Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Lecture 9 – Confounding factors and batch effects

Sara Mostafavi

Feb 5, 2020

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

Today

Part (1): Definitions and concepts

Part (2): In-class activity (project discussion)

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

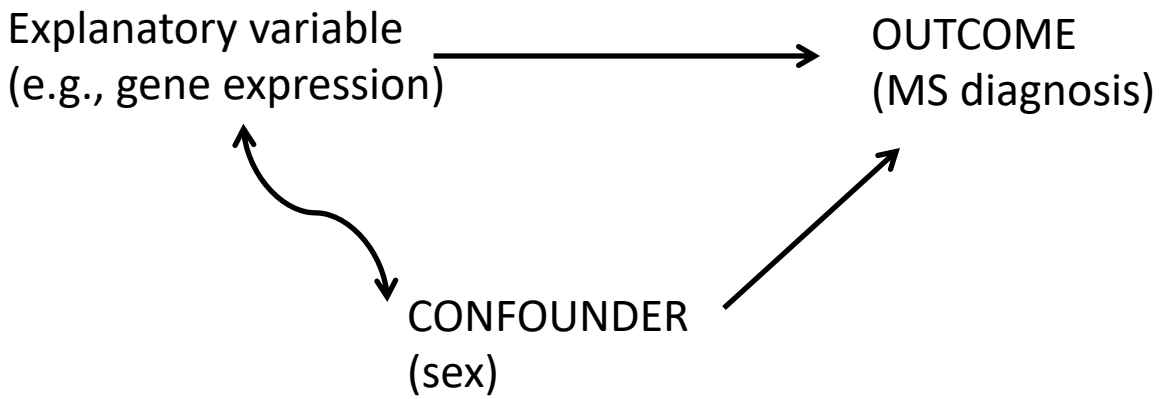
- MS more prevalent in females:
  - Cases: females with MS
  - Controls: males without MS
- What do you expect to find in gene expression analysis?

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.



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

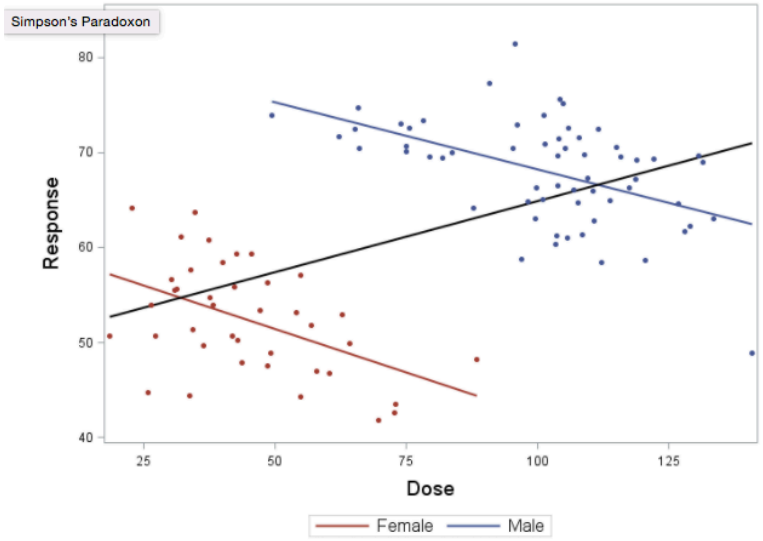
- Age
- Sex
- Smoking status
- Medication intake
- ….

