Dealing with batch effects with high-throughput genomic data

I-Shou Chang

National Cancer Institute, Division of Biostatistics and Bioinformatics, and Center for Biomedical Resources
National Health Research Institutes

Batch effects

- Accuracy of measurements depend on reagents, hardware, highly trained personnel.
- In high-throughput experiments, many of quantities being measured are **simultaneously** affected by both biological and non-biological factors.
- Batch effects are **subgroups** of measurements that have **qualitatively** different behavior across conditions and are **unrelated** to the biological variables in a study.

Relief and Confidence

- Although batch effects are difficult to detect in low-dimensional assay, high-throughput technologies provide the opportunity to detect and remove them.

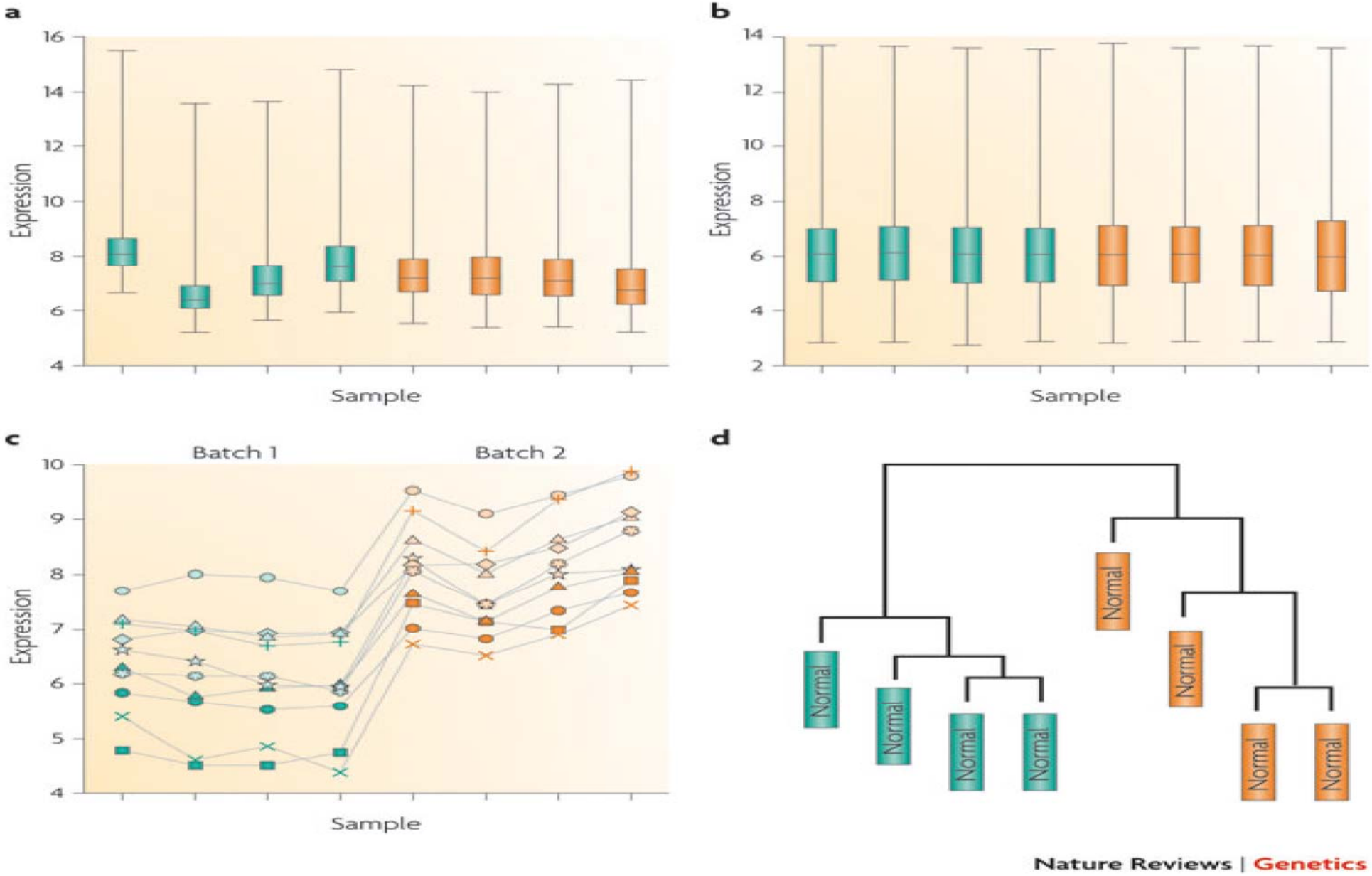
Systematic bias in microarray

- sample preparation, hybridization, measurement of expression;
- batch to batch variation in array manufacture;
- Day to day variation in laboratory conditions

