Human Molecular Genetics, 2007, Vol. 16, No. 4 ***364–373***
doi:10.1093/hmg/ddl456
Advance Access published on December 12, 2006

Replicated effects of sex and genotype on gene expression in human lymphoblastoid cell lines

Allan F. McRae[1,8,*,†], Nicholas A. Matigian[2,3,†], Lata Vadlamudi[4,5], John C. Mulley[6,7], Bryan Mowry[3], Nicholas G. Martin[1], Sam F. Berkovic[4], Nicholas K. Hayward[2] and Peter M. Visscher[1,8]

[1]Genetic Epidemiology Group and [2]Human Genetics Laboratory, Queensland Institute of Medical Research, Herston, QLD 4029, Australia, [3]Queensland Centre for Mental Health Research, The Park Centre for Mental Health, Wacol, QLD 4076, Australia, [4]Epilepsy Research Centre, Department of Medicine (Neurology), University of Melbourne, Austin Health, Heidelberg, Victoria, Australia, [5]Academic Unit of Medicine (Neurology), The Australian National University, Canberra, Australia, [6]Department of Genetic Medicine, Women's and Children's Hospital, North Adelaide, Australia, [7]Department of Molecular Biosciences, University of Adelaide, Adelaide, South Australia, Australia and [8]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK

Received July 4, 2006; Revised September 26, 2006; Accepted December 4, 2006

The expression level for 15 887 transcripts in lymphoblastoid cell lines from 19 monozygotic twin pairs (10 male, 9 female) were analysed for the effects of genotype and sex. On an average, the effect of twin pairs explained 31% of the variance in normalized gene expression levels, consistent with previous broad sense heritability estimates. The effect of sex on gene expression levels was most noticeable on the X chromosome, which contained 15 of the 20 significantly differentially expressed genes. A high concordance was observed between the sex difference test statistics and surveys of genes escaping X chromosome inactivation. Notably, several autosomal genes showed significant differences in gene expression between the sexes despite much of the cellular environment differences being effectively removed in the cell lines. A publicly available gene expression data set from the CEPH families was used to validate the results. The heritability of gene expression levels as estimated from the two data sets showed a highly significant positive correlation, particularly when both estimates were close to one and thus had the smallest standard error. There was a large concordance between the genes significantly differentially expressed between the sexes in the two data sets. Analysis of the variability of probe binding intensities within a probe set indicated that results are robust to the possible presence of polymorphisms in the target sequences.

INTRODUCTION

The use of whole-genome gene expression studies has received considerable attention in recent years because of their potential to provide a greater understanding of the biology of complex diseases (1–3). Initial investigations have provided an understanding of the natural variation in human gene expression levels and have demonstrated that a significant proportion of this is heritable (4–6). However, there have been no comparisons made regarding the concordance of the heritability estimates across studies.

The best methodology for the analysis of gene expression data has not been determined, with many new methods, or variants of old methods, being proposed on a regular basis (7,8). The choice of analysis methodology is further complicated by the differences in results obtained with different methods (9). Comparisons between different methods of data analysis are hampered by the fact that the majority of gene expression data sets are generated without prior knowledge of the underlying determinants of expression levels. Thus, the comparison of positive results obtained by different methods is not possible. Some progress has been made in this area through the

*To whom correspondence should be addressed. Tel: +61 733620190; Fax: +61 733620101; Email: allan.mcrae@qimr.edu.au
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

used of controlled experiments where the expression levels of various transcripts are adjusted manually (9,10). However, such experiments do not capture the biological interactions that occur between gene expression levels and thus conclusions based on such studies need to be treated with caution.

A novel alternative is to use a well-studied biological phenomenon as a baseline for comparisons. Ideally, a candidate would be reasonably ubiquitous so that gene expression data generated for other purposes could be used in model testing. One such biological phenomenon is X chromosome inactivation. It is well known that one of the X chromosomes in mammalian females is silenced as a method of dosage compensation (11,12). However, a few genes escape X inactivation to varying extents (13,14). Genes that escape X inactivation are therefore expressed at higher levels in females than in males. The relative expression levels of females may be below twice that of males (which may be expected given twice the copies of the gene), as the escape from inactivation is not necessarily complete. The utility of gene expression microarrays in the examination of X chromosome inactivation has previously been demonstrated (15). Recently, a comprehensive analysis of expression of X-linked genes has been performed by examining the expression of allelic variants in fibroblast lines (16). Given the underlying nature of X inactivation, it appears reasonable that these results will extend to other cell lines and biological samples where these genes are expressed. Thus, this survey provides an ideal data set to assess the quality of results obtained from the analysis of microarray data.

In this study, the analysis of gene expression levels is performed using linear mixed models. Mixed models are a class of powerful and versatile models that simultaneously fit observed 'fixed' effects (such as sex and age) that cause mean differences between samples and unobserved 'random' (or latent) effects that cause correlation between samples. These models allow the estimation of systematic effects while simultaneously partitioning the remaining variation into sources of underlying causal effects. Mixed linear models have been used extensively in biology research (17) and are implemented in widely available statistical packages. Mixed model methodology is applicable to a wide variety of experimental designs. Correlations caused by genetic relatedness between samples are readily accounted for in the mixed model framework (18,19) and an option to perform such an analysis through the specification of a pedigree file is often included in software packages that implement mixed models. For all of these reasons, mixed models are appealing for use in the analysis of gene expression data. The flexibility of the mixed model framework for the analysis of gene expression data was recently demonstrated in an analysis that included terms to model across-species differences in probe binding efficiency (20).

Here, we estimate the effect of sex and genotype on gene expression in lymphoblastoid cell lines (LCLs) from a sample of monozygotic (MZ) twin pairs using linear mixed models. The results are compared with those obtained using a second publicly available data set of gene expression from the CEPH families that contained a subset of the genes analysed.

