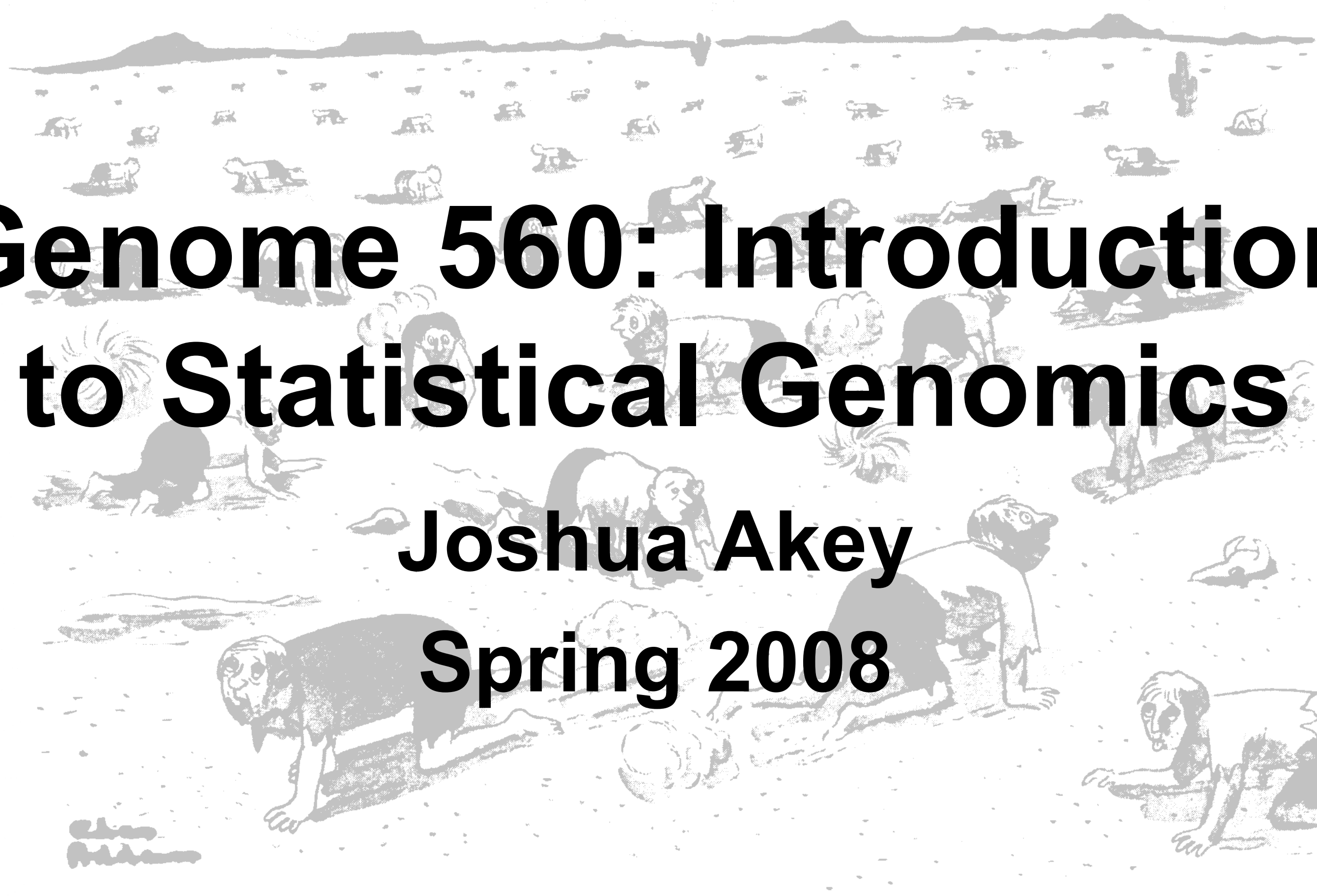


Genome 560: Introduction to Statistical Genomics

Joshua Akey

Spring 2008

Why You Are Taking This Course

- It’s required - now shut up and sit down

- Because I’m a sadist: You’re about to be confused, made to feel stupid, and bored for an entire quarter (shut up and sit down)

Why You Are Taking This Course

- ***Data*** are interesting, and they are interesting because they help us understand the world

- *Genomics = Massive Amounts of Data Data*

- Statistics is fundamental in genomics because it is integral in the ***design***, ***analysis***, and ***interpretation*** of experiments

