

Batch effects, technical variables, and unwanted variation

Neo Christopher Chung

Lecture 4, 1000-719bMSB

Project, report, and presentation

Study a specific molecular system and a biological/medical question

Be inspired by biological functions, diseases, modeling approaches

Use the modern research practices (GitHub, reproducible codes, etc)

Have a specific hypothesis or an exploratory goal

Replicate an interesting research

Experiment with how an analysis is done

Improve methods and algorithms

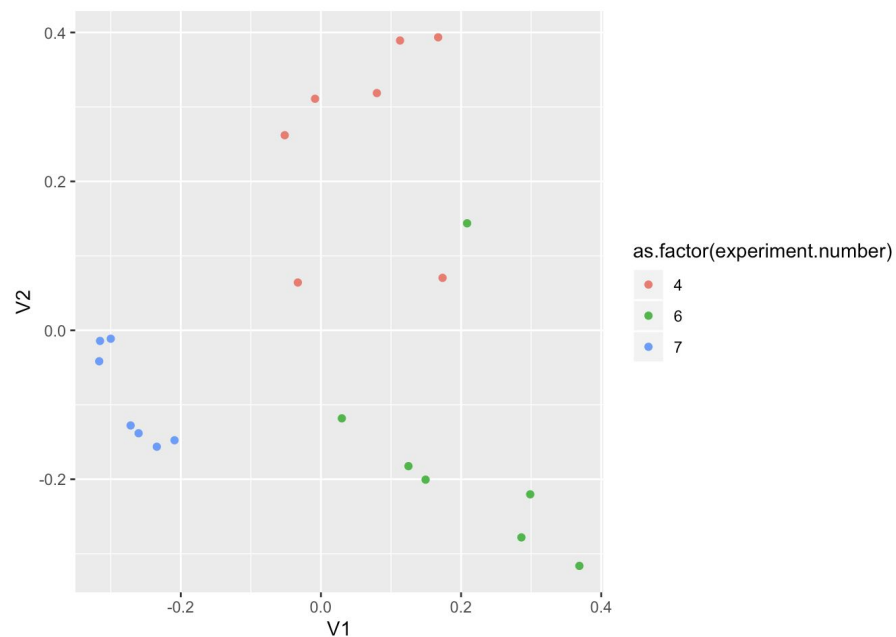
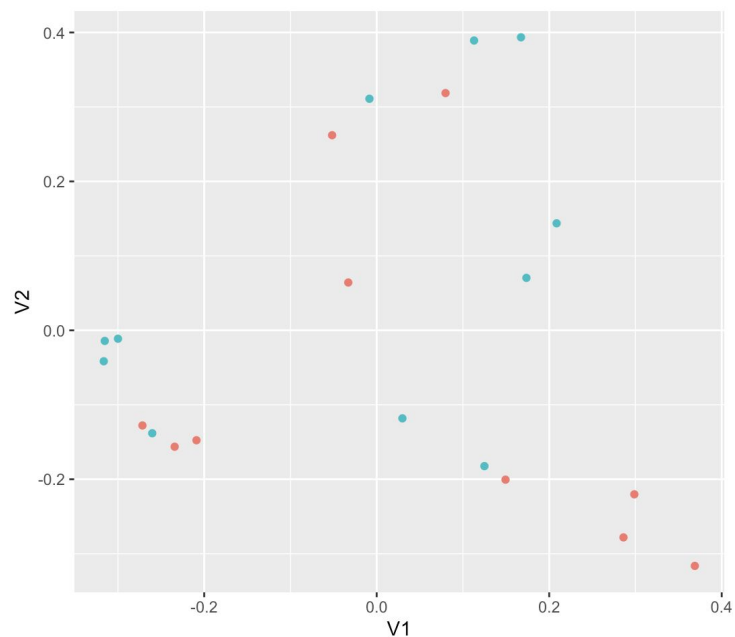
Biological signals

We have focused on modeling and estimating biological effects, such as

- Cell cycle in yeast gene expression data
- Genetic strains and their impacts on mouse RNA-seq data
- Population structure estimated from SNP data

In the last week, looking at the Bottomly et al. (2011) data using 2 distinct mouse genetic strains, we noticed that “experiment numbers” can segregate the scatter plot of PC1 vs. PC2.

Scatter plot of PC1 vs. PC2 in Bottomly et al.



PC1 and PC2 explain more than 90% of variance in the data!

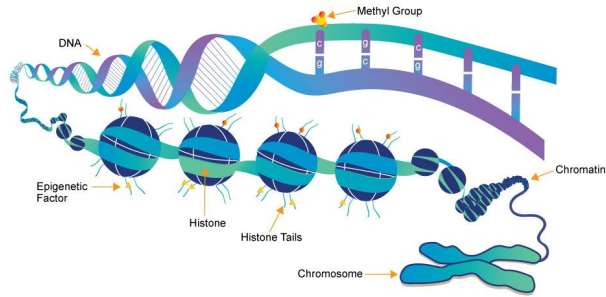
Batch effects

Unwanted variation due to suboptimal study designs and technical factors (e.g., experiment numbers and sequencing lanes in Bottomly data) cause BATCH EFFECTS

Sometimes it is inevitable that some batch effects exist even with the best efforts in balanced and randomized design.

Let's look at how batch effects have reduced the impacts and biological implications of major studies

Sources of batch effects



Genetics/epigenetics



Environments/Experimental

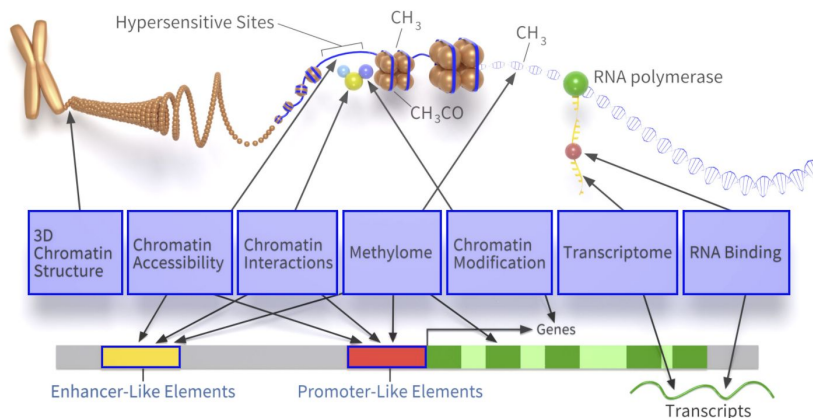


Technical

Case Study 1

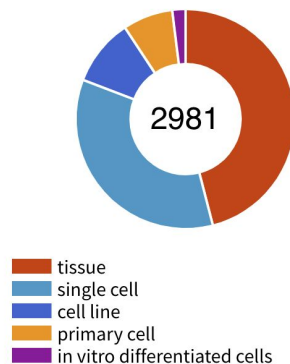
Comparison of the transcriptional landscapes between human and mouse tissues

Mouse ENCODE (Encyclopedia of DNA Elements), Lin et al. (2014) Proceedings of the National Academy of Sciences of the United States of America (IF=11)



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Biosample Type



Assay Categories

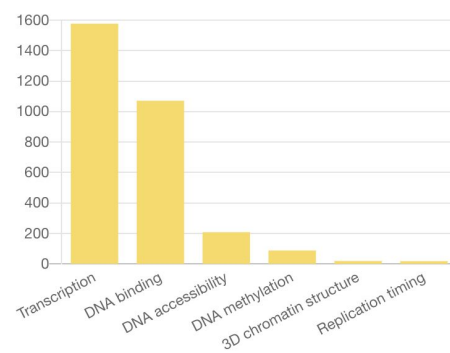


Image from <https://www.encodeproject.org>

Mouse ENCODE

To study gene expression levels, the Consortium collected RNA sequencing data from multiple tissues from human and mouse.