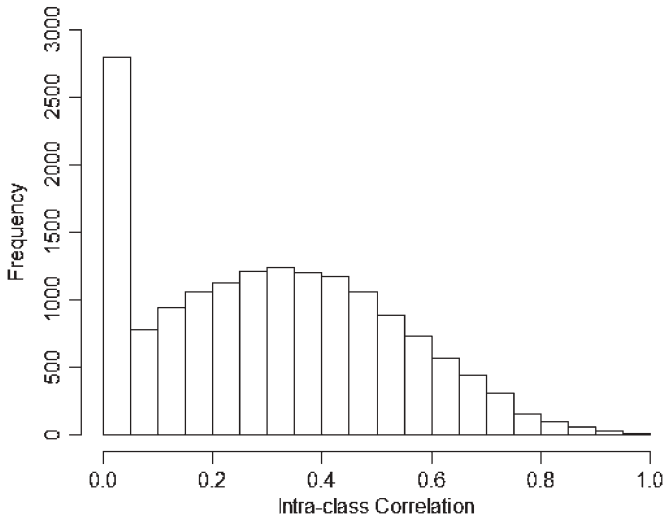

Figure 1. The distribution of estimated intra-class correlation of expression levels in MZ twin pairs.

RESULTS

Data pre-processing and normalization

Gene expression levels in LCLs from 19 MZ twin pairs were measured using Affymetrix Human Genome U133 plus 2.0 Gene Chips. Of the 56 675 transcripts whose expression levels were measured on the chip, 15 887 were determined to be expressed across 100% of samples. The normalization of the data across chips and genes demonstrated that the vast majority (97%) of the variance in the transformed expression levels was due to differences in average expression level across genes. As expected, the across chip variance was negligible due to scaling performed during the data pre-processing stages (see Materials and Methods).

Partitioning of the variance in expression levels

The variance in gene expression levels for a particular gene was partitioned using a linear mixed model. Out of the 15 887 genes analysed in the single-gene analysis, 2106 (13%) provided a zero estimate for the proportion of the variance in gene expression levels explained by pair (Fig. 1). Although this value is biased upwards due to the lack of power afforded by the sample size, it is significantly lower than the 50% expected under the null hypothesis of no effect of pair (21). On an average, the pair intra-class correlation is 0.31 or, stated alternatively, the twin pairings explained 31% of the variance in the normalized expression levels. In the absence of common environmental influences, this is a measure of the broad sense heritability of a gene's expression level. The intra-class correlation is measuring the proportion of variance explained by twin pairings and thus is bound between 0 and 1, unlike a standard correlation that has a range of -1 to 1. This constraint is standard in a mixed model framework and occurs commonly in biological applications such as heritability estimation and linkage mapping of quantitative trait loci, where the proportion of variance attributable to additive genetic effects has a lower bound of

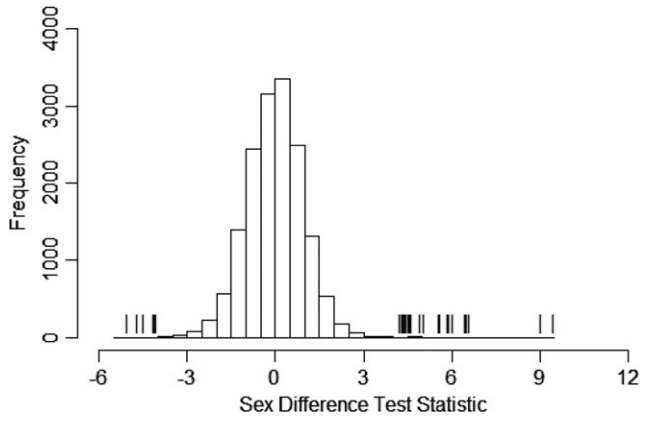

Figure 2. The distribution of the test statistic for the effect of sex (female–male) on gene expression. Vertical bars give the positions of test statistics with absolute value greater than four.

zero. From the distribution of non-zero intra-class correlations of MZ twin pairs, a reasonably symmetric decay around the mode of the distribution is apparent and indicates that the extent of this bias in the average proportion of variance explained by twin pairings caused by the imposed lower bound is limited.

Effect of sex on expression levels

The distribution of test statistics for the difference in average expression levels between male and female MZ twin pairs is given in Figure 2. These are calculated such that a positive test statistic represents an increased level of expression in females when compared with males. As expected when performing a large number of tests on genes that will, in the majority of cases, not be differentially expressed between the sexes, the central portions of the distribution closely match the expected normal distribution. The distribution is also consistent with the hypothesis that a large proportion of the differences are due to escape from X-inactivation, with the distribution of test statistics having a positive mean (0.06; $P < 0.05$) and skew (0.21; $P < 10^{-5}$).

Table 1 summarizes all probe sets with an absolute test statistic for differential expression between the sexes greater than four. Assuming normality of test statistics, this corresponds to a point significance of $\sim$0.00006. While this is not as stringent as a Bonferroni correction on the 15 887 tests being performed, many of these tests are non-independent due to correlations among expression levels across genes, thus using Bonferroni would result in an over-correction and subsequent loss of power to detect differentially expressed genes. In total, 31 gene-probes were detected to be differentially expressed between the sexes. These represented 20 distinct genes (and one unannotated target) with *DDX3X* represented four times, *UTX* represented three times and *EIF1AX*, *U2AF1L2*, *RBBP7*, *CD99* and *EIF2S3* represented twice. The genes in Table 1 were represented by a further 10 genes that were not detected to be significantly differently between the sexes. However, five of the 10 had test statistics of magnitude greater than three and the 10 genes had an average test statistic magnitude of 2.63, much greater than expected by random chance ($P < 10^{-6}$). Of the 20 distinct genes, 15 are located on the X chromosome, a markedly larger proportion than expected by chance. Of the five autosomal probe sets that detect differential expression between the sexes, two are for predicted genes (*C18orf1*, *ERICH1*) and the remaining three (*UTP15*, *METT5D1* and *FEZ1*) are not obvious candidates for genes causative of sexual differentiation.

