Challenges in Experimental Design and Data Analysis for Modern Biological Data

David M. Rocke
Distinguished Professor of Biostatistics and Biomedical Engineering
University of California, Davis

SSR Workshop on Emerging Technologies/Big Data
San Diego, CA
July 16, 2016

What is Big Data?

- Many observations, well structured
  - Framingham Heart Study, original cohort 5,209 subjects.
  - National Health and Nutrition Examination Survey (NHANES)—around 5,000 subjects surveyed per year.
  - US Census—various surveys and the decennial census, thousands to millions of respondents.
- Many observations, chaotically structure.
  - Twitter posts
  - Facebook posts
  - Mobile phone traffic

What is Big Data?

- Few observations, many variables
  - This is probably the kind of "big data" we mostly encounter in the biomedical sciences
  - Gene expression by arrays or RNA-Seq
  - Large-scale proteomics
  - Metabolomics
  - Many other high-throughput assays.
- This kind of big data is harder to analyze than the many-observations kind.

Design Issues in Omics Studies

- Selection of control group (clinical/population)
- Sample size
- Replicates and pseudo-replicates
- Randomization
- Blind evaluation
- Overfitting
- What is a real change?
- False positives
- Compare gene expression for a single gene across samples?
- Compare gene expression across genes within a sample?

Diagnostic studies

- A possible source of biomarkers is in glycomics
  - Measurement of relative amounts of glycans
  - Glycans re sugars adducted to proteins to create glycoproteins
- If we want glycomic markers of breast cancer in serum, we could take patients referred for treatment for breast cancer to UC Davis Medical Center
- Potential control group is women undergoing routine exams

