# Assessing Individual Differences in Genome-Wide Gene Expression in Human Whole Blood: Reliability Over Four Hours and Stability Over 10 Months

Emma L. Meaburn, Cathy Fernandes, Ian W. Craig, Robert Plomin, and Leonard C. Schalkwyk

*Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, United Kingdom*

Studying the causes and correlates of natural variation in gene expression in healthy populations assumes that individual differences in gene expression can be reliably and stably assessed across time. However, this is yet to be established. We examined 4-hour test–retest reliability and 10 month test–retest stability of individual differences in gene expression in ten 12-year-old children. Blood was collected on four occasions: 10 a.m. and 2 p.m. on Day 1 and 10 months later at 10 a.m. and 2 p.m. Total RNA was hybridized to Affymetrix-U133 plus 2.0 arrays. For each probeset, the correlation across individuals between 10 a.m. and 2 p.m. on Day 1 estimates test–retest reliability. We identified 3,414 variable and abundantly expressed probesets whose 4-hour test-retest reliability exceeded .70, a conventionally accepted level of reliability, which we had 80% power to detect. Of the 3,414 reliable probesets, 1,752 were also significantly reliable 10 months later. We assessed the long-term stability of individual differences in gene expression by correlating the average expression level for each probe-set across the two 4-hour assessments on Day 1 with the average level of each probe-set across the two 4-hour assessments 10 months later. 1,291 (73.7%) of the 1,752 probe-sets that reliably detected individual differences across 4 hours on two occasions, 10 months apart, also stably detected individual differences across 10 months. Heritability, as estimated from the MZ twin intraclass correlations, is twice as high for the 1,752 reliable probesets versus all present probesets on the array (0.68 vs 0.34), and is even higher (0.76) for the 1,291 reliable probesets that are also stable across 10 months. The 1,291 probesets that reliably detect individual differences from a single peripheral blood collection and stably detect individual differences over 10 months are promising targets for research on the causes (e.g., eQTLs) and correlates (e.g., psychopathology) of individual differences in gene expression.

**Keywords:** blood, human, individual differences, gene expression, reliability, genomewide

Oligonucleotide microarrays have made it possible to study gene expression at the genome-wide level of the transcriptome, which is the first step between the genome and the many paths leading to the phenome. Thousands of genome-wide expression (GWE) studies have begun to chart gene expression at a species-wide level across tissues and across development, and within a species, to compare *mean* expression levels for conditions (e.g., before and after administration of a drug, Yuferov et al., 2005) and for groups (e.g., cases versus controls; Konradi, 2005; and inbred strains of mice; Fernandes et al., 2004; Korostynski et al., 2006).

More recently, GWE research has progressed from studying average differences between groups to describe the extent of normal variation in gene expression among healthy individuals (Eady et al., 2005; Radich et al., 2004; Whitney et al., 2003). A large-scale study of individual differences in genomewide gene expression across diverse populations is currently underway (Nica & Dermitzakis, 2008). One direction for GWE research is to study the causes (i.e., genetic, environmental and epigenetic factors) and correlates (e.g., psychopathology) of these individual differences. For example, human and animal research has moved toward treating gene expression as a complex quantitative trait and identifying DNA variation (quantitative trait loci; QTLs) associated with individual differences in gene expression (expression QTLs; eQTLs; Breitling et al., 2008; Dixon et al., 2007; Emilsson et al., 2008; Goring et al., 2007; Nica & Dermitzakis, 2008; Rockman & Kruglyak, 2006; Stranger et al., 2007).

Individual differences research is more statistically demanding than mean differences research. Means analysis treats individual differences in gene expression as an error term; in contrast individual differences

*Received 17 December, 2008; accepted 19 March, 2009.*

*Address for correspondence: Emma L Meaburn, SGDP, PO 82, Institute of Psychiatry, King's College London, De Crespigny Park, London, UK, SE5 8AF. E-mail: e.meaburn@iop.kcl.ac.uk.*

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

research treats gene expression as a quantitative trait and focuses on the variation of gene expression between individuals. Reliability and long-term stability — that is, maintenance of the rank order of individual differences over time — is a prerequisite for analyses of the causes and correlates of individual differences of GWE. The most stringent test of reliability is the correlation between individuals on separate measurement occasions, called test-retest reliability. Usually test–retest reliability is assessed over a few hours or few days or at most over a few weeks. Test–retest *stability* assesses the extent to which reliable individual differences are maintained over longer periods of time. Test–retest reliability and stability are rooted in psychometric research and assesses the extent to which the rank order of individuals is maintained despite momentary 'state' sources of variance.

Although biological (e.g., sex, tissue, age) and technical (e.g., sample processing) sources of variation have been studied exhaustively by correlating GWE profile estimates across microarrays (Bakay et al., 2002; Dumur et al., 2004), we are not aware of research that has investigated the *test-retest reliability* or *stability* of individual differences in GWE — correlating gene expression values for each probeset across individuals whose RNA was obtained on more than one measurement occasion — even though several studies have obtained repeated blood samples (Calvano et al., 2005; Radich et al., 2004; Whitney et al., 2003).