Recently, a comprehensive survey of activation status of genes on the X chromosome was performed in fibroblast cell lines (16). Approximately, 70% of the genes on the X chromosome detected as being expressed in the LCLs used in this study were also detected in the fibroblast cell lines used by Carrel and Willard. Figure 3 compares Carrel and Willard's observed X inactivation status with the test statistic obtained from the analysis of MZ pairs. These show a sizeable concordance with the genes showing high test statistics also being detected by Carrel and Willard as escaping X inactivation frequently. This observation may not be surprising given that complete escape from X inactivation would result in the (approximate) doubling of expression levels in females when compared with males. However, the detection of these differences even with the sample size used in this study demonstrates the potential to uncover differences in gene expression levels between two or more groups.

Validation using CEPH family data

The publicly available expression data from CEPH families (5) provides replication of 4061 of the 15 887 (26%) genes analysed in the MZ twin pairs. Figure 4 plots the intra-class correlation of gene expression levels in MZ twin pairs against the heritability of gene expression estimated from the CEPH data. On an average, the twin intra-class correlation is greater than the CEPH heritability estimates. This is expected as the twin intra-class correlation includes variance from dominance and genetic interaction. It is also possible that common environmental influences are stronger between MZ twin pairs than individuals in a three generation pedigree, which would further inflate the MZ intra-class correlations. However, it is uncertain whether such influences remain in LCLs. Despite these differences, the MZ intra-class correlations and CEPH heritabilities show a highly significant correlation of 0.20 ($P < 10^{-15}$). The relationship is particularly strong when restricting the data set to MZ intra-class correlations greater than 0.8. This is a consequence of the reduced standard error of heritability estimates in this region.

The test statistics for differential expression between the sexes obtained from the MZ pairs and the CEPH families are compared in Figure 5. As expected under the assumption that the majority of genes are not differentially expressed between the sexes, the majority of test statistics are scattered in a circle around the central area. However, the majority of the large test statistics are concordant in both samples. Table 2 lists all probe sets that had an absolute test statistic greater than four in the CEPH families. All of the 14 genes detected as differentially expressed were for unique genes, only one of which was not located on the X chromosome. Of these 14 genes, 10 were previously detected as differentially expressed in the MZ twin sexes. The large amount of

Table 1. Summary of probes with an absolute test statistic for differential expression between the sexes greater than four

Systematic	Gene symbol	Chromosomal location	Test statistic	Gene description
203992_s_at	*UTX*	Xp11.2	9.49	Ubiquitously transcribed tetratricopeptide repeat, X chromosome
212515_s_at	*DDX3X*	Xp11.3-p11.23	9.05	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked
212514_x_at	*DDX3X*	Xp11.3-p11.23	6.64	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked
201018_at	*EIF1AX*	Xp22.12	6.51	Eukaryotic translation initiation factor 1A, X-linked
201211_s_at	*DDX3X*	Xp11.3-p11.23	6.49	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked
201210_at	*DDX3X*	Xp11.3-p11.23	6.48	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked
213876_x_at	*U2AF1L2*	Xp22.1	6.07	U2(RNU2) small nuclear RNA auxiliary factor 1-like 2
239106_at	*CA5BL*	Xp22.2	5.92	Carbonic anhydrase VB-like
228043_at	*UTP15*	5q13.2	5.90	UTP15, U3 small nucleolar ribonucleoprotein, homolog (yeast)
1557954_at	*RBBP7*	Xp22.2	5.63	Retinoblastoma binding protein 7
208174_x_at	*U2AF1L2*	Xp22.1	5.59	U2(RNU2) small nuclear RNA auxiliary factor 1-like 2
203990_s_at	*UTX*	Xp11.2	5.09	Ubiquitously transcribed tetratricopeptide repeat, X chromosome
201029_s_at	*CD99*	Xp22.32; Yp11.3	−5.02	CD99 antigen
202383_at	*SMCX*	Xp11.22-p11.21	4.96	Smcy homolog, X-linked (mouse)
209573_s_at	*C18orf1*	18p11.2	−4.69	Chromosome 18 open reading frame 1
203974_at	*HDHD1A*	Xp22.32	4.64	Haloacid dehalogenase-like hydrolase domain containing 1A
1554447_at	*LOC554203*	Xq13.2	4.63	Hypothetical LOC554203
203991_s_at	*UTX*	Xp11.2	4.62	Ubiquitously transcribed tetratricopeptide repeat, X chromosome
227520_at	*RBBP7*	Xp22.2	4.60	Retinoblastoma binding protein 7
216342_x_at	–	–	4.56	–
214678_x_at	*ZFX*	Xp21.3	4.55	Zinc finger protein, X-linked
201028_s_at	*CD99*	Xp22.32; Yp11.3	−4.48	CD99 antigen
204061_at	*PRKX*	Xp22.3	4.48	Protein kinase, X-linked
207551_s_at	*MSL3L1*	Xp22.3	4.44	Male-specific lethal 3-like 1 (Drosophila)
224935_at	*EIF2S3*	Xp22.2-p22.1	4.40	Eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa
201016_at	*EIF1AX*	Xp22.12	4.36	Eukaryotic translation initiation factor 1A, X-linked
224936_at	*EIF2S3*	Xp22.2-p22.1	4.36	Eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa
201589_at	*SMC1L1*	Xp11.22-p11.21	4.25	SMC1 structural maintenance of chromosomes 1-like 1 (yeast)
238773_at	*METT5D1, METT5D2*	11p14.1, 3q25.31	−4.11	Methyltransferase 5 domain containing 1, methyltransferase 5 domain containing 2
203562_at	*FEZ1*	11q24.2	−4.07	Fasciculation and elongation protein zeta 1 (zygin I)
1563315_s_at	*ERICH1*	8p23.3	−4.05	Glutamate-rich 1

replication between the two data sets demonstrates that although an exact significance threshold was not determined for these data sets, the chosen thresholds are suitably stringent. Further evidence for differential expression can be obtained by combining the test statistics from the MZ and CEPH analyses. A large positive value from the multiplication of the two test statistics implies concordance between the test statistics. A significance threshold can be obtained using the magnitude of the most negative combined test statistic, as the distribution of the combined test statistic is expected to be symmetrical under the null hypothesis of no effect of sex on gene expression. This approach relies on the lower end of the distribution of test statistics being accurately estimated from that data and should be satisfied with this data set containing almost 16 000 data points. As the most negative combined test statistics is −5.64, any combined test statistic greater than 5.64 can be considered significant. Six further genes are detected as differentially expressed using the combined analysis (Table 3). Three of these genes with the largest test statistics were located on the X chromosome and the remaining three are situated on autosomal chromosomes.

