Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 9 – Resampling, Batch Effects, and Normalization

Sara Mostafavi

January 31 2018

**Slide credits: Drs. Jenny Bryan; Su-In Lee; Doug Fowler*

Today:

- Review of multiple testing

- Resampling
  - Bootstrap
  - Permutation testing

- Batch effect and confounding problems
  - Statistical approaches for adjusting for them

Problem:

- In a typical gene expression study, we measure ~20K probes/genes and thus test ~20K hypotheses. If we use a p-value of 0.05, even if there are no effects (e.g., comparing WT vs WT), we end up with 0.05*20K = 1000 "significant" hits.

- Different solutions for addressing this problem:
  - Family Wise Error Rate control (e.g., Bonferroni)
  - False Discovery Rate control (e.g., BH procedure)

Review: False Discovery rate

	Null	Alternative	
	True	True	Total
Not Called Significant	U	T	$m - R$
Called Significant	V	S	R
	m_0	$m-m_0$	m

V = # Type I errors [false positives]

Total number of “discoveries”

Review: Benjamini and Hochberg FDR

- Proposed the idea of controlling FDR

$$FDR = E(\frac{V}{R}) \approx \frac{E(V)}{E(R)}$$

- Proposed a procedure for doing so
  - Note that we know R but we don’t know V
- Procedure: to control FDR with level q*
  1. Order the unadjusted p-values: $p_1 \leq p_2 \leq \ldots \leq p_m$
  2. Then find the test with the highest rank, j, for which the p-value, p_j,is less than or equal to (j/m)xq
  3. Declare ranks 1 through j as significant

Controlling the FDR at q* 0.05

Rank (j)	P-value	(j/m)× q*	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

moving BH procedure towards a “p-value adjustment procedure”

Basically BH boils down to this:

Call a "hit" if $\text{p-value} \le \dfrac{\text{rank of p-value}}{m} q^*$

Let's try to get in a more practical form:

Call a "hit" if $\text{q-value} \le q^*$

That implies this definition of a q-value:

$$\text{q-value} \equiv \text{p-value} \frac{m}{\text{rank of p-value}}$$

Controlling the FDR at q* 0.05

Rank (j)	P-value	(j/m)× q*	Reject H_0 ?	q-value
1	0.0008	0.005	1	0.0080
2	0.009	0.010	1	0.0450
3	0.165	0.015	0	0.5500
4	0.205	0.020	0	0.5125
5	0.396	0.025	0	0.7920
6	0.450	0.030	0	0.7500
7	0.641	0.035	0	0.9157
8	0.781	0.040	0	0.9763
9	0.900	0.045	0	1.0000
10	0.993	0.050	0	0.9900

Resampling approaches

- Parametric hypothesis tests:
  - Assumptions about the nature/distribution of data
  - Statistically more powerful (p(TP|true))
- Nonparametric tests:
  - Fewer assumptions about data distribution (i.e., population doesn't have to follow a specific parametric distribution)
  - Price: reduced statistical power
- Resampling approaches:
  - Best of both worlds!

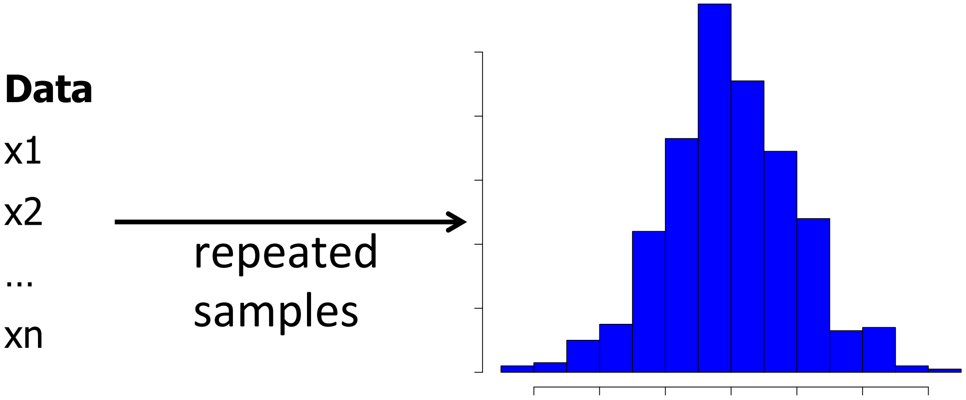
- Still everything assumes IID (unless you specifically design your resampling method to handle this)

What is resampling?

- Resampling procedure involves computing the test-statistic for many different arrangement of the data (i.e., recreating the sampling distribution “manually/computationally”).
  - We are assuming that the sampled data represents the “population”.

What is Resampling?

- Resampling procedure involves computing the test-statistic for many different arrangement of the data (i.e., recreating the sampling distribution “manually/computationally”).
- By doing so, we recreate the sampling distribution by “pretending” that we repeated the experiment many times.



