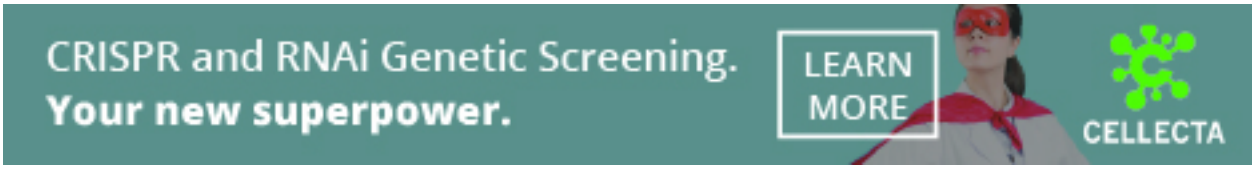
Allelic Variation in Gene Expression Is Common in the Human Genome

H. Shuen Lo, Zhining Wang, Ying Hu, et al.

Genome Res. 2003 13: 1855-1862
Access the most recent version at doi:10.1101/gr.1006603



Letter

Allelic Variation in Gene Expression Is Common in the Human Genome

H. Shuen Lo, Zhining Wang, Ying Hu, Howard H. Yang, Sheryl Gere, Kenneth H. Buetow, and Maxwell P. Lee[1]

Laboratory of Population Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA

Variations in gene sequence and expression underlie much of human variability. Despite the known biological roles of differential allelic gene expression resulting from X-chromosome inactivation and genomic imprinting, a large-scale analysis of allelic gene expression in human is lacking. We examined allele-specific gene expression of 1063 transcribed single-nucleotide polymorphisms (SNPs) by using Affymetrix HuSNP oligo arrays. Among the 602 genes that were heterozygous and expressed in kidney or liver tissues from seven individuals, 326 (54%) showed preferential expression of one allele in at least one individual, and 170 of those showed greater than fourfold difference between the two alleles. The allelic variation has been confirmed by real-time quantitative PCR experiments. Some of these 170 genes are known to be imprinted, such as *SNRPN*, *IPW*, *HTR2A*, and *PEG3*. Most of the differentially expressed genes are not in known imprinting domains but instead are distributed throughout the genome. Our studies demonstrate that variation of gene expression between alleles is common, and this variation may contribute to human variability.

[Supplemental material is available online at www.genome.org.]

Polymorphism and variations in gene expression provide the genetic basis for human variation. Mendelian inheritance assumes that genes from maternal and paternal chromosomes contribute equally to human development. X-chromosome inactivation silences gene expression from one of the two X chromosomes, thus providing an exception to mendelian inheritance (Gartler and Goldman 2001). In addition, ~50 human autosomal genes are known to be imprinted and thus are expressed from only one chromosome (Tycko and Morison 2002). However, it is unknown whether variations in allelic gene expression affect only the X chromosome and imprinted genes or whether it affects human genes generally. Recently, a group from Johns Hopkins University reported that six out of 13 genes showed significant difference in gene expression between the two alleles, and this variation in allelic gene expression was transmitted by mendelian inheritance (Yan et al. 2002b). Furthermore, they had previously shown that the allelic variation in the APC gene expression plays a critical role in colon cancer (Yan et al. 2002a). To address the issue of whether allelic variation in gene expression is a widespread phenomenon, we modified an existing genotyping technology, the Affymetrix HuSNP chip system, to analyze allele-specific gene expression.

The HuSNP chip was designed for simultaneous typing of 1494 SNPs of the human genome. It has been successfully applied to study loss of heterozygosity in lung cancer (Lindblad-Toh et al. 2000). The HuSNP chip contains 16 probes for each SNP locus (see Methods), with four matching perfectly to allele A and four matching perfectly to allele B. The other eight probes are identical to the first eight but with each having one mismatched base in the center of the probe. In this report, we performed both genotyping and allele-specific gene expression by using HuSNP chips. Our result shows that the HuSNP chip system is a reliable way to simultaneously measure allele-specific gene expression for hundreds of genes.

[1]Corresponding author.
E-MAIL leemax@mail.nih.gov; FAX (301) 402-9325.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1006603.

RESULTS

HuSNP Chips Can Reliably Analyze Allele-Specific Gene Expression

To measure allele-specific gene expression quantitatively, we first needed to find out (1) which of the 1494 SNPs on the chip are located in a transcribed region, and (2) whether the system can measure allele-specific expression accurately. By using BLAST searches and annotations in dbSNP, we found that 1063 SNPs are located in transcribed regions. To address the second issue, we developed a computational method to extract the fluorescent intensity for each probe and to quantify the ratio of expression of the two alleles. To assess the precision of the system, we performed experiments in duplicates for both genomic DNA and for cDNA derived from polyA RNA from three fetuses. As shown in Figure 1A, the correlation between the repeated experiments was very high, with Pearson correlation coefficients of 0.98 for genomic DNA and 0.95 for RNA. The *P* values for both correlation coefficients were <0.0001. Our analysis indicated that we could reliably identify differences between the expression of two alleles that differ by greater than twofold (for details, see Methods).

Altogether, we performed genotyping and allele-specific gene expression in kidney and liver for seven fetuses. Genotype calls were obtained by using the Affymetrix MAS 4.0 software, and quantitative allele-specific gene expression was obtained as described in the Methods. The average call rate for genomic DNAs from seven individuals was 71% of the 1494 SNPs on the chip. To be included in our analysis, each SNP had to meet the following criteria: (1) At least one fetus is heterozygous for the SNP; (2) the SNP is among the 1063 mapped within a transcribed region, and (3) the gene containing the SNP is expressed in kidney or liver. We found that 602 SNPs met all three criteria. RNA from kidney and liver of each individual was used to synthesize cDNA, which was hybridized to HuSNP chips. We computed the relative expression of the two alleles.

