Mixed-Model Reanalysis of Primate Data Suggests Tissue and Species Biases in Oligonucleotide-Based Gene Expression Profiles

Wen-Ping Hsieh,* Tzu-Ming Chu,† Russell D. Wolfinger† and Greg Gibson*,[1]

**Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695 and*
†*SAS Institute, Cary, North Carolina 27513*

Manuscript received March 26, 2003
Accepted for publication June 18, 2003

ABSTRACT

An emerging issue in evolutionary genetics is whether it is possible to use gene expression profiling to identify genes that are associated with morphological, physiological, or behavioral divergence between species and whether these genes have undergone positive selection. Some of these questions were addressed in a recent study (Enard *et al.* 2002) of the difference in gene expression among human, chimp, and orangutan, which suggested an accelerated rate of divergence in gene expression in the human brain relative to liver. Reanalysis of the Affymetrix data set using analysis of variance methods to quantify the contributions of individuals and species to variation in expression of 12,600 genes indicates that as much as one-quarter of the genome shows divergent expression between primate species at the 5% level. The magnitude of fold change ranges from 1.2-fold up to 8-fold. Similar conclusions apply to reanalysis of Enard *et al.*'s (2002) parallel murine data set. However, biases inherent to short oligonucleotide microarray technology may account for some of the tissue and species effects. At high significance levels, more differences were observed in the liver than in the brain in each of the pairwise species comparisons, so it is not clear that expression divergence is accelerated in the human brain. Further, there is an apparent bias toward upregulation of gene expression in the brain in both primates and mice, whereas genes are equally likely to be up- or downregulated in the liver when these species diverge. A small subset of genes that are candidates for adaptive divergence may be identified on the basis of a high ratio of interspecific to intraspecific divergence.

ONE of the most interesting applications of gene expression profiling in evolutionary genetics is the comparison of transcript abundance among closely related species. Given that studies of yeast, flies, and killifish have each suggested that between 10 and 25% of the transcriptome differs significantly in expression level between any two individuals of the same species (Cavalieri *et al.* 2000; Jin *et al.* 2001; Cheung and Spielman 2002; Oleksiak *et al.* 2002), there is an expectation that a similar fraction of the transcriptome may differ between sibling species. Some of these differences will be associated with morphological, physiological, and behavioral diversification and if causally related to the divergence may also provide signatures of natural selection. Quantification of transcript abundance within and between species thus has much to contribute to our understanding of the evolutionary forces acting at the level of gene expression.

The first effort to address these questions in relation to human evolution was recently published by Pääbo and co-workers (Enard *et al.* 2002). The centerpiece of their study was a comparison of 12,600 gene expression profiles of left prefrontal lobe brain samples (Brodmann area 9) from three humans, three chimpanzees, and an orangutan, each of which had died of natural causes, using Affymetrix U95A oligonucleotide gene chips. They also examined liver samples from the same specimens and conducted a parallel series of experiments with *Mus musculus*, *M. spretus*, and *M. caroli*, three mouse species that show similar levels of genetic divergence. After computing a pairwise distance matrix on the basis of the average level of expression for each gene on their arrays, they drew neighbor-joining trees that summarize the overall divergence in transcript abundance for each tissue between the triplets of species. Their major finding was that the branch joining the three human samples to the central node on their brain expression tree was almost twice as long in relative terms as the same branch on the liver tree or in either of the murine trees. The same result was obtained with a smaller experiment using cDNA microarrays. Hence the authors concluded that gene expression had diverged most rapidly in the human brain.

Although it is easy to criticize this study over concerns such as the small sample size, the suitability of senescent individuals, and the validity of extrapolating to general conclusions on the basis of a small section of the brain, it is also the case that the already rich data set will support further quantitative analyses that may be of

[1]*Corresponding author:* Department of Genetics, Gardner Hall, North Carolina State University, Raleigh, NC 27695-7614.
E-mail: ggibson@unity.ncsu.edu

interest. In the reanalysis of the data reported here, we sought to address the following questions: how many of the genes on the array are actually significantly more divergent between than within species; what is the mean magnitude of expression divergence between species; why did one of the human samples have an average difference from the other two that was as great as their overall divergence from the chimps; what is the nature of the genes that have diverged in expression; and do the same genes diverge between all three species? Our major finding is that gene expression actually diverges more between human and chimp liver samples than between human and chimp brain samples.

In the course of our analyses, we also noted biases in the directionality and significance of changes in expression that led us to question whether the Affymetrix technology is really suitable for interspecific comparisons. We implemented mixed-model analysis of variance in SAS (Wolfinger *et al.* 2001; Chu *et al.* 2002) to tease apart the contributions to transcript abundance of variance among individuals and between species. Fluorescence intensity from each individual perfect-match oligonucleotide probe was taken as the measure of expression, rather than the average difference between perfect and mismatch probes. A detailed analysis of a subset of the genes that showed strong species-by-probe interaction effects highlights some of the difficulties associated with the use of oligonucleotide arrays to compare genotypes that diverge at the nucleotide level. Consequently, our results also have implications for the interpretation of Affymetrix data for any comparisons of genetically polymorphic strains.

MATERIALS AND METHODS

Mixed-model analysis of variance: Variation in gene expression was assessed using a two-step strategy essentially as outlined in Chu *et al.* (2002). In the first step, each individual probe measurement was centered relative to the array mean by subtracting the $\log_2$-transformed value of the intensity from the mean $\log_2$ value for the probes on the array. We simply used the perfect-match (PM) intensities and ignored the mismatch (MM) values as we find that these statistically generally just add noise. Data quality was then checked by plotting pairwise scatter plots of the normalized probe intensities for each possible comparison of similar treatments. See supplementary Figure 1 (http://www.genetics.org/supplemental/) for the six human brain samples. All contrasts show good linear correlations as expected. Values range from −2 to +4, with the upper limit indicating saturated signal intensity. Almost 0.9% of the probes were at this level, undoubtedly reducing the power of comparisons involving highly expressed genes. As generally observed, transcript abundance is skewed toward a large number of genes showing relatively low abundance, hence the skewed distribution of intensity values about the mean. All pairwise comparisons contrasting individual 2 with individual 1 or 3 have broader scatter plots, reflecting the reported observation that this individual has more divergent expression in the brain than do the other two individuals. Similar saturation values were seen for the other tissues and species.