Gene expressions correlated with processing date (I)

- Dyrskjot, L. *et al*. Gene expression in the urinary bladder: a common carcinoma *in situ* gene expression signature exists disregarding histopathological classification. *Cancer Res.* **64**, 4040–4048 (2004).
- Zilliox, M. J. & Irizarry, R. A. A gene expression bar code for microarray data. *Nature Methods* **4**, 911–913 (2007).

- Microarray expression profiling was used to examine the gene expression patterns in superficial transitional cell carcinoma(sTCC) with or without surrounding carcinoma *in situ* (CIS).
- **Cluster analysis** based on microarray expression data separated the sTCC samples according to the presence or absence of CIS.
- However, the presence or absence of CIS was strongly ***confounded*** with processing date (Zilliox & Irizarry, 2007).



Nature Reviews Genetics 2010 (11), 733-739.

- For a published bladder cancer microarray data set obtained using an Affymetrix platform, we obtained 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. Hundreds of genes show similar behaviour but, for clarity, are not shown. **d** | Clustering of samples after normalization. Note that the samples perfectly cluster by processing date.

Nature Reviews Genetics 2010 (11), 733-739.

- Normalization helps reduce global differences among arrays, does not address batch effects.
- In gene expression studies, the greatest source of differential expression is nearly always across batches rather than across biological groups.

Gene expressions correlated with processing date (II)

- Spielman et al. 2007, Common genetic variants account for differences in gene expression among ethnic groups, Nature Genetics , 39, 2, 226-231.
- Akey et al. 2007, On the design and analysis of gene expression studies in human populations, Nature Genetics, 39, 7, 807-809.

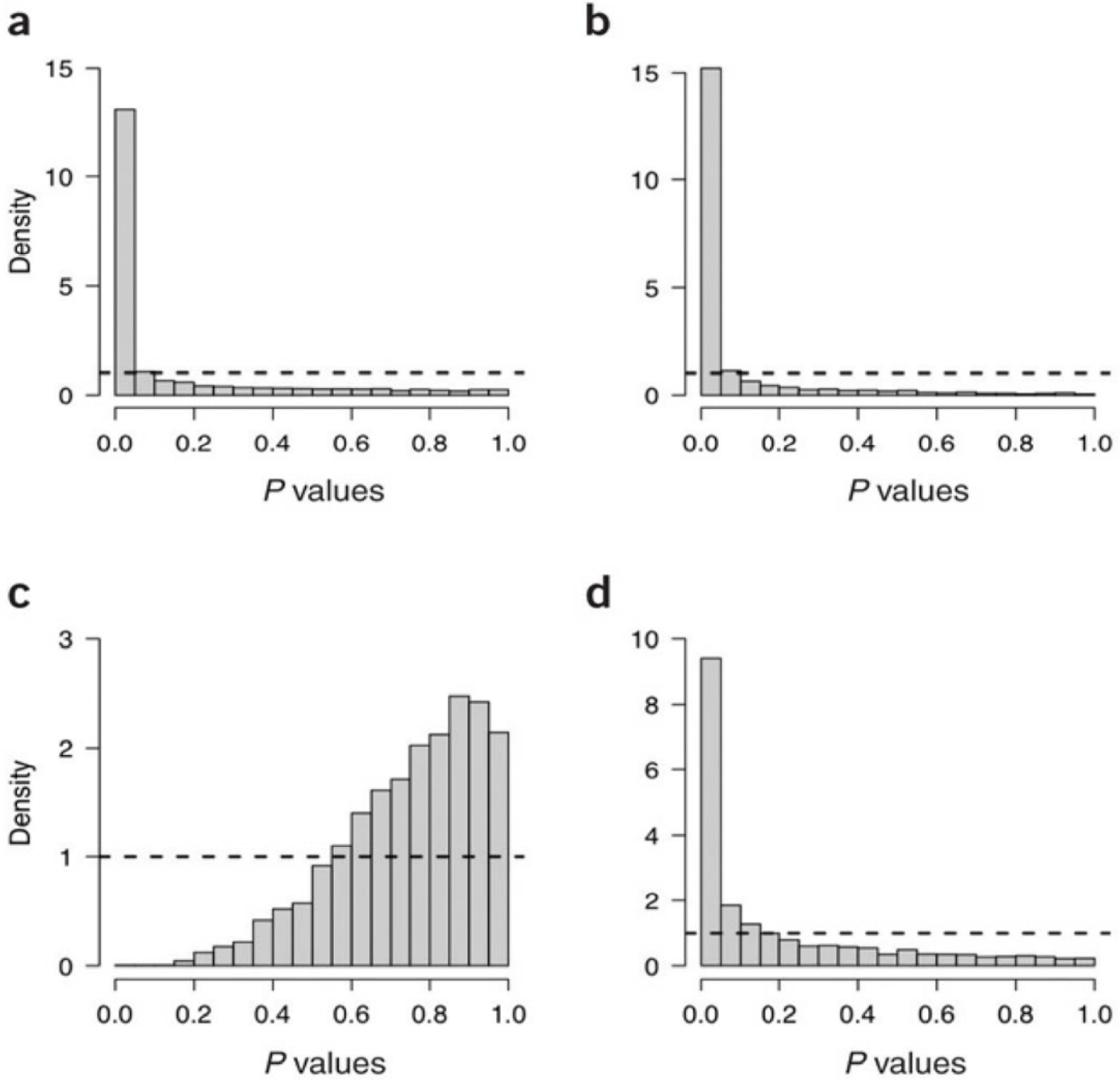
Gene expression, genetic variant, and ethnic group

