

Using high-density exon arrays to profile gene expression in closely related species

Lan Lin¹, Song Liu², Heather Brockway³, Junhee Seok⁴, Peng Jiang¹,
Wing Hung Wong⁵ and Yi Xing^{1,6,*}

¹Department of Internal Medicine, University of Iowa, Iowa City, IA 52242, ²Department of Biostatistics, Roswell Park Cancer Institute, The State University of New York at Buffalo, Buffalo, NY 14203, ³Interdisciplinary Graduate Program in Genetics, University of Iowa, Iowa City, IA 52242, USA, ⁴Department of Electrical Engineering, ⁵Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305 and ⁶Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242

Received January 11, 2009; Revised May 5, 2009; Accepted May 7, 2009

ABSTRACT

Global comparisons of gene expression profiles between species provide significant insight into gene regulation, evolutionary processes and disease mechanisms. In this work, we describe a flexible and intuitive approach for global expression profiling of closely related species, using high-density exon arrays designed for a single reference genome. The high-density probe coverage of exon arrays allows us to select identical sets of perfect-match probes to measure expression levels of orthologous genes. This eliminates a serious confounding factor in probe affinity effects of species-specific microarray probes, and enables direct comparisons of estimated expression indexes across species. Using a newly designed Affymetrix exon array, with eight probes per exon for approximately 315 000 exons in the human genome, we conducted expression profiling in corresponding tissues from humans, chimpanzees and rhesus macaques. Quantitative real-time PCR analysis of differentially expressed candidate genes is highly concordant with microarray data, yielding a validation rate of 21/22 for human versus chimpanzee differences, and 11/11 for human versus rhesus differences. This method has the potential to greatly facilitate biomedical and evolutionary studies of gene expression in nonhuman primates and can be easily extended to expression array design and comparative analysis of other animals and plants.

INTRODUCTION

Comparative genomic analysis of gene expression has become an important tool for studying mechanisms of gene regulation, evolution, and human diseases (1). A large number of studies have utilized microarray technology for global comparison of gene expression profiles between closely related species, such as humans and nonhuman primates (2). A typical gene expression array measures the expression levels of tens of thousands of genes simultaneously based on fluorescent intensities of probes complementary to specific gene targets (3). In past research, two microarray-based approaches were used for comparative analysis of gene expression (2). The first approach, often referred to as 'cross-species microarray hybridization', hybridizes RNAs from the species of interest to a microarray platform designed for a closely related species (4–9). For example, Khaitovich and colleagues hybridized human and chimpanzee RNAs to the Affymetrix human U133 Plus 2.0 arrays to examine within-species and between-species gene expression differences in five tissues (9). However, sequence divergence between orthologous genes poses a major problem for cross-species microarray hybridization (2,4,10). Microarray probes designed for a human gene may contain mismatches to orthologous transcripts from non-human primates. Although in principle it is possible to remove individual probes targeting non-conserved regions, the small number of probes per gene on conventional gene expression arrays significantly undermines the applicability of this filtering strategy (11). Based on the sequence divergence rate between human, chimpanzee and rhesus macaque genomes, Oshlack *et al.* estimated that an average of fewer than three probes per probeset

*To whom correspondence should be addressed. Tel: +1 319 384 3099; Fax: +1 319 384 3150; Email: yi-xing@uiowa.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2009 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

on the Affymetrix U133 Plus 2.0 array perfectly matched orthologous mRNA sequences from all three species (11). The second approach is to design species-specific microarray probes for every species being studied (10,12). For example, Blekhman and colleagues recently designed a NimbleGen microarray containing species-specific probes for mRNA sequences of humans, chimpanzees and macaques (13). However, it is well known that even microarray probes for the same mRNA target could have substantially different fluorescent intensities due to probe-by-probe variation in hybridization affinity (14,15). In comparative genomic studies using species-specific probes, as probes are designed independently for orthologous genes, probe affinity effects prevent direct comparisons of expression indexes across species (2). In fact, two studies show that the gene expression indexes in human tissues, as measured by an Affymetrix human 3' array, have poor correlation with expression indexes in corresponding mouse tissues measured by an Affymetrix mouse 3' array (16,17). After the calculation of expression indexes in individual species, complex and technically challenging statistical procedures are needed to correct for probe affinity effects before it is feasible to compare expression indexes across species (11,12).

In this work, we show that high-density exon arrays designed for a single reference genome can be used as a flexible platform for global comparisons of gene expression profiles between closely related species. With the increase of oligonucleotide probe density on microarrays, a new generation of expression arrays allocates multiple probes for every known and predicted exon in the genome (18). For example, the Affymetrix Human Exon 1.0 array has an average of four probes per exon and 147 probes per gene, including an average of 58 'core probes' per gene targeting exon regions supported by RefSeq transcripts (18,19). The new Affymetrix Human Exon Junction array (HJAY) has eight probes per exon for approximately 315 000 exons in the human genome (20,21), representing a 2-fold increase in the density of exon probes when compared to the Exon 1.0 array. The increased probe density of the HJAY array in well-annotated exon regions is achieved by removing Exon 1.0 array probes targeting computationally predicted transcripts. With the high probe density of these new arrays, there are a large number of perfectly matched probes between humans and closely related nonhuman primates. In this study, we assess the possibility of using high-density exon arrays of a single species for comparative analysis of gene expression profiles. We introduce a simple computational procedure to construct robust expression indexes of orthologous genes, which are not confounded by probe affinity effects and can be directly compared across multiple species. We test whether this approach can reliably detect between-species differences in gene expression levels, using the HJAY array and quantitative real-time PCR analysis of corresponding human, chimpanzee, and rhesus macaque tissues. We also provide probe annotations and a computer program JETTA (Junction and Exon array Toolkit for Transcriptome Analysis) to support exon array analysis of gene expression in nonhuman primates.