The second modeling step was to fit gene-specific mixed models using PROC MIXED in SAS as follows:

$$\log_2(\mathrm{PM}_{ijkl}) = S_i + T_j + P_k + ST_{ij} + TP_{jk} + SP_{ik} + R_{l(ij)} + \varepsilon_{ijkl}.$$

PM_{ijkl} denotes the perfect-match expression measurement of the kth probe of the lth individual for ith species (human, chimp, or orangutan) in the jth tissue (brain or liver). S, T, and P represent the fixed effects of species, tissues, and probes, respectively. Individual effects within species were specified as random effects and assumed to be independent and identically distributed according to a normal distribution with mean zero and variance σ_r^2. The ε_{ijkl}'s were also specified as independent and identical normal distributions with mean zero and variance σ^2 that are independent of the $R_{l(ij)}$'s. For the comparison in Figure 6, variance components of species effects were zero for a large fraction of genes, so we instead present results of a simplified general linear model run with PROC GLM in SAS on the reduced data consisting of human and chimp brain arrays only.

Correlation coefficients for filtering probes: Correlation coefficients were calculated between $\log_2$-transformed human brain PM intensity values (a_{ijk}) and corresponding chimp brain values (b_{ijk}) from the ith individual, jth sample, and kth probe associated with a particular gene. The expression profile for each species was first calculated as the average probe measure among samples: $a_{..k} = \Sigma_j\Sigma_i a_{ijk}/(I \times J)$, $b_{..k} = \Sigma_j\Sigma_i b_{ijk}/(I \times J)$. Subsequently, the correlation coefficient between the expression profiles of those two species was computed as $(\Sigma_k a_{..k}b_{..k} - K\Sigma a_{...}b_{...})/\left(\sqrt{(\Sigma_k a_{..k}^2 - Ka_{...}^2)}\ \sqrt{(\Sigma_k b_{..k}^2 - Kb_{...}^2)}\right)$, where $a_{...}$ is the average of $a_{..k}$ and $b_{...}$ is the average of $b_{..k}$.

Outlier probes were then deleted systematically until the correlation exceeded 0.95. For example, removal of the single inconsistent probes in Figure 4, C and D, results in a large increase in the overall correlation between human and primate data. The species effect for the remaining probes is consistent and likely represents a better measure of the true difference in gene expression.

Neighbor-joining trees: Euclidean distance matrices were computed for each pair of arrays on the least-squares mean gene expression measures from the mixed-model analysis and rescaled to fit the format required by the package PHYLIP (Felsenstein 1989). Neighbor-joining trees were generated by the NEIGHBOR option with default settings. This analysis is similar to that of Enard *et al.* (2002), except that they used average distance measures computed by Affymetrix software.

Crude estimation of the fraction of genes diverging under positive selection: Following Rifkin *et al.* (2003) and Lynch and Hill (1986), under mutation-drift equilibrium, the expected squared difference between species for each gene expression level is $\sigma_m^2 t$, where σ_m^2 is the mutational variance and t the time in generations since separation, and the expected level of intraspecific variance, which is assumed to have remained constant in both lineages since divergence, is $2N_e\sigma_m^2$, where N_e is the effective population size. Then the ratio of mean square estimates of the species and individual within-species effects is $F_{\text{human-chimp}} \sim [\mathrm{MS}_{\text{species}}/\mathrm{MS}_{\text{Ind(species)}}] \times [\,2N_e\sigma_m^2/\sigma_m^2 t]$. The mutational variances cancel out, so that the relationship between the observed and expected ratio of divergence to polymorphic variance is scaled by the ratio $2N_e/t$. Assuming an N_e of 10,000 individuals and one generation every 15 years in the 6–7 million years since divergence between human and chimp, the expected distribution of F ratios is expected to be 20–23 times the standard $F_{1,2}$ distribution (with 1 d.f. for the species comparison and 2 d.f. for the three individuals within each species). The outer 2.5% tail for this comparison must exceed an F value of 39; hence under these conservative conditions only ratios $>(39 \times 20)$, ~800, provide clear evidence

for a rate of expression divergence greater than that expected under this simple neutral model. Only 17 genes satisfy these criteria, but relaxation of the population size to 100,000 individuals and number of generations to 100,000 reduces the expected rate of neutral divergence, and almost 500 of the 12,600 genes (4%) would fall into the unexpectedly rapidly divergent class. This analysis serves primarily to highlight the conclusion that even high ratios of between-species to among-individual variance need not imply the action of positive (diversifying) selection.

RESULTS

Mixed-model analysis of the Affymetrix data: The primate data set reported by Enard *et al.* (2002) consists of 28 gene chips, including 14 for each tissue (brain or liver), and two replicates of each of the seven individuals. Each of the ~12,600 genes was represented by up to 20 unique probes, although these often overlap as described below. These data were analyzed using mixed-model analysis of variance (Wolfinger *et al.* 2001; Chu *et al.* 2002) as described in materials and methods and briefly here. The data were first centralized simply by taking the logarithm of each probe fluorescence intensity on the base 2 scale and then subtracting the mean value for the particular gene chip. The relative fluorescence intensity, $\log_2(PM_{ijkl})$, thus represents a measure of transcript abundance observed for the kth perfect-match probe for the lth individual within the ith species (human, chimp, or orangutan) sampled for the jth tissue (brain or liver). A value of 0 corresponds to a gene expressed at the sample mean, −1 or 1 to a gene that is one-half or twofold greater than the sample mean, −2 or 2 to a gene that is one-quarter or fourfold greater, and so on. If a gene has a mean value on this normalized scale of 2 in one species and −1 in another, we can conclude that there is an eightfold difference in gene expression between the species. Other methods of normalization have been proposed (Kerr *et al.* 2000; Quackenbush 2002), but we consider just this log-linear normalization strategy here. We next fit a mixed model with fixed main effects of species (human, chimp, or orangutan), tissue (brain or liver), and probe and individual within species as a random effect. Expression differences and significance were estimated for each effect, as well for each species and tissue comparison.

Several checks of data quality were performed. Figure 1 shows a "submarine" scatter plot of standardized residuals (the estimated residuals ε_{ijkl} divided by the square root of the variance of these residuals for each gene model) against predicted value. While there appear to be a large number of outliers, actually just 0.5% of the probes have standardized residuals >3. Many of these can be attributed to data saturation. Testing for the normality of the distribution of residuals for each gene-specific model indicated that as many as 39% of the genes did not reach the conservative 0.05 significance level. As discussed below, biases in the data due to probe effects may have a particularly large impact on interpretation of contrasts among species.

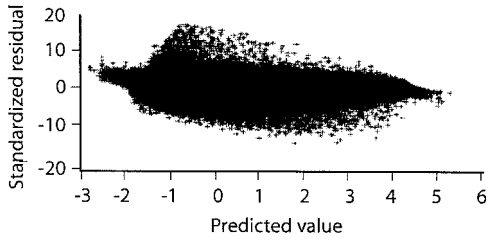

Figure 1.—Submarine plot of standardized residuals against predicted values for $\log_2$-transformed signal intensity measurements of each individual oligonucleotide in the primate data set. The shape of the plot is fairly typical for gene expression data, but asymmetry above and below the horizontal testifies to several percent of probes showing saturation or failure of hybridization.

