Observational Studies

Controlled experiment:

The investigator chooses who receives the treatment.

Observational study:

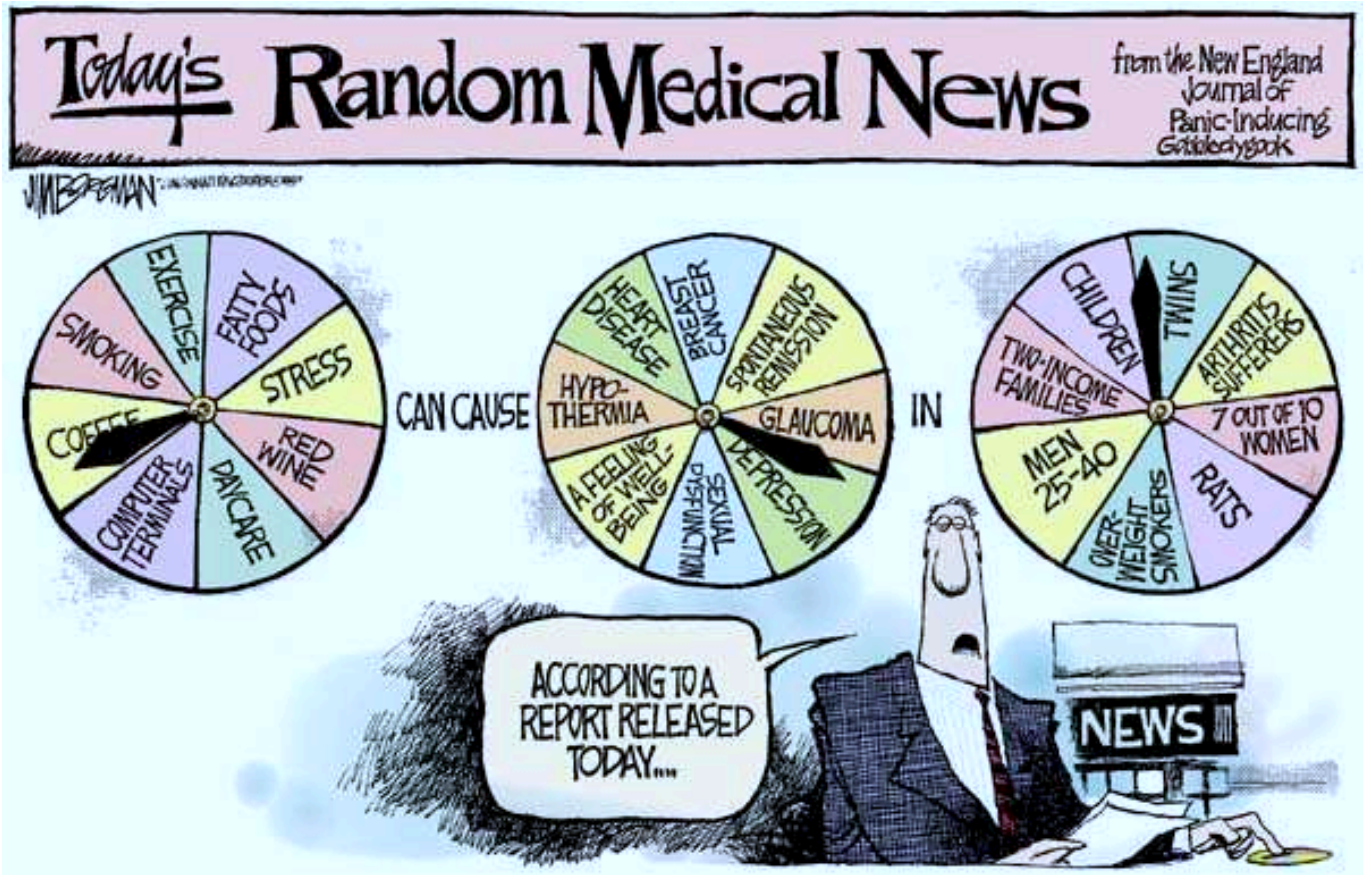
The investigator doesn't choose who receives the treatment.

The subjects themselves might choose whether they receive the treatment).

