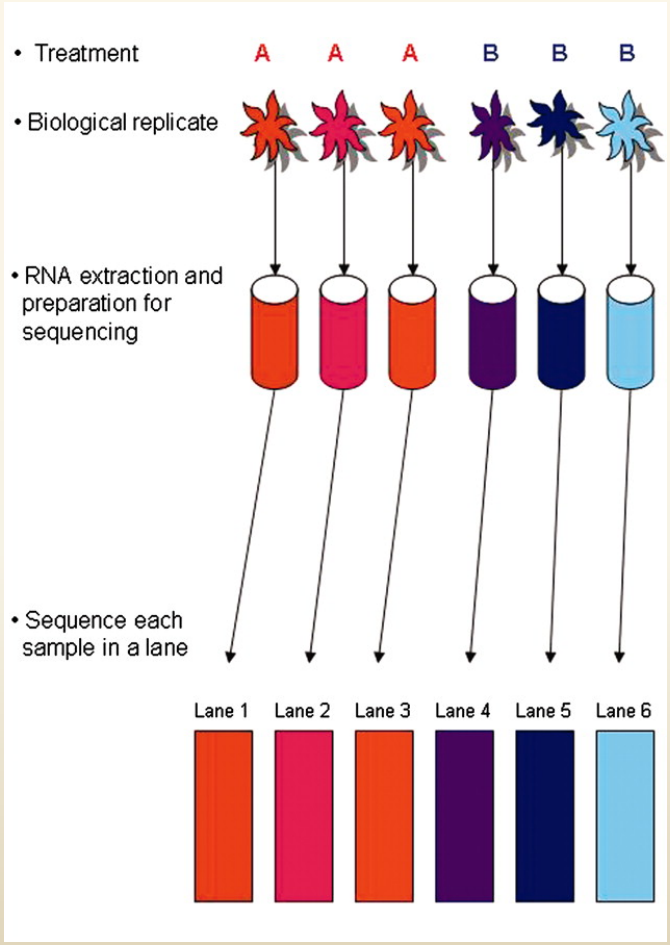
2) Technical variance
   - Imprecision observed between repeated measurements of the same sample

Sources of variation

2) Technical variance
   - Imprecision observed between repeated measurements of the same sample
   - Multiple sources:
     - Random sampling noise (e.g. <0.01% RNA sequenced)
     - Sample collection, storage, and processing
     - Library preparation: PCR biases, sample handling
     - Sequencing: flow cell and lane biases

Lane bias

Confounded design



Auer & Doerge 2010

- Systematic variations between sequencing lanes
- Actually lane bias at every step that occurs on plates (e.g. RNA isolation, library prep)

Sources of variation

3) Biological variance
   - Natural variation observed among samples due to environmental or genetic differences
   - Usually the greatest source of within-group variance
     - Lower for cell-lines and inbred animal strains (BCV≤0.2)
     - Higher for wild populations (BCV>0.3)

*BCV = Biological Coefficient of Variation

Sources of variation

- Tend to increase with sample size, transcriptome size and complexity
- How can we control for it?

Controlling for variation

1) Randomization
2) Blocking
3) Replication

...To reduce confounding sources of variation and more accurately estimate variation that is not of interest (i.e. error)

Ronald Fisher
"The Design of Experiments"
(1935)

Randomization

- Randomize treatments during sample collection, storage, handling, and processing whenever possible