Levels of divergence within and between species: Direct visualizations of the significance and magnitude of effects in the primate comparisons are provided by volcano plots for each pairwise species contrast and each tissue in Figure 2. Note that the main-effect estimates are averaged over and adjusted for all of the different oligonucleotide probes for each gene, and significance is assessed in the mixed model, taking into account among-probe variance. Volcano plots contrast significance on the $-\log_{10}(p)$ scale against expression difference on the $\log_2$ scale. Genes toward the left and right on each plot show a large expression difference, and those toward the top have high significance, with values of 2, 3, etc., representing P values of 10^{-2}, 10^{-3}, etc.

Two features of these plots stand out. First, the number of genes toward the top of each plot is greater for the liver than for the brain contrasts. For example, comparing human and chimp at the 5% significance level, 25% of the genes show evidence for differential transcript abundance in the brain and up to 35% in the liver, with a mean of just a 1.2-fold change in either direction. The numbers increase slightly for the contrasts involving the other species. We confirmed this observation on the human-chimp contrast using the analytical approach implemented in dChip software (Li and Wong 2002), which gave similar results (data not shown). Second, whereas the liver plots are fairly symmetrical, the brain plots are highly asymmetrical: in each case, those genes on the left-hand side of each plot are more dispersed across the range of expression differences. Since the plots were drawn with expression difference expressed as chimp minus human, orangutan minus human, and orangutan minus chimp, this means that apparently many more genes are upregulated than downregulated in the range of 2- to 4-fold in the human brain relative to the other species. Similarly, the chimp shows an apparent bias toward upregulation relative to the orangutan.

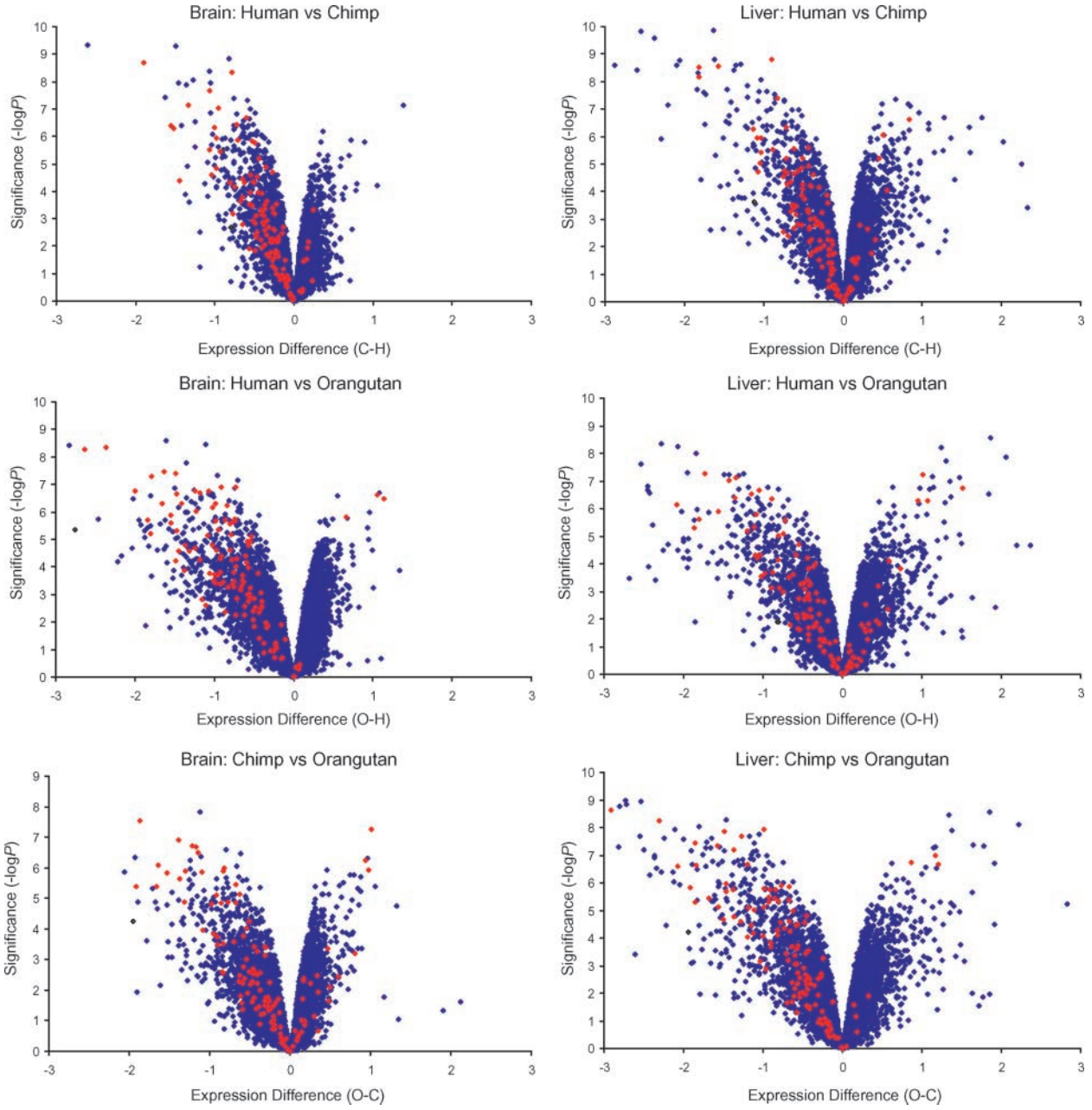

FIGURE 2.—Volcano plots of significance against fold change in expression for each primate species comparison in brain (left-hand side) and liver (right-hand side). Each point represents a single gene analyzed by mixed-model ANOVA. Highly significant values are toward the top, and small expression difference is at the center of each plot. Expression difference is plotted as difference in the least-squares means of $\log_2$-normalized expression values for chimp minus human ($C - H$), orangutan minus human ($O - H$), or orangutan minus chimp ($O - C$). The red points are the genes with the most significant (top 1%) species × probe interaction effects: these are clearly asymmetrically distributed in favor of higher apparent expression in the species expected to show the closest sequence homology to the human probes.