- Allele frequency differences between populations often have highly significant phenotypic consequences.
- The proportion of gene expression phenotypes differs significantly between populations and to what extent the phenotypic differences are attributable to specific genetic polymorphism.

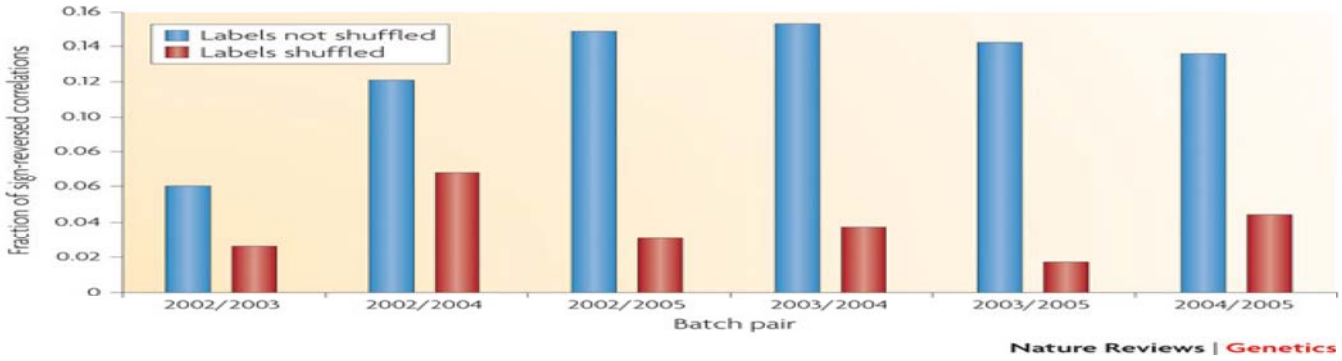
- Between European-derived and Asian-derived populations, expression phenotypes differs significantly for 1097 of 4197 genes at **p-value** less than 10^{-5}, based on cell lines from 60 CEU and 41 CHB and 41 JPT of the **HapMap Project**. (Spielman et al., 2007)
- Storey and coworkers think this is a too stringent criterion. Using the **complete distribution of P-values**, they found the proportion is 78%.

- Storey et al.(2007) found about **17**% of the genes are differentially expressed between individuals of European and African ancestry, based on 8 CEU and 8 Yoruban using the **complete distribution of P-values**.
- For comparison, Storey randomly chose 8 CEU and 8 ASN and estimated the proportion of differentially expressed genes using the **complete distribution of P-values**; they did the study 1000 times and found the average proportion is **43**%(s.e.=8%).