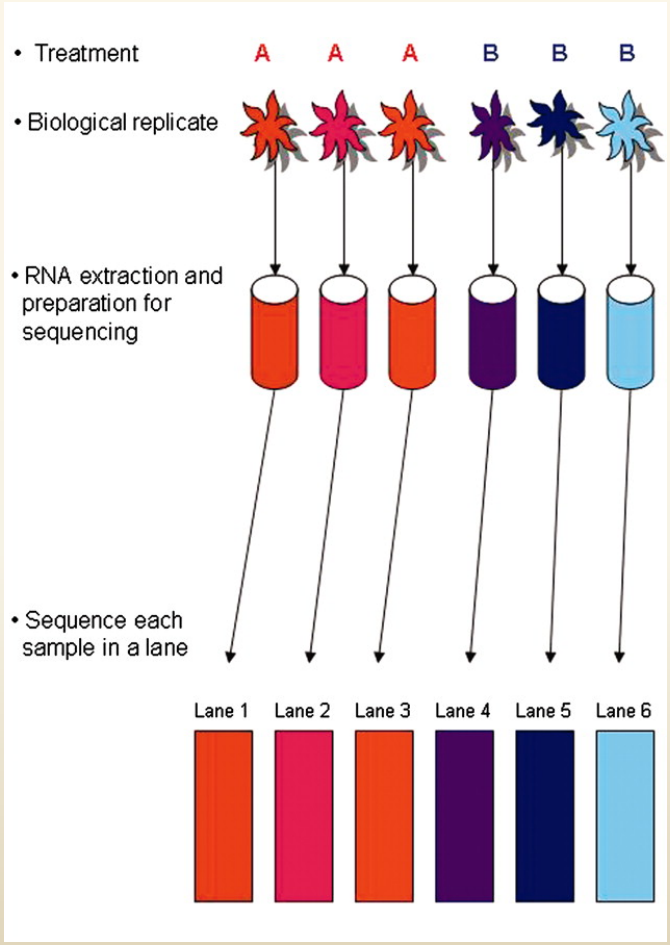
Blocking

- When every level of the factor of interest occurs the same number of times with the "nuisance" factor