Assessment of the significance of expression differences is complicated by the large number of contrasts that are performed as well as the variable residual variance for each gene. If two genes have the same fold difference between species, but one has higher among-individual variance within species than the other, the significance of the species difference will be elevated for the second gene. Further, the more genes that are assessed, the more likely it is that genes exceed a low significance threshold by chance. Consequently, we present the number of genes that are significant and the associated fold increase or decrease in expression between species at three different significance levels in Table 1. These are 4×10^{-6} (the conservative Bonferroni-adjusted contrast, calculated as 0.05/12,600, and reflected in a negative $\log_{10} P$value >5.4), 0.001 ($-\log_{10}$

TABLE 1

Fraction of genes showing expression differences among primate species

Comparison	SigLevel[a]	N(Up)[b]	×Up[c]	N(Dn)[b]	×Dn[c]	×All[c]	% Up[d]	% genes[e]
				Raw data				
Brain: $H - C$	Bonferroni	86	1.75	5	1.67	1.74	95	0.7
Brain: $H - C$	0.001	522	1.41	173	1.24	1.37	75	5.5
Brain: $H - C$	0.05	1520	1.26	1734	1.12	1.18	47	25.8
Liver: $H - C$	Bonferroni	126	2.01	41	1.75	1.95	75	1.3
Liver: $H - C$	0.001	614	1.49	449	1.33	1.42	58	8.4
Liver: $H - C$	0.05	1777	1.27	2664	1.16	1.20	40	35.2
Brain: $C - O$	Bonferroni	31	2.17	11	1.71	2.04	74	0.3
Brain: $C - O$	0.001	411	1.62	352	1.27	1.44	54	6.1
Brain: $C - O$	0.05	1685	1.37	3301	1.16	1.22	34	39.6
Liver: $C - O$	Bonferroni	72	2.48	33	2.28	2.41	69	0.8
Liver: $C - O$	0.001	528	1.72	369	1.46	1.61	59	7.1
Liver: $C - O$	0.05	1586	1.39	2184	1.23	1.29	42	29.9
Brain: $H - O$	Bonferroni	91	2.27	13	1.62	2.17	88	0.8
Brain: $H - O$	0.001	772	1.71	823	1.23	1.44	48	12.7
Brain: $H - O$	0.05	2120	1.44	4466	1.17	1.25	32	52.3
Liver: $H - O$	Bonferroni	139	2.68	31	2.19	2.57	82	1.3
Liver: $H - O$	0.001	647	1.82	407	1.44	1.66	61	8.4
Liver: $H - O$	0.05	1608	1.48	2334	1.21	1.31	41	31.3
				Filtered data				
Brain: $H - C$	Bonferroni	37	2.08	5	1.54	2.01	88	0.3
Brain: $H - C$	0.001	193	1.48	158	1.24	1.37	55	2.8
Brain: $H - C$	0.05	854	1.24	1878	1.12	1.16	31	21.7

[a] Significance levels: Bonferroni = $-\log P > 5.4$; 0.001 = $-\log P > 3.0$; 0.05 = $-\log P > 1.301$.
[b] Number of genes up- or downregulated at the indicated significance level.
[c] Magnitude of fold change up (greater in left-hand species) or down (opposite direction) based on the raw (unfiltered) data.
[d] Percentage of genes that are significantly differentially expressed that are upregulated.
[e] Percentage of all genes on the microarrays that are differentially expressed.

$P > 3.0$) and 0.05 ($-\log_{10} P > 1.301$). We also present the average expression difference for both up- and downregulated genes, the percentage of genes that are apparently upregulated, and the percentage of all genes that are differentially expressed for each contrast.

The murine data set consisted of 14 Affymetrix GeneChips, each containing up to 20 independent oligonucleotide probes for each of ~12,488 genes derived from *M. musculus* sequences. *M. musculus* and *M. spretus* were each represented by three individuals, with a single hybridization for each of the two tissues (hence six arrays each), while *M. caroli* was represented by a single individual (two arrays). We analyzed the data according to the same model as for the primates. The three data quality checks indicated that the data were slightly more favorable for analysis of variance. Only 0.3% of the data points had standardized residuals >3, while 86% of the genes passed the normality test for residuals from the mixed model. However, since there were no replicates of each individual, significance tests are not as powerful as for the primate data. Nevertheless, the overall nature of the analyses is remarkably similar, as documented in Figure 3 and Table 2. Between the two most closely related species, *M. musculus* and *M. spretus*, ~10% of the genes showed significantly different transcript abundance at the 5% significance level, with an average of almost 1.3-fold change in either direction for both brain and liver. The same biases toward greater divergence in the liver and asymmetric upregulation in the brain favoring *M. musculus* over *M. spretus* over *M. caroli* are observed, though not as strongly as for the primate data.

From both Tables 1 and 2, it can be seen that the fraction of genes that appear to be upregulated (that is, expression is greater in species A than in species B) is consistently reduced as the significance level is relaxed (for example, from 95 to 47% for the human-chimp brain contrast). This implies that there is a systematic tendency for overestimation of the expression level for genes in the order human > chimp > orangutan (or underestimation in the opposite order). A similar tendency was observed in the murine data set (*M. musculus* > *M. spretus* > *M. caroli*), and in all cases the consequent apparent bias toward upregulation is observed in the species genetically closest to *M. musculus*, from which the probe sequences derive.

Probe effects in the context of genetic divergence: This suggests the hypothesis that apparent upregulation is due to stronger hybridization to individuals of one