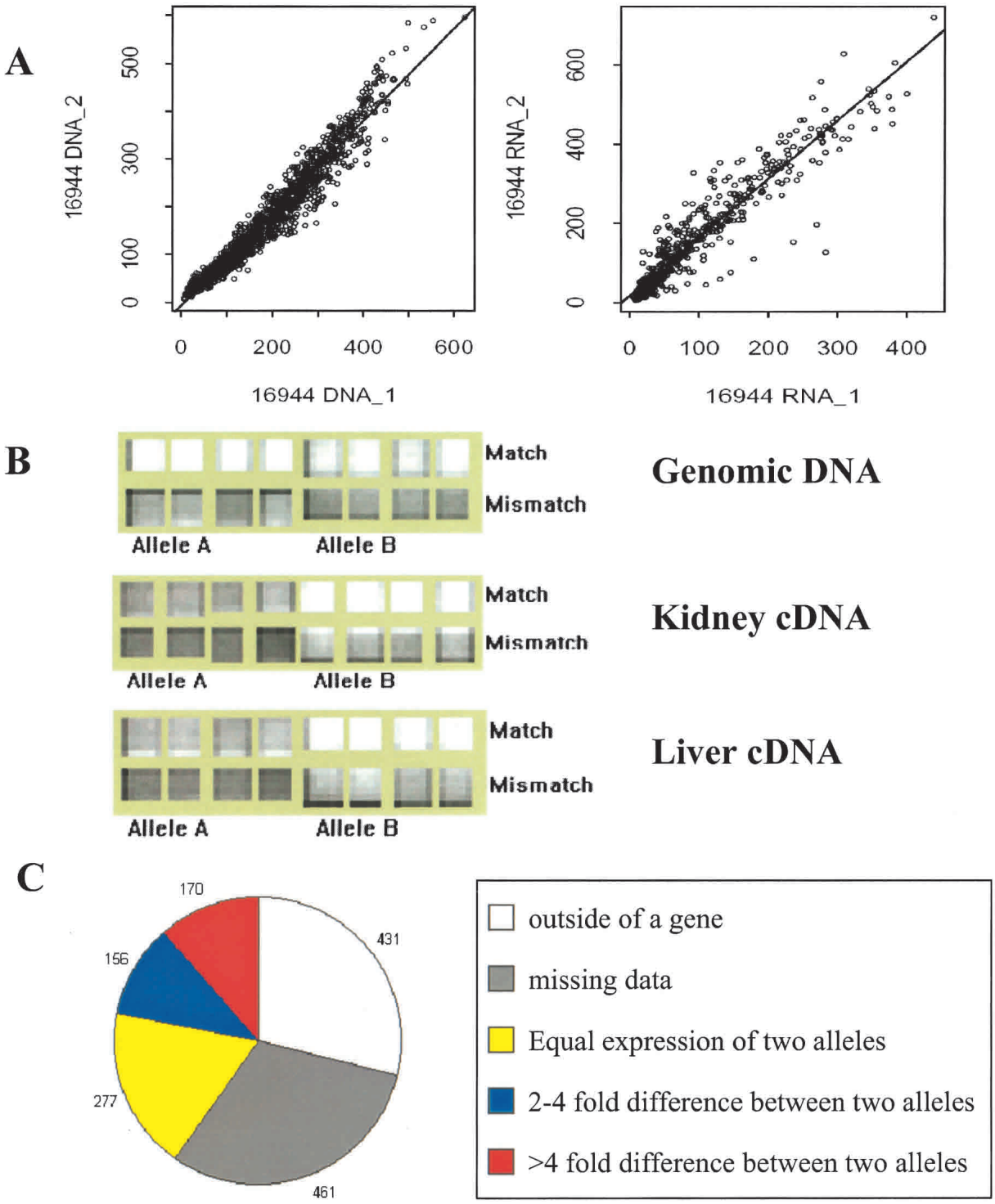


Figure 1 Evaluation of Affymetrix HuSNP chip for analysis of allelic gene expression. (*A*) Scatter plots for duplicate experiments. In this example, we performed duplicate experiments by using the same genomic DNA (*left*) or cDNA made from kidney RNA (*right*). Each *circle* represents a pair of intensity values in the two experiments for one SNP. The mean fluorescent intensities of the perfect match probe minus the mismatch probe from experiment 2 are plotted against the mean fluorescent intensity of the perfect match probe minus the mismatch probe from experiment 1. The Pearson correlation coefficients of the duplicate experiments for genomic DNAs and cDNA are 0.98 and 0.95, respectively. The *P* values for both correlation coefficients are <0.0001. In other duplicate sample tests, Pearson correlation coefficients ranged from 0.98–0.99 for genomic DNA and from 0.88 to 0.96 for cDNA, and the *P* values for these correlation coefficients are also <0.0001. (*B*) Probe images of the *PEG3* gene (SNP rs3143). The probe images were generated by using the Affymetrix MAS 4.0 software. The genomic DNA, kidney cDNA, and liver cDNA from the same fetus are each represented by a set of 16 hybridization signals. Within each set, each individual grid corresponds to a probe. The eight probes for allele A are on the *left*, and the eight probes for allele B are on the *right*. The *top* eight probes are for perfect match probes, whereas the *bottom* eight probes are for mismatch probes. The genomic DNA hybridized strongly to the perfect match probes for both alleles, whereas cDNAs hybridized strongly to the perfect match probes of allele B only. (*C*) SNPs/genes on the HuSNP chip are summarized in this diagram. There are 1494 SNPs on the HuSNP chip. We mapped 1063 SNPs to transcribed regions and 431 SNPs outside of the transcribed regions. Of the 1063 SNPs, 602 SNPs were analyzed (for selection criteria, see Methods). Among these 602 SNPs, 277 of them showed almost equal expression levels between the two alleles, whereas 156 SNPs had a ratio of gene expression between twofold and fourfold, and 170 SNPs had a ratio exceeding fourfold for at least one individual.

