Lecture 12:
More mistakes in published papers

BMI 713/GEN 229
November 30, 2016
Peter J Park

BMC Bioinformatics

BioMed Central

Correspondence

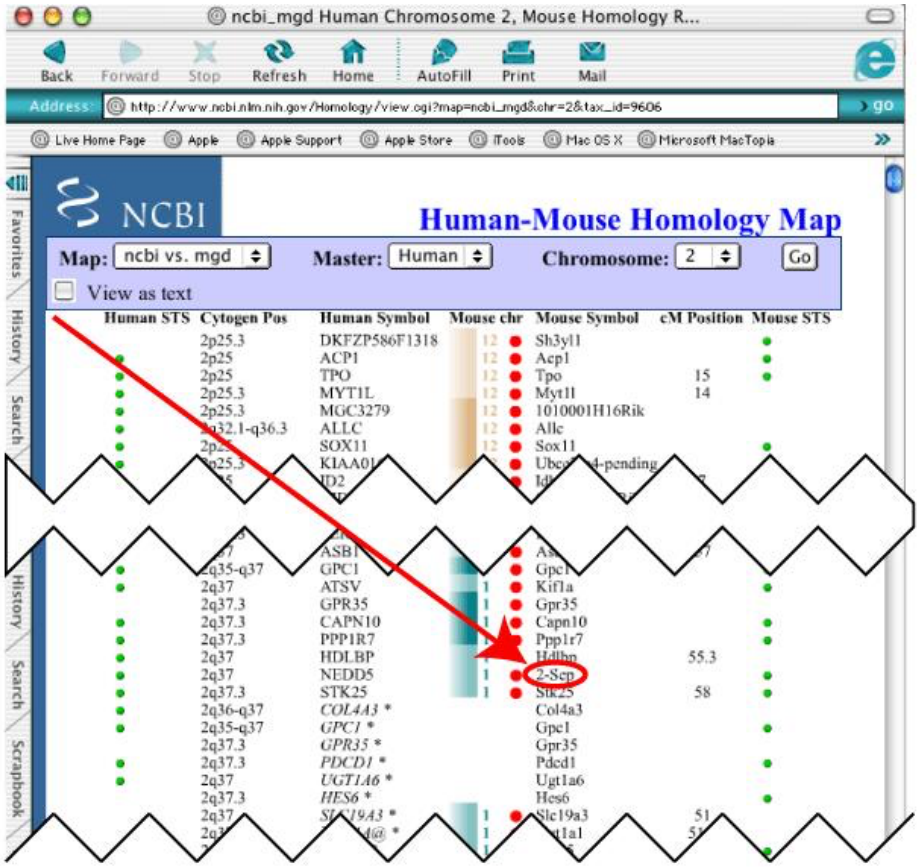
Open Access

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg[†1], Joseph Riss[†2], David W Kane[3], Kimberly J Bussey[1], Edward Uchio[4], W Marston Linehan[4], J Carl Barrett[2] and John N Weinstein[*1]

SEPT2 (Septin 2) -> '2-Sep'
MARCH1 -> '1-Mar'.
DEC1 -> '1-Dec'

at least 30 genes



Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

- Downloaded supplementary files from 18 journals (2005-2015) using a suite of shell scripts.
- **35,175** supplementary Excel files, finding 7467 gene lists attached to 3597 papers.
- Confirmed gene name errors in 987 files from 704 articles
- Overall rate: 19.6 %

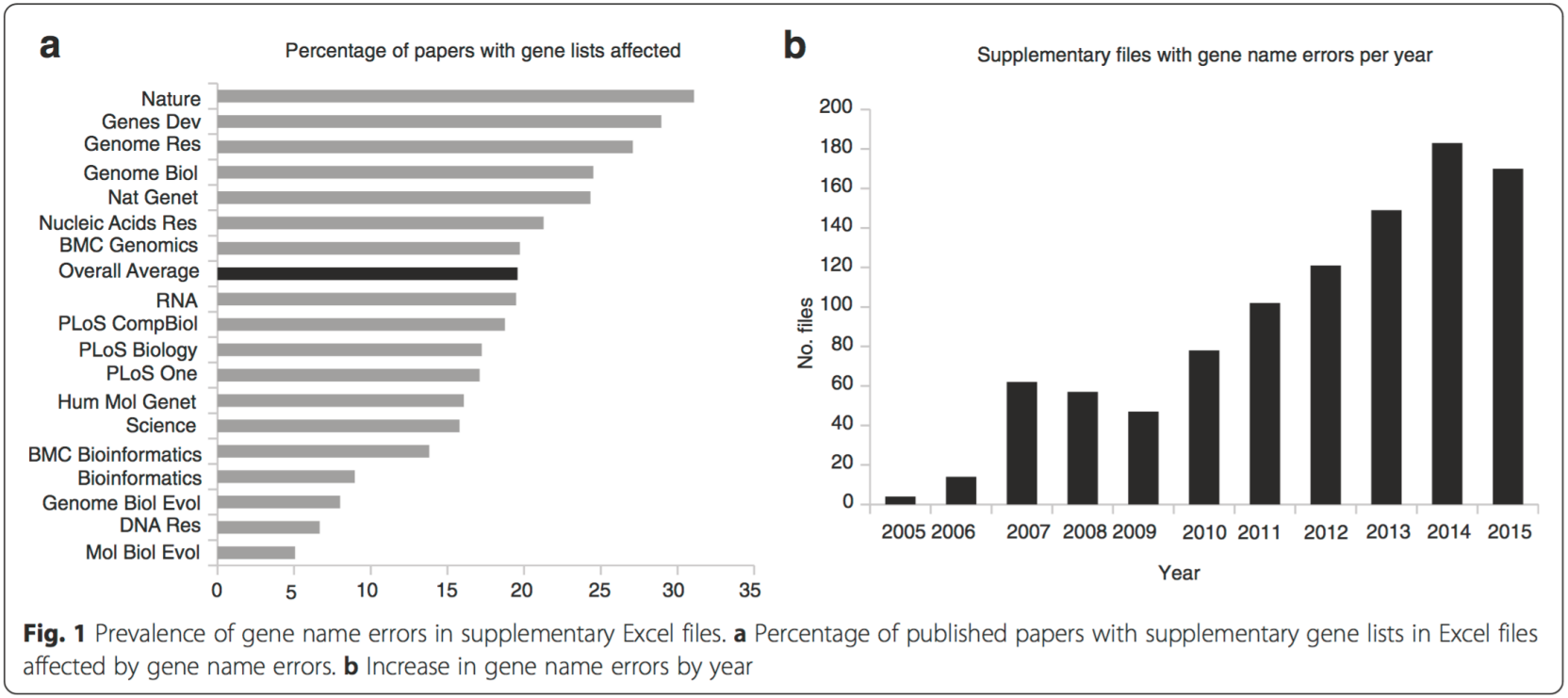


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

TCGA data - Illumina 450K arrays

```
genes <- as.data.frame(table(tcga$gene.symbol))
head(genes, 20)

      Var1   Freq
# 1          119652
# 2  10-Mar      5
# 3  11-Mar     21
# 4  11-Sep     32
# 5  13-Sep     18
# 6  14-Sep      4
# 7   1-Dec      8
# 8   1-Mar     30
# 9   1-Sep     10
# 10  3-Mar     33
# 11  4-Mar     25
# 12  5-Mar      4
# 13  5-Sep     12
# 14  6-Mar     21
# 15  7-Mar     13
# 16  8-Sep      2
# 17  9-Mar      6
# 18  9-Sep     16
# 19   A1BG      6
# 20   A1CF      5
```

GEO file

```
> library(GEOquery)
> x <- getGEO("GPL13534")
File stored at:
/tmp/RtmphmfKfm/GPL13534.soft
> geo.ann <- Table(x)
> grep("10-Mar|11-Mar|11-Sep|13-Sep|14-Sep|1-Dec|1-Mar|1-Sep", geo.ann$UCSC_RefGene_Name, val=TRUE)
  [1] "1-Mar"  "11-Sep" "1-Mar"  "11-Sep" "11-Sep" "11-Sep" "11-Sep" "11-Sep"
  [9] "1-Mar"  "11-Sep" "1-Mar"  "1-Mar"  "1-Mar"  "11-Sep" "1-Mar"  "11-Sep"
 [17] "1-Mar"  "1-Mar"  "11-Sep" "1-Mar"  "1-Mar"  "1-Mar"  "11-Sep" "11-Sep"
 [25] "11-Sep" "1-Mar"  "1-Mar"  "1-Mar"  "1-Mar"  "1-Mar"  "11-Sep" "11-Sep"
 [33] "1-Mar"  "11-Sep" "11-Sep" "11-Sep" "1-Mar"  "11-Sep" "11-Sep" "1-Mar"
 [41] "11-Sep" "1-Mar"  "1-Mar"  "11-Sep" "1-Mar"  "1-Mar"  "11-Sep" "11-Sep"
 [49] "1-Mar"  "1-Mar"  "1-Mar"  "11-Sep" "1-Mar"  "11-Sep" "1-Mar"  "11-Sep"
 [57] "11-Sep" "1-Mar"  "11-Sep" "11-Sep" "11-Sep" "11-Sep" "11-Mar" "11-Mar"
 [65] "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar"
 [73] "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar" "11-Mar"
 [81] "11-Mar" "11-Mar" "11-Mar" "14-Sep" "13-Sep" "13-Sep" "13-Sep" "14-Sep"
 [89] "14-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep"
 [97] "14-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep" "13-Sep"
[105] "13-Sep" "1-Dec"  "1-Dec"  "1-Dec"  "1-Dec"  "1-Dec"  "1-Dec"  "1-Dec"
[113] "1-Dec"  "1-Sep"  "1-Sep"  "1-Sep"  "1-Sep"  "1-Sep"  "1-Sep"  "1-Sep"
[121] "1-Sep"  "1-Sep"  "1-Sep"  "10-Mar" "10-Mar" "10-Mar" "10-Mar" "10-Mar"
>
```