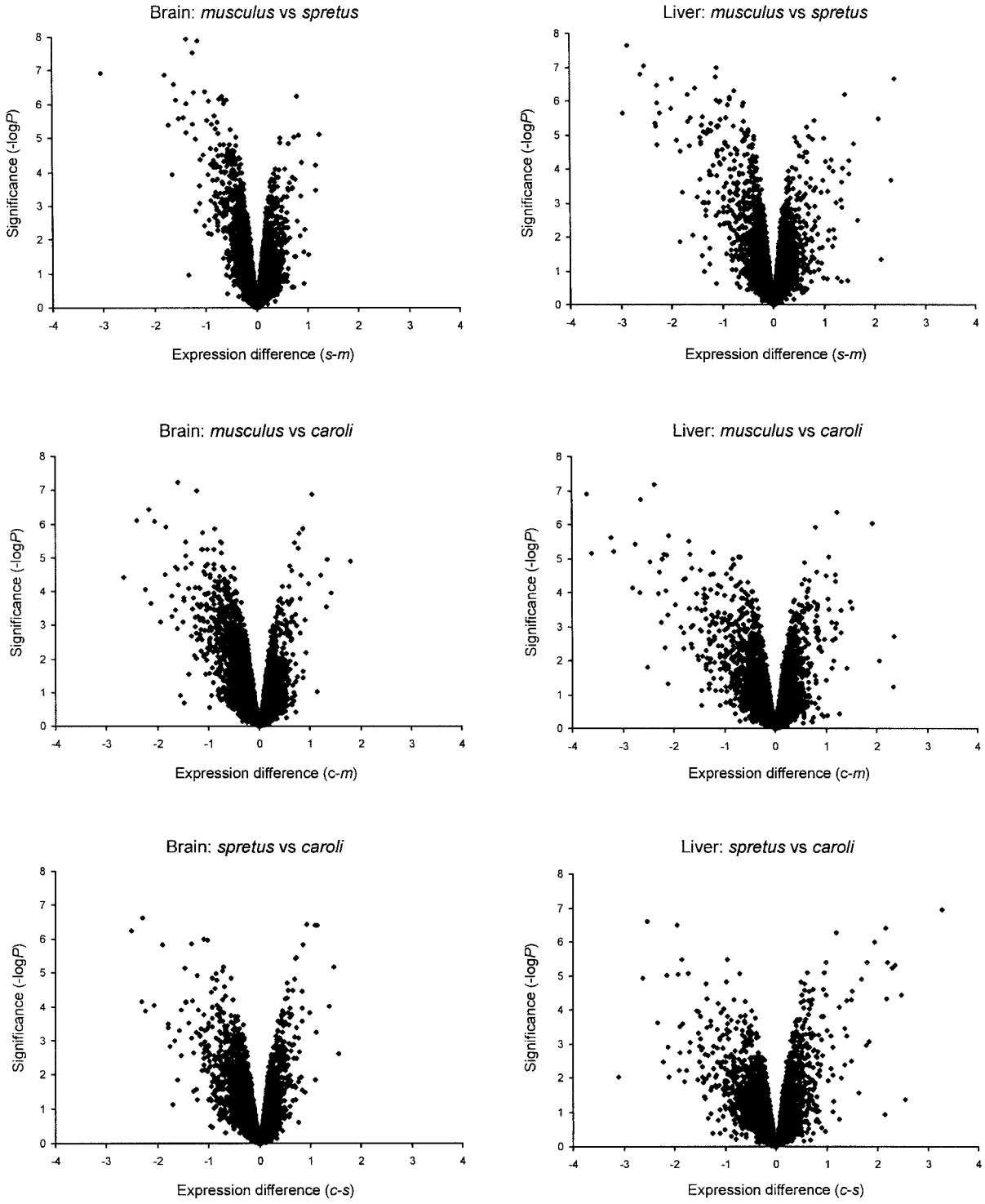

FIGURE 3.—Volcano plots of significance against fold change in expression for each murine species comparison in brain and liver. Layout is essentially the same as in Figure 2. Species comparisons are *Mus spretus* minus *M. musculus* ($s - m$), *M. caroli* minus *M. musculus* ($c - m$), and *M. caroli* minus *M. spretus* ($c - s$).

species over another. At a genome-wide rate of sequence divergence of 1%, if the probes were nonoverlapping, then only one-quarter of them should have any nucleotide differences between species, and only a fraction of these would be near the center of the probe where they are most likely to affect hybridization. Nevertheless, small differences in 2 or 3 probes out of 20 could be sufficient to yield an apparent upregulation of ~1.2-fold. It is also noteworthy that the estimated magnitude of downregulation is always less than the estimated mag-

TABLE 2

Fraction of genes showing expression differences among murine species

Comparison	SigLevel[a]	N(Up)[b]	×Up[c]	N(Dn)[b]	×Dn[c]	×All[c]	% Up[d]	% genes[e]
				Raw data				
Brain: $M - S$	Bonferroni	23	2.27	1	1.74	2.25	96	0.2
Brain: $M - S$	0.001	190	1.53	53	1.41	1.49	78	1.9
Brain: $M - S$	0.05	767	1.28	534	1.21	1.25	59	10.4
Liver: $M - S$	Bonferroni	27	2.81	4	3.23	2.87	87	0.2
Liver: $M - S$	0.001	186	1.80	64	1.72	1.78	74	2.0
Liver: $M - S$	0.05	738	1.38	812	1.23	1.30	48	12.4
Brain: $S - C$	Bonferroni	6	3.23	6	1.88	2.46	50	0.1
Brain: $S - C$	0.001	112	1.82	48	1.55	1.73	70	1.3
Brain: $S - C$	0.05	698	1.41	555	1.23	1.33	56	10.0
Liver: $S - C$	Bonferroni	5	4.89	5	4.47	4.69	50	0.1
Liver: $S - C$	0.001	72	2.08	80	1.84	1.95	47	1.2
Liver: $S - C$	0.05	582	1.51	497	1.37	1.44	54	8.6
Brain: $M - C$	Bonferroni	12	3.20	4	1.80	2.77	75	0.1
Brain: $M - C$	0.001	202	1.83	42	1.61	1.79	83	2.0
Brain: $M - C$	0.05	919	1.44	672	1.21	1.34	58	12.7
Liver: $M - C$	Bonferroni	7	6.23	3	2.50	4.72	70	0.1
Liver: $M - C$	0.001	127	2.19	58	1.64	2.00	69	1.5
Liver: $M - C$	0.05	646	1.56	538	1.29	1.43	55	9.5

[a] Significance levels: Bonferroni = $-\log P > 5.4$; 0.001 = $-\log P > 3.0$; 0.05 = $-\log P > 1.301$.
[b] Number of genes up- or downregulated at the indicated significance level.
[c] Magnitude of fold change up (greater in left-hand species) or down (opposite direction), based on the raw (unfiltered) data.
[d] Percentage of genes that are significantly differentially expressed that are upregulated.
[e] Percentage of all genes on the microarrays that are differentially expressed.

nitude of upregulation at the same significance level (hence, the absolute value of the fold change is always less than the magnitude of upregulation). This is consistent with the idea that reduced hybridization to a few probes in the divergent species contributes to apparent upregulation.

Significance levels are affected by a balance between the fold change averaged across probes (tending to make more genes appear to be upregulated) and the increase in among-probe variance due to sequence divergence for some probes (tending to reduce the significance of contrasts). We thus asked whether the species-by-probe interaction effect in the mixed model for each gene is more likely to be significant for upregulated genes. This effect is small in magnitude, but it is significant for more than half of the genes (see supplementary information at http://www.genetics.org/supplemental/). The red points in the volcano plots in Figure 2 indicate the genes with the top 1% of the most significant species-by-probe interaction effects, and these are almost all apparently upregulated. This result is consistent with the hypothesis that the overwhelming bias toward apparent upregulation in the brain in the phylogenetically closest species, which is expected to show the least sequence divergence, might be attributed to loss of hybridization to a subset of probes.