The purpose of the present study was to estimate four-hour test-retest reliability and 10-month test–retest stability for individual differences in gene expression analyzed as quantitative traits, using Affymetrix HG-U133 plus 2.0 expression arrays that assess 54,675 probesets throughout the genome. Although gene expression is tissue specific, we chose to study peripheral blood, as although invasive, it is the most accessible tissue for the large sample sizes needed to power studies of individual differences. Investigating human gene expression in the brain is limited to the use of postmortem brain tissue, which is not suitable for identifying subtle gene expression effects in large human samples. Moreover, a surprising degree of similarity between gene expression in blood and brain has been reported, although the validity of using blood as a surrogate for the brain will depend on the context of the research (Gladkevich et al., 2004; Mohr & Liew, 2007; Nicholson et al., 2004; Pahl, 2005; Sharp et al., 2006). For many genes, expression will be responsive to the environment, but we did not attempt to control for environment because a single uncontrolled measurement occasion for collecting blood would be the most useful design for large human samples, even though it is the most difficult condition for achieving reliability.

Using whole blood obtained on four occasions, we investigated, for the first time, test–retest reliability and test-retest stability of individual differences in GWE in order to identify a core of 'reliably' stable transcripts that can be used to inform and evaluate future substantive studies of the causes and correlates of individual differences in GWE.

## Materials and Methods

### Sample

The sampling frame for this study was the Twins Early Development Study (TEDS), a longitudinal study of behavioral development in a representative sample of twins born in 1994, 1995 and 1996 who have been followed from infancy through adolescence (Oliver & Plomin, 2007). From a sample of healthy 1,000 pairs of 12-year-old monozygotic (MZ) twins, five pairs were selected — four female pairs and one male pair.

### Blood Collection

All 10 subjects visited the Institute of Psychiatry on two occasions with a 10-month delay between each visit. On each visit, venous blood samples were collected at 10 a.m. and again at 2 p.m. using a standard phlebotomy protocol in conjunction with the PAXgene Blood RNA System (Becton & Dickinson, Oxford), which allows the collection, stabilization and transportation of a whole blood cellular RNA sample in a closed evacuated system. For each subject, four PAXgene blood tubes each containing 2.5mL of blood were collected at each of the two occasions. In addition, 3mL blood in an EDTA tube was also collected from each subject at 10am to assess differences in cell sub-type compositions. All cell sub-type counts were in the normal range and comparable across subjects.

### Isolation of Total RNA From Whole Blood

Total RNA was isolated from the PAXgene blood samples using the PAXgene Blood RNA Kit protocol (PreAnalytiX GmbH, Feldbachstrasse, CH-8634 Hombrechtikon). Total RNA yield (µg) and purity (260nm:280nm) were determined using a spectrophotometer. Integrity of ribosomal RNA (rRNA) bands was confirmed by running 10µl of purified RNA on a 1.2% agarose gel.

### cDNA and cRNA Synthesis, Labeling and Hybridization

Expression profiles were generated by hybridizing cRNA derived from 5 g of total RNA to Affymetrix U133 Plus 2.0 Arrays (Affymetrix, Santa Clara, CA) in accordance with the Affymetrix Eukaryote One-Cycle protocol with integrated globin reduction (see Affymetrix GeneChip Globin-reduction Kit Handbook and Affymetrix GeneChip Expression Analysis technical manual). The Affymetrix U133 Plus 2.0 Array has been shown to be reliable (Robinson & Speed, 2007).

Total RNA was concentrated (GeneChip blood RNA concentration kit; PN 900585) and 5 µg used to generate first-strand cDNA synthesis with integrated globin reduction using peptide nucleic acid (PNA) oligonucleotides in order to block reverse transcription of globin mRNA (GeneChip Globin-Reduction RNA controls; PN 900586, GeneChip® Expression 3' Amplification One-Cycle Target Labeling and Control

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

Reagents; PN 900493). After second-strand cDNA synthesis, biotinylated cRNA was generated, fragmented and hybridized to Affymetrix U133 Plus 2.0 Arrays for 16 hours at 45°C in an Affymetrix hybridization oven 640. Arrays were then washed and stained on an Affymetrix fluidic station 450 (protocol FS450_0001).

### Microarray Analysis and Quality Control

Each array was scanned using an Affymetrix GeneChip Scanner 3000 and GeneChip Operating Software (GCOS) version 1.4 was used to obtain fluorescence intensities. The data is MIAME compliant and is available to download at the Gene Expression Omnibus website (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE14844.

The arrays were processed together using Robust Multiarray Average (RMA; Irizarry et al., 2003), implemented in the 'affy' package in the statistical software environment R (http://www.r-project.org/), to produce normalized, background-adjusted, perfect-match, log-transformed probe set summaries.

