Epistatic and Environmental Control of Genome-Wide Gene Expression

Timothy P. York,[1,4] Michael F. Miles,[2] Kenneth S. Kendler,[3] Colleen Jackson-Cook,[4] Melissa L. Bowman,[2] and Lindon J. Eaves[1,3,4]

1 Massey Cancer Center, Virginia Commonwealth University, Richmond, Virginia, United States of America

2 Departments of Pharmacology/Toxicology, Neurology and the Center for Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, United States of America

3 Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, United States of America

4 Department of Human Genetics, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, United States of America

All etiological studies of complex human traits focus on analyzing the causes of variation. Given this complexity, there is a premium on studying those processes that mediate between gene products and cellular or organismal phenotypes. Studies of levels of gene expression could offer insight into these processes and are likely to be especially useful to the extent that the major sources of their variation are known in normal tissues. The classical study of monozygotic (MZ) and dizygotic (DZ) twins was employed to partition the genetic and environmental influences in gene expression for over 6500 human genes measured using microarrays from lymphoblastoid cell lines. Our results indicate that mean expression levels are correlated about .3 in monozygotic (MZ) and .0 in dizygotic (DZ) twins suggesting an overall epistatic regulation of gene expression. Furthermore, the functions of several of the genes whose expression was most affected by environmental effects, after correction for measurement error, were consistent with their known role in mediating sensitivity to environmental influences.

The measurement of gene expression levels for a very large number of genes through microarray technology offers the hope of a better understanding of some of the primary mechanisms through which both genes and environment affect susceptibility to complex multifactorial disorders. If microarray technology is to provide a means for the assessment of the endophenotype in studies of disease etiology, it will be necessary to characterize the differential roles of genetic and environmental influences on levels of gene expression.

The study of gene expression levels in samples of monozygotic (MZ) and dizygotic (DZ) twin pairs may offer a valuable overview of the contribution of genetic variation and environmental exposures to gene expression, and identifies clusters of genes whose expression in specific tissues is more or less sensitive to influences of genetic or environmental variation. The study of MZ and DZ twins is a well-tested approach to partition the genetic, shared and nonshared environmental factors in multifactorial traits (Jinks & Fulker, 1970). Modern refinements of statistical methods for the analysis of multivariate twin data (Neale & Cardon, 1992) have made it possible for the twin study to resolve genetic (or environmental) effects that contribute to the clustering of multiple phenotypes and to the sequence of developmental events that lead ultimately to clinically significant complex outcomes.

Numerous studies of human disease have utilized microarray technology to simultaneously measure thousands of gene expression profiles between groups of individuals or cell types with the goal of identifying candidate disease genes or classifying tissue into disease subgroups. It is not clear whether or not the interindividual variation associated with a particular outcome is due to genetic or random environmental causes (Oleksiak et al., 2002). Genetic sources of gene expression variation are likely due to DNA sequence polymorphisms, whereas all nongenetic components can be generally classified as random environmental sources. Recently, investigators have studied the extent to which gene expression levels are under genetic control for a limited number of genes (Cheung et al., 2003; Yan et al., 2002). In another report, a significant heritable component for many genes was suggested by mapping both cis- and trans-acting loci that effect gene expression (Morely et al., 2004).

The examination of interindividual variation in the level of gene expression in normal tissue is a nec-

Received 23 November, 2004; accepted 2 December, 2004.

Address for correspondence: Timothy P. York, Virginia Institute for Psychiatric & Behavioral Genetics, Virginia Commonwealth University, PO Box 980003, Richmond, VA 23298-0003, USA. E-mail: tpyork@vcu.edu

essary prerequisite for interpreting alterations in gene expression profiles that are causally associated with disease tissue. Peripheral blood is a readily accessible source of cells for the study of this relationship and has the potential for acting as a surrogate tissue for diagnostic purposes since it is exposed to many of the same environmental substances as target tissues (Whitney et al., 2003). Recent studies have demonstrated that stable gene expression profiles can be monitored despite temporal changes in blood chemistry (Radich et al., 2004) and signatures of gene expression due to environmental exposures can be reliably identified (Lampe et al., 2004).

Ultimately, the same methodology may be extended to quantify the genetic and environmental sources of variation at the intermediary stage of gene expression and present a clearer picture of the etiology and molecular pathology of multifactorial disease.

To provide proof of principle for the use of twins in the study of gene expression, we have examined the patterns of correlation in expression levels of over 22,000 genes in cultured lymphoblastoid cell lines derived from peripheral blood lymphocytes from 10 pairs of MZ and 5 pairs of DZ female twins that were obtained from Coriell Cell Repositories (Camden, NJ). The mean twin pair age was 31 (± 12) years old. Individuals had no reported history of major illness (Appendix Table 1). Gene expression was measured on oligonucleotide microarrays (Human GeneChip™ HG133-A, Affymetrix) which contain over 22,000 named genes and expressed sequence tags (ESTs). Labeled cRNA from each individual was hybridized to a single microarray. On average, 41% of genes present on the microarrays were expressed in the sample cell lines. Genes without a MAS 5 'present' or 'marginally present' call in all samples were removed from analysis (Affymetrix, 1999). A total of 6578 genes met this very conservative filtering criteria and were included in this study.

Materials and Methods

Sample

Cultured lymphoblastoid cell lines were obtained from 10 pairs of MZ and 5 pairs of DZ twins. These cells were obtained from Coriell Cell Repositories (Camden, NJ). Zygosity was confirmed by genotyping a panel of 20 microsatellite markers. There were no discordant markers within MZ twin pairs. The cell cultures were maintained in RPMI 1640 growth media (supplemented with 10% fetal bovine serum, 1% Penicillin/Streptomycin, and 0.5% L-glutamine) and were incubated at 37°C with 5% CO_2 and 95% humidity. When sufficient growth was observed the cultures were subdivided. For each twin, a total of four subcultured flasks (T75 flasks in 50 ml supplemented media) were prepared. We recognize that variation in the level of gene expression explained by genetic or environmental factors may be specific to lymphoblastoid cell lines. Future studies are needed to understand the correlation in gene expression between transformed cells and peripheral blood leukocytes.