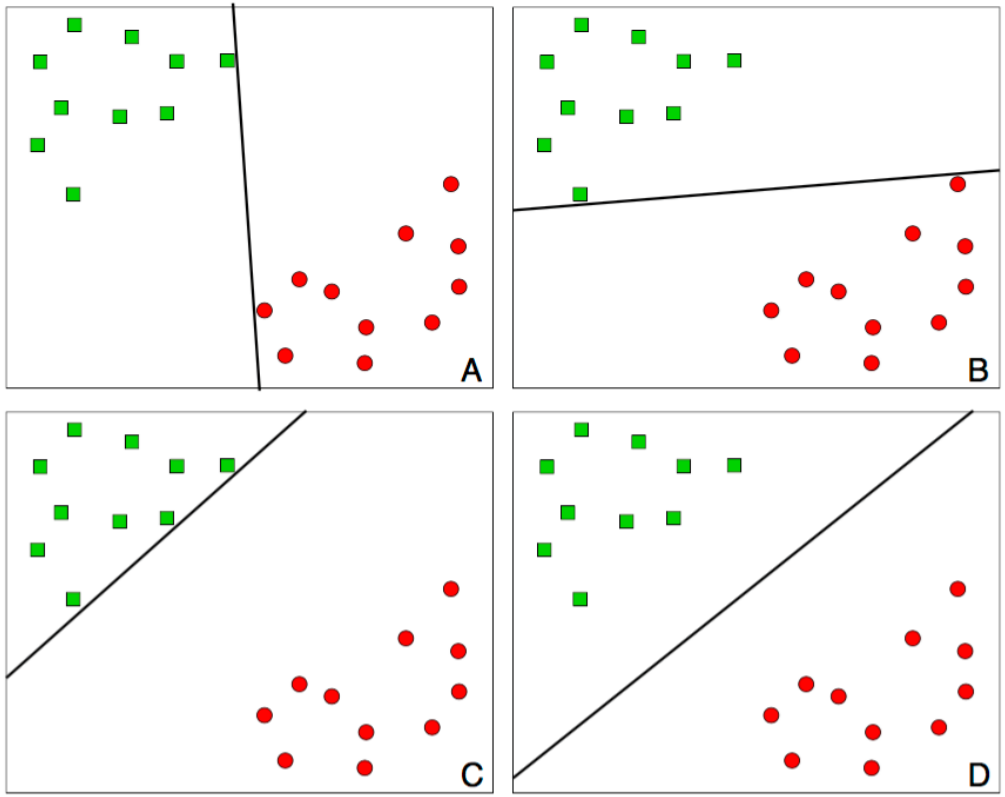
https://nsaunders.wordpress.com/2012/10/22/gene-name-errors-and-excel-lessons-not-learned/

Mistakes in classification

- Overfitting
- Data normalization
- Problems in proper training / testing

Many methods for classification

- k-nearest neighbors classifier
- logistic regression
- Bayes classifiers
- support vector machines
- neural networks
- decision trees
- random forest



Do you model the whole data set or just the boundary?

How do you penalize for errors?

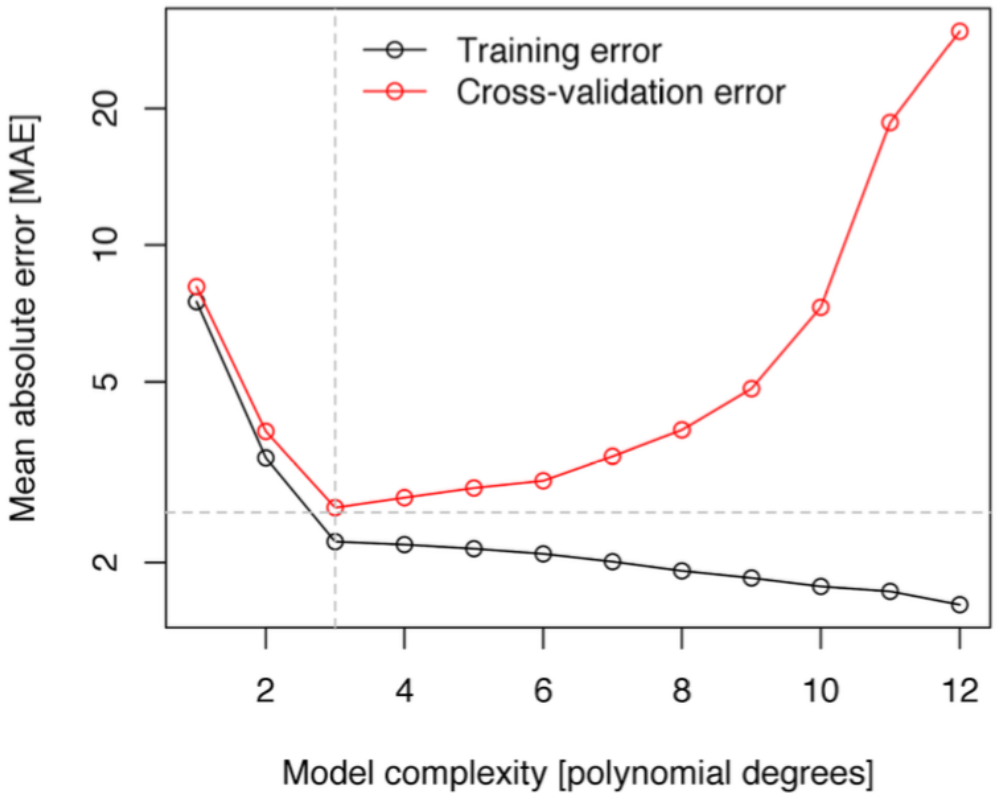
Cross-validation

- You often don't have enough samples for training and testing
- *N*-fold cross-validation: split the data into *N* pieces and use (*N*-1) for training and 1 for testing. Repeat *N* times.
- A special case: Leave-one-out cross-validation (LOOCV)

- Important: at what stage do you "leave out" the data?

Overfitting

- If there are more variable than data points, you can build a model that fits the data perfectly
- But it may not do well in prediction.
- Cross-validation



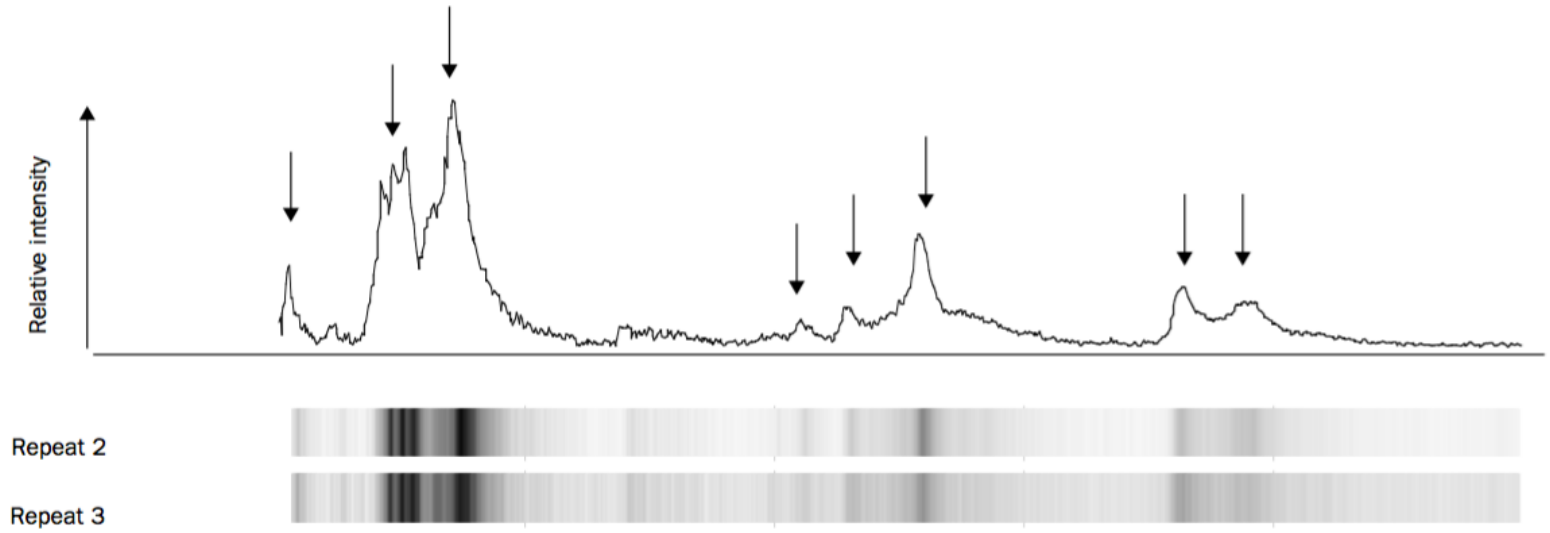
http://menugget.blogspot.com/2014/05/evaluating-model-performance-practical.html