Total RNA isolation, preparation, and microarray hybridization experiments were processed in two batches. In order to avoid introducing unwanted batch effects, samples were split evenly across both batches, with different samples being allocated to each batch during the isolation, preparation and hybridization steps.

In order to check RNA sample and microarray experiment quality, quality control was performed in three stages (see Affymetrix manual: data analysis fundamentals). First, the probe array images (.dat files) were inspected for the presence of image artefacts (e.g., high/low intensity spots). Second, using the R package 'affyQCReport' (http://www.bioconductor.org/packages/2.2/bioc/vignettes/affyQCReport/inst/doc/affyQCReport.pdf) each array was examined and compared for signal quality differences, average background intensity, scaling factor, percent present call rate and 3'/5' hybridization intensity ratios. Finally, probe-level model fitting was performed using the Bioconductor package 'affyPLM' (http://www.bioconductor.org/packages/2.3/bioc/html/affyPLM.html) to assess relative Log expression (RLE) values and unscaled standard errors (NUSE). Following quality control assessment, four arrays were found to be outliers (TD36282 at 2 p.m. from Day 1, and TD19901 at 10 a.m., TD23462 at 10 a.m. and TD36282 at 10 a.m. from Day 2) and were excluded.

After normalization and probeset summarization with RMA, the signal intensities for the 54,675 probesets were highly similar across the 36 arrays, with Pearson correlations ranging from 0.966 to 0.996 between arrays. Such 'profile' correlations between arrays are largely an indication of technical quality because they reflect the characteristic profile of expression across genes on the array.

Although it is possible that low-signal probesets (low abundance transcripts) may detect individual differences reliably and stably, we present results only for probesets that are detectable above background noise. Low intensity probesets that were called absent in 50% of arrays by the MAS5 algorithm were discarded, leaving 25,864 probesets per array. Results are robust to different definitions of low intensity probesets (data not shown).

### Statistical Analysis

***Test–retest reliability.*** Four-hour test–retest reliability of individual differences in gene expression was assessed by calculating Pearson's product-moment correlation ($r$) of expression values across individuals between 10 a.m. and 2 p.m. for each probeset.

***Test–retest stability.*** 10-month test–retest stability of individual differences in gene expression was assessed by averaging the two 4-hour assessments on each measurement occasion (original dataset and 10-month follow-up dataset) and correlating expression values across individuals between datasets for each probeset.

***Heritability.*** MZ twin intraclass correlations were calculated using the R package 'psy' to provide 'upper limit' estimates for heritability of gene expression. An ICC consistency estimate was calculated for Day 1 at 10 a.m. (five twin pairs), Day 1 at 2 p.m. (four twin pairs) and Day 2 at 2 p.m. (five twin pairs), for each probeset. An average of the three time points was used. Day 2 at 10 a.m. was not used due to array exclusions.

***Function and network analysis.*** A core analysis was performed using Ingenuity Pathway Analysis (IPA) to identify the top functions and pathways associated with our dataset (right-tailed Fisher's exact test with Benjamini and Hochberg method for correction for multiple testing). Affymetrix probeset IDs were used as identifiers and the Human Genome U133 plus2 array was used as the reference set. Probesets were annotated using Affymetrix's NetAffx resource (Annotation Release 27).

## Results

### Four-Hour Test–Retest Reliability of Individual Differences in Gene Expression

As mentioned above, one array was dropped; the other nine individuals were available for analysis. For each probeset, the Pearson product–moment correlation ($r$) across the nine individuals between 10 a.m. and 2 p.m. estimates test–retest reliability of individual differences in gene expression. We calculated 4-hour test–retest reliability for each of the 25,864 probesets detectable above background noise ('present' probesets). The mean test–retest reliability was 0.338. Test–retest reliability was statistically significant ($r \geq 0.55$; $p < .05$, one-tailed) for 9,238 (35.7%) of the 25,864 probesets; 1,293 (5.0%) would be expected to be significant by chance alone. Raising the bar for test–retest reliability to 0.70, which we could detect with 80% power ($p$ = .05, one-tailed) with our sample size of 10, 5,339 (20.6%) probesets met this criterion for reliability. See Figure 1.