MATERIALS AND METHODS

Identification of Human U133 Plus 2.0, Exon 1.0 and HJAY array probes targeting conserved regions between humans and nonhuman primates

Gene and probe annotations of the Affymetrix Human U133 Plus 2.0 array (GEO platform ID: GPL570) and the Exon 1.0 array (GEO platform ID: GPL5175) were downloaded from Affymetrix (www.affymetrix.com/products_services/arrays/specific/hgu133plus.affx and <http://www.affymetrix.com/support/technical/byproduct.affx?product=huexon-st>). The Affymetrix HJAY array (GEO platform ID: GPL8444) was purchased from Affymetrix as a Technology Access product. Gene and probe annotations of HJAY arrays were provided by Affymetrix.

For each probe, we obtained the coordinate of its target sequence from the hg18 assembly of the human genome. Using UCSC pairwise genome alignments of the human genome (hg18) to the genomes of chimpanzee (panTro2), orangutan (ponAbe2) and rhesus macaque (rheMac2) (22,23), we compiled the list of probes whose 25mer target regions were perfectly conserved in nonhuman primates for each array platform.

We used SeqMap (24) to search 25mer sequences of all probes against the genomes of human (hg18), chimpanzee (panTro2), orangutan (ponAbe2) and rhesus macaque (rheMac2). From these results, we identified probes for each platform that matched a single unique location in the human, chimpanzee, orangutan or rhesus macaque genome. By combining UCSC pairwise genome alignment results and SeqMap mapping results, we compiled the list of probes that perfectly matched the human genome and the genomes of nonhuman primates at a single unique location for all platforms in the study.

Human exon array data of 11 human tissues

We downloaded a public Affymetrix Exon 1.0 array data set of 11 human tissues (breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testes and thyroid), with three replicates per tissue (http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx).

Total RNA preparation and exon array profiling of human, chimpanzee and rhesus macaque tissues

Frozen cerebellums and livers from three chimpanzees and frozen cerebellums from three rhesus macaques were generously provided by Southwest National Primate Research Center (San Antonio, TX). Total RNA was extracted using TRIzol (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Total human cerebellum RNA (pool of 24 male and female donors) was purchased from Clontech (Mountain View, CA). Single-pass cDNA was synthesized using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA) according to manufacturer's instructions.

We used the Affymetrix Human Exon 1.0 array to profile cerebellum and liver tissues from chimpanzees, with biological replicates from three separate animals.

We also used the Affymetrix HJAY array to profile cerebellum tissues from humans (three technical replicates of the pooled cerebellum RNA), chimpanzees and rhesus macaques (biological replicates from three separate animals of each species). Detailed information (e.g. age, gender) of all RNA samples is described in Supplementary Table 1. Sample preparation and hybridization were identical for each platform. RNA samples were prepared using the GeneChip Whole Transcript Sense Target Labeling Assay (Affymetrix). For each sample, 2 µg of total RNA was subjected to ribosomal RNA reduction. Following rRNA reduction, double-stranded cDNA was synthesized with random hexamers tagged with a T7 promoter sequence. The double-stranded cDNA was used as a template for amplification with T7 RNA polymerase to create antisense cRNA. Next, random hexamers were used to reverse transcribe the cRNA to produce single-stranded sense strand DNA. The DNA was fragmented and labeled with terminal deoxynucleotidyl transferase. A hybridization cocktail was prepared, hybridized to the arrays and scanned.

Calculation of gene expression indexes of humans and nonhuman primates

We developed the JETTA program (Junction and Exon array Toolkit for Transcriptome Analysis, <http://gluegrant1.stanford.edu/~junhee/JETTA/>) to calculate gene expression indexes from Affymetrix Human Exon 1.0 array data of chimpanzee and human tissues (two chimpanzee tissues and 11 human tissues, each with three replicates). To calculate the expression index, we first predicted the background intensities of individual Exon 1.0 array probes, using a sequence-specific linear model trained from 'anti-genomic' background probes on the Human Exon 1.0 array (19). These 'anti-genomic' background probes were selected by Affymetrix to avoid a broad range of animal, plant and bacterial genomes (see http://www.affymetrix.com/support/technical/datasheets/exon_arraydesign_datasheet.pdf). For every probe, the predicted background intensity was an estimate for the amount of non-specific hybridization to the probe. This background intensity was subtracted from the observed probe intensity before downstream analysis. Second, we calculated gene expression indexes in human and chimpanzee samples. For each gene, starting with all core probes that perfectly matched human and chimpanzee genomes at a single unique location, we used a correlation-based iterative probe selection algorithm (25) to select a subset of probes with highly correlated intensities across all samples. This probe selection algorithm was developed to remove exon array probes that may not reflect overall gene expression levels, such as those targeting alternative exons or putative exon predictions, as well as low-affinity or cross-hybridizing probes (25). Our previous studies show that this probe selection algorithm produces robust expression indexes (19,26,27). The selected probes were regarded as reliable indicators of overall gene expression levels. In genes with at least six selected probes, the background-corrected intensities of selected probes were fitted to the Li-Wong model (14) as in (19,25) to construct

robust estimates of gene expression indexes. Finally, the expression indexes of all human and chimpanzee samples were normalized using quantile normalization.

We used the same procedure to calculate gene expression indexes from the HJAY array data of human, chimpanzee and rhesus macaque cerebellums. The background model was trained from 'anti-genomic' background probes on the HJAY array. Expression indexes were calculated from all HJAY probes that perfectly matched human/chimpanzee genomes or human/rhesus genomes at a single unique location, using the correlation-based iterative probe selection algorithm described above (25).

Correlation analysis of Human Exon 1.0 array profiles of human and chimpanzee tissues

For each of the two chimpanzee tissues and 11 human tissues, we first calculated average gene expression indexes of three replicates. For all possible pairs of tissues, we calculated the Spearman correlation coefficient using the expression indexes of 2165 genes with large variation in gene expression levels across tissues. These genes were selected by requiring a coefficient of variation (CV) in gene expression indexes of at least 0.8, and expression indexes of over 100 in at least 20% samples. For each gene, the coefficient of variation of its expression indexes is calculated as the standard deviation of expression indexes divided by the mean of expression indexes in all samples.

