Experimental Design

Design principles and artifacts in genomic data

Basic principles of experimental design

- Questions / goals of the experiment.
- Comparison / control.
- Replication.
- Randomization.
- Stratification (aka blocking).
- Factorial experiments.

Confounding

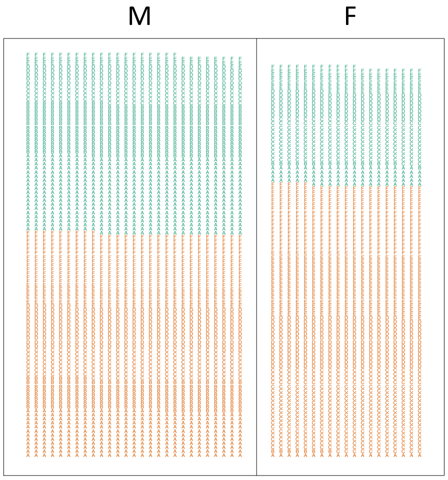
UC Berkeley 1973 Admission Data

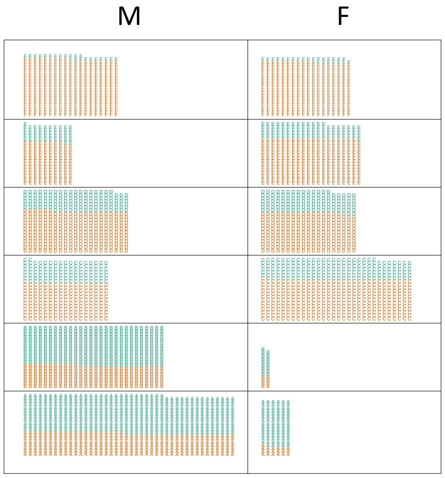
	Accepted	Rejected
Males	1,198	1,493
Females	557	1,278

Stratified by major

Major	Male	Female
A	62%	82%
B	63%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%

Confounding



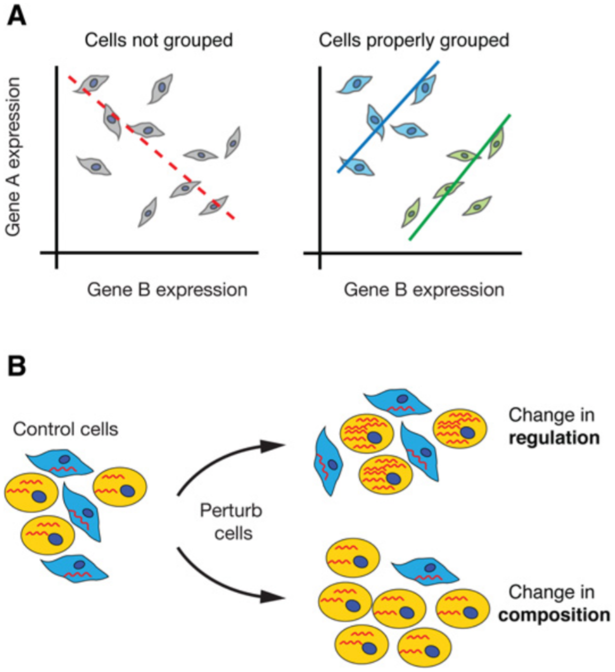


Confounding

Player	1995	1996	Combined
Derek Jeter	.250 (12/48)	.314 (183/582)	.310 (195/630)
David Justice	.253 (104/411)	.321 (45/140)	.271 (149/551)

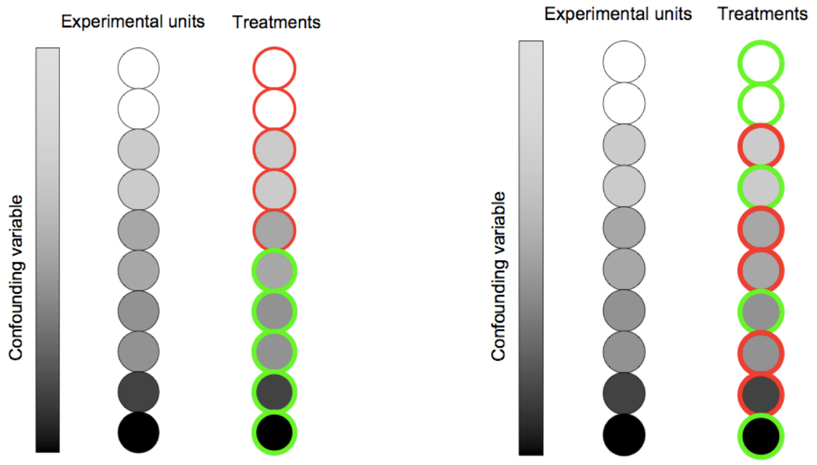
[RI]

Confounding

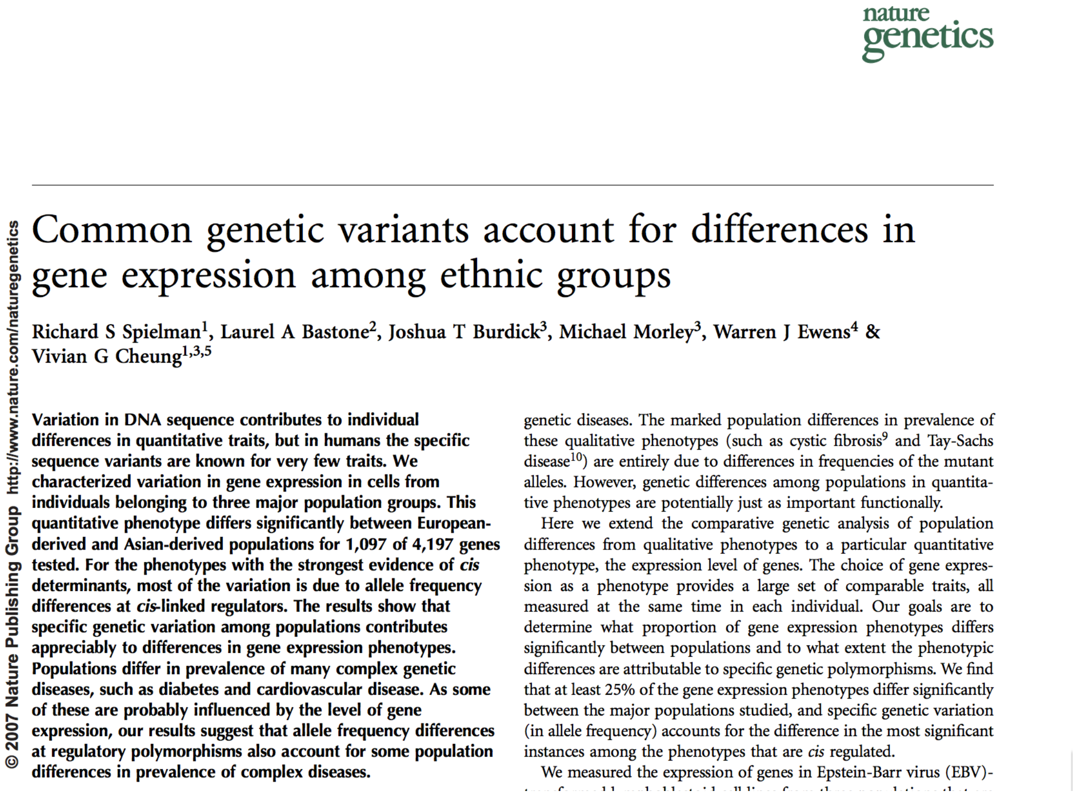


PMID 26430159

Confounding



Without randomization of “treatments”, the confounding variable may be correlated with the treatment. Any observed association between treatment and covariates will be difficult to distinguish from an association between the confounding variable and the covariates.



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]

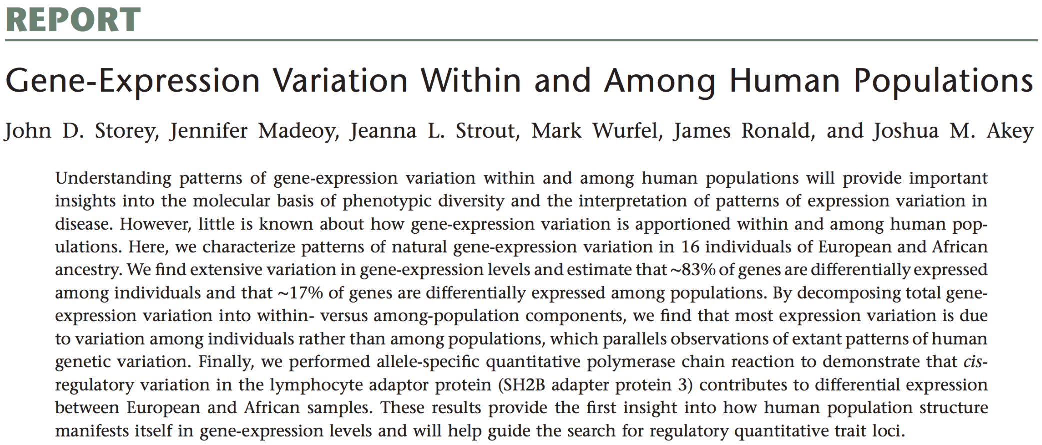
Variation in DNA sequence contributes to individual differences in quantitative traits, but in humans the specific sequence variants are known for very few traits. We characterized variation in gene expression in cells from individuals belonging to three major population groups. This quantitative phenotype differs significantly between European-derived and Asian-derived populations for 1,097 of 4,197 genes tested. For the phenotypes with the strongest evidence of *cis* determinants, most of the variation is due to allele frequency differences at *cis*-linked regulators. The results show that specific genetic variation among populations contributes appreciably to differences in gene expression phenotypes. Populations differ in prevalence of many complex genetic diseases, such as diabetes and cardiovascular disease. As some of these are probably influenced by the level of gene expression, our results suggest that allele frequency differences at regulatory polymorphisms also account for some population differences in prevalence of complex diseases.

genetic diseases. The marked population differences in prevalence of these qualitative phenotypes (such as cystic fibrosis[9] and Tay-Sachs disease[10]) are entirely due to differences in frequencies of the mutant alleles. However, genetic differences among populations in quantitative phenotypes are potentially just as important functionally.

Here we extend the comparative genetic analysis of population differences from qualitative phenotypes to a particular quantitative phenotype, the expression level of genes. The choice of gene expression as a phenotype provides a large set of comparable traits, all measured at the same time in each individual. Our goals are to determine what proportion of gene expression phenotypes differs significantly between populations and to what extent the phenotypic differences are attributable to specific genetic polymorphisms. We find that at least 25% of the gene expression phenotypes differ significantly between the major populations studied, and specific genetic variation (in allele frequency) accounts for the difference in the most significant instances among the phenotypes that are *cis* regulated.

We measured the expression of genes in Epstein-Barr virus (EBV)-

PMID 17206142



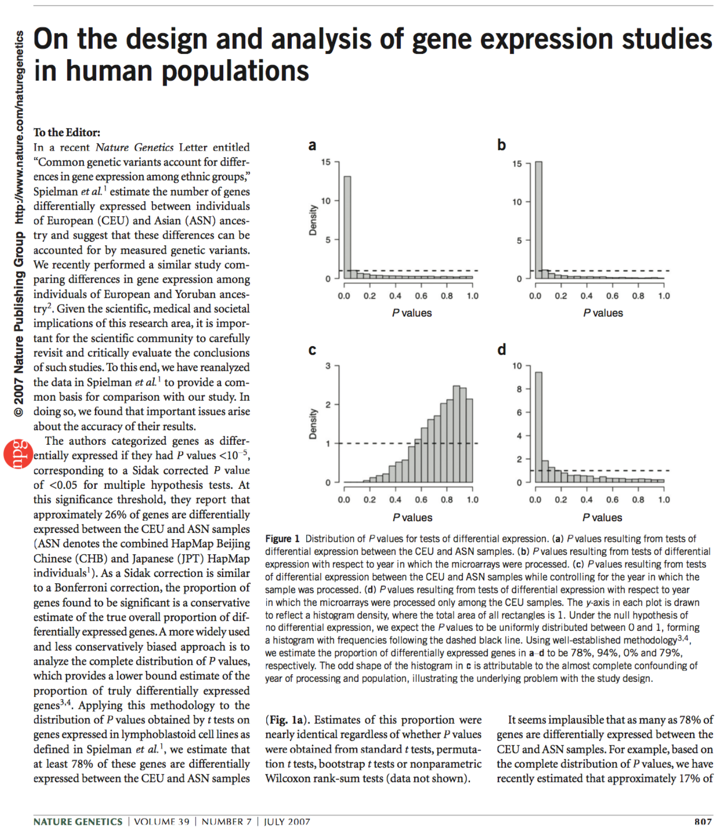
REPORT

Gene-Expression Variation Within and Among Human Populations

John D. Storey, Jennifer Madeoy, Jeanna L. Strout, Mark Wurfel, James Ronald, and Joshua M. Akey

Understanding patterns of gene-expression variation within and among human populations will provide important insights into the molecular basis of phenotypic diversity and the interpretation of patterns of expression variation in disease. However, little is known about how gene-expression variation is apportioned within and among human populations. Here, we characterize patterns of natural gene-expression variation in 16 individuals of European and African ancestry. We find extensive variation in gene-expression levels and estimate that ~83% of genes are differentially expressed among individuals and that ~17% of genes are differentially expressed among populations. By decomposing total gene-expression variation into within- versus among-population components, we find that most expression variation is due to variation among individuals rather than among populations, which parallels observations of extant patterns of human genetic variation. Finally, we performed allele-specific quantitative polymerase chain reaction to demonstrate that *cis*-regulatory variation in the lymphocyte adaptor protein (SH2B adapter protein 3) contributes to differential expression between European and African samples. These results provide the first insight into how human population structure manifests itself in gene-expression levels and will help guide the search for regulatory quantitative trait loci.