THE LANCET • Vol 359 • February 16, 2002

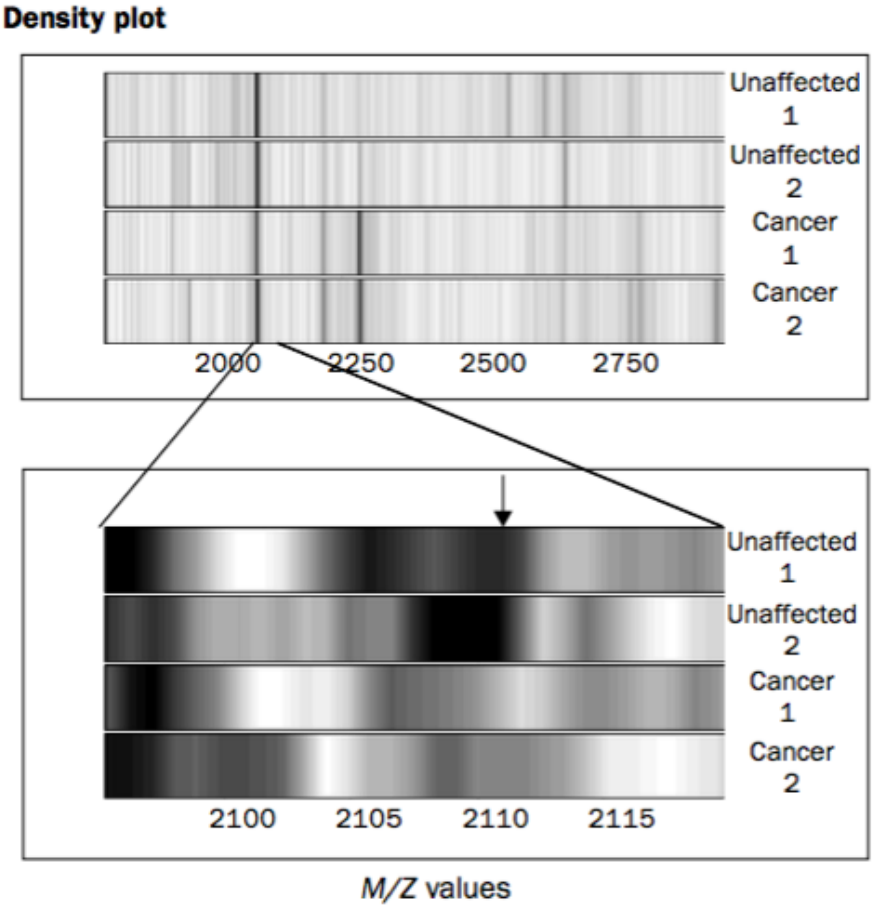
Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- Detection of early-stage ovarian cancer using proteomic patterns in serum?
- Generated proteomic spectra (SELDI-TOF)
- Training set: 50 unaffected women and 50 patients with ovarian cancer
- “an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer.”
- Testing set: 116 masked samples, 50 with ovarian cancer, 66 from unaffected or those with non-malignant disorders.
  - Correctly identified all 50 cases, including all 18 stage I cases.
  - 63 out of 66 recognized as not cancer.



sensitivity: 100%
specificity: 95%
positive predictive value: 94%



Detection of cancer-specific markers amid massive mass spectral data

Wei Zhu*†, Xuena Wang*, Yeming Ma‡, Manlong Rao*, James Glimm*§, and John S. Kovach¶

*Department of Applied Mathematics and Statistics, and ¶Long Island Cancer Center, State University of New York, Stony Brook, NY 11794; and ‡Medical Department, and §Center for Data Intensive Computing, Brookhaven National Laboratory, Upton, NY 11973

Edited by Richard V. Kadison, University of Pennsylvania, Philadelphia, PA, and approved October 8, 2003 (received for review April 16, 2003)

We propose a comprehensive pattern recognition procedure that will achieve best discrimination between two or more sets of subjects with data in the same coordinate system. Applying the procedure to MS data of proteomic analysis of serum from ovarian cancer patients and serum from cancer-free individuals in the Food and Drug Administration/National Cancer Institute Clinical Proteomics Database, we have achieved perfect discrimination (100% sensitivity, 100% specificity) of patients with ovarian cancer, including early-stage disease, from normal controls for two independent sets of data. Our procedure identifies the best subset of proteomic biomarkers for optimal discrimination between the groups and appears to have higher discriminatory power than other methods reported to date. For large-scale screening for diseases of relatively low prevalence such as ovarian cancer, almost perfect specificity and sensitivity of the detection system is critical to avoid unmanageably high numbers of false-positive cases.

What is wrong with this paper?

- PPV calculation
- There are no specific proteins identified

- m/z not in the typical range - not tumor-derived at all?
- if the women were at high-risk, stress hormones?

Their claim: sensitivity: 100%
specificity: 95%
positive predictive value: 94%

- Positive predictive value (PPV): the probability that subjects with a positive screening test truly have the disease.
- What is the incidence of ovarian cancer? 1 in 2500
- If you screen 2500 women, the test would identify 1 TP and 125 FP results (5% of 2500). PPV: 1/126
- PPV depends on prevalence but they used their training set numbers to calculate incidence/PPV!!
- Authors' reply: "a test with 100% sensitivity and 95% specificity is still not suitable to screen the general population for ovarian cancer… As an optimal standard, of course, we should strive for a 100% specific and 100% sensitive screening test for ovarian cancer, given its low prevalence. Proteomic pattern analysis may be able to reach this seemingly elusive goal." [we repeated the experiment on a better platform, we obtain 100% sensitivity/specificity]

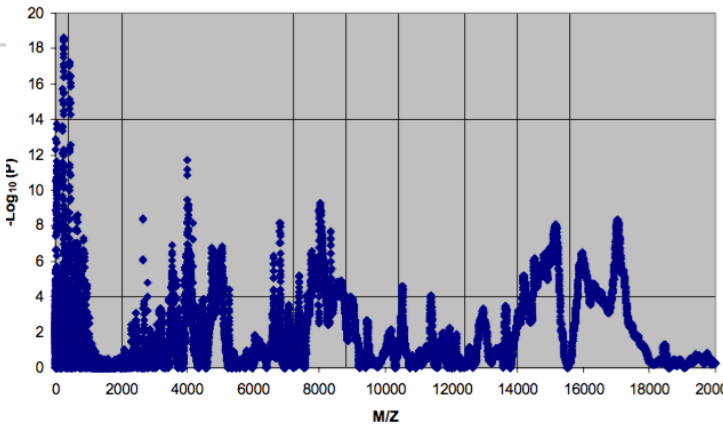
BMC Bioinformatics

Research article

A data review and re-assessment of ovarian cancer serum proteomic profiling

James M Sorace*[1,2,3] and Min Zhan[4]

Wilcoxon p-values between training & testing sets



- The region containing the M/Z (mass to charge) values of greatest statistical difference between cancer and controls occurred at M/Z values <500; these low values are considered to be noise.
- Some low values (e.g., 2.79 and 245.5) could be used to distinguish cancer and control in the testing set perfectly
- -> There may be significant non-biologic experimental bias between these two groups