Microarrays

Total RNA was isolated according to the STAT-60™ protocol. RNA concentration was determined by absorbance at 260 nm, and RNA quality was analyzed by agarose gel electrophoresis and 260/280 absorbance ratios. Total RNA (7 µg) derived from each sample was reverse transcribed into double-stranded cDNA using the Invitrogen Superscript II System (Invitrogen, Carlsbad, CA). Biotin-labeled cRNA was synthesized from this cDNA using BioArray High Yield RNA Transcript Labeling Kit (ENZO Diagnostics, Farmingdale, NY) according to the manufacturer's instructions, purified using the RNAeasy Mini Kit (Qiagen, Mountain View, CA), and quantified by absorbance at 260 nm. Labeled cRNA samples were hybridized to oligonucleotide microarrays (Human GeneChip™ HG-133A, Affymetrix, Santa Clara, CA) containing probes for over 22,000 genes and ESTs. Array hybridization, scanning and quality control assessment were performed according to the manufacturer's protocol and as described previously (Hassan et al., 2003; Thibault et al., 2000). Background correction, normalization, and expression summary were conducted with the robust multiarray average (RMA; Irizarry et al., 2003) implemented in the Bioconductor R packages (Gentleman et al., 2004).

Statistical Methodology

We derived the variance components separately for both zygosities to test the hypothesis that no differences exist within and between pairs across all genes (probesets) measured. The model was $y_{ijk} = m + P_i + P{:}I_{i(j)} + G_k + (P{*}G)_{ik} + (P{:}I{*}G)_{i(j)k}$. The response, $y_{i(j)k}$, is the level of intensity for the j^{th} individual in the i^{th} pair and k^{th} gene. The variable m is the mean gene expression level across all individuals. The variation between pairs is listed as P_i and the within pair variance is represented by the nested term $P{:}I_{i(j)}$. G_k is the average gene expression intensity across pairs and individuals within pairs. The interaction terms, $(P{*}G)_{ik}$ and $(P{:}I{*}G)_{i(j)k}$, account for the gene dependent variation between and within pairs respectively. These terms are used to gauge the overall importance and underlying mechanism of genetic sources of variation. The $(P{:}I{*}G)_{i(j)k}$ term also contains the estimate of measurement error.

Evidence for epistatic mechanisms responsible for the overall control of gene expression can be observed if the average MZ intraclass correlation is significantly more than twice the average DZ intraclass correlation. A test statistic was calculated by

$$t = \left(\bar{r}_{MZ} - 2 * \bar{r}_{DZ}\right) \Big/ \left(\sqrt{\left(s^2_{MZ}/N\right) + 4 * \left(s^2_{DZ}/N\right)}\right)$$

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

where the average MZ and DZ intraclass correlations were computed over 6578 genes.

Broad sense heritabilities, which are linear combinations of the MZ and DZ intraclass correlations, were estimated for each gene to measure the proportion of genetic influences on gene expression. The conservative approach was taken of only estimating heritabilities for those genes whose ratio of MZ to DZ intraclass correlations were between 1 and 4.

The overall contributions of environmental factors to gene expression were estimated for individual genes using ANOVA methods. Environmental sources of variation can be calculated by estimating the within pair variance in MZ twins and correcting for experimental error. The model used was $y_{i(j)k} = m + P_i + I{:}P_{i(j)} + R_k + (P{*}R)_{ik} + (P{:}I{*}R)_{i(j)k}$, and fitted separately to all genes in the filtered gene set. In this model $y_{i(j)k}$ is the gene expression level for the j^{th} individual in the i^{th} pair and k^{th} probe processed using the RMA algorithm. The variable R_k is the average intensity level for the perfect match probe k across pairs and individuals within pairs. The interaction terms, $(P{*}R)_{ik}$ and $(P{:}I{*}R)_{i(j)k}$, account for the probe dependent variation between and within pairs respectively. Using this model, the contribution of environment for each gene can be calculated as mean squares due to $I{:}P_{i(j)}$ minus the mean squares due to $(P{:}I{*}R)_{i(j)k}$ divided by the number of probesets used to measure gene activity. This value was recorded as a percent of total variation.

Genes were ranked by the contribution of genetic and environmental percentage of explained variation. The top 100 genes due to each source were then examined for enriched gene themes using Expression Analysis Systematic Explorer (EASE v 1.21; Hosack et al., 2003). Themes tested for included: chromosome, SwissProt keyword, PIR keyword, GenMAPP pathway, KEGG pathway, PFAM domain, SMART domain, Gene Ontology Consortium Biological Process, Molecular Function, and Cellular Component. Results were filtered to remove themes with less than 4 probesets or an EASE score less than 5%.

Results

The scatter plots in Figure 1 show that MZ twins are more similar in gene expression profiles than DZ twins for a substantial proportion of the 6578 genes sampled. The genes falling above the solid line in figure 1A and 1B are those expression scores that are more similar within twin pairs than between twin pairs, which suggest a correlation in gene expression. This is in contrast to the genes that fall at or below the solid line. Here, the within- compared to between- twin pair variance has equal or greater variance which suggests uncorrelated gene expression levels.

