



Technical Note

■ GeneChip® Human Exon 1.0 ST Array and GeneChip® WT Sense Target Labeling Assay for Genome-Wide, Exon-Level Expression Analysis

The recent development at Affymetrix of the new Whole Transcript (WT) Assay, featuring a random-priming strategy in combination with a novel, end-point DNA fragmentation and labeling method, has enabled the generation of biotinylated DNA targets along the entire length of the RNA transcripts. The first iteration of the WT Assay, the WT Sense Target Labeling Assay, as well as the associated reagents, is optimized specifically for the GeneChip® brand Exon Array type of design, as a complete system for genome-wide, exon-level expression and alternative splicing analysis.

This Technical Note describes the basic principles of the new GeneChip® WT Sense Target Labeling Assay, as well as its performance characteristics when used in combination with the GeneChip Human Exon 1.0 ST Array. New quality assessment metrics based on the Quality Reports generated by the Affymetrix Exon Array Computational Tool (ExACT) are discussed.

The results demonstrate that the assay is robust and generates sufficient targets for array hybridization. In addition, the combination of the assay and array produces sensitive and reproducible data, providing a new insight for researchers to understand the complexity of biology. Refer to the Technical Note "GeneChip® Exon Array Design" for more detailed information on the array design (P/N 702026, available online at www.affymetrix.com).

GeneChip® WT Sense Target Labeling Assay Principles

To meet the requirements of generating targets for analysis of individual exons along the entire length of the transcripts while achieving the desired sensitivity from limited starting materials, the GeneChip® WT Sense Target Labeling Assay and Reagents have been developed and optimized specifically for use with the Exon Arrays. Some of the key highlights of this assay include:

- The majority of the ribosomal RNA (rRNA) is removed from the total RNA samples prior to target labeling for increased sensitivity.
- *In vitro* transcription-based linear amplification is incorporated to allow reduction of the input starting total RNA to 1 µg.
- Single-stranded DNA targets are generated in the sense orientation, eliminating the complexity of data interpretation where a double-stranded target is used to hybridize to the array, since a portion of the genomic locus may have transcription activity from both strands.
- A novel, end-point DNA fragmentation strategy is utilized replacing the alternative DNase fragmentation method for robust and reproducible results.
- All key target labeling reagents after the rRNA removal step are provided by Affymetrix in the GeneChip® WT Sense Target Labeling and Control Reagents (P/N 900652, for 30 reactions), providing users with convenience, consistency, and a higher success rate.

A schematic representation of the basic steps of the assay is shown in Figure 1. The detailed protocol can be found in the GeneChip® WT Sense Target Labeling Assay Manual (P/N 701880 Rev. 2), available online at www.affymetrix.com.