DISCUSSION

In this study, the effects of genotype and sex on gene expression levels in human LCLs have been examined. Initially, a novel data set of the gene expression levels from 19 MZ twin pairs were analysed. The results from this analysis were then validated using a publicly available data from CEPH families that contained expression information on a subset of the genes analysed in the twin pairs.

The analysis of the effect of twin pair on the effect of expression levels in this study has demonstrated a significant

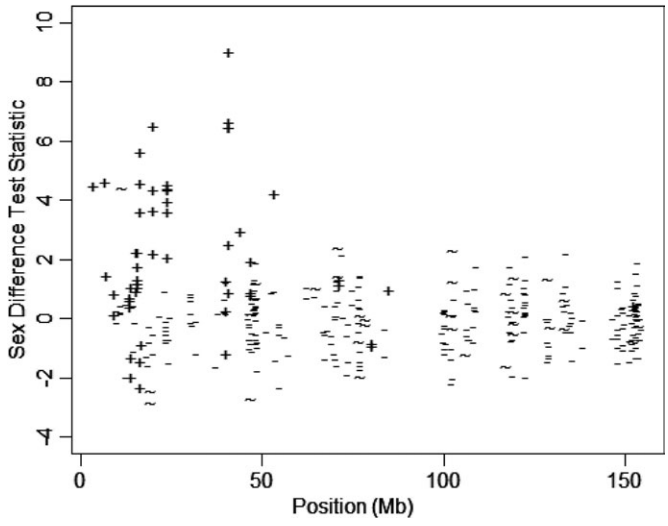

Figure 3. The distribution of test statistics for sex (female−male) across the X chromosome. Symbols represent the proportion of genes escaping X inactivation in fibroblast cell lines as presented by Carrel and Willard (2005). Genes are divided into groups where greater than two-thirds of their samples demonstrated escape from X inactivation (+), those showing between one and two-thirds escaping (∼) and those where less than one-third of samples escape (−).

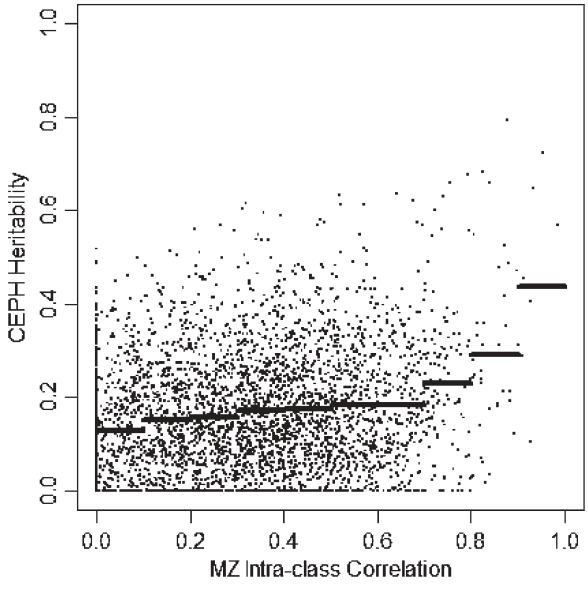

Figure 4. Comparison of intra-class correlation of expression levels in MZ twins with narrow sense heritabilities estimated from CEPH families. Horizontal lines represent the average heritability estimate in the CEPH family for probes in the relevant intra-class correlation window in the MZ twins.

between pair variance for the majority of genes. This strongly suggests a significant genetic component, and thus heritability, for gene expression levels. However, care needs to be taken in making such a conclusion, as the between pair variance of MZ twin pairs includes both additive and non-additive genetic variance as well as potential common environmental variance (22). An analysis of expression levels in a sample of 10 MZ and five dizygotic (DZ) pairs

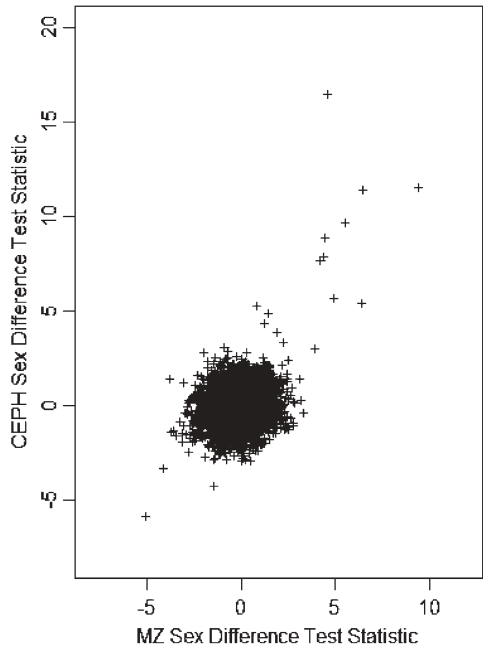

Figure 5. Comparison of test statistics for differential gene expression between males and females in the MZ pairs and CEPH families.

