Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays

Anu Sharma, Vineet K. Sharma, Shirley Horn-Saban, Doron Lancet, Srinivasan Ramachandran and Samir K. Brahmachari

Physiol. Genomics 21:117-123, 2005. First published 11 January 2005;
doi:10.1152/physiolgenomics.00228.2003

Physiological Genomics publishes results of a wide variety of studies from human and from informative model systems with techniques linking genes and pathways to physiology, from prokaryotes to eukaryotes. It is published quarterly in January, April, July, and October by the American Physiological Society, 9650 Rockville Pike, Bethesda MD 20814-3991. ISSN: 1094-8341, ESSN: 1531-2267. Visit our website at http://www.the-aps.org/.

Physiol Genomics 21: 117–123, 2005.
First published January 11, 2005; doi:10.1152/physiolgenomics.00228.2003.

Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays

Anu Sharma,[1] Vineet K. Sharma,[2] Shirley Horn-Saban,[3] Doron Lancet,[4] Srinivasan Ramachandran,[1,2] and Samir K. Brahmachari[1,2]

[1]*Functional Genomics Unit and* [2]*G. N. Ramachandran Knowledge Center for Genome Informatics, Institute of Genomics and Integrative Biology, Delhi, India;* [3]*Microarray Facility, Department of Biological Services, and* [4]*Department of Molecular Genetics and Crown Human Genome Center, The Weizmann Institute of Science, Rehovot, Israel*

Submitted 29 December 2003; accepted in final form 9 January 2005

Sharma, Anu, Vineet K. Sharma, Shirley Horn-Saban, Doron Lancet, Srinivasan Ramachandran, and Samir K. Brahmachari. Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays. *Physiol Genomics* 21: 117–123, 2005. First published January 11, 2005; doi:10.1152/physiolgenomics.00228.2003.—Quantitative variation in gene expression in humans is the outcome of various factors, including differences in genetic background, gender, age, and environment. However, the extent of the influence of these factors on gene expression is not clear. We attempted to address this issue by carrying out gene expression profiling in blood leukocytes with 13 individuals (including 5 pairs of monozygotic twins) on 10,000 genes using HG-U95Av2 oligonucleotide microarrays. The proportion of differentially expressed genes between monozygotic twins was low (up to 1.76%). Most of the variations belonged to the least variable category. These genes, exhibiting "random variations," did not show clear preference to any functional class, although "signaling and communication" and "immune and related functions" generally topped the list. The extent of variation in gene expression increased in comparisons between unrelated individuals (up to 14.13%). Most of the genes (89%) exhibiting random variations in twins also varied in expression in unrelated individuals. As with twins, signaling and communication topped the list, and substantial variations were observed in all three categories: least variable, moderately variable, and most variable. An important outcome of this study was that the housekeeping genes were nearly insensitive to random variations but appeared to be more susceptible to genetic differences. However, the highly expressed housekeeping genes exhibited low variation and appeared to be insensitive to all known factors. Gene expression profiling in monozygotic twins can provide useful data for the assessment of natural variation in gene expression in humans.

GeneChip; microarrays; twins; differential gene expression; housekeeping genes

THE ROLE OF VARIATION in gene expression due to sequence polymorphisms in humans is largely unknown. Several yr ago, it was pointed out that, in humans and in their evolutionarily closely related primates, phenotypic differences could arise from quantitative differences in gene expression rather than structural changes in protein (17). However, natural variation in gene expression between healthy human individuals has been largely unexplored. Comparatively, variations in the DNA in the form of single nucleotide polymorphisms, length polymorphisms in simple sequence repeats (expansion or contraction), and insertion/deletion polymorphisms have been comprehensively studied (2). To understand the genetic basis of variation in gene expression between normal human individuals, we need to obtain genome-wide expression data from various populations.

The natural variation in gene expression is an outcome of the complex interplay of genetic polymorphisms (acting in *cis* or in *trans*), physiological variations (such as time of day and gender), and environmental factors (12). One approach to address this complexity is the use of model systems, including animals, insects, or lower eukaryotes. In these cases, conditions can be chosen to minimize the contribution of nongenetic variables. Such studies in yeast (*Saccharomyces cerevisiae*), fruitfly (*Drosophila melanogaster*), and fish (genus *Fundulus*) allowed inferences on global patterns of variation in gene expression that could be correlated to genetic differences (6, 16, 22). Although these data are very useful, it is desirable that, in parallel, estimation of natural variation in gene expression in humans be carried out directly.

Minimizing the contributions of nongenetic factors in humans is inherently difficult. Therefore, estimation of variation in gene expression due to genetic differences will have to be addressed from a different angle. Studies in monozygotic twins could enable us to estimate the size of the contribution of genetic and environmental factors to the natural variation in gene expression, because phenotypic differences within monozygotic twin pairs are due to environmental effects alone, as they uniquely share their entire genetic background (20). Therefore, differentially expressed genes between monozygotic twins can be classified as "genes whose expression varies randomly due to environmental factors."