Detection of differentially expressed genes between human, chimpanzee and rhesus macaque cerebellums using HJAY array data

Using expression indexes calculated from HJAY array data, we performed a pairwise comparison of gene expression levels in human and chimpanzee cerebellums using Significance Analysis of Microarrays (SAM) (28). We filtered genes whose maximum expression indexes were <100 in the three chimpanzee samples and three human samples. SAM analysis was performed with a log transformation of the gene expression indexes. We used the default setting of SAM to identify significantly differentially expressed genes with a minimum fold change of 2.0. We used the same procedure to identify differentially expressed genes from the HJAY array data on the human and rhesus macaque cerebellums.

Quantitative real-time PCR validation of differentially expressed genes between human, chimpanzee and rhesus macaque cerebellums

Quantitative real-time polymerase chain reaction (qRT-PCR) was performed using Power SYBR Green PCR Master Mix (Applied Biosystems, Foster City, CA). For qPCR analysis of expression differences between human and chimpanzee cerebellums, a single primer set that perfectly matched both human and chimpanzee mRNAs was designed using PRIMER3 (29). Primer sequences are described in Supplementary Table 2. Using these primers, qPCR was conducted on extracted RNA from human and chimpanzee cerebellums. Two micrograms of total RNA were used for each 20 µl cDNA synthesis reaction. Using a

mathematical method described by Pfaffl (30), we calculated the average expression fold change in the pooled human cerebellum sample over each of the three chimpanzee cerebellum samples. All tested mRNA concentrations were normalized to HPRT1 as the reference gene. Similar results were obtained using β -actin as the reference gene (data not shown).

To determine the differences in expression between human and rhesus macaque cerebellums, qPCR primers were separately designed using PRIMER3 (29) to amplify orthologous regions in the mRNA. If the human and rhesus primer sets had a difference in amplification efficiency of >10% [as estimated by a standard curve analysis (30)], we designed and tested additional primers to select the human primer set and the rhesus primer set with similar amplification efficiency. Primer sequences are described in Supplementary Table 3. Two micrograms of total RNA were used for each 20 μ l cDNA synthesis reaction. Using a mathematical method described by Pfaffl (30), we calculated the average expression fold change in the pooled human cerebellum sample over each of the three rhesus cerebellum samples. All tested mRNA concentrations were normalized to HPRT1 as the reference gene.

RESULTS

Analysis of Human U133 Plus 2.0, Exon 1.0 and HJAY array probes targeting conserved regions between humans and nonhuman primates

We analyzed three generations of Affymetrix human expression arrays (U133 Plus 2.0 array, Exon 1.0 array and Exon Junction (HJAY) array) to determine the extent of their probe coverage for expression profiling of closely related nonhuman primates. The Human U133 Plus 2.0 array is the latest and most popular version of Affymetrix 3'-biased expression arrays. It uses sets of 11 perfect-match probes complementary to the 3' ends of mRNA. Individual genes may have multiple probesets that target different regions within the 3'-end or alternative 3'-ends. The Human Exon 1.0 array, released in 2005, is the first generation of Affymetrix exon arrays. This array averages four probes per exon and 58 'core probes' per gene. The HJAY array (Human Exon Junction array) is a second generation of Affymetrix exon array. This array has eight probes per exon for approximately 315 000 exons in the human genome (20,21) and also includes probes for exon-exon junctions. The two-fold increase in exon probe density on HJAY arrays is achieved by removing Exon 1.0 array probes targeting computationally predicted transcripts.

For each 25mer probe, we used pairwise alignments of the UCSC human and nonhuman primate genomes to determine if a probe was a perfect-match for its orthologous target region in chimpanzees, orangutans and rhesus macaques (see 'Materials and Methods' section). As conventional Affymetrix 3'-biased expression arrays (including the U133 Plus 2.0 array) have 11 perfect-match probes per probeset (14), we asked how many genes on the Exon 1.0 array and the HJAY array

have at least 11 or at least 6 perfect-match probes for their orthologous regions in nonhuman primates. In comparison, for the U133 Plus 2.0 array, we counted the number of probesets with at least 6 or 11 perfect-match probes for nonhuman primates. For genes with multiple U133 Plus 2.0 probesets, we also combined probes from multiple probesets (regardless of whether these probesets target distinct alternative transcripts) to count the maximum number of probes that perfectly matched nonhuman primates.

Our analysis indicates that the HJAY array and the Exon 1.0 array have a much higher number of probes that perfectly match nonhuman primate genomes, when compared to the U133 Plus 2.0 array. As summarized in Table 1, on the HJAY array, the number of genes with at least 11 perfect-match probes in chimpanzees, orangutans and rhesus macaques is 16402, 15322 and 14360, with a median count of 84, 61 and 48 probes per gene, respectively. On the Exon 1.0 array, the number of genes with at least 11 perfect-match probes in chimpanzees, orangutans and rhesus macaques was 16885, 14488 and 12824, with a median count of 41, 32 and 28 probes per gene, respectively. In contrast, the number of U133 Plus 2.0 probesets with at least 11 perfect-match probes was 4213, 735 and 282 for these three species. When we combined multiple probesets for the same gene on the U133 Plus 2.0 array, the number was 10488, 6978 and 4241 for chimpanzee, orangutan and rhesus genomes, with a median count of 20, 17 and 15 probes per gene, respectively. The same trend was observed when we counted the number of genes with at least six perfect-match probes in nonhuman primates (see Table 1). In fact, 12481 genes on the HJAY array and 10106 genes on the Exon 1.0 array had at least 11 probes that perfectly matched all four genomes (human, chimpanzee, orangutan and rhesus genomes), with a median count of 38 probes and 23 probes per gene on these two array platforms. By contrast, only 2281 genes on the U133 Plus 2.0 array had more than 11 probes that matched all four genomes, with a median count of 15 probes per gene. It should be noted that our estimate of probe counts for the U133 Plus 2.0 array is an upper bound estimate, since many genes on the U133 Plus 2.0 array have multiple probesets targeting distinct alternative transcripts which should not be combined in the counting of perfect-match probes (see http://www.affymetrix.com/support/technical/technotes/hgu133_p2_technote.pdf).