Figure 2A displays the results of 6578 intraclass correlations for gene expression levels for MZ and DZ twin pairs. The average MZ intraclass correlation is significantly more than twice the DZ mean which supports an epistatic model for gene expression regulation ($p < .01$). MZ twins had greater intraclass correlations than DZ twins (where the MZ correla-

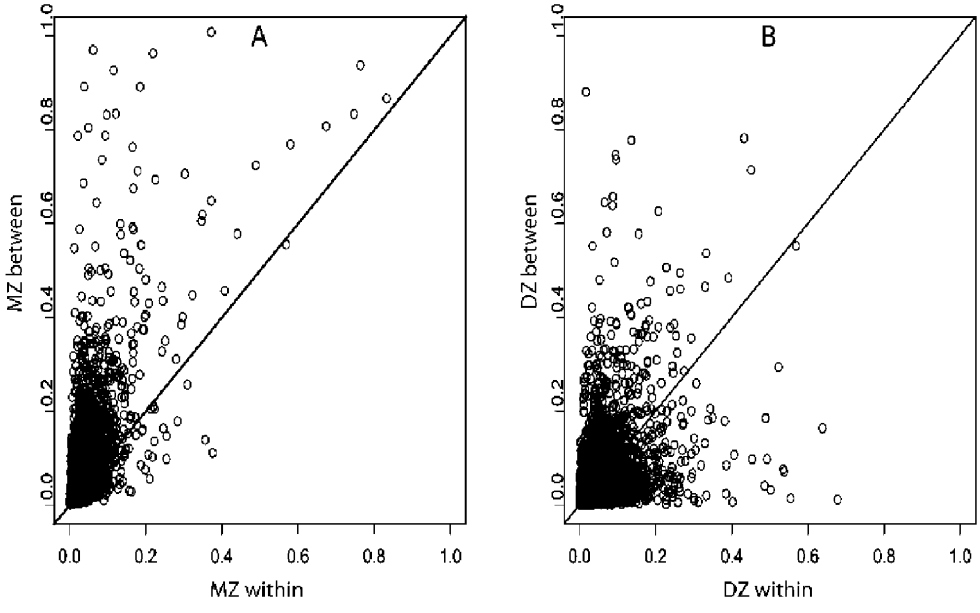

Figure 1

Scatter plots A. and B. are, respectively, the MZ and DZ twin within by between components of variance (mean squares) for 6578 genes (probesets). The solid line represents a between by within ratio of 1. Approximately 78% genes fall above the solid line in A, compared to 44% of genes in B, suggesting that MZ twins are more highly correlated for a greater proportion of gene expression levels than DZ twins.

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

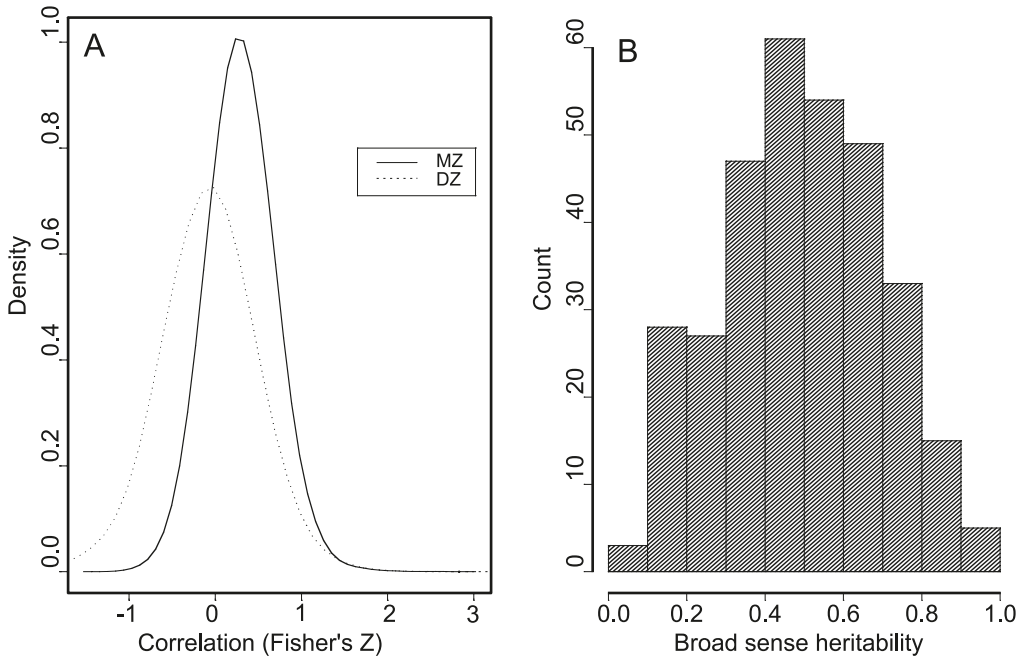

Figure 2.

A. Distribution of 6578 intraclass gene correlations for MZ and DZ twins. MZ $M = 0.30$, $SD = 0.38$ and DZ $M = -0.08$, $SD = 0.54$. B. Estimates of broad sense heritability for 322 genes (Appendix Table 2).

tion was in the positive range) in approximately 63% of genes measured. The estimates of broad sense heritability (which includes additive, dominance and epistatic components) are shown for 322 genes whose MZ to DZ intraclass correlation ratio was between 1 and 4 (Figure 2B and Appendix Table 2).

The apparent global genetic effects seen in Figure 1 and Figure 2 can be more formally tested using analysis of variance (ANOVA) techniques. ANOVA models were applied to derive the estimated components of variance for each zygosity separately using the terms listed in Table 1. As expected the largest component depends on average differences in expression between genes across all subjects (the 'genes' item). This term may also contain confounding effects of the probe sets which are known to operate at different levels of efficiency. The MZ to DZ ratio of the Gene × Between pair term indicates that MZ twins are 11.2 times more correlated than DZ pairs in their idiosyncratic profiles of gene expression. The highly significant difference between the Genes × Within-pairs mean squares for DZ versus MZ twins ($F_{32885, 65770}$, $p < 10^{-6}$) confirms the overall significance of a genetic contribution to intrinsic differences in gene expression profiles. The finding that the overall correlation in gene expression profiles of MZ twins, on average, are substantially more than twice that in DZ twins, albeit in this small sample, reinforces a *prima facie* case for the role of epistatic gene action in the regulation of gene expression.