13 human tissues, ENCODE

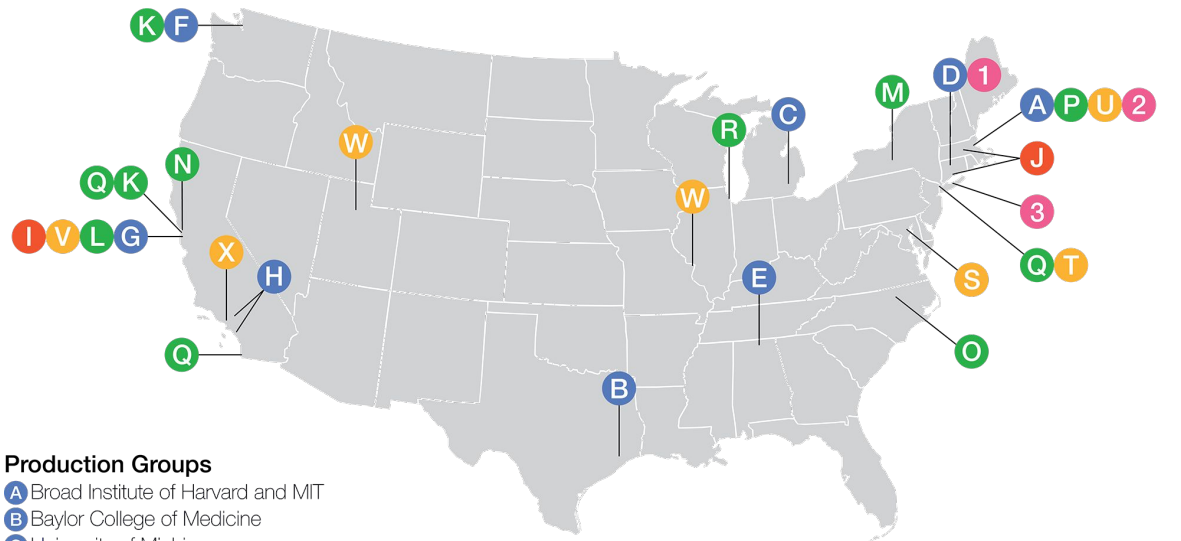
11 human tissues, REMC

13 mouse tissues, mouse ENCODE

Human BodyMap 2.0 (HBM)

In total, “93 datasets encompassing the most tissue-diverse RNA-seq dataset to date”

Additional 294 RNA-seq datasets from the Genotype-Tissue Expression (GTEx) project



Production Groups

- A** Broad Institute of Harvard and MIT
- B** Baylor College of Medicine
- C** University of Michigan
- D** The Jackson Laboratory
- E** HudsonAlpha Institute for Biotechnology, University of Alabama in Huntsville
- F** Altius Institute for Biomedical Sciences
- G** Stanford University
- H** California Institute of Technology, University of California, Irvine

Data Coordination Center

- I** Stanford University

Data Analysis Center

- J** University of Massachusetts Medical School; Yale University

Characterization Centers

- K** University of California, San Francisco; University of Washington
- L** Stanford University
- M** Cornell University
- N** Lawrence Berkeley National Laboratory
- O** Duke University
- P** Broad Institute of Harvard and MIT
- Q** University of California, San Francisco; University of California, San Diego; Ludwig Institute for Cancer Research
- R** University of Chicago

Computational Analysis Groups

- S** Johns Hopkins University
- T** Memorial Sloan Kettering Cancer Center
- U** Harvard University; Brigham and Women's Hospital
- V** Stanford University
- W** Washington University; University of Utah
- X** University of California, Los Angeles

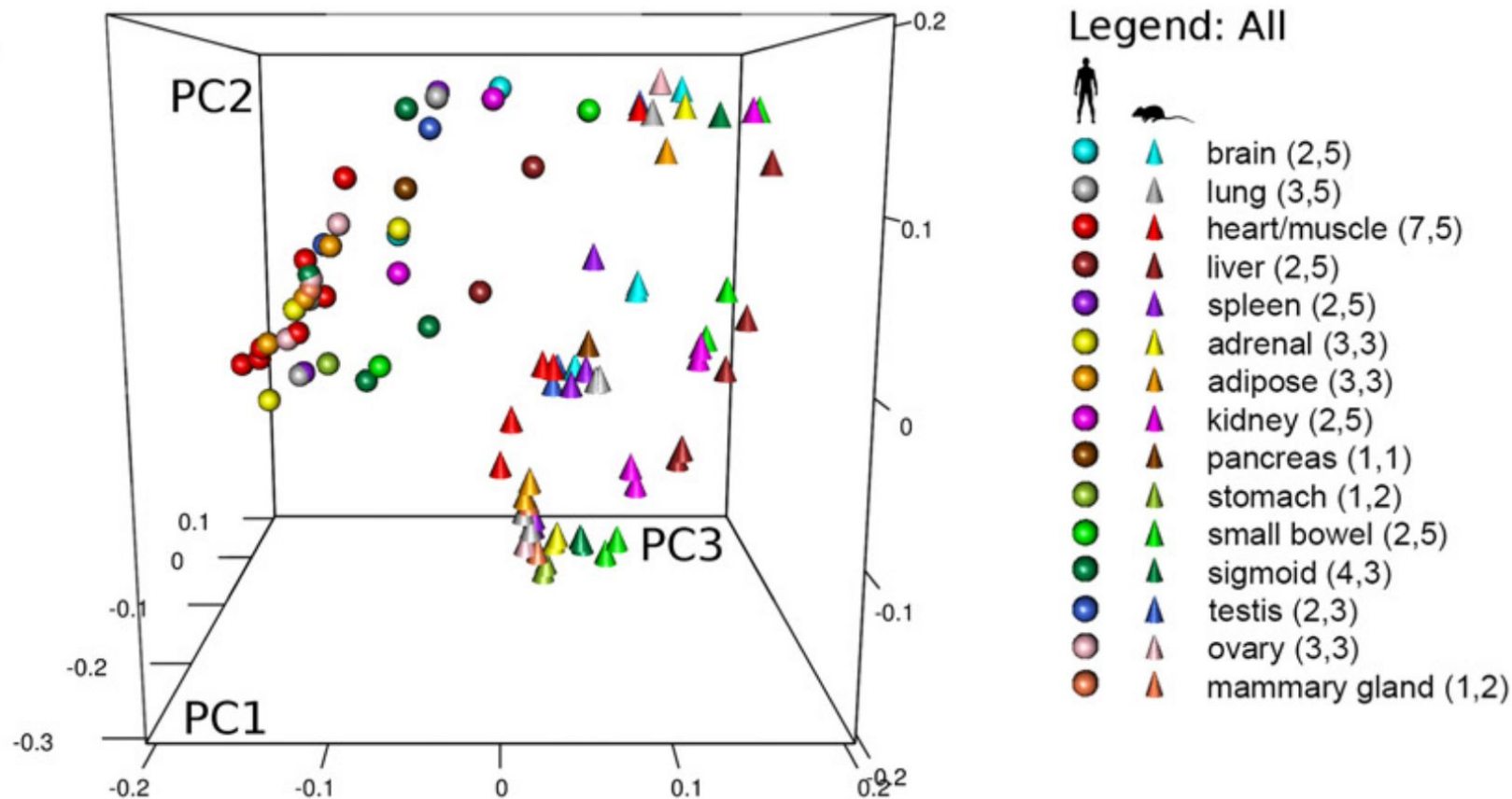
Affiliated Groups

- 1** University of Connecticut Health Center
- 2** Dana-Farber Cancer Institute
- 3** Cold Spring Harbor Laboratory

Mouse ENCODE abstract

[We] performed a comparison of the expression profiles of 15 tissues by deep RNA sequencing and examined the similarities and differences in the transcriptome for both protein-coding and -noncoding transcripts. Although commonalities are evident in the expression of tissue-specific genes between the two species, **the expression for many sets of genes was found to be more similar in different tissues within the same species than between species.** These findings were further corroborated by associated epigenetic histone mark analyses. We also find that many noncoding transcripts are expressed at a low level and are not detectable at appreciable levels across individuals. Moreover, **the majority lack obvious sequence homologs between species, even when we restrict our attention to those which are most highly reproducible across biological replicates.** Overall, our results indicate that there is considerable RNA expression diversity between humans and mice, well beyond what was described previously, likely **reflecting the fundamental physiological differences between these two organisms.**

A



Loading plots from PCA on human and mouse gene expression data. (A) **PCA is performed on the combined Stanford (human, mouse), Salk (human), HBM (human), LICR (mouse), and CSHL (mouse) expression datasets using 15 tissue type**

Clustering by species vs by tissue

This research, supported by an international team and numerous fundings, caused a big controversial in the community.