We further searched 25mer probe sequences against human and nonhuman primate genomes and removed those that matched multiple locations in the genomes. For example, in the human versus rhesus genome alignment analysis, we removed probes that matched multiple locations in either human or rhesus genomes (see 'Materials and Methods' section). As summarized in Table 2, 14037 genes had more than 11 HJAY array probes that perfectly matched human and rhesus genomes at a single unique location, with a median count of 47 probes per gene. On the Exon 1.0 array, the number was 12250 genes with a median count of 28 probes per gene. In contrast, on the U133 Plus 2.0 array, only 3865 genes had more than 11 probes that perfectly matched

Table 1. Number of genes on HJAY array, Exon 1.0 array and U133 Plus 2.0 array with at least 11 or at least six probes that perfectly match human, chimpanzee, orangutan and rhesus genomes

	Human	Chimpanzee	Orangutan	Rhesus
Human HJAY array				
≥6 PM probes	17414 ^a (102 ^b)	16980 (81)	16 127 (58)	15 498 (45)
≥11 PM probes	16 774 (104)	16 402 (84)	15 322 (61)	14 360 (48)
Human Exon 1.0 array				
≥6 PM probes	18 473 (48)	17 989 (39)	16 258 (29)	15 169 (24)
≥11 PM probes	17 991 (49)	16 885 (41)	14 488 (32)	12 824 (28)
Human U133 Plus 2.0 array (Probe-set)				
≥6 PM probes	35 660 (11)	31 519 (9)	17 008 (7)	8 162 (7)
≥11 PM probes	24 432 (11)	4 213 (11)	735 (11)	282 (11)
Human U133 Plus 2.0 array (Gene)				
≥6 PM probes	18 225 (20)	16 979 (15)	12 437 (12)	8 984 (10)
≥11 PM probes	15 264 (22)	10 488 (20)	6 978 (17)	4 241 (15)

^aNumber of genes.^bMedian count of perfect match probes.

human and rhesus genomes at a single unique location, with a median count of 15 probes per gene (see Table 2).

Together, these results suggest that we can use a single high-density human exon array to measure expression levels of the vast majority of genes in a variety of non-human primates. Compared to the U133 Plus 2.0 array, the second generation of Affymetrix exon array (the HJAY array) has a substantial increase in the number of perfect-match probes for nonhuman primates. For instance, for genes with at least 11 perfect-match probes in all four genomes, the HJAY array has a 5.5-fold higher gene coverage than the U133 Plus 2.0 array, and a 2.5-fold higher probe density per gene. As expected, the HJAY array also has a higher probe density for orthologs of RefSeq human genes in nonhuman primates when compared to the Exon 1.0 array (see Tables 1 and 2).

Correlation analysis of human exon 1.0 array profiles of human and chimpanzee tissues

In comparative analyses of gene expression using species-specific arrays, variation in probe affinity is a major confounding factor in comparing expression indexes across species as probes are designed independently for multiple species (16,17). Our proposed approach using high-density exon arrays should not be affected by such probe effects, as we select identical sets of perfect-match probes to measure expression levels of orthologous genes. Thus, we expect a high correlation between the expression profiles of corresponding tissues from different species. To confirm this, we took advantage of a large preexisting Exon 1.0 dataset of 11 human tissues (including cerebellum and liver, see 'Materials and Methods' section), and generated triplicate Exon 1.0 array data of chimpanzee cerebellum and liver RNAs for comparisons between humans and chimpanzees. For each gene, starting with all core probes that perfectly matched human and chimpanzee genomes at a single unique location, we used a correlation-based iterative probe selection algorithm to select

Table 2. Number of genes on HJAY array, Exon 1.0 array and U133 Plus 2.0 array with at least 11 or at least six probes that perfectly match both the human genome and the genome of chimpanzees, orangutans or rhesus macaques at a single unique location

	Human	Chimpanzee	Orangutan	Rhesus
Human HJAY array				
≥6 PM probes	16 974 ^a (96 ^b)	16 745 (77)	15 955 (56)	15 226 (43)
≥11 PM probes	16 329 (100)	16 151 (80)	15 119 (59)	14 037 (47)
Human Exon 1.0 array				
≥6 PM probes	17 587 (46)	17 343 (37)	15 832 (29)	14 604 (23)
≥11 PM probes	16 881 (48)	16 066 (40)	14 023 (32)	12 250 (28)
Human U133 Plus 2.0 array (Probe-set)				
≥6 PM probes	32 949 (11)	28 524 (9)	15 637 (7)	7 373 (7)
≥11 PM probes	18 758 (11)	3 513 (11)	629 (11)	241 (11)
Human U133 Plus 2.0 array (Gene)				
≥6 PM probes	16 813 (19)	15 689 (14)	11 724 (11)	8 267 (10)
≥11 PM probes	13 139 (21)	9 579 (20)	6 485 (17)	3 865 (15)

^aNumber of genes.^bMedian count of perfect match probes.

reliable indicators of overall expression levels (see 'Materials and Methods' section). Requiring that at least six probes were selected for a gene, we calculated expression indexes of 15143 genes in human and chimpanzee tissues.

From the computed expression indexes, we investigated the similarity of expression profiles between human and chimpanzee tissues. We selected 2165 genes with large variations in expression levels across all samples and calculated their average expression indexes in two chimpanzee and 11 human tissues (see 'Materials and Methods' section). For each pair of tissues, we calculated the Spearman correlation coefficient of expression indexes of these 2165 genes as the metric of similarity in expression profiles. Our analysis indicates that the expression profiles of human cerebellum and liver are closest to their chimpanzee counterparts as opposed to any other human tissue (Figure 1). We obtained a Spearman correlation coefficient of 0.936 between human and chimpanzee cerebellums, and 0.887 between human and chimpanzee livers. In contrast, the correlation coefficient was -0.159 between human cerebellum and human liver, and -0.127 between chimpanzee cerebellum and chimpanzee liver. These results contrasted with previous analyses of human and mouse tissues using species-specific expression arrays, where probe affinity effects largely obscured the similarity of expression profiles of orthologous tissues (16,17).