BIOINFORMATICS

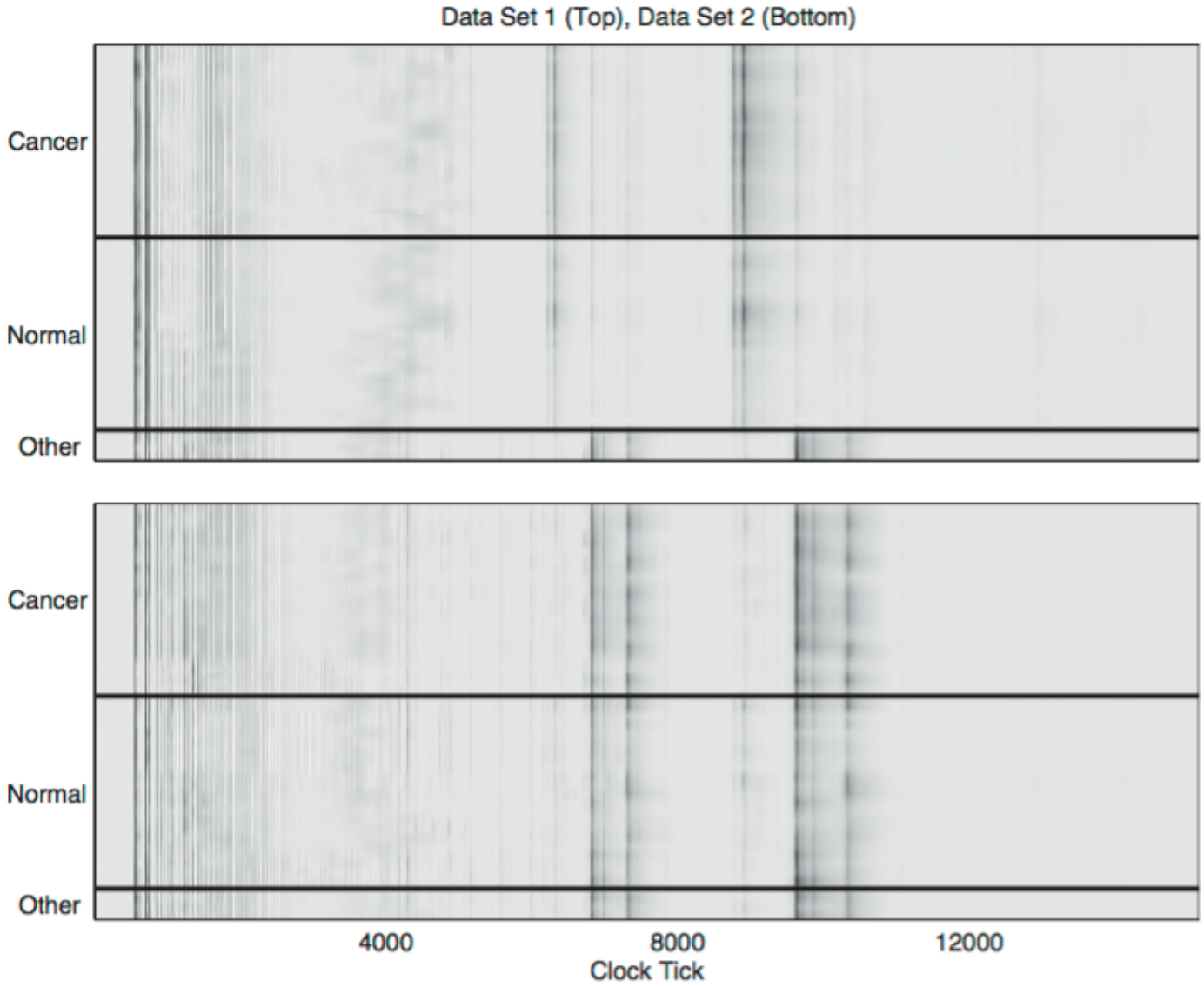
Vol. 20 no. 5 2004, pages 777–785
DOI: 10.1093/bioinformatics/btg484

Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments

Keith A. Baggerly, Jeffrey S. Morris and Kevin R. Coombes*

- Dataset 1: original data, 216 samples
- Dataset 2: same samples run on a newer array

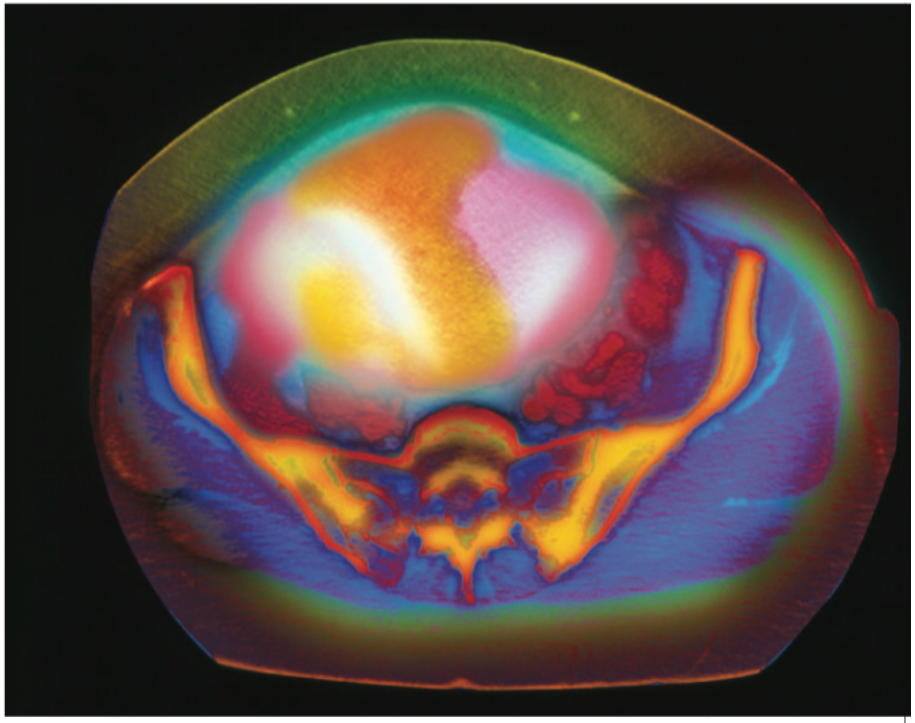
- Protocol change
- Discriminating features are not reproducible



news feature

Running before we can walk?

Two years ago, a new proteomic test was heralded as the future of cancer diagnostics. But since then, doubts about its effectiveness have begun to grow. Erika Check reports.



On target: can proteins in the blood reveal ovarian tumours (pink/yellow) before they reach this stage?

Correlogic Files for Chapter 11 Bankruptcy Protection

Jul 23, 2010

The article is updated to clarify the history of the OvaCheck test and changes made to it.

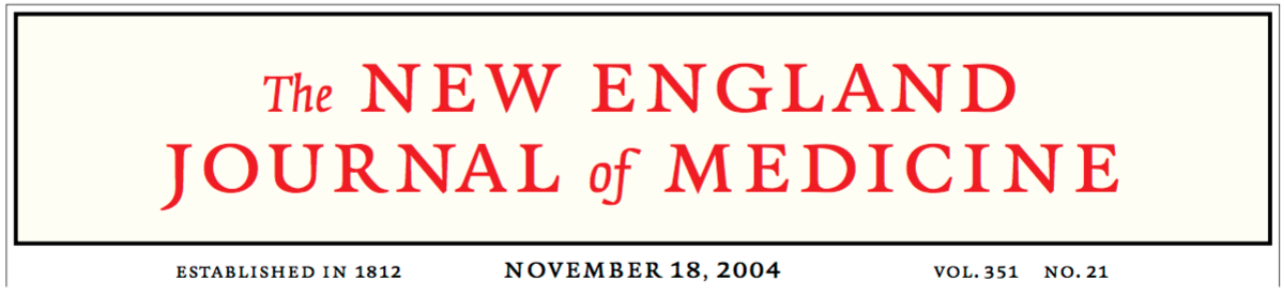
By Tony Fong

NEW YORK (GenomeWeb News) – Correlogic Systems has filed for Chapter 11 reorganization in the hopes of securing funding for its troubled OvaCheck test and eventually obtaining clearance from the US Food and Drug Administration.

The privately held Germantown, Md., firm filed its petition for Chapter 11 protection last week in US Bankruptcy Court in the District of Maryland, and has asked the court to allow it to "reject" licensing agreements with Quest Diagnostics and the Lab Corporation of America.

Some issues are subtle

- What is the proper way to choose training and testing sets?