Genes showing primarily genetic or environmental influences were examined to determine whether they differed with respect to their Gene Ontology (GO) Consortium categories (Ashburner et al., 2000) or other characteristics. The top 100 genes were ranked according to their proportion of genetic or environmental contributions to the total observed variance (Appendix Tables 2 and 3). Environmental variation was measured using the probe level data to estimate the contribution of error present within an individual of a MZ pair. Genetic contribution was determined by ranking the broad sense heritability estimates. The EASE program was used to identify significant gene themes in these two groups compared to the entire population of 6578 genes (Table 2). The categories of genes identified as being most heavily influenced by environmental factors were consistent with their biological functions in cells and included genes related to

Table 1

Variance Components Derived for Both Zygosities Separately Using ANOVA

Component	MZ	DZ
Genes (main effect)	2.8765	2.9113
Between pairs	0.0016	–0.0014
Within pairs	0.0030	0.0049
Genes × between pairs	0.0146	0.0013
Genes × within pairs (includes error)	0.0266	0.0403
Total	2.9223	2.9564

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

Table 2

Enriched Gene Categories Identified for the Top 100 Gene Expression Levels Explained by Environmental or Genetic Factors

System	Gene category	List% (*n*)	Chip% (*n*)	*p*-value
Environmental mediation				
GO Biological process	Response to biotic stimulus	24.7 (20)	6.6 (263)	5.90E-07
GO Biological process	Immune response	22.2 (18)	5.5 (221)	1.08E-06
GenMAPP pathway	Hs glycolysis and gluconeogenesis	53.8 (7)	4.5 (24)	3.55E-06
SwissProt keyword	Glycolysis	10.1 (7)	0.8 (23)	1.29E-05
KEGG pathway	Glycolysis/gluconeogenesis — Hs	36.8 (7)	3.5 (27)	1.51E-05
GO biological process	Organismal physiological process	23.5 (19)	7.5 (300)	1.77E-05
GO biological process	Glycolysis (includes glucose, monosaccharide, hexose, carbohydrate and alcohol catabolism/metabolism)	8.6 (7)	0.7 (29)	1.79E-05
PIR keyword	Gluconeogenesis	10.9 (5)	0.7 (8)	1.23E-04
GO biological process	Development	24.7 (20)	10.2 (407)	3.15E-04
GO biological process	Response to pest, pathogen or parasite	12.3 (10)	3.0 (118)	5.13E-04
GO molecular function	Heat shock protein activity	6.2 (5)	0.5 (21)	6.68E-04
PIR keyword	Glycolysis	10.9 (5)	1.0 (12)	7.78E-04
GO biological process	Organogenesis	13.6 (11)	4.3 (170)	1.90E-03
SwissProt keyword	Cytokine	7.2 (5)	0.8 (23)	2.09E-03
SwissProt keyword	Heat shock	5.8 (4)	0.4 (12)	2.67E-03
GO Molecular function	Cytokine activity	6.2 (5)	0.8 (31)	3.02E-03
GO biological process	Humoral immune response	7.4 (6)	1.4 (54)	4.23E-03
GO biological process	Neurogenesis	8.6 (7)	2.2 (86)	7.13E-03
GO biological process	Chemotaxis	4.9 (4)	0.7 (26)	1.46E-02
GO molecular function	Isomerase activity	6.2 (5)	1.2 (49)	1.55E-02
GO cellular component	Extracellular space	6.2 (5)	1.4 (54)	2.36E-02
GO biological process	Cell communication	27.2 (22)	17.0 (678)	2.36E-02
SwissProt keyword	Polymorphism	26.1 (18)	16.3 (450)	4.07E-02
GO biological process	Cell adhesion	7.4 (6)	2.4 (95)	4.10E-02
GO biological process	Response to wounding	6.2 (5)	1.7 (68)	4.67E-02
GO biological process	Response to stress	13.6 (11)	7.0 (277)	4.88E-02
Genetic mediation				
GO biological process	Antigen processing & presentation	4.6 (4)	0.4 (17)	6.28E-03
GO cellular component	Plasma membrane	20.5 (18)	10.5 (413)	8.19E-03
GO cellular component	Integral to plasma membrane	14.8 (13)	6.6 (259)	1.11E-02
SwissProt keyword	MHC II	4.5 (3)	0.2 (8)	1.43E-02
GO biological process	Antigen processing, exogenous antigen via MHC class II	3.4 (3)	0.2 (9)	1.50E-02
GO biological process	Antigen presentation, exogenous antigen	3.4 (3)	0.2 (9)	1.50E-02
GO molecular function	Receptor activity	13.6 (12)	6.2 (248)	1.67E-02
GO molecular function	Signal transducer activity	19.3 (17)	10.7 (429)	1.96E-02
SwissProt keyword	Signal	19.4 (13)	9.6 (264)	2.01E-02
GO cellular component	Membrane	42.0 (37)	31.5 (1232)	3.08E-02
Chromosome	Homo sapiens 7p	5.4 (5)	1.5 (68)	4.78E-02

Note: Genes in some categories overlap, as GO organizes gene groups in a pseudo-hierarchical format where each node corresponds to a distinct gene category. Child nodes are selected before parent nodes. If significant, parent nodes were included if 10% of genes differed from child node. All categories are listed for gene sets if they appeared in multiple systems. *List%* = percentage of 'environmental' or 'genetic' gene lists in gene category out of total list in system, *Chip%* = percentage of filtered set of genes (*n* = 6578) in gene category present in system, *n* = number of genes (probesets) in gene category.

defense and immune response; heat shock protein activity; cytokine activity; response to pest, pathogen, or parasite; and those functioning in carbohydrate and alcohol processing, glycolysis, and gluconeogenesis pathways. Categories influenced largely by genetic factors included signal transducer activity, integral to the plasma membrane, and MHC class II genes. The included MHC class II genes, HLA-DPA1, HLA-DPB1 and HLA-DRB1, are known to be highly variable at the sequence level, which raises the ques-

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

tion of whether this specificity also translates to the level of gene expression resulting in the observed increased MZ twin correlation.

