

LETTER

Assessing the Conservation of Mammalian Gene Expression Using High-Density Exon Arrays

Yi Xing,^{*†} Zhengqing Ouyang,[‡] Karen Kapur,[†] Matthew P. Scott,^{§||} ¶ and Wing Hung Wong[†]

^{*}Department of Internal Medicine, Roy J. and Lucille A. Carver College of Medicine, University of Iowa; [†]Department of Statistics, Stanford University; [‡]Department of Biological Sciences, Stanford University; [§]Department of Developmental Biology, Stanford University School of Medicine; ^{||}Department of Genetics, Stanford University School of Medicine; [¶]Department of Bioengineering, Stanford University School of Medicine; and [#]Howard Hughes Medical Institute, Chevy Chase, Maryland

Microarray data from multiple species have been used to study evolutionary constraints on gene expression. Expression measurements from conventional microarray platforms such as the 3' expression arrays are strongly affected by platform-dependent probe effects that may introduce apparent but misleading discrepancies between species. In this manuscript, we assess the conservation of mammalian gene expression in adult tissues using data from a high-density exon array platform. The exon arrays have more than 6 million probes on a single array targeting all exons in a genome. We find that, unlike 3' array data, gene expression measurements from exon arrays reveal patterns of gene expression that are highly conserved between humans and mice in multiple tissues. Our analysis provides strong evidence for widespread stabilizing selection pressure on transcript abundance during mammalian evolution.

Introduction

There has been strong interest in characterizing the selection pressure on gene expression using high-throughput genomic data such as microarray expression profiles (Gilad et al. 2006; Khaitovich et al. 2006). A central question is whether tissue or organ-specific patterns of gene transcription tend to be preserved over evolutionary time. Recent studies have generated contradictory results (Khaitovich et al. 2004; Yanai et al. 2004; Jordan et al. 2005; Liao and Zhang 2006). By comparing expression data from Affymetrix 3' array technology for human and mouse adult tissues (Su et al. 2004), one study concludes that the evolution of mammalian gene expression is largely without selective constraints (Yanai et al. 2004), whereas others find evidence for stabilizing selection pressure (Jordan et al. 2005; Liao and Zhang 2006). Unfortunately, conventional expression microarray platforms, such as Affymetrix 3' expression arrays, have inherent limitations for comparative genomics analyses because of platform-dependent probe effects (Irizarry et al. 2005). Because of variations in probe affinity, different microarray probes detecting the same transcript can show very different intensities (Li and Wong 2001). Because 3' expression microarrays use a small number of probes for each gene's 3' end, typically 11, probe effects can significantly affect the estimated expression indexes. Therefore, it is misleading to directly compare absolute expression estimates between human and mouse 3' arrays, which have completely independent probe designs for orthologous genes (Irizarry et al. 2005; Liao and Zhang 2006).

In this letter, we reassess the conservation of gene expression levels between human and mouse tissues, using data from a new microarray platform—the Affymetrix Exon Arrays. Exon Arrays have over 6 million probes targeting annotated and predicted exons in a genome (Affymetrix 2005a). Most exons are targeted by at least 4 probes. In well-annotated human genes with RefSeq mRNAs, the aver-

age number of probes is 147, including an average of 58 “core probes” per gene that target high-confidence (i.e., RefSeq supported) exon annotations. Although Exon Arrays were designed for genome-wide analyses of alternative RNA splicing, the high density and even spacing of exon array probes for each gene also enables accurate measures of overall gene expression levels (Xing et al. 2006).

We used a public Exon Array data set for 6 adult human tissues and their corresponding adult mouse tissues, each with 3 replicates (see Methods). We calculated gene expression indexes for 10,480 pairs of orthologs using our GeneBase program (see details in Methods). For a comparison to 3' expression array data, we also obtained estimated expression indexes for 11,580 pairs of orthologs in 6 tissues from the Novartis Gene Expression Atlas (see Methods).