Data

x1

x2

…

xn

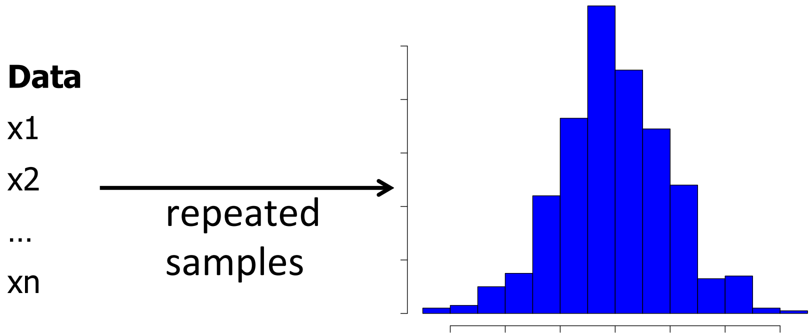
repeated samples

What is Resampling?

- Resampling procedure involves computing the test-statistic for many different arrangement of the data (i.e., recreating the sampling distribution "manually/computationally").

- By doing so, we recreate the sampling distribution by "pretending" that we repeated the experiment many times.

- We make this "empirically driven" sampling distribution to make inferences.



Data

x1

x2

…

xn

repeated samples

Resampling

- Bootstrap
  - Focus specifically on parameter estimation (e.g., confidence intervals, hypothesis testing)

- Permutation testing
  - Unconcerned about the actual value of parameters (estimates): mainly focused on hypothesis testing

Simple example

- You measure gene expression levels in two different tissues.

- Bootstrap: what's the variance of the population parameter of interest (e.g., mean)

- Permutation: is the mean expression of a gene different in the two tissues?

(you could use parametric approaches to answer the above, the point is that with resampling you don't make any of the assumptions that such test would make)

Why use resampling?

- Useful if we don’t know anything about the distribution of our population, and don’t want to rely on non-parametric test as they’d hurt the statistical power.
- Useful when we know the assumption of the test has been violated.
- Conceptually easy.
- Can create the null distribution for our a test-statistic that fits our problem.

Permutation testing

Observed data

Sample ID	Group	Expression of Gene A
1	W	0.6
2	W	0.5
3	W	0.9
4	KO	0.2
5	KO	0.4
6	KO	0.1

$|\mu_{1-} \mu_2|=0.47$

- Consider the observed data as one of many possible ways of "arrangements"

Permutation testing

Observed data

Sample ID	Group	Expression of Gene A
1	W	0.6
2	W	0.5
3	W	0.9
4	KO	0.2
5	KO	0.4
6	KO	0.1

$|\mu_{1-} \mu_2|=0.47$

Experiment 1:
Permute the assignment of groups

Sample ID	Group	Expression of Gene A
1	KO	0.6
2	W	0.5
3	KO	0.9
4	W	0.2
5	KO	0.4
6	W	0.1

$|\mu_{1-} \mu_2|=0.33$

- Consider the observed data as one of many possible ways of "arrangements"

Permutation testing

Observed data

Sample ID	Group	Expression of Gene A
1	W	0.6
2	W	0.5
3	W	0.9
4	KO	0.2
5	KO	0.4
6	KO	0.1

$|\mu_{1-}\,\mu_2|=0.47$

- Consider the observed data as one of many possible ways of "arrangements"

Experiment 1:
Permute the assignment of groups

Sample ID	Group	Expression of Gene A
1	KO	0.6
2	W	0.5
3	KO	0.9
4	W	0.2
5	KO	0.4
6	W	0.1

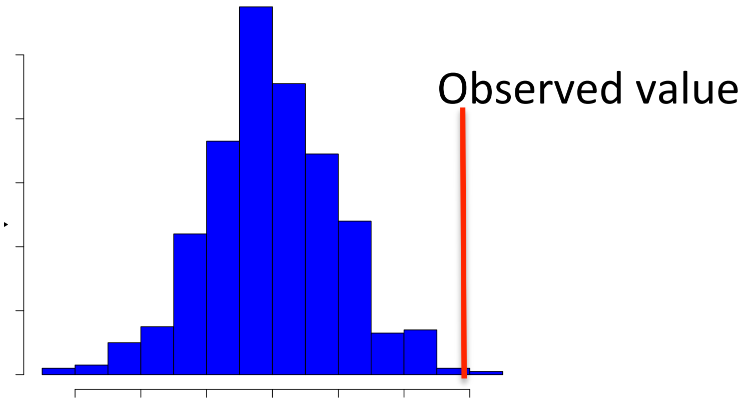
$|\mu_{1-}\,\mu_2|=0.33$

Experiment 2:
Permute the assignment of groups

Sample ID	Group	Expression of Gene A
1	KO	0.6
2	KO	0.5
3	W	0.9
4	W	0.2
5	W	0.4
6	KO	0.1

$|\mu_{1-}\,\mu_2|=0.066$

(hypothetical) histogram of mean differences from all permutation



$$p\text{-value}=\frac{1+\#\,\text{perm\_test} \leq \text{observed}}{1+\#\text{permutations}}$$

In the simple example there are exactly 20 permutations. So histogram of test statistics looks “sparse”.

Some notes on permutation testing

- In most cases, you can't enumerate all possible arrangements. So have to randomly select X.