Key issues:

- Correlation (association) is not causation.
- Confounding
- Simpson's paradox

Researchers have shown that...



Causes of association

Suppose A is associated with B.

This may be because:

- A causes B
- B causes A
- X is associated with both A and B.

Confounding

In the association of A and B, X is a confounder if it is associated with both A and B.

X need not be a cause of either A or B.

For example, in the consideration of smoking and lung cancer, a gene which

- causes smoking but is not associated with lung cancer is not a confounder.
- causes lung cancer but is not associated with smoking is not a confounder.
- causes lung cancer and is associated with smoking is a confounder.

Controlling for a confounder

The problem with observational studies is that subjects differ among themselves in crucial ways besides the treatment.

We deal with this by controlling for the confounding variable(s)—we compare smaller, more homogeneous subgroups.

- Anticipate, measure, and control for possible confounders.
- Think about other possible confounders that were not considered.
- Never draw very strong conclusions from a single observational study.

Sex bias in graduate admissions

Observational study on sex bias in graduate admissions at the University of California, Berkeley.

During the study period:

- 44% (of 8,442) men were admitted
- 35% (of 4,321) women were admitted

Does this indicate a sex bias?

Sex bias in graduate admissions

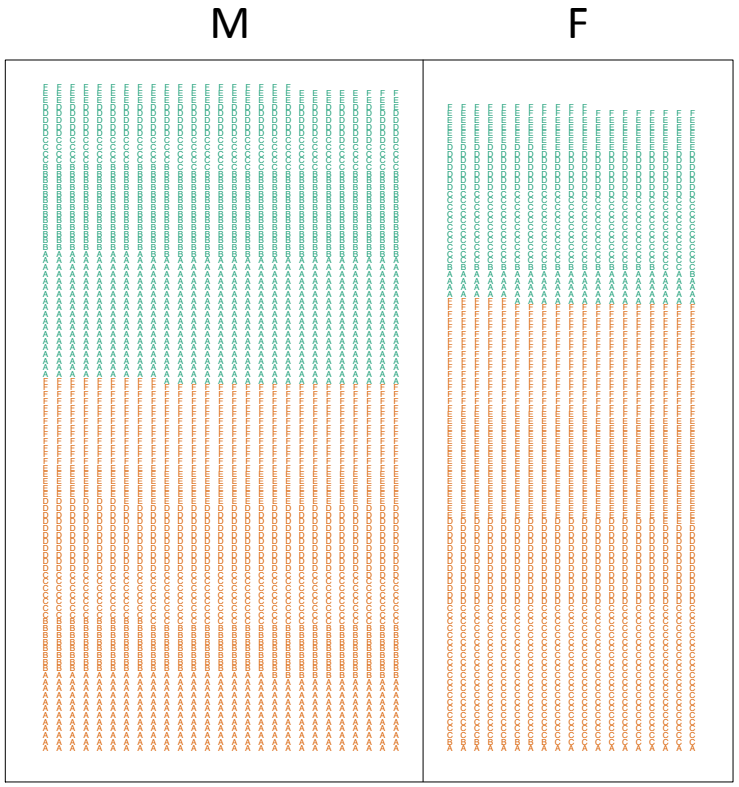
For the six largest majors:

	Men		Women	
Major	Number of applicants	Percent admitted	Number of applicants	Percent admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Simpson's paradox:

Relationships between variables within subgroups can be reversed when the subgroups are combined.

Confounding



[RI]

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)

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.