The Roots of Modern Statistics Emerged From Genetics

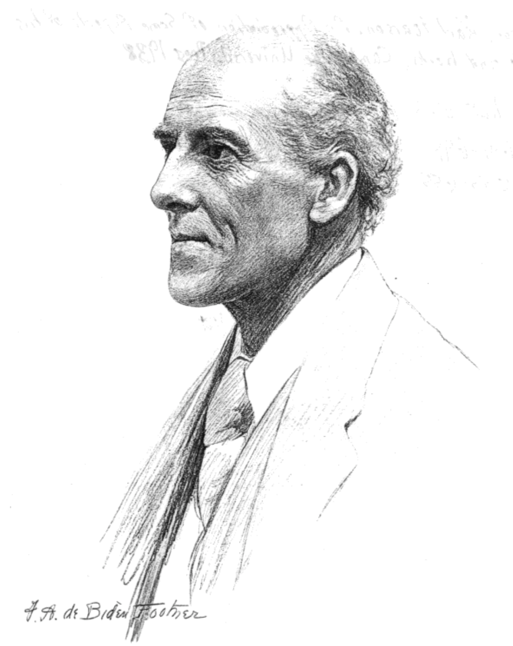
Sir Francis Galton

Inventor of fingerprints,
study of heredity of quantitative traits

Regression & correlation

Also: efficacy of prayer,
attractiveness as function of
distance from London

Karl Pearson



Polymath-

Studied genetics

Correlation coefficient
χ^2 test
Standard deviation

Sir Ronald Fisher

The Genetical Theory of
Natural Selection
Founder of population genetics

Analysis of variance
likelihood
P-value
randomized experiments
multiple regression
etc., etc., etc.

Why I Am The Rightful Heir of Statistical Genomics

Fisher

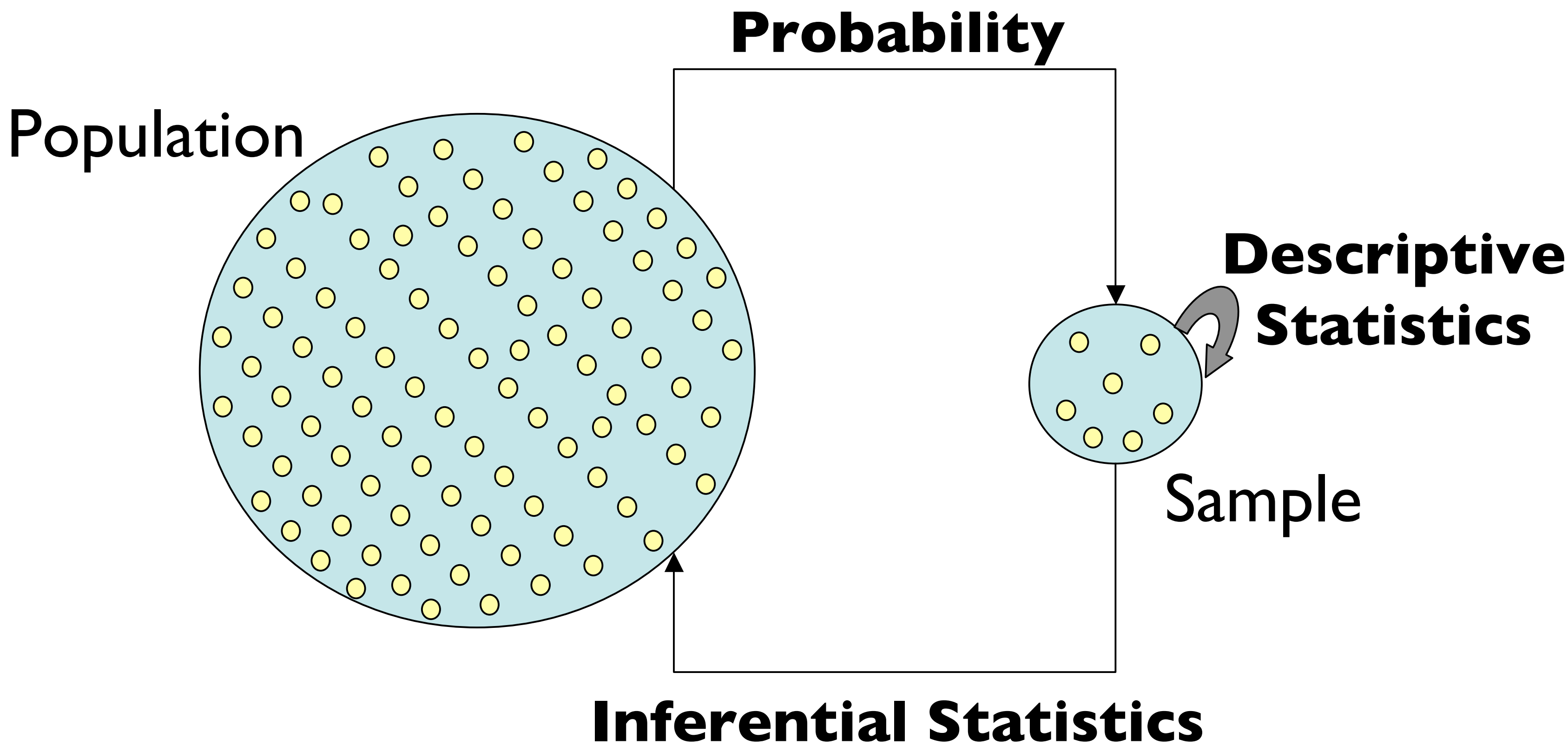
Rao

Chakraborty

Jin

Akey

“Central Dogma” of Statistics



Course Stuff

Syllabus:

Date	Topic
April 1	Collecting Data and Experimental Design
April 3	Descriptive Statistics and Visualizing Data
April 8	Randomness and Probability
April 10	Distributions
April 15	Estimating Parameters
April 17	Hypothesis testing 1: Inferences based on one or two samples
April 22	Hypothesis testing 2: ANOVA
April 24	Linear Regression
April 29	Analysis of Categorical Data
May 1	Assessing Significance in High-Dimensional Space

Grading:

5 problem sets (20% each) - sorry, really no other way

Books and Resources

- No required text

- Good on-line resources

  - http://www.math.wm.edu/~trosset/Courses/351/book.pdf
  - http://www.statsoft.com/textbook/stathome.html
  - http://www.stat.berkeley.edu/~stark/SticiGui/Text/toc.htm

- Some good books if you ever have some extra $$$:

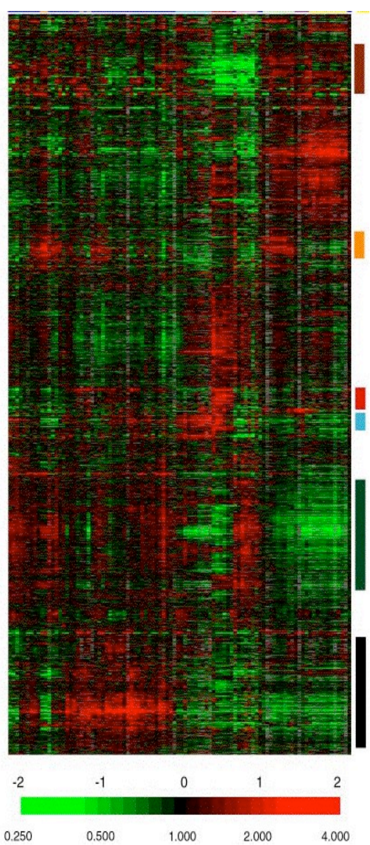
  - Probability and Statistics for Engineering and the Sciences 6th ed. Jay L. Devore. (2004). Duxbury press, Thompson-Brooks/Cole.
  - Statistical Inference. Casella, G. and Berger, R. L. (1990). Wadsworth, Belmont, CA.

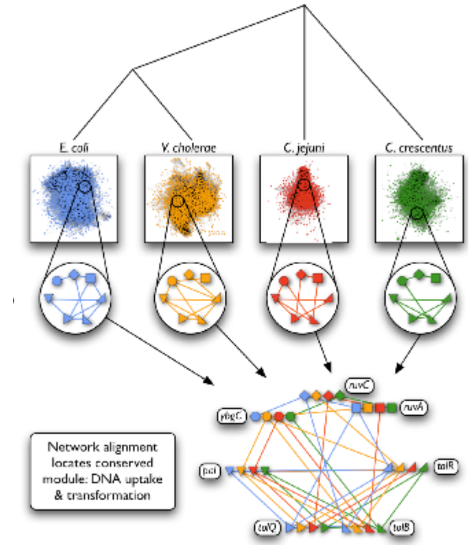
Collecting Data and Experimental Design

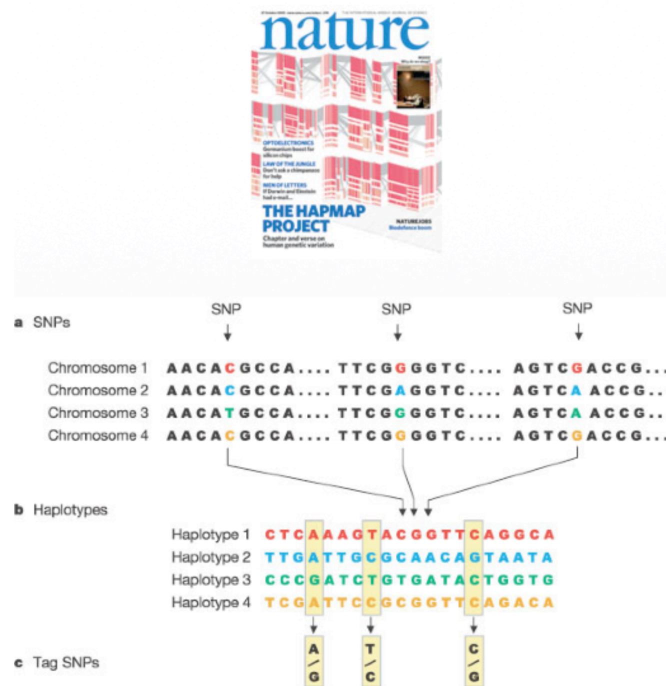
"[Experimental design] encompasses the myriad details that constitute the substance of the actual planning, conduct, and interpretation of a research study"