Differential Allelic Gene Expression Is Common in the Human Genome

Among the 602 genes analyzed, we found that 326 genes (54%) displayed a significant difference (at least a twofold difference; for details, see Methods) in at least one individual. Moreover, 170 genes (28%) showed a difference greater than fourfold. The number of genes that could be analyzed and showed differential expression between the two alleles is summarized in Figure 1C. We compiled a complete list of the allelic expression of all 326 genes identified (Supplemental Table 1, available online at www.genome.org). A subset of the genes showing differential allelic expression is given in Table 1. For 119 genes that showed differential allelic expression, there was more than one heterozygous fetus for the SNP. The degree of difference in the expression between the two alleles varied from individual to individual (Table 1). This result is consistent with a recent article that reported that six out of 13 genes had significant differences between expression of the two alleles, but these differences varied among individuals (Yan et al. 2002b).

The overall distribution of variation in gene expression between the two alleles of all 602 genes examined in this study is shown in Figure 2. The log of the ratios between the intensity of the two alleles in genomic DNA centers at zero (Fig. 2, solid line). The log of the ratios between gene expression of the two alleles in the cDNAs also center at 0 but with a much wider distribution of the data points (Fig. 2, dashed lines), indicating that there were many genes that showed allelic variation in gene expression.

Imprinted genes and genes subject to X-chromosome inactivation are expected to display skewed allelic expression. We compiled a list of 44 known imprinted genes from the literature (Supplemental Table 2). The HuSNP chip contains SNPs in six known imprinted genes: *DLK1*, *HTR2A*, *PEG3*, *SNRPN*, *WT1*, and *IPW*. Five of them had heterozygous fetuses among the seven fetuses analyzed. *SNRPN*, *IPW*, and *PEG3* showed mono-allelic expression in both kidney and liver, whereas *HTR2A* showed mono-allelic expression in liver. *WT1* showed bi-allelic expression, which is consistent with the previous reports that *WT1* imprinting is restricted to certain tissues such as placenta and brain (Little et al. 1992; Nishiwaki et al. 1997). The probe images for *PEG3* are shown in Figure 1B. Genomic DNA gave a uniformly strong signal to the probes for both alleles of *PEG3*, whereas cDNAs from kidney and liver hybridized strongly only to allele B.

Some of the genes that displayed preferential expression of one allele are located in regions that contain other imprinted genes. *TRIM22*, for example, showed mono-allelic expression in both kidney and liver, and it is located in a cluster of at least 10 imprinted genes at 11p15. *LOC145622*, which is mapped next to *SNRPN*, an imprinted gene located in the imprinting domain at 15q11, showed mono-allelic expression in kidney. Four of the seven genes on the X chromosome also displayed skewed allelic expression (Supplemental Fig. 1).

Mapping of Genes That Show Differential Allelic Gene Expression

To understand the genomic organization of allelic gene expression, we mapped all SNPs onto the human chromosomes. Chromosomes 9, 13, and 15 are shown (Fig. 3), and the complete map for the entire genome can be found in the Supplemental Figure 1. Some of the genes that showed skewed allelic expression are located next to each other, and a subset of these genes is located in known imprinting domains. *HTR2A*, *LOC51131*, and *FLJ13639* are located at 13q14, and all three show mono-allelic expression (Table 1). *SNRPN*, *IPW*, and *LOC145622* are located in the imprinting domain at 15q12, and all three genes preferentially ex-

Table 1. A Subset of Genes That Displayed Allelic Variation in Expression in HuSNP Experiments

				Kidney			Liver		
Gene	SNP	Alleles	Location	Mean ± SD	Min–Max	No. of fetuses	Mean ± SD	Min–max	No. of fetuses
DCTD	rs978	A, G	4q35.1	2.25 ± 1.70	1.30–5.28	5	4.27 ± 2.70	1.17–6.85	5
UGDH	rs1450	C+, T	4p15.1	5.54 ± 6.36	1.04–10.03	2	1.54 ± 0.27	1.35–1.74	2
RAI14	rs1390	A, G+	5p13.3-p13	3.38 ± 3.28	1.23–1.75	3	5.64 ± 3.06	2.81–8.88	3
RASGRF2	rs1589	C, T+	5q13	4.53 ± 4.62	1.26–7.80	2	1.58	1.58–1.58	1
TAP2	rs17034	A, G	6p21.3	3.40 ± 2.97	1.40–7.79	4	1.96 ± 0.61	1.53–2.40	2
DKFZP727G051	rs1837	C, T	9q33.3	2.00 ± 1.46	1.17–4.20	4	1.68 ± 0.66	1.13–2.65	4
PAPPA	rs1405	C, T	9q33.1	2.77 ± 2.00	1.01–4.93	4	6.85	6.85–6.85	1
GRF2	rs1772	A, G+	9q34.3	1.47	1.47–1.47	1	6.08	6.08–6.08	1
VAV2	rs16763	A, C+	9q34.1	3.14 ± 1.85	1.12–4.74	3	4.18 ± 2.13	1.31–6.44	4
TRIM22	rs2179	C, G	11p15	4.37 ± 4.18	1.51–9.18	3	3.54 ± 2.53	1.13–6.29	4
DSCAML1	rs16867	C, T	11q23	6.06 ± 3.37	2.94–9.62	3	5.45	5.45–5.45	1
VELI1	rs1537	C, G	12q21	1.64 ± 0.14	1.54–1.73	2	5.05 ± 2.87	3.01–8.33	3
FLJ13639	rs2735	C, G+	13q14.2	3.17 ± 2.03	1.74–4.61	2	4.57	4.57–4.57	1
HTR2A	rs3125	C+, G	13q14-q21	1.06	1.06–1.06	1	16.46	16.46–16.46	1
LOC51131	rs2980	A+, G	13q14.12	1.29	1.29–1.29	1	6.96	6.96–6.96	1
SNRPN	rs705	C+, T	15q12	5.4	5.40–5.40	1	5.21	5.21–5.21	1
IPW	rs691	C, T+	15q12	6.44 ± 4.24	1.88–10.26	3	3.73	3.73–3.73	1
LOC145622	rs17068	C+, G	15q11.2	2.41	2.41–2.41	1	1.85	1.85–1.85	1
ELAC2	rs2523	C, T+	17p11.2	4.07	4.07–4.07	1	2.28	2.28–2.28	1
KRT13	rs1031	G, T	17q21-q23	3.92 ± 3.25	1.05–7.45	3	4.00 ± 2.55	1.12–5.98	3
ZIM2	rs3143	A, G+	19q13.4	4.59 ± 2.34	2.93–6.24	2	5.72	5.72–5.72	1
ARHGAP8	rs33329	A, G	22q13.31	2.40 ± 0.48	1.79–2.95	4	1.95 ± 1.27	1.18–3.86	4
RANGAP1	rs1953	A, G	22q13	3.75 ± 3.83	1.08–9.26	4	2.32 ± 1.22	1.09–3.80	4