Identification of differentially expressed genes between monozygotic twins could allow us to determine the contribution of environmental factors, if a given twin pair can be sampled at the same time. Comparison between unrelated individuals can be carried out by considering various factors such as differences in gender, age, and time of day (27) and examining the characteristics of the housekeeping genes, since these genes are expressed constitutively in all tissues to maintain cellular functions.

Here we report the gene expression analysis of five pairs of monozygotic twins and three unrelated individuals using HG-U95Av2 microarrays. Our results serve to expand the current understanding of natural variation in gene expression in humans and suggest the use of monozygotic twins for comparative analysis in these investigations.

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: S. Ramachandran, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India (E-mail: ramu@igib.res.in; ramucbt@yahoo.com).

MATERIALS AND METHODS

Volunteers, Blood Samples, and Haplotyping

Blood samples were drawn around midday in all cases to reduce the potential contribution of the variation in gene expression during different times of day. Normal healthy twin pairs were recruited for the study. Three pairs of female twins belonged to the age group 20–23 yr, and two pairs of male twins were 25 and 37 yr of age. All of the female (F) twin pairs (F1:F2, F5:F6, and F7:F8) considered for the study incidentally lived close to each other, pursued a similar kind of profession, and had similar nutrition habits at the time of sampling. In the case of the male (M) twin pairs, one of the twin pairs (M1:M2) lived far apart, in different geographical locations with very different climates (coastal-humid vs. inland-dry), and had different occupations and different nutrition habits at the time when sample was drawn. The other male twin pair (M4:M5) lived separately but in similar regions and had similar professions and nutrition.

Three more normal individuals, including two females and one male, were recruited. Their ages were 23, 34, and 37 yr, respectively. Informed consent was obtained from all. About 20 ml of blood were drawn by vein puncture and immediately processed for nucleic acid isolation. Three-quarters of the isolated blood was used for total RNA isolation, and the rest was used for isolating genomic DNA. Twelve highly polymorphic microsatellite markers located on eight different chromosomes (Linkage panel set, version 2; Perkin Elmer Applied Biosystems, Foster City, CA) were used for haplotyping of genomic DNA from twins to assess their monozygosity.

Isolation of Total RNA and Genomic DNA from Blood Leukocytes

Total RNA was isolated from blood leukocytes after the red blood cells (RBCs) were lysed in 1× RBC lysis buffer (150 mM NH_4Cl, 10 mM NaHCO, and 1 mM EDTA prepared in diethylpyrocarbonate-treated water). The blood leukocytes were recovered by centrifugation at 250 *g*, and total RNA was isolated with an EZ-RNA isolation kit (Biological Industries, Kibbutz Beth Haemek, Israel). The quality of total RNA was examined by gel electrophoresis. Samples with either DNA contamination or degradation were discarded. The genomic DNA was isolated with the salting-out procedure (21).

Preparation of cDNA and In Vitro Transcription and Labeling

The amount of RNA taken from each sample was equalized, based on absorbance at 260 nm. Double-stranded cDNA was synthesized from 8 μg of total RNA by reverse transcription, using T7-$(dT)_{24}$ primer and the Superscript Choice cDNA synthesis system (Invitrogen). In vitro transcription of the cDNA was carried out with the use of an Enzo Bioarray High Yield RNA transcript labeling kit (Affymetrix) to prepare biotin-labeled cRNA. The labeled cRNA was cleaned, using RNeasy columns (Qiagen). The labeled target was fragmented, and a hybridization cocktail was prepared including fragmented cRNA, probe array controls, BSA, and Herring sperm DNA.

GeneChip Processing

GeneChips were processed (HG-U95Av2 arrays, Affymetrix) under the same set of experimental conditions. First, labeled products were hybridized with the Affymetrix GeneChip Test3 arrays. If the results were judged satisfactory, hybridization was subsequently carried out with the HG-U95Av2 arrays as per the manufacturer's instructions. Arrays were hybridized at 45°C for 16 h. After hybridization, arrays were washed using an automated GeneChip Fluidics Station 400. After the washing, the array was stained with streptavidin-phycoerythrin and scanned with an HP Gene Array Scanner. Data analysis was carried out using Affymetrix Microarray Suite Software (MAS 5.0). All GeneChip experiments were performed at the Weizmann Institute of Science (Rehovot, Israel).

Data Analysis

The HG-U95Av2 array consists of 12,626 probe sets (including controls) for ~10,000 genes. Global scaling was carried out to reliably compare the data from multiple arrays. The raw data from the GeneChip experiments have been submitted to Gene Expression Omnibus (GEO; **http://www.ncbi.nlm.nih.gov/geo**) under the following accession numbers: GSM14477, GSM14478, GSM14479, GSM14480, GSM14481, GSM14482, GSM14483, GSM14485, GSM20645, GSM29053, GSM29054, GSM29055, GSM29056, GSM29057, and GSM29058.