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

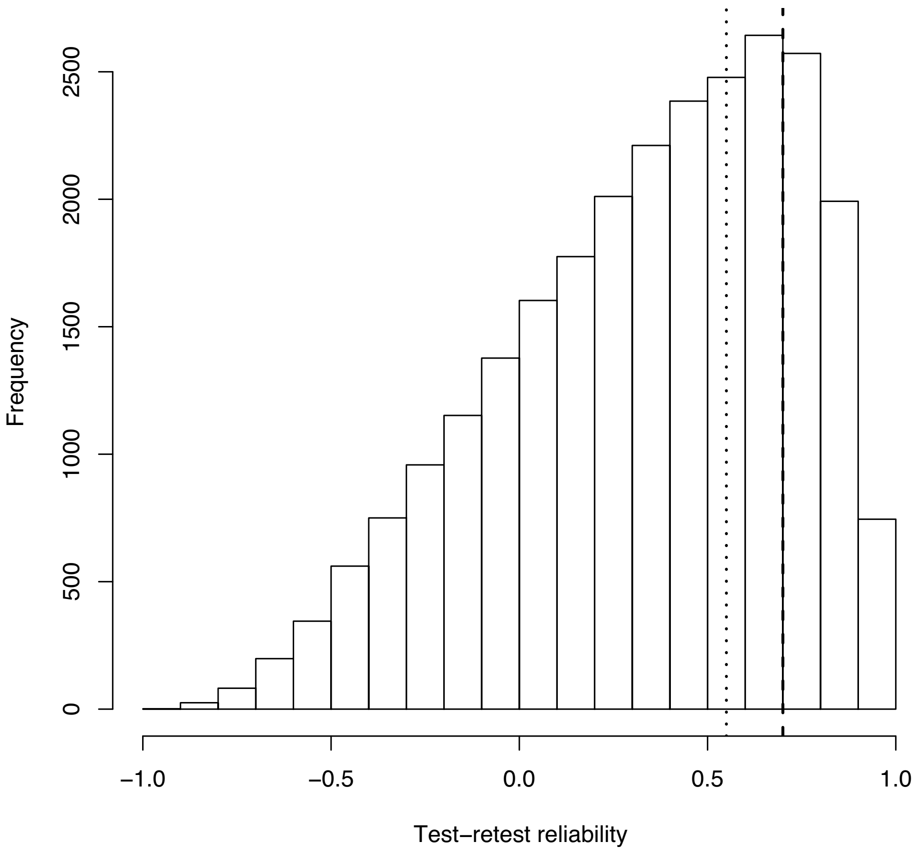

**Figure 1**

Distribution of test–retest reliabilities for 25,864 'present' probesets. The dotted line indicates a test–retest correlation of .55 ($p < .05$, 50% power), the dashed line designates a test–retest correlation of .70 ($p < .05$, 80% power).

It should be emphasized that test–retest reliability focuses on detection of differences between individuals and not just technical reproducibility. Test–retest reliability will only exist if there are individual differences. For this reason we explored the relationship between test–retest reliability and variability. Because variance increases with the mean, we used as an index of variability the coefficient of variation (CV), which is the ratio of the standard deviation to the mean. CV was calculated for each of the 25,864 probesets. The median CV was 0.034, ranging from 0.002 to 0.42. Selecting the most variable probesets using a median split of CV (CV ≥ .034, $N$ = 13,007), increases the average test–retest reliability from .338 to .359. Furthermore, a greater proportion of the 13,007 variable probesets (3414, 26.3%) meet our criterion for reliability ($r > 0.70$).

### Four-Hour Test–Retest Reliability and Transcript Abundance

Low abundance of transcripts could be a source of low reliability, even though transcripts of low abundance may be of biological importance. We therefore examined the relationship between test–retest reliability and probeset signal intensity for the 25,864 'present' probesets in greater detail by correlating test–retest reliability with the average probeset signal across the 18 arrays. Across all 25,864 'present' probesets, reliability was moderately correlated with probeset signal intensity ($r = 0.21$, $p < .001$) indicating that reliability is related to abundance even among abundant transcripts.

### Four-Hour Test–Retest Reliability of Individual Differences in Gene Expression 10 Months Later

We repeated the 4-hour test–retest reliability analysis in the 10 month follow-up dataset to confirm the reliability of the 3,414 probesets that detect individual differences over a 4-hour period reliably. Due to the exclusion of outlier arrays (see Materials and Methods: Microarray Analysis and Quality Control section) the number of paired arrays available for analysis was seven.

We calculated 4-hour test–retest reliability for each of the 25,864 'present' probesets in the 10-month follow-up dataset; the mean test-retest reliability was 0.474 and was statistically significant ($r \geq 0.55$; $p < .05$, one tailed) for 13,156 (50.9%) probesets.

Selecting variable probesets with a CV greater than the median (CV ≥ .033) yields 13,054 probesets, of which 5,174 (39.6%) met our criterion for reliability ($r > 0.70$).

Of the 5,174 probesets that reliably detect individual differences in gene expression over 4 hours in