HJAY array detection and real-time qPCR validation of differentially expressed genes between human, chimpanzee and rhesus macaque cerebellums

A key goal of this study is to assess whether high-density exon arrays could be used to detect expression differences of orthologous genes in corresponding tissues. To test this, we used the HJAY array to generate expression profiles of human, chimpanzee and rhesus macaque cerebellums, with three replicates per species. We chose the HJAY

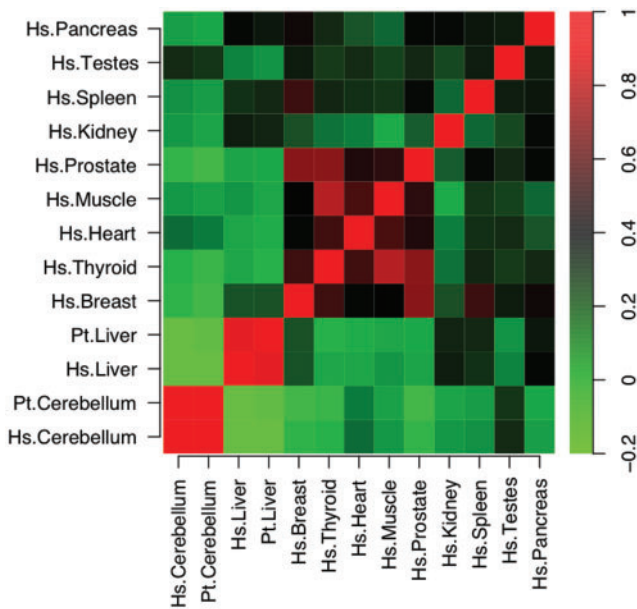


Figure 1. Correlation of Exon 1.0 array profiles of human (Hs) and chimpanzee (Pt) tissues. The heatmap shows that expression profiles of human cerebellum and liver are closest to their chimpanzee counterparts as opposed to any other human tissue.

array for this analysis, because it has a higher probe density for orthologs of RefSeq human genes in nonhuman primates. It should be noted that the HJAY array and the Exon 1.0 array use identical sample preparation and hybridization protocols.

We first tested HJAY array detection of expression differences between human and chimpanzee cerebellums. Using HJAY exon probes that perfectly matched human and chimpanzee genomes at a single unique location, we calculated expression indexes of 14884 genes. We used Significance Analysis of Microarrays (SAM) (28) under its default settings (see ‘Materials and Methods’ section) and identified 916 genes with a minimum of 2-fold change in expression levels between human and chimpanzee cerebellums, including 453 genes with increased expression in humans and 463 genes with decreased expression in humans. We randomly selected 22 differentially expressed genes for validation by SYBR Green real-time qPCR. Among the 22 genes selected for qPCR validation, 10 genes had increased expression and 12 genes had decreased expression in the human cerebellum according to HJAY array expression indexes. The genes selected for validation span a broad spectrum of functional categories and estimated expression indexes. QPCR analysis was performed on the same samples used for HJAY array profiling. Real-time qPCR data of 21 genes indicated at least 2-fold change in expression levels between human and chimpanzee cerebellums and were concordant with the microarray data (Table 3). The only exception was NT5C, for which the HJAY array and qPCR data both indicated a decreased expression level in the human cerebellum, but the fold-change estimated by qPCR was only 1.3. Therefore, using a qPCR fold change of 2.0 as the criteria for positive validation,

21 out of 22 candidate genes were validated by qPCR. We also plotted the log₂ expression fold changes of these 22 genes between human and chimpanzee cerebellums as estimated by the HJAY array and qPCR. We observed a strong positive correlation between the HJAY array data and qPCR data, with a Spearman correlation coefficient of 0.90 (see Figure 2). Morey and colleagues suggest that a correlation of over 0.8 indicates strong qPCR validation of microarray results (31). It should be noted that the fold change estimated by the HJAY array was typically smaller than the fold change estimated by qPCR (see Table 3). This is expected, as saturation of oligonucleotide probes at high mRNA concentration is known to compress the fold change estimates of differentially expressed genes. Taken together, our results provide strong evidence that the human versus chimpanzee expression differences detected by HJAY arrays are accurate and reliable.

To assess whether the HJAY array can also detect expression differences of orthologous genes from more distantly related species, we compared HJAY-based expression indexes of 12473 genes between human and rhesus cerebellums. Using SAM, we identified 893 genes with increased expression levels in humans and 789 genes with decreased expression levels in humans. From the 22 genes in Table 3, we selected 11 that also had significant differences between human and rhesus cerebellums as detected by HJAY data, and examined their expression levels using real-time qPCR. All 11 genes had more than two-fold change between human and rhesus cerebellums according to qPCR (see Table 4), yielding a validation rate of 11/11. The Spearman correlation coefficient between HJAY-estimated fold changes and qPCR estimated fold changes was 0.85.

DISCUSSION

The transition from conventional ‘probe-poor’ expression arrays to a new generation of ‘probe-rich’ exon arrays marks a major shift in the design strategy of gene expression arrays. In this manuscript, we present a flexible and intuitive approach for comparative analysis of gene expression between closely related species, using high-density exon arrays designed for a single reference genome. Our approach builds on previous work that uses microarrays to examine evolutionary differences in gene expression (5,6,9,10,12,13,32), and is intended to overcome limitations in past research using cross-species microarray hybridization or species-specific microarrays (see ‘Introduction’ section). For example, sequence divergence between species is a serious problem for cross-species hybridization to conventional expression arrays (4). Oshlack *et al.* estimated that an average of fewer than three probes per probeset on Affymetrix 3’ arrays perfectly match orthologous transcripts from human, chimpanzee and rhesus genomes (11). In this study, we analyzed probe sequences of three generations of Affymetrix expressions arrays, including the 3’ biased U133 Plus 2.0 array and two generations of exon arrays (Exon 1.0 array and HJAY array). Our results indicate that high-density exon arrays,

Table 3. HJAY array and qPCR data of 22 genes in human and chimpanzee cerebellums