Comparative analyses were performed by considering probe sets with "present" (P) call with a *P* value of 0–0.04. Our goal was to identify differentially expressed genes above experimental noise. First, we compared duplicate experiments using the same RNA sample to obtain a cutoff limit of signal log ratio to identify differentially expressed genes above experimental noise. We observed that none of the ~10,000 genes was differentially expressed at a signal log ratio >1.585 in duplicate experiments. Therefore, the differentially expressed genes in pairwise comparisons were identified by selecting the probe sets with "change" call "I" or "D" and a signal log ratio >1.585.

Functional Classification of Differentially Expressed Genes

To examine the correlation of functional classification of genes with their variability in expression, we first categorized the differentially expressed genes into three categories: least variable (absolute signal log ratio value: 1.6–2.3), moderately variable (absolute signal log ratio value: 2.3–3), and most variable (absolute signal log ratio value: >3). Subsequently, they were classified according to function into six functional classes, based on the scheme described by Adams et al. (1) and Hsiao et al. (14). The genes belonging to replication, transcription, and translation have been collectively grouped into the "information" class, as described by Andrade et al. (4). We show (see Figs. 3 and 5, vertical bars) the number of genes exhibiting fold change variation in the three categories and in each functional class: information (IN; includes replication, transcription, and translation), "signaling and communication" (SC), "immune and related functions" (IR), "metabolic processes" (MP), "cell cycle" (CC), and "structure and motility" (SM). Signaling and communication include receptors, protein modification, hormone/growth factors, intracellular transducers, effectors/modulators, metabolism, cell adhesion, and channels/transport proteins. Information includes protein synthesis, translation factors, ribosomal proteins, posttranslational modification/targeting, protein degradation, tRNA synthesis/metabolism, RNA synthesis, transcription factors, RNA polymerase, RNA processing, RNA degradation, DNA synthesis/replication, and DNA repair. Metabolic processes include amino acids, nucleotides, sugars, lipids, cofactors, protein modification, energy, and carrier proteins/membrane transport. Cell cycle/cell division includes cell cycle, apoptosis, chromosomal structure, and DNA repair. Structure and motility include cytoskeletal, microtubule-associated proteins/motors, and extracellular matrix. Immune and related functions include immunology, homeostasis, and carrier proteins/membrane transport and stress response.

NetAffx (version dated 23 June 2004, **http://www.affymetrix.com**) was mainly used for annotation and functional classification of the differentially expressed genes (19). Supplementary information was obtained from GeneCards (**http://bioinformatics.weizmann.ac.il/cards**) (23) and LocusLink (**http://www.ncbi.nlm.nih.gov/LocusLink**).

Housekeeping Genes

The reference data set of 575 housekeeping genes compiled by Eisenberg and Levanon (10) was used for comparative analysis (**http://www.compugen.co.il/supp_info/Housekeeping_genes.html**) (25). A total of 475 housekeeping genes were identified as meeting the

criteria of P call in at least 9 of 13 arrays (70%). Expression patterns of housekeeping genes were examined by computing their mean expression and coefficient of variation (CV), as suggested previously (14). Mean expression was computed by logarithmic transformation (base 10) of the signal values from all 13 arrays. Probe sets without P calls were not considered. The CV was computed as standard deviation (SD)/mean.

Statistical Analysis

Preferences in distribution of the differentially expressed genes in different functional classes for each category of variation (least, moderate, and most variable) were tested, using the chi-square (χ^2) test. To compute expected occurrence, the total number of genes was equally distributed in each of the six functional classes. Equal occurrence of genes in the different functional classes is expected when variation in gene expression occurs solely because of random fluctuations. A statistical test was carried out only for those cases where substantially high numbers of genes varied in expression.

RESULTS

Twins

Confirmation of monozygosity. All five twin pairs had identical alleles for the 12 repeat markers of high heterozygosity index. Because the probability of monozygosity is >99.9% when more than five highly polymorphic markers have identical size distribution within a twin pair (3), our data confirm that the five pairs of twins are monozygotic.

Differentially Expressed Genes

Female twin pairs. The scatter plots of the gene expression levels measured by "signal" values for female twin pairs are shown in Fig. 1, *A–C*. It is evident that, in all cases, gene expression is highly similar between monozygotic twins. None of the genes was observed to be differentially expressed in the pair F1:F2. The number of differentially expressed genes was 19 in the pair F5:F6 and 24 in the pair F7:F8. The majority of the differentially expressed genes belonged to the least variable category: 15/19 in the pair F5:F6 and 14/24 in the pair F7:F8. Functional classification revealed that IR topped the list (28%) of differentially expressed genes in F5:F6, whereas SC topped the list (39%) in the pair F7:F8. The distribution in other classes was nearly equal in both pairs. The only exception was the low representation of SM function in the pair F7:F8, at 5%.