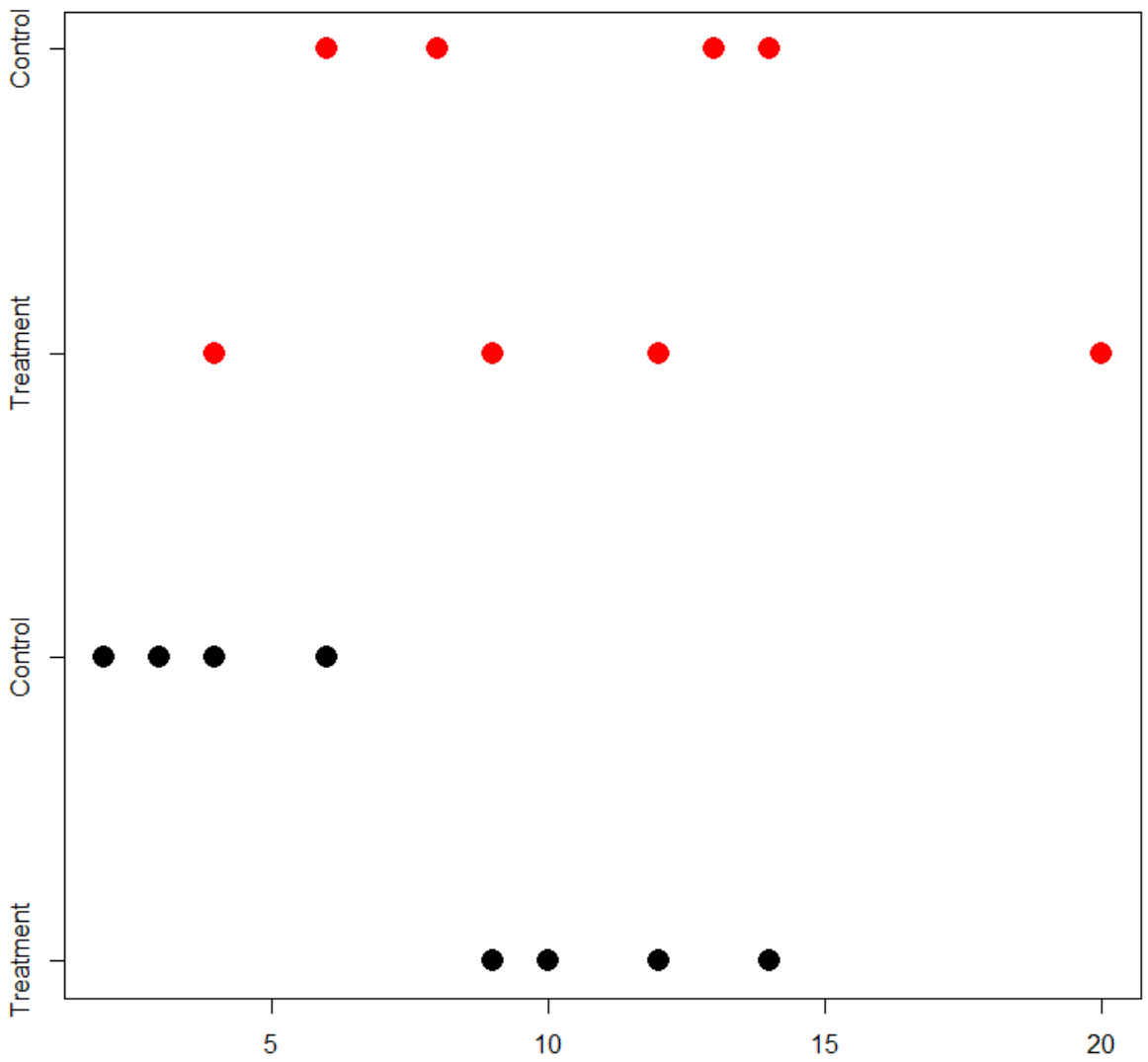
Possible Issues

- We must age match the treatment and control groups. Could glycans change with menopause?
- Women referred for breast cancer are more of a cross-section of the community than women in routine exams (at least in the US).
- Can differences in diet or environment affect glycomic profile?
- How do we make sure that the two groups are essentially identical except for the disease?

Sample Size

- How to compute sample size for an RNA-Seq experiment?
- Divide the experimental budget by the cost per assay?
  - Measuring lots of things does not make it easier to get by with a small sample—it makes it harder!
  - This is because we have to use stricter criteria for significance to avoid too many false positives.
- Technical replicates are mostly useless except for initial studies of errors of analysis.

- What can we tell if we have one treatment sample and one control sample?
- NOTHING!
- If I want to know if expression of certain genes changes in a particular keratinocyte cell line when TNF-α is applied, I need to measure multiple wells/plates of control and multiple wells/plates of treated.
- If the expression of CASP9 in one treatment sample is 12 and on control sample is 6, what do I have?
- NOTHING!
- Treatment {12, 10, 14, 9} control {6, 4, 2, 3} is one case
- Treatment {12, 20, 4, 9} control {6, 14, 8, 13} is another


Control
Treatment
Control
Treatment
5
10
15
20

Fold change

- What is the significance of a fold change of 2?
- It has no biological or statistical meaning.
- It is only meaningful as a comparison of treatment vs control if compared to variation with the groups.
- Mean(2, 4, 3) = 3, mean(6, 7, 5) = 6
  t = 3.6742, p = 0.021
- Mean(1, 5, 3) = 3, mean(6, 8, 4) = 6
  t = 1.8371, p = 0.140
- Fold change apart from statistical analysis is meaningless

One Sample

- Some RNA-Seq software will produce p-values even with one sample under each condition
- What hypothesis is this testing?
- If I want to know if men weigh more than women (on the average) what do I have if I weigh one man and one woman?
- What do I have if I weigh one man 10 times and one woman 10 times?
- (Nothing, in both cases)

Freezer

Cells grown to 50% confluence in 8 dishes

4 treated with TNF-α
4 controls

Actual transcript count for one gene in 8 dishes

RNA-Seq mapped read count for one gene in 8 dishes

Other (biological) variability

Technical variability

Sources of Variability

- Transcript abundance (etc.) in biological samples differs even if the conditions have been the same. This is *biological variability*.
- Library construction and sequencing (or other procedures) adds *technical variability*, so that the relative fragment abundance is different from the relative transcript abundance in the sample.
- Two measurements of aliquots of the same RNA sample differ by technical variability. Replicates differ by the sum of biological and technical variability.