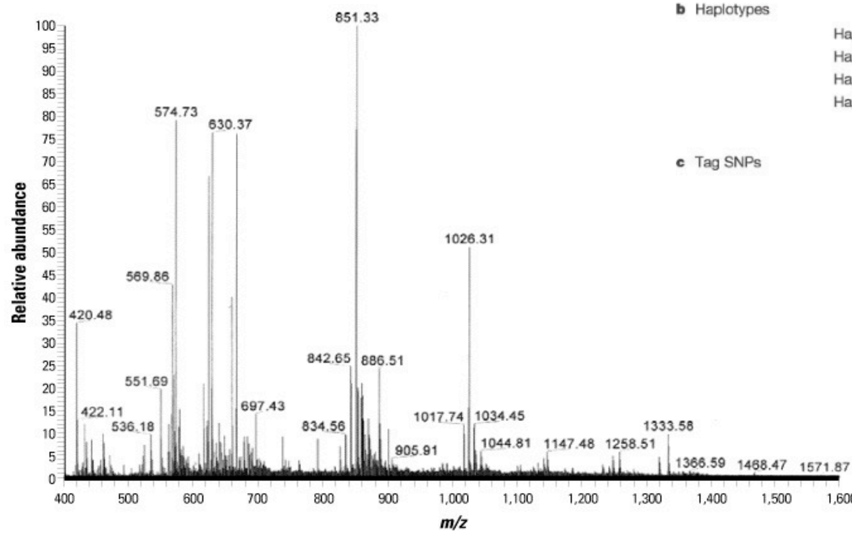
- Ransohoff (2007) Journal of Clinical Epidemiology, 60:1205

Typical Genomics Data









Bad Things Can Happen With Bad Experimental Design

Common genetic variants account for differences in gene expression among ethnic groups

Richard S Spielman[1], Laurel A Bastone[2], Joshua T Burdick[3], Michael Morley[3], Warren J Ewens[4] & Vivian G Cheung[1,3,5]

- Compared gene expression levels between 60 CEU and 82 ASN HapMap individuals

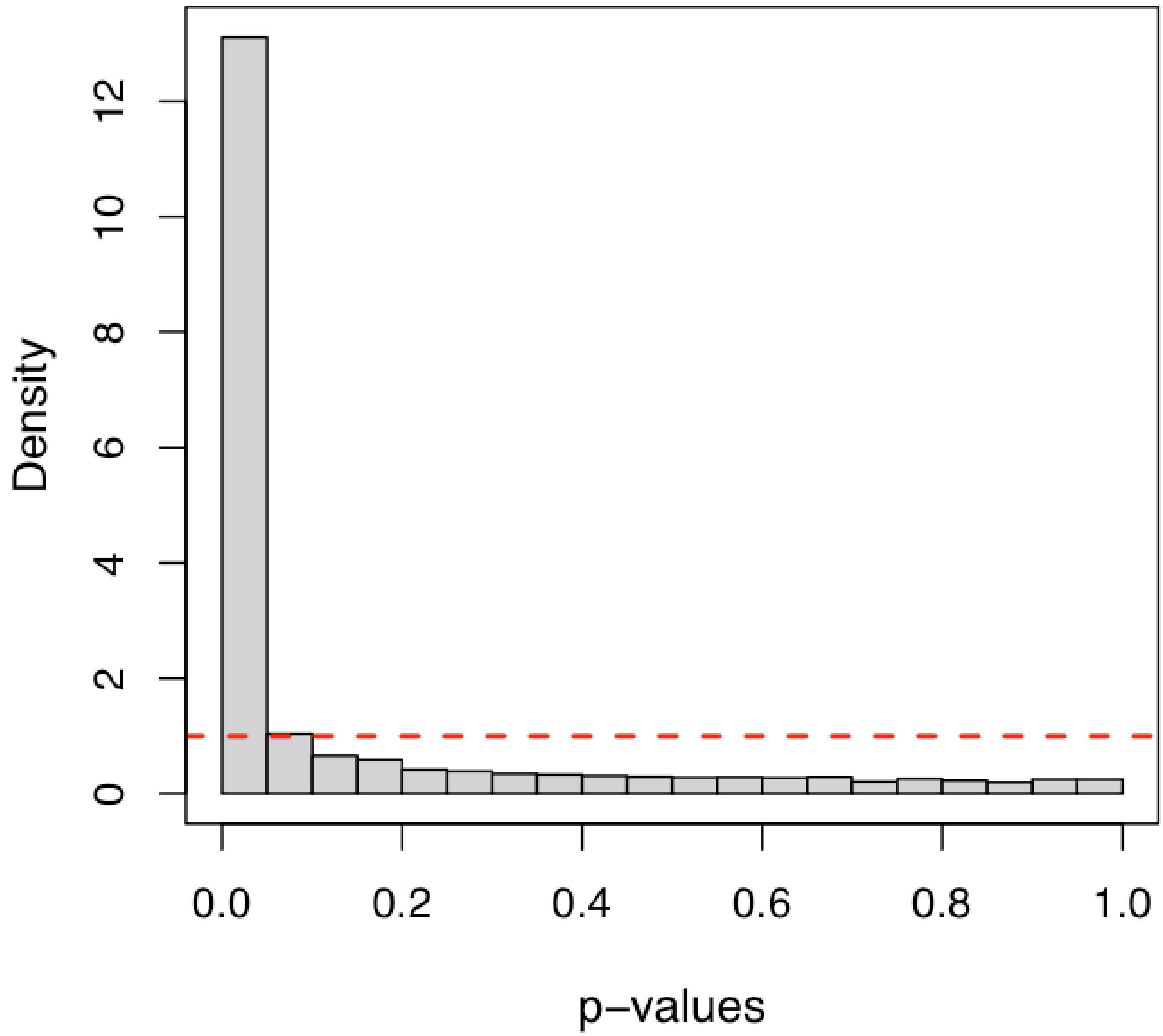
- Tests of differential expression performed by ***parametric t-tests*** and adjustment for ***multiple testing*** through Sidak corrections