Male twin pairs. The scatter plots of gene expression levels in the twin pairs M1:M2 and M4:M5 are displayed in Fig. 2, *A* and *B*. Compared with the female twins, the points are more widespread in male twins. A sum of 176 genes was differentially expressed in the pair M1:M2, whereas 18 genes were differentially expressed in the pair M4:M5. As in the case of female twins, the majority of the differentially expressed genes belonged to the least variable category (108/176 in M1:M2 and 10/18 in M4:M5). Functional classification uncovered that SC and IR were nearly equally represented at 29 and 27%, respectively, in the pair M1:M2 (Fig. 3). The distribution in other classes followed the order IN (18%), MP (13.6%), CC (6.5%), and SM (5.2%). In the pair M4:M5, IR topped the list at 28.6%. The distribution in the classes IN, MP, and SM were equal at 21.4%. SC was lowly represented at 7%, and none of the genes of CC was differentially expressed.

Because the number of differentially expressed genes in pair M1:M2 was high compared with other pairs, we assessed the

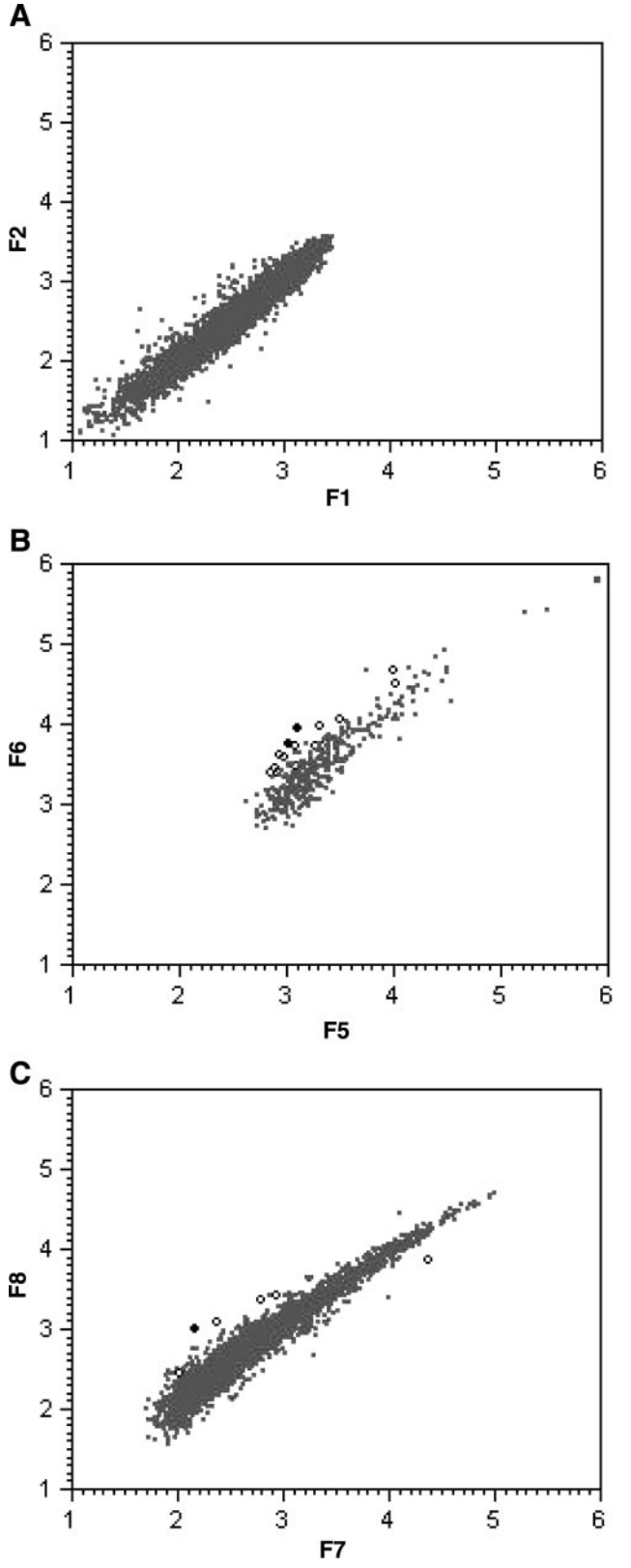

Fig. 1. Gene expression pattern between female (F) monozygotic twin pairs. Log_{10}(signal) values are plotted. Genes not differentially expressed are shaded light gray. Differentially expressed genes are classified into 3 categories: most variable (>3 signal log ratio, x), moderately variable (2.3–3 signal log ratio, ●), and least variable (1.6–2.3 signal log ratio, ○). *A*: *x*-axis, F1; *y*-axis, F2. *B*: *x*-axis, F5; *y*-axis, F6. *C*: *x*-axis, F7; *y*-axis, F8.