WHY?

Clustering by species vs by tissue

This research, supported by an international team and numerous fundings, caused a big controversial in the community.

WHY?

Homology: the central concept for all of biology (Wake, 1994 in Science)

Generally, modern biology is built upon the empirical observation that **homologous gene regulatory networks establish the identities of homologous cell-types, tissues, and organs across species**

Manfred D. Laubichler (2000) Homology in Development and the Development of the Homology Concept.

<https://academic.oup.com/icb/article/40/5/777/157228>

Gilad and Mizrahi-Man (2015) F1000Research

Recently, the Mouse ENCODE Consortium reported that comparative gene expression data from human and mouse tend to cluster more by species rather than by tissue. This observation was surprising, as it contradicted much of the comparative gene regulatory data collected previously, as well as the common notion that major developmental pathways are highly conserved across a wide range of species, in particular across mammals.

Here we show that the Mouse ENCODE gene expression data were collected using a flawed study design, which confounded sequencing batch [...] with species.

When we account for the batch effect, the corrected comparative gene expression data from human and mouse tend to cluster by tissue, not by species.

Reconstructed study design

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● Human
testis		pancreas		● Mouse

Sequencing batches as inferred based on the sequence identifiers of the RNA-Seq reads.

Without accounting for batch effects

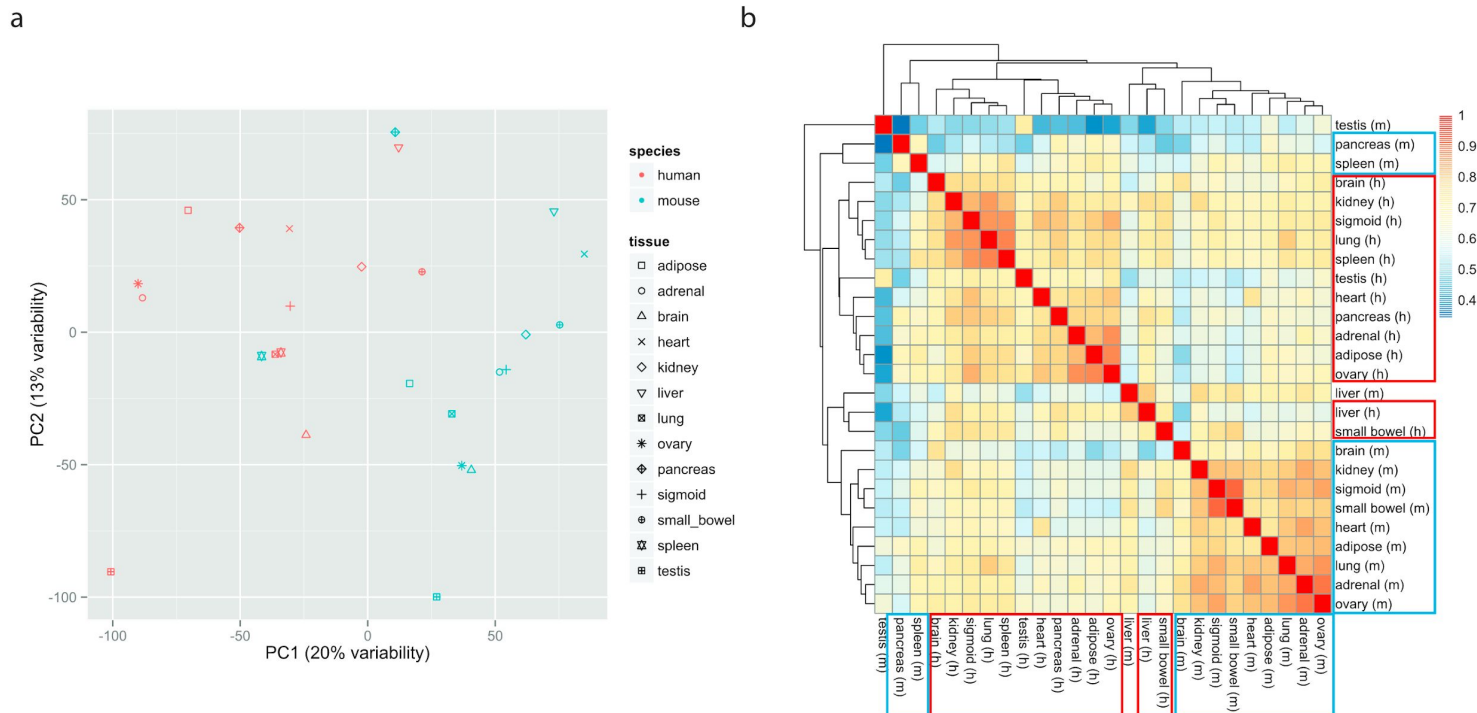


Figure 2. Recapitulating the patterns reported by the mouse ENCODE papers. **a.** Two-dimensional plots of principal components calculated by performing PCA of the transposed log-transformed FPKM values (from 14,744 orthologous gene pairs) for the 26 samples, after removal of invariant columns (genes). **b.** Heatmap based on pairwise Pearson correlation of expression data used in panel **a**. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.

Gilad and Mizrahi-Man method

1. Remove the 30% of genes with the lowest expression
2. Remove reads that map to the 12 mitochondrial genes
3. Remove the GC bias (human vs. mouse), by within-column normalization
4. Normalize for the library sizes for the samples using the trimmed mean of M-values
5. Log2-transformation
6. Account for the study design by fitting 'ComBat' with batches, species and tissue.