Simpson's paradox

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

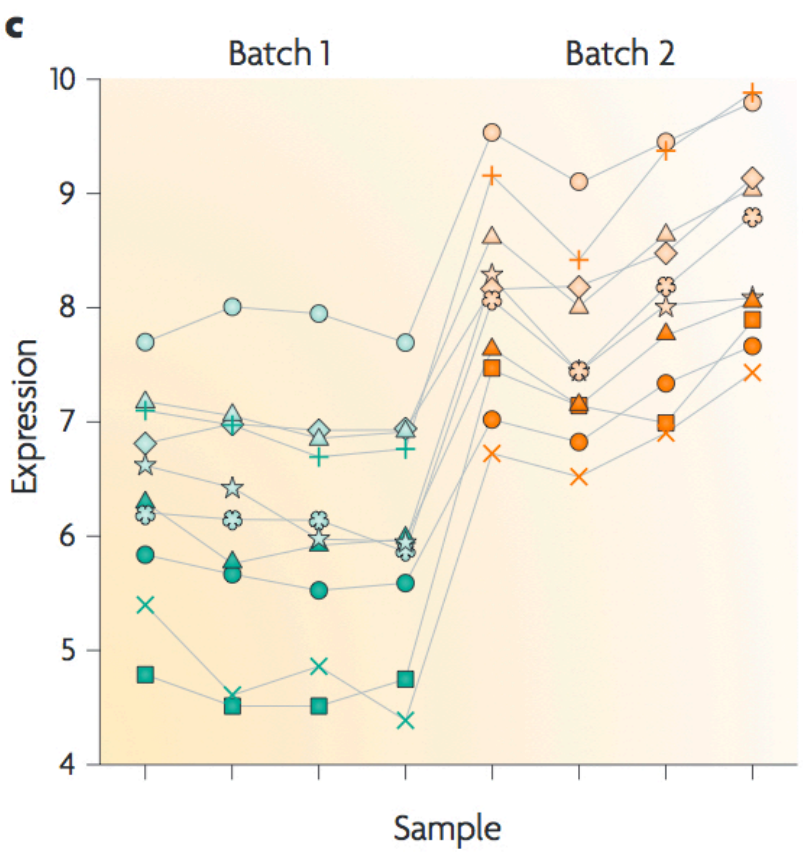
Tackling the widespread and critical impact of batch effects in high-throughput data

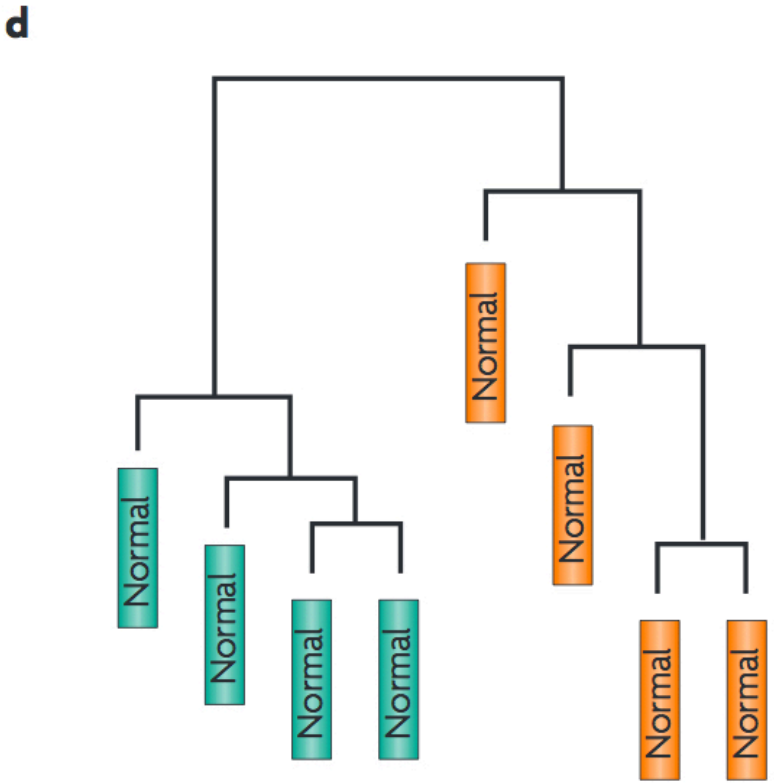
Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly & Rafael A. Irizarry

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

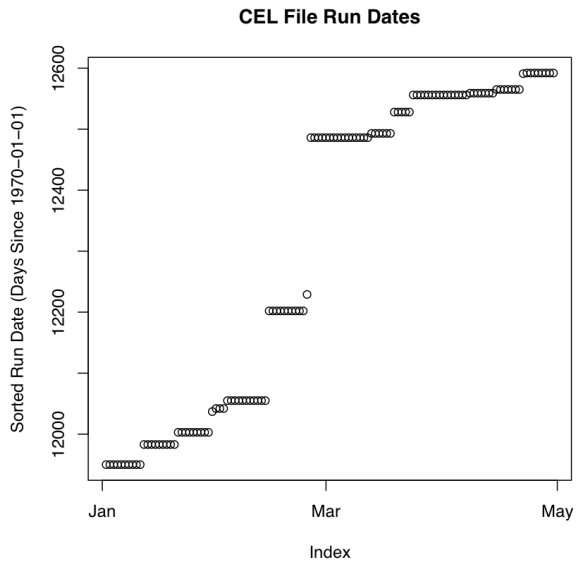
- Reduced statistical power: false negative
- Confounding and hence false positives

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...