has previously been presented (6). It was shown that the expression levels were more similar within MZ pairs than within DZ pairs, thus demonstrating a genetic component to expression levels. As the distributions of Fisher's Z transformed intra-class correlations for MZ twin pairs is similar in both this study and the study by York *et al*. (this study: mean = 0.34, SD = 0.27; York *et al*.: mean = 0.30, SD = 0.38), their conclusion of a significant genetic component to gene expression levels is likely to extend to this study as well, although the observation that DZ twin pairs show an average zero intra-class correlation indicating that an overall epistatic control of gene expression are unable to be supported by this study. Here, the intra-class correlations from MZ twin pairs were compared with narrow-sense heritability estimates from data on 14 CEPH families. The twin intra-class correlations include effects due to non-additive genetic variation and common environmental variation in addition to additive genetic variance and thus are inflated when compared with narrow-sense heritability estimates. This is reflected in the average heritability estimate from the CEPH families being lower than the intra-class correlation for the heritability. However, a significant correlation was observed between these estimates, particularly when the intra-class correlation and heritability estimates were high and thus have smaller standard errors. A further study has examined the effect of family relationship on variation in gene expression using a sample of CEPH families (4). The distribution of significantly heritable genes (Fig. 1) shows marked similarity to the upper end of the distribution of the proportion of variance attributable to the pairing of the twins from this study (Fig. 1).

A potential source of bias in estimating the heritability of expression level is sequence variation in the investigated individuals for the target sequence of the probe (23,24). If such

Table 2. List of probes showing differential expression between males and females in the CEPH families

Systematic	Gene symbol	Chromosomal location	Test statistic	Gene description
203974_at	*HDHD1A*	Xp22.32	16.54	Haloacid dehalogenase-like hydrolase domain containing 1A
203992_s_at	*UTX*	Xp11.2	11.61	Ubiquitously transcribed tetratricopeptide repeat, X chromosome
201018_at	*EIF1AX*	Xp22.12	11.49	Eukaryotic translation initiation factor 1A, X-linked
208174_x_at	*U2AF1L2*	Xp22.1	9.76	U2(RNU2) small nuclear RNA auxiliary factor 1-like 2
204061_at	*PRKX*	Xp22.3	8.96	Protein kinase, X-linked
207551_s_at	*MSL3L1*	Xp22.3	7.91	Male-specific lethal 3-like 1 (Drosophila)
201589_at	*SMC1L1*	Xp11.22-p11.21	7.73	SMC1 structural maintenance of chromosomes 1-like 1 (yeast)
201029_s_at	*CD99*	Xp22.32; Yp11.3	−5.77	CD99 antigen
202383_at	*SMCX*	Xp11.22-p11.21	5.71	Smcy homolog, X-linked (mouse)
201210_at	*DDX3X*	Xp11.3-p11.23	5.48	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked
201099_at	*USP9X*	Xp11.4	5.35	Ubiquitin specific peptidase 9, X-linked (fat facets-like, Drosophila)
203767_s_at	*STS*	Xp22.32	4.92	Steroid sulfatase (microsomal), arylsulfatase C, isozyme S
200933_x_at	*RPS4X*	Xq13.1	4.38	Ribosomal protein S4, X-linked
204161_s_at	*ENPP4*	6p21.1	−4.18	Ectonucleotide pyrophosphatase/phosphodiesterase 4 (putative function)

Probes replicating results from the analysis of MZ twin pairs are given in bold.

Table 3. Summary of additional probes showing evidence for differential expression between the sexes when combining results from MZ pairs and CEPH families

Systematic	Gene symbol	Chromosomal location	Test statistic	Gene description
217176_s_at	*ZFX*	Xp21.3	12.20 (+)	Zinc finger protein, X-linked
219351_at	*TRAPPC2*	Xp22	7.79 (+)	Trafficking protein particle complex 2
200964_at	*UBE1*	Xp11.23	7.68 (+)	Ubiquitin-activating enzyme E1 (A1S9T and BN75 temperature sensitivity complementing)
201925_s_at	*DAF*	1q32	6.58 (−)	Decay accelerating factor for complement (CD55, Cromer blood group system)
208770_s_at	*EIF4EBP2*	10q21-q22	6.42 (+)	Eukaryotic translation initiation factor 4E binding protein 2
212878_s_at	*KNS2*	14q32.3	5.82 (−)	Kinesin 2

The sign besides the combined test statistic indicates the direction of the average expression level in females relative to males.

variation is present, estimates of expression differences between individuals may not reflect actual differences but rather sequence variation in the population. In the results presented here, the between pair variance could be inflated, biasing the intra-class correlation upwards as any sequence variation is also grouped with twin pairs. However, it is unlikely that this source of bias will affect the estimates of differences due to sex as sequence variation will be random across the sexes. The extent of this potential bias was addressed by examining individual probe levels within a probe set. A linear mixed model was used to model the deviation of all background-corrected probes for a particular target sequence from their median within each individual. This measure is used as it removes the overall expression differences between individuals and follows the approach used by the MAS 5.0 algorithm used in calculating the relative weight of each probe when summarizing the probe set expression levels (25). As with the overall expression levels, these were used to estimate an intra-class correlation for the deviations in probe level. The model used a probe by pair interaction to allow individual probes to deviate independently within a pair. A similar approach is used by Doss *et al*. (24), who fit a probe by strain interaction to model SNPs in mouse lines. The premise of this approach is that if sequence variation is affecting probe binding, then the between pair variance (and thus intra-class correlation) will be increased.

Figure 6 shows the comparison of the intra-class correlations for the probe deviation to that of the summarized expression level. The data show a weak correlation of 0.17, indicating that the differences in probe binding across pairs is responsible for only 2.8% of the variation in expression level. The results from correcting for probe binding differences across pairs using a simple linear regression showed little deviation from the distribution of heritabilities presented in Figure 1, with the overall mean reducing from 0.31 to 0.28. The correlation between the probe deviation heritability and the absolute value of the test statistic for sex differences is 0.01, which is not significantly different from zero ($P = 0.15$), as expected.