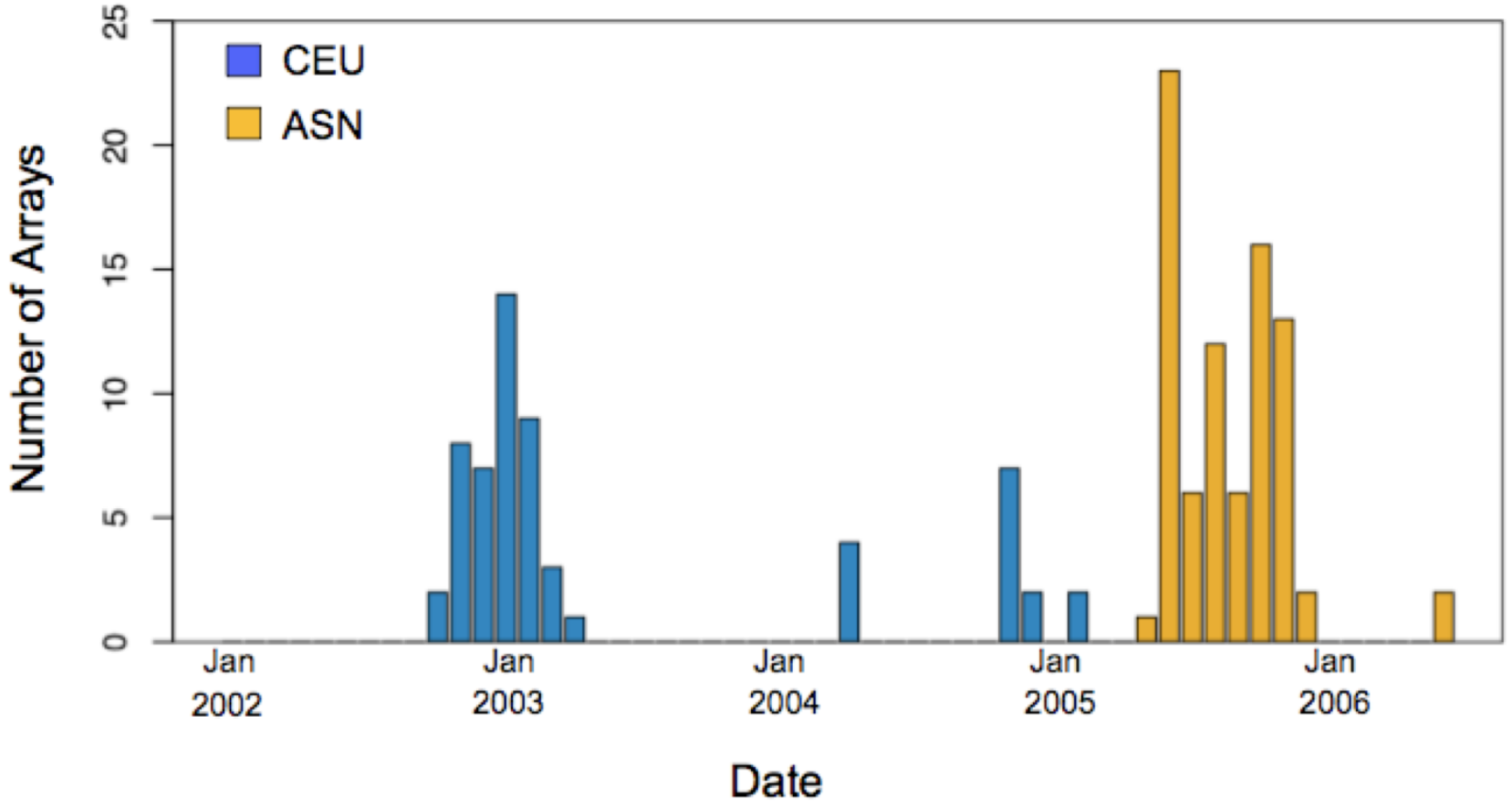
genetic diseases. The marked population differences in prevalence of these qualitative phenotypes (such as cystic fibrosis[9] and Tay-Sachs disease[10]) are entirely due to differences in frequencies of the mutant alleles. However, genetic differences among populations in quantitative phenotypes are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

Confounding of population and processing time



[JL]

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).