Codes and commands are all available: <http://dx.doi.org/10.5281/zenodo.17606>

With accounting for batch effects

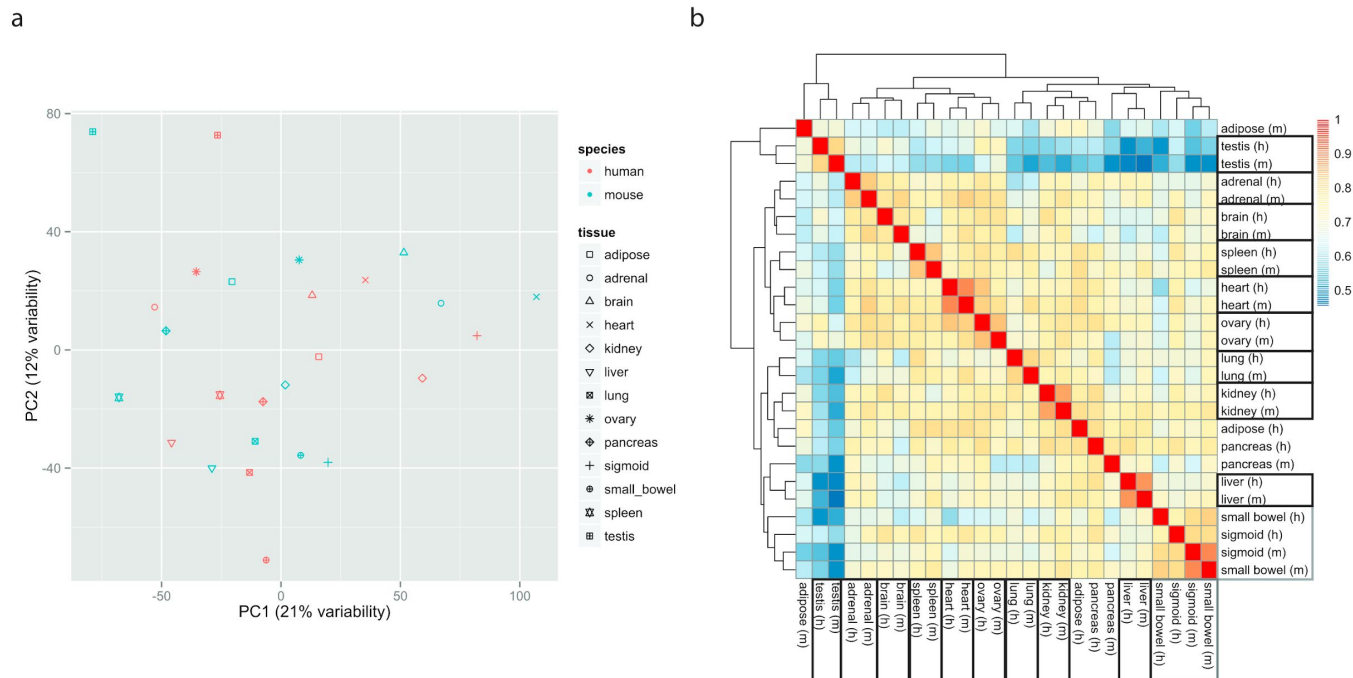


Figure 3. Clustering of data once batch effects are accounted for. **a.** Two-dimensional plots of principal components calculated by applying PCA to the transposed matrix of batch-corrected log-transformed normalized fragment counts (from 10,309 orthologous gene pairs that remained after the exclusion steps described in the results) for the 26 samples, after removal of invariant columns (genes). **b.** Heatmap based on pairwise Pearson correlation of the expression data used in panel **a**. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.

Discussion on this study designs

See post-publication reviews on <https://f1000research.com/articles/4-121>



Lin: Provide feedbacks & additional experiments in support of the original claims

Salzberg: Because this batch effect is almost completely confounded with the main effect reported (the clustering by species), it's nearly impossible to separate the two.

Gilad: Our principal claim is NOT that the data cluster by tissue rather than by species.

Rather, our claim is that your study design does not allow you to address the question of whether the data cluster better by tissue or species.

Codes and commands are all available



UploadCommunities

Log inSign up

May 14, 2015


SoftwareOpen Access

Data files and codes used in the reanalysis of the mouse encode comparative gene expression data


Orna Mizrahi-Man; Yoav Gilad

We provide supplementary files of the python codes used to process and prepare the data for analysis with R, and the data files for the python codes. We also provide the R codes we used to perform the different analyses as supplementary files, as well as the input for the R codes. Please see supplementary text files for more details.


Preview

 R_input_files.zip


R_input_files

 .Rhistory


29.2 kB

 Stanford_datasets.txt


1.3 kB

 Stanford_datasets_fpkMat.txt

2.9 MB

 Stanford_datasets_rawCountsMat.txt

1.3 MB

 ortholog_GC_table.txt

456.5 kB

<http://dx.doi.org/10.5281/zenodo.17606>

837

views

191

downloads

[See more details...](#)

Indexed in

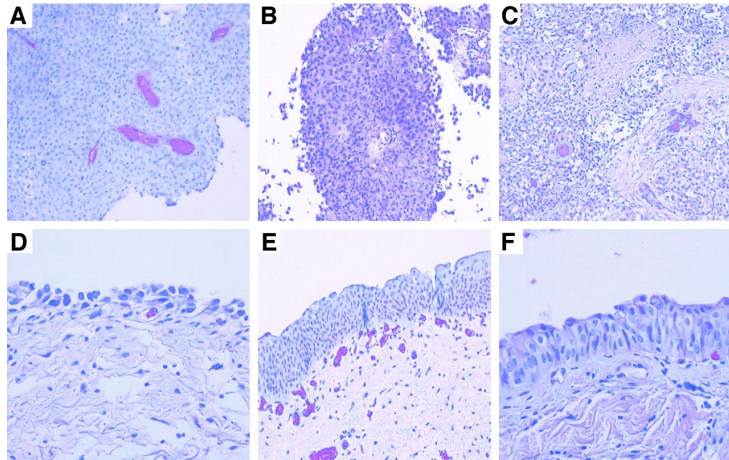
**Publication date:**
May 14, 2015**DOI:**
[DOI 10.5281/zenodo.17606](https://doi.org/10.5281/zenodo.17606)**Keyword(s):**
[ENCODE](#) [RNA-Seq](#)**Published in:**
F1000Research: 4 (2015).**Related identifiers:**
Supplement to
[10.12688/f1000research.6536.1](https://doi.org/10.12688/f1000research.6536.1)**Communities:**
[F1000Research](#)**License (for files):**
[MIT License](#)

Case Study 2

Gene expression in the urinary bladder: a common carcinoma in situ (CIS) gene expression signature exists disregarding histopathological classification.

Dyrskjot, L. et al. (2004) Cancer Research

The second-most frequently cited cancer journal in the world, with an impact factor of > 11



Stained tissue sections showing typical histological appearance of the different groups of samples used.

A, superficial Ta tumor from a bladder with no CIS

B, superficial Ta tumor from a bladder with surrounding CIS

C, muscle invasive T2+ tumor

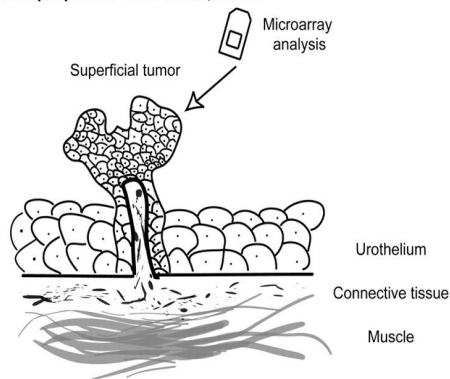
D, biopsy with CIS lesion from a bladder removed by cystectomy

E, biopsy with normal appearing urothelial cells from a bladder with CIS lesions removed by cystectomy

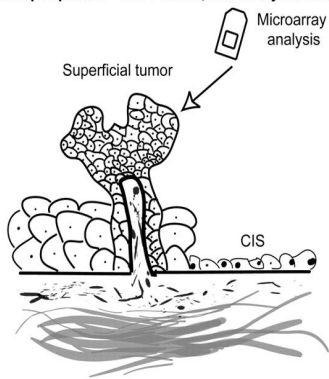
F, biopsy from a normal bladder showing normal urothelium, no bladder cancer history.

Motivation: The presence of carcinoma in situ (CIS) lesions in the urinary bladder is associated with a high risk of disease progression to a muscle invasive stage. Non-invasive urinalysis assessment strategies would benefit patients.