A complete list of all genes can be found in the Supplemental Table. The values are the ratios (allele A/allele B) between the two alleles. The values were inverted if less than one (allele B/allele A, when allele B was preferentially expressed). The allele A and allele B are the first base and second base, respectively. *The bases are the ones defined in WIAF markers. They are complementary to the bases in rs if sequences described in WIAF are in opposite orientation to the sequences described in rs.* The preferentially expressed allele is labeled with "+," which is selected if the allele is preferentially expressed in at least 80% of samples.

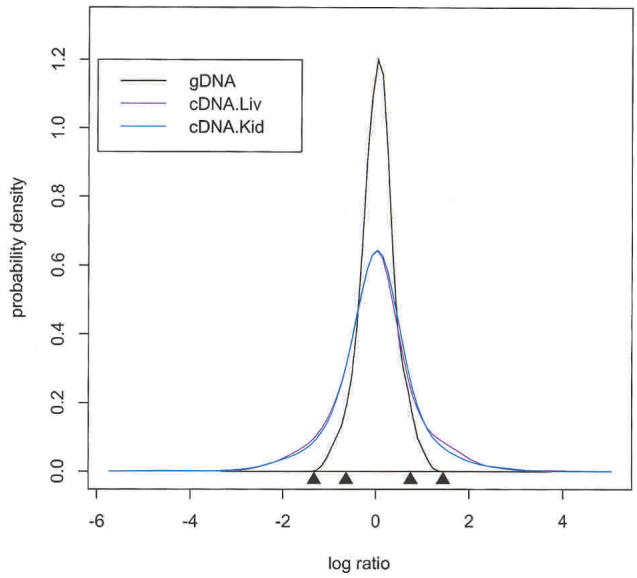


Figure 2 Distribution of ratios of the fluorescent intensities between the two alleles for genomic DNA and cDNA. The ratios were computed as (PMA − MMA)/(PMB − MMB) for each SNP for every sample. From the ratios in genomic DNA samples, the 1-SD interval around the mean is between −1.27 and 1.17 in log scale. The interval in log scale corresponds to the interval between 0.28 and 3.22 for the ratios. We selected 602 SNPs for analysis (for selection criteria, see Methods). To compare the distributions of ratios in genomic DNA and cDNA, we plotted frequency of samples against the log ratio. Density functions for genomic DNA, kidney cDNA, and liver cDNA are represented by a black line, blue line, and purple line, respectively. Black *triangles*, from *left* to *right*, indicate *X* coordinates at log(0.25), log(0.5), log(2), and log(4). The coordinates at log(0.5) and log(2) represent twofold ratios, and log(0.25) and log(4) represent fourfold ratios. The density functions for the kidney cDNA ratios and the liver cDNA ratios are similar. Both have a wider spread compared to the density function for the genomic DNA.

press one allele (Fig. 3; Table 1). There are also regions that contain several allele-biased genes outside of the known imprinting domains. For example, *RASGRF2* and *RAI14* are located at 5q13 (Table 1), and *GRF2* and *VAV2* are located at 9q34 (Fig. 3; Table 1). Thus, these might be novel regions undergoing allele-specific gene regulation. It is also interesting to note that *RASGRF2* and *GRF2* both showed mono-allelic expression, whereas a homologous gene, *RASGRF1*, is a known imprinted gene (Plass et al. 1996). However, the vast majority of the genes that show preferential expression of one allele are scattered (Fig. 3), indicating that allelic variation occurs throughout the human genome.

Validation of Allelic Variation in Gene Expression by Real-Time Quantitative PCR

To validate the results of the HuSNP experiment, we performed allele-specific quantitative PCR for seven genes: two known imprinted genes (*PEG3* and *SNRPN*), four genes (*TAP2*, *ELAC2*, *DKFZP727G051*, and *UGDH*) that displayed allelic variation in gene expression, and one gene (*C11orf23*) that expressed almost equally between two alleles in the HuSNP experiment. We first performed genotyping by using TaqMan genotyping assay for the seven fetuses analyzed in the HuSNP experiment. The resulting genotype calls were identical to the HuSNP experiment. We then performed genotyping for additional fetuses in order to identify more samples that are heterozygous or homozygous. An example of the genotyping calls for *ELAC2* is shown in Figure 4A. We then mixed genomic DNAs of homozygous AA and BB individuals with seven different ratios in TaqMan assays to establish a linear regression line for the log of fluorescent intensity ratio (FAM/VIC) versus the log of allele ratio for each gene (see Methods). Figure 4B shows an example of such a linear regression line for *ELAC2*. For each of the seven genes, we established these standard curves. By using real-time quantitative PCR and these standard curves, we were able to deduce the ratios of gene expression between the two alleles by measuring the fluorescent intensity of the two alleles in cDNA samples as shown in Figure 4, C and D. The real-time quantitative PCR results for the seven genes are summarized in Table 2. The two known imprinted genes displayed more than eightfold difference between two alleles (the linear regression lines were established for difference within eightfold). The allele ratio for *C110rf23*, which displayed nearly equal expression in the HuSNP experiment, also showed nearly equal expression between the two alleles (ratio between 1.2 and 1.5; Table 2). *UGDH* did not show much difference between the two alleles, although a difference between the two alleles was detected in the HuSNP experiments. *TAP2*, *ELAC2*, and *DKFZP727G051* showed significant differences in gene expression between the two alleles, which confirmed the results of the HuSNP experiment. Thus, the results of real-time quantitative PCR are consistent with the HuSNP experiments in six out of seven genes (Table 2).