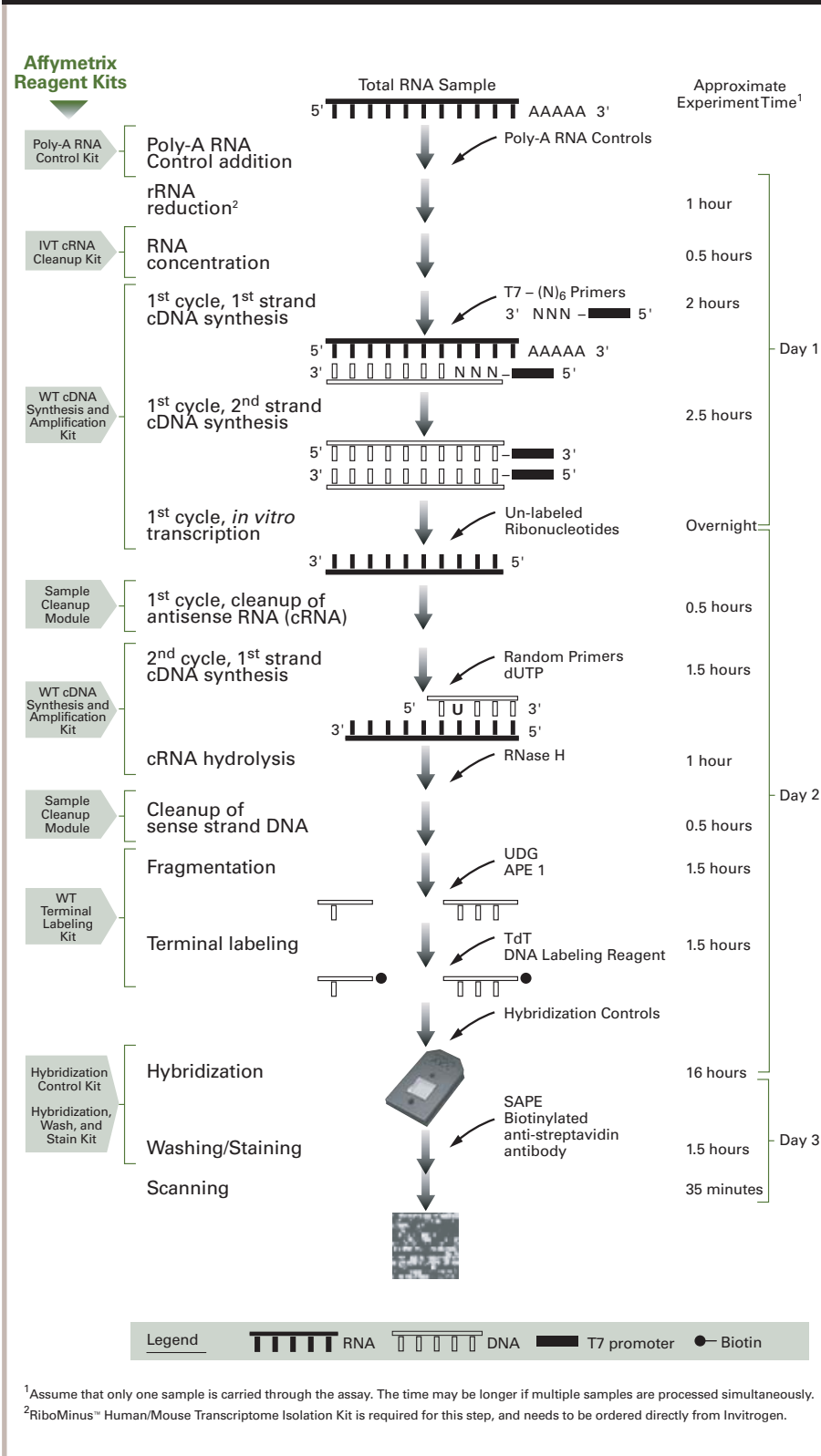
Ribosomal RNA Removal

In order to effectively label the entire length of the mRNA transcripts, a random-priming strategy is required. On average, the 18S and 28S ribosomal RNAs account for approximately 90 percent of any eukaryotic total RNA sample whereas the mRNA population makes up less than 2 percent in a given sample. Therefore, as a side effect of the random-priming strategy, the rRNA species are amplified and labeled together with the mRNA transcripts, introducing significant background, reducing the overall assay sensitivity, and unnecessarily consuming costly labeling enzyme and reagents.

To circumvent the undesirable effect and to reduce the assay complexity, Affymetrix has introduced an rRNA removal step prior to target labeling utilizing the RiboMinus™ Human/Mouse Transcriptome Isolation Kit from Invitrogen (P/N K1550-01). The efficiency of the rRNA removal procedure can be assessed by applying the samples before and after the treatment on a Bioanalyzer 2100 using an RNA 6000 Nano LabChip. Typical results are shown in Figure 2, where approximately 60 to 80 percent of rRNA is depleted.

It should be noted that the experimental procedures have been modified from the original vendor recommendations when

Figure 1. Schematic representation of the WT Sense Target Labeling Assay.



using the RiboMinus Kit for preparing the samples for the WT Sense Target Labeling Assay. Therefore, follow the protocol outlined in the GeneChip WT Sense Target Labeling Assay Manual for optimal results on the GeneChip® Exon Arrays. The RiboMinus Kit has also been tested for use with mouse and rat total RNA samples and the results are satisfactory (data not shown).

cRNA and cDNA Yields

During the course of assay development, a tissue panel was run to demonstrate the performance of the GeneChip WT Sense Target Labeling Assay and GeneChip Human Exon 1.0 ST Array. This data set also illustrates the range of performance on different tissue types.