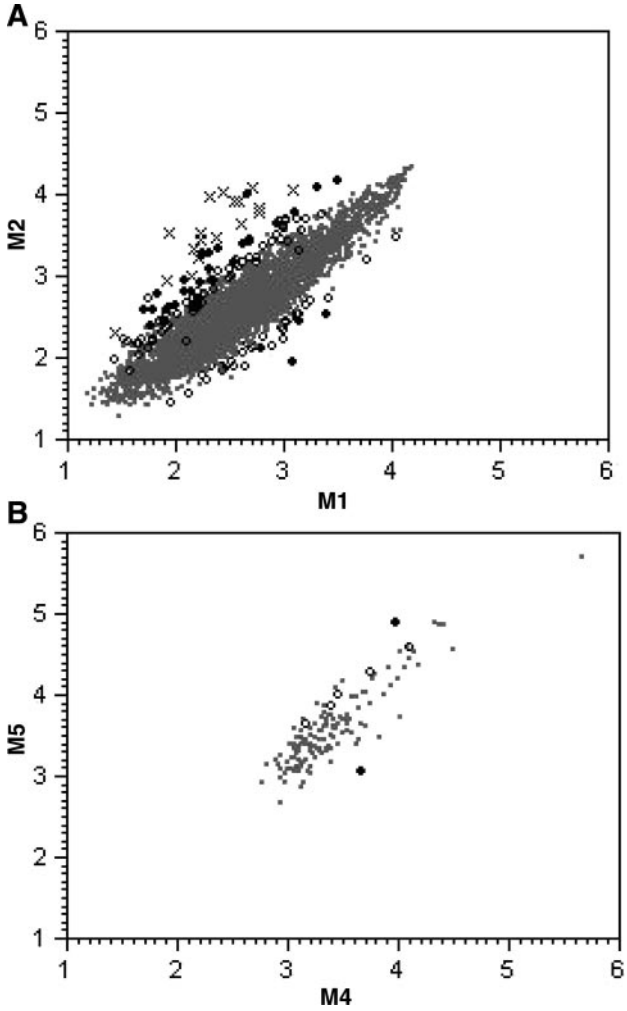

Fig. 2. Gene expression pattern between male (M) monozygotic twins. Log_{10}(signal) values are plotted. For symbol descriptions of the 3 categories of differentially expressed genes, see legend to Fig. 1. *A*: *x*-axis, M1; *y*-axis, M2. *B*: *x*-axis, M4; *y*-axis, M5.

statistical significance of the differences in the representation in six functional classes. It was apparent that the deviation from equal representation was statistically significant in all three categories of variation (most variable, $P < 0.0001$; moderately variable, $P < 0.001$; and least variable, $P < 0.0001$). It was interesting to note the presence of several genes belonging to IR functions grouping into the most variable category.

Overall, differentially expressed genes between monozygotic twins was low (0–1.76%), and the majority of them belonged to the least variable category in all pairs. In general, there appears to be no clear preference for any of the functional classes, although genes of SC and IR classes generally tend to top the list of differentially expressed genes. A sum of 214 genes (nonredundant set) was differentially expressed in all pairwise comparisons of monozygotic twins.

Housekeeping genes. In the backdrop of differences in gene expression, analysis of the expression patterns in housekeeping genes is an important step to characterize differentially expressed genes. The number of differentially expressed housekeeping genes between monozygotic twins was very low. The results are displayed in Table 1. No clear preference to any of the functional classes was observed among the differentially expressed housekeeping genes. These observations mirror the global pattern of distribution of differentially expressed genes between monozygotic twins.

Interestingly, these observations suggest that the housekeeping genes are generally not susceptible to random variations in expression due to environmental factors. Furthermore, we observed that none of the housekeeping genes coding for basal transcriptional machinery, ribosomal proteins, and DNA replication was found to belong to the most variable category between the twins.

Comparisons Between Unrelated Individuals

Differentially expressed genes among unrelated individuals. To further elaborate on the influence of genetic and environmental factors on gene expression, we carried out comparative gene expression analysis between unrelated individuals of the same gender and similar age to minimize the contribution of other factors. A total of eighteen pairs of comparisons between seven unrelated female individuals of similar age were carried out meeting these criteria (Fig. 4). The number of differentially expressed genes in the pairs ranged from 37 to 1,413, corresponding to an extent of variation from 0.37 to 14.13%. This range is higher than that observed between monozygotic twins. The total number of these genes in all 18 pairs was 3,057. These genes were distributed as 46% in least variable, 31% in moderately variable, and 23% in most variable categories. This distribution differs from the pattern between monozygotic twins, wherein we observed that a majority of the differentially expressed genes belonged to the least variable category. These observations indicate that the variability in the expression of genes increases with genetic distance.

Interestingly, 191 of the 214 genes (89%) differentially expressed between monozygotic twins also varied in expres-

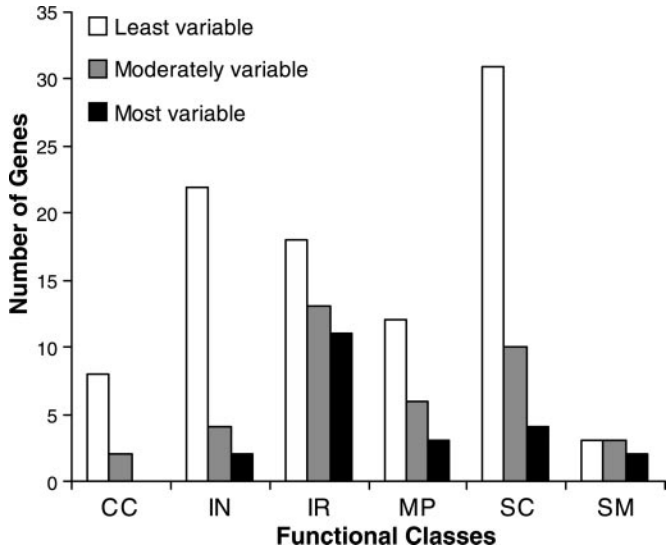

Fig. 3. Distribution of differentially expressed genes in the 6 functional classes between the male twin pair M1:M2. Note the presence of several genes belonging to "immune and related functions" (IR) grouping into the most variable category. CC, cell cycle; IN, information; MP, metabolic processes; SC, signaling and communication; SM, structure and motility.