- The smallest p-value you can achieve in X permutations is 1/(1+X).

- We can use permutation (resampling) testing to compute sampling distribution/p-values for our standard test statistics – more statistically powered than non-parametric test but still doesn't make assumptions about data distributions (other than IID)

- Down side: computationally expensive, especially if you want p-value with high resolution (e.g., $<10^{-6}$ requires 1M permutations)

The Bootstrap

- Used to assign measure of accuracy (e.g., confidence intervals, variance) to sample estimates of parameters.

- Bootstrap is useful for estimating sampling distribution of a statistic without using parametric distributions (e.g., z-stat, t-stat)

- As always: assumes IID

The Bootstrap - Procedure

- As in resampling/permutation, we treat the sample as the "population"

- Then we resample **with replacement** from the same population and recalculate our test statistic

- Finally w use the resampled distribution to draw conclusions (e.g., compute variance of the test statistic)

The Bootstrap

- Sample with replacement

Original data

Sample ID	Expression of Gene A
1	0.6
2	0.5
3	0.9
4	0.2
5	0.4
6	0.1
7	0.6
8	0.5
9	0.9
10	0.2
11	0.4
12	0.1

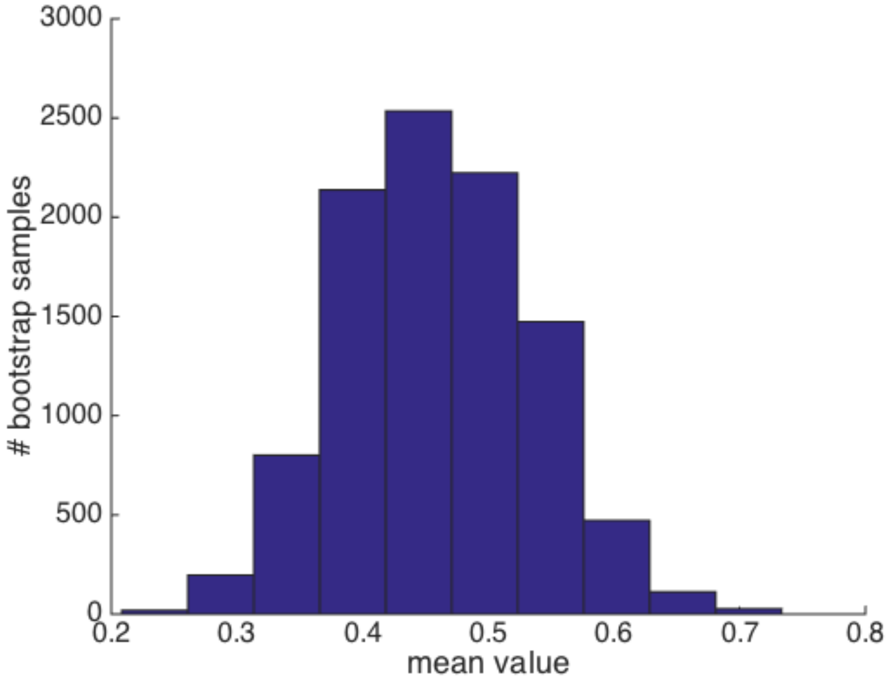
μ=0.45

Bootstrap 1: sample 12 “data points”

Sample ID	Expression of Gene A
1	0.6
5	0.4
3	0.9
12	0.1
5	0.4
5	0.4
7	0.6
9	0.9
9	0.9
12	0.1
11	0.4
2	0.1

μ=0.52

The Bootstrap - example



	Basic Assumptions	Power	Conceptual complexity	Ease of computation
Parametric	Strong	Best	High	Easy
Nonparametric	Weak	Less	Intermediate	Intermediate
Resampling	Weaker	Nearly as good as parametric	Low	Hard

Part II – Confounding factors and batch effects

Confounding

- Confounding: a situation in which a measure of association or relationship between response and explanatory variables is distorted by presence of another variable.

- Confounder: an extraneous variable that wholly or partially accounts for your observed effect.

Hypothetical example

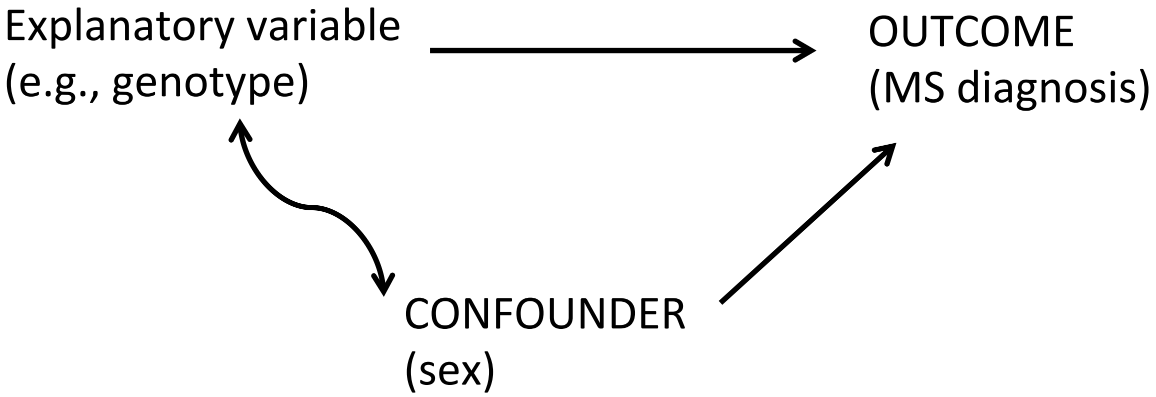
- MS more prevalent in females:
  - Cases: females with MS
  - Controls: males without MS
- What do you expect to find in gene expression analysis?