Gene	Transcript cluster ID	Gene description	Gene ID	Chimpanzee cerebellum			Human cerebellum			Human vs chimpanzee change		Human vs chimpanzee fold change	
				#100	#327	#487	A	B	C	HJAY array	qPCR	HJAY array	qPCR array
BCLAF1	812232	BCL2-associated ion factor 1	9774	1794.8	1831.4	1943.0	638.5	627.2	647.8	Decrease	Decrease	-2.91	-3.32
CABYR	829453	Calcium binding tyrosine-(Y)-phosphorylation regulated	26 256	791.7	1052.4	868.2	86.2	75.3	84.3	Decrease	Decrease	-11.03	-14.27
CENPT	827 391	Centromere protein T	80 152	138.2	94.1	116.5	415.1	431.9	398.3	Increase	Increase	3.57	7.36
CHL1	805763	Cell adhesion molecule with homology to L1CAM	10 752	3185.7	3027.8	3220.8	829.0	767.4	886.9	Decrease	Decrease	-3.84	-11.38
COL6A1	832836	Collagen, type VI, alpha 1	1291	53.1	71.8	53.5	368.1	415.8	383.9	Increase	Increase	6.55	14.71
CRYM	827150	Crystallin, mu	1428	41.5	41.6	38.6	1038.4	1042.6	996.5	Increase	Increase	25.28	49.57
DNTTIP2	802388	Deoxynucleotidyltransferase, terminal, interacting protein 2	30 836	1150.6	1167.4	1134.4	243.8	128.2	182.3	Decrease	Decrease	-6.23	-13.50
DSEL	829865	Dermatan sulfate epimerase-like	92 126	2180.7	1907.5	2288.1	806.9	818.5	810.5	Decrease	Decrease	-2.62	-2.67
EPHA6	806177	EPH receptor A6	285 220	810.7	752.6	823.3	156.0	130.9	131.7	Decrease	Decrease	-5.70	-4.54
FOS	824317	Proto-oncogene protein c-fos	2353	90.1	147.2	152.2	666.1	742.7	666.6	Increase	Increase	5.33	7.19
GSTM5	800 796	Glutathione S-transferase M5	2949	11.4	27.8	21.4	487.0	454.1	472.6	Increase	Increase	23.35	2.05
HYDIN	827427	Hydrocephalus inducing homolog (mouse)	54 768	24.6	31.7	27.5	265.3	255.9	256.2	Increase	Increase	9.28	3.79
JMJD1C	819286	Jumonji domain-containing protein 1C	221 037	2859.7	2792.5	2711.5	960.1	872.7	933.2	Decrease	Decrease	-3.02	-8.67
KTN1	824190	Kinesin (Kinesin receptor)	3895	1061.1	921.3	969.0	185.8	135.0	172.5	Decrease	Decrease	-5.98	-9.71
LPXN	821047	Leupaxin	9404	277.9	389.5	306.7	1751.4	1801.6	1739.6	Increase	Increase	5.43	2.19
NR4A1	821945	Nuclear receptor subfamily 4, group A, member 1	3164	54.6	69.4	98.7	200.3	177.1	165.2	Increase	Increase	2.44	2.88
NRG4	826086	Neuregulin 4	145 957	1527.5	1641.8	1736.3	702.8	618.6	651.7	Decrease	Decrease	-2.49	-3.04
NT5C	829187	5', 3'-nucleotidase, cytosolic	30 833	1650.5	1450.2	1464.1	379.6	572.5	773.6	Decrease	Decrease	-2.65	-1.30
SNTG2	803360	Gamma-2-syntrophin	54 221	209.3	235.1	213.3	69.3	62.6	63.3	Decrease	Decrease	-3.37	-2.01
SYNGR4	830567	Synaptogyrin 4	23 546	61.6	74.1	63.6	291.7	246.5	259.6	Increase	Increase	4.00	3.70
TMF1	807072	TATA element modulatory factor 1	7110	980.1	1004.7	1082.0	294.3	243.2	236.4	Decrease	Decrease	-3.96	-5.87
ZP2	827149	Zona pellucida glycoprotein 2 (sperm receptor)	7783	45.5	40.8	45.8	723.2	795.1	749.5	Increase	Increase	17.17	79.06

in particular the HJAY array, have high probe coverage for measuring gene expression in closely related nonhuman primates. The expression indexes constructed from exon array data have two desirable features. First, for each gene the expression indexes are computed from the signals of a large number of probes tiled over its entire transcribed region. In our HJAY array analysis of human and chimpanzee tissues, on average 80 probes per gene were used in the estimation of expression levels. The increased probe density is likely to produce more accurate gene expression indexes as demonstrated by previous studies (19,33). Second, probe affinity effects do not confound between-species comparisons of expression levels, as identical sets of perfect-match probes are used for constructing expression indexes of orthologous genes. Thus, the

computed expression indexes from multiple species can be directly imported into standard software tools for high-level analysis of expression data, such as detection of differential expression and hierarchical clustering. The elimination of probe affinity effects during the calculation of expression indexes greatly simplifies downstream data analysis.

Our approach is expected to have false negatives and false positives. Even among genes with sufficient probe coverage in nonhuman primates, false negatives could arise due to poor probe affinity or various types of microarray artifacts. As in most microarray experiments, we were unable to systematically assess the false negative rate in our study due to the lack of a large gold-standard for differentially expressed genes between these human

and nonhuman primate RNA samples. In the future, generation of spike-in data set may allow us to evaluate false negatives of the between-species HJAY array analysis. On the other hand, false discovery rate (i.e. the fraction of false positives among all reported positives) is widely accepted as the most crucial metric to evaluate