Replicates

- Suppose I have two mice treated with a new drug and two control mice.
- I extract RNA from two blood samples in each mouse, both taken on the same day.
- I run an RNA-Seq analysis in duplicate on each of the eight samples.
- Question: How big is “n”?
- Answer: 2+2.
- Even though I have 16 observations on each gene, we should average both replicates for each mouse, or use a fancy hierarchical regression model.

Replicates

- I have an antibody to caspase 9 with attached fluorophore and examine 30 cells in a well and measure the fluorescence of each cell
- I do this with one well treated with TNF-α and one untreated well.
- Question: What is the sample size?
- Answer: two, one treatment and one control
- I have no evidence of anything.
- You have to average the fluorescent intensity of all the cells in a well and use that as the measurement (or else use a hierarchical model).

Replicates

- I have 6 control mice and 6 treated mice with RNA from each one.
- Question: Which is better,
  - Submitting the 12 samples separately to RNA-Seq or gene expression arrays or
  - Pooling the control samples and pooling the treatment samples so only two analyses are needed.
- Answer: The latter strategy costs 1/6 as much but generates no useful data.
- Pooling is almost never useful.

Overfitting

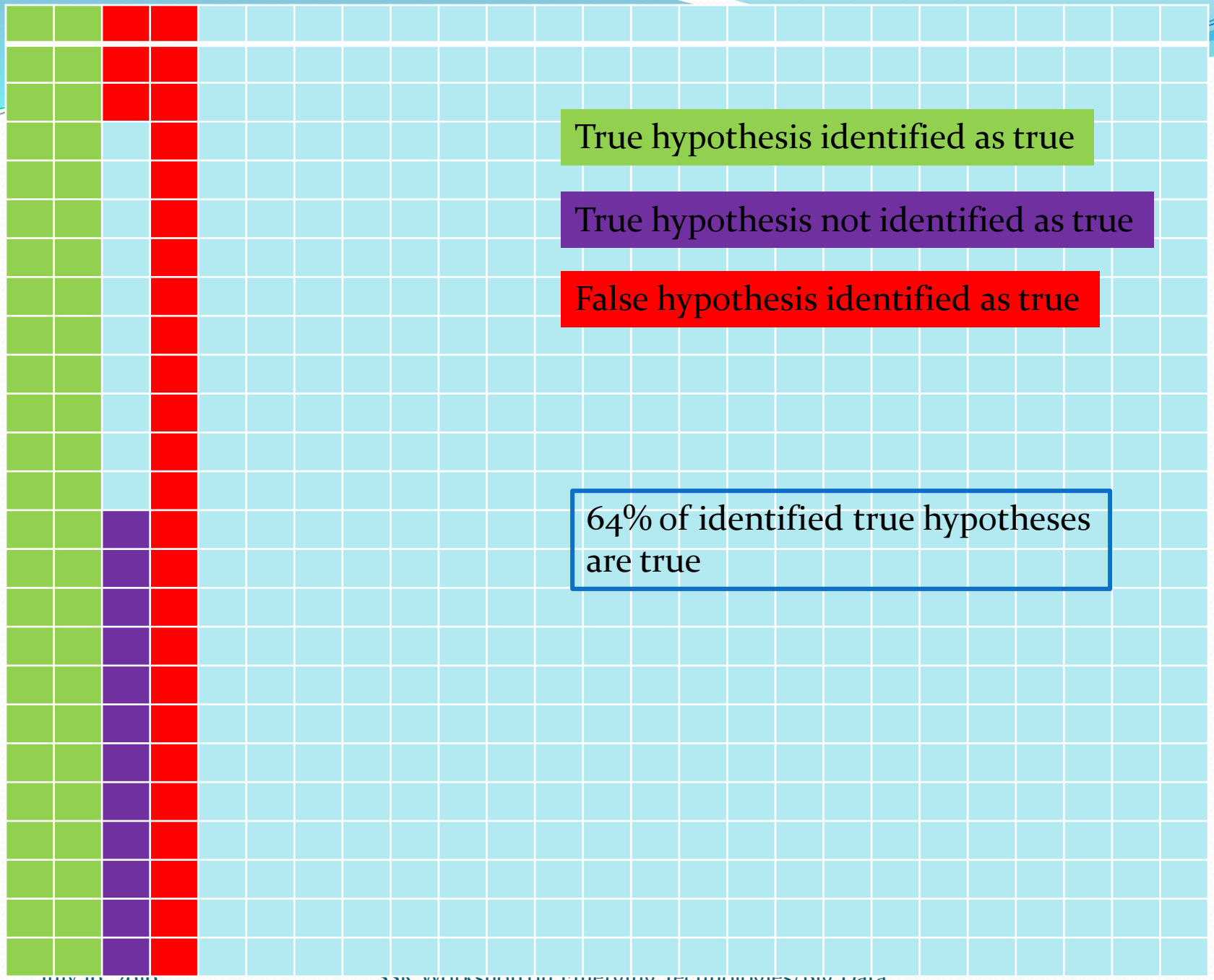
- If I have 20,000 measurements (say gene expression) on each sample (say control and treatment), then I can choose a subset of the measurements to use to predict treatment or control.
- Even if the 20,000 measurements are completely random, it will be possible to choose a small number of measurements (maybe 10 or 20 out of 20,000) which collectively predict perfectly.
- In addition, even if the variables are pre-specified, fitting the coefficients of the predictors will result in often highly optimistic bias.
- Measurements of prediction accuracy must always be done on samples that were not used in any way to compute the predictor.