- Estimate ~26% of genes to be differentially expressed

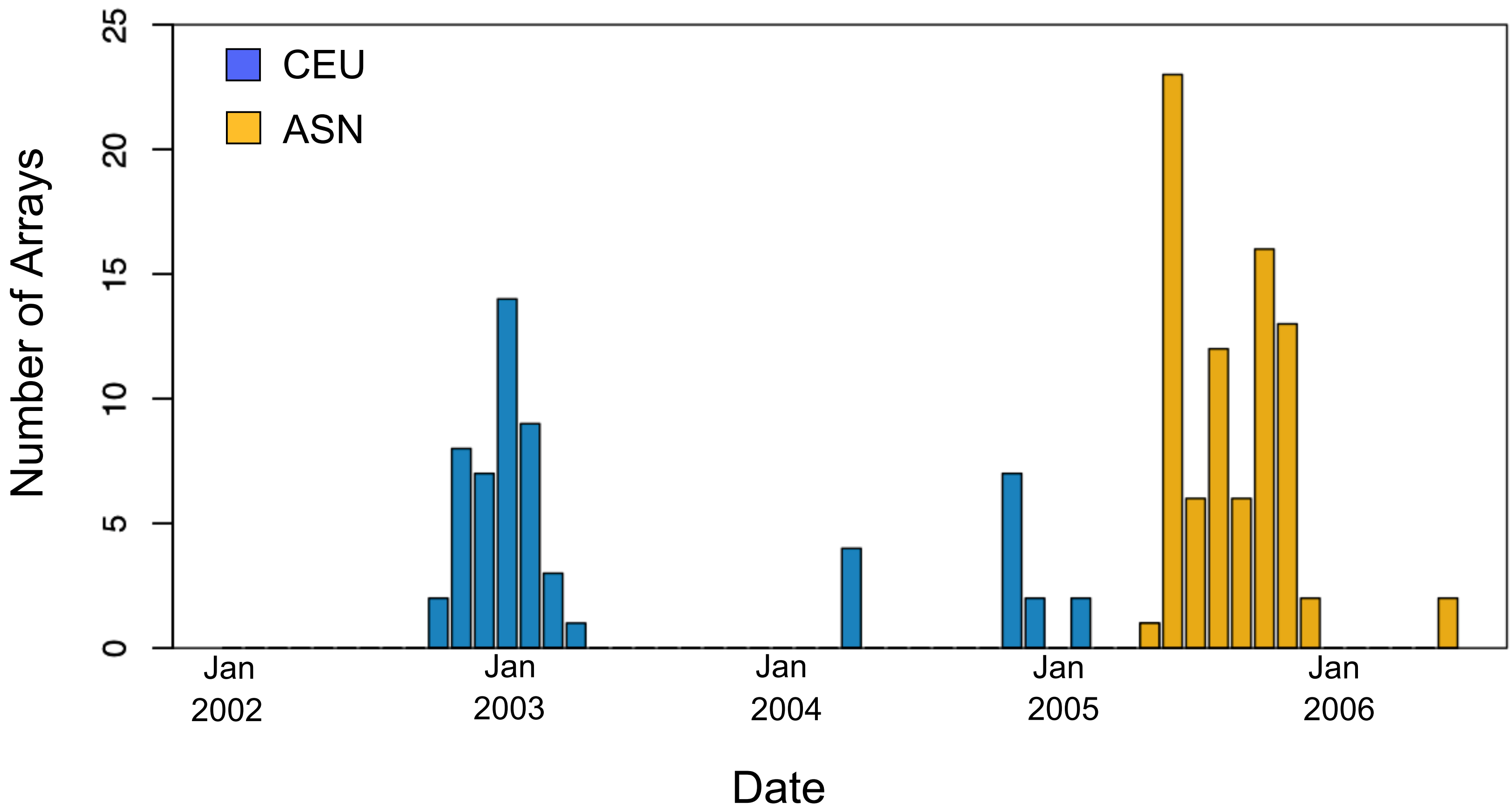
On the design and analysis of gene expression studies in human populations