Table 1. *Differentially expressed housekeeping genes between monozygotic twins*

Pair	No. of Differentially Expressed Genes	Gene Symbol (Function)*	Functional Class
F1:F2	0		—
F5:F6	5	*EEF1A1* (eukaryotic translation elongation factor-1, alpha-1)	Information
		B2M (beta-2-microglobulin)	Immune and related functions
		TALDO1 (transaldolase-1)	Metabolic processes
		ACTG1 (actin, gamma-1)	Structure and motility
		ACTB (actin, beta)	Structure and motility
F7:F8	2	*YWHAZ* (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide)	Metabolic processes
		RTN4 (reticulon-4)	Signaling and communication
M1:M2	4	*EEF1D* (eukaryotic translation elongation factor-1, delta)	Information
		ATP51 (ATP synthase, H^+ transporting, mitochondrial F0 complex, subunit e)	Metabolic processes
		JUND (jun D protooncogene)	Signaling and communication
		LAMP1 (lysosomal-associated membrane protein-1)	Unknown function
M4:M5	9	*SNRP70* (small nuclear ribonucleoprotein 70-kDa polypeptide)	Information
		HNRPH1 (heterogeneous nuclear ribonucleoprotein H1)	Immune and related functions
		FCGR2A (Fc fragment of IgG, low affinity IIa, receptor for CD32)	Immune and related functions
		MT3{metallothionein-3 [growth inhibitory factor (neurotrophic)]}	Metabolic processes
		GM2A (GM2 ganglioside activator)	Metabolic processes
		YWHAZ (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide)	Metabolic processes
		NXF1 (nuclear RNA export factor-1)	Signaling and communication
		RAB8A (RAB8A, member RAS oncogene family)	Signaling and communication
		KIAA0515	Unknown function

F, female; M, male. *Gene symbols are according to the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) nomenclature.

sion when compared between unrelated individuals. This observation supports the prediction that random variations in gene expression due to environmental factors tend to vary with no apparent relationship to genetic distances between individuals in human populations.

The distribution of 3,057 differentially expressed genes from 18 pairs in the six functional classes is shown in Fig. 5. The top ranking class was SC (31%), followed by IN (24%), MP (20%), IR (12%), CC (7%), and SM (6%). It was apparent that the deviation from equal representation was statistically significant in all three categories of variation (most variable, $P < 0.0001$; moderately variable, $P < 0.0001$; and least variable, $P < 0.0001$; χ^2 test).

Housekeeping genes. The number of differentially expressed housekeeping genes between unrelated females varying in genetic relationship and environment was in the range 0–159, summing to 303 nonredundant entries. This observation suggests that the proportion of differentially expressed housekeeping genes increases with genetic distance. These genes span a wide range of functional classes, including SC (86), MP (71), IN (67), SM (21), IR (19), and CC (12). Twenty-seven genes were of unknown function. The number of differentially expressed housekeeping genes increased further to 351 when our comparisons included difference in age and gender.

In the backdrop of housekeeping genes varying in expression, an important goal is to identify the most highly expressed housekeeping genes. We ranked them according to their mean

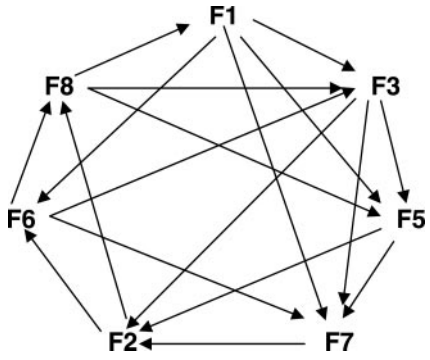

Fig. 4. Schematic representation of pairwise comparisons between unrelated female individuals of similar age. F3 is an unrelated singleton individual of the same age group. Eighteen pairwise comparisons were considered. In this scheme, the following comparisons between related individuals (F1:F2, F5:F6, and F7:F8) were excluded.