PMID 17273971



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

To the Editor:

In a recent *Nature Genetics* Letter entitled "Common genetic variants account for differences in gene expression among ethnic groups," Spielman *et al.*[1] estimate the number of genes differentially expressed between individuals of European (CEU) and Asian (ASN) ancestry and suggest that these differences can be accounted for by measured genetic variants. We recently performed a similar study comparing differences in gene expression among individuals of European and Yoruban ancestry[2]. Given the scientific, medical and societal implications of this research area, it is important for the scientific community to carefully revisit and critically evaluate the conclusions of such studies. To this end, we have reanalyzed the data in Spielman *et al.*[1] to provide a common basis for comparison with our study. In doing so, we found that important issues arise about the accuracy of their results.

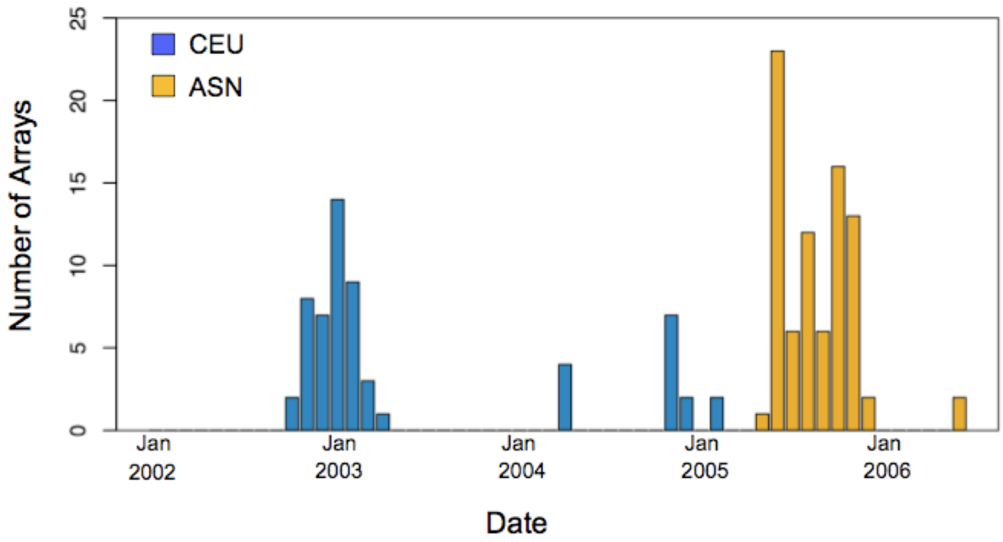
The authors categorized genes as differentially expressed if they had P values $<10^{-5}$, corresponding to a Sidak corrected P value of <0.05 for multiple hypothesis tests. At this significance threshold, they report that approximately 26% of genes are differentially expressed between the CEU and ASN samples (ASN denotes the combined HapMap Beijing Chinese (CHB) and Japanese (JPT) HapMap individuals[1]). As a Sidak correction is similar to a Bonferroni correction, the proportion of genes found to be significant is a conservative estimate of the true overall proportion of differentially expressed genes. A more widely used and less conservatively biased approach is to analyze the complete distribution of P values, which provides a lower bound estimate of the proportion of truly differentially expressed genes[3,4]. Applying this methodology to the distribution of P values obtained by t tests on genes expressed in lymphoblastoid cell lines as defined in Spielman *et al.*[1], we estimate that at least 78% of these genes are differentially expressed between the CEU and ASN samples (**Fig. 1a**). Estimates of this proportion were nearly identical regardless of whether P values were obtained from standard t tests, permutation t tests, bootstrap t tests or nonparametric Wilcoxon rank-sum tests (data not shown).

Figure 1 Distribution of P values for tests of differential expression. (**a**) P values resulting from tests of differential expression between the CEU and ASN samples. (**b**) P values resulting from tests of differential expression with respect to year in which the microarrays were processed. (**c**) P values resulting from tests of differential expression between the CEU and ASN samples while controlling for the year in which the sample was processed. (**d**) P values resulting from tests of differential expression with respect to year in which the microarrays were processed only among the CEU samples. The y-axis in each plot is drawn to reflect a histogram density, where the total area of all rectangles is 1. 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. Using well-established methodology[3,4], 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.

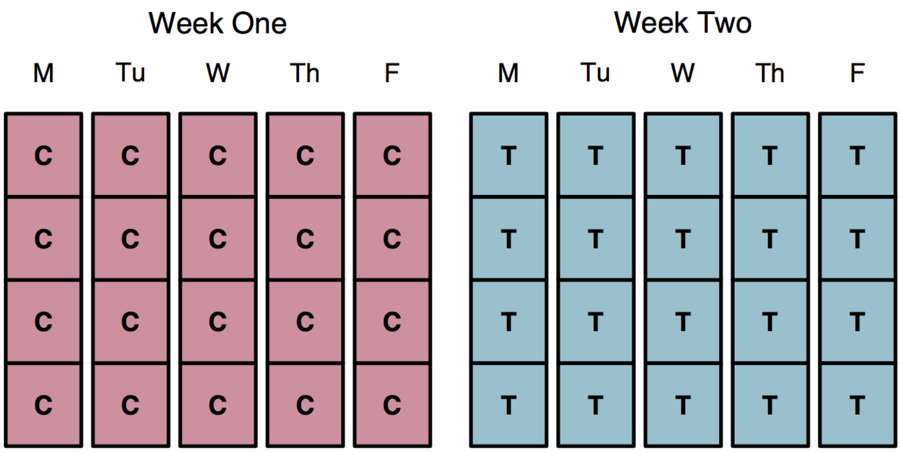
It seems implausible that as many as 78% of genes are differentially expressed between the CEU and ASN samples. For example, based on the complete distribution of P values, we have recently estimated that approximately 17% of

PMID 17597765

Confounding of population and processing time

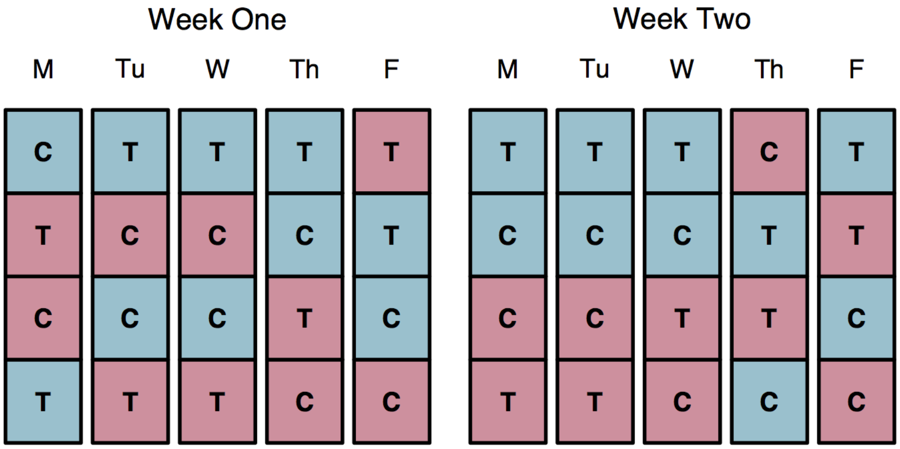


[JL]



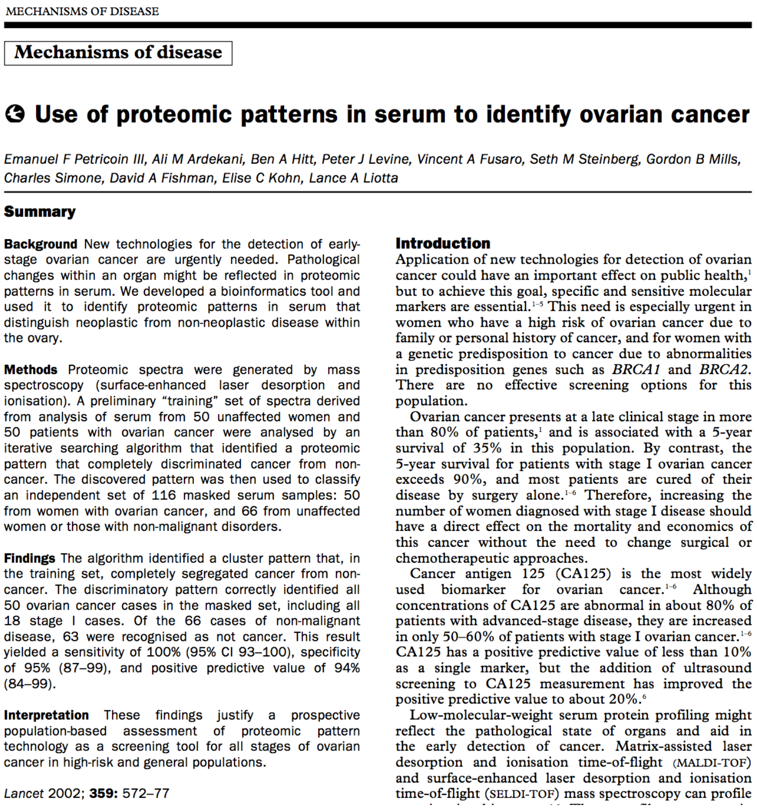
Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T

T = treated, C = control, pink = female, blue = male



Week One M	Week One Tu	Week One W	Week One Th	Week One F	Week Two M	Week Two Tu	Week Two W	Week Two Th	Week Two F
C 🟦	T 🟦	T 🟦	T 🟦	T 🟥	T 🟦	T 🟦	T 🟦	C 🟥	T 🟦
T 🟥	C 🟥	C 🟥	C 🟦	T 🟦	C 🟦	C 🟦	C 🟦	T 🟦	T 🟥
C 🟥	C 🟦	C 🟦	T 🟥	C 🟦	C 🟥	C 🟥	T 🟥	T 🟥	C 🟦
T 🟦	T 🟥	T 🟥	C 🟥	C 🟥	T 🟥	T 🟥	C 🟥	C 🟦	C 🟥

T = treated, C = control, pink = female, blue = male



MECHANISMS OF DISEASE

Mechanisms of disease

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

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary “training” set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognised as not cancer. This result yielded a sensitivity of 100% (95% CI 93–100), specificity of 95% (87–99), and positive predictive value of 94% (84–99).

Interpretation These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

Lancet 2002; **359:** 572–77

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,[1] but to achieve this goal, specific and sensitive molecular markers are essential.[1–5] This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as *BRCA1* and *BRCA2*. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients,[1] and is associated with a 5-year survival of 35% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and most patients are cured of their disease by surgery alone.[1–6] Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or chemotherapeutic approaches.

Cancer antigen 125 (CA125) is the most widely used biomarker for ovarian cancer.[1–6] Although concentrations of CA125 are abnormal in about 80% of patients with advanced-stage disease, they are increased in only 50–60% of patients with stage I ovarian cancer.[1–6] CA125 has a positive predictive value of less than 10% as a single marker, but the addition of ultrasound screening to CA125 measurement has improved the positive predictive value to about 20%.[6]

Low-molecular-weight serum protein profiling might reflect the pathological state of organs and aid in the early detection of cancer. Matrix-assisted laser desorption and ionisation time-of-flight (MALDI-TOF) and surface-enhanced laser desorption and ionisation time-of-flight (SELDI-TOF) mass spectroscopy can profile

RETRACTED 22 JULY 2011; SEE LAST PAGE

Sciencexpress Report

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani,[1]* Nadia Solovieff,[1] Annibale Puca,[2] Stephen W. Hartley,[1] Efthymia Melista,[3] Stacy Andersen,[4] Daniel A. Dworkis,[3] Jemma B. Wilk,[5] Richard H. Myers,[5] Martin H. Steinberg,[6] Monty Montano,[3] Clinton T. Baldwin,[6,7] Thomas T. Perls[4]*

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. [2]IRCCS Multimedica, Milano, Italy; Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, 20122, Italy. [3]Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. [4]Section of Geriatrics, Department of Medicine, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [5]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA. [6]Departments of Medicine and Pediatrics, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [7]Center for Human Genetics, Boston University School of Medicine, Boston, MA 02118, USA.

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)

Healthy aging is thought to reflect the combined influence of environmental factors (lifestyle choices) and genetic factors. To explore the genetic contribution, we undertook a genome-wide association study of exceptional longevity (EL) in 1055 centenarians and 1267 controls. Using these data, we built a genetic model that includes 150 single nucleotide polymorphisms (SNPs) and found that it could predict EL with 77% accuracy in an independent set of centenarians and controls. Further in-silico analysis revealed that 90% of centenarians can be grouped into 19 clusters characterized by different combinations of SNP genotypes—or genetic signatures—of varying predictive value. The different signatures, which attest to the genetic complexity of EL, correlated with differences in the prevalence and age of onset of age-associated diseases (e.g., dementia, hypertension, and cardiovascular disease) and may help dissect this complex phenotype into subphenotypes of healthy aging.