Discussion

We provide evidence that epistasis is a ubiquitous component in the regulation of gene expression which conforms to the idea that epistasis is a natural property of transcriptional control (Gibson, 1996). If these results can be replicated in larger samples this observation is not without implications for the study of genetic influences on quantitative traits, since most twin studies show that, at a more gross phenotypic level, genetic effects appear to be largely additive. Thus, what is seen at the level of aggregate genetic influences may obscure more specific patterns of interaction at the level of gene expression. The study of profiles of gene expression offers a new approach that may enhance our ability to elucidate complex pathways through which naturally occurring genetic polymorphisms and environmental exposures interact more specifically to create differences in risk to common disease.

Mapping genetic determinants that contribute to gene expression variation will be important for understanding the genetic basis of human disease. Although relating variation in common disease to DNA sequence variation has been difficult at best, this may be relatively less problematic at the more proximal level of transciptional control and could serve as a model for more complex processes. Further quantification of the genetic and environmental sources of gene expression variation in normal and disease tissues will be necessary to guide the genetic analysis of disease etiology.

This study integrates the classical genetic approach of evaluating traits in unselected MZ and DZ twins with expression microarray technology. Future studies may selectively sample twin pairs who are discordant for exposure to a risk factor. Different patterns of gene expression in exposed and nonexposed twins could allow for the identification of those aspects of gene expression that mediate sensitivity to specific environmental risk factors.

Acknowledgments

We thank J. L. Ware for helpful discussions and comments. This work was supported by grants from the National Cancer Institute (5R25CA93423, for T. P. Y.) and the National Institute on Alcohol Abuse and Alcoholism (AA13678, M.F.M.).

References

Affymetrix (1999). *Affymetrix Microarray Suite User Guide*. Santa Clara, CA: Affymetrix.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25*, 25–29.

Cheung, V. G., Conlin, V. G., Weber, T. M., Arcaro, M., Jen, K. Y., Moreley, M., & Speilman, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics, 33*, 422–425.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., & Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinfomatics. *Genome Biology, 5*, R80.

Gibson, G. (1996). Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theoretical Population Biology, 49*, 58–89.

Hassan, S., Duong, B., Kim, K. S., & Miles, M. F. (2003). Pharmacogenomic analysis of mechanisms mediating ethanol regulation of dopamine beta-hydroxylase. *The Journal of Biological Chemistry,* 278, 38860–38869.

Hosack, D. A., Dennis, G., Jr., Sherman, B. T., Lane, H. C., & Lempicki, R. A. (2003). Identifying Biological Themes within Lists of Genes with EASE. *Genome Biology, 4*, R70.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics, 4*, 249–264.

Jinks, J.L., & Fulker, D.W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behaviour. *Psychological Bulletin, 73*, 311–349.

Lampe, J. W., Stepaniants, S. B., Mao, M., Radich, J. P., Dai, H., Linsley, P. S., Friend, S. H., & Potter, J. D. (2004). Signatures of environmental exposures using peripheral leukocyte gene expression: Tobacco smoke. *Cancer Epidemiology, Biomarkers & Prevention*, *13*, 445–453.

Morely, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature, 430*, 743–747.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for Genetic Studies of Twins and Families*. Dordrecht, the Netherlands: Kluwer Academic.

Oleksiak, M. F., Churchill, G. A., & Crawford, D. L. (2002). Variation in gene expression within and among natural populations. *Nature Genetics, 32*, 261–266.

Radich, J. P., Mao, M., Stepaniants, S., Biery, M., Castle, J., Ward, T., Schimmack, G., Kobayashi, S., Carleton, M., Lampe, J., & Linsley, P. S. (2004). Individual-specific variation of gene expression in peripheral blood leukocytes. *Genomics, 83*, 980–988.

Thibault, C., Lai, C., Wilke, N., Duong, B., Olive, M. F., Rahman, S., Dong, H., Hodge, C. W., Lockart, D. J., & Miles, M. F. (2000). Expression profiling of neural cells reveals specific patterns of ethanol-responsive gene expression. *Molecular Pharmacology, 58*, 1593–1600.

Whitney, A. R., Diehn, M., Popper, S. J., Alizadeh, A. A., Boldrick, J. C., Relman, D. A., & Brown, P. O. (2003). Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences USA, 100*, 1896–1901.

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science, 297*, 1143.

Appendix

Table 1

Characteristics of Individual Samples Obtained From Coriell Cell Repositories (Camden, NJ) Used in Microarray Experiments

Repository number	Pair	Zygosity	Age	BMI	Race
GM14405	1	MZ	41	22	White
GM14406	1	MZ	41	26	White
GM14437	2	DZ	48	29	White
GM14438	2	DZ	48	30	White
GM14439	3	MZ	45	18	Black
GM14440	3	MZ	45	19	Black
GM14464	4	MZ	40	–	Black
GM14465	4	MZ	40	26	Black
GM14474	5	MZ	20	23	Black
GM14475	5	MZ	20	22	Black
GM14476	6	MZ	30	26	Black
GM14477	6	MZ	30	26	Black
GM14495	7	MZ	46	30	Black
GM14496	7	MZ	46	28	Black
GM14503	8	MZ	20	24	Black
GM14504	8	MZ	20	24	Black
GM14511	9	DZ	18	21	Black
GM14512	9	DZ	18	20	Black
GM14520	10	MZ	22	32	White
GM14521	10	MZ	22	30	White
GM14529	11	DZ	18	23	Black
GM14530	11	DZ	18	22	Black
GM14532	12	MZ	23	27	Black
GM14533	12	MZ	23	25	Black
GM14535	13	MZ	26	18	Black
GM14536	13	MZ	26	19	Black
GM14548	14	DZ	48	22	White
GM14549	14	DZ	48	26	White
GM14632	15	DZ	21	21	Black
GM14633	15	DZ	21	21	Black

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

Table 2

Top 100 Gene Expression Levels Explained by Genetic Factors Ordered by Broad Sense Heritability Estimates