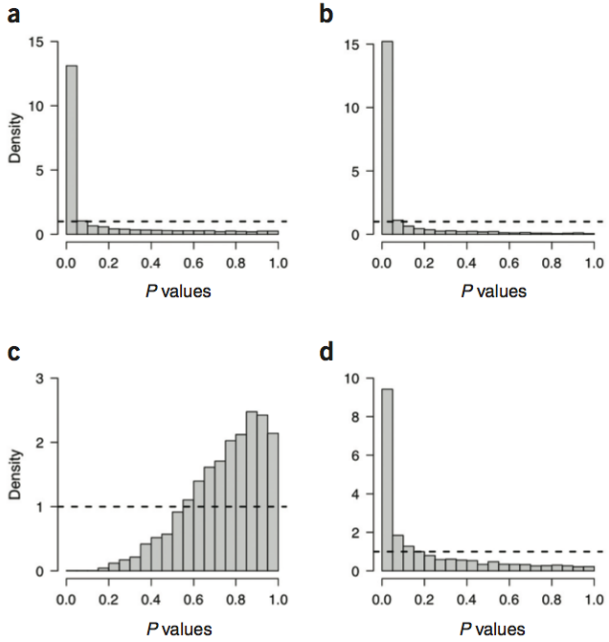
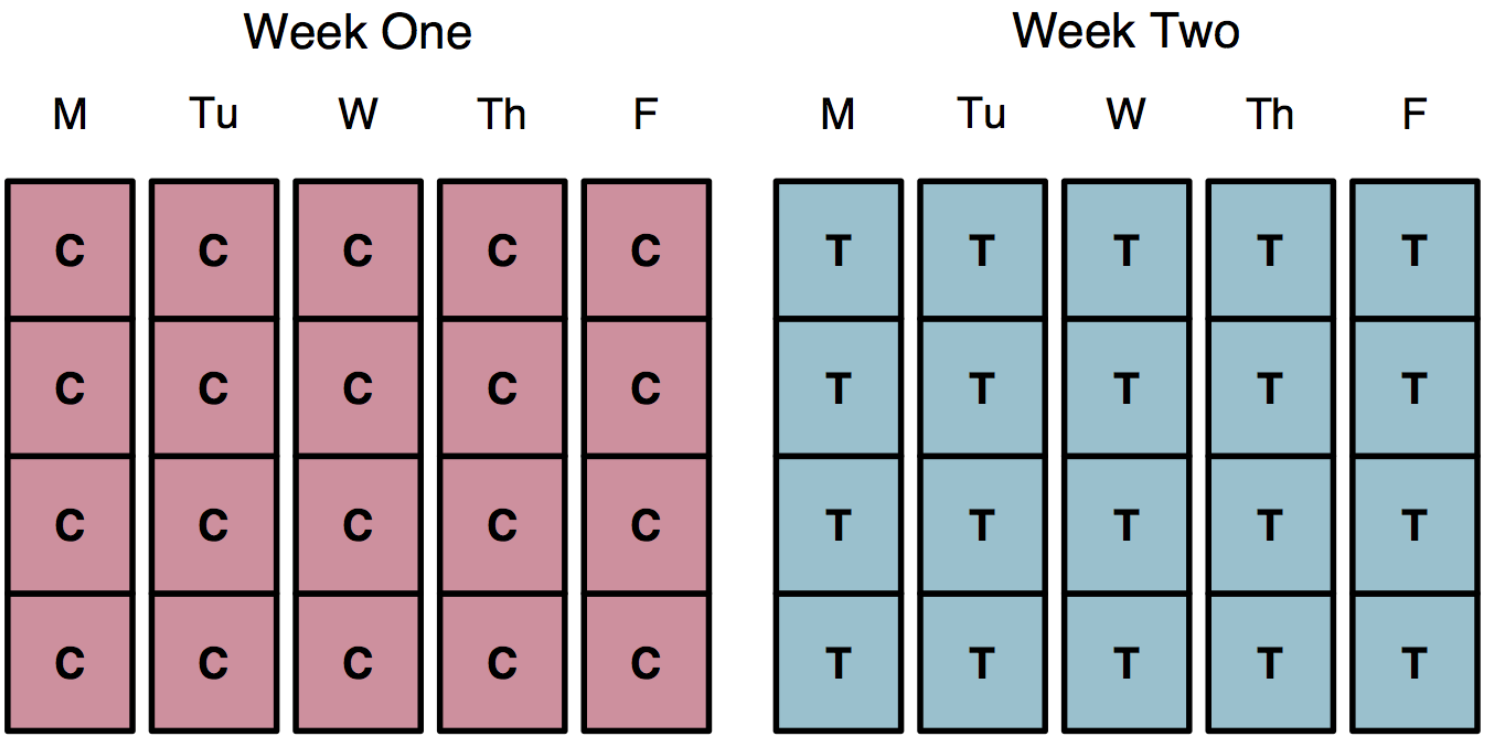