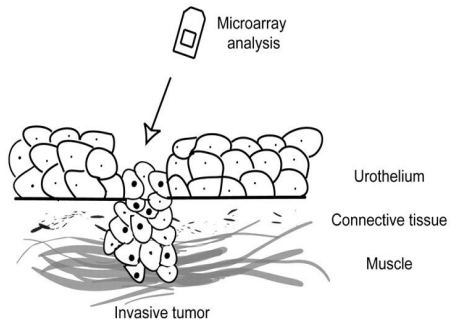
Group A patient - Ta/T1 tumor, no CIS



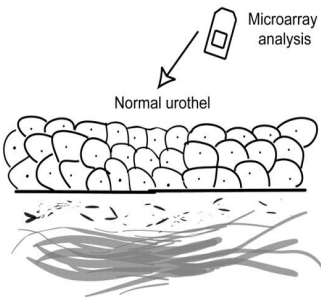
Group B patient - Ta/T1 tumor, CIS in adjacent mucosa



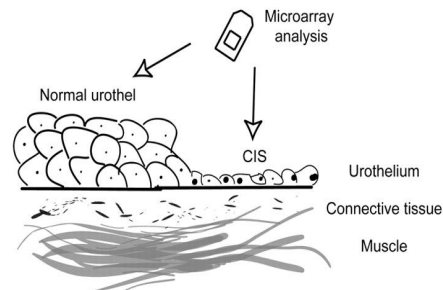
Group C patient - Invasive tumor



Group D patient - Normal urothelium from patient with no bladder cancer history



Group E patient - Cystectomy specimen from patient with CIS



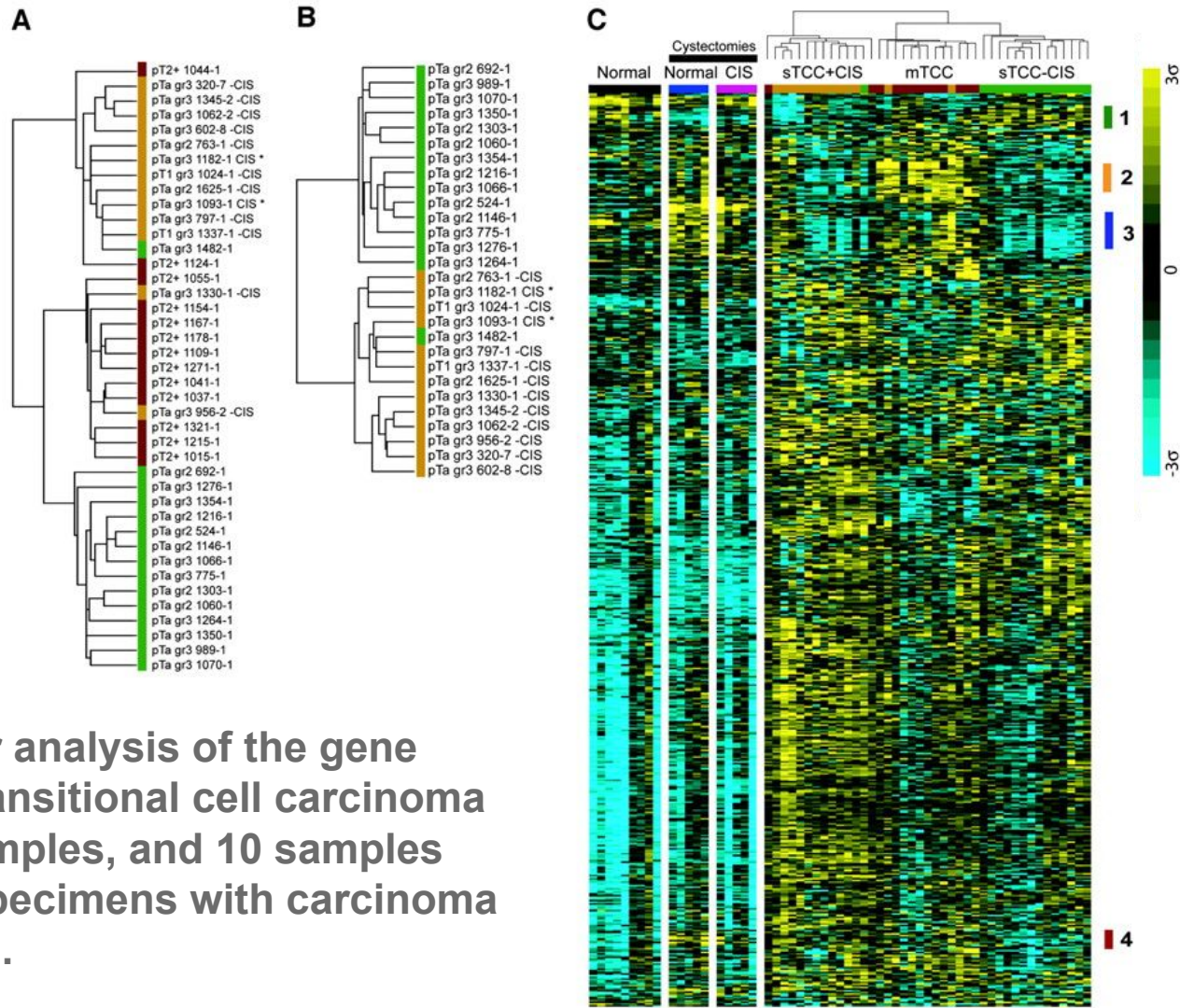
Experiments: Gene expression microarray analysis to exfoliated urothelia recovered from bladder washes obtained from 46 patients with subsequently confirmed presence or absence of bladder cancer.

Urothelia: a specialized type of tissue inside of an urinary tract

Dyrskjot, L. et al. (2004) Cancer Res

Analysis: Data from microarrays containing 56,000 targets was subjected to a panel of statistical analyses to identify bladder cancer-associated gene signatures. Hierarchical clustering and supervised learning algorithms were used to classify samples on the basis of tumor burden.

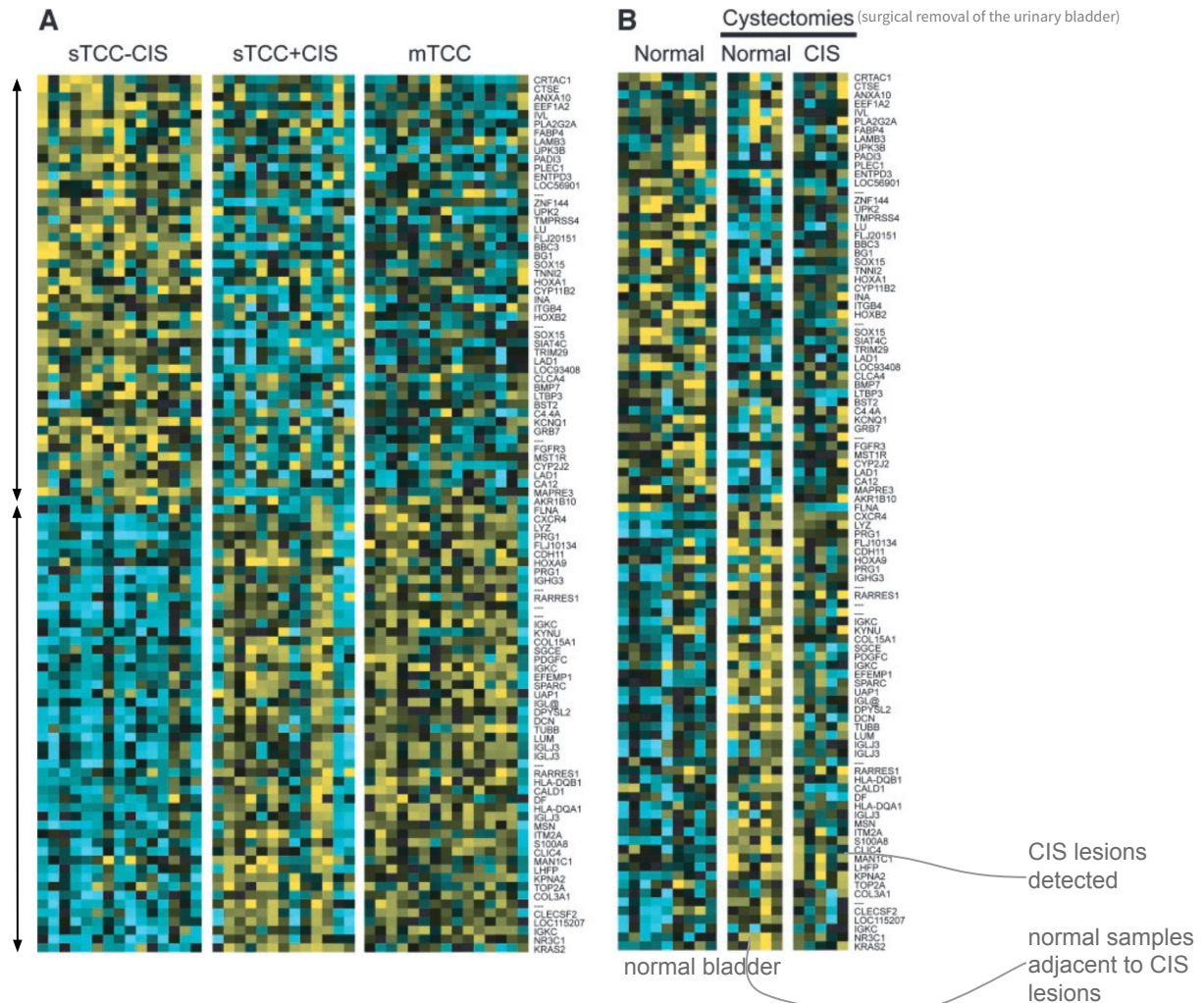
Results: A differentially expressed geneset of 319 gene probes was associated with the presence of bladder cancer ($P < 0.01$), and visualization of protein interaction networks revealed VEGF and AGT as pivotal factors in tumor cells. Supervised machine learning and a cross-validation approach were used to build a **14-gene molecular classifier** that was able to classify patients with and without bladder cancer with an overall **accuracy of 76%**.