Definition of a confounder

- For a variable to be a confounder it should meet three conditions:
    1. The factor must be associated with the exposure being investigated
    2. Must be independently associated with the outcome being investigated
    3. Not be in the causal pathway between exposure and outcome.

Definition of a confounder

- For a variable to be a confounder it should meet three conditions:
  1. The factor must be associated with the exposure being investigated
  2. Must be independently associated with the outcome being investigated
  3. Not be in the causal pathway between exposure and outcome.



Explanatory variable
(e.g., genotype)

OUTCOME
(MS diagnosis)

CONFOUNDER
(sex)

Confounding factors in genomics studies

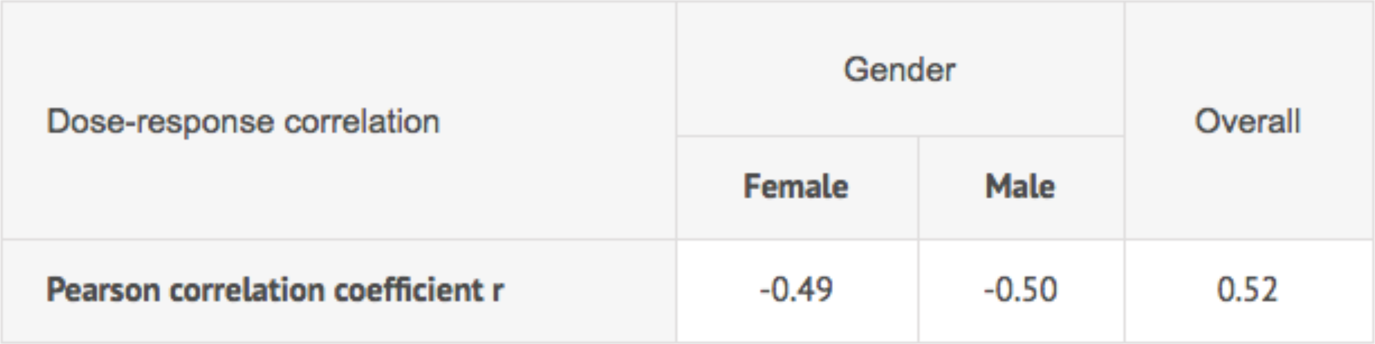
- Observational studies:
  - “Independent” variable is not under the control of the researcher (e.g., ethical reasons).
  - E.g., case/control study: which subjects are case and which are controls are out of the control of the investigator (e.g., SCZ study)
  - Typically many variables/factors are correlated with the “independent” variable of interest. The selection bias problem.

- Interventional studies:
  - E.g., randomized study: Investigator can randomly assign individuals to groups and so control the assignment of the independent variable.
  - Minimizes the **selection bias** problem.

Example: confounding factors in genomics study of MS

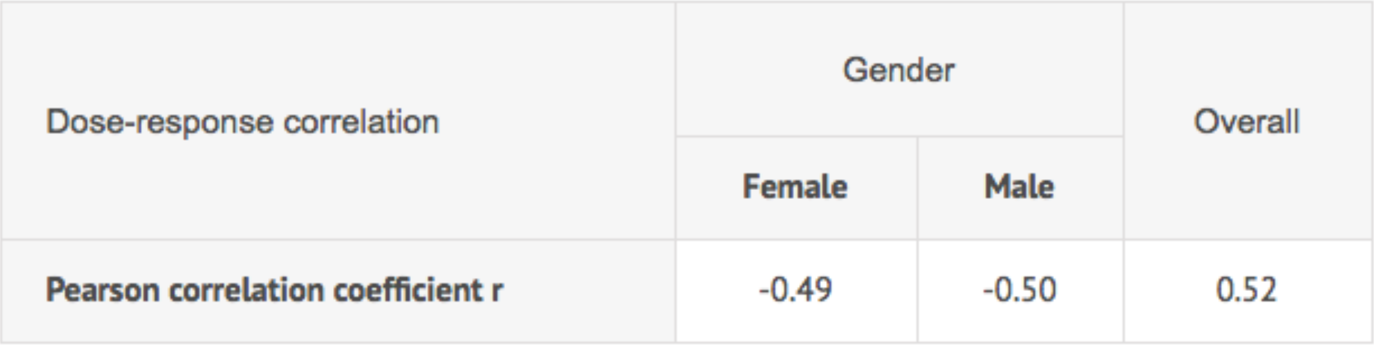
- Age
- Sex
- Smoking status
- Medication intake
- ….