The analysis of 3' array data and of Exon Array data leads to strikingly different results about the conservation of mammalian gene expression. Consistent with previous analyses (Yanai et al. 2004; Jordan et al. 2005), expression indexes derived from 3' arrays were indeed poorly correlated between corresponding human and mouse tissues. For example, the correlation between human testis and mouse testis was ~ 0.4 (see Fig 1A; Spearman rank correlation, 0.37; Pearson correlation, 0.42). The correlation was also low for other tissues, such as muscle (Fig 1B) and liver (Fig 1C). By contrast, Exon Array expression indexes were highly correlated between human and mouse in testis, muscle, and liver (see Fig 1D–F). The correlation between human and mouse testis was ~ 0.7 (Spearman rank correlation, 0.69; Pearson correlation, 0.68). We observed a similarly high level of correlation in Exon Array expression indexes of orthologous genes in kidney, spleen, and heart (data not shown). Thus, unlike 3' expression arrays, exon arrays show highly correlated expression levels for orthologous genes in corresponding human and mouse tissues, suggesting a strong stabilizing selective pressure on transcript abundance.

Next, we assess whether the overall gene expression profiles across multiple tissues are significantly more similar for human–mouse orthologs than what would be expected by random chance. Do orthologous genes share similar tissue-specific expression patterns? To measure the similarity of expression profiles between species, for each ortholog pair we calculated

Key words: evolution, gene expression, selection pressure, microarray, exon array.

E-mails: yi-xing@uiowa.edu; whwong@stanford.edu.