DISCUSSION

Affymetrix HuSNP chip array provides a very effective platform for the simultaneous analysis of large numbers of genes to analyze allele-specific gene expression. Our study indicates that variation in expression between two alleles is common and that these genes are distributed throughout the entire genome, although some of them are clustered. Variations in allelic expression can be caused by genomic imprinting, X-chromosome inactivation, or other mechanisms. One example of this latter class is provided by a recent study (Yan et al. 2002b), which demonstrated that allelic variation can be transmitted by mendelian inheritance. Their earlier work also linked the reduced expression of an affected allele of the *APC* gene to colon cancer (Yan et al. 2002a).

Our studies identified 326 genes that showed preferential expression of one allele, with 170 of those showing greater than fourfold difference between the two alleles. There are six imprinted genes with SNPs represented on the HuSNP chip, and five of them had at least one heterozygous fetus among the samples that we analyzed. Among these five genes, four of them showed differential expression between the two alleles, whereas the fifth gene, *WT1*, is known not to be imprinted in fetal kidney or liver tissues. The fact that these known imprinted genes were identified by our method indicates that additional novel imprinted genes can be identified from our list of genes that showed differential expression between the two alleles. Thus, these genes provide a rich source to identify novel imprinted genes and to study the role of allelic variation in gene expression in normal physiology and in diseases.

Real-time quantitative PCR is an established method for measuring quantitatively gene expression and genotyping. By mixing DNAs with various ratios from homozygous AA and BB individuals, it is possible to define a region of linear response between the log of allele ratio and the log of fluorescent intensity ratio and to use this linear regression line to determine allele-specific gene expression. By using real-time quantitative PCR, we found that the status of allelic gene expression variation in six of the seven genes was in agreement with what we found in the HuSNP experiments (Table 2).

The consistency of the HuSNP experiment system has been demonstrated in Figure 1A. High degree of correlation between