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

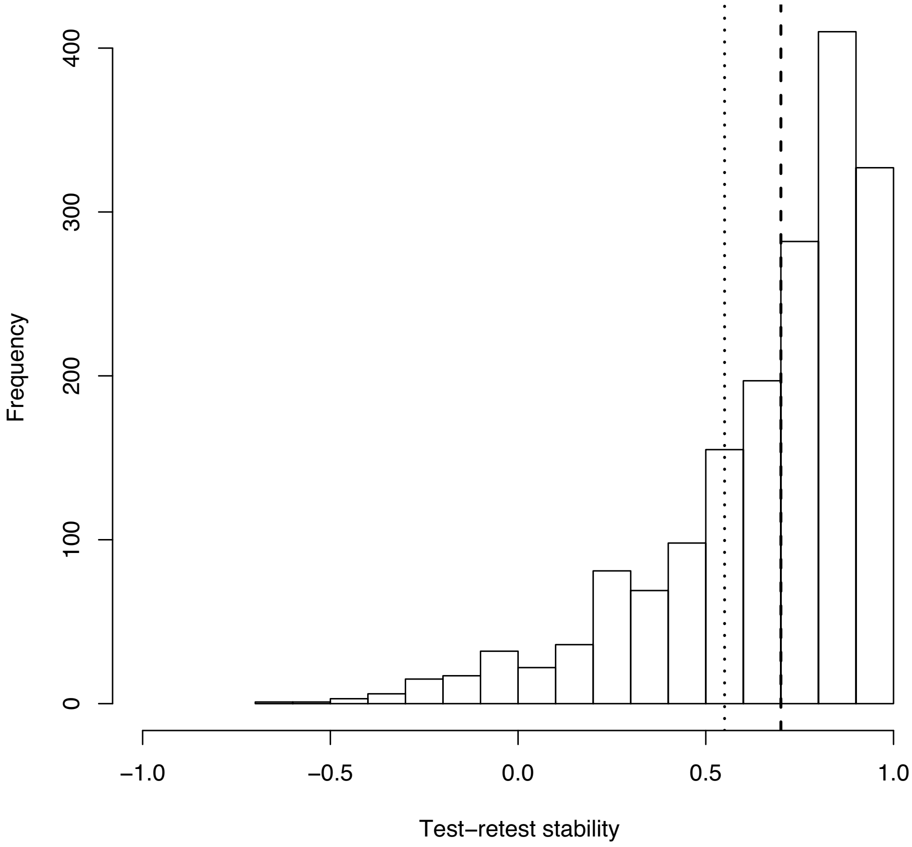

**Figure 2**

Distribution of 10-month stability correlations for individual differences in gene expression for 1,752 probesets that reliably detect individual differences in gene expression across a 4-hour period from blood collection on two occasions 10 months apart. The dotted line indicates test–retest correlation of .55, the dashed line indicates a test–retest correlation of .70.

the 10-month follow-up dataset, 1,752 (33.9%) reliably detected individual differences over 4 hours 10 months earlier.

### Ten-Month Stability of Individual Differences in Gene Expression

The purpose of this analysis is to determine the extent to which the 1,752 probesets that reliably detect individual difference in gene expression from a single blood collection across a 4-hour period on two occasions 10 months apart, also stably detect individual difference in gene expression across 10 months.

In order to increase reliability, the expression level of each probeset was, where available, averaged across the two 4-hour assessments (10 a.m. and 2 p.m.) in the original dataset and in the 10-month follow-up dataset. For each averaged probe-set expression level, the correlation across individuals between the original dataset and the 10-month follow-up dataset was used to estimate 10-month stability of individual differences in gene expression.

The test–retest 10-month stability was statistically significant ($r \geq 0.55$; $p < .05$, one-tailed) for 1,291 (73.9%) of the 1,752 probesets that showed reliable individual differences across a 4-hour period on two occasions, 10 months apart. The mean test–retest stability for the 1,752 probesets was 0.667. (See Figure 2.)

Raising the bar for test–retest stability to 0.70, which we could detect with 80% power ($p = .05$, one-tailed), 1,019 (58.2%) probesets met this criterion.

### Reliability, Stability and Heritability

Reliability and stability of detection of individual differences create a ceiling for heritability. A twin intraclass correlation for the MZ twin pairs was calculated for each probeset as an ‘upper-limit’ estimate of heritability. Although five MZ twin pairs only provide power to detect correlations greater than 0.70 as significant ($p < .05$), our goal is to examine ‘heritability’ estimates as a function of the reliability and stability of individual differences in gene expression over time.

The average MZ twin ICC was 0.34 for all 25,864 ‘present’ probesets. However, for the 3,414 ‘present and variable’ probesets that detect individual differences in gene expression reliably, the average MZ twin correlation was 0.60. Focusing on the 1,752 ‘present and variable’ probesets that detect individual differences reliably on *two* occasions, 10 months apart, the average MZ twin correlation increases to 0.68. Turning to 10-month test–retest stability, heritability is greatest for the 1,291 probesets that ‘stably’ reliably detect individual differences in gene expression; the average twin intraclass correlation is 0.76.