To further explore whether this is the case, we next examined the actual profiles of fluorescence intensity for representative genes. Figure 4 shows plots of relative fluorescence intensity for human and chimp brain arrays for each probe for a set of six representative genes. The order and spacing of probes along the abscissa is proportional to the number of bases offset along the gene sequence for each probe. Human intensity values are indicated as large open diamonds, and chimp values as small solid boxes. Gene A is an example of a "well-behaved" probe set: despite absolute differences in intensity for each probe, all probes indicate a similar magnitude of upregulation in the human relative to chimp. Gene B by contrast is "poorly behaved" in so far as each probe predicts a different magnitude for the species difference. Gene C is an example of a locus where a single probe that shows much-reduced hybridization to chimp cDNA (the far right probe) would be sufficient to suggest an overall 1.2-fold upregulation in humans relative to chimps. This situation was also occasionally seen in the reverse direction (one probe gives a stronger chimp signal) as shown for gene D. However, many of the cases of strong species-by-probe interaction effects involved multiple probes, as seen for genes E and F. *LAMP1* is apparently upregulated in humans, but only one-half of the probes showed the difference, and all of these eight probes overlap with their 5′-most nucleotides separated by just 14 bases. The

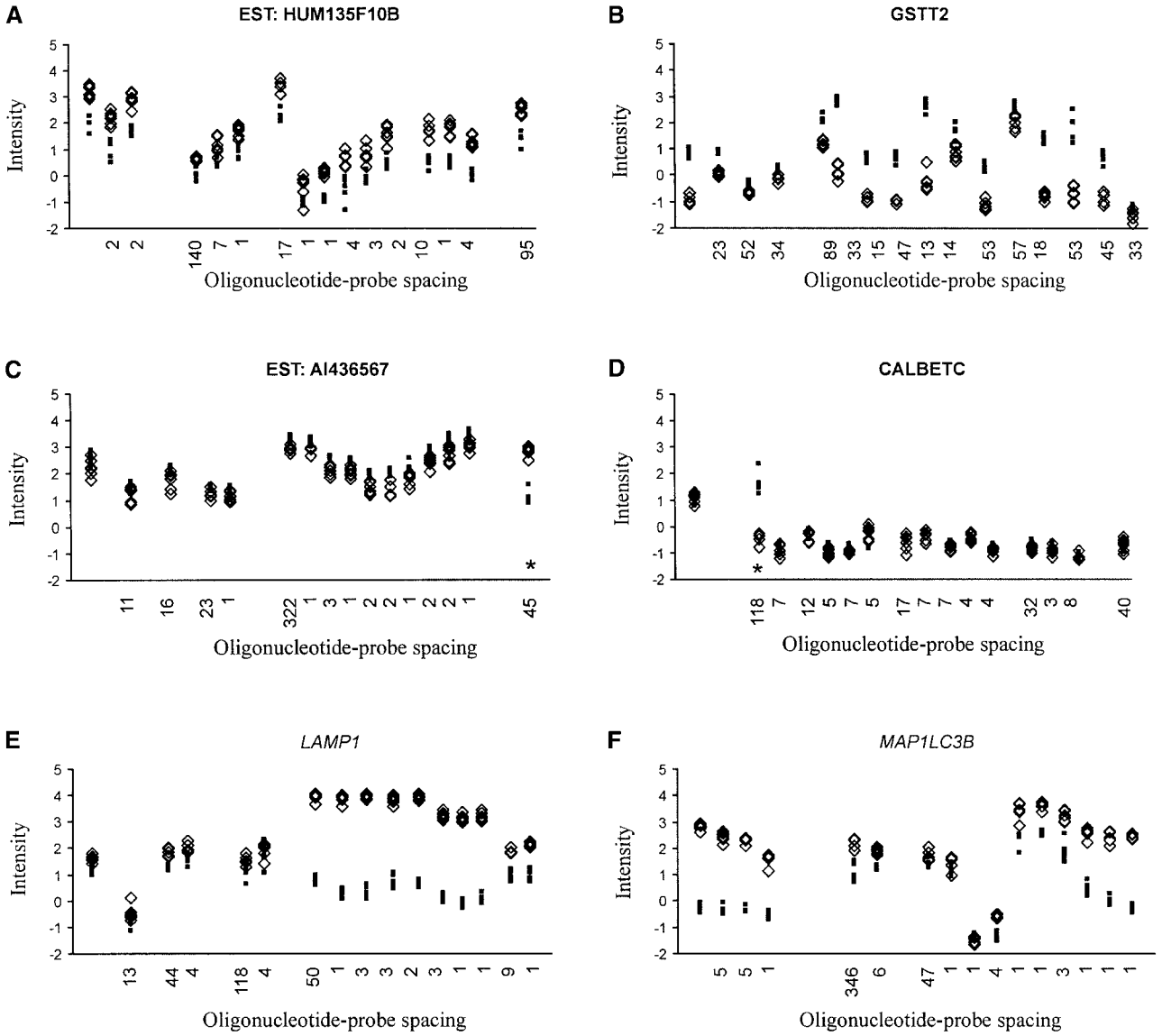

FIGURE 4.—Parallel plots of individual oligonucleotide measurements for human and chimp brain samples for six genes. Each of the 16 oligonucleotides is plotted in proportion to the spacing between first nucleotides from 5′ to 3′: numbers below each plot show the number of nucleotides between these sites. Thus a spacing of 1 represents oligonucleotide probes that overlap by 24 of 25 bases, while a spacing of 45 represents nonoverlapping probes. Normalized $\log_2$ expression levels for the perfect-match probe on the *y*-axis are shown as open diamonds (human) and solid squares (chimp). Gene or expressed sequence tag names correspond to GenBank accessions D54318, L38503, AI36567, M92302, J04182, and W28807 for A–F, respectively. (A) A well-behaved gene with similar differences between species for each probe; (B) poorly behaved gene with variable differences; (C and D) genes where an overall expression difference is contributed almost entirely by a single probe indicated by the asterisks; (E and F) genes where two classes of expression difference, largely but not completely corresponding to overlapping probes, are observed.

next two probes, just 9 and 10 bases farther toward the 3′, show much reduced species difference. *MAP1LC3B* gave a similar result, except that the species difference was seen in two nonoverlapping sets of probes. It is sobering in this case that even probes that overlap by all but one nucleotide give 10-fold differences in signal intensity for both species, and severalfold differences between species.

In an attempt to filter out the probe-by-species interactions, we imposed a constraint that genes should be included in the analysis only if the correlation between human and chimp fluorescence intensity exceeded 0.95. So as to include all genes, we wrote a script to systematically remove outlier probes for each gene until this condition was met. Typically this meant removal of just two to five probes per gene, but more than half of the genes showed the high correlation without removing any probes. A plot of the expression difference be-