Three independent target preparation replicates per tissue were carried out starting with 1 µg of total RNA. The recommended hybridization quantity of fragmented, labeled sense DNA target is 5.5 µg. Table 1 illustrates the range of cRNA and cDNA yields obtained from the 11 human tissues used in this study. Only 8 µg of the resultant cRNA were used in the second cycle of the cDNA synthesis reaction.

The average yields over the tissue panel for the cRNA step ranged from 16.2 to 32.9 µg. Average yields for the sense DNA target produced in the second cycle ranged from 6.0 to 7.3 µg, exceeding the 5.5 µg requirement for hybridization to a single array.

Reproducibility of DNA Fragmentation

A key advancement of the WT Sense Target Labeling Assay is the introduction of a novel, robust DNA fragmentation method. With an optimized ratio of dUTP to dTTP in the reverse transcription reaction of the second-cycle cDNA synthesis, as illustrated in Figure 1, the sense DNA strand is generated with dU incorporated at pre-defined intervals. The subsequent treatment with a combination of Uracil-

Figure 2. Ribosomal RNA depletion before and after treatment with the Invitrogen RiboMinus Human/Mouse Transcriptome Isolation Kit. Human brain total RNA (1 μg) was treated with the RiboMinus Kit according to the GeneChip[®] WT Sense Target Labeling Assay Manual. One μL out of the 8 μL of eluted RNA sample, after rRNA removal, was run on a Bioanalyzer 2100 using an RNA 6000 Nano LabChip. For comparison, the untreated starting material was also run on the LabChip, and the same proportion of the total sample was loaded for direct comparison (1/8 of the starting 1 μg of total RNA, therefore, 125 ng of starting sample was loaded on the LabChip). The ribosomal 18S and 28S peaks before (green) and after (blue, purple, and maroon) treatment with the RiboMinus Kit are shown.

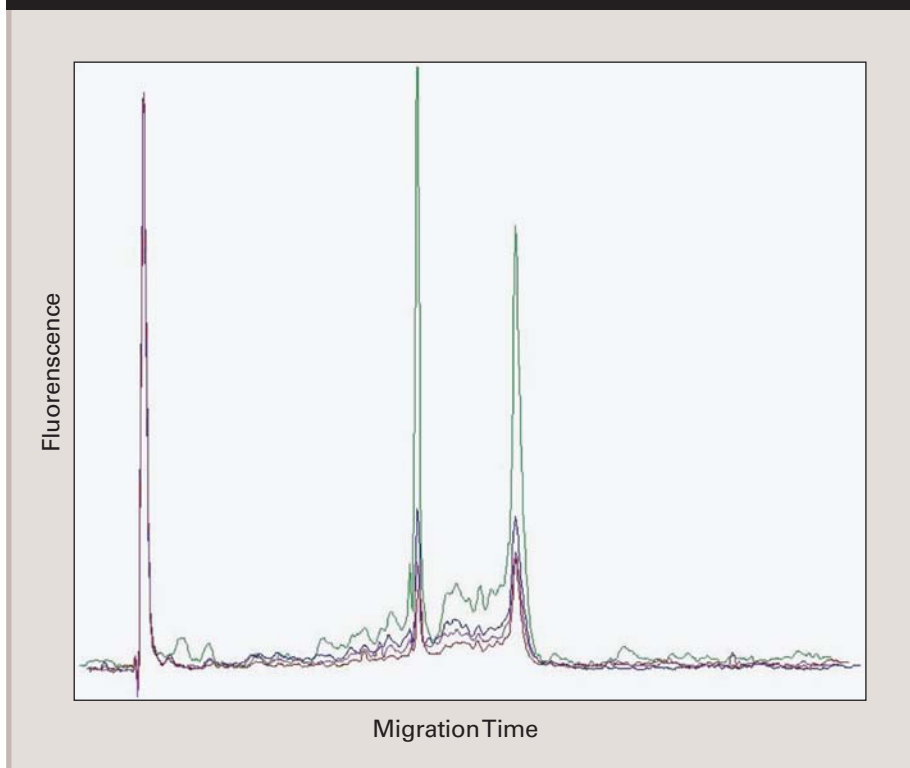


Table 1. cRNA and Sense DNA Target Yields from Human Tissues. Sense DNA target was prepared according to the protocol described in the GeneChip[®] WT Sense Target Labeling Assay Manual using 1 μg each from human thyroid, breast, pancreas, liver, heart, spleen, testes, kidney, skeletal muscle, prostate, and cerebellum total RNA. Three sample preparation replicates were carried out for each starting total RNA. Table shows average yields for each tissue. Standard Deviation is shown in parentheses.