Based upon the hypothesis that exceptionally old individuals are carriers of multiple genetic variants that influence human lifespan (*4*), we conducted a genome-wide association study (GWAS) of centenarians. Centenarians are a model of healthy aging, as the onset of disability in these individuals is generally delayed until they are well into their mid-nineties (*5*, *6*). We studied 801 unrelated subjects enrolled in the New England Centenarian Study (NECS) and 926 genetically matched controls. NECS subjects were Caucasians who were born between 1890 and 1910 and had an age range of 95 to 119 years (median age 103 years). Figure S1 in the Supporting Online Material (*7*) describes the age distribution. Approximately one-third of the NECS sample included centenarians with a first-degree relative also achieving EL, thus enhancing the sample's power (*8*). Controls included 243 NECS referent subjects who were spouses of centenarian offspring or children of parents who died at the mean age of 73 years, and genome-wide SNP data

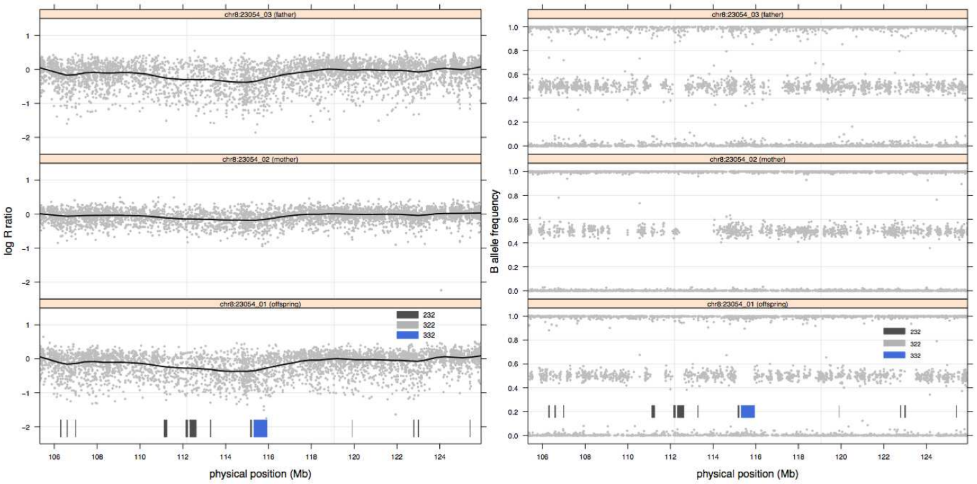
Raw high throughput genomic data always contain artifacts. No exceptions. Really.

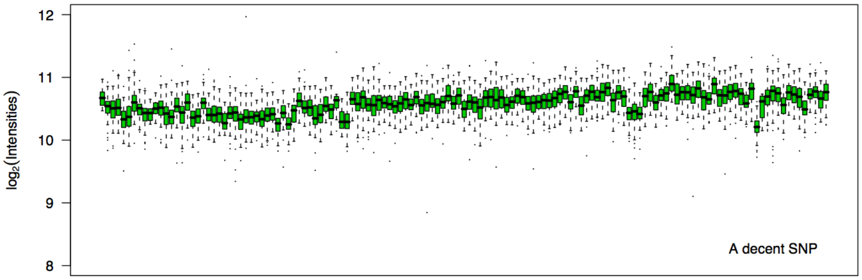
Identifying genomic signatures is a super hard problem. Technical artifacts in the data are often much larger than any biological signal.

Not addressing those artifacts can have nasty consequences, in particular when coupled with poor experimental design.

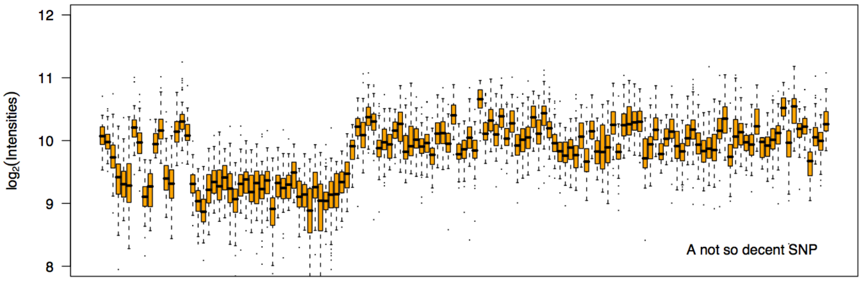
These artifacts include:

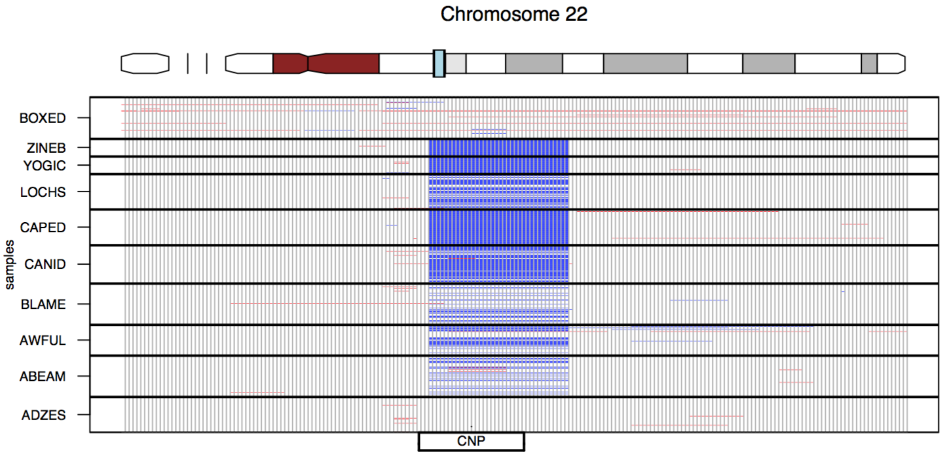
- Known systematic biases.
  For example genomic waves due to GC content.
- Random but possibly reproducible biases.
  For example laboratory specific artifacts.
- Random non-reproducible biases.
  For example plate and batch effects.

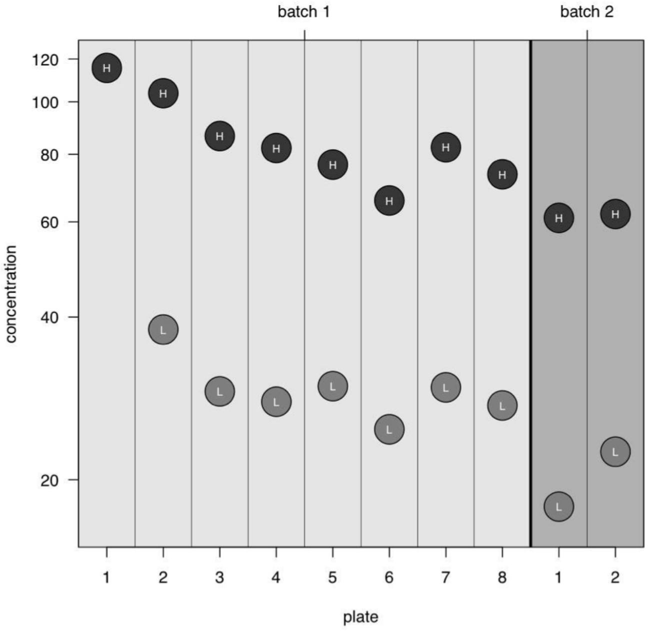


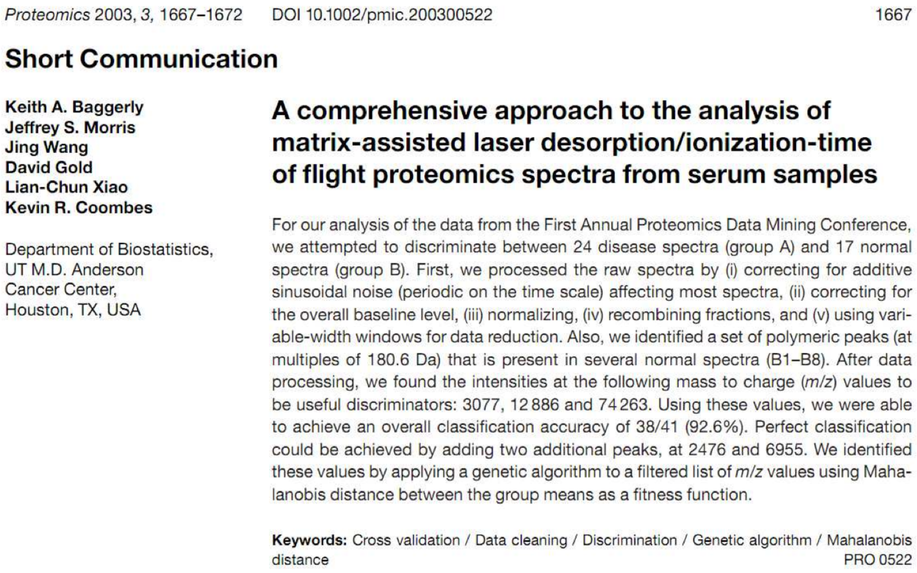


← Plates ordered by date of first run →




Chromosome 22
samples
BOXED
ZINEB
YOGIC
LOCHS
CAPED
CANID
BLAME
AWFUL
ABEAM
ADZES
CNP


batch 1
batch 2
concentration
120
100
80
60
40
20
H
L
plate
1
2
3
4
5
6
7
8
1
2



Proteomics 2003, *3,* 1667–1672 DOI 10.1002/pmic.200300522 1667

Short Communication

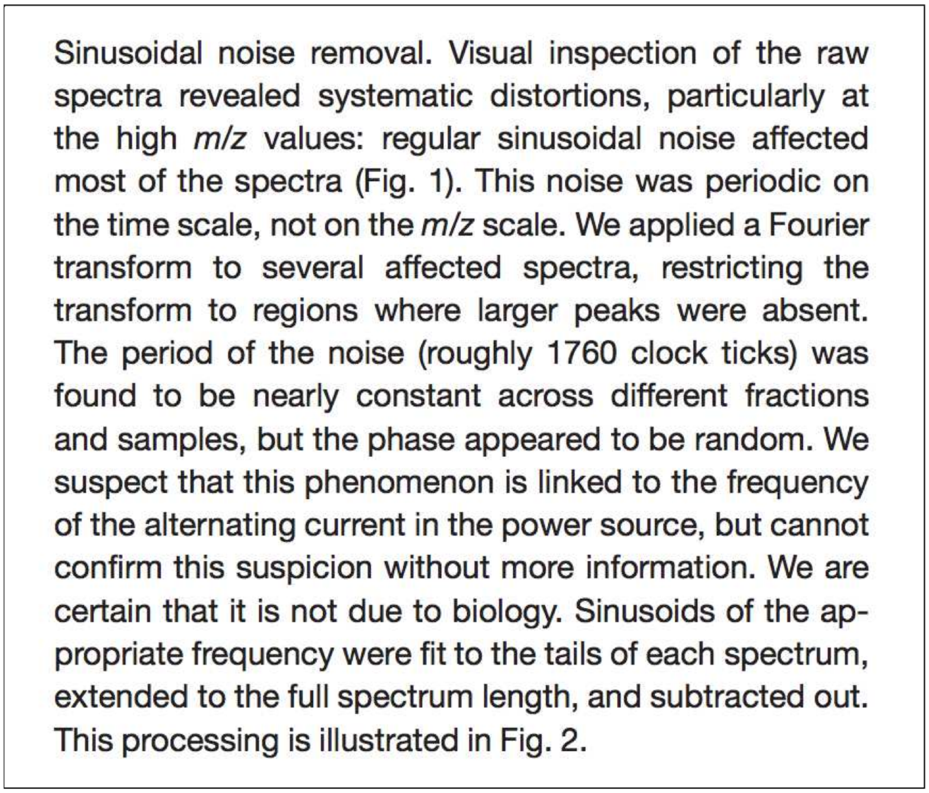
Keith A. Baggerly
Jeffrey S. Morris
Jing Wang
David Gold
Lian-Chun Xiao
Kevin R. Coombes

Department of Biostatistics,
UT M.D. Anderson
Cancer Center,
Houston, TX, USA

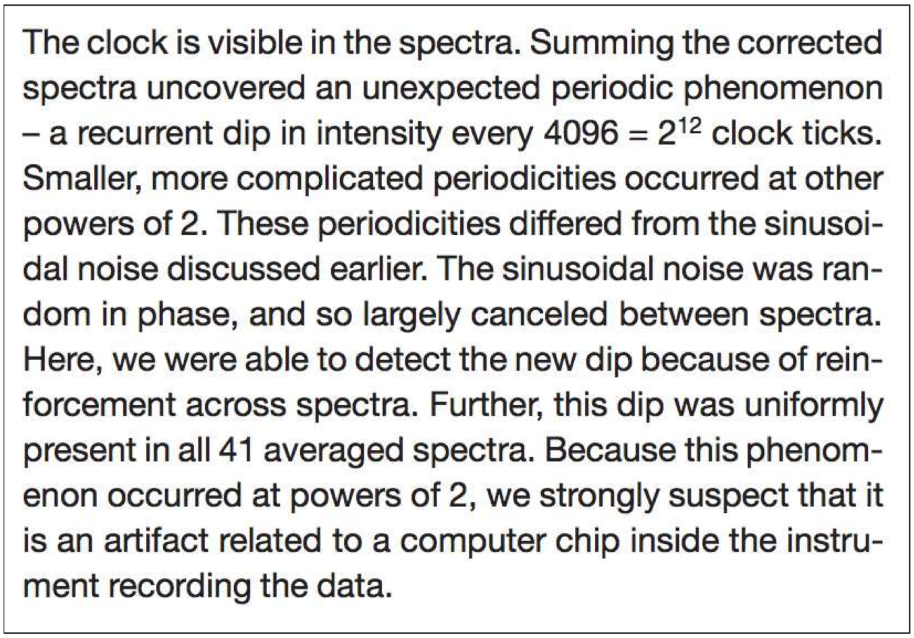
A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples

For our analysis of the data from the First Annual Proteomics Data Mining Conference, we attempted to discriminate between 24 disease spectra (group A) and 17 normal spectra (group B). First, we processed the raw spectra by (i) correcting for additive sinusoidal noise (periodic on the time scale) affecting most spectra, (ii) correcting for the overall baseline level, (iii) normalizing, (iv) recombining fractions, and (v) using variable-width windows for data reduction. Also, we identified a set of polymeric peaks (at multiples of 180.6 Da) that is present in several normal spectra (B1–B8). After data processing, we found the intensities at the following mass to charge (*m/z*) values to be useful discriminators: 3077, 12886 and 74263. Using these values, we were able to achieve an overall classification accuracy of 38/41 (92.6%). Perfect classification could be achieved by adding two additional peaks, at 2476 and 6955. We identified these values by applying a genetic algorithm to a filtered list of *m/z* values using Mahalanobis distance between the group means as a fitness function.