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

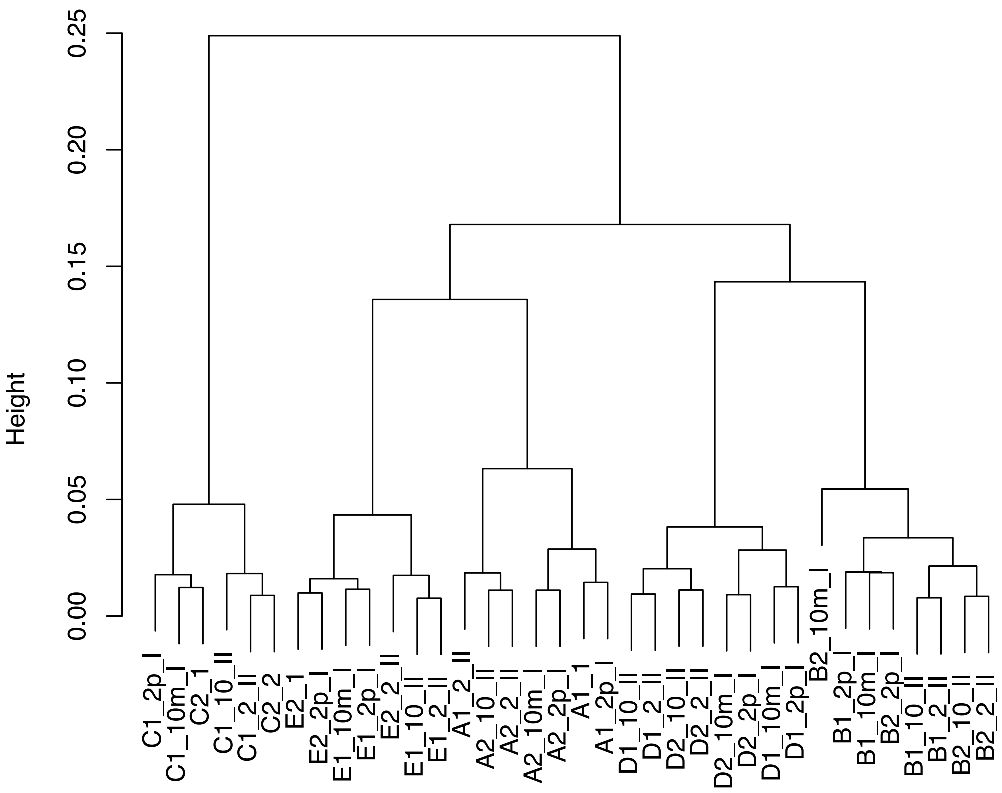

**Figure 3**

Hierarchical clustering across all 36 arrays for the 1,291 probesets that reliably and stably detect individual differences in gene expression. Five distinct branches can be seen, each representing a twin pair (pairs A to E).

The heritability of the 1,291 probesets that reliably and stably detect individual differences in gene expression can be visualized by hierarchical clustering, using 1-R as the distance measure (where R represents the profile correlations of the gene expression signal; see Figure 3.) The five twin pairs can clearly be distinguished, indicating the pervasive heritability of gene expression for these probesets.

### Reliability, Stability and Function

The 1,291 probesets that reliably and stably detect individual differences in gene expression over time represent 775 Entrez genes which are distributed widely across the genome (several transcripts are represented by multiple probesets on the arrays) and expressed widely across tissues. Nearly all (88.8%) of the genes are expressed in brain as well as blood (Zhang et al., 2005).

No particular functional themes were identified. Ingenuity Pathways Analysis (IPA) of the 1,291 probesets was performed to obtain a high-level overview of the general biology associated with their networks and functions. 1,029 of the 1,291 probesets were mapped. The top-associated IPA networks were connective tissue disorders, inflammatory disease, skeletal and muscular disorders, skeletal and muscular system development and function, tissue development, and cell-to-cell signaling and interaction. The top five molecular and cellular functions associated with the 1,291 probesets were cellular growth and proliferation, cell-to-cell signaling and interaction, cell death, cell signaling and molecular transport. This is similar to the profile of the present probesets without reliable or stable individual differences.

Our website (http://sgdp.iop.kcl.ac.uk/oleo/meaburn/) lists details for the 1,291 probesets that reliably and stably detect individual differences in gene expression.

## Discussion

When assessed at a single time point, GWE differs between individuals. Some of these differences are due to transient intra-individual differences in gene expression — that is, variance at time 1 that does not covary with variance at time 2. Transitory differences in GWE are expected because gene expression is labile and state specific. However, the usefulness of GWE for investigating individual differences — such as the genetics of GWE or the relationship between GWE and individual differences in outcomes measures — depends on reliable individual differences.

Large samples are needed for individual differences research due to small expected effect sizes. Assessing genomewide gene expression at a single time point in uncontrolled circumstances is most practical for individual differences research using large samples,

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

even though these are the most difficult conditions for detecting individual differences reliably. Using our microarray platform and tissue in our sample of 12-year-olds, we have shown that 26.2% of the 13,007 probesets that detect individual differences above background noise reach our criterion of 0.70 for reliability detecting individual differences across four hours, which we had 80% power to detect. Furthermore, 51.3% of these probesets reliably detect individual differences across four hours on a second occasion, 10 months later, of which 73.7% stably detect individual differences in gene expression across 10 months.

This represents 1,291 reliable and stable probesets that can be used in GWE analyses of individual differences with a single collection of peripheral blood.