Simpson’s paradox: an extreme case of confounding

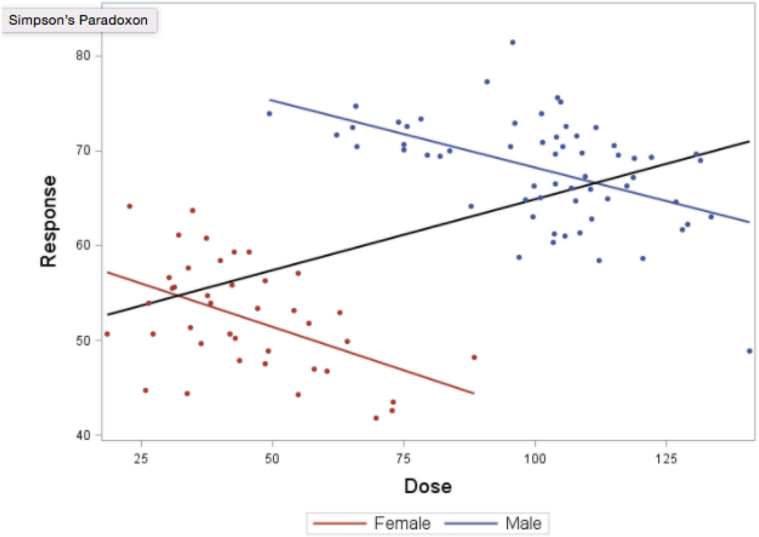


Dose-response correlation	Gender		Overall
	Female	**Male**	
Pearson correlation coefficient r	-0.49	-0.50	0.52

Simpson’s paradox: an extreme case of confounding



Dose-response correlation	Gender		Overall
	Female	**Male**	
Pearson correlation coefficient r	-0.49	-0.50	0.52



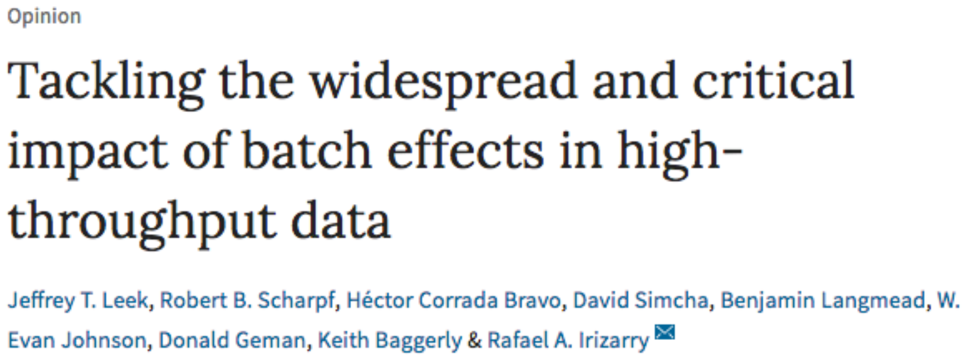
Types of confounding

- Experimental (batch effects):
  - E.g., heterogeneity of technical and biological replicates
- Demographical heterogeneity
  - E.g., sex, age…
- Environmental heterogeneity
  - E.g., smoking, alcohol use, …
- Genetic heterogeneity
  - E.g., population stratification

Batch effects are a huge problem in genomics study:

- Generation of data depends on: complicated reagents + software used by highly trained personnel

- If some of these conditions vary in the course of experiment: measurements for MANY genes/features will be effected.
  - E.g., subset of experiments were run on Monday and rest on Wed.