- Example:
  - Treatment = of interest
  - Sequencing lane = nuisance

Blocking

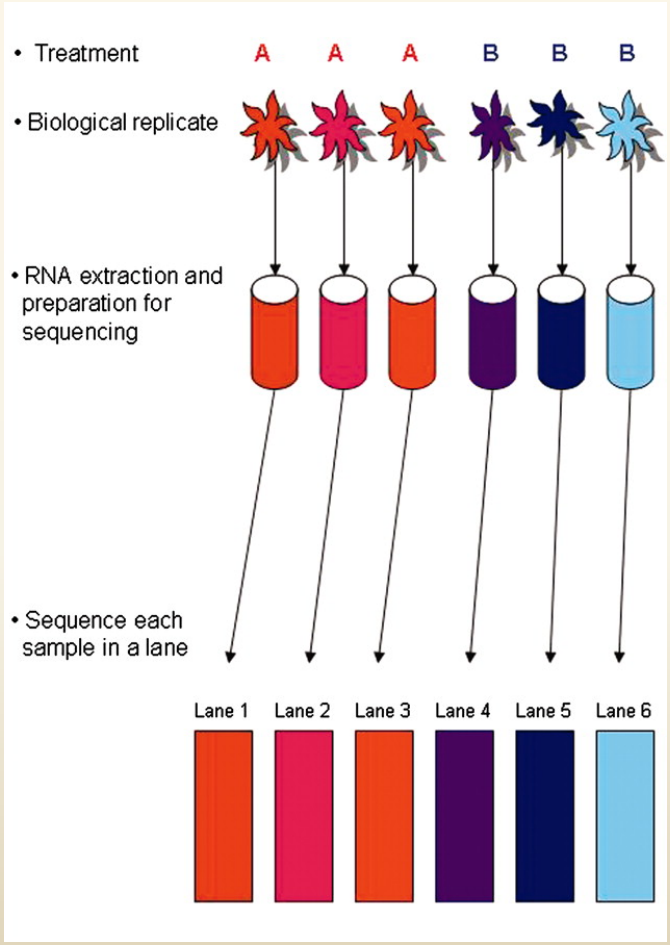
Confounded design



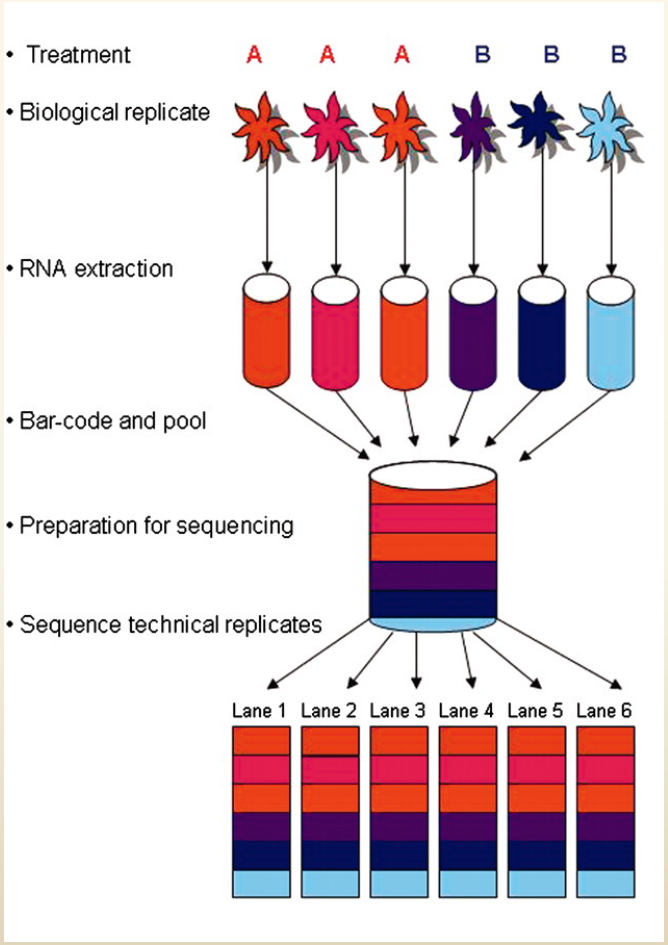
Auer & Doerge 2010

Blocking

Confounded design



Balanced blocked design

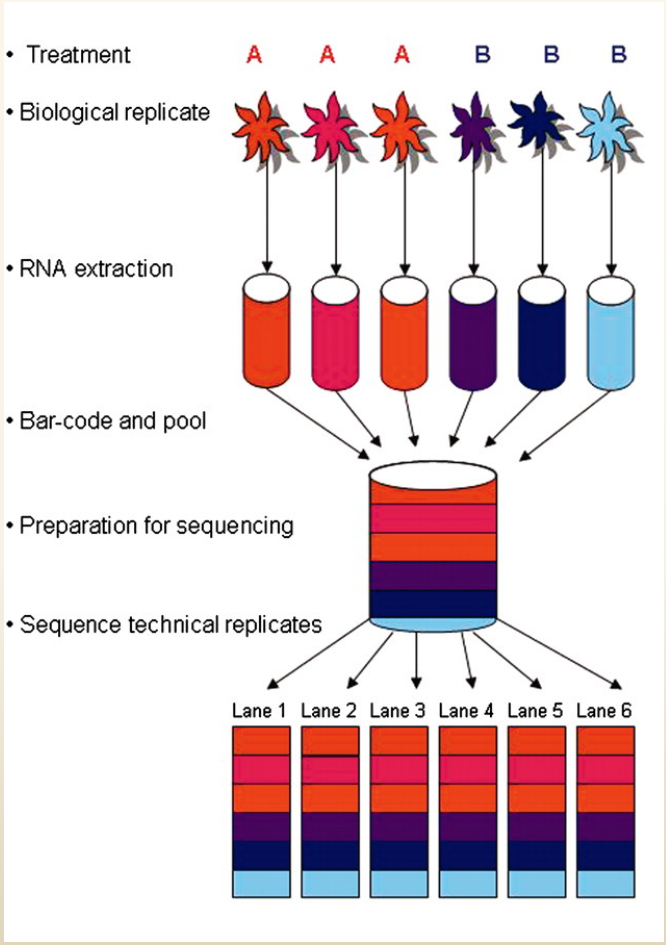


Auer & Doerge 2010

Replication

- Technical replicates
  - No longer necessary for standard experiments
  - RNA-seq is highly replicable
- Biological replicates
  - The only way to quantify biological variation
  - Improves estimates of all sources of variance