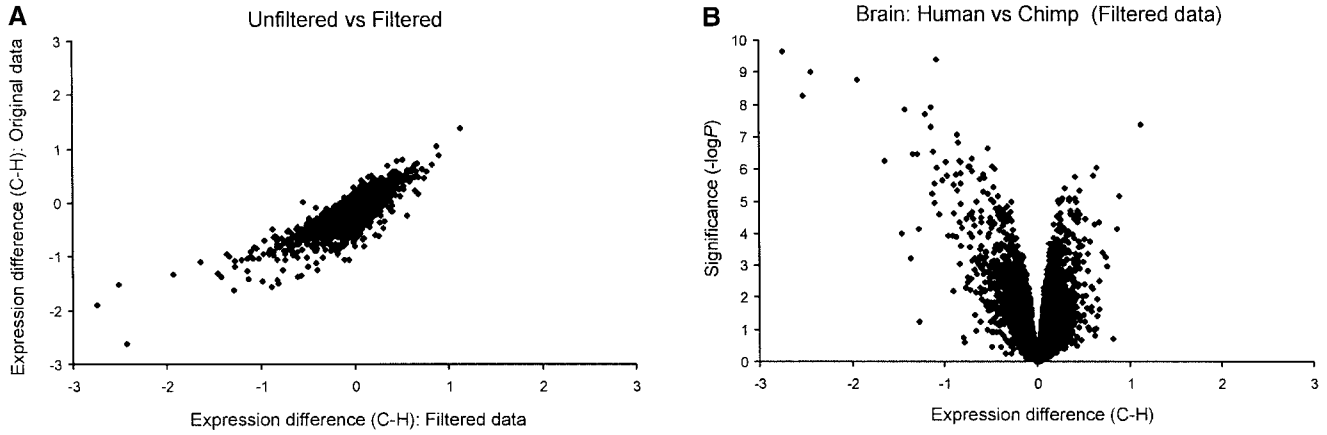

FIGURE 5.—Effect of filtering outliers on inference of expression difference between human and chimp brain. (A) Subtraction of human from chimp expression value tends to produce more negative values on the original data than on the filtered data: most points lie at or below an imagined diagonal line running through points for which filtering has no effect. (B) Volcano plot after filtering: compared with the top left plot in Figure 2, this plot is considerably more symmetric, due to removal of probes that contribute to the large species × interaction effect. Both plots are for just the brain data.

fore filtering against after filtering in Figure 5A shows many more points below the diagonal than above, indicating that the effect of filtering is typically to reduce the magnitude of the apparent upregulation in human brains, as expected. However, the volcano plot for the human *vs.* chimp brain comparison in Figure 5B remains somewhat asymmetric, and the overall tendency for more genes to be differentially expressed in the liver than in the brain when comparing human and chimp is still apparent (see also Table 1 and supplementary information).

DISCUSSION

Possible biases in oligonucleotide expression data: Our mixed-model analyses of the primate and murine gene expression data lead to conclusions that are not necessarily consistent with those reported by the original authors (ENARD *et al.* 2002) in so far as there is little evidence for accelerated divergence in gene expression in the human brain. Whichever method of analysis is used, the interpretation should be tempered to some extent by our finding of potential species-specific biases in the magnitude of inferred transcript abundance. Since in all cases more genes were seen to be upregulated in the species that is closest to the one whose sequence was used to generate the probes (that is, *Homo sapiens* or *M. musculus*), the most straightforward explanation is that this bias reflects differential hybridization due to loss of perfect sequence matching.

However, three lines of evidence lead us to question this explanation. The first is that detailed analysis of numerous genes that showed a species-by-probe interaction effect (that is, variable differences in transcript abundance among probes within a gene) indicated a complex relationship between sequence and signal. Overlaying the mismatch probe data on the perfect match data does not help at all as it just increases the noisiness of the results (data not shown; many mismatches hybridize as strongly as the match and the difference between match and mismatch also varies greatly by probe within each gene). Many factors, presumably including amount of cross-hybridization, alternative splicing, and sequence divergence, must contribute to probe effects, and it is not obvious how to deal with these statistically. The fact that Affymetrix's probe selection algorithm tends to choose clusters of sequences that differ by just a few bases also introduces a correlation structure to the data that formally but impractically should be dealt with on a gene-by-gene basis. We and others (CHU *et al.* 2002; LI and WONG 2002; SASIK *et al.* 2002) have demonstrated that modeling gene expression profiles by probe within gene is generally much more accurate than using the average difference measure, but it is also clear that genotypic differences can affect the results in ways that are difficult to control.

The second line of evidence arguing against sequence divergence accounting for all of the biases toward upregulation is that the effect appears to be much greater in the brain samples than in the liver. This could imply that brain proteins are diverging at a faster rate than liver proteins. Comparative sequence analyses will soon resolve this issue. ENARD *et al.* (2002) also provided two-dimensional gel electrophoresis evidence for divergence in protein sequence and abundance between human and chimp brain, but it is not yet possible to assess whether differential sequence divergence is responsible for at least some of the apparent upregulation of a large number of human genes. The third line of evidence is that the upward bias is observed only for the 10–20% of genes that show the most significant divergence in gene expression. Below the 5% significance level there

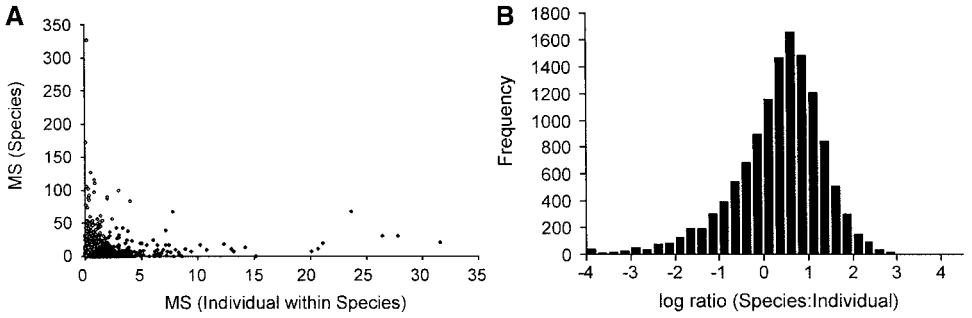

FIGURE 6.—Contrast of contributions of species and individual within species to expression variance between human and chimp brain. (A) Plot of mean square from a general linear ANOVA for the species and individual within-species terms for each gene. Open diamonds toward the left show genes with a significant *F*-ratio, indicating significant divergence between species relative to variation within. Note the large number of genes (solid diamonds, mostly toward the right of the plot) with much greater variation within than between species. (B) Histogram of frequency of $\log_{10}$ ratio of mean-square species:mean-square individual within-species values from A. Only a few genes have a ratio approaching 1000 (that is, three on the log scale), whereas almost 25% of the genes have a ratio >10.

are essentially equivalent numbers of genes up- and downregulated in each comparison. Attempts to filter out the largest probe-by-specific effects had little impact on the overall conclusions, arguing that many of the observed differences in gene expression are real and that there may in fact be a biological basis to the tendency toward increased gene expression in humans over chimps and orangutan. Whether this relates to increased size and/or complexity of the brain remains to be seen.