To the Editor:
In a recent *Nature Genetics* Letter entitled "Common genetic variants account for differences in gene expression among ethnic groups," Spielman *et al.*[1] estimate the number of genes

78% of Genes Are Estimated To Be Differentially Expressed

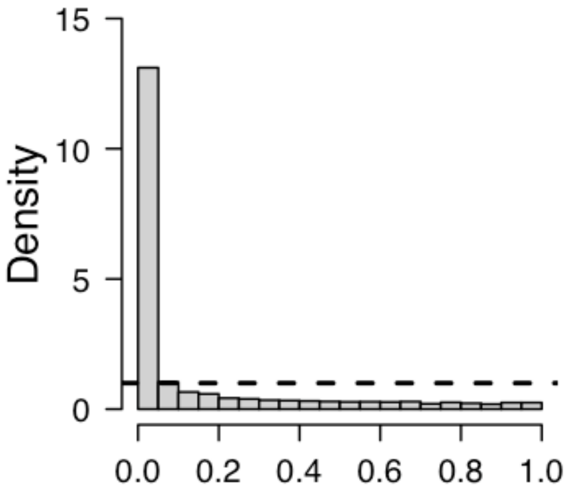


Population and Time of Processing Are Confounded



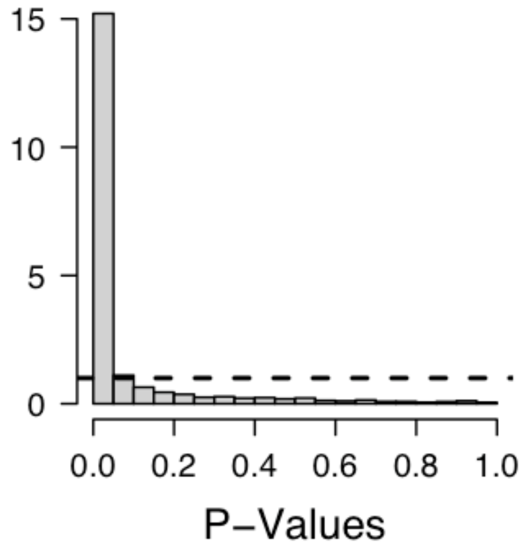
Batch Effects Can Completely Account For Differential Expression

Between Population



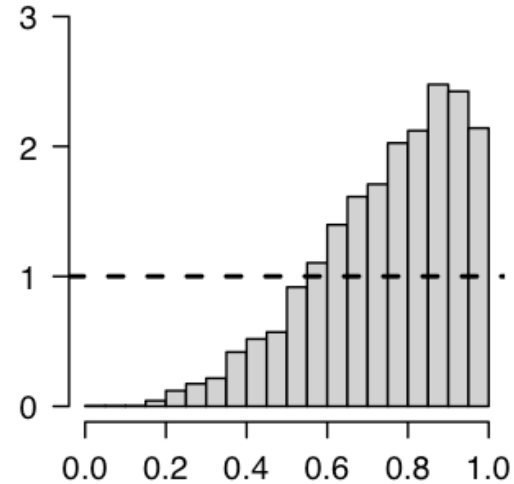
78% of genes estimated to be differentially

Between Years



96% of genes estimated to be differentially

Between Populations, Adjusting For Years



0% of genes estimated to be differentially

Elements of Good Experimental Design

- Experimental design is a whole sub-discipline in statistics research

- For genetics/genomics studies the two most important ideas are:

    1. ***Randomization***

    2. ***Control***

- Randomization and control are essential for making valid statistical inferences and minimizing bias caused by confounding variables

Some Jargon

- **Units:** the basic objects on which the experiment is done
- **Variable:** a measured characteristic of a unit
- **Treatment:** any specific experimental condition applied to the units. A treatment can be a combination of specific values (called *levels*) of each experimental factor.
- **Bias:** consistent divergence between the value of a variable in a sample from the corresponding value in a population