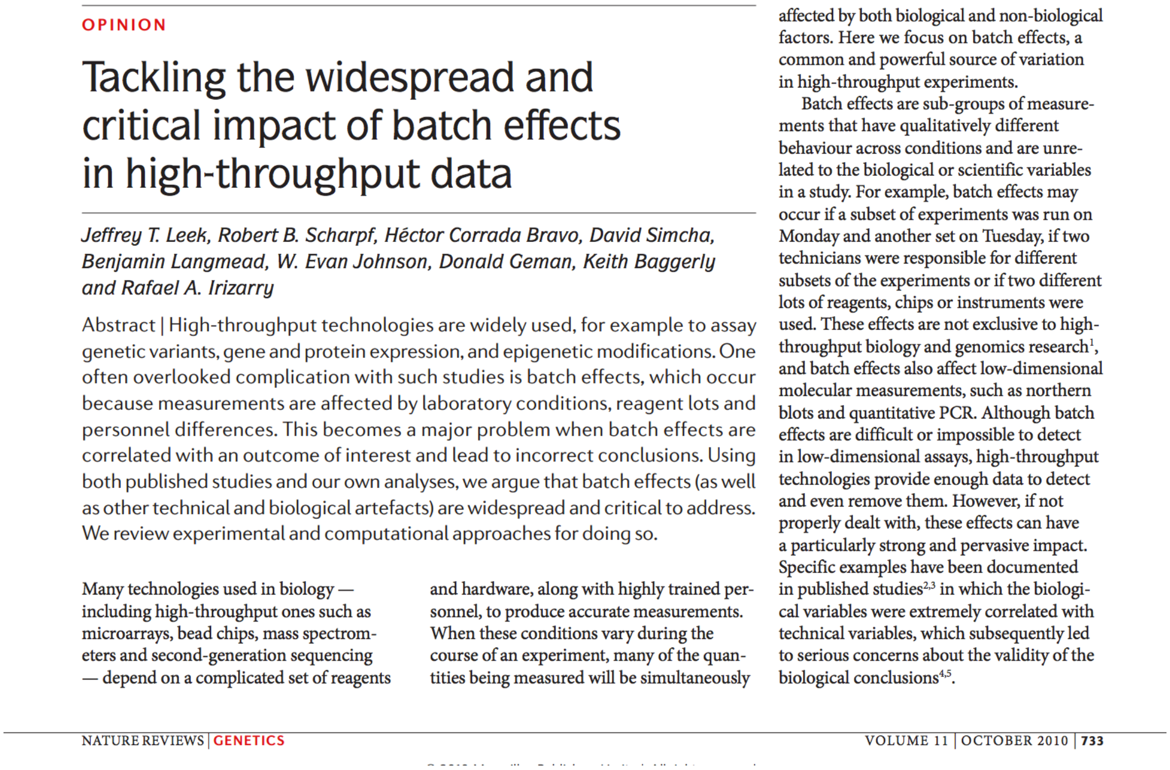
Keywords: Cross validation / Data cleaning / Discrimination / Genetic algorithm / Mahalanobis distance PRO 0522



Sinusoidal noise removal. Visual inspection of the raw spectra revealed systematic distortions, particularly at the high *m/z* values: regular sinusoidal noise affected most of the spectra (Fig. 1). This noise was periodic on the time scale, not on the *m/z* scale. We applied a Fourier transform to several affected spectra, restricting the transform to regions where larger peaks were absent. The period of the noise (roughly 1760 clock ticks) was found to be nearly constant across different fractions and samples, but the phase appeared to be random. We suspect that this phenomenon is linked to the frequency of the alternating current in the power source, but cannot confirm this suspicion without more information. We are certain that it is not due to biology. Sinusoids of the appropriate frequency were fit to the tails of each spectrum, extended to the full spectrum length, and subtracted out. This processing is illustrated in Fig. 2.



The clock is visible in the spectra. Summing the corrected spectra uncovered an unexpected periodic phenomenon – a recurrent dip in intensity every 4096 = 2^{12} clock ticks. Smaller, more complicated periodicities occurred at other powers of 2. These periodicities differed from the sinusoidal noise discussed earlier. The sinusoidal noise was random in phase, and so largely canceled between spectra. Here, we were able to detect the new dip because of reinforcement across spectra. Further, this dip was uniformly present in all 41 averaged spectra. Because this phenomenon occurred at powers of 2, we strongly suspect that it is an artifact related to a computer chip inside the instrument recording the data.



OPINION

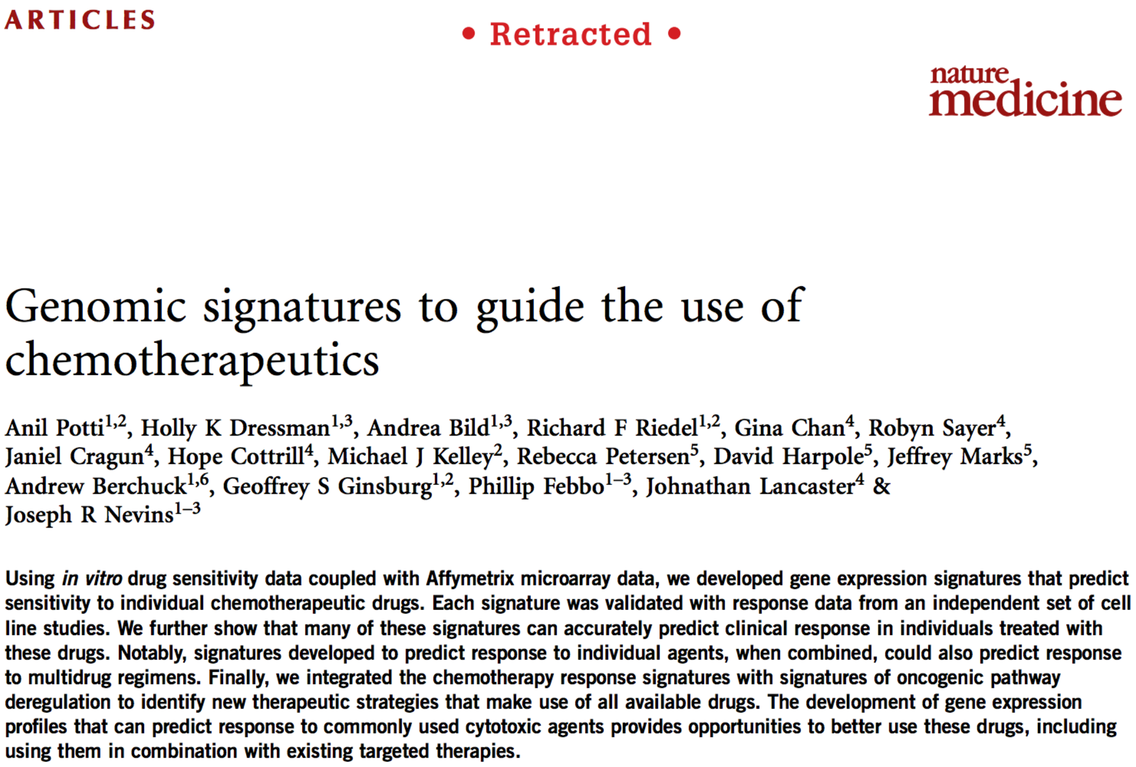
Tackling the widespread and critical impact of batch effects in high-throughput data

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

Abstract | High-throughput technologies are widely used, for example to assay genetic variants, gene and protein expression, and epigenetic modifications. One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address. We review experimental and computational approaches for doing so.

Many technologies used in biology — including high-throughput ones such as microarrays, bead chips, mass spectrometers and second-generation sequencing — depend on a complicated set of reagents and hardware, along with highly trained personnel, to produce accurate measurements. When these conditions vary during the course of an experiment, many of the quantities being measured will be simultaneously affected by both biological and non-biological factors. Here we focus on batch effects, a common and powerful source of variation in high-throughput experiments.

Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study. For example, batch effects may occur if a subset of experiments was run on Monday and another set on Tuesday, if two technicians were responsible for different subsets of the experiments or if two different lots of reagents, chips or instruments were used. These effects are not exclusive to high-throughput biology and genomics research[1], and batch effects also affect low-dimensional molecular measurements, such as northern blots and quantitative PCR. Although batch effects are difficult or impossible to detect in low-dimensional assays, high-throughput technologies provide enough data to detect and even remove them. However, if not properly dealt with, these effects can have a particularly strong and pervasive impact. Specific examples have been documented in published studies[2,3] in which the biological variables were extremely correlated with technical variables, which subsequently led to serious concerns about the validity of the biological conclusions[4,5].



ARTICLES

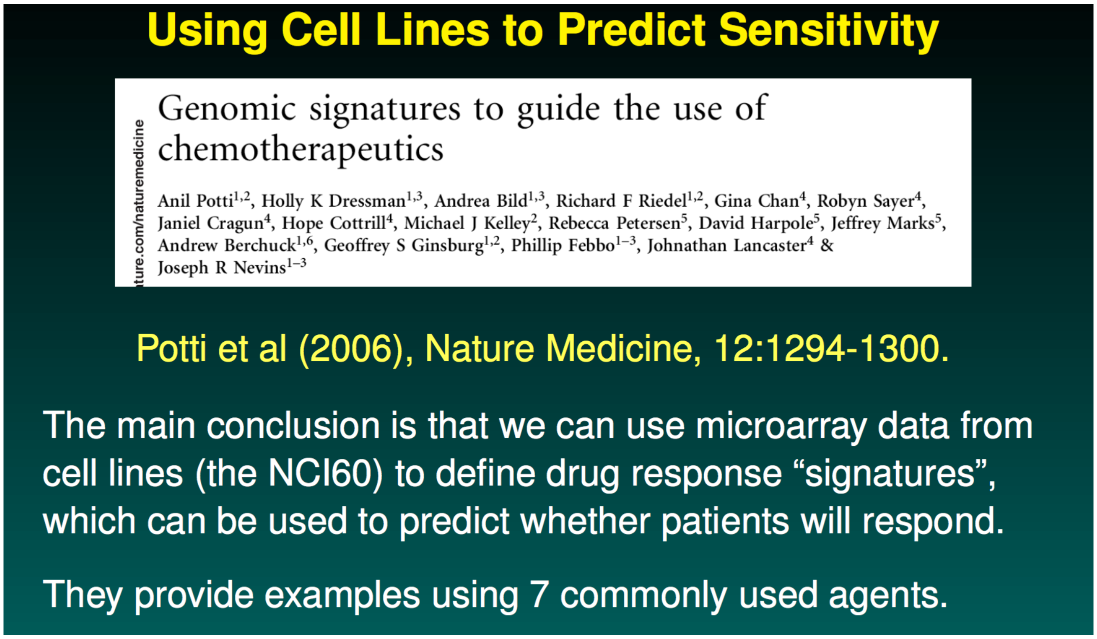
• Retracted •

nature medicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to commonly used cytotoxic agents provides opportunities to better use these drugs, including using them in combination with existing targeted therapies.