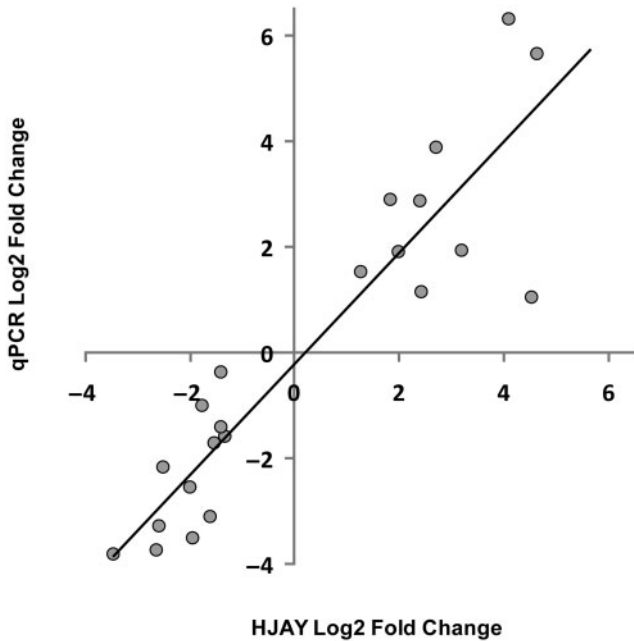


Figure 2. Correlation of expression fold change between human and chimpanzee cerebellums measured by HJAY array and real-time qPCR. X-axis: log₂-fold change of human expression level over chimpanzee expression level measured by HJAY array. Y-axis: log₂ fold change of human expression level over chimpanzee expression level measured by real-time qPCR.

genome-wide studies such as microarrays (34). Our qPCR validation suggests a low false discovery rate for HJAY array detection of differentially expressed genes between humans and nonhuman primates (1/22 for human versus chimpanzee differences; 0/11 for human versus rhesus differences). Moreover, the fold change values estimated by qPCR are highly concordant with the fold change values estimated by the HJAY arrays, with a Spearman correlation coefficient of 0.90 in the human versus chimpanzee comparison, and 0.85 in the human versus rhesus comparison. Collectively, we demonstrate that high-density exon arrays represent a cost-effective and high-throughput tool for detecting expression differences between closely related species. Also, although this study focuses on between-species comparisons of gene expression, high-density exon arrays can be used for standard microarray expression profiling within a single closely related species, such as the comparison between diseased animals and healthy controls. This could circumvent the need for designing custom arrays when expression array platform for a given species of interest is unavailable.

In this work, we use exon array probes that perfectly match orthologous regions of human exons to estimate gene expression levels in nonhuman primates. This assumes that the orthologous regions of human exons are also exons in other primate species. While this assumption is generally true, we know that a small percentage of human exons have had altered splicing patterns during primate evolution (e.g. recent creation of new exons) (35,36). Although evolutionary changes in alternative splicing have extremely interesting implications for function and evolution of eukaryotic genomes (37), such exons will introduce bias into the estimate of overall gene

Table 4. HJAY array and qPCR data of 11 genes in human and rhesus cerebellums

Gene	Transcript cluster ID	Gene description	Gene ID	Rhesus cerebellum			Human cerebellum			Human vs rhesus change		Human vs rhesus fold change	
				#453	#759	#775	A	B	C	HJAY array	qPCR	HJAY array	qPCR
CABYR	829453	Calcium binding tyrosine-(Y)-phosphorylation regulated	26256	282.9	247.1	327.3	89.3	68.7	78.5	Decrease	Decrease	-3.63	-4.34
CENPT	827391	Centromere protein T	80152	52.4	50.8	60.3	390.2	409.7	361.6	Increase	Increase	7.1	4.41
COL6A1	832836	Collagen, type VI, alpha 1	1291	44.8	46.1	49.4	391.4	431.6	374.2	Increase	Increase	8.54	217.74
CRYM	827150	Crystallin, mu	1428	67.8	61.9	54.0	1134.1	1178.4	1104.6	Increase	Increase	18.6	55.35
EPHA6	806177	EPH receptor A6	285220	35.8	32.6	30.4	205.9	162.0	143.0	Increase	Increase	5.17	5.42
HYDIN	827427	Hydrocephalus inducing homolog (mouse)	54768	29.5	29.3	20.9	265.7	240.6	234.5	Increase	Increase	9.29	16.56
JMJD1C	819286	Jumonji domain containing 1C	221037	2433.3	2411.0	2484.5	1000.7	912.7	979.7	Decrease	Decrease	-2.53	-7.92
KTN1	824190	Kinesin 1 (kinesin receptor)	3895	996.0	1129.5	1054.4	188.1	132.8	176.4	Decrease	Decrease	-6.39	-21.88
NT5C	829187	5', 3'-nucleotidase, cytosolic	30833	67.1	55.8	69.8	173.9	174.1	160.9	Increase	Increase	2.64	19.63
TMF1	807072	TATA element modulatory factor 1	7110	1075.0	1030.2	1051.9	287.8	232.6	241.0	Decrease	Decrease	-4.15	-4.53
ZP2	827149	Zona pellucida glycoprotein 2 (sperm receptor)	7783	33.3	34.1	44.0	611.0	668.0	605.0	Increase	Increase	16.91	15173.79

expression levels. Our correlation-based probe selection algorithm will help guard against this scenario, as it is designed to remove probes exhibiting substantially different splicing levels across samples (25). In the future, it will be possible to use transcript sequence data (e.g. cDNAs and mRNA-seq reads) of nonhuman primates to refine the selection of probes.

Software/data availability

We developed the JETTA program (<http://gluegrant1.stanford.edu/~junhee/JETTA/>) to support gene-level and exon-level analysis of HJAY array and Exon 1.0 array data. Probe annotations for HJAY array and Exon 1.0 array analysis of nonhuman primates can be downloaded from <http://www.medicine.uiowa.edu/Labs/Xing/Primate-microarray/>. These probe annotations can be used directly by JETTA to calculate gene expression indexes of nonhuman primates.

Affymetrix Human Exon 1.0 array data of chimpanzee cerebellum/liver and Affymetrix HJAY data of human, chimpanzee and rhesus cerebellums have been deposited to the NCBI GEO database under the accession number GSE15666.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank Jerilyn Pecotte, Mary Jo Aivaliotis, Garry Hauser, Elizabeth Zuo, Seiko Sato and Beverly Davidson for assistance. We thank David Eichmann, Ben Rogers and the University of Iowa Institute for Clinical and Translational Science (NIH grant UL1 RR024979) for computer support. This study used biological materials obtained from the Southwest National Primate Research Center, which is supported by NIH-NCRR grant P51 RR013986.