Adjusting for multiple comparisons

- If I test 20,000 genes for differential expression at the 5% level of significance, there should be around 1000 false positives.
- If we have 2000 significant genes, half will be false discoveries
- It is possible to choose a significance level that is low enough that no gene will be significant most of the time when the null hypothesis is true, though that makes it hard to detect real differences.
- Instead, most investigators control the false discovery rate (FDR) at, say, 10%.
- This means that if we tag 2000 genes as "significant" then we estimate that 200 are false positives and the rest true.

Is a Published Result True?

- Suppose that a hypothesis that drug A has effect B has an a priori likelihood of 0.10 to be true.
- If it is false, then the null hypothesis that drug A has no effect may be true, and the chance of a statistically significant result is 0.05.
- If it is true, suppose the study has 80% power, so the chance of a statistically significant result is 0.80.
- Out of 10,000 such hypotheses, 1000 are true. Of those 1000, 800 will have significant p-values.
- Out of the 9,000 hypotheses that are not true, 450 will have significant p-values.
- So the chance that the result is true given that $p < 0.05$ is 800/1250 = 64%


True hypothesis identified as true
True hypothesis not identified as true
False hypothesis identified as true
64% of identified true hypotheses are true

Is a Published Result True?

- If there are 20 research groups pursuing this hypothesis, and if the result is published if any of the groups gets a significant p-value, then all 1000 of the true hypotheses will be verified, but also 5,774 of the 9,000 false hypotheses will be "proven"
- The chance that a published result is true would then be 1000/6774 = 15%.
- This gets worse if 1) there are more groups, 2) the finding is more surprising, 3) the power is lower.
- For example, with 50 groups, and a priori probability of 1%, and 30% power, the chance that the result is true is 1%.
- Ioannidis (2005)

Valid comparisons

- Suppose I have metagenomic sequence data from stool samples of 32 infants targeting the 16s ribosomal RNA subunit genes.
- From this, I can obtain counts for particular bacterial types (e.g., *Bacteroidetes*).
- I can validly ask if one group of infants (C-section) differ in the count of *Bacteroidetes*.
- But asking if for any one infant there is more of one bacterial type than another is fraught with difficulties.
- There are many known sequence biases, and many likely unknown ones.
- The same is true of microarrays and can even be true for PCR.

Personalized Medicine

- Most medical research is focused on groups of patients with (supposedly) similar conditions.
- An example would be a clinical trial to help decide if chemotherapy is helpful after surgery for stage I and stage II gastric cancer patients.
- Personalized medicine would mean making this decision for each patient separately, based on (usually) genetic, genomic, or biomarker data.
- Although this sounds like a good idea, implementation is very difficult and has been achieved only in a few cases.