Opinion

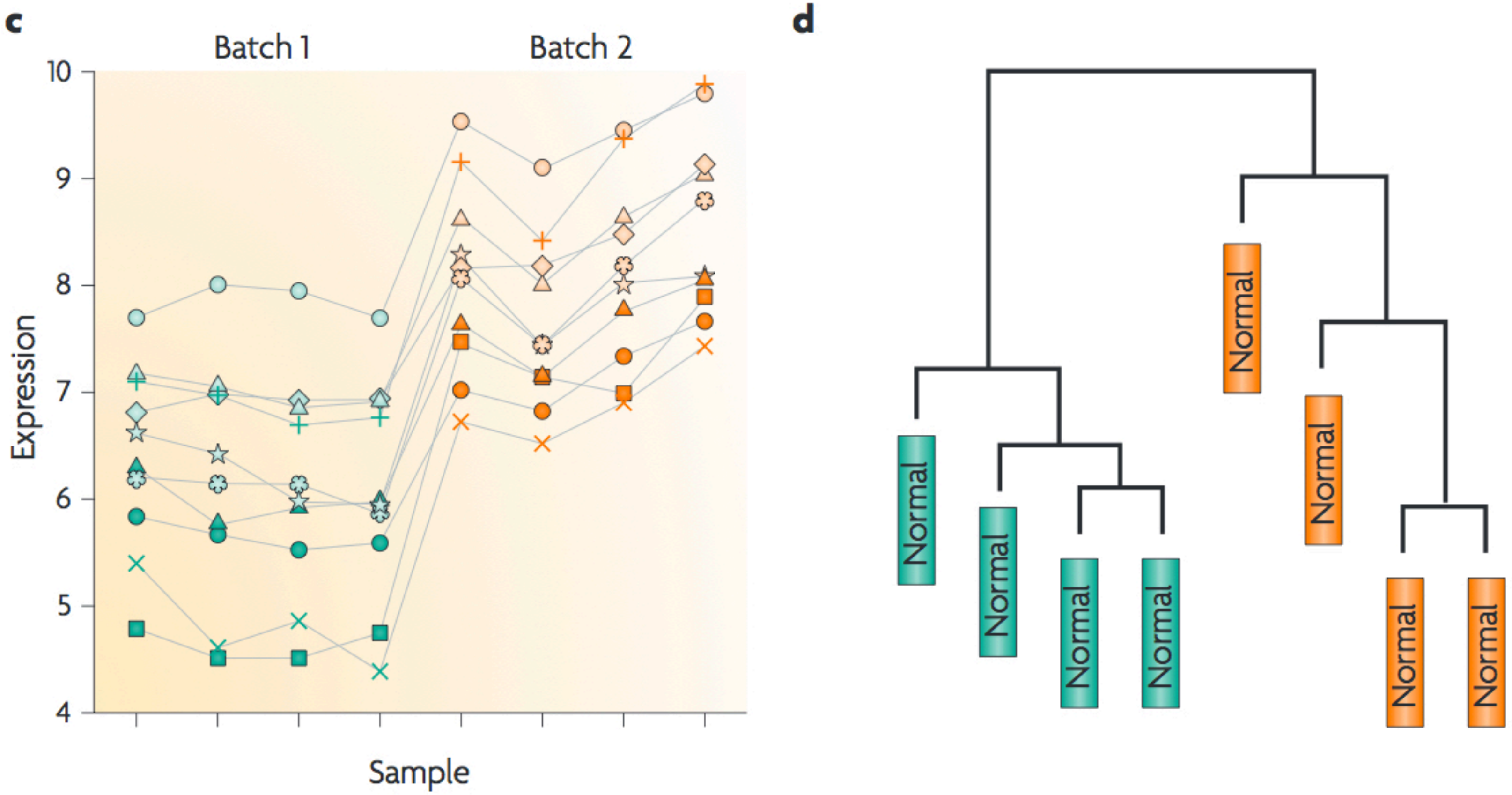
Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly & Rafael A. Irizarry

Opinion

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly & Rafael A. Irizarry



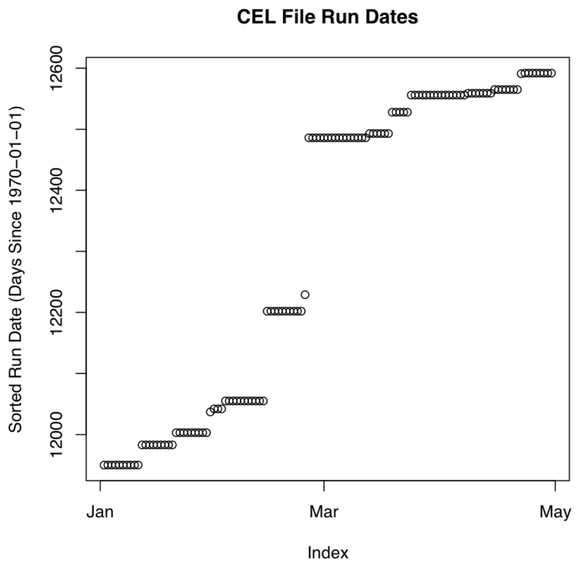
Consequence of batch effects

- Reduced statistical power: false negative
- Confounding and hence false positives

An Integrated Genomic-Based Approach to Individualized Treatment of Patients With Advanced-Stage Ovarian Cancer

Holly K. Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer...

Looked at profiles of 119 patients with ovarian cancer and signatures of response to cisplatin-based chemo.



Clinical data

Date	cancer stage
2392	Early Stage
2393	Early Stage
1772	Long
1773	Long
1774	Long
1775	Long
1776	Long
1777	Long
1778	Long
1779	Long
1780	Long
1781	Long
1900	Long

Survival is confounded with the date of sample collection!!!