Researchers interested in the causes and correlates of individual differences in GWE will profit from focusing on these probesets that detect individual differences reliably and stably. For example, our results indicate that heritability estimates are much higher for the 3,414 'variable, present and reliable' probesets — 60% on average — as compared to 34% for all the 'present' probesets. Moreover, heritability is greatest (76%) for those probesets that stably detect individual differences over 10 months. Previous studies of heritability of GWE did not take reliability or stability into account and report heritability estimates of about .30 (Cheung et al., 2003; Dixon et al., 2007; Emilsson et al., 2008; Goring et al., 2007; Morley et al., 2004; Stranger et al., 2007). We predict that these studies would yield much higher heritability estimates for probesets that show 'reliably stable' individual differences. It should be noted that our estimate of heritability is an upper-limit based on the MZ correlation alone which could be inflated by shared environmental influences (Plomin et al., 2008). Much larger studies of both MZ and DZ twins are needed to provide more precise estimates of heritability.

Although the small sample size is a limitation of our study, a sample size of 10 provides 80% power to detect correlations of 0.70 ($p$ = .05, one-tailed). In the field of psychometry, test–retest reliability of 0.70 is traditionally viewed as an acceptable level of reliability.

As mentioned, a limitation of our study is the relatively uncontrolled circumstances of blood collection. Reliability might be increased by controlling for, or accounting for, variables such as distance traveled to the laboratory, health, amount of sleep, food intake, hormonal influence and mood. As such, our results represent a 'lower-limit' of reliable and stable individual differences. It would be possible to increase the number of probesets that detect reliable and stable individual differences by obtaining blood on multiple measurement occasions. However, because the average reliability is only 0.338 for the 'present' probesets, many repeated measurements would be required to reach reliability for even 50% of the probesets.

In addition to this specific limitation, we recognize the general limitations of high-throughput gene expression as assessed by microarrays such as platform differences, difficulties in detecting low abundance genes, and sensitivity and specificity (Draghici et al., 2006; Wang et al., 2006). Another possible limitation of our study is its use of whole blood rather than specific cell types such as leukocytes. Our rationale for using whole blood was to avoid restricting our analyses to those transcripts expressed in a particular cell type and to avoid systematic effects on transcripts during the invasive process of extracting lymphocytes from blood. As a control measure, we used blood cell counts to assess relative numbers of cell populations and to control for infection status (Eady et al., 2005).

We also recognize that some of the 1,291 probesets that we identify as showing reliable and stable individual differences will not necessarily be the same for other tissues, other populations, different time points, or other microarrays. In order to explore this issue further, we searched for comparable datasets of reasonable size (i.e., that have repeated measures of peripheral blood, in healthy subjects with no drug intervention and for which the CEL files are available) in GEO (Barrett et al., 2007). A recently published study using whole blood (Dusek et al., 2008) tested 21 individuals twice, separated by an eight-week period. In our reanalysis of these data, after quality control exclusions, 434 (33.6%) of the 1,291 probesets that we identify as reliable and stable are also reliable in the Dusek et al. study, demonstrating that at least a proportion of the probesets we identify as showing reliable and stable individual differences are confirmed across samples and laboratories and can be used in research on the causes and correlates of individual differences in gene expression.

Our samples were closely matched for age but were of mixed sex. However, of the 775 known genes represented by the 1,291 probesets, one is located on the Y chromosome and as the individual difference here is sex, this probeset should be excluded. It is also likely that some of the probesets located on the X chromosome escape X-inactivation and so should also be excluded. In a larger sample a linear mixed effects model could be performed to account for factors such as age and sex. No obvious functional or biological themes were apparent in the 775 genes, but we might predict that they represent genes that can tolerate a large degree of stable variation in expression levels without grossly affecting behavior or physiology.

## Conclusions

From genome-wide gene expression arrays we have identified probesets whose individual differences are reliable over 4 hours and stable over 10 months. Although the proportion of transcripts in which we see reliable and stable individual differences is modest, the result is that there are at least a thousand such transcripts expressed in blood. Use of these transcripts

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

is likely to improve the results of studies of the causes and correlates of gene expression. Transcripts whose expression is not stable over time are of course potentially interesting in other ways, especially as an index of environmental effects.

## Acknowledgments

The Twins Early Development Study (TEDS) is supported by a program grant from the UK Medical Research Council (Grant no. G0500079) and the US National Institute of Child Health and Human Development (Grant no. HD49861). The study was approved by the Institute of Psychiatry/South London and Maudsley Research Ethics Committee (CREC/06/07-5) and informed consent was obtained from the parents of the children.

## References

Bakay, M., Chen, Y. W., Borup, R., Zhao, P., Nagaraju, K., & Hoffman, E. P. (2002). Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics, 3*, 4.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., & Edgar, R. (2007). NCBI GEO: Mining tens of millions of expression profiles — Database and tools update. *Nucleic Acids Research, 35*, D760–D765.

Breitling, R., Li, Y., Tesson, B. M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L. V., de, H. G., Su, A. I., & Jansen, R. C. (2008). Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics, 4*, e1000232.