- CEU individuals were primarily processed from 2003 to 2004 and ASN individuals were all in 2005-2006.


a
b
c
d
Density
P values
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
1
2
3
4
6
8

- (**a**) *P* values comparing CEU and ASN samples. (**b**) *P* values comparing samples having different microarrays processing year. (**c**) *P* values comparing CEU and ASN samples, controlling for the sample processing year. (**d**) *P* values comparing samples having different microarrays processing year among the CEU individuals. Under the null hypothesis of no differential expression, we expect the *P* values to be uniformly distributed between 0 and 1, forming a histogram with frequencies following the dashed black line. We estimate the proportion of differentially expressed genes in **a**–**d** to be 78%, 94%, 0% and 79%, respectively. The odd shape of the histogram in **c** is attributable to the almost complete confounding of year of processing and population, illustrating the underlying problem with the study design.



- We normalized every gene in the second gene expression data set in Table 1 to mean 0, variance 1 within each batch. (The 2006 batch was omitted owing to small sample size.) We identified all significant correlations ($p < 0.05$) between pairs of genes within each batch using a linear model. We looked at genes that showed a significant correlation in two batches and counted the fraction of times that the correlation changed between the two batches. A large percentage of significant correlations reversed signs across batches, suggesting that the correlation structure between genes changes substantially across batches. To confirm this phenomenon is due to batch, we repeated the process — looking for significant correlations that changed sign across batches — but with the batch labels randomly permuted. With random batches, a much smaller fraction of significant correlations change signs. This suggests that correlation patterns differ by batch, which would affect rank-based prediction methods as well as system biology approaches that rely on between-gene correlation to estimate pathways.

Nature Reviews Genetics 2010 (11), 733-739.

Some remarks

- Batch effects appear quite frequently.
- Try best to avoid batch effects in the first place.
- In any case, the following data analyses strategies/workflow are recommended.

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

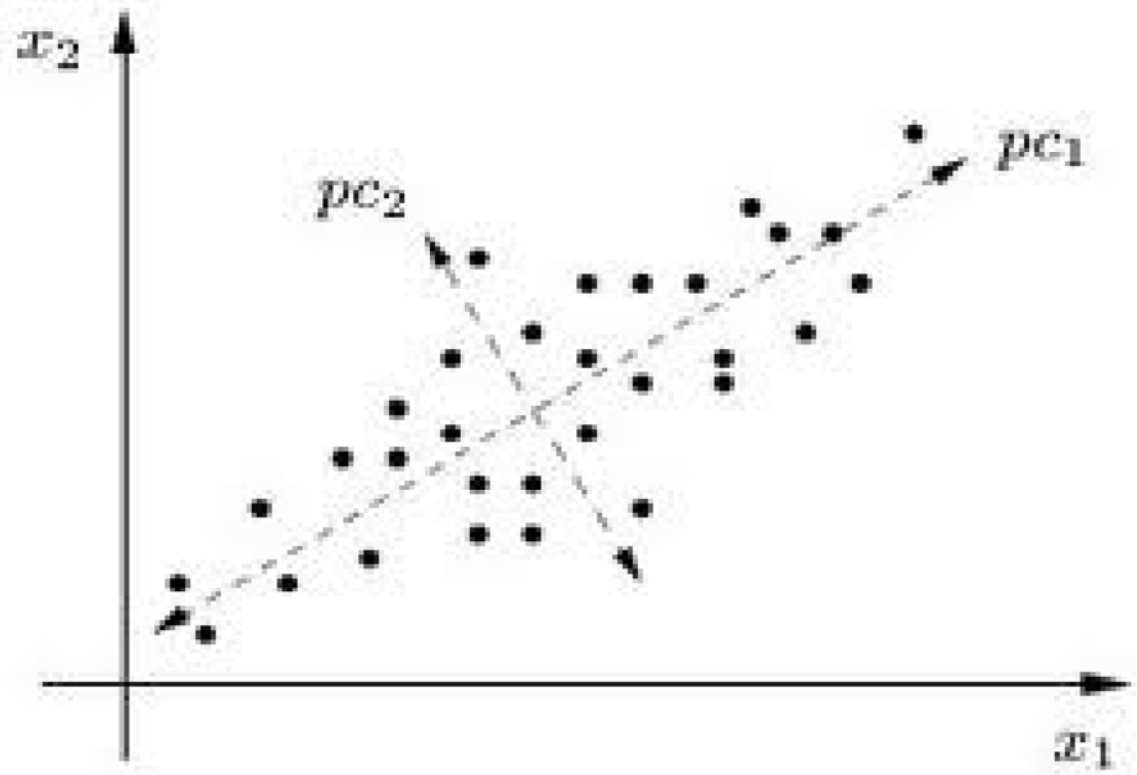
Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects

Nature Reviews Genetics 2010 (11), 733-739.