Tissue	Input Total RNA	cRNA Yield (μg) (Std. Dev.)	cRNA input (μg)	cDNA Yield (μg) (Std. Dev.)
Thyroid	1 μg	27.7 (1.6)	8.0	7.3 (0.9)
Breast	1 μg	31.6 (6.1)	8.0	6.5 (0.5)
Pancreas	1 μg	32.9 (9.2)	8.0	7.0 (0.8)
Liver	1 μg	28.9 (9.1)	8.0	7.1 (0.1)
Heart	1 μg	31.8 (3.2)	8.0	6.3 (1.3)
Spleen	1 μg	21.0 (10.9)	8.0	6.9 (0.9)
Testes	1 μg	25.4 (2.4)	8.0	6.8 (0.8)
Kidney	1 μg	23.0 (8.1)	8.0	6.8 (1.3)
Skeletal Muscle	1 μg	26.2 (2.2)	8.0	6.4 (1.1)
Prostate	1 μg	28.7 (5.4)	8.0	6.0 (0.4)
Cerebellum	1 μg	16.2 (3.8)	8.0	6.7 (0.2)

DNA Glycosylase (UDG) and human apurinic/apyrimidinic endonuclease (APE 1) fragments the DNA where dU has been incorporated. The consistency and reproducibility of the fragmentation reaction is shown in Figure 3, with the peak fragment size at approximately 50 bases.

Target Labeling Representing the Entire Length of the Transcript

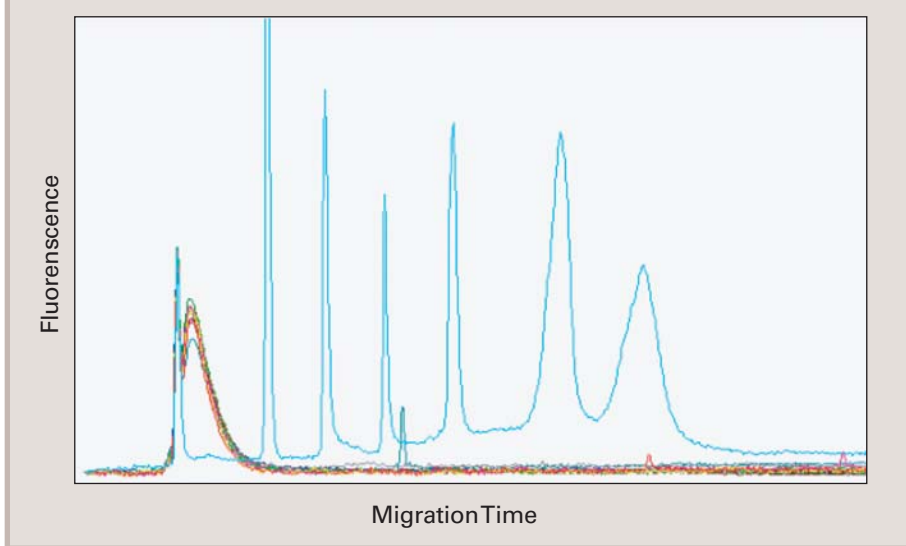
One key requirement for this assay is the full representation of the entire length of the mRNA in order to examine potential alternative splicing events occurring in various regions of the transcript. For an example of a global view demonstrating the detection of probe sets throughout the transcript, five full-length human transcripts were evaluated in triplicate on the Human Exon 1.0 ST Arrays.

Human transcripts from cDNA clones, ranging from 2,000 to 3,000 nucleotides, were spiked into human cervical cell line HeLa total RNA prior to carrying out the WT Sense Target Labeling Assay. Transcripts were spiked at 1:25,000 (ratio of copy number relative to mRNA) in the sample. These clones were previously determined not to be expressed in the HeLa complex background by quantitative RT-PCR.