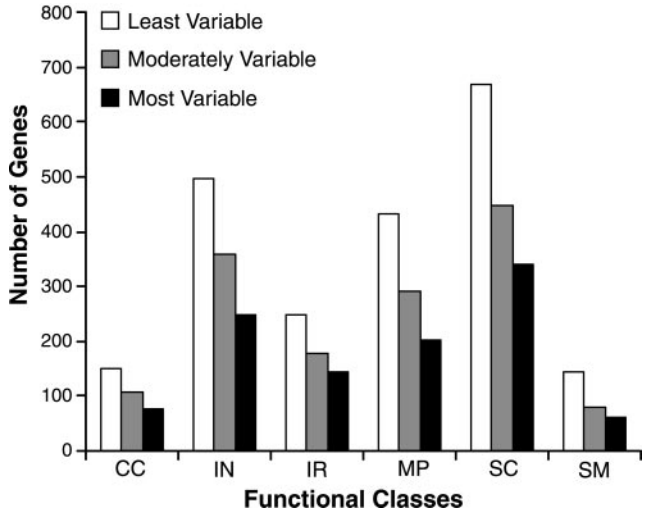

Fig. 5. Distribution of differentially expressed genes in the 6 functional classes among the unrelated female individuals of similar age.

Table 2. *Fifteen most highly expressed housekeeping genes*

Gene Symbol*	Description	CV†	Mean†
HLA-C	major histocompatibility complex, class I, C	0.07	26,607
RPL13A	ribosomal protein L13A	0.15	7,926
RPL13	ribosomal protein L13	0.11	7,294
RPS12	ribosomal protein S12	0.13	5,461
RPL8	ribosomal protein L8	0.10	5,272
MYL6	myosin, light polypeptide 6, alkali, smooth muscle and nonmuscle	0.26	5,155
FAU	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV); ribosomal protein S30	0.12	5,017
RPL38	ribosomal protein L38	0.12	4,660
RPS5	ribosomal protein S5	0.10	4,656
ATP51	ATP synthase, H^+ transporting, mitochondrial F0 complex, subunit e	0.12	4,634
RPLP1	ribosomal protein, large, P1	0.17	4,533
UBC	ubiquitin C	0.06	4,422
RPL37	ribosomal protein L37	0.06	4,419
JUND	jun D protooncogene	0.09	4,350
RPS19	ribosomal protein S19	0.13	4,310

Genes are ranked by decreasing order of mean expression values. *HGNC gene symbols are shown. †Mean expression levels and coefficients of variation (CV; SD/mean) across the array experiments in which these genes were detected as "present" were computed as described in MATERIALS AND METHODS.

expression levels in our experiments. The top 15 highly expressed housekeeping genes are listed in Table 2. It is evident that several of these highly expressed genes (9/15) are ribosomal protein coding genes that carry out important cellular functions. It is also interesting to note that the CV in expression across individuals varying in genetic background, age, gender, and environment is low among the highly expressed housekeeping genes.

DISCUSSION

Natural variation in human gene expression has begun to be explored only recently (8, 11, 27). Peripheral blood leukocytes are a readily accessible source of cells to investigate the natural variation in gene expression in humans. However, this tissue consists of a diverse population of cell types such as neutrophils, eosinophils, basophils, monocytes, and lymphocytes. The use of monozygotic twins in studying natural variation in gene expression offers a unique advantage, since they share an identical genetic background. Therefore, studies in monozygotic twins offer the possibility of functional dissection of the influence of genetic and environmental factors (20).

Overall, we found very low variation in gene expression between monozygotic twins (0–1.76%). The high variation observed in gene expression between the monozygotic pair M1:M2 could be attributed to significant differences in the environment to which they were exposed. These differences were comprised of diverse climates, nutrition habits, and professions. Compared with this pair, the other twin pairs either lived closely or lived in similar geographical locations and generally had similar nutrition habits and professions.

We also observed that most of the differentially expressed genes between monozygotic twins belonged to the least variable category. Furthermore, we noted that there was no clear preference for these genes to belong to any of the six functional classes, although the genes belonging to SC and IR tended to top the list. Thus random variation in gene expression due to environmental factors is more likely to be found among the genes belonging to SC and IR classes. This is perhaps due to the characteristic role of these genes to function at the interface between body and environment.

Examination of the expression of housekeeping genes between monozygotic twins indicated very low variation. Because housekeeping genes carry out essential functions for the maintenance of cellular physiology, it appears that environmental differences only play a minor role when the underlying genetic background is identical. None of the genes coding for basal transcription machinery, ribosomal proteins, and DNA replication was found to be highly variable in expression between monozygotic twins. Perhaps this is due to the generally observed high level of sequence conservation and the ancient characteristics of these genes (24, 25).

Compared with monozygotic twins, the variation in gene expression between unrelated individuals of the same gender and similar age exhibited a higher range. Furthermore, the substantial representation of differentially expressed genes that was observed in all three categories of variation was distinctly different from that observed between monozygotic twins. Our results are in agreement with independent observations made by Cheung et al. (8), who observed that genes showed less variability in expression between closely related individuals compared with unrelated individuals. Taken together, it appears that differences in genetic background are primary contributors to variation in gene expression in humans, while environmental effects may play a minor role. Because genes belonging to SC and IR functions tend to top the list between unrelated individuals, similar to monozygotic twins, it appears that SC and IR genes are highly sensitive to genetic and environmental differences.