Balanced blocked design

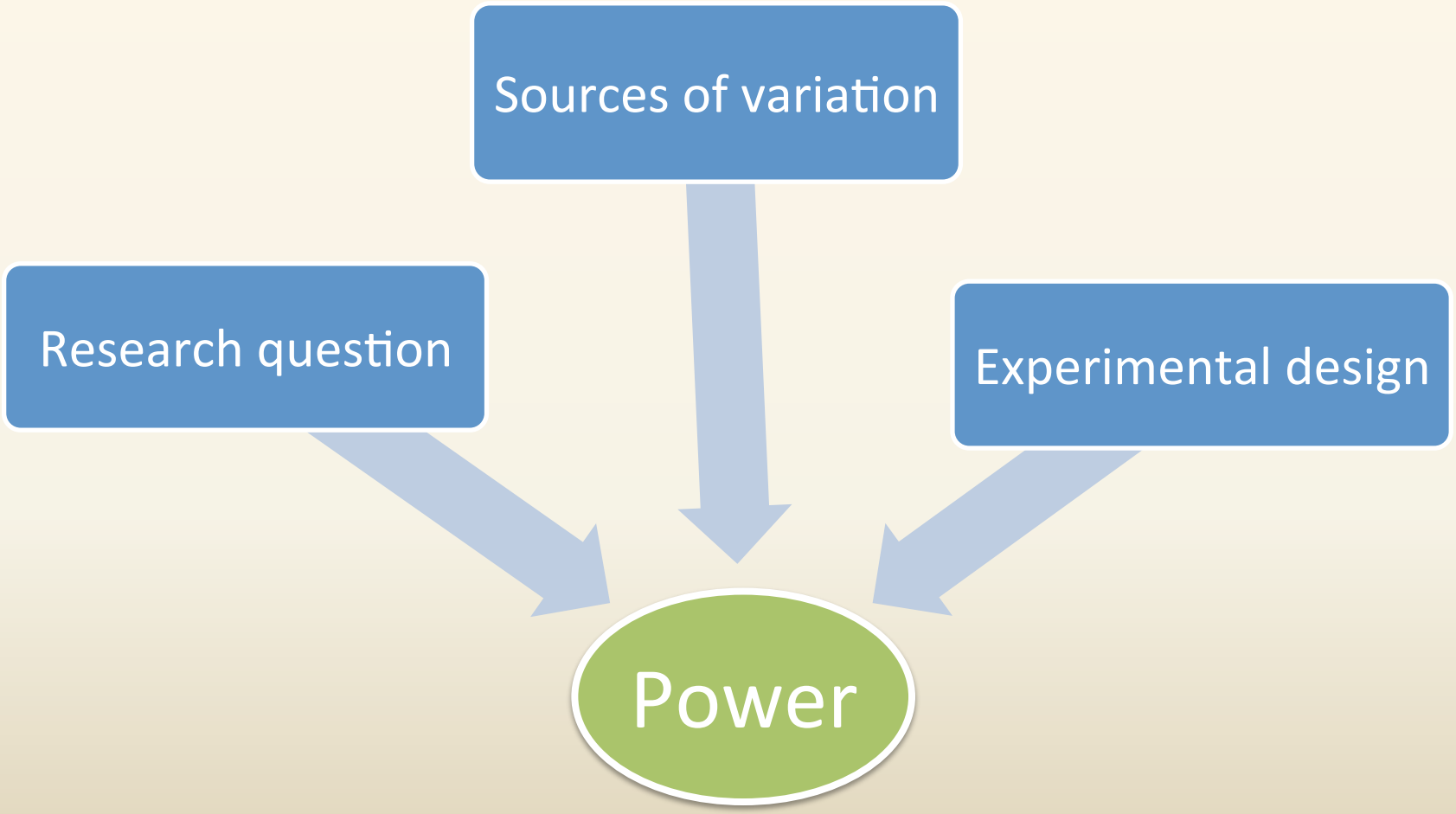


Auer & Doerge 2010

+ Sequencing depth

- Reduces Poisson noise and random sampling error
- Improves detection for transcripts that are lowly expressed, have low fold changes, or higher variance
- BUT:
  - Advantages plateau at an average of ~10 mapped reads per transcript
  - 5-20 million mapped reads generally sufficient

Power of DE detection



Other factors

Time

Cost

Manpower

Tools

Less time/manpower

- Model organism
- High quality RNA
- Plain treatment/control setup or simple time-series
- Dedicated person with support system

More time/manpower

- Non-model organism
- Low quality RNA
- Complex multifactorial designs
- No single dedicated person/lack of support

Toolkit

- Tools
  - Model organisms
    - Genome information
    - Several pipelines
    - Lots of tools for cool visualizations
  - Non-model organisms
    - No genome (or have to make your own)
    - Few tools designed for this
- Computing power
  - One strong computer (model organism)
  - Access to a cluster