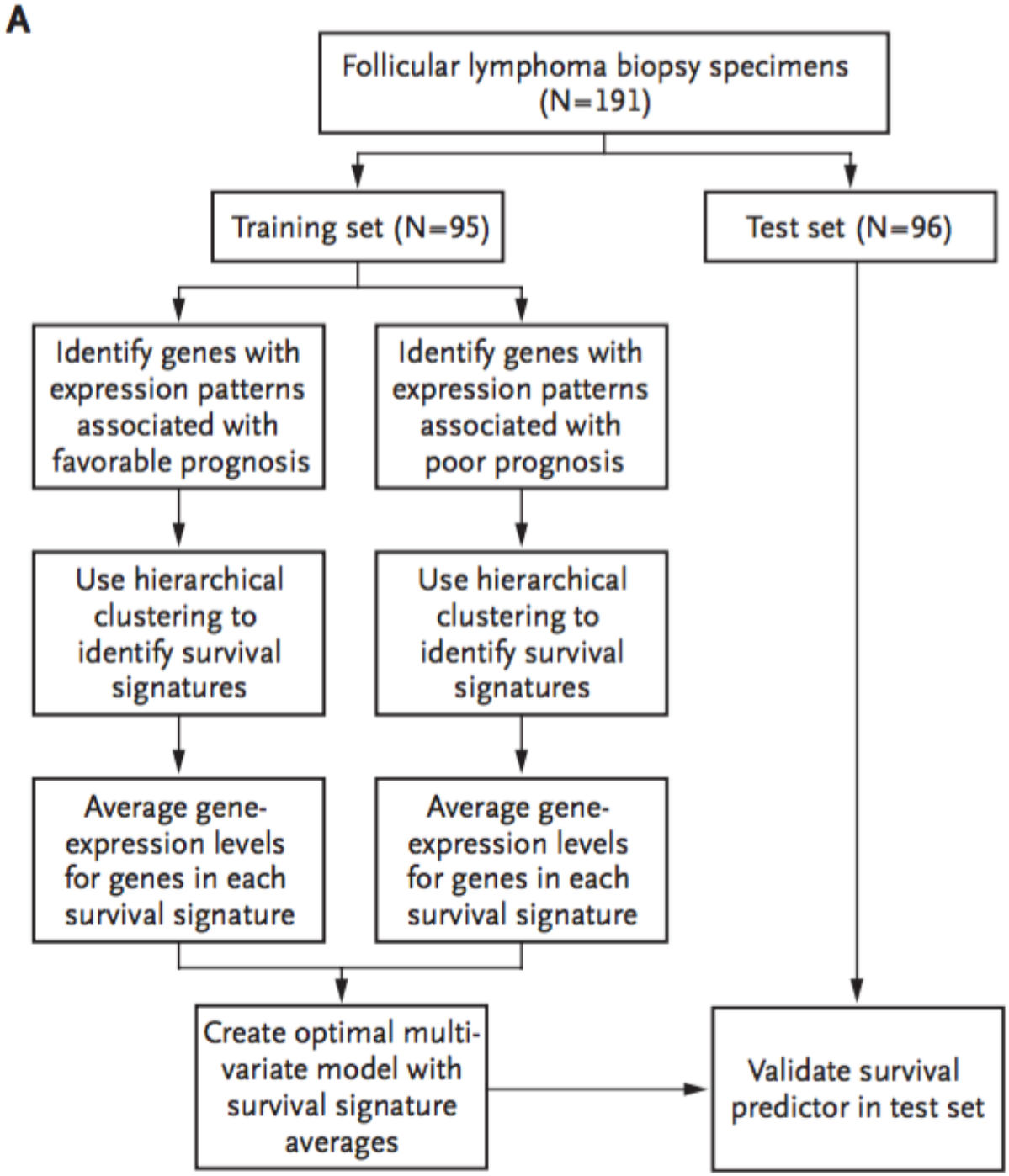


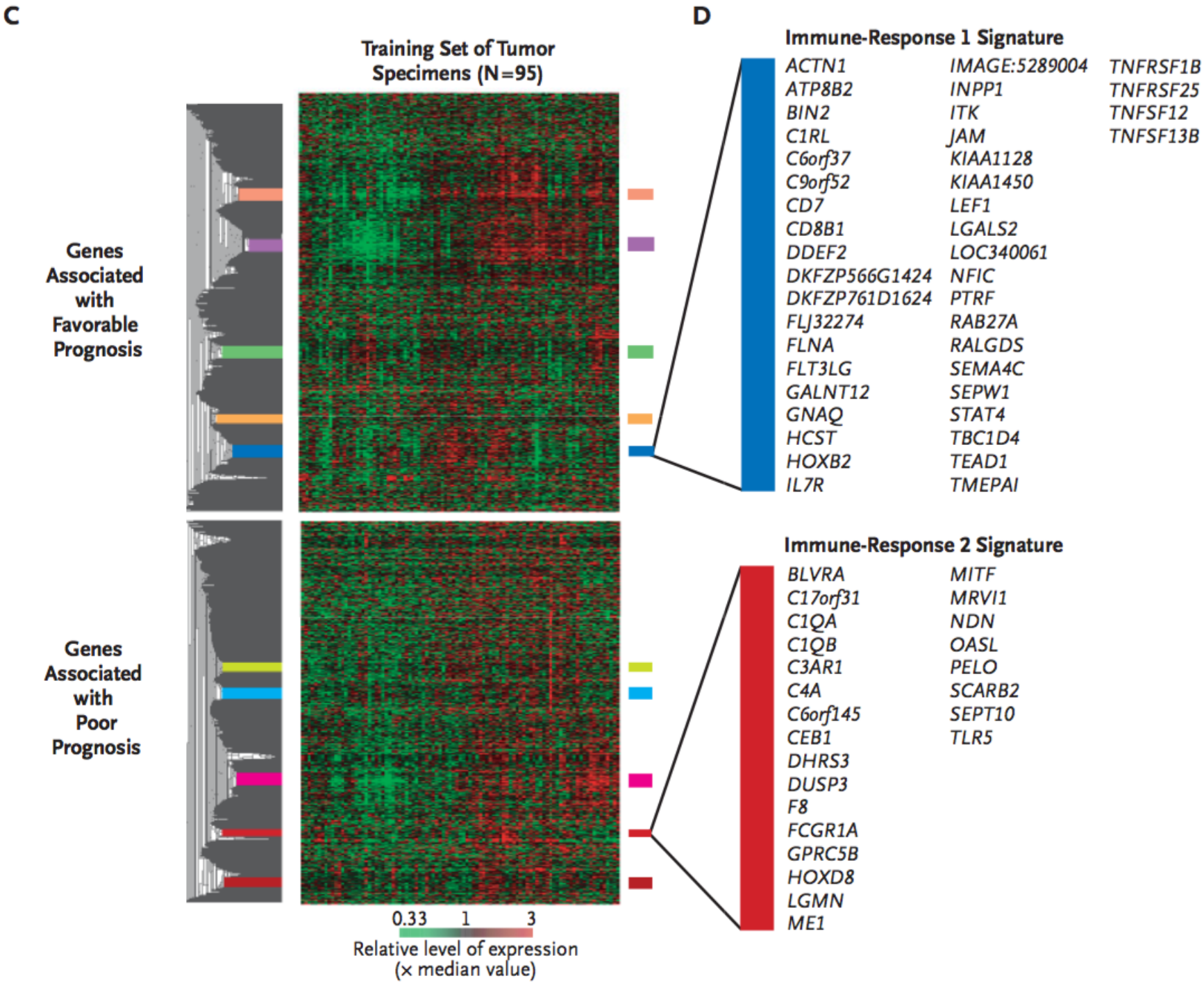
The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 18, 2004 VOL. 351 NO. 21

Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells

Sandeep S. Dave, M.D., George Wright, Ph.D., Bruce Tan, M.D., Andreas Rosenwald, M.D., Randy D. Gascoyne, M.D., Wing C. Chan, M.D., Richard I. Fisher, M.D., Rita M. Braziel, M.D., Lisa M. Rimsza, M.D., Thomas M. Grogan, M.D., Thomas P. Miller, M.D., Michael LeBlanc, Ph.D., Timothy C. Greiner, M.D., Dennis D. Weisenburger, M.D., James C. Lynch, Ph.D., Julie Vose, M.D., James O. Armitage, M.D., Erlend B. Smeland, M.D., Ph.D., Stein Kvaloy, M.D., Ph.D., Harald Holte, M.D., Ph.D., Jan Delabie, M.D., Ph.D., Joseph M. Connors, M.D., Peter M. Lansdorp, M.D., Ph.D., Qin Ouyang, Ph.D., T. Andrew Lister, M.D., Andrew J. Davies, M.D., Andrew J. Norton, M.D., H. Konrad Muller-Hermelink, M.D., German Ott, M.D., Elias Campo, M.D., Emilio Montserrat, M.D., Wyndham H. Wilson, M.D., Ph.D., Elaine S. Jaffe, M.D., Richard Simon, Ph.D., Liming Yang, Ph.D., John Powell, M.S., Hong Zhao, M.S., Neta Goldschmidt, M.D., Michael Chiorazzi, B.A., and Louis M. Staudt, M.D., Ph.D.