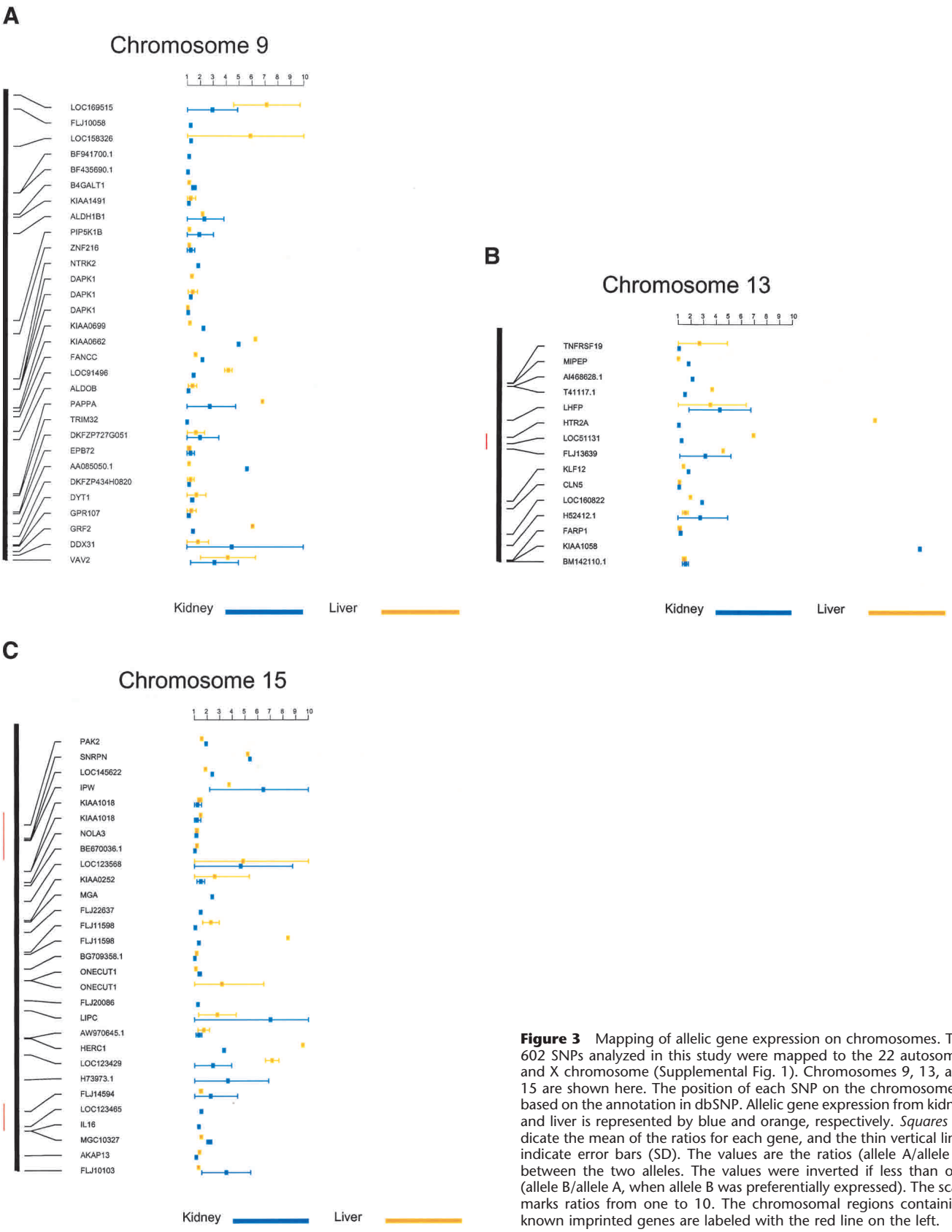


Figure 3 Mapping of allelic gene expression on chromosomes. The 602 SNPs analyzed in this study were mapped to the 22 autosomes and X chromosome (Supplemental Fig. 1). Chromosomes 9, 13, and 15 are shown here. The position of each SNP on the chromosome is based on the annotation in dbSNP. Allelic gene expression from kidney and liver is represented by blue and orange, respectively. *Squares* indicate the mean of the ratios for each gene, and the thin vertical lines indicate error bars (SD). The values are the ratios (allele A/allele B) between the two alleles. The values were inverted if less than one (allele B/allele A, when allele B was preferentially expressed). The scale marks ratios from one to 10. The chromosomal regions containing known imprinted genes are labeled with the red line on the left.

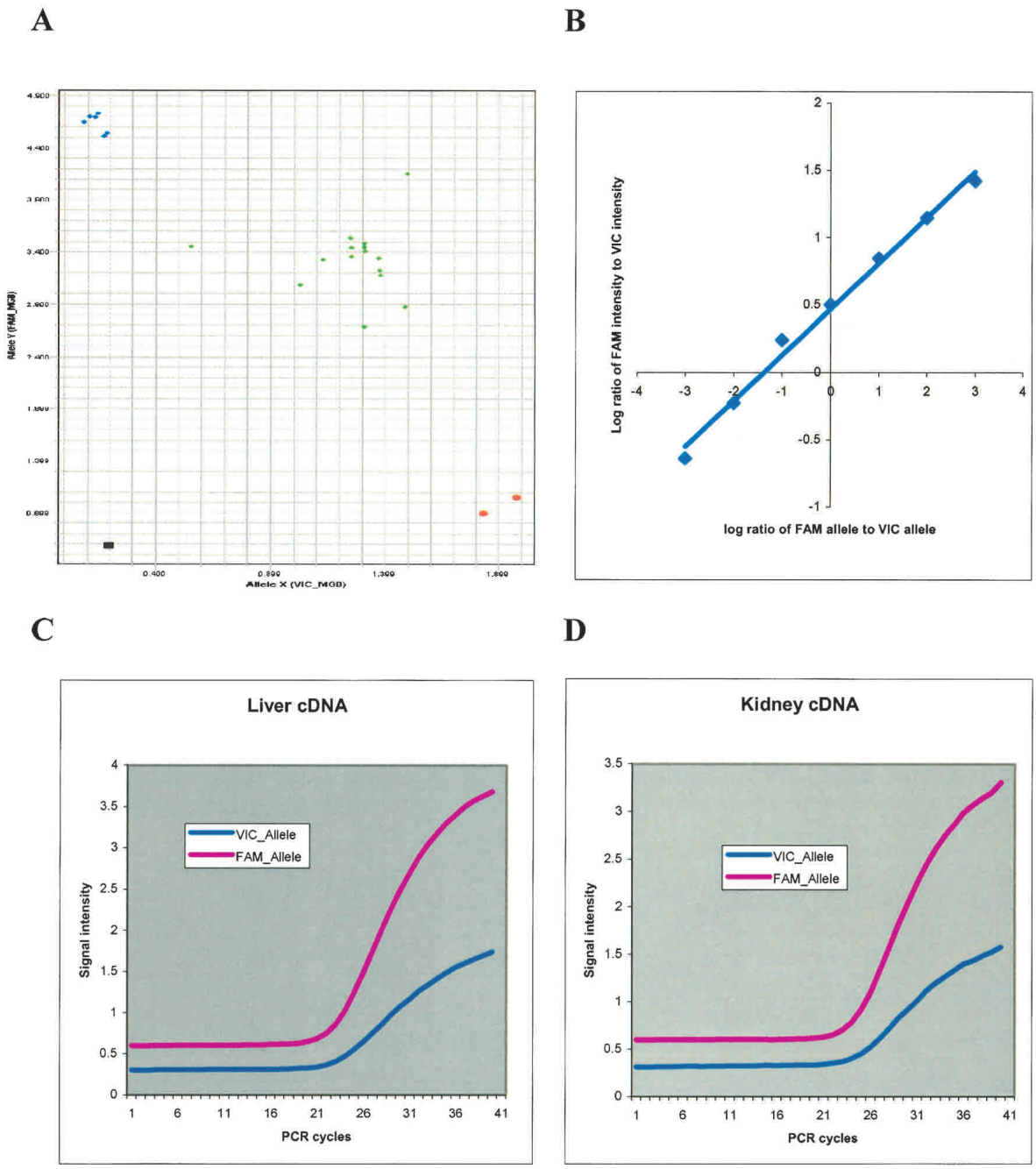


Figure 4 Validation of allele-specific gene expression using real-time quantitative PCR. (*A*) Genotyping of *ELAC2* in 23 fetuses. Genomic DNAs from homozygous AA fetuses are at *top left* corner (blue), and genomic DNAs from homozygous BB fetuses are at *bottom right* corners (red). Genomic DNAs from heterozygous fetuses are located near the diagonal line (green). The black *square* represents no template control (NTC). The *X*-axis is for allele labeled by the VIC dye, and the *Y*-axis is for allele labeled by the FAM dye. (*B*) The $\log_2$ of (FAM intensity/VIC intensity) for *ELAC2* was plotted against $\log_2$ of (FAM allele/VIC allele) of mixing homozygous DNAs at seven different ratios (8 : 1, 4 : 1, 2 : 1, 1 : 1, 1 : 2, 1 : 4, 1 : 8; VIC allele/FAM allele). (*C*) Real-time quantitative PCR amplification of a cDNA sample from liver for *ELAC2*. The *X*-axis is the number of PCR amplification cycles, and the *Y*-axis is the fluorescence intensity. The red and blue curves represent alleles labeled with FAM and VIC, respectively. (*D*) Same as *C* except that the data are from kidney.