Using Cell Lines to Predict Sensitivity

Genomic signatures to guide the use of chemotherapeutics

ture.com/naturemedicine

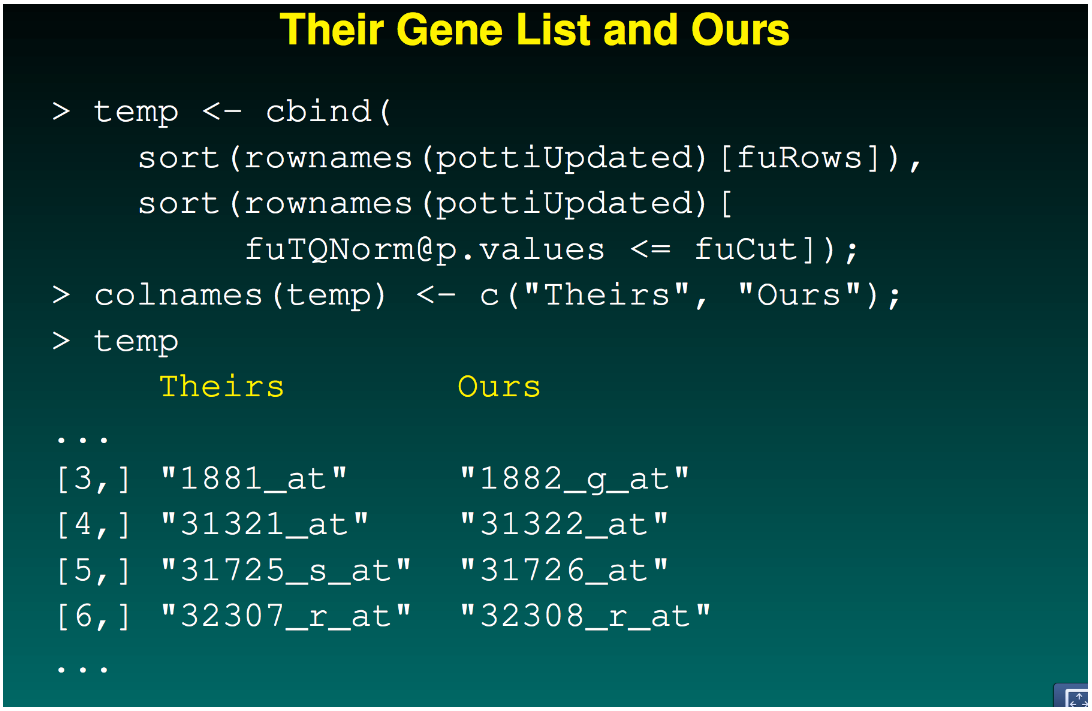
Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Potti et al (2006), Nature Medicine, 12:1294-1300.

The main conclusion is that we can use microarray data from cell lines (the NCI60) to define drug response “signatures”, which can be used to predict whether patients will respond.

They provide examples using 7 commonly used agents.

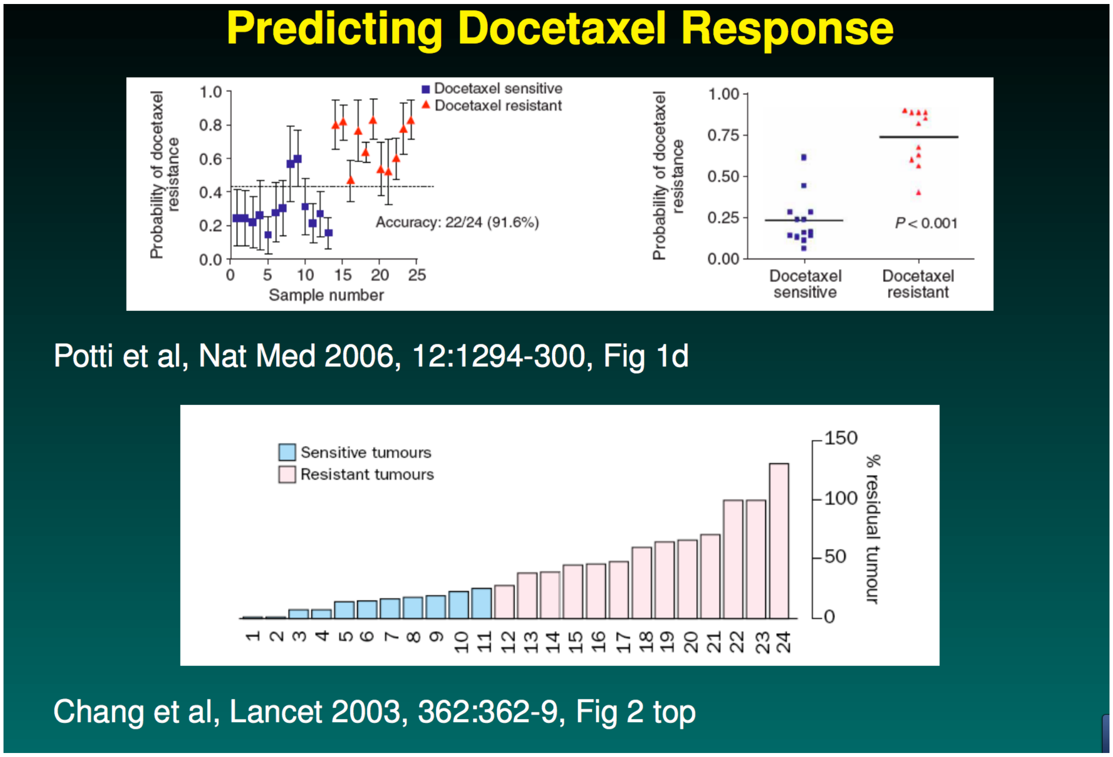
[Keith Baggerly]



Their Gene List and Ours

```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
         fuTQNorm@p.values <= fuCut]);
> colnames(temp) <- c("Theirs", "Ours");
> temp
      Theirs         Ours
...
[3,] "1881_at"     "1882_g_at"
[4,] "31321_at"    "31322_at"
[5,] "31725_s_at"  "31726_at"
[6,] "32307_r_at"  "32308_r_at"
...
```

[Keith Baggerly]

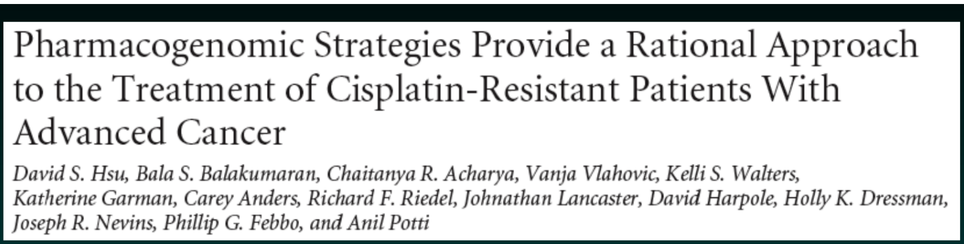


Predicting Docetaxel Response

Potti et al, Nat Med 2006, 12:1294-300, Fig 1d

Chang et al, Lancet 2003, 362:362-9, Fig 2 top

[Keith Baggerly]



Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo, and Anil Potti

J Clin Oncol, Oct 1, 2007, 25:4350-7.

Same approach, using Cisplatin and Pemetrexed.

For cisplatin, U133A arrays were used for training. ERCC1, ERCC4 and DNA repair genes are identified as “important”.

[Keith Baggerly]

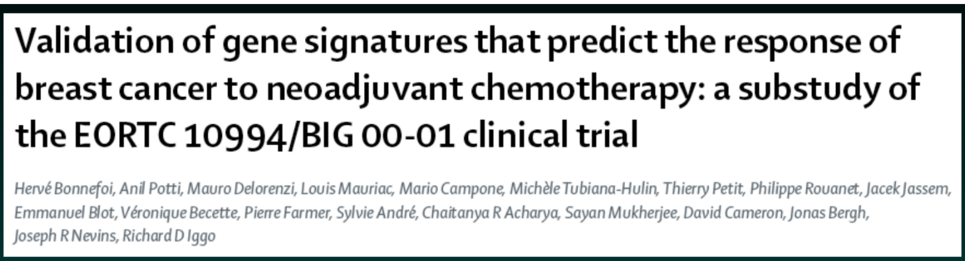
The 4 We Can’t Match (Reply)

203719_at, ERCC1,
210158_at, ERCC4,
228131_at, ERCC1, and
231971_at, FANCM (DNA Repair).

Another problem –

The last two probesets aren’t on the U133A arrays that were used. They’re on the U133B.

[Keith Baggerly]



Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Campone, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D Iggo

Lancet Oncology, Dec 2007, 8:1071-8. (early access Nov 14)

Similar approach, using signatures for Fluorouracil, Epirubicin, Cyclophosphamide, and Taxotere to predict response to combination therapies: FEC and TET.

[Keith Baggerly]

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to TET and FEC. Let P() indicate prob sensitive. The rules used are as follows.

$$P(TFAC) = P(T)+P(F)+P(A)+P(C)-P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

$$P(FEC) = \frac{5}{8}[P(F)+P(E)+P(C)] - \frac{1}{4}.$$

Each rule is different.

[Keith Baggerly]

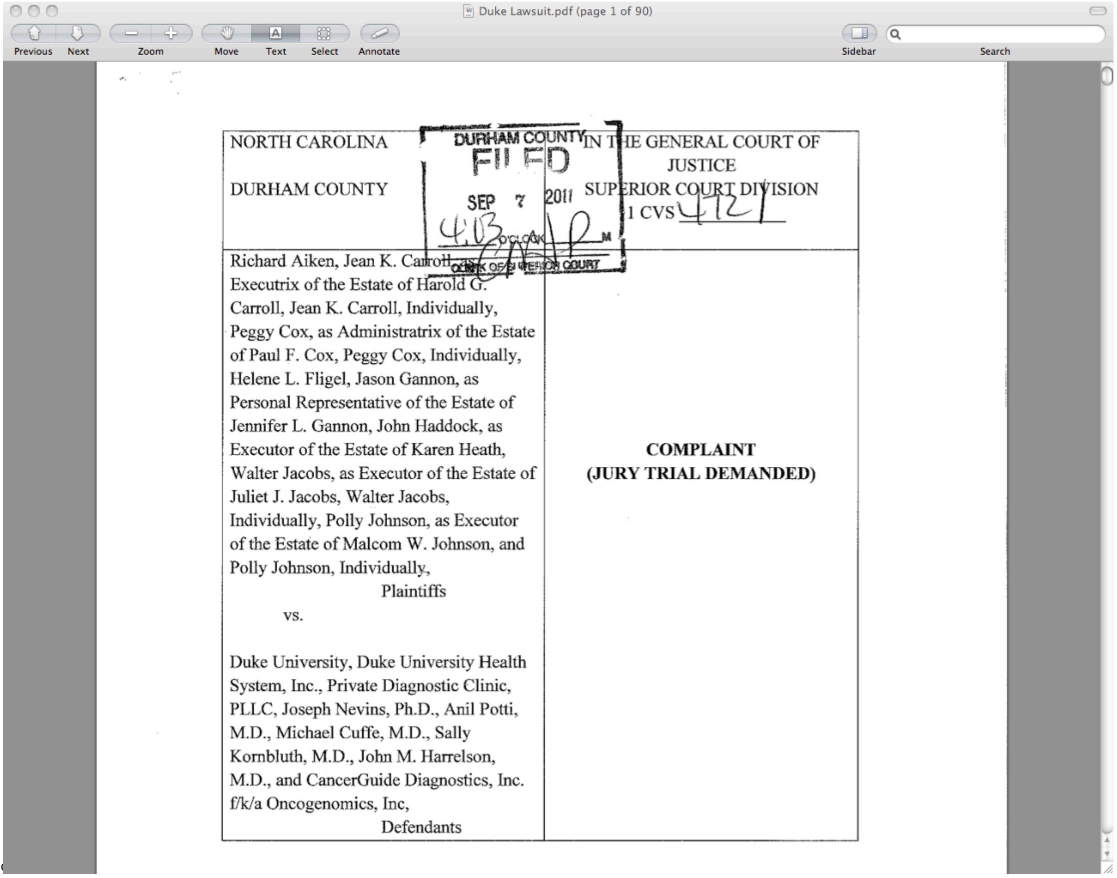
DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY* AND KEVIN R. COOMBES†