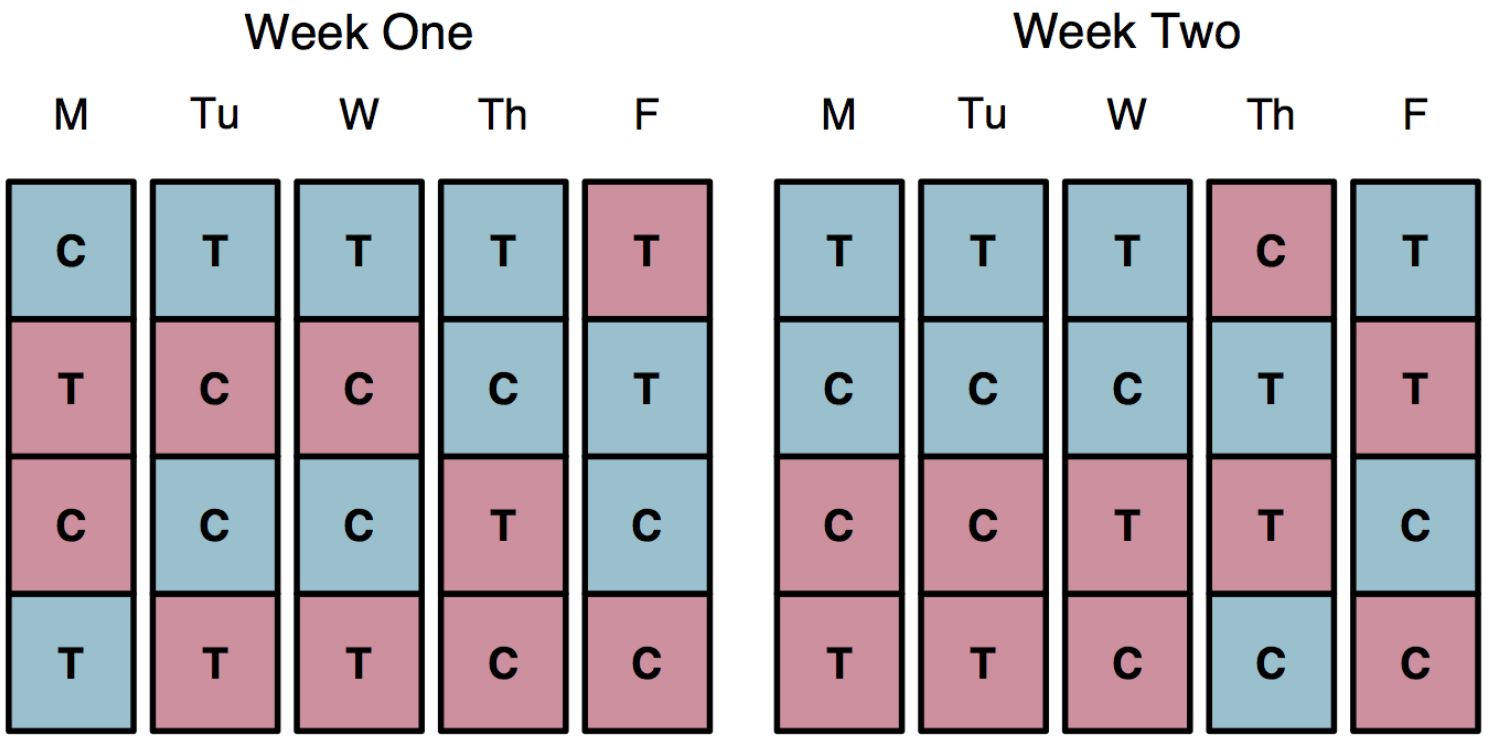
Figure 1 Distribution of *P* values for tests of differential expression. (**a**) *P* values resulting from tests of differential expression between the CEU and ASN samples. (**b**) *P* values resulting from tests of differential expression with respect to year in which the microarrays were processed. (**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



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



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

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

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