Cost

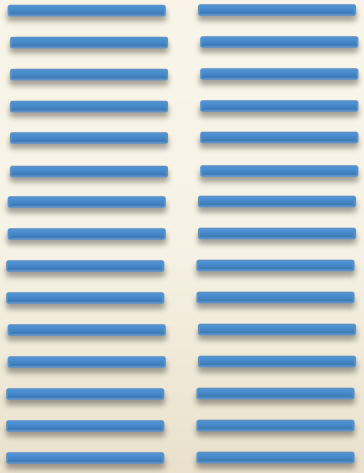
- Affects choice of:
  - Biological replicates
  - Technical replicates
  - Sequencing depth
  - Sequencing technology
  - Computing resources
  - Manpower
  - … and more!

Cost

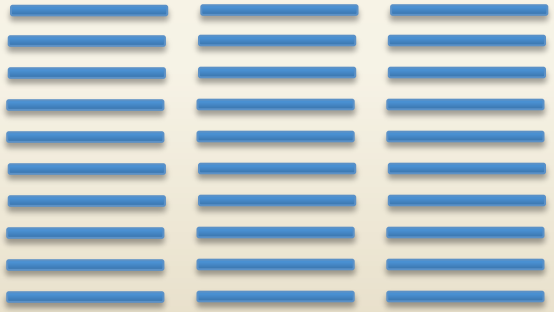
- Tradeoff between sequencing depth and replication
  - More power comes from biological replication!

Cost

- Tradeoff between sequencing depth and replication
  - More power comes from biological replication!



15 M reads x 2 reps

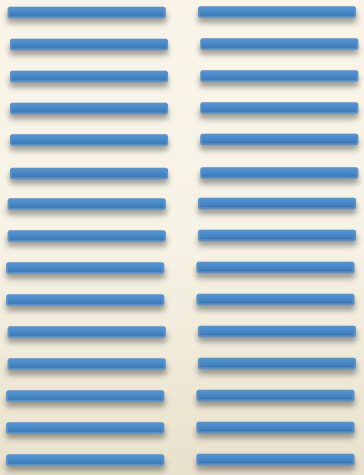


10 M reads x 3 reps

Liu et al. 2014

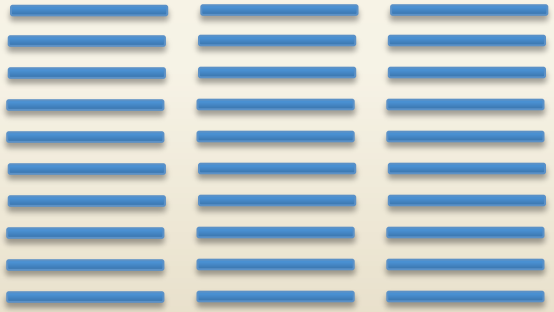
Cost

- Tradeoff between sequencing depth and replication
  - More power comes from biological replication!