Principal component analysis

- Change the coordinates to best represent the data, singular value decomposition, analytic geometry
- An important topic in linear algebra and multivariate analysis for **data reduction**
- Detect the **hidden** population substructure in genetic studies
- Detect the **unmeasured** batch effects in expression array

Principal component analysis



Principal component analysis

- A layman’s introduction to PCA, Youtube

Principal component analysis

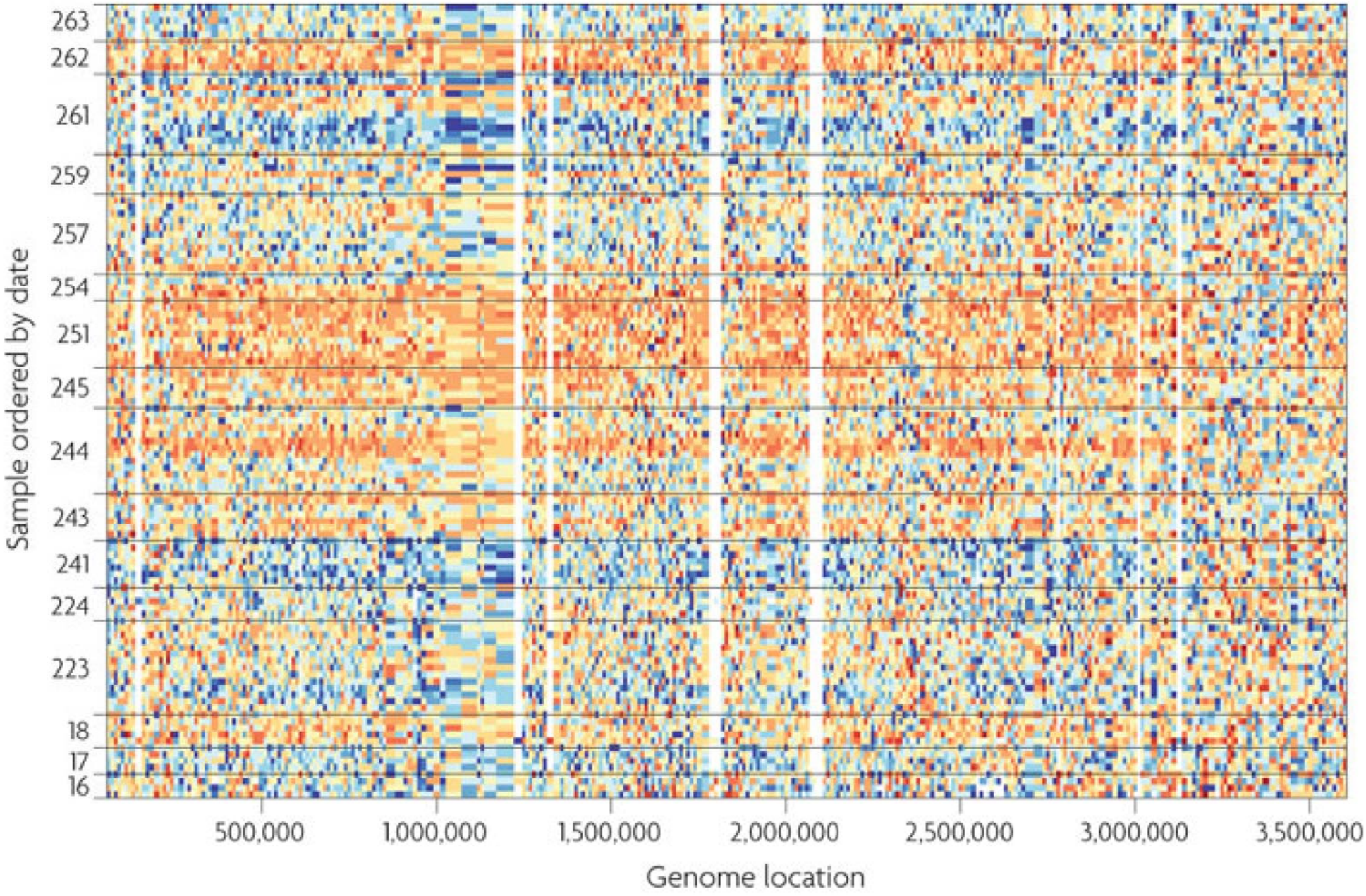
- Jeffrey T. Leek, John D. Storey (2007). Capturing Heterogeneity in Gene Expression Studies by **Surrogate Variable Analysis**, PLoS genetics
- Reich, Price, Patterson (2008), Principal component analysis of genetic data *Nature Genetics* 40, 491 – 492
- Ringnér (2008), What is principal component analysis? *Nature Biotechnology* 26, 303 - 304