Affymetrix probeset ID	Gene symbol	Broad sense heritability	Gene title
205048_s_at	PSPHL	0.967010798	phosphoserine phosphatase-like
211990_at	HLA-DPA1	0.958281642	major histocompatibility complex, class II, DP alpha 1
217225_x_at	PM5	0.925262662	pM5 protein
201533_at	CTNNB1	0.910001878	catenin (cadherin-associated protein), beta 1, 88kDa
201297_s_at	C2orf6	0.909119219	chromosome 2 open reading frame 6
205859_at	LY86	0.897232798	lymphocyte antigen 86
208727_s_at	CDC42	0.884619025	cell division cycle 42 (GTP binding protein, 25kDa)
221536_s_at	FLJ11301	0.874073612	hypothetical protein FLJ11301
217972_at	CHCHD3	0.873312532	coiled-coil-helix-coiled-coil-helix domain containing 3
214894_x_at	MACF1	0.856527663	microtubule-actin crosslinking factor 1
215193_x_at	HLA-DRB1	0.840831221	major histocompatibility complex, class II, DR beta 1
201137_s_at	HLA-DPB1	0.831226939	major histocompatibility complex, class II, DP beta 1
217317_s_at	MN7	0.828276734	D15F37 (pseudogene)
200715_x_at	RPL13A	0.827501355	ribosomal protein L13a
211009_s_at	ZNF271	0.821063501	zinc finger protein 271
212063_at	CD44	0.818270192	CD44 antigen (homing function and Indian blood group system)
201920_at	SLC20A1	0.818160338	solute carrier family 20 (phosphate transporter), member 1
217777_s_at	HSPC121	0.816762604	butyrate-induced transcript 1
201201_at	CSTB	0.809766203	cystatin B (stefin B)
219874_at	SLC12A8	0.800083483	solute carrier family 12 (potassium/chloride transporters), member 8
200025_s_at	RPL27	0.799604033	ribosomal protein L27
217523_at	CD44	0.79401146	CD44 antigen (homing function and Indian blood group system)
209433_s_at	PPAT	0.788330471	phosphoribosyl pyrophosphate amidotransferase
216942_s_at	CD58	0.78693124	CD58 antigen, (lymphocyte function-associated antigen 3)
211077_s_at	TLK1	0.780708897	tousled-like kinase 1
220177_s_at	TMPRSS3	0.776727045	transmembrane protease, serine 3
203139_at	DAPK1	0.769516064	death-associated protein kinase 1
212483_at	IDN3	0.76782345	IDN3 protein
216591_s_at	SDHC	0.764084596	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa
203819_s_at	IMP-3	0.760648769	IGF-II mRNA-binding protein 3
201900_s_at	AKR1A1	0.754658289	aldo-keto reductase family 1, member A1 (aldehyde reductase)
218386_x_at	USP16	0.75448844	ubiquitin specific protease 16
201807_at	VPS26	0.751888156	vacuolar protein sorting 26 (yeast)
215333_x_at	GSTM1	0.747823305	glutathione S-transferase M1
209233_at	C2F	0.746148618	C2f protein
201669_s_at	MARCKS	0.737309858	myristoylated alanine-rich protein kinase C substrate
203470_s_at	PLEK	0.737231849	pleckstrin
203787_at	SSBP2	0.732987323	single-stranded DNA binding protein 2
204142_at	HSRTSBETA	0.727635761	rTS beta protein
206245_s_at	IVNS1ABP	0.726917756	influenza virus NS1A binding protein
210912_x_at	GSTM4	0.72559181	glutathione S-transferase M4
218090_s_at	WDR11	0.721311625	WD repeat domain 11
202869_at	OAS1	0.719699282	2′,5′-oligoadenylate synthetase 1, 40/46kDa
201516_at	SRM	0.717782614	spermidine synthase
209036_s_at	MDH2	0.716564086	malate dehydrogenase 2, NAD (mitochondrial)
200904_at	HLA-E	0.713181962	major histocompatibility complex, class I, E
214501_s_at	H2AFY	0.710967245	H2A histone family, member Y
202203_s_at	AMFR	0.709459621	autocrine motility factor receptor
204362_at	SCAP2	0.706961663	src family associated phosphoprotein 2
208908_s_at	CAST	0.705909876	calpastatin

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

Table 2 (continued)

Top 100 Gene Expression Levels Explained by Genetic Factors Ordered by Broad Sense Heritability Estimates