Breast Cancer

- Mutations in the tumor suppressor genes BRACA1 and BRACA2 are associated with elevated risk of breast cancer.
- Breast tumors can be evaluated by several markers that affect treatment and prognosis.
- Estrogen/progesterone receptor (ER/PR) tumors can be treated with tamoxifen or aromatase inhibitors.
- Her2/neu positive can be treated with Herceptin and related drugs.
- Triple negative tumors can be treated only with standard, less targeted chemotherapy.

Personalized Medicine Successes

- Treatment of breast cancer depends not only on the stage, but also on genetic characteristics of the tumor.
- This type of personalized medicine has been successful in cases in which the molecular mechanisms are well understood.
- Other successes come from sequencing the tumor genome and looking for mutations in genes that relate to specific therapies.

The Search for Criteria

- Much of the research in personalized medicine depends on gathering large amounts of information about patients and correlating the information with outcome, with response to therapy, or with diagnosis.
- For example, one may test serum or tissue samples using a gene expression array, which can have as many as 50,000 measurements.
- Sequencing transcripts or the genome of the patient or of a tumor also can yield tens of thousands of measurements.

The Problem of Scale

- Oncotype DX is a proprietary predictive signature based on the levels of 21 transcripts which supposedly predicts the risk of recurrence in breast cancer.
- It was derived from 250 candidate genes in a trial of 400 patients. The number of sets of 21 genes from 250 is 2×10^{30}, which is a large number (larger than a trillion trillion, which is only 2×10^{24}
- The number of gene sets of 21 or fewer is larger by a factor of 2 million
- If the genes had been selected from the full array of 25,000 to 50,000, the number is 10^{72} to 10^{79}, nearly the number of particles in the universe.

Biomarkers for Personalized Medicine

- Without exercising very great care, when one starts with 50,000 measurements on each patient, there is a large chance of accidental correlations that will not extend to a new group of patients, and is thus useless.
- We understand much about methods for preventing false discoveries, but there is still much to learn.
- It seems that many to most studies that purport to make such discoveries are not ever validated, and perhaps cannot be.

Rules for Diagnostic Markers

- Many studies use a comparison group that is essentially arbitrary and may differ from the patients in many ways other than the disease.
- Only effects that are so large that they could not be due to measured or unmeasured differences in the groups can be trusted.
- These almost never occur, since effects that large likely would already have been discovered.
- These and other biomarker studies are only convincing when repeatedly replicated in different settings by different investigators.