Mol. Biol. Evol. 24(6):1283–1285. 2007

doi:10.1093/molbev/msm061

Advance Access publication March 25, 2007

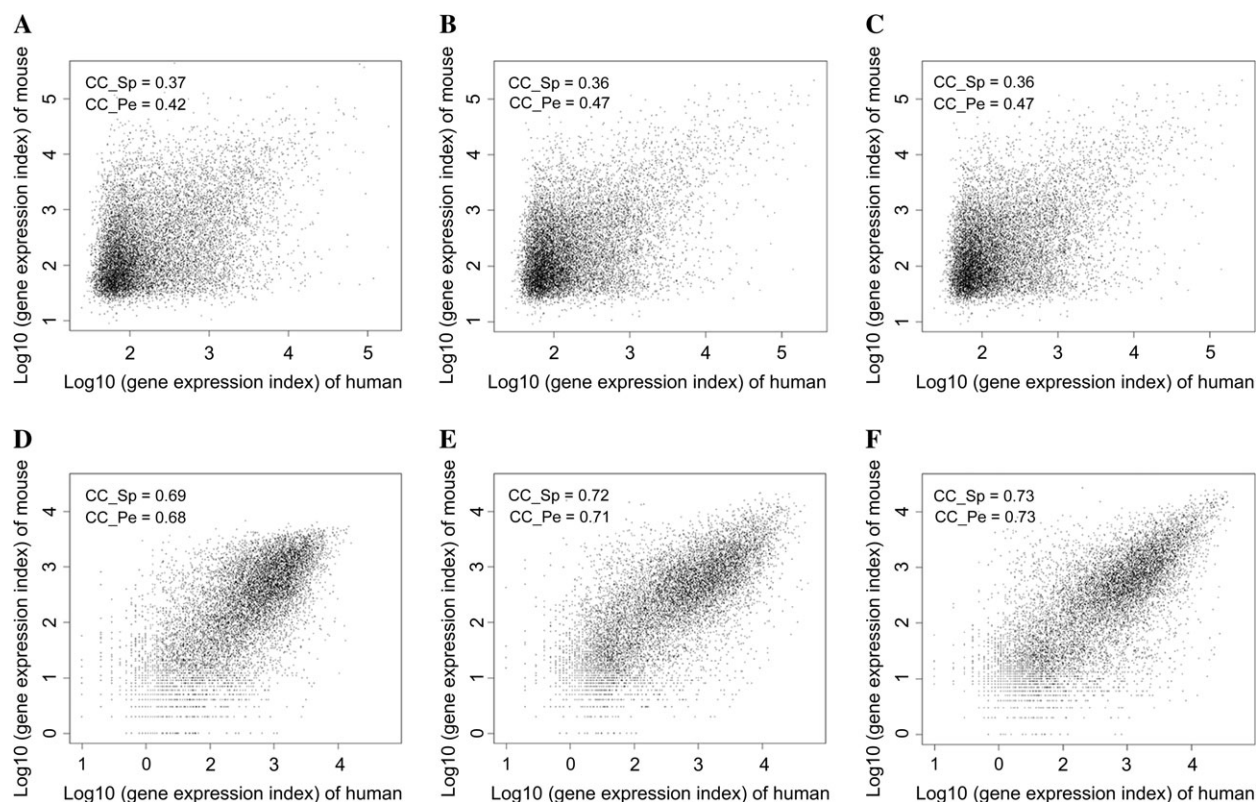


FIG. 1.—Correlations of gene expression indexes between corresponding human and mouse tissues. Top row shows 3' array data: (A) testis; (B) muscle; and (C) liver. Bottom row shows Exon Array data: (D) testis; (E) muscle; and (F) liver. The x axis shows gene expression indexes in human, and y axis shows gene expression indexes in mouse. CC_Sp: Spearman's rank correlation coefficient. CC_Pe: Pearson's correlation coefficient.

the Pearson correlation coefficient (PCC) of expression indexes over 6 corresponding tissues. We used random human–mouse gene pairs to approximate the neutral rate of expression profile divergence (Jordan et al. 2005; Liao and Zhang 2006). As expected, the expression profiles of random human–mouse gene pairs had no similarity. The median PCC of random gene pairs is -0.03 on 3' arrays and 0.02 on exon arrays (see dashed lines in Fig 2). For true human–mouse orthologs, the expression indexes derived from 3' expression arrays over 6 tissues also showed a relatively low similarity, with a median PCC of 0.18 (see Fig 2A), similar to the observation by Liao and Zhang (Liao and Zhang 2006). By contrast, when we used expression indexes derived from Exon Arrays, the distribution of PCC was significantly shifted (Fig 2B). The median PCC was 0.54 in these 6 tissues. Therefore, Exon Arrays indicate that overall expression profiles across multiple tissues (i.e., tissue-specific expression patterns) are strongly conserved.

Our study provides direct evidence for widespread stabilizing selection pressure on gene expression during mammalian evolution. Unlike 3' expression arrays, Exon Arrays indicate strong conservation of both absolute transcript abundance in individual tissues and relative transcript abundance across different tissues. Our analysis also demonstrates the power of high-density Exon Array technology, in particular for evolutionary studies of gene expression.

Methods

We downloaded the Affymetrix Exon Array tissue-panel data from Affymetrix (<http://www.affymetrix.com/>

[support/technical/sample_data/exon_array_data.affx](http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx)). We calculated gene expression indexes using our GeneBASE program for Exon Array analysis (<http://biogibbs.stanford.edu/~kkapur/genebase/>). Briefly, raw probe intensities were background adjusted using the MAT (model-based analysis of tiling-arrays) model (Johnson et al. 2006) trained from Exon Array background probes. We used a probe selection algorithm (Xing et al. 2006) to calculate expression indexes from a subset of Exon Array probes of each gene, excluding probes targeting alternative exons and putative exon predictions, as well as low-affinity or cross-hybridizing probes. We also calculated expression indexes using the PLIER algorithm (Affymetrix 2005b) of the Affymetrix Exon Array Computation Tool and obtained similar results in the subsequent human–mouse comparison (data not shown). Six tissues (heart, kidney, liver, muscle, spleen, and testis) present on both human and mouse Exon Array tissue panels were included in our comparative analyses. The Exon Array expression indexes are available as online supplementary data on the *Molecular Biology and Evolution* Web site. Additional annotations can be downloaded from <http://biogibbs.stanford.edu/~yxing/MBE/>. For the 3' array data, we used the Novartis Gene Expression Atlas (<http://wombat.gnf.org/>), in which the expression indexes were computed by the background-adjusted robust multiarray analysis (Wu and Irizarry 2005). The data sets were based on the combination of the Affymetrix HG_U133A chip and the GNF1H chip for human and the GNF1M chip for mouse (Su et al. 2004). We mapped the probe sets in the HG_U133A chip to their corresponding