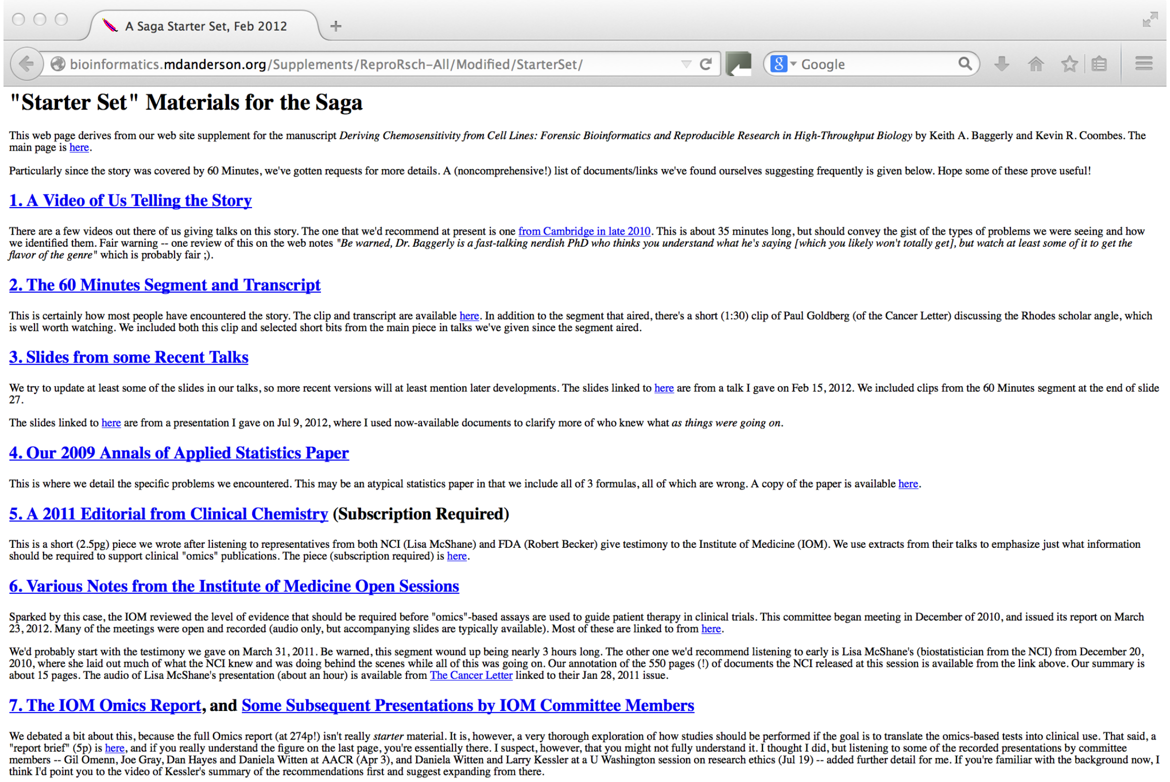
U.T. M.D. Anderson Cancer Center

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

projecteuclid.org/euclid.aoas/1267453942



NORTH CAROLINA DURHAM COUNTY	IN THE GENERAL COURT OF JUSTICE SUPERIOR COURT DIVISION 11 CVS 4721
Richard Aiken, Jean K. Carroll, as Executrix of the Estate of Harold G. Carroll, Jean K. Carroll, Individually, Peggy Cox, as Administratrix of the Estate of Paul F. Cox, Peggy Cox, Individually, Helene L. Fligel, Jason Gannon, as Personal Representative of the Estate of Jennifer L. Gannon, John Haddock, as Executor of the Estate of Karen Heath, Walter Jacobs, as Executor of the Estate of Juliet J. Jacobs, Walter Jacobs, Individually, Polly Johnson, as Executor of the Estate of Malcom W. Johnson, and Polly Johnson, Individually, Plaintiffs vs. Duke University, Duke University Health System, Inc., Private Diagnostic Clinic, PLLC, Joseph Nevins, Ph.D., Anil Potti, M.D., Michael Cuffe, M.D., Sally Kornbluth, M.D., John M. Harrelson, M.D., and CancerGuide Diagnostics, Inc. f/k/a Oncogenomics, Inc, Defendants	**COMPLAINT** **(JURY TRIAL DEMANDED)**


A Saga Starter Set, Feb 2012
bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/StarterSet/
Google

"Starter Set" Materials for the Saga

This web page derives from our web site supplement for the manuscript *Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology* by Keith A. Baggerly and Kevin R. Coombes. The main page is here.

Particularly since the story was covered by 60 Minutes, we've gotten requests for more details. A (noncomprehensive!) list of documents/links we've found ourselves suggesting frequently is given below. Hope some of these prove useful!

1. A Video of Us Telling the Story

There are a few videos out there of us giving talks on this story. The one that we'd recommend at present is one from Cambridge in late 2010. This is about 35 minutes long, but should convey the gist of the types of problems we were seeing and how we identified them. Fair warning -- one review of this on the web notes "*Be warned, Dr. Baggerly is a fast-talking nerdish PhD who thinks you understand what he's saying [which you likely won't totally get], but watch at least some of it to get the flavor of the genre*" which is probably fair ;).

2. The 60 Minutes Segment and Transcript

This is certainly how most people have encountered the story. The clip and transcript are available here. In addition to the segment that aired, there's a short (1:30) clip of Paul Goldberg (of the Cancer Letter) discussing the Rhodes scholar angle, which is well worth watching. We included both this clip and selected short bits from the main piece in talks we've given since the segment aired.

3. Slides from some Recent Talks

We try to update at least some of the slides in our talks, so more recent versions will at least mention later developments. The slides linked to here are from a talk I gave on Feb 15, 2012. We included clips from the 60 Minutes segment at the end of slide 27.

The slides linked to here are from a presentation I gave on Jul 9, 2012, where I used now-available documents to clarify more of who knew what *as things were going on*.

4. Our 2009 Annals of Applied Statistics Paper

This is where we detail the specific problems we encountered. This may be an atypical statistics paper in that we include all of 3 formulas, all of which are wrong. A copy of the paper is available here.

5. A 2011 Editorial from Clinical Chemistry (Subscription Required)

This is a short (2.5pg) piece we wrote after listening to representatives from both NCI (Lisa McShane) and FDA (Robert Becker) give testimony to the Institute of Medicine (IOM). We use extracts from their talks to emphasize just what information should be required to support clinical "omics" publications. The piece (subscription required) is here.

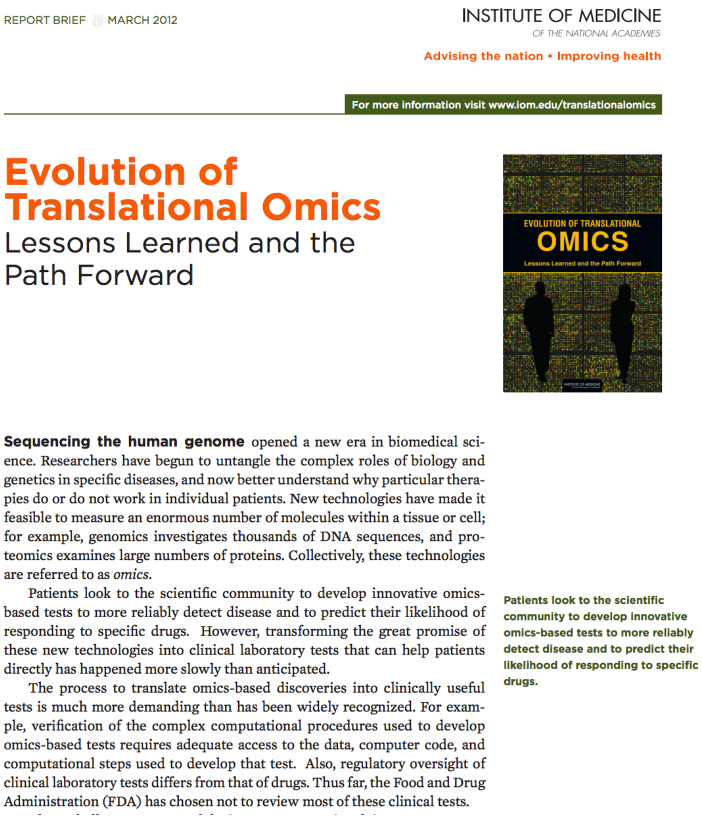
6. Various Notes from the Institute of Medicine Open Sessions

Sparked by this case, the IOM reviewed the level of evidence that should be required before "omics"-based assays are used to guide patient therapy in clinical trials. This committee began meeting in December of 2010, and issued its report on March 23, 2012. Many of the meetings were open and recorded (audio only, but accompanying slides are typically available). Most of these are linked to from here.

We'd probably start with the testimony we gave on March 31, 2011. Be warned, this segment wound up being nearly 3 hours long. The other one we'd recommend listening to early is Lisa McShane's (biostatistician from the NCI) from December 20, 2010, where she laid out much of what the NCI knew and was doing behind the scenes while all of this was going on. Our annotation of the 550 pages (!) of documents the NCI released at this session is available from the link above. Our summary is about 15 pages. The audio of Lisa McShane's presentation (about an hour) is available from The Cancer Letter linked to their Jan 28, 2011 issue.

7. The IOM Omics Report, and Some Subsequent Presentations by IOM Committee Members

We debated a bit about this, because the full Omics report (at 274p!) isn't really *starter* material. It is, however, a very thorough exploration of how studies should be performed if the goal is to translate the omics-based tests into clinical use. That said, a "report brief" (5p) is here, and if you really understand the figure on the last page, you're essentially there. I suspect, however, that you might not fully understand it. I thought I did, but listening to some of the recorded presentations by committee members -- Gil Omenn, Joe Gray, Dan Hayes and Daniela Witten at AACR (Apr 3), and Daniela Witten and Larry Kessler at a U Washington session on research ethics (Jul 19) -- added further detail for me. If you're familiar with the background now, I think I'd point you to the video of Kessler's summary of the recommendations first and suggest expanding from there.


EVOLUTION OF TRANSLATIONAL
OMICS
Lessons Learned and the Path Forward
INSTITUTE OF MEDICINE

REPORT BRIEF | MARCH 2012

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES
Advising the nation • Improving health

For more information visit www.iom.edu/translationalomics

Evolution of Translational Omics

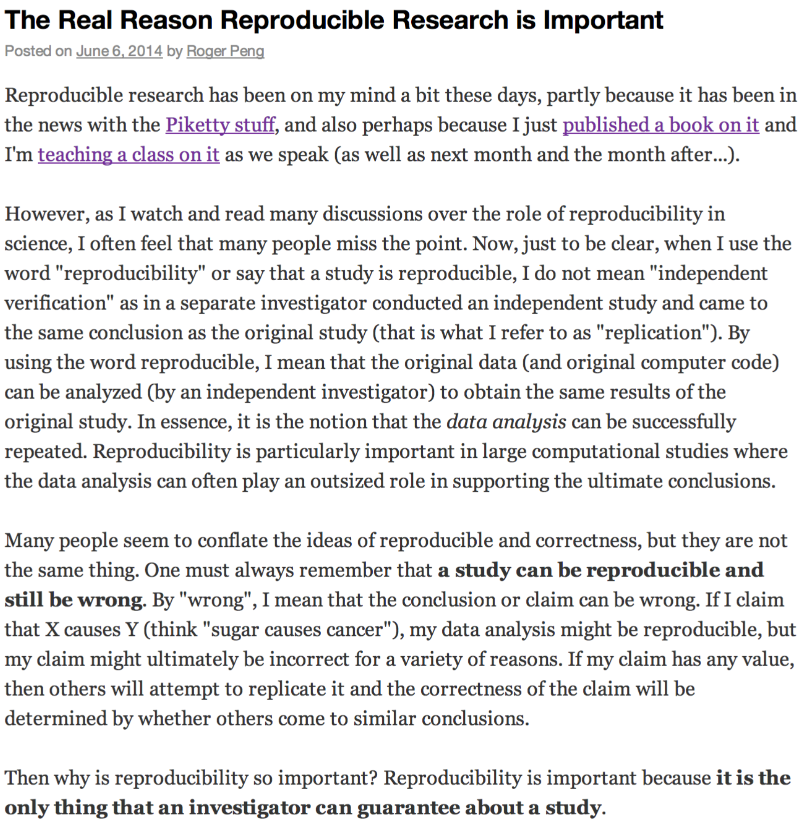
Lessons Learned and the Path Forward

Sequencing the human genome opened a new era in biomedical science. Researchers have begun to untangle the complex roles of biology and genetics in specific diseases, and now better understand why particular therapies do or do not work in individual patients. New technologies have made it feasible to measure an enormous number of molecules within a tissue or cell; for example, genomics investigates thousands of DNA sequences, and proteomics examines large numbers of proteins. Collectively, these technologies are referred to as *omics*.

Patients look to the scientific community to develop innovative omics-based tests to more reliably detect disease and to predict their likelihood of responding to specific drugs. However, transforming the great promise of these new technologies into clinical laboratory tests that can help patients directly has happened more slowly than anticipated.

The process to translate omics-based discoveries into clinically useful tests is much more demanding than has been widely recognized. For example, verification of the complex computational procedures used to develop omics-based tests requires adequate access to the data, computer code, and computational steps used to develop that test. Also, regulatory oversight of clinical laboratory tests differs from that of drugs. Thus far, the Food and Drug Administration (FDA) has chosen not to review most of these clinical tests.

Patients look to the scientific community to develop innovative omics-based tests to more reliably detect disease and to predict their likelihood of responding to specific drugs.



The Real Reason Reproducible Research is Important

Posted on June 6, 2014 by Roger Peng

Reproducible research has been on my mind a bit these days, partly because it has been in the news with the Piketty stuff, and also perhaps because I just published a book on it and I'm teaching a class on it as we speak (as well as next month and the month after...).