Design Standards

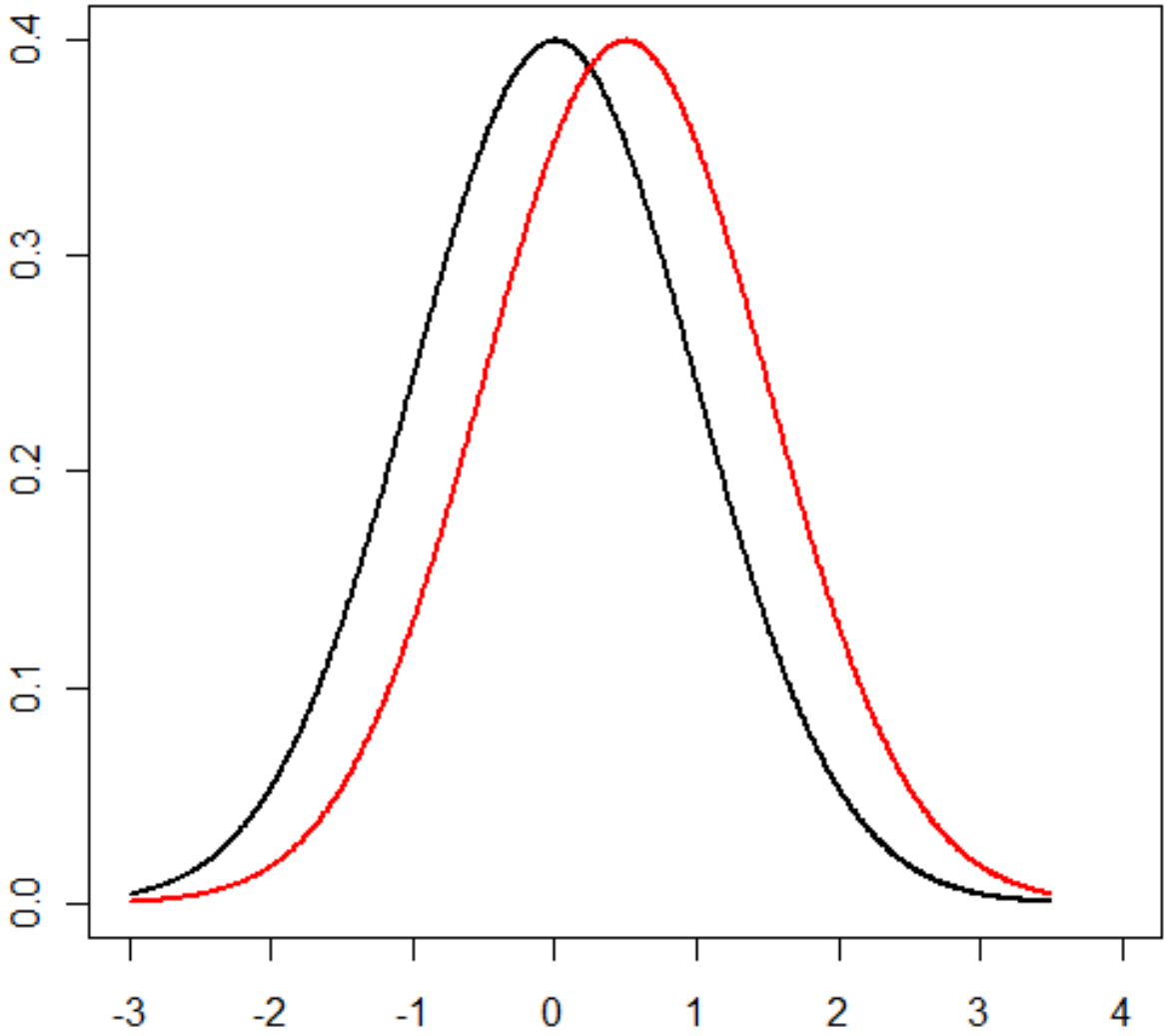
- The most believable studies start with a cohort containing the groups who will eventually develop, as the disease is diagnosed, or as the treatment is successful or unsuccessful.
- Ideally, the measurements are taken before the outcomes are known.
- Any candidate predictor is later validated blind.
- Early Detection Research Network (EDRN)

What is Needed for Valid Studies

- Besides good design, an important contributor to valid studies is good statistical/bioinformatics methods.
- We know a lot about what not to do (assessing fit on the same data that was used by the fit, data mining without consideration of the size of the problem).
- We need to put much effort into developing better methods of analysis, so that we can retain the sensitivity of broader methods while avoiding false discoveries as much as possible.