Why Randomize

- Breaks the association between potential ***confounding variables*** and the explanatory variables

- Confounding variables: variables whose effects cannot be distinguished from one another

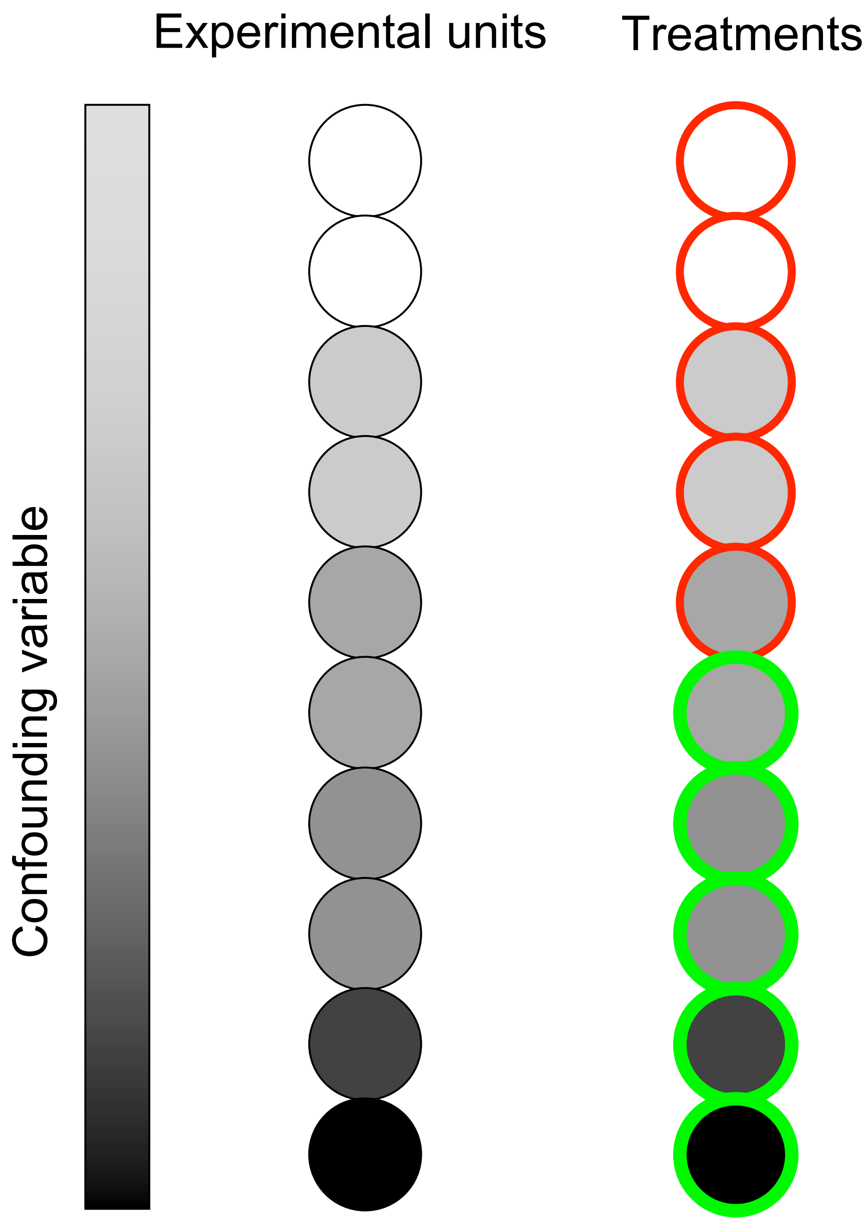
- Helps to avoid hidden sources of bias:

  - Association of shoe size (S) and literacy (L) in kids

Randomization

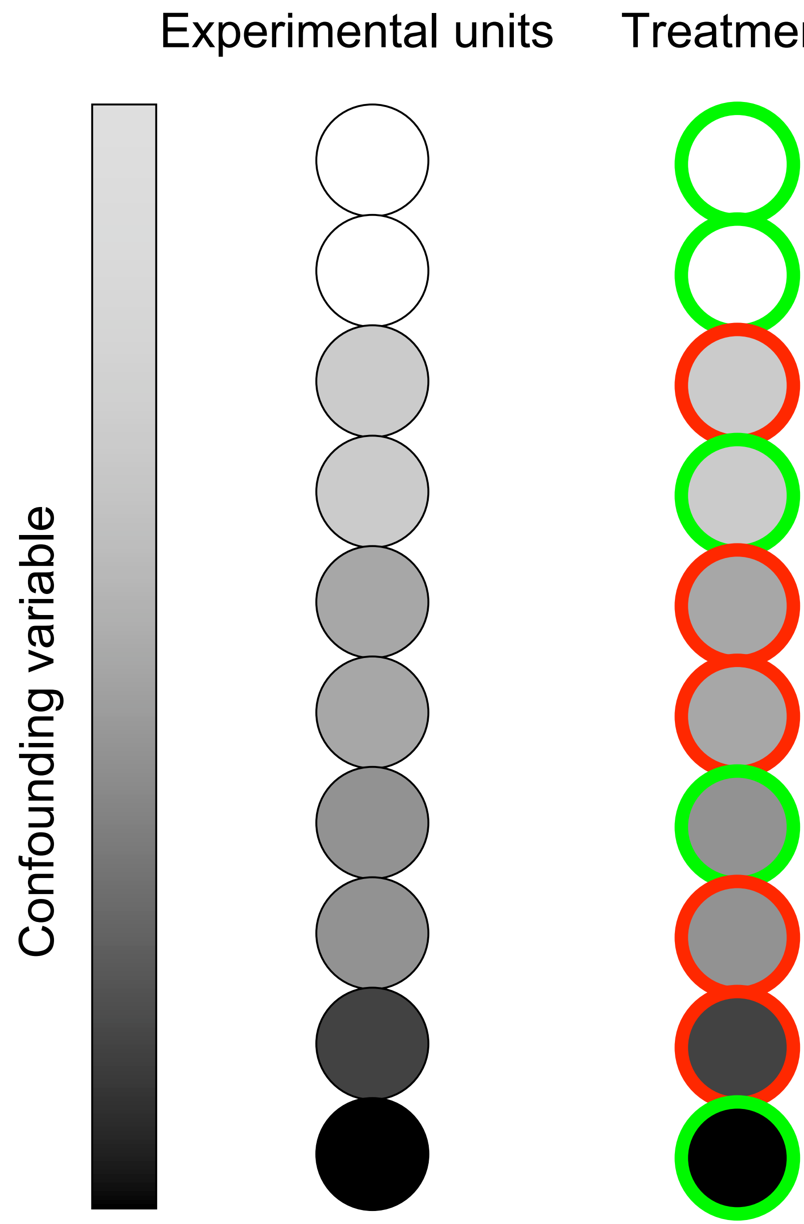
- Randomization can be implemented in multiple ways depending on the particular experiment
  - Randomly selecting individuals from a population
  - Randomly assigning treatments to units in an experiment
  - Randomization in the technical aspects of how an experiment is performed

Confounding



Without randomization, the confounding variable differs among treatments

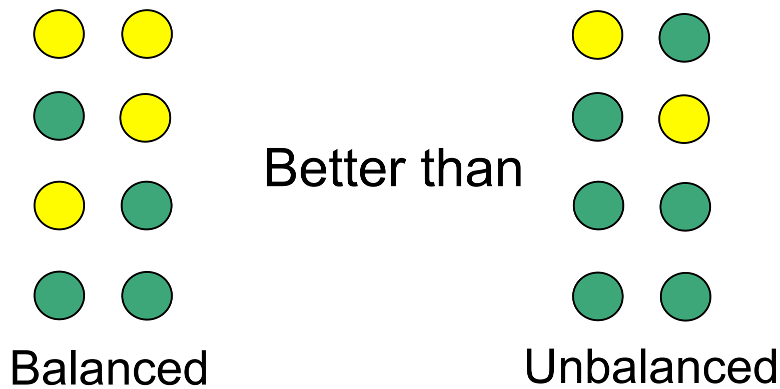
Confounding



With randomization, the confounding variable does not differ among treatments

Other Aspects of Good Experimental Design

- Balanced experimental design: all treatments have equal sample size



- Replication: reduces and allows estimates of variation
  - Technical versus Biological

Eliminating Bias: Controls

- A control group is a group of subjects left untreated for the treatment of interest but otherwise experiencing the same conditions as the treated subjects

- Example: one group of patients is given an inert placebo

Thought Question

Using the principles we just discussed, how would you design the gene expression study of Cheung et al. discussed previously?

Summary: Elements of Good Experimental Design

- Allow unbiased estimation of treatment effects
- Allow estimation of underlying variability
- Control for known sources of extraneous variation
- Allocate treatments to units randomly
- As simple as possible

What is R?

- The R statistical programming language is a ***free open source*** package based on the S language developed by Bell Labs
- Many statistical functions are already built in
- Contributed packages expand the functionality to cutting edge research
- Amazing graphics
- Widely used in genetics, genomics, bioinformatics: Learn it, love it, use it…

R Resources

- Windows, Mac, and Linux binaries available at:

  http://www.r-project.org

- Extensive resources at the above web-site, in particular see:

  http://cran.r-project.org/other-docs.html

Goals of Our R Tutorial

- Installing R
- Using R as a fancy calculator
- Data structures: scalars, vectors, data frames, matrices
- Reading in data from a file
- Subsetting and extracting data
- Writing and executing simple R scripts