Hierarchical cluster analysis of the gene expression in 41 transitional cell carcinoma (TCC), 9 normal samples, and 10 samples from cystectomy specimens with carcinoma *in situ* (CIS) lesions.

the 50 best
up-regulated marker
genes in sTCC
without CIS are
shown at *top*

the 50 best
up-regulated marker
genes in transitional
cell carcinoma with
CIS



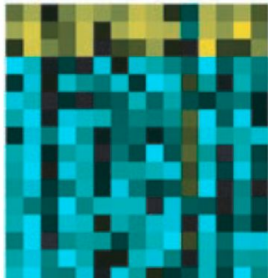
Construction of a Molecular CIS Classifier

Our first approach was to try to classify sTCC with or without CIS in the surrounding mucosa, based on tissue from the sTCC. The best classifier performance (one error) was obtained in cross-validation loops using 25 genes

16 of these were included in 70% of the cross-validation loops, and these were selected to represent our final classifier for CIS diagnosis. Permutation analysis showed that 13 of these were significant at a 1% confidence level; the remaining 3 genes were above a 10% confidence level.

A

sTCC-CIS



sTCC+CIS



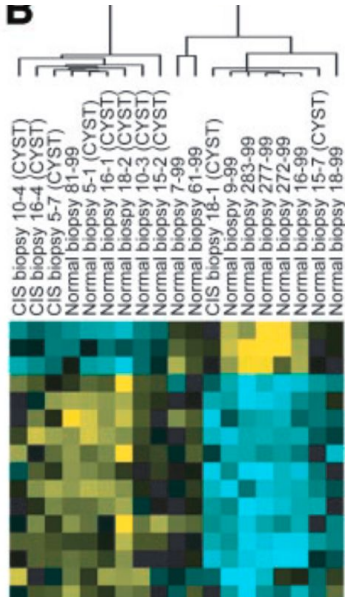
SH3BP1
CIC
MINK
DKFZP434C212
SHOC2

FBXL5
MBD4
PPP2R5C
ERBB2IP
ARL5
IL13RA1
SDCBP
BIRC2
SPOP
KIAA1028

28
28
23
28
28
28
28
28
28
28
28
28
26
22
21

CV-loops

B

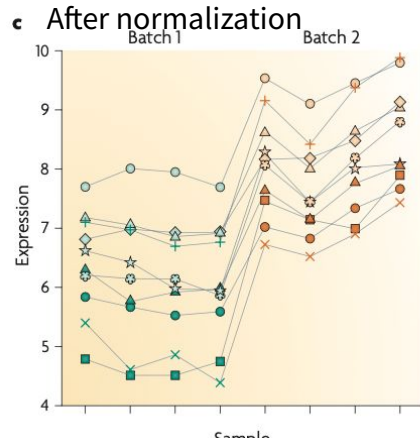
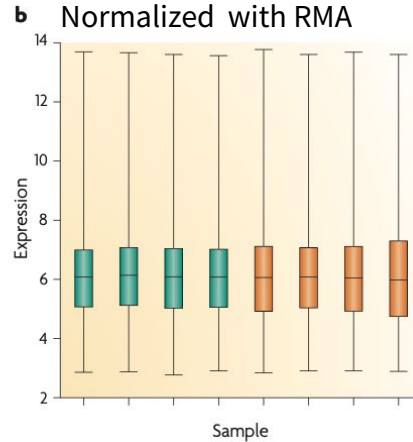
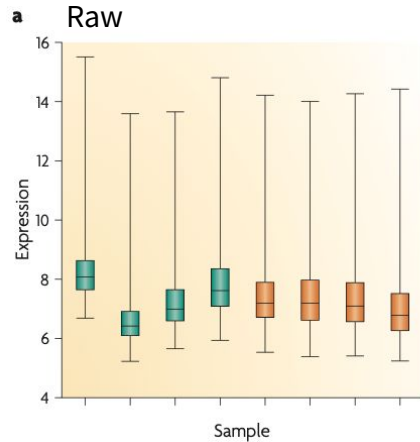


SH3BP1
CIC
MINK
DKFZP434C212
SHOC2

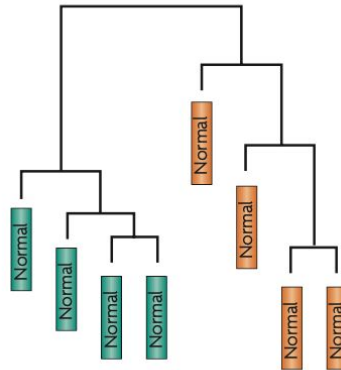
FBXL5
MBD4
PPP2R5C
ERBB2IP
ARL5
IL13RA1
SDCBP
BIRC2
SPOP
KIAA1028

Batch effects in Dyrskjot, L. et al. (2004)

Leek et al. 2010 Tackling the widespread and critical impact of batch effects in high-throughput data



d Clustering of samples after normalization



The raw data for only the normal samples. Here, green and orange represent **two different processing dates**. a | Box plot of raw gene expression data (log base 2). b | Box plot of data processed with RMA, a widely used preprocessing algorithm for Affymetrix data. RMA applies quantile normalization — a technique that forces the distribution of the raw signal intensities from the microarray data to be the same in all samples. c | Example of ten genes that are susceptible to batch effects even after normalization. d | Clustering of samples after normalization. Note that the samples perfectly cluster by processing date.