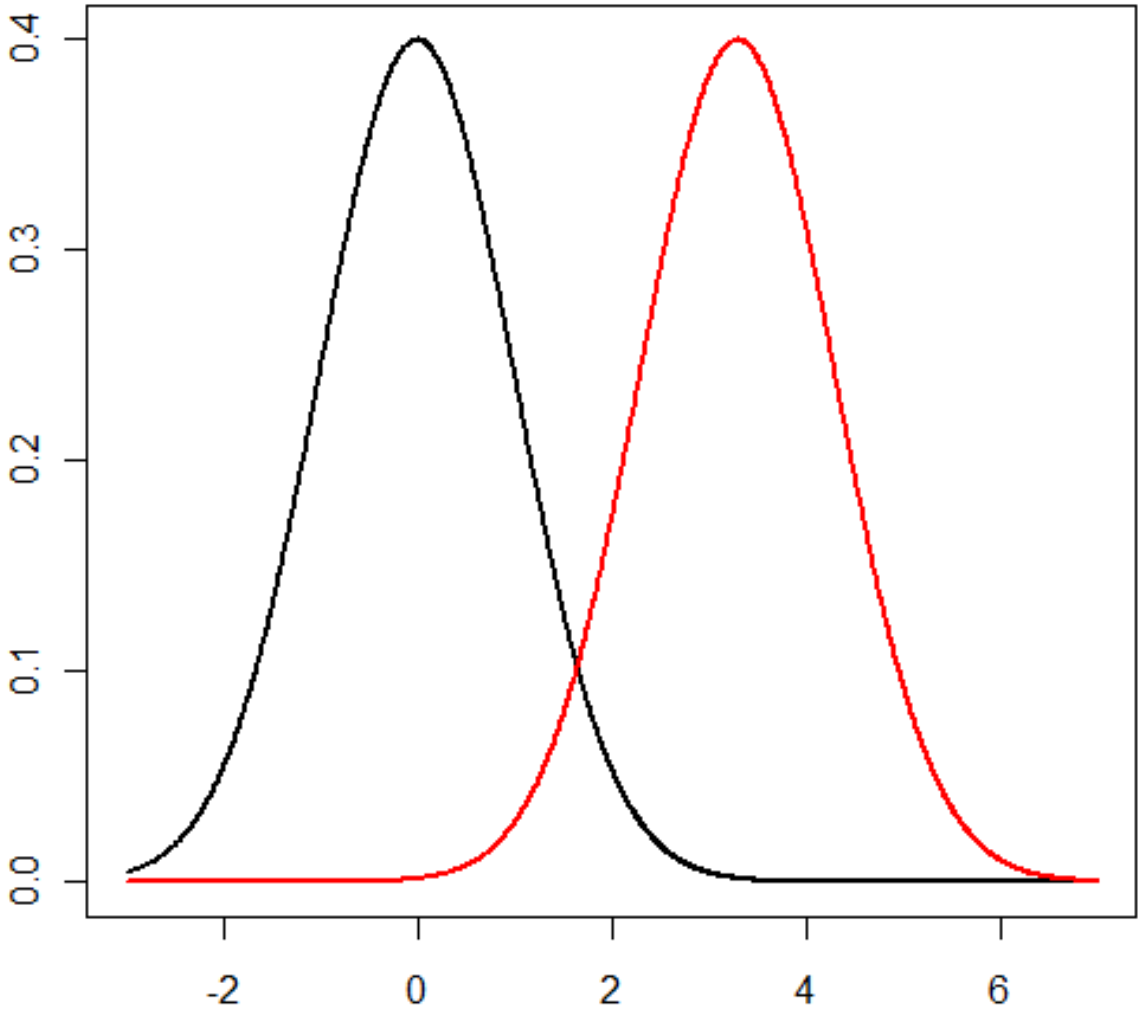
Statistical Significance and Classification Success

- It is easier for a variable to be statistically significant than for the classification using that variable to be highly accurate, measured, for example, by the ROC curve.
- Suppose we have 100 patients, 50 in each group (say disease and control).
- If the groups are separated by 0.5 times the within group standard deviation, then the p-value for the test of significance will be around 0.01 but the classification will only be 60% correct, and we can get 50% correct by saying nobody has the disease.


0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
4

Statistical Significance and Classification Success

- If the classification is to be correct 95% of the time, then the groups need to be separated by 3.3 times the within group standard deviation, and then the p-value for the test of significance will be around essentially 0.


0.0
0.1
0.2
0.3
0.4
-2
0
2
4
6

A Bad Example from Duke

- In 2006, Potti et al. introduced a method to combine array studies on the NCI160 cell lines with drug sensitivity assays to derive "signatures" of sensitivity to specific drugs.
- Coombes, Wang, and Baggerly (2007) at MD Anderson uncovered significant errors in the work, but Potti et al. did not acknowledge them, and stuck to their conclusions.
- They designed a clinical trial which "assigned subjects to either pemetrexed/gemcitabine or cisplatin/gemcitabine therapy using a genomic based platinum predictor to determine chemotherapy sensitivity and predict response to chemotherapy for first-line therapy in advanced non-small cell lung cancer."