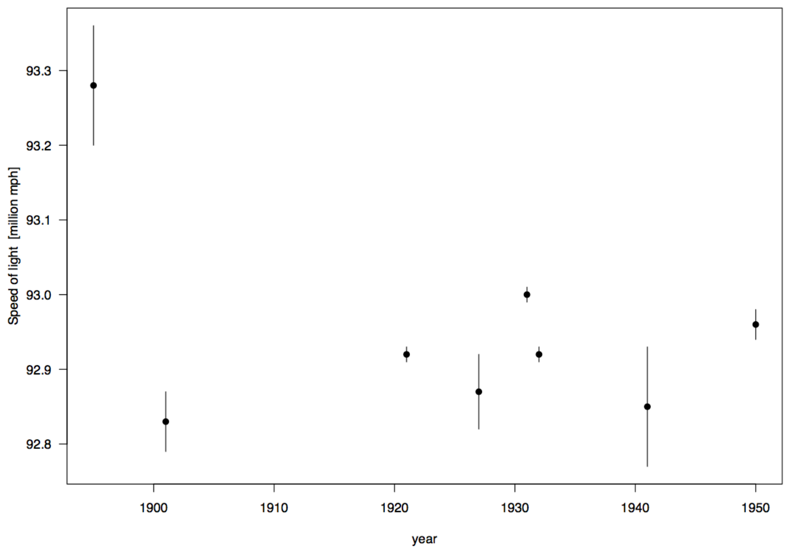
However, as I watch and read many discussions over the role of reproducibility in science, I often feel that many people miss the point. Now, just to be clear, when I use the word "reproducibility" or say that a study is reproducible, I do not mean "independent verification" as in a separate investigator conducted an independent study and came to the same conclusion as the original study (that is what I refer to as "replication"). By using the word reproducible, I mean that the original data (and original computer code) can be analyzed (by an independent investigator) to obtain the same results of the original study. In essence, it is the notion that the *data analysis* can be successfully repeated. Reproducibility is particularly important in large computational studies where the data analysis can often play an outsized role in supporting the ultimate conclusions.

Many people seem to conflate the ideas of reproducible and correctness, but they are not the same thing. One must always remember that **a study can be reproducible and still be wrong**. By "wrong", I mean that the conclusion or claim can be wrong. If I claim that X causes Y (think "sugar causes cancer"), my data analysis might be reproducible, but my claim might ultimately be incorrect for a variety of reasons. If my claim has any value, then others will attempt to replicate it and the correctness of the claim will be determined by whether others come to similar conclusions.

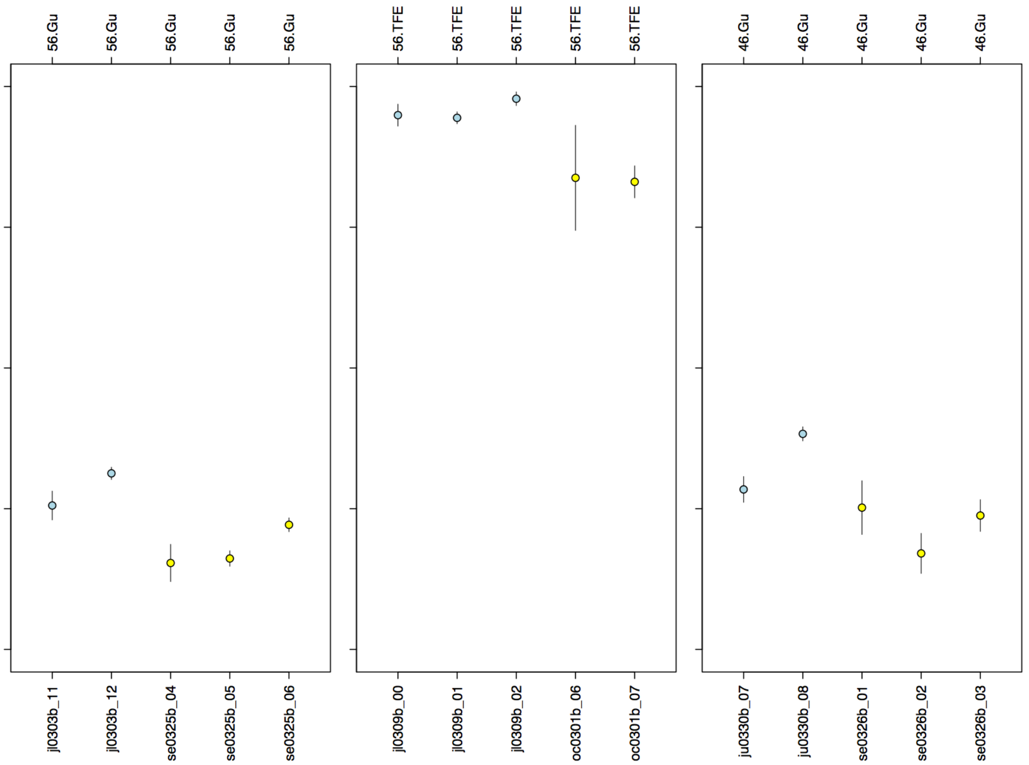
Then why is reproducibility so important? Reproducibility is important because **it is the only thing that an investigator can guarantee about a study**.

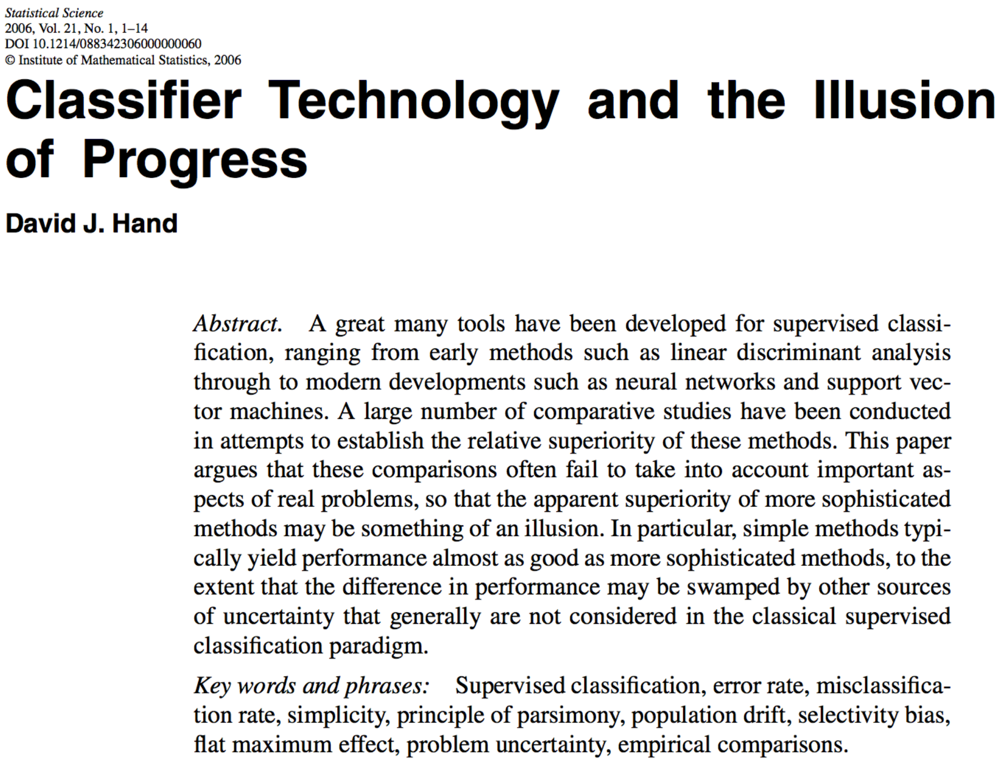
simplystatistics.org

Estimates of the speed of light, with “confidence intervals”.



Youden W (1972). *Technometrics* 14: 1-11.





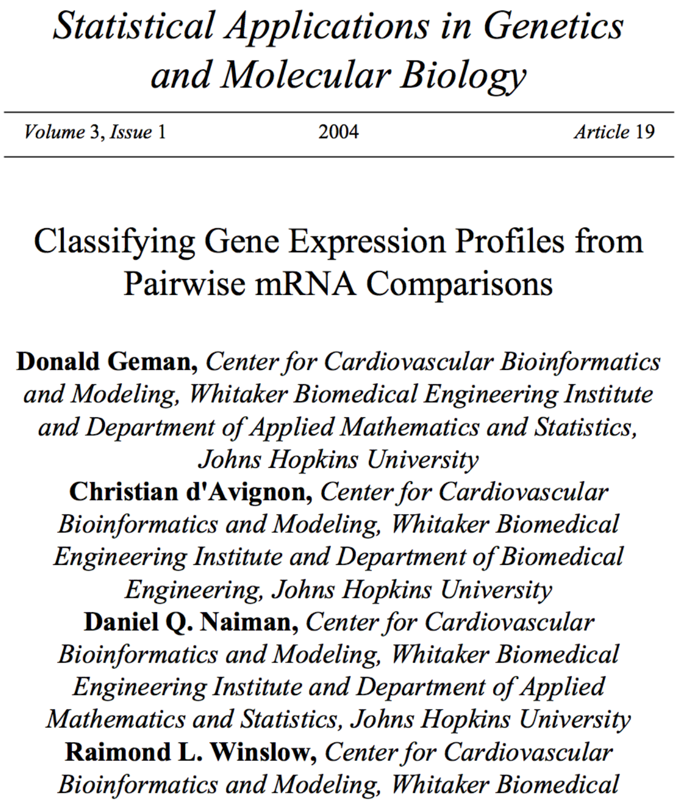
Statistical Science
2006, Vol. 21, No. 1, 1–14
DOI 10.1214/088342306000000060
© Institute of Mathematical Statistics, 2006

Classifier Technology and the Illusion of Progress

David J. Hand

Abstract. A great many tools have been developed for supervised classification, ranging from early methods such as linear discriminant analysis through to modern developments such as neural networks and support vector machines. A large number of comparative studies have been conducted in attempts to establish the relative superiority of these methods. This paper argues that these comparisons often fail to take into account important aspects of real problems, so that the apparent superiority of more sophisticated methods may be something of an illusion. In particular, simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.

Key words and phrases: Supervised classification, error rate, misclassification rate, simplicity, principle of parsimony, population drift, selectivity bias, flat maximum effect, problem uncertainty, empirical comparisons.

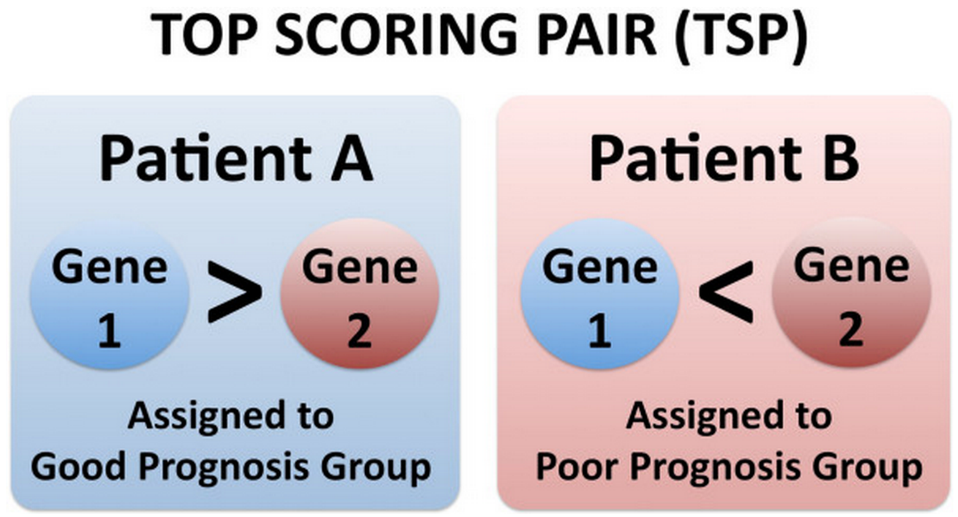


Statistical Applications in Genetics and Molecular Biology

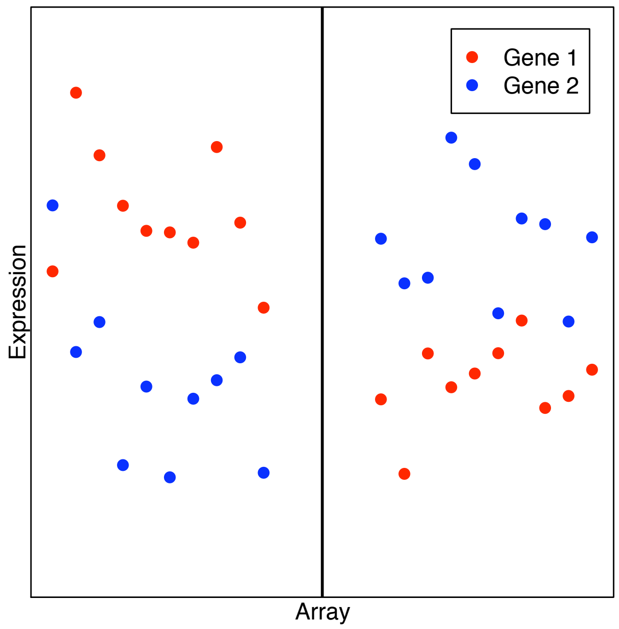
Volume 3, *Issue* 1 2004 *Article* 19

Classifying Gene Expression Profiles from Pairwise mRNA Comparisons

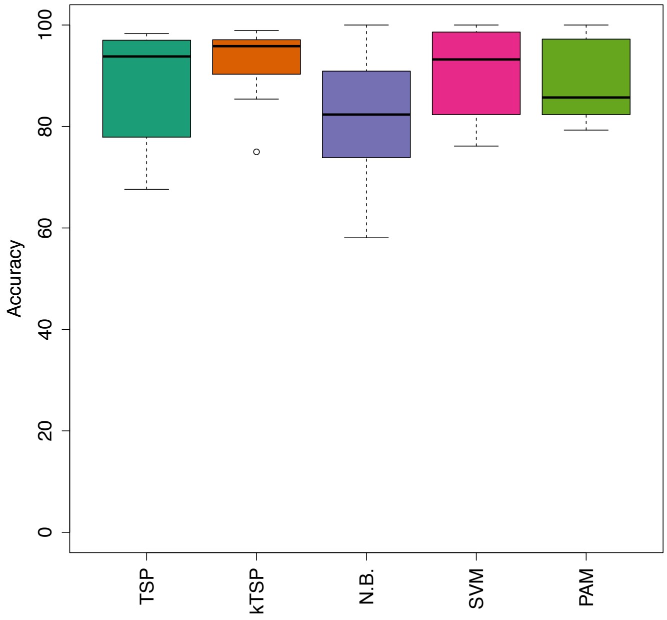
Donald Geman, *Center for Cardiovascular Bioinformatics and Modeling, Whitaker Biomedical Engineering Institute and Department of Applied Mathematics and Statistics, Johns Hopkins University*
Christian d'Avignon, *Center for Cardiovascular Bioinformatics and Modeling, Whitaker Biomedical Engineering Institute and Department of Biomedical Engineering, Johns Hopkins University*
Daniel Q. Naiman, *Center for Cardiovascular Bioinformatics and Modeling, Whitaker Biomedical Engineering Institute and Department of Applied Mathematics and Statistics, Johns Hopkins University*
Raimond L. Winslow, *Center for Cardiovascular Bioinformatics and Modeling, Whitaker Biomedical*