Leek et al. 2010 Tackling the widespread and critical impact of batch effects in high-throughput data

Table 1 | **Batch effects seen for a range of high-throughput technologies**

Study description*	Known variable used as a surrogate			Principal components used as a surrogate			Association with outcome Significant features (%) ^{††}	Refs
	Surrogate [‡]	Confounding (%) [§]	Susceptible features (%)	Principal components rank of surrogate (correlation) [¶]	Principal components rank of outcome (correlation) [#]	Susceptible features (%) ^{**}		
Data set 1: gene expression microarray, Affymetrix ($N_p = 22,283$)	Date	29.7	50.5	1 (0.570)	1 (0.649)	91.6	71.9	9
Data set 2: gene expression, Affymetrix ($N_p = 4167$)	Date	77.6	73.7	1 (0.922)	1 (0.668)	98.5	62.2	2
Data set 3: mass spectrometry ($N_p = 15,154$)	Processing group	100	51.7	2 (0.344)	2 (0.344)	99.7	51.7	3
Data set 4: copy number variation, Affymetrix ($N_p = 945,806$)	Date	29.2	99.5	2 (0.921)	3 (0.485)	99.8	98.8	16
Data set 5: copy number variation, Affymetrix ($N_p = 945,806$)	Date	12.2	83.8	1 (0.553)	1 (0.137)	99.8	74.1	17
Data set 6: gene expression, Affymetrix ($N_p = 22,277$)	Processing group	NA	83.8	5 (0.369)	NA	97.1	NA	18
Data set 7: gene expression, Agilent ($N_p = 17,594$)	Date	NA	62.8	2 (0.248)	NA	96.7	NA	18
Data set 8: DNA methylation, Agilent ($N_p = 27,578$)	Processing group	NA	78.6	3 (0.381)	NA	99.8	NA	18
Data set 9: DNA sequencing, Solexa ($N_p = 2,886$)	Date	24.2	32.1	2 (0.846)	2 (0.213)	72.7	16.9	1000 Genomes Project

The first three rows represent studies for which batch effects have been described in the literature.

Rows four and five are from genome-wide association study data sets.

Rows six to eight represent data from The Cancer Genome Atlas (TCGA).

Finally, the last row represents second-generation sequencing data from the 1000 Genomes Project.

How to account for batch effects?

Exploration

Visualize

Modeling

Estimation

Overall, there are MANY methods. We look at 2 most popular methods.

Exploratory analyses

Hierarchically cluster the samples and label them with biological variables and batch surrogates (such as laboratory and processing time)



Plot individual features versus biological variables and batch surrogates



Calculate principal components of the high-throughput data and identify components that correlate with batch surrogates

Downstream analyses

Do you believe that measured batch surrogates (processing time, laboratory, etc.) represent the only potential artefacts in the data?

Yes

Use measured technical variables as surrogates for batch and other technical artefacts

No

Estimate artefacts from the high-throughput data directly using surrogate variable analysis (SVA)

Perform downstream analyses, such as regressions, t-tests or clustering, and adjust for surrogate or estimated batch effects. The estimated/surrogate variables should be treated as standard covariates, such as sex or age, in subsequent analyses or adjusted for use with tools such as ComBat

Diagnostic analyses

Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects

Linear model

Conventionally, a linear model is used for relating a biological variable and an observed genomic data:

$$Y = BX + E$$

Y = Observed genomic data, containing m variables (rows) and n observations (cols)

X = Biological variables

E = Independently and identically distributed (i.i.d.) noise

Linear model, with dependent noise

$$Y = BX + E$$

When we fail to model batch effects in this linear model, there is a dependence across the noise term E (e.g., no longer i.i.d.)

If we know the technical variable, we may include them as covariates in a linear model.

Linear model, with technical variables

$$Y = BX + E$$

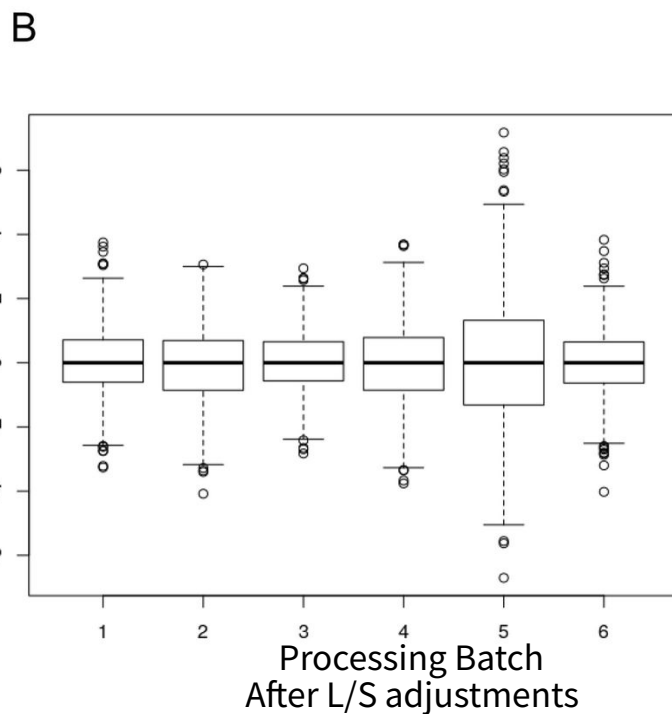
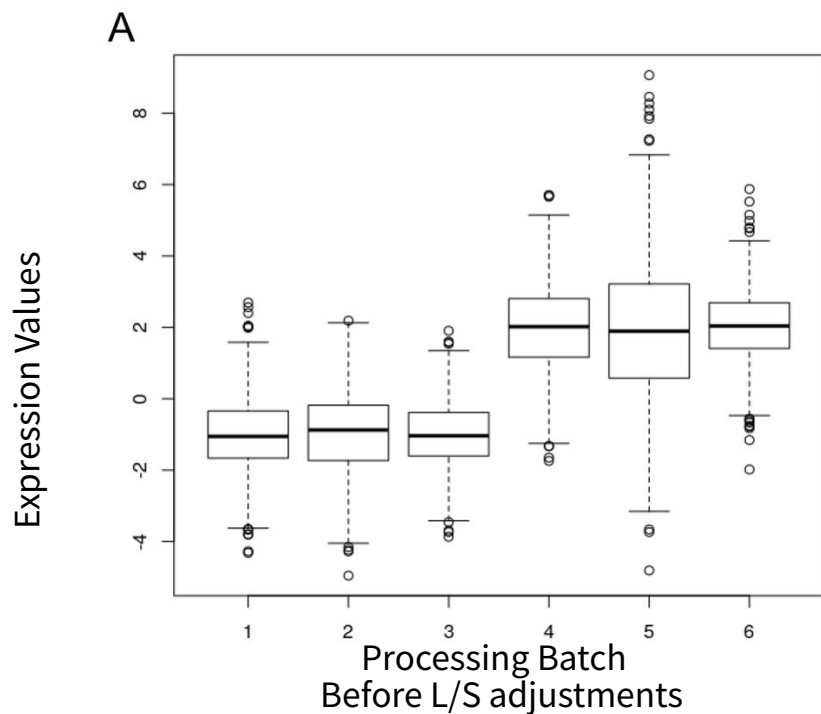
$$= BX + \Gamma G + U$$

X = Biological variables

G = Technical variables

U = i.i.d. Noise

Boxplots of data, within each batch



Location and scale (L/S) adjustments

A wide family of adjustments in which one assumes a model for the location (mean) and/or scale (variance) of the data within batches and then adjusts the batches to meet assumed model specifications.

Therefore, the batch effects can be modeled out by standardizing means and variances across batches

The simplest approach for L/S batch adjustment is to mean center and standardize the variance of each batch for each gene independently.