Individual probe sets were divided into bins according to their distance from the 5' end of the transcript. Each bin represents 500 nucleotides. For example, the first bin includes all probe sets from the five transcripts that are within 500 nucleotides from the 5' end. The next bin includes all probe sets from 501 to 1,000 nucleotides from the 5' end, and so on up to 3,000 nucleotides. Information regarding the genes encoding the spiked transcripts, as well as the lengths of the transcripts and numbers of probe sets in each bin, are shown in Table 2.

PLIER signal values were calculated using the Exon Array Computational Tool (ExACT, available for free download from the Affymetrix web site). Signal values of the binned probe sets from all five tran-

Figure 3. Reproducibility of DNA fragmentation. Single-stranded DNA prepared according to the recommended protocol was fragmented with UDG and APE 1. Fragmented samples were run on an RNA 6000 Nano LabChip on a Bioanalyzer. The RNA 6000 Nano LabChip was found to work better with single-stranded DNA than a DNA LabChip. A total of twelve replicate fragmentations using two different reagent lots are shown. The light blue trace indicates the size marker and ladders with the following sizes from the left to the right: 25 base marker, and 0.2, 0.5, 1.0, 2.0, 4.0, and 6.0 kb ladders.



scripts are plotted in Figure 4. It is important to note that signals from different probe sets should not be compared directly with each other. However, in a qualitative fashion, Figure 4 demonstrates that different regions of the transcripts are detectable and represented using the WT Sense Target Labeling Assay.

The median signal for each bin across the transcripts ranged from 296-701, centering around 400-500. Note that the box for the last bin (2,501-3,000) represents the 3 replicates from only one probe set from transcript BC000249 and is, therefore, much smaller than the boxes repre-

sending multiple probe sets from different transcripts.

Another way to examine the full-length representation of the target labeling assay is to evaluate signal values from various probe sets throughout a specific transcript, which can be viewed in the Integrated Genome Browser (IGB, freely downloadable from Affymetrix web site). An example, phosphodiesterase 6B (PDE6B), is shown in Figure 5. Despite differences in the PLIER signal values for each probe set, the figure illustrates that there is signal above background for probe sets throughout this transcript, from 5' to 3'.

Assay Sensitivity Analysis

The overall sensitivity of the assay and the array was assessed with spike-in transcripts. With this new array design, two levels of analysis may be carried out – at the gene level (using meta-probe sets) and at the exon level (using probe sets).

Because most exons are relatively short in size with the median length of approximately 100 bp, even the best-performing probes selected to represent each exon are constrained in their base composition (see the “GeneChip® Exon Array Design Technical Note” for more detail). In addition, the Human Exon 1.0 ST Array has a reduced number of probes with four probes for most probe sets. Combining these two factors, it is anticipated that the sensitivity at the exon level will be lower than that achieved at the gene level.

In contrast, since most of the full length RefSeq sequences are composed of multiple exons, on average, 30 to 40 probes may be used on the array to represent RefSeq supported exons at that genomic locus, enhancing the detection at the gene level.