A
Follicular lymphoma biopsy specimens (N=191)
Training set (N=95)
Test set (N=96)
Identify genes with expression patterns associated with favorable prognosis
Identify genes with expression patterns associated with poor prognosis
Use hierarchical clustering to identify survival signatures
Use hierarchical clustering to identify survival signatures
Average gene-expression levels for genes in each survival signature
Average gene-expression levels for genes in each survival signature
Create optimal multivariate model with survival signature averages
Validate survival predictor in test set

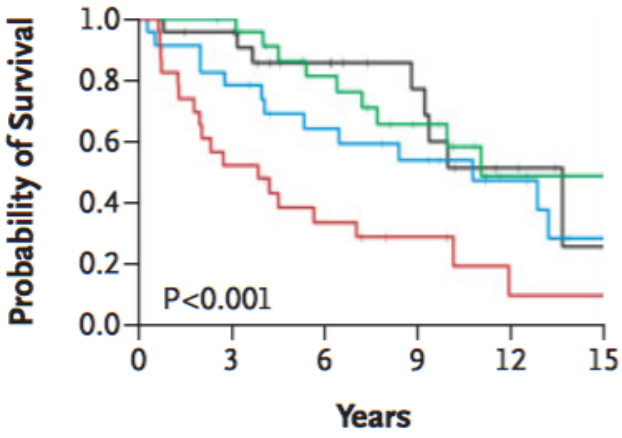

C
Training Set of Tumor Specimens (N=95)
Genes Associated with Favorable Prognosis
Genes Associated with Poor Prognosis
0.33
1
3
Relative level of expression (× median value)
D
Immune-Response 1 Signature
ACTN1
ATP8B2
BIN2
C1RL
C6orf37
C9orf52
CD7
CD8B1
DDEF2
DKFZP566G1424
DKFZP761D1624
FLJ32274
FLNA
FLT3LG
GALNT12
GNAQ
HCST
HOXB2
IL7R
IMAGE:5289004
INPP1
ITK
JAM
KIAA1128
KIAA1450
LEF1
LGALS2
LOC340061
NFIC
PTRF
RAB27A
RALGDS
SEMA4C
SEPW1
STAT4
TBC1D4
TEAD1
TMEPAI
TNFRSF1B
TNFRSF25
TNFSF12
TNFSF13B
Immune-Response 2 Signature
BLVRA
C17orf31
C1QA
C1QB
C3AR1
C4A
C6orf145
CEB1
DHRS3
DUSP3
F8
FCGR1A
GPRC5B
HOXD8
LGMN
ME1
MITF
MRVI1
NDN
OASL
PELO
SCARB2
SEPT10
TLR5

A simple model

- Identified 10 clusters; tried multivariate models
- One model with two signatures (both “immune response”) was highly predictive of survival in the training set
- Predictive of survival in the training set (P<0.001) and the test set (P<0.003)

Table 2. Predictive Power of Gene-Expression Signatures in Follicular Lymphoma.*

Gene-Expression Signature	P Value for Contribution to Model in Test Set	Relative Risk of Death (95% CI)*	Effect of Increased Gene Expression on Survival
Immune-response 1	<0.001	0.15 (0.05–0.46)	Favorable
Immune-response 2	<0.001	9.35 (3.02–28.90)	Unfavorable



No. at Risk

SPS quartile 1	23	21	14	9	4	1
SPS quartile 2	24	23	16	10	3	2
SPS quartile 3	24	18	13	10	6	2
SPS quartile 4	23	12	7	4	1	1

A re-analysis by Tibshirani *et al*

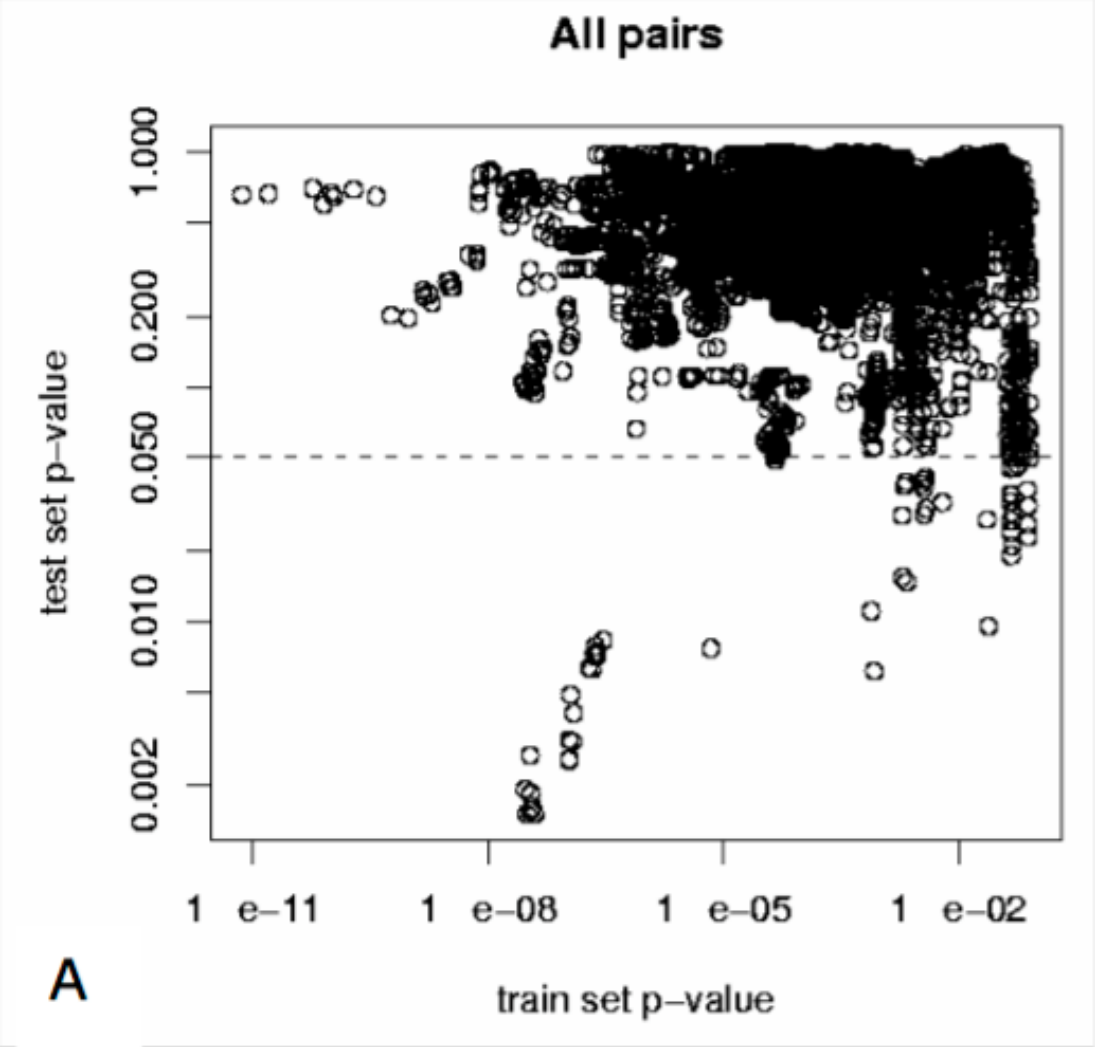
"I re-analyzed their data and found that it was not reproducible. In particular when I applied their algorithm as they described it, I could not get the same results as they did on their training and test sets (although there are many models that have low p-values on both the training and test sets).

Furthermore when their equal-sized training and test sets are swapped, and their model-building procedure is re-applied, virtually nothing is significant in the test set.

Also, when a small change is made to their model-building recipe (changing the allowable cluster size range from [25,50] to [30,60]) with either the original or swapped datasets, again, very little of significance emerges. This and other analyses suggest that their result occurred by chance and is not robust or reproducible."

http://statweb.stanford.edu/~tibs/FL/report/

Randomly split testing/training sets



Reproducibility

- “Most readers do not have the time to spend weeks reconstructing a data analysis (as I did here). For this reason, it would be very helpful if authors in general provided not only details of their analysis, but a software script that carries it out. That way the reader can assess for himself the fragility of the results.”

Batch effect

- instrumentation calibration
- laboratory condition (train; power)
- protocol differences
- biological variability
- reagent lots
- personnel differences
- data processing differences
- **Numerous multi-laboratory studies: laboratory-specific effect is the strongest!**