15 M reads x 2 reps

35% higher DE detection



10 M reads x 3 reps

Liu et al. 2014

All aspects are connected

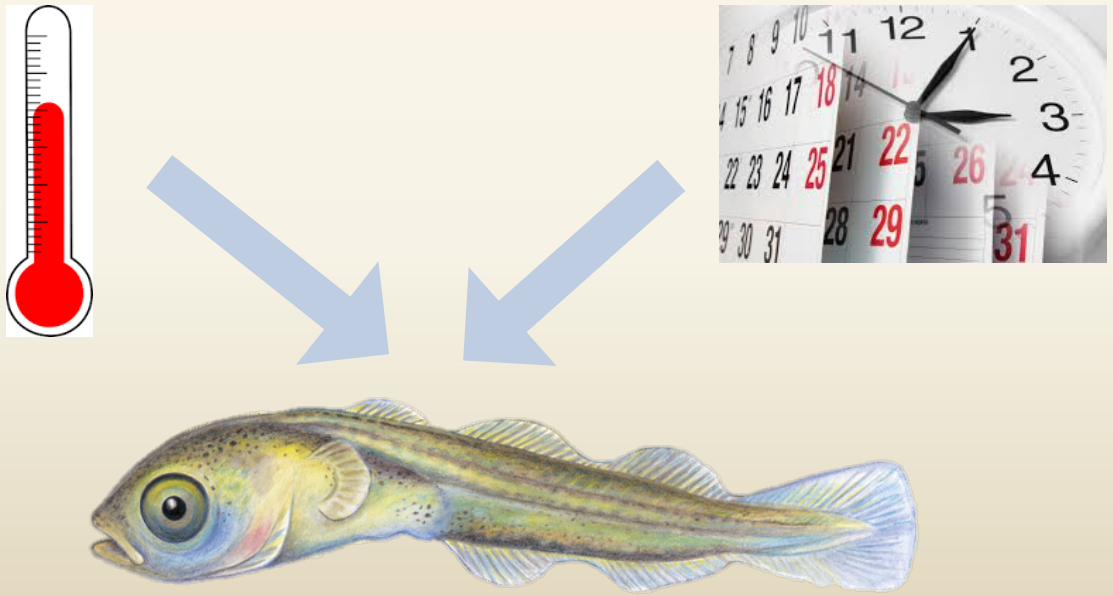
- Sample preparation must reflect experimental design!
- Otherwise this will be your outcome:

Exercise: Experimental Design

Exercise: Experimental Design

Scenario 1 – Exploratory, unlimited

- Question: How do cod respond to temperature during early development?



Exercise: Experimental Design

Scenario 1 – Exploratory, unlimited

- Question: How do cod respond to temperature during early development?
- Design your experiment:
  - Treatments?
  - # replicates?
  - Sequencing depth?
  - Analysis pipeline?
  - Type I error rate?

Exercise: Experimental Design

Scenario 1 – Exploratory, unlimited

- Question: How do cod respond to temperature during early development?
- Illumina Hiseq 4000
  - Cost per lane = 22,000 NOK
  - Reads per lane = 280 million
- How much would your experiment cost?

Exercise: Experimental Design

Scenario 2 – Exploratory, limited

- Question: How do cod respond to temperature during early development?
- Budget = 66,000 NOK

Exercise: Experimental Design

Scenario 2 – Exploratory, limited

- Question: How do cod respond to temperature during early development?
- Budget = 66,000 NOK
- Design your experiment:
  - Treatments?
  - # replicates?
  - Sequencing depth?

Technical biases

- Make a plate map

Exercise: Experimental Design

Scenario 3 - Aquaculture

- Question: What genes are the most differentially expressed in response to temperature in cod?
- How might you change your design?

Exercise: Experimental Design

Scenario 4 - Fisheries management in response to climate change

- Question: What genes are involved in temperature adaptation in cod?
- How might you change your design?

Exercise: Experimental Design

Scenario 5 – Model species

- Question: What genes are associated with temperature adaptation in zebrafish?
- How might you change your design?

Exercise: Scotty

Scotty

- Online tool for calculating power in RNA-seq experiments based on model or pilot datasets

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 5 2013, pages 656–657
doi:10.1093/bioinformatics/btt015

Gene expression Advance Access publication January 12, 2013

Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression

Michele A. Busby, Chip Stewart, Chase A. Miller, Krzysztof R. Grzeda and Gabor T. Marth*
Department of Biology, Boston College, Chestnut Hill, MA 02467, USA

Associate Editor: Ivo Hofacker

Scotty

- Go to: http://scotty.genetics.utah.edu
- Run optimization of:
  1) Model dataset (Human liver – Blekhman)
  2) Nonmodel dataset (Atlantic cod larvae – Oomen)

     ~/data/RNAseq/scotty/Trinity_genes.counts.matrix.cvsd28t13.scotty
- Evaluate and compare results

Rules of thumb

1) Average transcript coverage >10
2) No less than 3 biological replicates
3) Increase # replicates rather than sequencing depth
4) Conduct a pilot sequencing experiment!

MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 1224–1241 doi: 10.1111/mec.13526

INVITED REVIEWS AND SYNTHESES

The power and promise of RNA-seq in ecology and evolution

ERICA V. TODD,* MICHAEL A. BLACK† and NEIL J. GEMMELL*

**Department of Anatomy, University of Otago, PO Box 913, Dunedin 9054, New Zealand, †Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9054, New Zealand*