Affymetrix probeset ID	Gene symbol	Broad sense heritability	Gene title
212692_s_at	LRBA	0.704286755	LPS-responsive vesicle trafficking, beach and anchor containing
202797_at	SACM1L	0.702451044	SAC1 suppressor of actin mutations 1-like (yeast)
209682_at	CBLB	0.700275446	Cas-Br-M (murine) ecotropic retroviral transforming sequence b
201676_x_at	PSMA1	0.694887117	proteasome (prosome, macropain) subunit, alpha type, 1
218460_at	FLJ20397	0.693353047	hypothetical protein FLJ20397
219806_s_at	FN5	0.689276381	FN5 protein
213239_at	PIBF1	0.688322436	progesterone-induced blocking factor 1
208905_at	CYCS	0.68015715	cytochrome c, somatic
204088_at	P2RX4	0.679961082	purinergic receptor P2X, ligand-gated ion channel, 4
212308_at	CLASP2	0.679777809	cytoplasmic linker associated protein 2
214097_at	RPS21	0.673936281	ribosomal protein S21
207467_x_at	CAST	0.670215002	calpastatin
208986_at	TCF12	0.666841383	transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)
213320_at	PRMT3	0.666507556	protein arginine N-methyltransferase 3(hnRNP methyltransferase S. cerevisiae)-like 3
203992_s_at	UTX	0.661897438	ubiquitously transcribed tetratricopeptide repeat gene, X chromosome
202534_x_at	DHFR	0.661266741	dihydrofolate reductase
212534_at	ZNF24	0.654640425	zinc finger protein 24 (KOX 17)
201097_s_at	ARF4	0.649977923	ADP-ribosylation factor 4
206653_at	RPC32	0.647245202	polymerase (RNA) III (DNA directed) (32kD)
212694_s_at	PCCB	0.644030039	propionyl Coenzyme A carboxylase, beta polypeptide
209704_at	M96	0.640450062	likely ortholog of mouse metal response element binding transcription factor 2
201444_s_at	ATP6AP2	0.639797251	ATPase, H+ transporting, lysosomal accessory protein 2
56256_at	CGI-40	0.63865428	CGI-40 protein
212331_at	RBL2	0.638380828	retinoblastoma-like 2 (p130)
202804_at	ABCC1	0.637178492	ATP-binding cassette, sub-family C (CFTR/MRP), member 1
214452_at	BCAT1	0.634751582	branched chain aminotransferase 1, cytosolic
203304_at	BAMBI	0.634003317	BMP and activin membrane-bound inhibitor homolog (Xenopus laevis)
202211_at	ARFGAP3	0.63253629	ADP-ribosylation factor GTPase activating protein 3
203380_x_at	SFRS5	0.632492993	splicing factor, arginine/serine-rich 5
203156_at	AKAP11	0.632059988	A kinase (PRKA) anchor protein 11
215380_s_at	C7orf24	0.630936938	chromosome 7 open reading frame 24
202272_s_at	KIAA0483	0.630923695	KIAA0483 protein
202069_s_at	IDH3A	0.630596904	isocitrate dehydrogenase 3 (NAD+) alpha
221499_s_at	STX16	0.62872761	syntaxin 16
208798_x_at	GOLGIN-67	0.623047766	golgin-67
212352_s_at	TMP21	0.620022335	transmembrane trafficking protein
214042_s_at	RPL22	0.61772918	ribosomal protein L22
208968_s_at	LOC57019	0.617063889	hypothetical protein LOC57019
203286_at	RNF44	0.616125564	ring finger protein 44
218150_at	ARL5	0.615516839	ADP-ribosylation factor-like 5
218310_at	RABGEF1	0.614414977	RAB guanine nucleotide exchange factor (GEF) 1
210479_s_at	RORA	0.610021021	RAR-related orphan receptor A
203306_s_at	SLC35A1	0.608230235	solute carrier family 35 (CMP-sialic acid transporter), member A1
1007_s_at	DDR1	0.607352748	discoidin domain receptor family, member 1
207469_s_at	PIR	0.606894302	Pirin
212560_at	SORL1	0.60633017	sortilin-related receptor, L(DLR class) A repeats-containing
200891_s_at	SSR1	0.605792645	signal sequence receptor, alpha (translocon-associated protein alpha)
209251_x_at	TUBA6	0.602902357	tubulin alpha 6
200086_s_at	COX4I1	0.602329072	cytochrome c oxidase subunit IV isoform 1
220175_s_at	CBWD2	0.601711632	COBW domain containing 2 COBW-like placental protein COBW-like protein dopamine responsive protein

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

Table 3

Top 100 Gene Expression Ordered by Percent of Variance Explained by Environmental Factors

Affymetrix probeset ID	Gene symbol	Percent environmental variation	Gene title
214677_x_at	IGLJ3	50.78294	immunoglobulin lambda joining 3
209374_s_at	IGHM	37.85213	immunoglobulin heavy constant mu
201841_s_at	HSPB1	31.98722	heat shock 27kDa protein 1
202581_at	HSPA1A	31.1781	heat shock 70kDa protein 1A
215118_s_at	MGC27165	29.46612	hypothetical protein MGC27165
208308_s_at	GPI	28.12264	glucose phosphate isomerase
213872_at	C6orf62	26.75021	chromosome 6 open reading frame 62
212599_at	AUTS2	24.27974	autism susceptibility candidate 2
200664_s_at	DNAJB1	24.16818	DnaJ (Hsp40) homolog, subfamily B, member 1
200666_s_at	DNAJB1	24.10276	DnaJ (Hsp40) homolog, subfamily B, member 1
200968_s_at	PPIB	22.71453	peptidylprolyl isomerase B (cyclophilin B)
200650_s_at	LDHA	22.40607	lactate dehydrogenase A
200866_s_at	PSAP	21.59879	prosaposin (variant Gaucher disease and variant metachromatic leukodystrophy)
201426_s_at	VIM	20.92698	vimentin
200800_s_at	HSPA1A	20.83912	heat shock 70kDa protein 1A
202551_s_at	CRIM1	20.68993	cysteine-rich motor neuron 1
209457_at	DUSP5	20.56571	dual specificity phosphatase 5
200799_at	HSPA1A	20.43088	heat shock 70kDa protein 1A
200869_at	RPL18A	19.62439	ribosomal protein L18a
200055_at	TAF10	19.02615	TAF10 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 30kDa
209318_x_at	PLAGL1	17.47877	pleiomorphic adenoma gene-like 1
201243_s_at	ATP1B1	17.39659	ATPase, Na+/K+ transporting, beta 1 polypeptide
221523_s_at	RRAGD	16.52905	Ras-related GTP binding D
203752_s_at	JUND	16.22446	jun D proto-oncogene
204331_s_at	MRPS12	15.98433	mitochondrial ribosomal protein S12
208152_s_at	DDX21	15.72878	DEAD (Asp-Glu-Ala-Asp) box polypeptide 21
204961_s_at	NCF1	15.64712	neutrophil cytosolic factor 1 (47kDa, chronic granulomatous disease, autosomal 1)
208730_x_at	MYH9	15.46814	myosin, heavy polypeptide 9, non-muscle
206975_at	LTA	15.03013	lymphotoxin alpha (TNF superfamily, member 1)
212587_s_at	PTPRC	15.00539	protein tyrosine phosphatase, receptor type, C
204674_at	LRMP	14.88316	lymphoid-restricted membrane protein
208934_s_at	LGALS8	14.60094	lectin, galactoside-binding, soluble, 8 (galectin 8)
207861_at	CCL22	14.07008	chemokine (C-C motif) ligand 22
213655_at	YWHAE	13.73642	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide
209995_s_at	TCL1A	13.39615	T-cell leukemia/lymphoma 1A
210356_x_at	MS4A1	13.06091	membrane-spanning 4-domains, subfamily A, member 1
207419_s_at	RAC2	12.889	ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding protein Rac2)
200737_at	PGK1	12.83654	phosphoglycerate kinase 1
206641_at	TNFRSF17	12.7961	tumor necrosis factor receptor superfamily, member 17
204480_s_at	C9orf16	12.75505	chromosome 9 open reading frame 16
209417_s_at	IFI35	12.74506	interferon-induced protein 35
217967_s_at	C1orf24	12.54847	chromosome 1 open reading frame 24
205081_at	CRIP1	12.51425	cysteine-rich protein 1 (intestinal)
204103_at	CCL4	11.96613	chemokine (C-C motif) ligand 4
204082_at	PBX3	11.88659	pre-B-cell leukemia transcription factor 3
202022_at	ALDOC	11.83344	aldolase C, fructose-bisphosphate
201242_s_at	ATP1B1	11.51903	ATPase, Na+/K+ transporting, beta 1 polypeptide