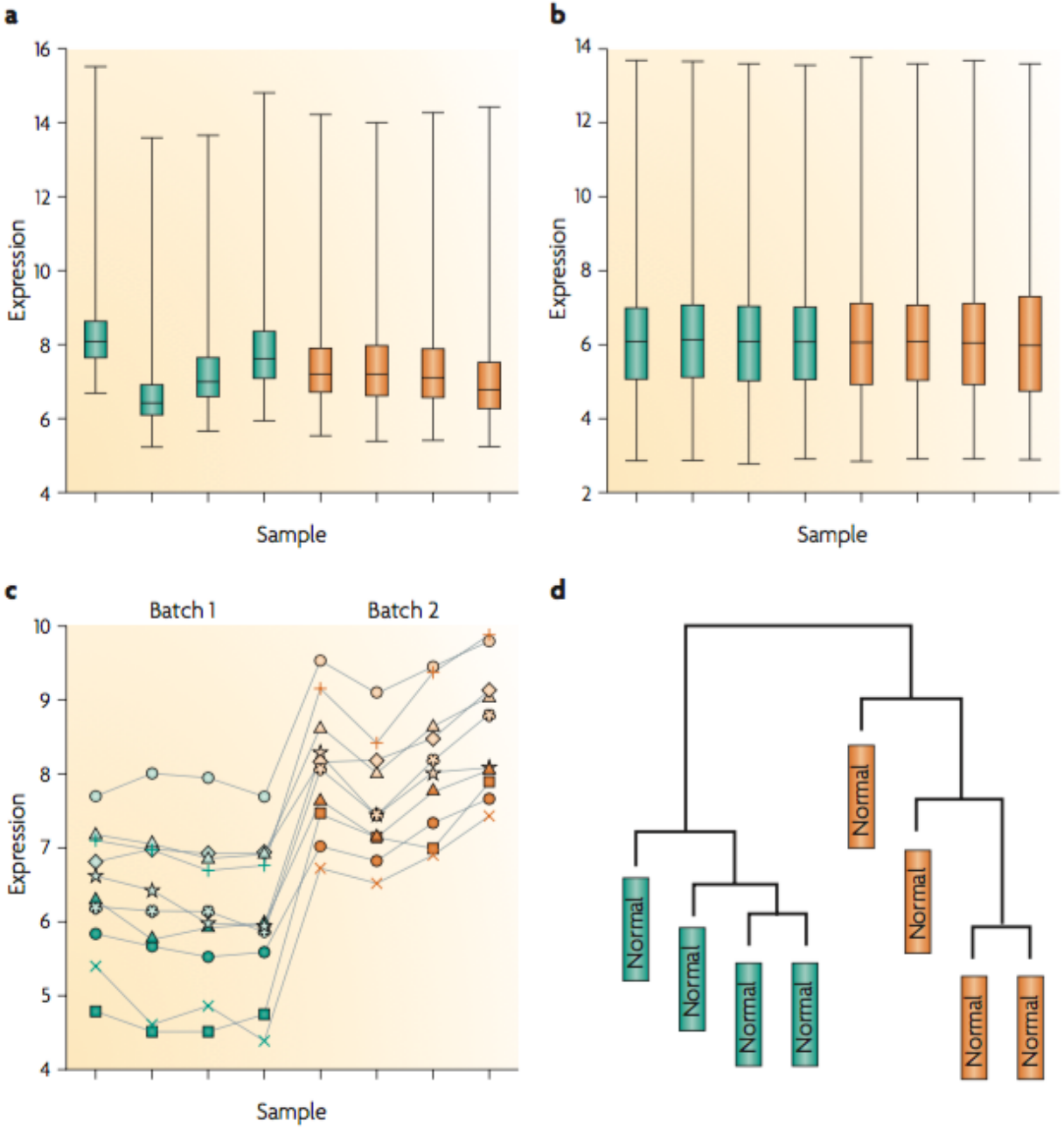
Batch effect

- Many features will be different with statistical significance between batches.
- If batch effect is confounded with biological effect, it will be impossible to distinguish the two.

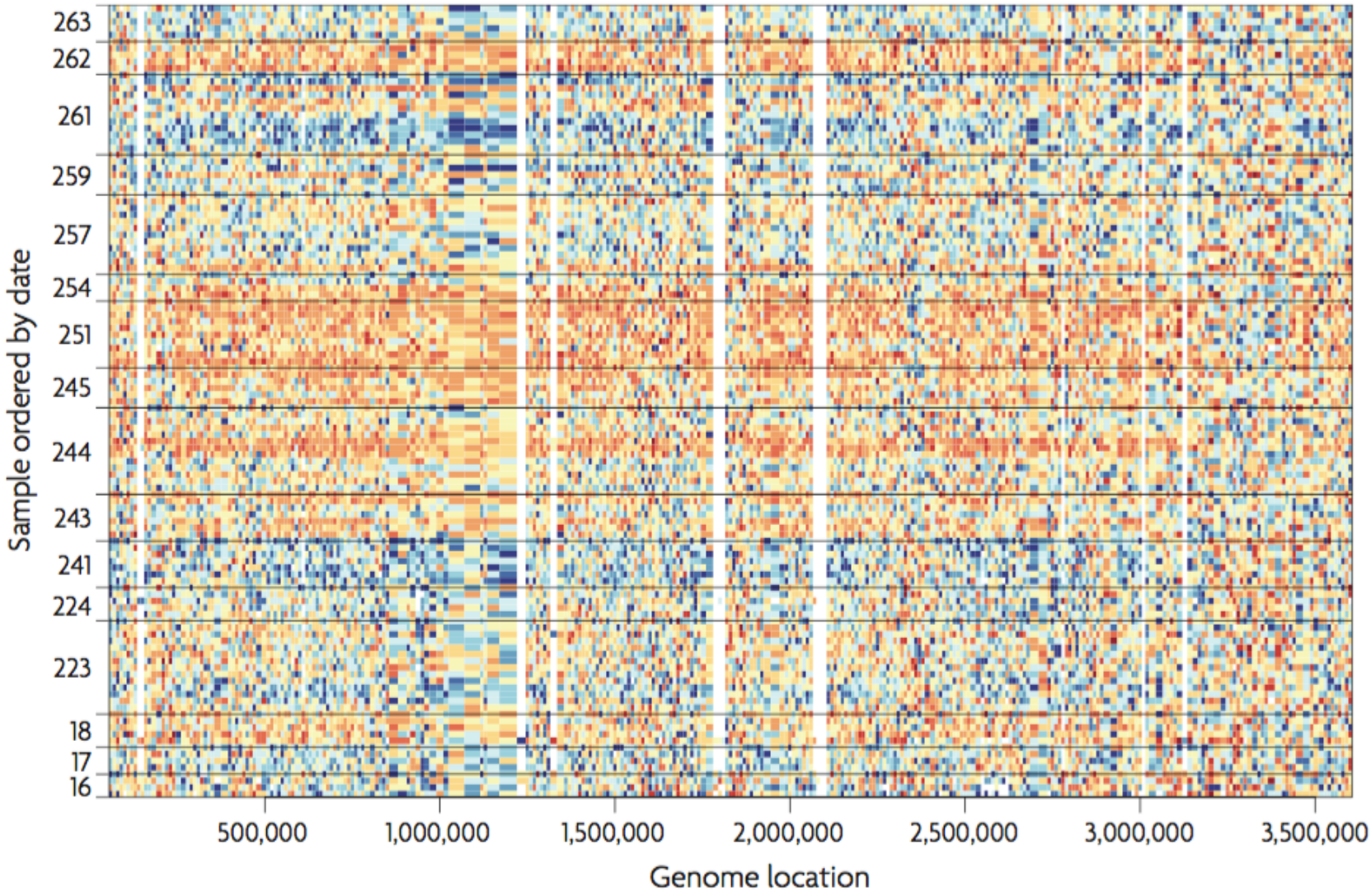
OPINION

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 and Rafael A. Irizarry



- Gene expression study (bladder cancer)
- Green/orange - processing dates
- Even quantile normalization does not fix the problem!



1000 Genomes data (3.5Mb from chr 16):
Each row - HapMap samples processed in the same facility with the same platform
Coverage data from each feature were standardized across samples: blue represents three standard deviations below average and orange represents three standard deviations above average.

Leek et al, Nat Reviews Genetics, 2010

LETTERS

nature genetics

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]

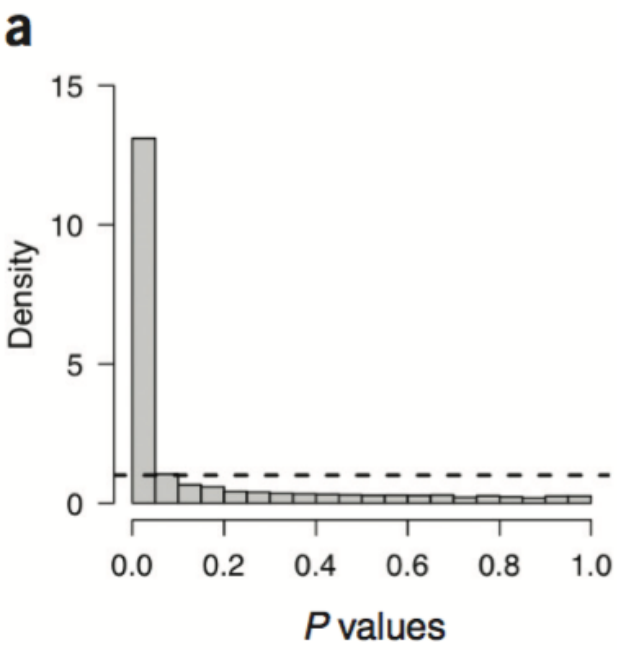
Claim: mean gene expression differed significantly between European-derived and Asian-derived populations for approximately 25% of 4,197 genes tested.