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., & Lowry, S. F. (2005). A network-based analysis of systemic inflammation in humans. *Nature, 437,* 1032–1037.

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

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., & Cookson, W. O. (2007). A genome-wide association study of global gene expression. *Nature Genetics, 39*, 1202–1207.

Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics, 22*, 101–109.

Dumur, C. I., Nasim, S., Best, A. M., Archer, K. J., Ladd, A. C., Mas, V. R., Wilkinson, D. S., Garrett, C. T., & Ferreira-Gonzalez, A. (2004). Evaluation of quality-control criteria for microarray gene expression analysis. *Clinical Chemistry, 50*, 1994–2002.

Dusek, J. A., Otu, H. H., Wohlhueter, A. L., Bhasin, M., Zerbini, L. F., Joseph, M. G., Benson, H., & Libermann, T. A. (2008). Genomic counter-stress changes induced by the relaxation response. *PLoS ONE, 3,* e2576.

Eady, J. J., Wortley, G. M., Wormstone, Y. M., Hughes, J. C., Astley, S. B., Foxall, R. J., Doleman, J. F., & Elliott, R. M. (2005). Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. *Physiological Genomics, 22,* 402–411.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G. H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K. P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H. G., Stefansson, T., Leifsson, B. G., Thorsteinsdottir, U., Lamb, J. R., Gulcher, J. R., Reitman, M. L., Kong, A., Schadt, E. E., & Stefansson, K. (2008). Genetics of gene expression and its effect on disease. *Nature, 452*, 423–428.

Fernandes, C., Paya-Cano, J. L., Sluyter, F., D'Souza, U., Plomin, R., & Schalkwyk, L. C. (2004). Hippocampal gene expression profiling across eight mouse inbred strains: Towards understanding the molecular basis for behaviour. *European Journal of Neuroscience, 19,* 2576–2582.

Gladkevich, A., Kauffman, H. F., & Korf, J. (2004). Lymphocytes as a neural probe: Potential for studying psychiatric disorders. *Progress in Neuropsychopharmacol Biological Psychiatry, 28, 559–576.*

Goring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., Jowett, J. B., Abraham, L. J., Rainwater, D. L., Comuzzie, A. G., Mahaney, M. C., Almasy, L., MacCluer, J. W., Kissebah, A. H., Collier, G. R., Moses, E. K., & Blangero, J. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics, 39*, 1208–1216.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research, 31*, e15.

Konradi, C. (2005). Gene expression microarray studies in polygenic psychiatric disorders: Applications and data analysis. *Brain Research Review, 50*, 142–155.

Korostynski, M., Kaminska-Chowaniec, D., Piechota, M., & Przewlocki, R. (2006). Gene expression profiling in the striatum of inbred mouse strains with distinct opioid-related phenotypes. *BMC Genomics, 7,* 146.

Mohr, S., & Liew, C. C. (2007). The peripheral-blood transcriptome: new insights into disease and risk assessment. *Trends in Molecular Medicine, 13*, 422–432.

Morley, 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.

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press

Nica, A. C., & Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Humun Molecular Genetics, 17*, R129–R134.

Nicholson, A. C., Unger, E. R., Mangalathu, R., Ojaniemi, H., & Vernon, S. D. (2004). Exploration of neuroendocrine and immune gene expression in peripheral blood mononuclear cells. *Molecular Brain Research, 129*, 193–197.

Oliver, B. R. & Plomin, R. (2007). Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics, 10,* 96–105.

Pahl, A. (2005). Gene expression profiling using RNA extracted from whole blood: Technologies and clinical applications. *Expert Reviews in Molecular Diagnosis, 5*, 43–52.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). New York: Worth Publishers.

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.

Robinson, M. D., & Speed, T. P. (2007). A comparison of Affymetrix gene expression arrays. *BMC Bioinformatics, 8,* 449.

Rockman, M. V., & Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics, 7*, 862–872.

Sharp, F. R., Xu, H., Lit, L., Walker, W., Apperson, M., Gilbert, D. L., Glauser, T. A., Wong, B., Hershey, A., Liu, D. Z., Pinter, J., Zhan, X., Liu, X., & Ran, R. (2006). The future of genomic profiling of neurological diseases using blood. *Archives of Neurology, 63*, 1529–1536.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavare, S., Deloukas, P., & Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics, 39*, 1217–1224.

Wang, Y., Barbacioru, C., Hyland, F., Xiao, W., Hunkapiller, K. L., Blake, J., Chan, F., Gonzalez, C., Zhang, L., & Samaha, R. R. (2006). Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics, 7*, 59.

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 in National Academy of Sciences USA, 100,* 1896–1901.

Yuferov, V., Nielsen, D., Butelman, E., & Kreek, M. J. (2005). Microarray studies of psychostimulant-induced changes in gene expression. *Addiction Biology, 10*, 101–118.

Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research, 33*, W741–W748.

https://doi.org/10.1375/twin.12.4.372 Published online by Cambridge University Press