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press

Table 3 (continued)

Top 100 Gene Expression Ordered by Percent of Variance Explained by Environmental Factors

Affymetrix probeset ID	Gene symbol	Percent environmental variation	Gene title
212737_at	GM2A	11.49095	GM2 ganglioside activator protein
44783_s_at	HEY1	11.4397	hairy/enhancer-of-split related with YRPW motif 1
221488_s_at	C6orf82	11.43703	chromosome 6 open reading frame 82
221731_x_at	CSPG2	11.42384	chondroitin sulfate proteoglycan 2 (versican)
217294_s_at	ENO1	11.3838	enolase 1, (alpha)
213011_s_at	TPI1	11.28868	triosephosphate isomerase 1
211969_at	HSPCA	11.28417	heat shock 90kDa protein 1, alpha
210448_s_at	P2RX5	11.24721	purinergic receptor P2X, ligand-gated ion channel, 5
39729_at	PRDX2	11.18626	peroxiredoxin 2
220939_s_at	DPP8	11.10536	dipeptidylpeptidase 8
219841_at	AICDA	10.88388	activation-induced cytidine deaminase
219221_at	FLJ35036	10.83614	hypothetical protein FLJ35036
221253_s_at	TXNDC5	10.80888	thioredoxin domain containing 5
201811_x_at	SH3BP5	10.64359	SH3-domain binding protein 5 (BTK-associated)
205483_s_at	G1P2	10.5186	interferon, alpha-inducible protein (clone IFI-15K)
209138_x_at	IGLC2	10.26688	immunoglobulin lambda constant 2 (Kern-Oz- marker)
213294_at	FLJ38348	10.2401	hypothetical protein FLJ38348
203239_s_at	CNOT3	10.23926	CCR4-NOT transcription complex, subunit 3
218589_at	P2RY5	10.07417	purinergic receptor P2Y, G-protein coupled, 5
206255_at	BLK	9.981363	B lymphoid tyrosine kinase
203758_at	CTSO	9.961683	cathepsin O
212977_at	CMKOR1	9.924967	chemokine orphan receptor 1
203881_s_at	DMD	9.916813	dystrophin (muscular dystrophy, Duchenne and Becker types)
221607_x_at	ACTG1	9.770038	actin, gamma 1
200778_s_at	NEDD5	9.713052	neural precursor cell expressed, developmentally down-regulated 5
217826_s_at	UBE2J1	9.701461	ubiquitin-conjugating enzyme E2, J1 (UBC6 homolog, yeast)
201169_s_at	BHLHB2	9.605374	basic helix-loop-helix domain containing, class B, 2
210105_s_at	FYN	9.420399	FYN oncogene related to SRC, FGR, YES
200046_at	DAD1	9.393835	defender against cell death 1
205321_at	EIF2S3	9.324742	eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa
204849_at	TCFL5	9.267302	transcription factor-like 5 (basic helix-loop-helix)
221753_at	SSH1	9.267046	slingshot 1
217732_s_at	ITM2B	9.172218	integral membrane protein 2B
201170_s_at	BHLHB2	9.160878	basic helix-loop-helix domain containing, class B, 2
202856_s_at	SLC16A3	9.153007	solute carrier family 16 (monocarboxylic acid transporters), member 3
211645_x_at		9.076039	immunoglobulin kappa light chain variable region mRNA, partial cds
217759_at	TRIM44	9.025651	tripartite motif-containing 44
211919_s_at	CXCR4	8.993819	chemokine (C-X-C motif) receptor 4
216253_s_at	PARVB	8.978139	parvin, beta
200772_x_at	PTMA	8.962768	prothymosin, alpha (gene sequence 28)
219014_at	PLAC8	8.914034	placenta-specific 8
204439_at	C1orf29	8.860344	chromosome 1 open reading frame 29
213887_s_at	POLR2E	8.838963	polymerase (RNA) II (DNA directed) polypeptide E, 25kDa
203665_at	HMOX1	8.792702	heme oxygenase (decycling) 1
200967_at	PPIB	8.77222	peptidylprolyl isomerase B (cyclophilin B)
202233_s_at	UQCRH	8.761494	ubiquinol-cytochrome c reductase hinge protein
208656_s_at	CCNI	8.758128	cyclin I
217480_x_at	IGKV1OR15-118	8.717833	immunoglobulin kappa variable 1/OR15-118
200966_x_at	ALDOA	8.59418	aldolase A, fructose-bisphosphate
203562_at	FEZ1	8.578454	fasciculation and elongation protein zeta 1 (zygin I)
206693_at	IL7	8.577712	interleukin 7
221658_s_at	IL21R	8.570516	interleukin 21 receptor
201623_s_at	DARS	8.55971	aspartyl-tRNA synthetase

https://doi.org/10.1375/twin.8.1.5 Published online by Cambridge University Press