The effect of sex on gene expression in LCLs was examined in both the MZ twin pairs and CEPH families. The majority of the genes detected as differentially expressed were located on the sex chromosomes. Although it would be expected for many genes to be differentially expressed between the sexes

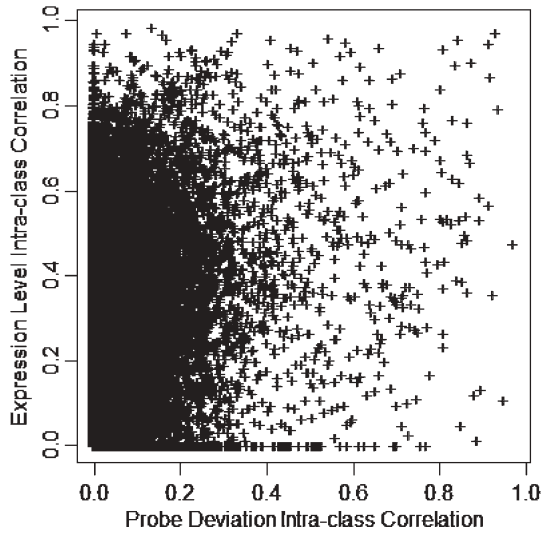

Figure 6. Comparison of MZ twin pair intra-class correlations for probes set expression level and individual probe deviations from the probe set median. The low correlation between the two measures indicates that bias in expression level intra-class correlations due to sequence variation between pairs is limited.

given the readily observable differences between males and females for a wide variety of traits (26), this effect is expected to be much reduced in cell-lines that are grown on a uniform media. Eleven of the 31 genes that were detected as being differentially expressed between the sexes in MZ twin pairs also were present in the CEPH data subset. Of these, 10 were replicated using the CEPH family data, with the remaining gene (*FEZ1*) having a test statistic (−3.29) approaching the chosen significance threshold in the CEPH families. However, of the four genes detected as differentially expressed between the sexes in the CEPH families but not in the MZ twin pairs, the test statistics in MZ pairs were relatively small, although all in the same direction. The combined MZ twin and CEPH family tests for differential expression detected six further genes as being differentially expressed.

Differences in expression levels between males and females on autosomal chromosomes is also of interest, although these genes are not able to be used to validate the performance of different analysis methods as the understanding of the underlying differences is, in general, incomplete. The effect of sex on gene expression levels in peripheral blood cells has been examined, identifying a number of genes as being differentially expressed between the sexes (27). The use of peripheral blood cells instead of cell lines introduces much more variation in genes that are differentially expressed between the sexes. This is reflected in that the majority of their significantly differentially expressed genes are not located of the sex chromosomes. However, none of these overlap with the autosomal genes detected as differentially expressed in this study.

The mixed model framework is an extremely versatile approach to the analysis of gene expression data. The ability to simultaneously fit observed fixed and random effects allows a wide variety of study designs to be analysed. Here, we included random effects to account for correlations in the data caused by genetic relatedness between sampled individuals and for the effect of replicated gene expression measurements. The testing of the fixed effect of sex on gene expression is exactly the same approach used to model gene expression differences between any two groups, perhaps of most interest being disease status. Although not examined in this study, testing the effects of continuous variables on gene expression is readily achieved in a similar manner. For example, testing the effect of age on the gene expression data from the MZ twin pairs used in this study provides evidence for the expression level of HLA-DQB1 and BCL11A decreasing with age and suggestive evidence of effects on several other genes (CDC42, IGLJ3, ITPK1, P2RY5, PLTP, S100A11). However, these results need to be treated with caution as the distribution of the ages in this sample has several outliers. The appeal of the mixed model framework has seen a rapid increase of its use in the analysis of gene expression data in the past few years to its current position as the gold-standard approach.

MATERIALS AND METHODS

Monozygous twin pairs

The sample consisted of participants with epilepsy recruited from the Australian Epilepsy Twin Database (AETD), participants with bipolar disorder, brief psychotic disorder and schizophrenia from the Queensland Centre for Mental Health Research Twins Database and control participants recruited from the Australian Twin Registry. The exact composition is: one pair discordant for schizophrenia, one pair discordant for brief psychotic disorder, three pairs discordant for bipolar disorder, five pairs discordant for epilepsy, four pairs concordant for epilepsy and five unaffected pairs. Hence, the cell lines are based upon 18 affected individuals (one schizophrenic, one person affected with brief psychotic disorder, three with bipolar disorder and 13 with epilepsy) and 20 unaffected individuals. Because of the mixture of disease status among the samples, the relatively balance nature of this status across pairs and sexes and the expectation that each disease will only alter the expression status of a very small fraction of genes, the effects of this sample heterogeneity on the results presented here is likely to be limited. This has been confirmed by modelling the effects of the most common disease in this sample (epilepsy). Adding disease status to the model used here has essentially no effect on the estimates of the fixed effects and twin correlation. The correlation of the sex test statistics for the models with and without disease status was 0.995 and the correlation between the estimates of twin correlation was 0.98. Zygosity was tested using an AmpFLSTR Profiler Plus PCR Amplification Kit (Applied Biosystems), and data was analysed using Genescan v3.7.1 software (Applied Biosystems) and Genotyper v2.5 software (Applied Biosystems) to confirm MZ twin status. In total, 19 MZ twin pairs were obtained, 10 male and nine female pairs. A male twin pair had blood taken on two separate occasions, 64 days apart. The twin pairs had an average age of 34 with an SD of 12 years.

Sample preparation

LCLs were established by Epstein-Barr virus transformation of lymphocytes (28). For RNA, cell lines were all grown under tightly controlled growth conditions in the same batch of RPMI 1640 media with 10% FCS and antibiotics, to limit the cell culture effects on RNA production. Total RNA was extracted from samples using Qiagen RNeasy Midi-Kits, when cells were in log phase growth. RNA from all samples was run on an Agilent Bioanalyzer to assure quality and to obtain concentration.