Show More

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



Clinical data

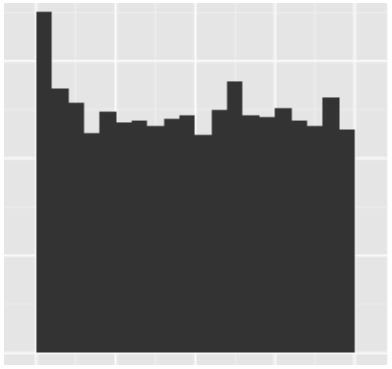
Date	cancer stage
2392	Early Stage
2393	Early Stage
1772	Long
1773	Long
1774	Long
1775	Long
1776	Long
1777	Long
1778	Long
1779	Long
1780	Long
1781	Long
1900	Long

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

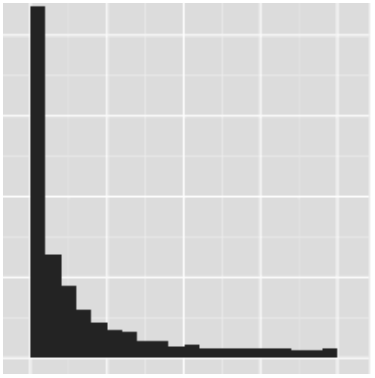
Diagnosing potential confounding effects

P-value distribution under the null should look uniform

P-value distribution under the null should look uniform



Too-much signal should alarm you



Don’t use “public” data on auto-pilot

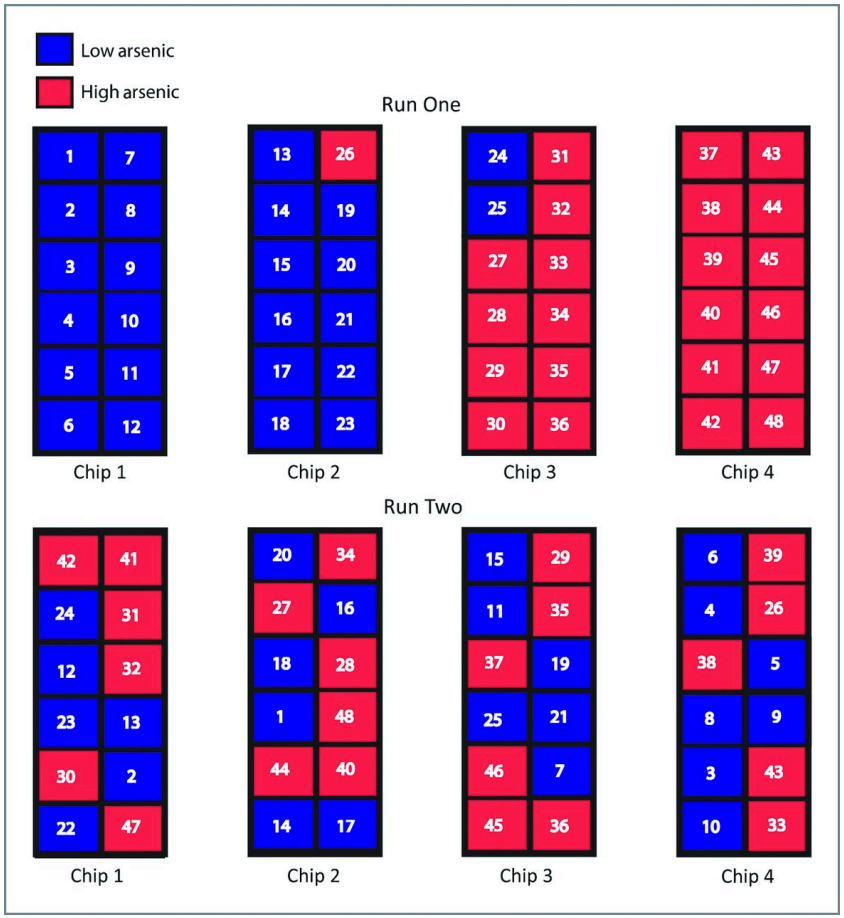
- TCGA: are samples pre- or post- chemo?? (not documented in most cases)
- MDD/SCZ expression profiling: who is taking which medication?
- Blood gene expression data: season??