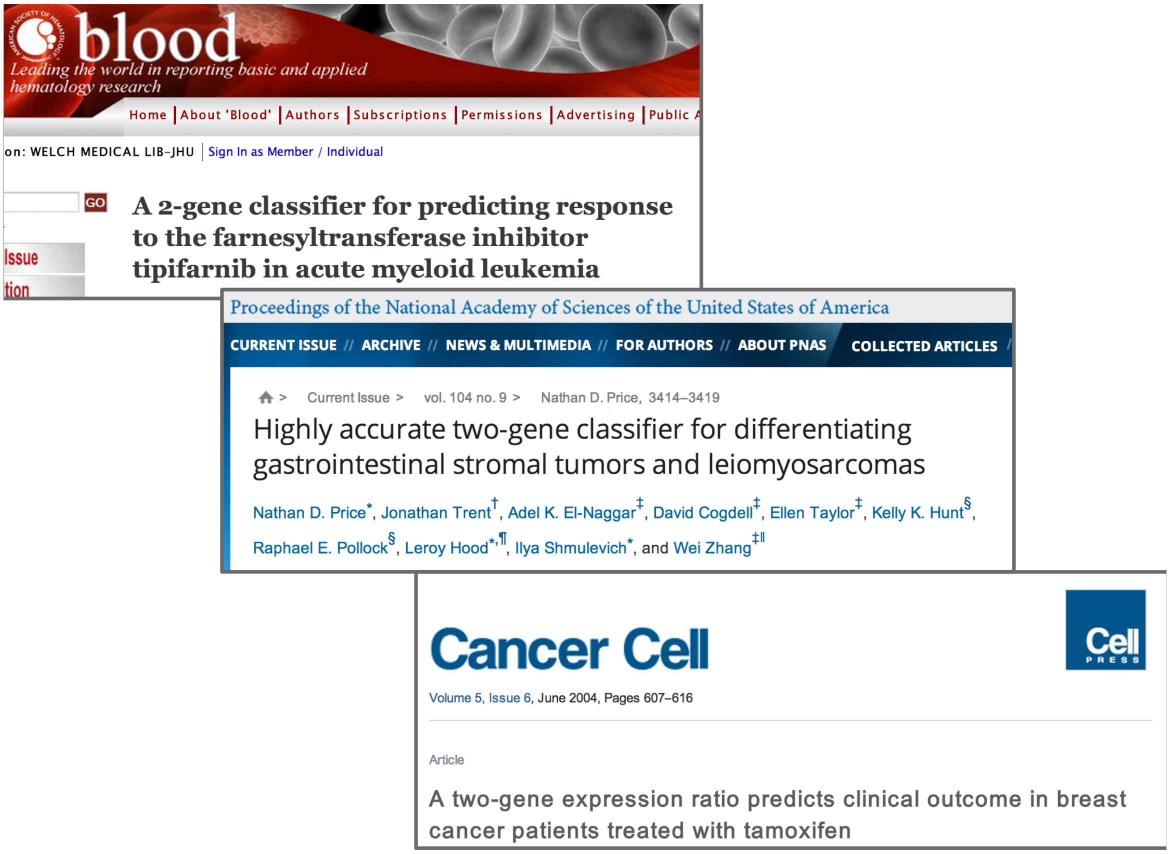
PMID 23682826


● Gene 1
● Gene 2
Expression
Array

[JL]


Accuracy
0
20
40
60
80
100
TSP
kTSP
N.B.
SVM
PAM

[JL]


blood
Leading the world in reporting basic and applied hematology research
Home | About 'Blood' | Authors | Subscriptions | Permissions | Advertising | Public A
on: WELCH MEDICAL LIB–JHU | Sign In as Member / Individual
GO
Issue
tion

A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia

Proceedings of the National Academy of Sciences of the United States of America
CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // FOR AUTHORS // ABOUT PNAS
COLLECTED ARTICLES
> Current Issue > vol. 104 no. 9 > Nathan D. Price, 3414–3419

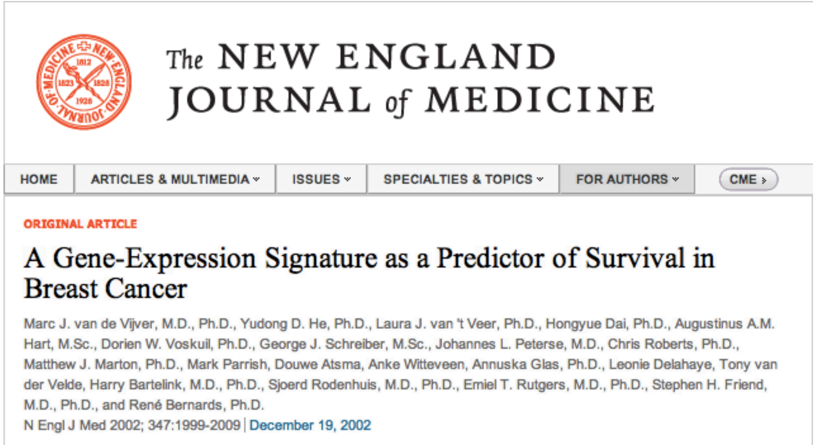
Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas

Nathan D. Price*, Jonathan Trent†, Adel K. El-Naggar‡, David Cogdell‡, Ellen Taylor‡, Kelly K. Hunt§, Raphael E. Pollock§, Leroy Hood*,¶, Ilya Shmulevich*, and Wei Zhang‡‖

Volume 5, Issue 6, June 2004, Pages 607–616

Article

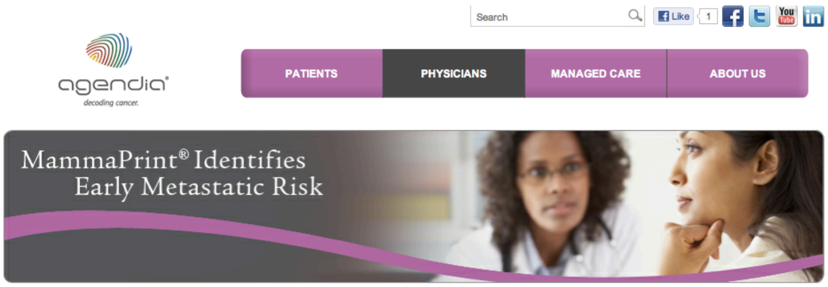
A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen

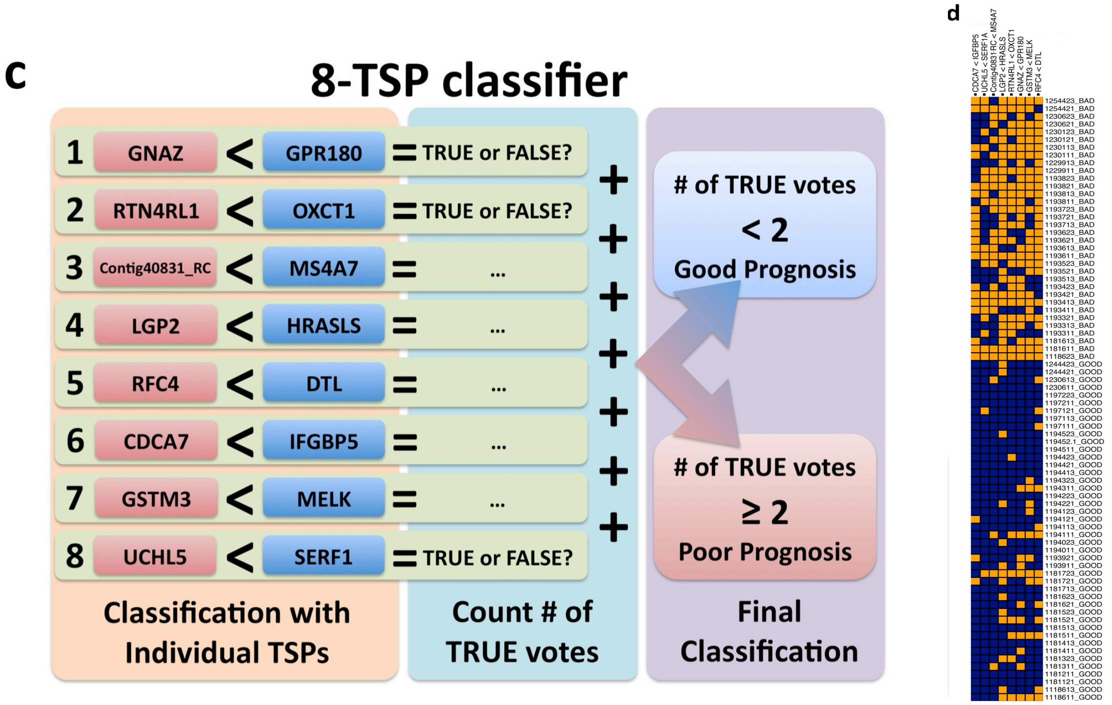


A Gene-Expression Signature as a Predictor of Survival in Breast Cancer

Marc J. van de Vijver, M.D., Ph.D., Yudong D. He, Ph.D., Laura J. van 't Veer, Ph.D., Hongyue Dai, Ph.D., Augustinus A.M. Hart, M.Sc., Dorien W. Voskuil, Ph.D., George J. Schreiber, M.Sc., Johannes L. Peterse, M.D., Chris Roberts, Ph.D., Matthew J. Marton, Ph.D., Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Ph.D., Leonie Delahaye, Tony van der Velde, Harry Bartelink, M.D., Ph.D., Sjoerd Rodenhuis, M.D., Ph.D., Emiel T. Rutgers, M.D., Ph.D., Stephen H. Friend, M.D., Ph.D., and René Bernards, Ph.D.

N Engl J Med 2002; 347:1999-2009 | December 19, 2002





c

8-TSP classifier

1 GNAZ < GPR180 = TRUE or FALSE? +

2 RTN4RL1 < OXCT1 = TRUE or FALSE? +

3 Contig40831_RC < MS4A7 = … +

4 LGP2 < HRASLS = … +

5 RFC4 < DTL = … +

6 CDCA7 < IFGBP5 = … +

7 GSTM3 < MELK = … +

8 UCHL5 < SERF1 = TRUE or FALSE?

# of TRUE votes < 2 Good Prognosis

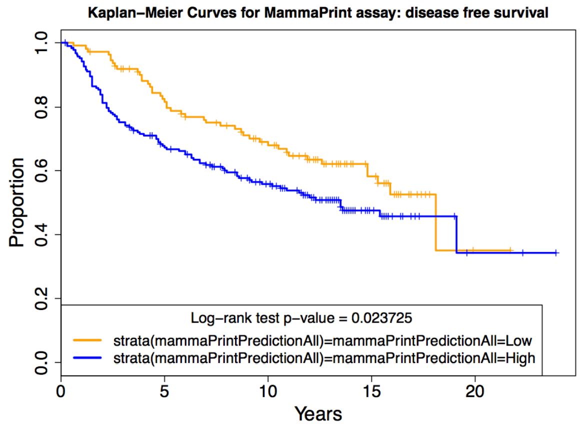
# of TRUE votes ≥ 2 Poor Prognosis

Classification with Individual TSPs

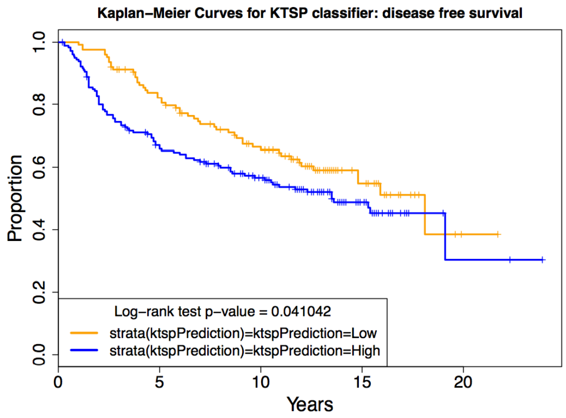
Count # of TRUE votes

Final Classification

PMID 23682826



Kaplan–Meier Curves for MammaPrint assay: disease free survival



Kaplan–Meier Curves for KTSP classifier: disease free survival

PMID 23682826