Microarray hybridization

Expression profiles were generated by hybridizing 5 μg of total RNA to Affymetrix Human Genome U133 plus 2.0 Gene Chips (HG U133plus 2.0) according to the Affymetrix Eukaryote One-cycle protocol. Briefly, 5 μg of total RNA were used to generate biotinylated cRNA, which was fragmented and hybridized to an Affymetrix whole genome chip, HG U133plus 2.0 for 16 h at 45°C in an Affymetrix Hybridization Oven 640. Gene Chips were then washed and stained on an Affymetrix Fluidics Station 450 and subsequently scanned on an Affymetrix GeneChip Scanner 3000 to obtain fluorescence intensities.

Data pre-processing

Relative expression values were generated for each transcript using Affymetrix MAS5.0 algorithm in the GeneChip® Operating Software (GCOS) version 1.2, with the average intensity of all transcripts on each array scaled to 150. Data were then filtered for transcripts which were present across 100% of samples according to the global-error threshold calculated by GeneSpring's (v7.2) cross-gene error model. An important consequence of only including genes detected as expressed in all samples is the removal of all Y chromosome transcripts.

CEPH family data

Several gene expression data sets from Centre d'Etude du Polymorphisme Humain (CEPH) Utah pedigrees (29) are publicly available. These can be used to provide validation of the genes determined to be significantly differentially expressed between the sexes and the estimates of heritability of gene expression. Here, the data set generated by Morley *et al*. (5) is chosen, as the chip used in that study used a subset of probes on the chip used in this study and the data pre-processing methods in that study are similar to those used here. This microarray data has GEO accession no. GSE1485. Briefly, the data were from members of 14 CEPH families (CEPH 1333, 1340, 1341, 1345, 1346, 1347, 1362, 1408, 1416, 1418, 1421, 1423, 1424 and 1454) with expression levels measured on Affymetrix Genome Focus Arrays. This data set had partial replication with a total of 277 chips on the 194 individuals. The probes were filtered to only include those that were selected in the MZ twin data set. All these probes would also have been selected using the 'always present' criterion originally used in filtering the MZ twin data set.

Data normalization

Pre-processed data were transformed using the generalized-logarithm transformation (30–32) to achieve a stabilized variance distribution across average expression levels. This transformation can be written in several forms and is presented here as:

$$y = \ln\left(\frac{x + \sqrt{x^2 + c}}{2}\right)$$

where x is the pre-processed expression level, y the transformed level and c a constant chosen for optimal variance stabilization. This is an attractive class of transformation as it approaches the widely used log transformation as x gets large relative to c, but allows an approximately linear transformation at low levels of x consistent with the observed data patterns. As the data has been filtered to only include genes which show significant gene expression in all samples, the usual approach of estimating c using a regression of the variance on the mean of a gene expression level is not appropriate. Instead, c was chosen such that the correlation of a gene expression level on the rank of its mean was zero. This approach minimizes the leverage of genes with high expression on the final transformation.

Further normalization was performed to allow expression levels to be compared across chips and genes. This was achieved using the mixed linear model.

$$y_{ij} = \mu + C_i + G_j + r_{ij}$$

where y_{ij} is the transformed expression level for individual i on chip j, C_i and G_j are random effects removing variation in the data due to chip and gene differences and r_{ij} is the residual. The between chip variance is expected to be small due to the scaling that was performed during the pre-processing of the data. The residuals from this model were used in all further analyses.

Analysis of the effect of genotype on expression levels

Linear mixed models were used to assess the effects of sex and genotype on the normalized gene expression levels. For the MZ pairs, the model used was:

$$r_{ijk} = \mu + S_{ij} + P_i + R_k + E_{ijk}$$

where the response variable r_{ijk} is the normalized expression level for the jth individual of the ith pair in replicate k. The variable μ represents the average expression level across all individuals and the fixed effect S_{ij} is the difference between the average expression levels of males and females. The remaining terms are the random effects that partition the remaining variance in the data with P_i being the variance between pairs, R_k the replication variance and E_{ijk} the residual variance. Parameters were estimated using residual maximum likelihood (REML) with the program ASReml (33). The intra-class correlation for each gene was calculated as $\sigma^2_P/(\sigma^2_P + \sigma^2_R + \sigma^2_E)$, where additional subscripts are removed for simplicity. This is simply the proportion of the variance in the data explained by pair and in the absence of common

environmental effects is a measure of the broad sense heritability of a genes expression level.

For the analysis of the CEPH family data, the pair variance was replaced with an additive genetic variance (18,19) in order to account for the family structure in the data. In this case, the proportion of the variance explained by the additive genetic effect [calculated as $\sigma_A^2/(\sigma_A^2+\sigma_R^2+\sigma_E^2)$ where σ_A^2 is the additive genetic variance] is the narrow-sense heritability of the gene expression level.

Analysis of the effect of sex on expression levels

The significance of the effect of sex on the expression level of an individual gene probe set was tested using the same mixed linear model framework as in the analysis of the effects of genotype. An approximate *t*-test was constructed by dividing the effect by its standard error. This is different from a standard *t*-test in that the variances in the model are estimated by REML instead of the usual full maximum likelihood and thus the degrees of freedom of the test statistic are in general unknown. However, using the usual residual degrees of freedom provides a close (though anti-conservative) approximation with moderate data sizes. An improvement can be made to the estimation of the test statistic for sex by noting that the variance used in its construction will itself have a large variance given the limited number of data points used in its estimation. Thus, under the assumption of a uniform variance across genes, combining the estimates of the variance of the sex effect will increase the accuracy of the estimated test statistic. This approach of combining information across genes has been termed 'shrinkage' in the literature (34). Given the transformation performed in data normalization aimed to stabilize the data variance across probes, the assumption of uniform variance seems reasonable. Let the variance of the estimated sex effect on probe i be X_i. The shrinkage estimator takes the form (35).

$$S_i = \prod_{i=1}^{n}(X_i)^{1/G}\exp\left[w(\ln(X_i) - \overline{\ln(X_i)})\right].$$

with

$$w = \left(1 - \frac{2/\nu}{\mathrm{var}(\ln(X_i))}\right)_+$$

where $(x)_+$ denotes $\max(0, x)$. Here, the various corrections for biases that occur only with extremely small samples sizes have been removed for simplicity. The weight, w falling between zero and one, determines the relative influence of individual and averaged variances. When w is zero, the shrinkage estimate is the geometric mean and w of one returns the original variance. In the MZ twin pairs, the shrinkage procedure reduced the largest standard errors by 14% (equivalent to a reduction of the variance of the effect size of 26%). However, in the CEPH families, where the samples size is much larger, the estimated standard errors are reduced at most by 2%.