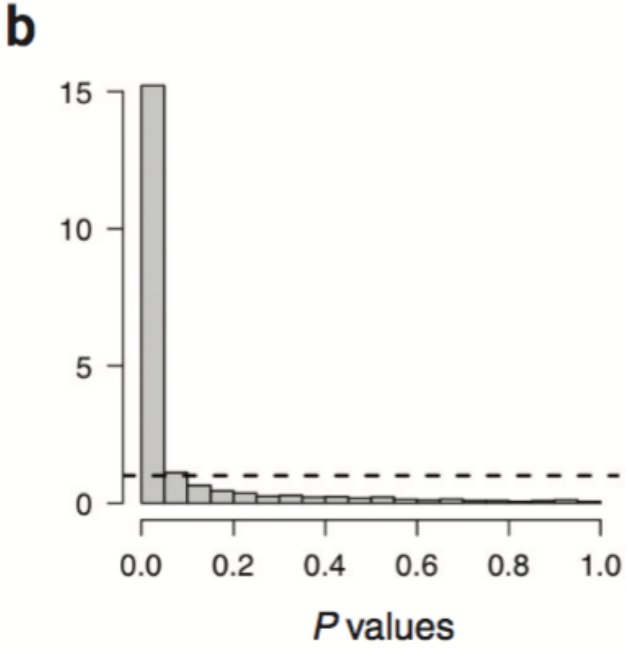
CORRESPONDENCE

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

Joshua M Akey[1], Shameek Biswas[1], Jeffrey T Leek[2] & John D Storey[1,2]

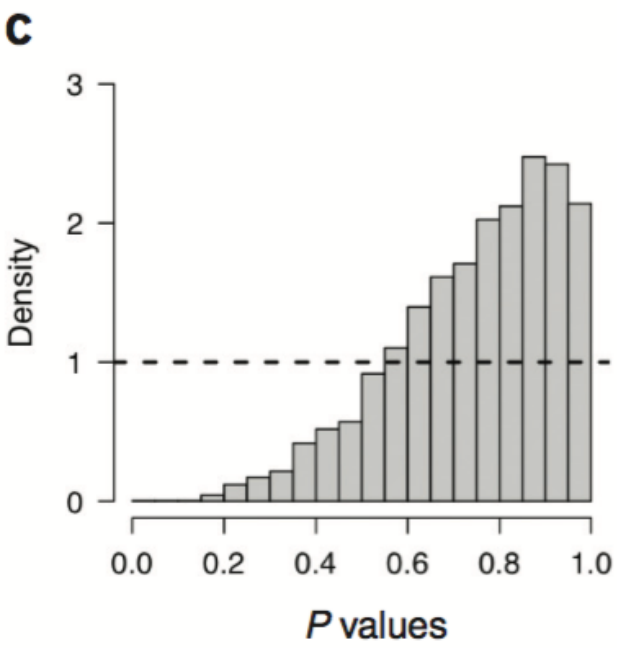
differential expression between CEU and ASN samples

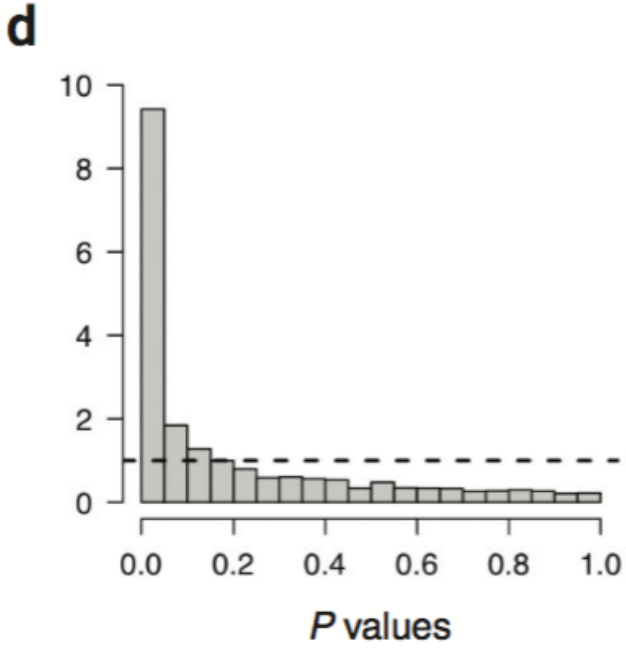




differential expression between samples of different processing year

differential expression between CEU and ASN samples **controlling for the year**





differential expression only among CEU samples **between the years**

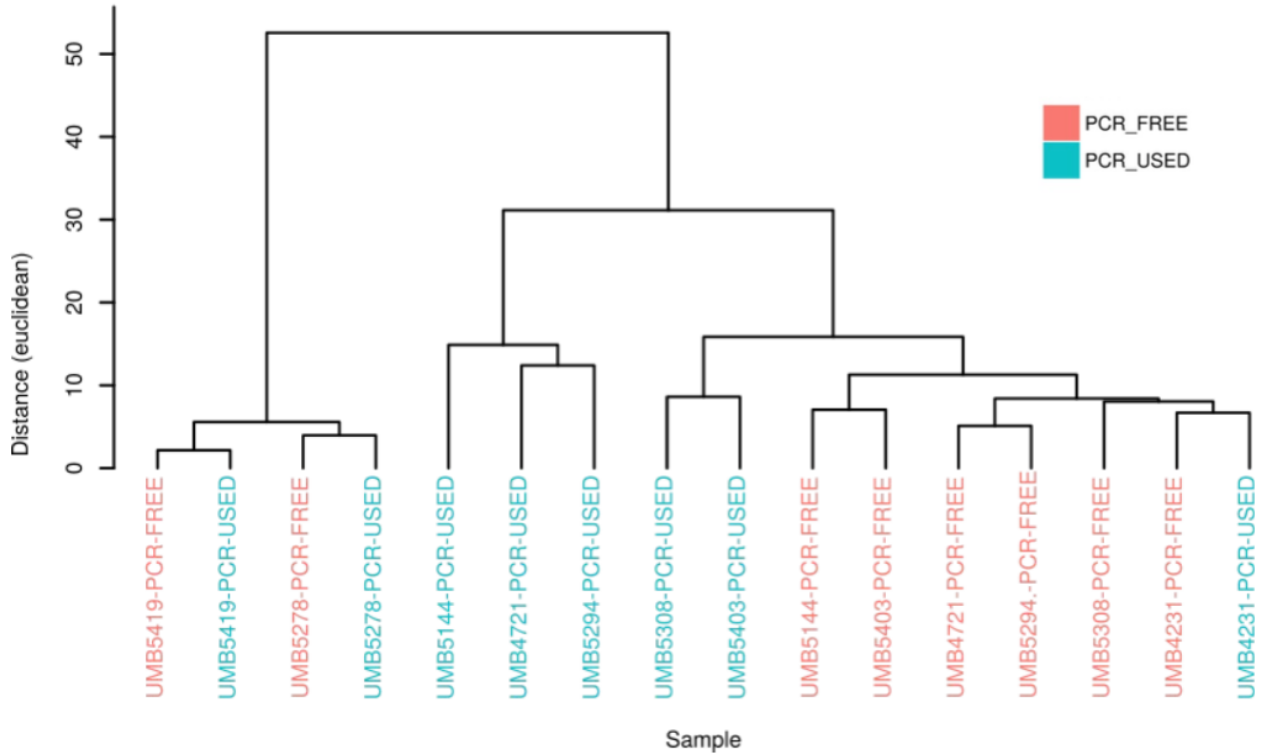
Authors' response

In our paper, we wrote (in the Methods section), "The growing and processing of the HapMap cell lines was randomized by population group to eliminate batch effects that may contribute to apparent population differences in gene expression." Because of the different dates of processing described in the previous paragraph, this was not actually done. (Of course, we did not intend

Nevertheless, data obtained in an independent study[3] strongly support our conclusions for the most extreme populatior differences. Stranger *et al.*[3] carried out gene expression analysis on cells independently prepared from the same HapMap samples we studied, and with different microarray technology. In Table 1 of our paper, we listed 35 genes whose mean expression level differed between CEU and CHB+JPT samples by a factor of 2 or more (and with $P < 0.05$ after multiple testing correction). We looked for those 35 genes in the data of Stranger *et al.*[3] Of the 35 genes, 32 were on their arrays. Among these 32, 30 were also significantly different between CEU and CHB+JPT in their analysis, and 29 of the 30 differed in the same direction as in our results. Thus, a large proportion (29 of 32) of the genes in Table 1 were also found to differ by Stranger *et al.*[3] In addition, we

From my laboratory:

- Two sequencing companies used different protocols for library construction (PCR vs PCR-free)



Steps to mitigate the problem

- **Have a proper study design!**
- **Exploratory analysis** - visualize the data; cluster samples with all known potential variables; plot individual features (see an earlier example where differences in individual features may be hidden in the bulk distribution).
- Principal component analysis (or other methods such as SVD)
- **Linear models** can remove some batch effects
- Packages: (especially for small sample sizes) ComBat, Surrogate Variable Analysis (SVA)

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