FUNDING

National Institutes of Health grant U54-GM062119 (to W.H.W.) and R01-HG004634 (to W.H.W. and Y.X.). University of Iowa research startup fund (to Y.X.). Funding for open access charge: National Institutes of Health grant R01-HG004634.

Conflict of interest statement. None declared.

REFERENCES

1. Khaitovich,P., Enard,W., Lachmann,M. and Paabo,S. (2006) Evolution of primate gene expression. *Nat. Rev. Genet.*, **7**, 693–702.
2. Gilad,Y. and Borevitz,J. (2006) Using DNA microarrays to study natural variation. *Curr. Opin. Genet. Dev.*, **16**, 553–558.
3. Allison,D.B., Cui,X., Page,G.P. and Sabripour,M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.
4. Bar-Or,C., Czosnek,H. and Koltai,H. (2007) Cross-species microarray hybridizations: a developing tool for studying species diversity. *Trends Genet.*, **23**, 200–207.
5. Caceres,M., Lachuer,J., Zapala,M.A., Redmond,J.C., Kudo,L., Geschwind,D.H., Lockhart,D.J., Preuss,T.M. and Barlow,C. (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl Acad. Sci. USA*, **100**, 13030–13035.
6. Enard,W., Khaitovich,P., Klose,J., Zollner,S., Heissig,F., Giavalisco,P., Nieselt-Struwe,K., Muchmore,E., Varki,A., Ravid,R. *et al.* (2002) Intra- and inter-specific variation in primate gene expression patterns. *Science*, **296**, 340–343.
7. Wang,Z., Lewis,M.G., Nau,M.E., Arnold,A. and Vahey,M.T. (2004) Identification and utilization of inter-species conserved (ISC) probesets on Affymetrix human GeneChip platforms for the optimization of the assessment of expression patterns in non human primate (NHP) samples. *BMC Bioinformatics*, **5**, 165.
8. Walker,S.J., Wang,Y., Grant,K.A., Chan,F. and Hellmann,G.M. (2006) Long versus short oligonucleotide microarrays for the study of gene expression in nonhuman primates. *J. Neurosci. Methods*, **152**, 179–189.
9. Khaitovich,P., Hellmann,I., Enard,W., Nowick,K., Leinweber,M., Franz,H., Weiss,G., Lachmann,M. and Paabo,S. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, **309**, 1850–1854.
10. Gilad,Y., Rifkin,S.A., Bertone,P., Gerstein,M. and White,K.P. (2005) Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.*, **15**, 674–680.
11. Oshlack,A., Chabot,A.E., Smyth,G.K. and Gilad,Y. (2007) Using DNA microarrays to study gene expression in closely related species. *Bioinformatics*, **23**, 1235–1242.
12. Gilad,Y., Oshlack,A., Smyth,G.K., Speed,T.P. and White,K.P. (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, **440**, 242–245.
13. Blekhan,R., Oshlack,A., Chabot,A.E., Smyth,G.K. and Gilad,Y. (2008) Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.*, **4**, e1000271.
14. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
15. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G., Geoghegan,J., Germino,G. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–350.
16. Yanai,I., Graur,D. and Ophir,R. (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS*, **8**, 15–24.
17. Liao,B.Y. and Zhang,J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.*, **23**, 530–540.
18. Clark,T.A., Schweitzer,A.C., Chen,T.X., Staples,M.K., Lu,G., Wang,H., Williams,A. and Blume,J.E. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, **8**, R64.
19. Kapur,K., Xing,Y., Ouyang,Z. and Wong,W.H. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biol.*, **8**, R82.
20. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
21. Yamamoto,M.L., Clark,T.A., Gee,S.L., Kang,J.A., Schweitzer,A.C., Wickrema,A. and Conboy,J.G. (2009) Alternative pre-mRNA splicing switches modulate gene expression in late erythropoiesis. *Blood*, **113**, 3363–3370.
22. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
23. Miller,W., Rosenbloom,K., Hardison,R.C., Hou,M., Taylor,J., Raney,B., Burhans,R., King,D.C., Baertsch,R., Blankenberg,D. *et al.* (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, **17**, 1797–1808.
24. Jiang,H. and Wong,W.H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **24**, 2395–2396.

25. Xing, Y., Kapur, K. and Wong, W.H. (2006) Probe selection and expression index computation of affymetrix exon arrays. *PLoS ONE*, **1**, e88.
26. Xing, Y., Ouyang, Z., Kapur, K., Scott, M.P. and Wong, W.H. (2007) Assessing the conservation of mammalian gene expression using high-density exon arrays. *Mol. Biol. Evol.*, **24**, 1283–1285.
27. Chiao, E., Elazar, M., Xing, Y., Xiong, A., Kmet, M., Millan, M.T., Glenn, J.S., Wong, W.H. and Baker, J. (2008) Isolation and transcriptional profiling of purified hepatic cells derived from human embryonic stem cells. *Stem Cells*, **26**, 2032–2041.
28. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
29. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
30. Pfaffl, M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.*, **29**, e45.
31. Morey, J.S., Ryan, J.C. and Van Dolah, F.M. (2006) Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol. Proced. Online*, **8**, 175–193.
32. Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigele, S., Do, H.H., Weiss, G., Enard, W. *et al.* (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.*, **14**, 1462–1473.
33. Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D.J., Jensen, R.V. and Majewski, J. (2008) Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC Genomics*, **9**, 529.
34. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
35. Sorek, R. (2007) The birth of new exons: mechanisms and evolutionary consequences. *RNA*, **13**, 1603–1608.
36. Calarco, J.A., Xing, Y., Caceres, M., Calarco, J.P., Xiao, X., Pan, Q., Lee, C., Preuss, T.M. and Blencowe, B.J. (2007) Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev.*, **21**, 2963–2975.
37. Xing, Y. and Lee, C. (2006) Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, **7**, 499–509.