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



[Harper, Peters, and Gamble, 2013]

Statistical approaches

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

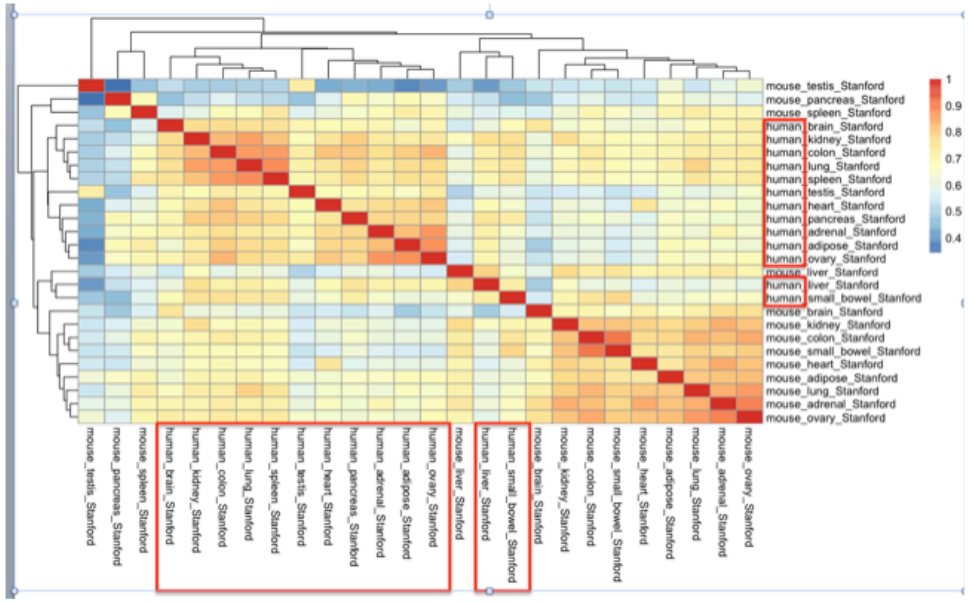
2. “Explore” the effect of batch based on data visualization and dimensionality reduction

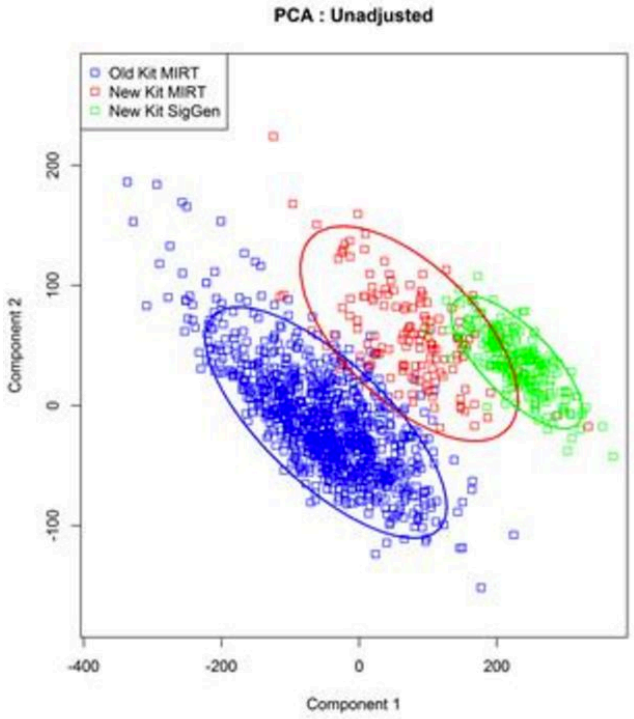
3. Model the effect of batch and “discount” it from the associated variability

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):





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!

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

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