duplicate experiments was observed (Fig. 1A). We have taken three approaches to further address the issue of consistency among individuals and tissues. First, we selected 51 SNPs for which there were at least four heterozygous individuals. Among the 51 SNPs, 29 (57%) showed skewed allelic expression in at least two fetuses, and 10 (20%) showed skewed allelic expression in at least three fetuses. To examine consistency of preferentially expressing the same allele, we compared the genes that contained the 29 SNPs and expressed preferentially one allele in both kidney and liver. We found that 26 sample sets had ratios for both kidney and liver, and 19 out of 26 (73%) preferentially expressed the same allele. Seven out of 26 (27%) expressed different alleles, and this appeared to be due to marginal hybridization signals in one of the tissues. This is not unexpected for this

Table 2. The Allelic Gene Expression Ratios for Seven Genes in Multiple Tissues Measured by Quantitative PCR

Genes	SNP ID	Liver	Kidney	Heart	Other[a]	No. of fetuses
ZIM2[b]	rs3143	>8.0	>8.0	>8.0	>8.0	3
SNRPN[b]	rs705	>8.0	>8.0	>8.0	>8.0	3
TAP2[b]	rs17034	3.22 ± 1.57	1.90 ± 0.88	4.67 ± 2.88	2.19 ± 0.89	7
ELAC2[b]	rs2523	2.33 ± 0.87	2.55 ± 0.54	2.92 ± 0.73	2.86 ± 0.52	7
DKFZP727G051[b]	rs1837	2.80 ± 1.05	2.32 ± 0.53	2.16 ± 0.41	2.05 ± 0.43	7
UGDH	rs1450	1.53 ± 0.72	1.44 ± 0.29	1.39 ± 0.14	1.49 ± 0.35	3
C11orf23	rs1996	1.52 ± 0.27	1.38 ± 0.39	1.52 ± 0.29	1.22 ± 0.16	2

The values are the ratios (allele A/allele B) between the two alleles. The values were inverted if less than one (allele B/allele A, when allele B was preferentially expressed). The mean ± SD is provided for allelic gene expression in the various tissues.
[a]Several different fetal tissues were used, including adrenal, limb, and lung.
[b]Allelic difference in gene expression is significantly deviated from equal expression (95% confidence interval). The confidence interval was generated from heterozygous DNAs.

type of experiment that performs hybridization for several hundreds of genes. Second, we extended the analysis to the 326 genes that showed preferential expression of one allele, and there were multiple samples in which allelic variation in gene expression was observed. We found that 272 (83%) out of the 326 genes preferentially expressed the same allele among different fetuses and tissues (preferentially expressed allele was labeled with "+" in Table 1 and in Supplemental Table 1). Third, we measured allelic gene expression in seven individuals in four tissues for three genes in TaqMan assay (Table 2). We found that the consistency for skewed allelic expression is 95% in cross-individual comparison and cross-tissue comparison.

Our method will enable large-scale analysis of allelic gene expression of clinical samples such as those from human cancers. The method can be easily scaled up with a higher density chip such as the 10K SNP chip, which may be available from Affymetrix in 2003. Our study indicates that the two alleles of human genes are not always expressed "equally." On the contrary, allelic variation in gene expression is common and may affect 20% to 50% of human genes. This may be the basis for variation in the transmission of some diseases, and it provides a potential mechanism for generating human variation.

METHODS

HuSNP Experiments

Fetal tissues were obtained from the Birth Defects Research Laboratory, University of Washington. The tissues were snap-frozen after surgery and were stored in liquid nitrogen. Five fetuses were male, and two were female. The ages of the fetuses ranged from 78 to 103 d. Fetal genomic DNA was prepared by using the QIAamp DNA mini kit (Qiagen, Inc.). RNAs were isolated from fetal tissues by using RNAzol B (Tel-Test, Inc.) according to the manufacturer's protocol. Poly-A RNAs were isolated by using the Micro-Fast Track kit (Invitrogen Corp.). cDNA was synthesized by using AMV reverse transcriptase (Invitrogen Corp.).

The subsequent steps of multiplex PCR amplification, chip hybridization, chip staining, and chip scanning were all conducted according to the GeneChip HuSNP mapping assay manual (P/N 700308, Affymetrix, Inc.). Briefly, 120 ng of genomic DNA or 6 ng of cDNA was used for each set of 24 multiplex PCR reactions, and the resulting biotinylated amplicons were concentrated to 60 µL. Half (30 µL) of the concentrated amplicon was used for hybridization to a HuSNP chip for 16 h at 44°C. The chip was then washed and stained with a complex of streptavidin phycoerythrin (SAPE) and biotinylated anti-streptavidin IgG antibody on a GeneChip fluidics station, followed by scanning in a HP GeneArray Scanner (Affymetrix, Inc.). Genotyping calls were made by using the Affymetrix MicroArray Suite (MAS) software, version 4.0.

Computational Analysis of the HuSNP Data

We downloaded the sequence of each of the 1494 SNPs and performed BLAST search against dbEST. We also mapped SNPs by using the annotation in dbSNP. We were able to map 1063 SNPs to the transcribed regions of genes. The criteria for mapping SNPs to the transcribed regions were (1) at least two EST hits, (2) *E* value $<10^{-10}$, and (3) alignment >40 bp.