Divergent gene expression among primates: Our reanalysis of ENARD *et al.*'s (2002) data on a gene-by-gene basis is in broad agreement with their analysis based on whole-transcriptome variation in several respects, but also allows quantification of the fraction of genes that contribute to within- and between-species differences. Both analyses indicate that there are significant differences in gene expression among species that are of a greater magnitude than the differences among individuals within a species and that there is a general increase in degree of transcriptional divergence as sequence (and hence temporal) divergence increases. As pointed out by ENARD *et al.* (2002), conclusions concerning relative rates of divergence are, however, quite sensitive to the metrics used.

Mixed-model analysis provides formal statistical support for 51 genes being differentially expressed between human and chimp Brodmann's area 9 after filtering outlier probes and just under twice this number if raw data is used. At the less conservative significance threshold of 0.001, 482 genes are differentially expressed with an average almost 1.4-fold change between human and chimp brain, compared with a chance expectation of just 13 genes at this level. Based on the raw data, this number increases to 695 genes and to 1595 genes when human is compared to orangutan. For the liver, 1063 genes are differentially expressed between human and chimp, also with an average 1.4-fold change, and 1054 genes between human and orangutan at a slightly higher mean fold change of 1.6. The chimp-orangutan comparisons are intermediate, with slightly more genes differentially expressed with a larger fold change in the liver than in the brain.

Most of our comparisons of species and tissue pairs suggest then that more genes are divergently expressed in the liver than in the brain and that the magnitude of change also tends to be greater in the liver. While it is clear that dramatic cognitive changes have occurred particularly in the human lineage, it also not surprising that transcription has evolved greatly in the liver, given the differences in diet and culture of the primate species. A possible reconciliation of our findings with the inference favored by ENARD *et al.* (2002) that "changes of gene expression in the brain may have been especially pronounced during recent human evolution" is the suggestion that much of that change has occurred on the human-orangutan axis. We also observe a relatively large branch length between all humans and the central node on neighbor-joining trees on the basis of transcriptome-wide average expression differences at each level of significance (see supplementary Figure 2 at http://www.genetics.org/supplemental/ and GU and GU 2003). It is noteworthy, though, that the relative length of this node, as well as the divergence of the second human individual from the others, is very much a function of the number of genes included in the analysis.

The nature of the differentially expressed genes is also of interest. Those that are significantly divergent between human and chimp brain and liver are tabulated in supplementary Figure 3 (http://www.genetics.org/supplemental/). A number of neuronal genes such as neurotransmitter receptors and channels are obvious in the brain list, as are detoxification enzymes such as cytochrome P450s on the liver list. However, the majority of genes have more general potential functions in regulation of cell growth and division and cell structure: members of most of the major gene ontology categories are represented in both lists.

Finally, we can also ask whether the divergence in gene expression is more likely attributable to drift or diversifying selection. A significantly elevated measure of divergence in expression between species, relative to

the observed level of among-individual within-species variation, is not *prima facie* evidence for selection. Figure 6A shows that the most significant genes in the human *vs.* chimp brain comparison both diverge between species and have relatively low levels of intraspecific variance. For a large number of genes this relationship is reversed. In fact, a histogram of the log ratios of the mean squares for the species and individual within-species components is slightly skewed toward low ratios, suggesting that many genes may be more variable within species than expected. Rifkin *et al.* (2003) have recently proposed, following Lynch and Hill's (1986) approach for phenotypic traits, that the expected degree of divergence under mutation-drift equilibrium can be formulated by scaling the ratio of mean squares for divergence and polymorphism by the ratio of twice the effective population size over the number of generations since divergence. Assuming a small effective population size for humans of 10,000 individuals and a mean generation time of 15 years over the 6–7 million years since the human and chimp lineages diverged (Brunet *et al.* 2002), the expected distribution of ratios for this data set is 20 to 23 times larger than the $F_{1,2}$ distribution. Consequently, only genes with a divergence to polymorphism ratio >800, a handful of just 15–20 genes in our analysis, clearly lie in the upper 2.5% tails of the expected level of divergence for these population parameters. Relaxation of these conservative assumptions provides suggestive evidence that 5% or more of the genes may be experiencing diversifying selection. Clearly more individuals need to be sampled at different ages and for more targeted tissue samples, but comparison of gene expression divergence, coincident with assessment of nucleotide sequence divergence, is a promising approach to identification of genes that may have contributed to human cognitive evolution.

We thank Svante Pääbo for correspondence and comments on the analysis of data generated in his laboratory and Bruce Weir, Zhao-Bang Zeng, Spencer Muse, and Maryellen Ruvolo for discussions. W.P.H. is supported by a genome science graduate fellowship from the Bioinformatics Program at North Carolina State University, and research in G.G.'s laboratory is supported in part by National Institutes of Health grant PO1-GM45344.

LITERATURE CITED

Brunet, M., F. Guy, D. Pilbeam, H. T. Mackaye, A. Likius *et al.*, 2002 A new hominid from the Upper Miocene of Chad, Central Africa. Nature **418:** 145–151.

Cavalieri, D., J. P. Townsend and D. L. Hartl, 2000 Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc. Natl. Acad. Sci. USA **97:** 12369–12374.

Cheung, V. G., and R. S. Spielman, 2002 The genetics of variation in gene expression. Nat. Genet. **32** (Suppl.): 522–525.

Chu, T.-M., B. Weir and R. D. Wolfinger, 2002 A systematic statistical linear modeling approach to oligonucleotide array experiments. Math. Biosci. **176:** 35–51.

Enard, W., P. Khaitovich, J. Klose, S. Zoellner, F. Heissig *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340–343.

Felsenstein, J., 1989 PHYLIP—phylogeny inference package (version 3.2). Cladistics **5:** 164–166.

Gu, J., and X. Gu, 2003 Induced gene expression in human brain after the split from chimpanzee. Trends Genet. **19:** 63–65.

Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat. Genet. **29:** 389–395.

Kerr, M. K., M. Martin and G. A. Churchill, 2000 Analysis of variance for gene expression microarray data. J. Comput. Biol. **7:** 819–837.

Li, C., and W. Wong, 2002 Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA **98:** 31–36.

Lynch, M., and W. G. Hill, 1986 Phenotypic evolution by neutral mutation. Evolution **40:** 915–935.

Oleksiak, M. F., G. A. Churchill and D. L. Crawford, 2002 Variation in gene expression within and among natural populations. Nat. Genet. **32:** 261–266.

Quackenbush, J., 2002 Microarray data normalization and transformation. Nat. Genet. **32** (Suppl.): 496–501.

Rifkin, S. A., J. Kim and K. P. White, 2003 Evolution of gene expression in the Drosophila melanogaster subgroup. Nat. Genet. **33:** 138–144.

Sasik, R., E. Calvo and J. Corbeil, 2002 Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. Bioinformatics **18:** 1633–1640.

Wolfinger, R. D., G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh *et al.*, 2001 Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8:** 625–637.

Communicating editor: M. Feldman