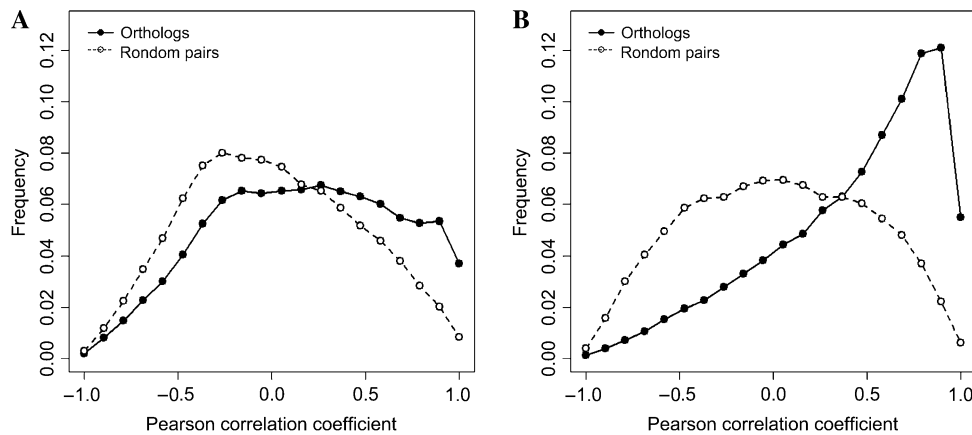


FIG. 2.—Similarity of gene expression profiles in 6 human tissues and 6 corresponding mouse tissues. For each ortholog pair, we calculated the PCC of expression indexes in 6 tissues (solid line). We also permuted ortholog relationships and calculated the PCC for random human–mouse gene pairs (dashed line). (A) PCC distribution based on 3′ array data. (B) PCC distribution based on Exon Array data.

human genes using the UCSC KnownGene mapping (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownToU133.txt.gz>). We selected 6 tissues shared by the human and mouse 3′ array data sets: heart, kidney, liver, muscle (listed as skeletal muscle), lung, and testis. For genes with multiple 3′ probe sets, we randomly selected a probe set following the procedure of Liao and Zhang 2006. We also selected the probe set with the highest gene expression index, and the results were similar. Orthologous genes between human and mouse on Exon Arrays and 3′ arrays were extracted from the current version (Build 53) of the HomoloGene database (<ftp://ftp.ncbi.nih.gov/pub/HomoloGene/>) (Wheeler et al. 2007).

For each tissue, we calculated the Spearman rank correlation and Pearson correlation of expression indexes (\log_{10} transformed) between human and mouse orthologs using R (<http://www.r-project.org>). To compare expression profiles in multiple tissues, for each ortholog pair, we calculated the PCC of expression indexes in 6 tissues. We also permuted ortholog relationships and calculated the PCC of random human–mouse gene pairs, following a procedure by Jordan et al. (2005).

Supplementary Material

Supplementary Tables 1–4 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We wish to thank Rafael Irizarry for discussions and Yan Zhang and Sam Stingley for assistance. This work was supported by National Science Foundation grant DMS0505732 and National Institutes for Health grant R01HG002341. M.P.S. is an Investigator of the Howard Hughes Medical Institute.

Literature Cited

Affymetrix. 2005a. Exon array design datasheet [Internet]. Available from: http://www.affymetrix.com/support/technical/datasheets/exon_arraydesign_datasheet.pdf

- Affymetrix. 2005b. Gene signal estimates from exon arrays [Internet]. Available from: http://www.affymetrix.com/support/technical/whitepapers/exon_gene_signal_estimate_whitepaper.pdf
- Gilad Y, Oshlack A, Rifkin SA. 2006. Natural selection on gene expression. *Trends Genet.* 22:456–461.
- Irizarry RA, Warren D, Spencer F, et al. (25 co-authors). 2005. Multiple-laboratory comparison of microarray platforms. *Nat Methods.* 2:345–350.
- Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS. 2006. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci USA.* 103:12457–12462.
- Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene.* 345:119–126.
- Khaitovich P, Enard W, Lachmann M, Paabo S. 2006. Evolution of primate gene expression. *Nat Rev Genet.* 7:693–702.
- Khaitovich P, Muetzel B, She X, et al. 2004. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* 14:1462–1473.
- Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA.* 98:31–36.
- Liao BY, Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol.* 23:530–540.
- Su AI, Wiltshire T, Batalov S, et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA.* 101:6062–6067.
- Wheeler DL, Barrett T, Benson DA, et al. (30 co-authors). 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35:D5–D12.
- Wu Z, Irizarry RA. 2005. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol.* 12:882–893.
- Xing Y, Kapur K, Wong WH. 2006. Probe selection and expression index computation of affymetrix exon arrays. *PLoS ONE.* 1:e88.
- Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics.* 8:15–24.

Kenneth Wolfe, Associate Editor

Accepted March 19, 2007