In more complex situations such as unbalanced designs or when incorporating numerical covariates, use a general L/S framework:

General L/S framework

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

Y_{ijg} : observed data -- the expression value for gene g for sample j from batch i

Contain m batches containing n_i samples within batch i for $i=1,\dots,m$, for gene and $g=1,\dots,G$

X : biological variables -- a design matrix for sample conditions

β_g : regression coefficients corresponding to X

γ_{ig} : additive batch effects of batch i for gene g

δ_{ig} : multiplicative batch effects of batch i for gene g

ε : noise with mean zero and variance σ_g^2

Empirical Bayes approach with ComBat

The most important disadvantage of many existing methods is that large batch sizes are required for implementation because such methods are not robust to outliers.

ComBat make this possible even for a smaller sample size by:

1. Estimating the L/S model parameters that represent the batch effects by “pooling information” across genes in each batch
2. Shrinking the batch effect parameter estimates toward the overall mean of the batch effect estimates (across genes)

ComBat algorithm

1. Standardize the data

Standardize gene-wise so that genes have similar overall mean and variance

Standardized data, Z_{ijg} , satisfy the distributional form, $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$

2. Batch effect parameter estimates using parametric empirical priors

Johnson et al (2007) uses the following priors $\gamma_{ig} \sim N(Y_i, \tau_i^2)$ and $\delta_{ig}^2 \sim \text{Inverse Gamma}(\lambda_i, \theta_i)$

Those hyperparameters are estimated from the standardized data, Z_{ijg}

Then, the posteriors are

$$\gamma_{ig}^* = \frac{n_i \bar{\tau}_i^2 \hat{\gamma}_{ig} + \delta_{ig}^{2*} \bar{\gamma}_i}{n_i \bar{\tau}_i^2 + \delta_{ig}^{2*}} \quad \text{and} \quad \delta_{ig}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2} \sum_j (Z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_i}{2} + \bar{\lambda}_i - 1}$$

3. Adjust the data for batch effects with

$$\gamma_{ijg}^* = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*} (Z_{ijg} - \hat{\gamma}_{ig}^*) + \hat{\alpha}_g + X \hat{\beta}_g$$

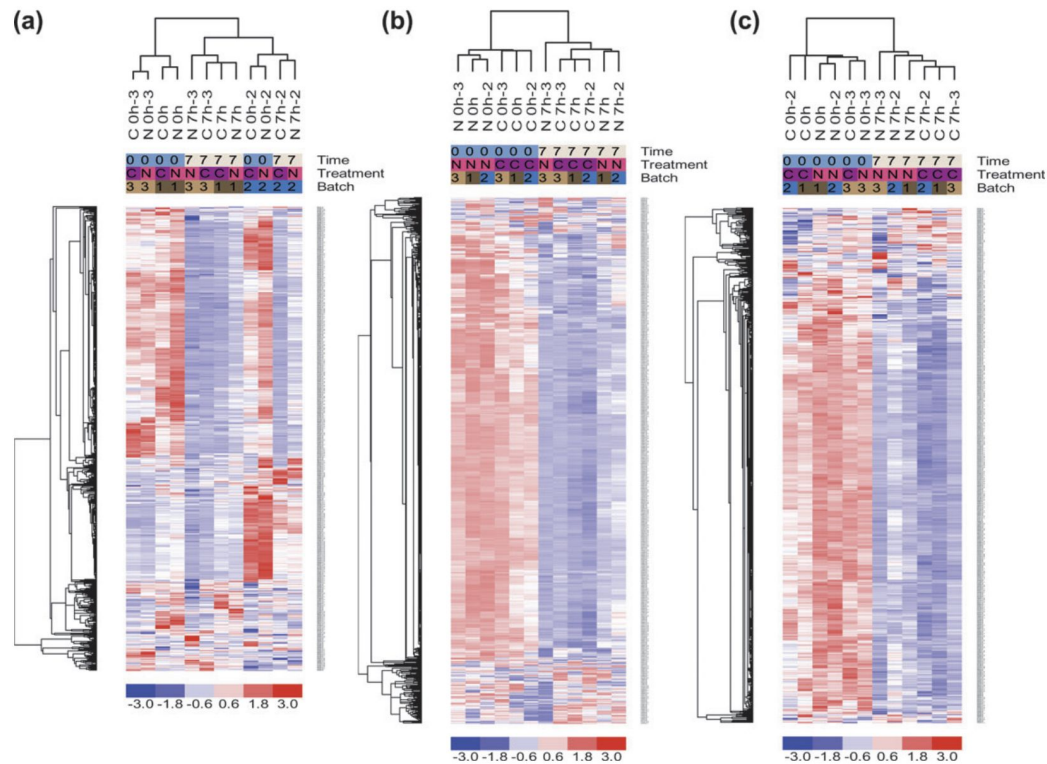


Fig. 1. Heat map clusterings for data set 1. The gene-wise expression values are used to compute gene and sample correlations and displayed in color scale, and the sample legends on the top are 0 (0 h), 7 (7.5 h), C (Control), and N (NO treated). (a) Expression for 628 genes with large variation across all the 12 samples. Note that the samples from the batch 2 cluster together and the baseline (time = 0) samples also cluster by batch 1 and 3; (b) 720 genes after applying “standardized separators” (which standardize each gene within each batch to have a mean 0 and variance of 1) for gene filtering and clustering in the dChip software; (c) 692 genes after applying the EB batch adjustments and then filtered for clustering. Note that there is no strong evidence of batch effects after adjustment in heat maps (b)–(c). The EB adjustment in (c) has the advantage of being robust to outliers in small sample sizes.

Shrinkage

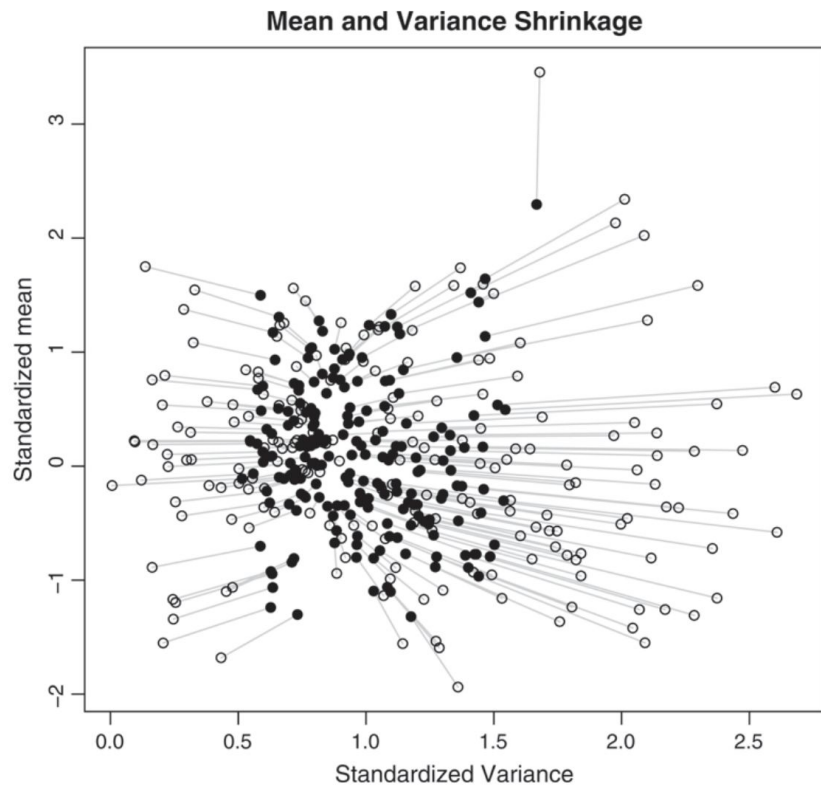


Fig. 3. Shrinkage plot for the first 200 probes from one of the batches in data set 1. The gene-wise and EB estimates of γ_{ig} and δ_{ig}^2 in Section 3.1 are plotted on the Y and X axis. Open circles are the gene-wise values and the solid are after applying the EB shrinkage adjustment.