Experimental Design Principles for Big Data Bioinformatics Analysis

Bruce A Craig
Department of Statistics
Purdue University
My Background

• Professor of Statistics (at Purdue 20+ years)
• Director of Purdue’s Statistical Consulting Service (SCS) since 2004
  – Free service available to all associated with Purdue
  – Balances research and teaching
    • Education and training of student consultants (and clients)
    • Improving the quality and productivity of research
• Had been involved in the design and analysis of spotted microarrays and QTL analysis in polyploids
• Now just try to keep up with the ever-changing technologies within my SCS role
Importance of experimentation

• “There are three principal means of acquiring knowledge... observation of nature, reflection, and experimentation. Observation collects facts; reflection combines them; experimentation verifies the result of that combination.” – Denis Diderot

• “The proper method for inquiring after the properties of things is to deduce them from experiments.” – Isaac Newton
Importance of experimental design

• “You can't fix by analysis what you bungled by design.” - Light, Singer and Willett

• “To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.” - Sir Ronald Aylmer Fisher
Importance of experimental design

• "All experiments are designed experiments, it is just that some are poorly designed and some are well-designed."

• “If your experiment needs statistics, you ought to have done a better experiment.” - Lord Ernest Rutherford

• “No experiment is ever a complete failure. It can always be used as a bad example.” – Paul Dickson
Experimental Design

• Does **NOT** simply involve statistics...

• Requires a combination of

  – Scientific/biologic insight
  – Scientific logic
  – Common sense
  – Planning
  – Statistics
  – Communication skills

• [https://www.youtube.com/watch?v=PbODigCZqL8](https://www.youtube.com/watch?v=PbODigCZqL8)

*Not a one person task!!!
Experimental Design

• Basic design principles do not change with number of outcomes (e.g., genes, proteins), advancement in technology, or size of the data

• “Keep it simple stupid” (KISS) principle often plays even more of a role in these bioinformatics studies due to
  – Costs involved
  – Complexity/size of data obtained
Basic Design Principles

• Randomization
  – “balancing out” effects of lurking/hidden variables

• Replication
  – Improving precision

• Blocking
  – Control the effects of nuisance factors that otherwise increase the noise in the experiment
Experimental Strategies

• One-factor-at-a-time (OFAT)
  – Favored when data cheap and abundant
  – Often very inefficient
  – Cannot investigate interaction

• Factorial treatment structure
  – Hidden replication advantage
  – Can assess interaction

Both perhaps involving a blocking factor
Factorial Treatment Structure

\[ n = 6 \quad \text{Assume variance is } \sigma^2 \]

Effect of A = \( \overline{A2B1} - \overline{A1B1} \)

\[ \text{Var(Effect of A)} = \frac{2\sigma^2}{2} = \sigma^2 \]

Effect of A = \( \overline{A2} - \overline{A1} \)

\[ \text{Var(Effect of A)} = \frac{2\sigma^2}{2} = \sigma^2 \]
Nuisance Factors

- Imperative to consider all possible factors that could obscure or alter your results
- Often rely on sound judgement of center generating the bioinformatics data (i.e., temp, humidity)
  - Block on day or technician(?)
- Controlling these factors allow results to be as generalizable as possible
  - Sample split in half $\rightarrow$ Trt1 to one half, Trt2 to other
- Randomization provides protection against the unknown....avoid possible biases
Replication

• Biological versus technical replicates
  – Technical replicate: Measuring same biological sample multiple times
  – Technical replicates assess non-biological variation
  – May be beneficial if this variation expected to be large
  – Biological replicates typically improve precision more and allow conclusions to be more generalizable

• Number of replicates depends on numerous factors
  – Cost and availability of resources
  – Desired precision / power
Replication

• Pooling
  – Used when sample material scarce
  – Larger sample $\rightarrow$ better precision of measurement
  – Danger in pooling
    • Measurement obtained from pooled sample can be different from the average of individual measurements
  – Better to avoid pooling if possible
    • More precision from multiple biological replicates
Calculating Power

• Numerous calculators available in software and online

• Be wary...you will always get numbers but whether they’re meaningful depends on the quality of the inputted values
  – Trusting the quality of the inputs requires a basic understanding of the process
Why Power Analysis?

• Research is expensive…wouldn’t want to conduct experiment with far...
  – too few experimental units (EUs)
    • Project won’t find important differences that exist
    • Not worth the time and money
  – too many experimental units (EUs)
    • Project is unnecessarily too expensive
• Typical funding agency requirement
  – Demonstrates thinking about plan and organization
A Simple Experiment

• Study the effect of cold on a fat gene in rat

• Use a Completely Randomized Design (CRD):
  – 6 rats are randomly assigned to one of two different environments
    • Trt 1: Normal environment (20°C)...\(n=3\)
    • Trt 2: Cold environment (5°C)...\(n=3\)

• Investigator expects lower expression of gene when under Trt 2

• Is \(n=3\) per trt enough to detect this difference?
Statistical Analysis

• Two competing hypotheses:
  \[ H_0: \log_2(\mu_1/\mu_2) = \log_2(\mu_1) - \log_2(\mu_2) = 0 \]
  \[ H_1: \log_2(\mu_1) - \log_2(\mu_2) > 0 \quad \text{i.e. one-tailed test (for now)} \]