ACKNOWLEDGEMENTS

We acknowledge the participation of individuals sampled in this work and their family members. We would like to thank the Peter MacCallum Cancer Institute Microarray Facility, in particular Dr Andrew Holloway and Dileepa Diyagama, for their technical assistance with the microarray hybridizations. This article was improved with the helpful comments of two referees and the editor. This research was supported by Australian National Health and Medical Research Council Grant (NHMRC) Grants 144105 and 389892.

Conflict of Interest statement. None declared.

REFERENCES

1. Lander, E.S. (1999) Array of hope. *Nat. Genet.*, **21** (suppl), 3–4.
2. Miklos, G.L. and Maleszka, R. (2004) Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.*, **22**, 615–621.
3. Cobb, J.P., Mindrinos, M.N., Miller-Graziano, C., Calvano, S.E., Baker, H.V., Xiao, W., Laudanski, K., Brownstein, B.H., Elson, C.M., Hayden, D.L. *et al*. (2005) Application to genome-wide expression analysis to human health and disease. *Proc. Natl Acad. Sci. USA*, **102**, 4810–4806.
4. Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwars, E., Phillips, J.W., Sachs, A. and Schadt, E. (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.*, **75**, 1094–1105.
5. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
6. York, T.P., Miles, M.F., Kendler, K.S., Jackson-Cook, C., Bowman, M.L. and Eaves, L.J. (2005) Epistatic and environmental control of genome-wide gene expression. *Twin Res. Hum. Genet.*, **8**, 5–15.
7. Parmigiani, G., Garret, E.S., Irizarry, R.A. and Zeger, S.L.(eds) (2003) *The Analysis of Gene Expression Data: Methods and Software*, Springer-Verlag, New York.
8. Holzman, T. and Kolker, E. (2004) Statistical analysis of global gene expression data: some practical considerations. *Curr. Opin. Biotechnol.*, **15**, 52–57.
9. Barash, Y., Dehan, E., Krupsky, M., Franklin, W., Geraci, M., Friedman, N. and Kaminski, N. (2004) Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. *Bioinformatics*, **20**, 839–846.
10. Rajagopalan, D. (2002) A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics*, **19**, 1469–1476.
11. Lyon, M.F. (1961) Gene action in the X chromosome of the mouse (*Mus musculus* L.). *Nature*, **190**, 372–373.
12. Russel, L.B. (1961) Genetics of mammalian sex chromosomes. *Science*, **133**, 1795–1803.
13. Disteche, C.M. (1995) Escape from X inactivation in human and mouse. *Trends Genet.*, **11**, 17–22.
14. Heard, E., Clerc, P. and Avner, P. (1997) X-chromosome inactivation in mammals. *Annu. Rev. Genet.*, **31**, 571–610.
15. Craig, I.W., Mill, J., Craig, G.M., Loat, C. and Schalkwyk, L.C. (2004) Application of microarrays to the analysis of the inactivation status of human X-linked genes expressed in lymphocytes. *Eur. J. Hum. Genet.*, **12**, 639–646.
16. Carrel, L. and Willard, H.F. (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, **434**, 400–404.
17. McCulloch, C.E. and Searle, S.R. (2001) *Generalized, Linear and Mixed Models*, John Wiley and Sons, New York.
18. Almasy, L. and Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.
19. George, A.W., Visscher, P.M. and Haley, C.S. (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, **156**, 2081–2092.

20. Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P. and White, K.P. (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, **440**, 242–245.
21. Self, S.G. and Liang, Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Am. Stat. Assoc.*, **82**, 605–610.
22. Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*, Sinauer Associates Inc., Sunderland.
23. Alberts, R., Terpstra, P., Bystrykh, L.V., de Haan, G. and Jansen, R.C. (2005) A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide array. *Genetics*, **171**, 1437–1439.
24. Doss, S., Schadt, E.E., Drake, T.A. and Lusis, A.J. (2005) *Cis*-acting expression quantitative trait loci in mice. *Genome Res.*, **15**, 681–691.
25. Affymetrix (2002) Statistical Algorithms Description Document. Affymetrix, Santa Clara, CA.
26. Weiss, L.A., Pan, P., Abney, M. and Ober, C. (2006) The sex-specific genetic architecture of quantitative traits in humans. *Nat. Genet.*, **38**, 218–222.
27. Whitney, A.R., Diehn, M., Popper, S.J., Alizadeth, A.A., Boldrick, J.C., Relman, D.A. and Brown, P.O. (2003) Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. USA*, **100**, 1896–1901.
28. Neitzel, H. (1986) A routine method for the establishment of permanent growing lymphoblastoid cell lines. *Hum. Genet.*, **73**, 320–326.
29. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M. and White, R. (1990) Centre d'Etude du Polymorphism Humain (CEPH): collaborative genetic mapping of the human genome. *Genomics*, **6**, 575–577.
30. Durbin, B.R., Hardin, J.S., Hawkins, D.M. and Rocke, D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18** (Suppl. 1), S105–S110.
31. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.
32. Munson, P.A. (2001) 'Consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. *In GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data*, Bethesda, MD, November 2001.
33. Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. (2002) *ASReml User Guide Release 1.0*, VSN International Ltd., Hemel Hempstead.
34. Allison, D.B., Cui, X., Page, G.P. and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.
35. Cui, X., Hwang, J.T.G., Qiu, J., Blades, N.J. and Churchill, G.A. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.