Don’t use “public” data on auto-pilot

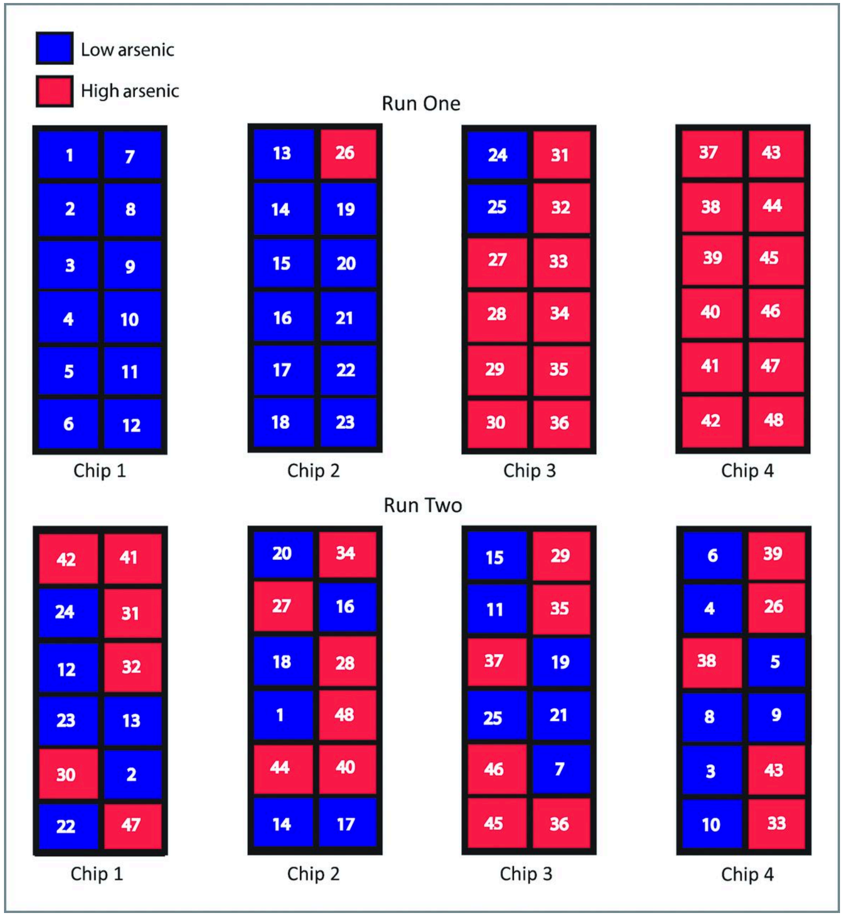
- TCGA: are samples pre- or post- chemo?? (not documented in most cases)
- MDD/SCZ expression profiling: who is taking which medication?
- Blood gene expression data: season??

Normalization vs experimental design

- Batch effects and confounding are an experimental design problem
- Normalization doesn't take care of confounding and can in fact exacerbate them.
  - Normalizing: modifying the scale or distribution of samples so they are comparable across the whole experiment.
  - E.g., normalizing to house keeping genes in qPCR, log transformation, variance stabilization, LOESS, quantile normalization…
- You need to explicitly address batch effects.
  - "pre data": Design of experiments that reduce potential for batch effects/confounders
  - "post hoc": Statistical adjustment

Experimental design solution

- Randomization
- Record keeping



🟦 Low arsenic

🟥 High arsenic

Run One

Chip 1		Chip 2		Chip 3		Chip 4	
🟦 1	🟦 7	🟦 13	🟥 26	🟦 24	🟥 31	🟥 37	🟥 43
🟦 2	🟦 8	🟦 14	🟦 19	🟦 25	🟥 32	🟥 38	🟥 44
🟦 3	🟦 9	🟦 15	🟦 20	🟥 27	🟥 33	🟥 39	🟥 45
🟦 4	🟦 10	🟦 16	🟦 21	🟥 28	🟥 34	🟥 40	🟥 46
🟦 5	🟦 11	🟦 17	🟦 22	🟥 29	🟥 35	🟥 41	🟥 47
🟦 6	🟦 12	🟦 18	🟦 23	🟥 30	🟥 36	🟥 42	🟥 48

Run Two

Chip 1		Chip 2		Chip 3		Chip 4	
🟥 42	🟥 41	🟦 20	🟥 34	🟦 15	🟥 29	🟦 6	🟥 39
🟦 24	🟥 31	🟥 27	🟦 16	🟦 11	🟥 35	🟦 4	🟥 26
🟦 12	🟥 32	🟦 18	🟥 28	🟥 37	🟦 19	🟥 38	🟦 5
🟦 23	🟦 13	🟦 1	🟥 48	🟦 25	🟦 21	🟦 8	🟦 9
🟥 30	🟦 2	🟥 44	🟥 40	🟥 46	🟦 7	🟦 3	🟥 43
🟦 22	🟥 47	🟦 14	🟦 17	🟥 45	🟥 36	🟦 10	🟥 33

[Harper, Peters, and Gamble, 2013]

Statistical approaches

1. Identify and collect “batch-related” variables.
   - Solutions for such information is missing (lecture on unsupervised learning)

2. “Explore” the effect of batch based on data visualization and dimensionality reduction