• Basis for choosing between the two
  – P-value of test relative to a declared significance level (\(\alpha\))...typically \(\alpha = 0.05\)
    • \(P \leq \alpha\) \(\rightarrow\) reject \(H_0\) and conclude mean larger in Trt 1
    • \(P > \alpha\) \(\rightarrow\) fail to reject \(H_0\), not enough evidence to conclude \(H_1\)

• There are two possible incorrect conclusions based on the analysis of the data
Type I and Type II errors

<table>
<thead>
<tr>
<th>True state</th>
<th>What the data indicate:</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₀: is true</td>
<td>Fail to reject H₀: (P &gt; α)</td>
</tr>
<tr>
<td></td>
<td>Reject H₀: (P ≤ α)</td>
</tr>
<tr>
<td>H₁: is true</td>
<td>No error</td>
</tr>
<tr>
<td></td>
<td>Type I error (Prob is α)</td>
</tr>
<tr>
<td></td>
<td>Type II error (Prob = β)</td>
</tr>
<tr>
<td></td>
<td>No error</td>
</tr>
</tbody>
</table>
So is n = 3 rats large enough?

• Rephrase: Do we have enough statistical power?
• Need to “know” several things
  – How large is the true mean difference (\( \delta = \log_2(\mu_1/\mu_2) \))?
    1) What do you anticipate?
    2) What would be scientifically/practically important?
      – Suppose researchers believe that \( \delta = 1 \) (fold change)
  – How much variability (\( \sigma \)) exists between rats within a grp?
    • Some prior information potentially available from previously published studies or small pilot study. Variability due to rat and method
    • May also have to guess
      – Suppose researchers believe that \( \sigma = 0.7 \)
  – Effect size (\( \delta/\sigma \)) often used when scale arbitrary
• Power analysis involves “educated guessing”
One way to elicit values for $\sigma$

• **Use an empirical rule:**
  Consider range of responses to be equal to $4\sigma$ or $6\sigma$

• **Question:** What would be the likely range (max-min) of log expression levels for rats within the same trt?
  – Suppose the answer was 2.8
    • $R = 2.8 \rightarrow \sigma = 0.7$

Can often find similar published studies with estimates of $\sigma$. Always round up to be a little conservative.
Two competing hypotheses:

- Under $H_0$: $\bar{y}_2 - \bar{y}_1 \sim N\left(0, \frac{2\sigma^2}{n}\right)$

- Under $H_1$: $\bar{y}_2 - \bar{y}_1 \sim N\left(\delta, \frac{2\sigma^2}{n}\right)$

• Conduct one-tailed $t$-test for a certain $\alpha$

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{2s^2/n}} > t_\alpha$$

Reject $H_0$: if

Need to study distribution of $t_0$ under $H_0$ and $H_1$
Distributions of $t_0$

Central and non-central $t$ distributions

Yikes!!
Power is 42%

$1 - \beta = \text{Power}$
Power Calculators

• Calculators available in software such as Minitab, SAS, JMP, and R
  – Be wary of calculators (such as PWR in R) that asks just for an effect size
  – Effect size essentially a signal versus noise ratio
  – Noise may be more than just biological variation

• Many calculators also take into account issue of multiple comparisons
Multiple Comparisons

- Testing changes in expression for thousands of features across several treatments

Probability of at least one false positive quickly rises
Multiple Comparisons

• Trade-off between control of false positive and false negative (power) rates

• Two common types of control on false positives
  
  – **Familywise error rate**: $P(\text{at least one}) \leq \alpha$
    
    Bonferroni (compare P-value to $\alpha/M$)
  
  – **False discovery rate**: Expected proportion of rejected hypotheses rejected incorrectly
False Discovery Rate

• Proposed by Benjamini and Hochberg (1995)

<table>
<thead>
<tr>
<th></th>
<th>Not Rejected</th>
<th>Rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Null</td>
<td>U</td>
<td>V</td>
<td>m₀</td>
</tr>
<tr>
<td>False Null</td>
<td>T</td>
<td>S</td>
<td>m-m₀</td>
</tr>
<tr>
<td></td>
<td>m-R</td>
<td>R</td>
<td>M</td>
</tr>
</tbody>
</table>

• FWER controls V
• FDR = E(V/R | R>0)Pr(R>0)
Microarray Myths and Truths
appeared: The Scientist, Vol. 16, No.17, p.22

• Myths:
  • “That complex classification algorithms such as neural networks perform better than simpler methods for class prediction
  
  • That **multiple testing issues can be ignored** without filling the literature with spurious results

  • That **prepackaged analysis tools are a good substitute for collaboration** with statistical scientists in complex problems.”
• Truths

  “The greatest challenge is organizing and training for a more multidisciplinary approach to systems biology. The greatest specific challenge is good practice in design and analysis of microarray-based experiments.

Comparing expression in two RNA samples tells you only about those samples and may relate more to sample handling and assay artifacts than to biology. Robust knowledge requires multiple samples that reflect biological variation.

Biologists need both good analysis tools and good statistical collaborators. Both are in short supply.”