- *Nature Reviews Genetics* **11**, 733-739 (October 2010) | doi:10.1038/nrg2825
- **Opinion: Tackling the widespread and critical impact of batch effects in high-throughput data**
- Jeffrey T. Leek, et al.

Study description*	Known variable used as a surrogate			Principal components used as a surrogate			Association with outcome	Refs
	Surrogate‡	Confounding (%)§	Susceptible features (%)‖	Principal components rank of surrogate (correlation)¶	Principal components rank of outcome (correlation)#	Susceptible features (%)**	Significant 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[4,5,10]. 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. Details for each data set and the analyses used to construct the table are included in Supplementary information S1 (box). *Study description includes the application, platform and number of features (N_p). ‡A known variable was used as a surrogate for batch effect. §Level of confounding between surrogate and biological outcome of interest. We use a generalized R^2 statistic for categorical data. The correlation ranges from 0% (no confounding) to 100% (completely confounded). ‖For each feature of the technology (for example, genes), we computed an *F*-statistic to test for association by stratifying measurements by the surrogate. *p*-values were obtained and, because of multiple comparisons, false discovery rates (FDRs) were obtained using the Benjamini–Hochberg procedure. A feature obtaining an FDR below 5% was considered susceptible to batch effects. ¶Principal components analysis was performed on the feature level data. The principal components were ranked in decreasing order of the variability that they explained. We computed the association (using R^2) between the surrogate and the first five principal components. We report the rank of the component with the highest correlation; the correlation is given in parenthesis. #As for ¶ but using the biological outcome of interest instead of the surrogate. **As for ¶ but using principal components to define batch. ‡‡As for ¶ but using biological outcome. NA, not available.

Nature Reviews Genetics 2010 (11), 733-739.



Nature Reviews Genetics 2010 (11), 733-739.

- Each row is a different HapMap sample processed in the same facility with the same platform. See Supplementary information S1 (box) for a description of the data represented here. The samples are ordered by processing date with horizontal lines dividing the different dates. We show a 3.5 Mb region from chromosome 16. **Coverage** data from each feature were standardized across samples: blue represents three standard deviations below average and orange represents three standard deviations above average. Various batch effects can be observed, and the largest one occurs between days 243 and 251 (the large orange horizontal streak).

Nature Reviews Genetics 2010 (11), 733-739.

Data pre-processing in our lab

- Data quality control/check according to wet lab/manufacture guidelines.
- Background correction.
- Adjustment by observed confounding variables.
- Principle component analysis to check any unobserved confounding variables.
- SVA (surrogate variable analysis) by Leek et al. (2012, 2008)

Data pre-processing in our lab

- Array based expression data, DNA methylation data, SNP data etc. (Illumina, Affymetrix, etc.)
- Sequence based data for SNP discovery.
- Technology dependent, experiment dependent, study dependent**. Case control or cohort design. Association or prediction**.
- Bayesian factor model approach vs principle component analysis

Acknowledgements

- Drs. Chung-Hsin Chen, Shih-Sheng Jiang, Chao Hsiung (Expression data and DNA methylation data)
- Drs. Wen-Chang Wang, Chao Hsiung (SNP for GWAS)
- Drs. Ying-Hsiang Chen, Chao Hsiung (Sequence data for fine mapping)
- Capable RA.
- TBI Bioinformatics core supported by NSC

Thank you for your attention