- All of these claims were examined by Baggerly, Coombes, and colleagues (2009) and found to be corrupted by data errors, methodological errors, and other problems to the extent that all of the apparent conclusions were actually unsupported by the data.
- These errors lay somewhere on the continuum from sloppiness to fraud.
- Nonetheless, it was difficult even with all this for Baggerly and Coombes to attract enough attention to stop the trial.
- Duke carried out a secret internal review and then restarted the trial.

Outcome

- It was only when Dr. Potti was found to have claimed fraudulently on his CV that he was a Rhodes scholar that the trial was stopped.
- Four major papers were retracted and Potti lost his position.
- Baggerly and Coombes were proved correct in every particular.
- The systems of review for publication and internal review at Duke were both subject to question.

Cautionary Notes

- Complex and/or large data require great care and often sophisticated analysis, and trained statisticians specialized in the area should always be involved.
- Reproducible research is vital. Ideally, the raw data, and all the code required should be made public on publication so that the data fed into the program yields the results in the paper.
- Journals need to take seriously disputes over serious errors, and not treat them as “cat fights”, aka dueling statisticians.

Cautionary Notes

- Ioannidis et al. (2009) showed that of 18 microarray articles examined from Nature Genetics, a top journal, only 2 could be reproduced in full, 6 could be reproduced in part with some discrepancies, and 10 could not be reproduced at all.
- Steen (2011, J. Medical Ethics) showed that between 2000 and 2010, at least 80,000 patients had been in clinical trials based on research that was incorrect and later retracted, so the Duke case is not an isolated instance.
- Personalized medicine has great potential, but only if the research is conducted with great care.
- Most things that are tried (in any field) don't work, so persistence is required.

Studies of Innovations

- Gilbert, Light, and Mosteller in the 1970's studied large numbers of published reports on innovations in medicine, surgery, and social programs.
- Positive publication bias, so this is optimistic.
- Yet, most reported innovations did not work.
- Those that did work, worked only a little, at least in the first form tried.
- All the investigators must have believed that the innovation would work.
- There was a strong relationship between the enthusiasm for the innovation expressed in the paper and lack of statistical control!
- One main purpose of good statistical analysis is revealing when a positive appearing effect could just be statistical noise.

Reproducible Research

- A published paper should be backed up with original data files (not processed) and computer code that converts the data file into the results in the paper.
- Documenting the exact version of software and the parameters used is critically important in advanced areas like RNA-Seq, where software and the default parameters can change sometimes weekly.
- How can you edit a figure in response to referee comments if you can't reproduce the figure?
- So even reproducing your own work can sometimes be difficult.

Experimental Design

- You need more biological replicates than you can afford!
- Don't confuse technical replicates with biological replicates—almost always technical replicates are a waste or resources.
- Analyze the data in accord with the design—if you have two treatments and two sampling times with two biological replicates for each of the four conditions, then use two way ANOVA possibly with interaction.
- If the two times are each from the same animal/subject, then this is analyzed differently than if they are different animals/subjects.

Clear Hypotheses

- If you measure 20,000 things per sample, and your hypothesis is that there is something unspecified that is different between the conditions, then avoiding false positives is very difficult, and usually will need to be confirmed by separate studies.
- Focused investigations are less risky, at least once a clear hypothesis exists.

Randomization and Blinding

- Randomize assignment of treatments to subjects/animals/cell cultures
- Randomize run order
- Blind assessment when not automated
- 48 ovarian cancer serum samples, 48 controls, in batches of 12 (6 each) balanced roughly by age, and blinded to the analytical chemist.
- Mass spec analysis of glycomic markers.
- Some were significant for diagnosis.
- But the largest source of variability was by batch!

Lessons from the Omics Wars

- Be realistic
  - Many things don't work as expected
  - Measuring lots of things does not automatically generate statistically significant differences.
- Document, document, document
- Design your study carefully
- Have clear hypotheses
- Randomization and blinding
- Don't be afraid to fail a few times—not every experiment is a paper