The number of housekeeping genes differing in expression between unrelated individuals was severalfold higher compared with monozygotic twins, indicating that differences in genetic background contribute substantially to this variability. However, the highly expressed housekeeping genes showed very low variation with apparent independence with respect to differences in genetics, environment, gender, and age. These results uphold the observations by Hsiao et al. (14). In summary, our study, although subject to the characteristics of experimental signal-to-noise ratio specific to GeneChip experiments, indicates that gene expression profiling in monozygotic twins could be very useful to identify genes the expression of which varies randomly with environmental factors, and this data can be used to assess natural variations in gene expression.

A data set of these genes across different populations could be used as a sieve to identify genes the expression of which primarily varies due to genetic differences in humans. Although our study is somewhat limited due to a small sample size, we envisage that similar studies conducted in other populations could define the extent and nature of normal variability in gene expression and provide insights to understand the genetic basis of the differences between individuals in a population.

ACKNOWLEDGMENTS

We thank the Twins Today Trust (Chennai, India) for help in providing access to twin samples, Vani Brahmachari and Mitali Mukerji for invaluable discussions, and Dr. K. K. Taneja for assistance.

GRANTS

A. Sharma and V. K. Sharma are recipients of a fellowship from the Council of Scientific and Industrial Research. We thank the Department of Biotechnology, Government of India, and the Ministry of Science, Israel, for a grant under Indo-Israel cooperation.

REFERENCES

1. **Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, et al.** Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377: 3–174, 1995.
2. **Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, and Lander ES.** An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407: 513–516, 2000.
3. **Amann ST, Gates LK, Aston CE, Pandya A, and Whitcomb DC.** Expression and penetrance of the hereditary pancreatitis phenotype in monozygotic twins. *Gut* 48: 542–547, 2001.
4. **Andrade MA, Ouzounis C, Sander C, Tamames J, and Valencia A.** Functional classes in the three domains of life. *J Mol Evol* 49: 551–557, 1999.
5. **Bortoluzzi S, D'Alessi F, Romualdi C, and Danieli GA.** Differential expression of genes coding for ribosomal proteins in different human tissues. *Bioinformatics* 17: 1152–1157, 2000.
6. **Brem RB, Yvert G, Clinton R, and Kruglyak L.** Genetic dissection of transcriptional regulation in budding yeast. *Science* 296: 752–755, 2002.
7. **Bustin MA.** Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 25: 169–193, 2000.
8. **Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, and Spielman RS.** Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* 33:422–425, 2003.
9. **Chin KV and Kong AN.** Application of DNA microarrays in pharmacogenomics and toxicogenomics. *Pharm Res* 19: 1773–1778, 2002.
10. **Eisenberg E and Levanon EY.** Human housekeeping genes are compact. *Trends Genet* 19: 362–365, 2003.
11. **Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, and Paabo S.** Intra- and inter-specific variation in primate gene expression patterns. *Science* 296: 340–343, 2002.
12. **Hamilton BA.** Variations in abundance: genome-wide responses to genetic variation and vice versa. *Genome Biol* 3: 1029.1–1029.3, 2002.
13. **Haverty PM, Weng Z, Best NL, Auerbach KR, Hsiao LL, Jensen RV, and Gullans SR.** HugeIndex: a database with visualization tools for high-density oligonucleotide array data for normal human tissues. *Nucleic Acids Res* 30: 214–217, 2002.
14. **Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, and Gullans SR.** A compendium of gene expression in normal human tissues. *Physiol Genomics* 7: 97–104, 2001.
15. **Huang SH, Triche T, and Jong AY.** Infectomics: genomics and proteomics of microbial infections. *Funct Integr Genomics* 1: 331–344, 2002.
16. **Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, and Gibson G.** The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat Genet* 29: 389–395, 2001.
17. **King MC and Wilson AC.** Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116, 1975.
18. **Lercher MJ, Urrutia AO, and Hurst LD.** Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* 31: 180–183, 2002.
19. **Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, and Siani-Rose MA.** NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* 31: 82–86, 2003.
20. **MacGregor AJ, Snieder H, Schork NJ, and Spector TD.** Twins. Novel uses to study complex traits and genetic diseases. *Trends Genet* 16: 131–134, 2000.
21. **Miller SA, Dykes DD, and Polesky HF.** A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 16: 1215, 1988.
22. **Oleksiak MF, Churchill GA, and Crawford DL.** Variation in gene expression within and among natural populations. *Nat Genet* 32: 261–266, 2002.
23. **Rebhan M, Chalifa-Caspi V, Prilusky J, and Lancet D.** GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 14: 656–664, 1998.
24. **Rivera MC, Jain R, Moore JE, and Lake JA.** Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 95: 6239–6244, 1998.
25. **Sharma VK, Rao CB, Sharma A, Brahmachari SK, and Ramachandran S.** $(TG/CA)_n$ repeats in human housekeeping genes. *J Biomol Struct Dyn* 21: 303–310, 2003.
26. **Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, and Hogenesch JB.** Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA* 99: 4465–4470, 2002.
27. **Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, and Brown PO.** Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci USA* 100: 1896–1901, 2003.