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).

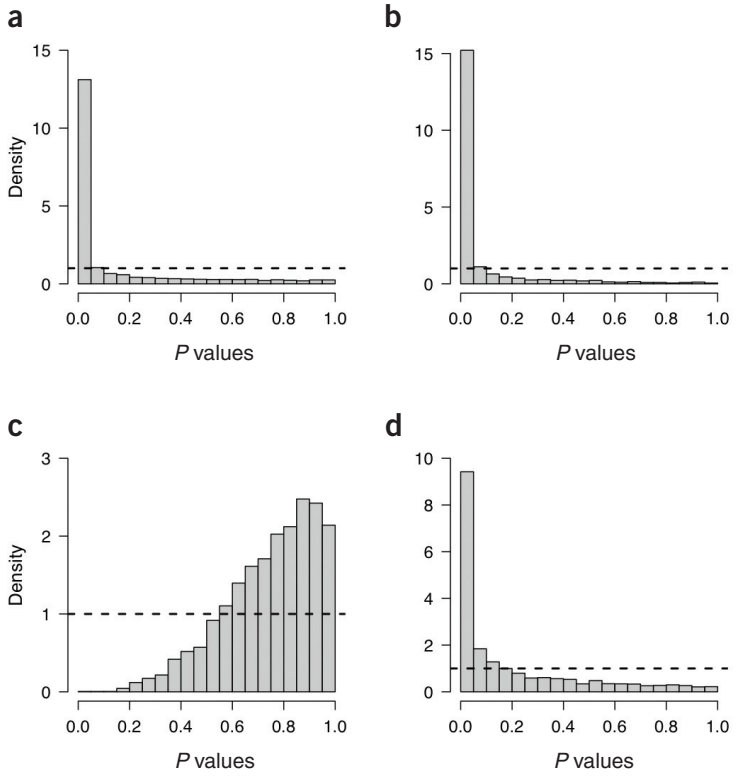

Figure 1 Distribution of *P* values for tests of differential expression. (**a**) *P* values resulting from tests of differential expression between the CEU and ASN samples. (**b**) *P* values resulting from tests of differential expression with respect to year in which the microarrays were processed.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

genes are differentially expressed between individuals of European and African ancestry[2]. To rule out the possibility that this difference is related to the larger sample size of Spielman *et al.*[1] and is thus not simply a power issue, we randomly sampled eight CEU and eight ASN individuals (corresponding to the sample size of Storey *et al.*[2]) 1,000 times, and for each sample we estimated the proportion of differentially expressed genes as described above. The average proportion was 43% (standard deviation (s.d.) = 8%), demonstrating that differences in power between our study and Spielman *et al.*[1] do not fully account for differences in the estimated proportion of differentially expressed genes among human populations.

A possible explanation for the pervasive signature of differential expression observed in Spielman *et al.* is a systematic bias introduced during sample preparation or microarray expression measurements. The authors clearly recognize the importance of controlling for systematic confounding variables, as they state that "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." In addition to sample processing, it is widely known that technical variation can also be introduced through batch-to-batch variation in microarray manufacturing and through day-to-day laboratory conditions under which hybridization is performed[5]. To explore these issues in more detail, we downloaded the raw CEL files from Gene Expression Omnibus (GSE5859) and extracted from the header line the date on which the file was created. Interestingly, the arrays used to measure expression for the CEU individuals were primarily processed from 2003 to 2004, whereas the arrays used to measure expression for the ASN individuals were all processed in 2005–2006.

We tested for differential expression with respect to the year in which the microarrays were processed and found that at least 94% of genes are estimated to be differentially expressed (**Fig. 1b**). Typically, one would take these batch effects into account before performing any differential expression analyses. When we used a standard method to do so[6], we find no evidence for differential expression between populations (**Fig. 1c**), which is not surprising, given that microarray batch effects seem to be completely confounded with population effects. Obviously, these findings do not mean that all differentially expressed genes in Spielman *et al.*[1] are due to batch effects; rather, the source of population differences in expression cannot be uniquely attributed to biological causes. To gain insight into the magnitude of the batch effects in Spielman *et al.*[1], we tested for differential expression within the CEU sample with respect to the year of processing. Strikingly, 79% of genes among CEU individuals are estimated to be differentially expressed between processing years (**Fig. 1d**). Collectively, these results suggest that the expression data analyzed in Spielman *et al.*[1] possess systematic and uncorrectable bias, raising concerns about the accuracy of their reported results.

The genotype-phenotype correlations made with the data set of Spielman *et al.* should also be viewed with caution. Specifically, because batch effect appears to be the major source of differential expression, any marker with allele frequency differences among batches is therefore also vulnerable to confounding when testing for genotype-phenotype correlations. Even though our primary purpose here was to explore potential explanations for the large number of genes estimated to be differentially expressed between the CEU and ASN samples, the consequences of microarray batch effects on the gene expression association results of Spielman *et al.*[1] also warrant further investigation.

In summary, characterizing patterns of gene expression variation within and among human populations is an important and interesting problem. However, it is critical that experimental design and statistical analyses be carefully thought out and implemented in order for accurate conclusions to be drawn. In particular, components of variation from both measured and unmeasured variables must be taken into account, for example, by balancing or randomizing the study design with respect to sex, time of sample preparation and processing and microarray batch.

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

[1]Department of Genome Sciences and [2]Department of Biostatistics, University of Washington, 1705 NE Pacific St., Seattle, Washington 98195, USA.
e-mail: akeyj@u.washington.edu or jstorey@u.washington.edu

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

1. Spielman, R.S. *et al. Nat. Genet.* **39**, 226–231 (2007).
2. Storey, J.D. *et al. Am. J. Hum. Genet.* **80**, 502–509 (2007).
3. Storey, J.D. *J. R. Stat. Soc. Ser. B* **64**, 479–498 (2002).
4. Storey, J.D. *et al. Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
5. Churchill, G.A. *Nat. Genet.* **32** (Suppl.), 490–495 (2002).
6. Jin, W. *et al. Nat. Genet.* **29**, 389–395 (2001).

Spielman and Cheung reply:

In our paper[1], we found that mean gene expression differed significantly between European-derived and Asian-derived populations for approximately 25% of 4,197 genes tested. For the expression phenotypes with the strongest evidence of polymorphic *cis* determinants, the differences in mean expression were largely explained by differences in the frequency of specific polymorphic variants.

Akey *et al.*[2] reanalyzed our data and asserted that their findings "suggest that the expression data of Spielman *et al.* possess systematic and uncorrectable bias, raising concerns about the accuracy of [the reported results of Spielman *et al.*]." To reach this conclusion, Akey *et al.*[2] restricted their analysis to just one part of our results[1]. Here, we first comment on their analysis and conclusions. We then describe results from a published study by others[3] that confirms our findings. Finally, we show how other aspects of our paper contradict the conclusions of Akey *et al.*[2] and support the "accuracy of [our] reported results."

Akey *et al.*[2] obtained our earlier data[1] from the Gene Expression Omnibus (GEO). They found that the expression arrays used for the HapMap CEU sample (of European ancestry) were processed in 2003–2004 (except for four individuals). In contrast, the arrays used for the HapMap CHB+JPT samples (their 'ASN' group, of Chinese and Japanese ancestry) were processed in 2005–2006. Akey *et al.* point out, correctly, that "microarray batch effects appear to be completely confounded with population effects."

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 to mislead. Other CEPH cell lines—not HapMap CEU—were grown in the same batch as CHB+JPT, which gave rise to our