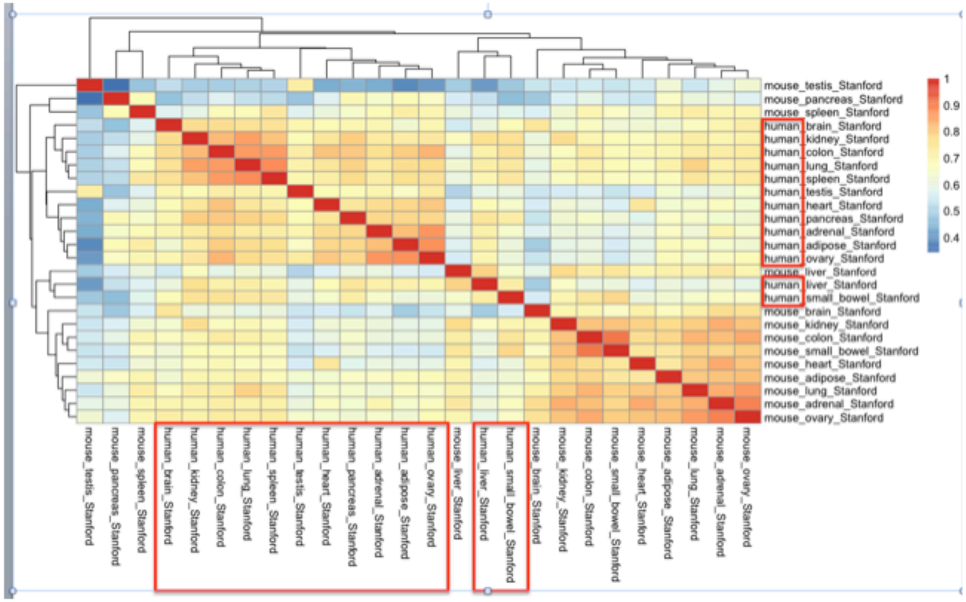
3. Model the effect of batch and “discount” it from the associated variability (lecture 16)

Visualizing batch effects

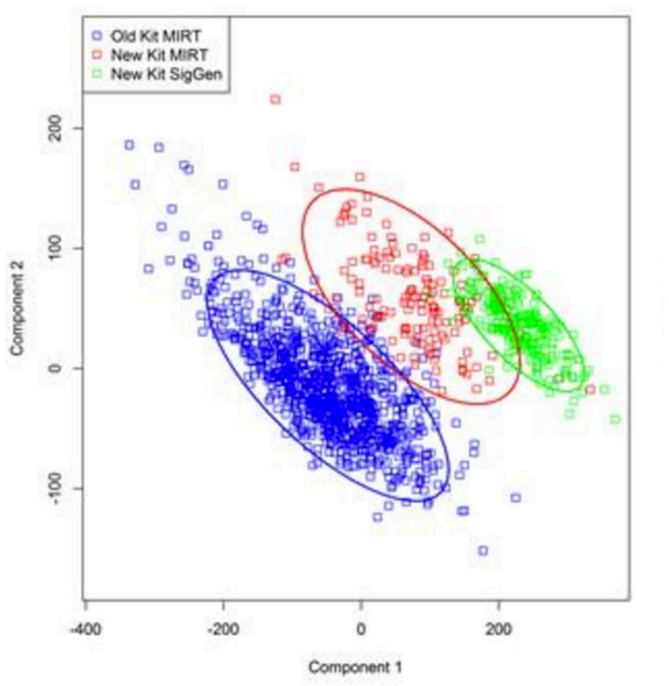
- Sample-sample covariance matrix (clustering)

- Principle Component Analysis (dimensionality reduction)

The original analysis in the paper, considering only the samples that were sequenced at Stanford (data cluster by species):



PCA : Unadjusted

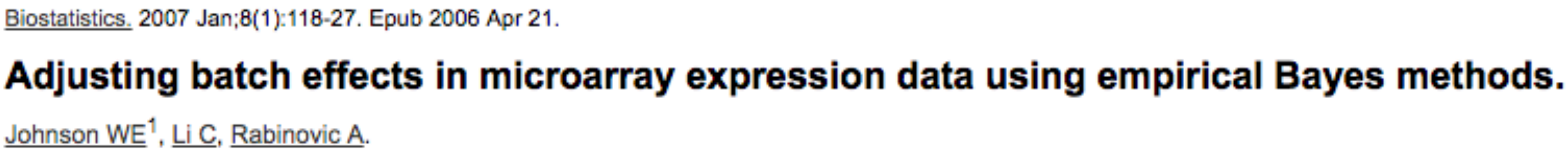


Statistical adjustment

- It's just a linear model again!
  - The two step approach
    - Fit a linear model to determine the effect of batch, use residual as the "batch corrected" data.
  - The "Combined" approach
    - A "batch" variable in your linear model
  - The "retainment" approach
    - Be careful!

- Active area of research and hence several different approaches
  - Not a big difference, main thing is to identify and somehow account for batch

Combat:



Biostatistics. 2007 Jan;8(1):118-27. Epub 2006 Apr 21.

Adjusting batch effects in microarray expression data using empirical Bayes methods.

Johnson WE[1], Li C, Rabinovic A.

- Model based location/scale (L/S) adjustments
  - Assume a model for the mean (location) and variance (scale) of the data and then normalizes across batches
  - E.g., standardize mean and standard deviation for each batch separately
    - Sensitive to unbalanced design

Sample i, batch j, gene g

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg},$$

Batch adjusted $$Y^{*}_{ijg} = \frac{Y_{ijg} - \widehat{\alpha}_g - X\widehat{\beta}_g - \widehat{\gamma}_{ig}}{\widehat{\delta}_{ig}} + \widehat{\alpha}_g + X\widehat{\beta}_g,$$