We extracted the intensity values for each probe from the .CEL files generated by Affymetrix MAS 4.0. The .CEL files contain the fluorescent intensity values for each of the probes. The HuSNP chip contains 16 probes for each SNP locus. Four of the 16 probes match perfectly to allele A, four match to allele B, four have one mismatch to allele A, and the other four have one mismatch to allele B. Allele A and allele B represent the two alleles of the SNP. Allele A and allele B are assigned alphabetically. For example, if a SNP has C and T bases, the C base is defined as A allele and the T base is defined as allele B (for allele information, see Supplemental Table 1). Each probe contains 20 nucleotides. The center of the nucleotide probes is located at positions −4, −1, 0, and 1 relative to the SNP. The four mismatch probes are identical to the perfect match probes, except for one mismatched base, which is always located in the center of the probe. There are typically four probe pairs for each of the allele A and the allele B, except for 95 SNPs that have five probe pairs. The value for each probe pair was computed by subtracting the mismatch intensity from the perfect match intensity. A *t* test was used to calculate a *P* value for the presence of signal (intensity greater than zero) for each allele of each SNP. We considered a signal to be present if at least one allele had signal ($P < 0.01$, *t* test). Affymetrix defines a mini-block as a group of four probes that include a perfect match probe for allele A (PMA), a mismatch probe for allele A (MMA), a perfect match probe for allele B (PMB), and a mismatch probe for allele B (MMB). We set (PMA − MMA) = 50 if (PMA − MMA) is <50 for each mini-block. Similarly, baseline for allele B was set at 50. An allele A fraction, defined as $f = (\text{PMA} - \text{MMA})/(\text{PMA} - \text{MMA} + \text{PMB} - \text{MMB})$, was computed for each mini-block, and the mean of the allele A fraction *f* from mini-blocks was computed for each SNP. The gene expression difference between the two alleles from a heterozygous individual can be quantified by using the ratio of allele A to allele B, computed from $f/(1 - f)$. For each chip, we have intensities from two scans called scan A and scan B. Generally, we used the intensity values from scan A. We used the intensity values from scan B if the *t* test showed that both alleles have no signal in scan A, and at least one of the alleles from scan B had signal. The ratio was further normalized by the ratio of genomic DNAs for the SNP. Among the 602 SNPs analyzed in our studies, 39 had at least five heterozygous fetuses. We computed the 95% confidence interval for the allelic ratio of genomic DNA for each of these 39 SNPs, and the average confidence interval was between 0.5 and 2.0. This value was used to select those genes that show significant difference in the expression between the two alleles.

To evaluate concordance between two duplicate experiments, we computed the Pearson correlation coefficient between the two experiments by using the mean intensity of the probe pairs from each allele of a SNP.

Real-Time Quantitative PCR Experiments

We used the ABI PRISM 7900HT Sequence Detection System and Assays-on-Demand SNP Genotyping products for genotyping and allele-specific gene expression (TaqMan assay). We followed the manufacturer's protocol for the preparation of the PCR reactions. Sequence Detection Systems software (SDS 2.0) was used to automatically collect and analyze the data and to generate the genotype calls. We mixed the genomic DNAs from the two homozygous individuals, one with genotype of AA and the other with genotype of BB, with the following ratios: 8 : 1, 4 : 1, 2 : 1, 1 : 1, 1 : 2, 1 : 4, and 1 : 8 (VIC allele/FAM allele). TaqMan assays were conducted, and the fluorescent intensity data were exported as tab-delimited text files from the SDS software. For each mixing ratio of a given gene, we calculated the log of (FAM intensity/VIC intensity) at the last PCR cycle (cycle 40). We generated a standard curve (linear regression line), $y = a + bx$, where y is the log of (FAM intensity/VIC intensity) at a given mixing ratio, x is the log of mixing ratio, a is the intercept, and b is the slope. We then measured allele-specific gene expression by using real-time quantitative PCR. We extrapolated the allele ratio on gene expression by intercepting log of (FAM intensity/VIC intensity) on the standard curve.

ACKNOWLEDGMENTS

We thank University of Washington Fetal Tissue Bank for providing fetal samples. We like to thank Dr. Jeffery Struewing and Jenny Kelley for critical reading of the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Gartler, S.M. and Goldman, M.A. 2001. Biology of the X chromosome. *Curr. Opin. Pediatr.* **13:** 340–345.

Lindblad-Toh, K., Tanenbaum, D.M., Daly, M.J., Winchester, E., Lui, W.O., Villapakkam, A., Stanton, S.E., Larsson, C., Hudson, T.J., Johnson, B.E., et al. 2000. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.* **18:** 1001–1005.

Little, M.H., Dunn, R., Byrne, J.A., Seawright, A., Smith, P.J., Pritchard-Jones, K., van, H.V., and Hastie, N.D. 1992. Equivalent expression of paternally and maternally inherited WT1 alleles in normal fetal tissue and Wilms' tumours. *Oncogene* **7:** 635–641.

Nishiwaki, K., Niikawa, N., and Ishikawa, M. 1997. Polymorphic and tissue-specific imprinting of the human Wilms tumor gene, *WT1*. *Jpn. J. Hum. Genet.* **42:** 205–211.

Plass, C., Shibata, H., Kalcheva, I., Mullins, L., Kotelevtseva, N., Mullins, J., Kato, R., Sasaki, H., Hirotsune, S., Okazaki, Y., et al. 1996. Identification of Grf1 on mouse chromosome 9 as an imprinted gene by RLGS-M. *Nat. Genet.* **14:** 106–109.

Tycko, B. and Morison, I.M. 2002. Physiological functions of imprinted genes. *J. Cell Physiol.* **192:** 245–258.

Yan, H., Dobbie, Z., Gruber, S.B., Markowitz, S., Romans, K., Giardiello, F.M., Kinzler, K.W., and Vogelstein, B. 2002a. Small changes in expression affect predisposition to tumorigenesis. *Nat. Genet.* **30:** 25–26.

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

WEB SITE REFERENCES

ftp://ftp.ncbi.nih.gov/snp/human; National Center for Biotechnology Information (NCBI) dbSNP FTP site.

Received November 18, 2002; accepted in revised form June 9, 2003.