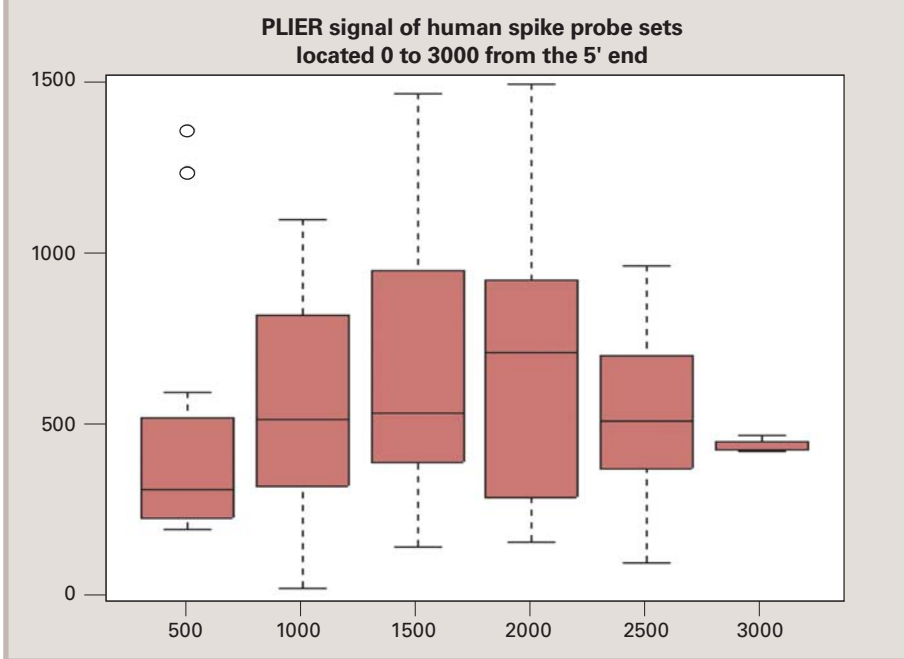
To assess the system’s capability to detect transcripts expressed at low levels, a group of forty exogenously generated transcripts (not endogenously expressed) were spiked into the HeLa total RNA at approximately 2 to 3 (1:100,000) or 4 to 6 (1:50,000) copies per cell.

As shown with the blue trace in Figure 6A, at the “core gene” level, the sensitivity for detecting two-fold changes in the spike-in transcripts at low levels

Table 2. Spike-in transcripts. Gene symbols, accession numbers and transcript length for each of the five spiked transcripts are given in the first three columns. The remaining columns list how many probe sets for each transcript fall into each of the 500-nucleotide bins representing distance from the 5' end of that transcript. Total number of probe sets in each bin for all transcripts is shown in the bottom row. Note that only one transcript was longer than 2,500 nucleotides, so there is only one probe set in the bin representing 2501-3000 nucleotides from the 5' end.

Gene	Acc. #	Transcript Length	0-500	501-1000	1001-1500	1501-2000	2001-2500	2501-3000
MYCN	BC002712	2458	2	2	4	3	1	0
PDE6B	BC000249	3109	1	5	5	4	7	1
ITGB8	BC002630	2068	0	3	3	4	0	0
CCKBR	BC000740	2004	2	2	1	0	0	0
SOX10	BC002824	2137	2	2	2	0	0	0
Total # of probe sets			7	14	15	11	8	1

Figure 4. PLIER signal for probe sets from 5' to 3' end of spike-in transcripts. Boxplot shows distribution of triplicate signal values of probe sets from five spike-in transcripts. Probe sets are binned by distance from the 5' end of each transcript. X-axis denotes bins in distance from 5' end: 0 - 500 nt from 5' end, 501 - 1,000 nt, 1,001 - 1,500 nt, 1,501 - 2,000 nt, 2,001 - 2,500 nt and 2,501-3,000 nt. Y-axis: PLIER signal for the three replicates of all probe sets in a bin. The lower and upper bounds of the boxes represent the 25th and 75th percentile, respectively. The line in the box represents the median. The box for the last bin (2,501-3,000) represents the 3 replicates from only one probe set from transcript BC000249 and is, therefore, much smaller than the boxes representing multiple probe sets from different transcripts. The variability within each box plot is driven by probe-specific effects.



(1:50,000 vs. 1:100,000 ratio copy number relative to mRNA) is quite high with approximately 90 percent sensitivity at 0 percent false positive. For more information on “core gene” annotations and gene-level signal estimates, see the “Gene Signal Estimates from Exon Arrays” v1.0 and the “Exon Probeset Annotations and Transcript Cluster Groupings” white papers available from the www.affymetrix.com, Support page.

At the probe set, or exon level, the detection sensitivity is modest, as shown with the blue trace in Figure 6B, where approximately 66 percent sensitivity is possible with no false positives. For researchers with limited samples, an alternative procedure may be followed bypassing the rRNA removal step, allowing the reduction of starting material from 1 µg in the standard recommended procedure

to 100 ng of total RNA. However, the trade-off for following the alternative protocol is some compromise in detection sensitivity. This is shown with the red traces in Figures 6A and 6B, for anticipated gene-level and exon-level performance, respectively.

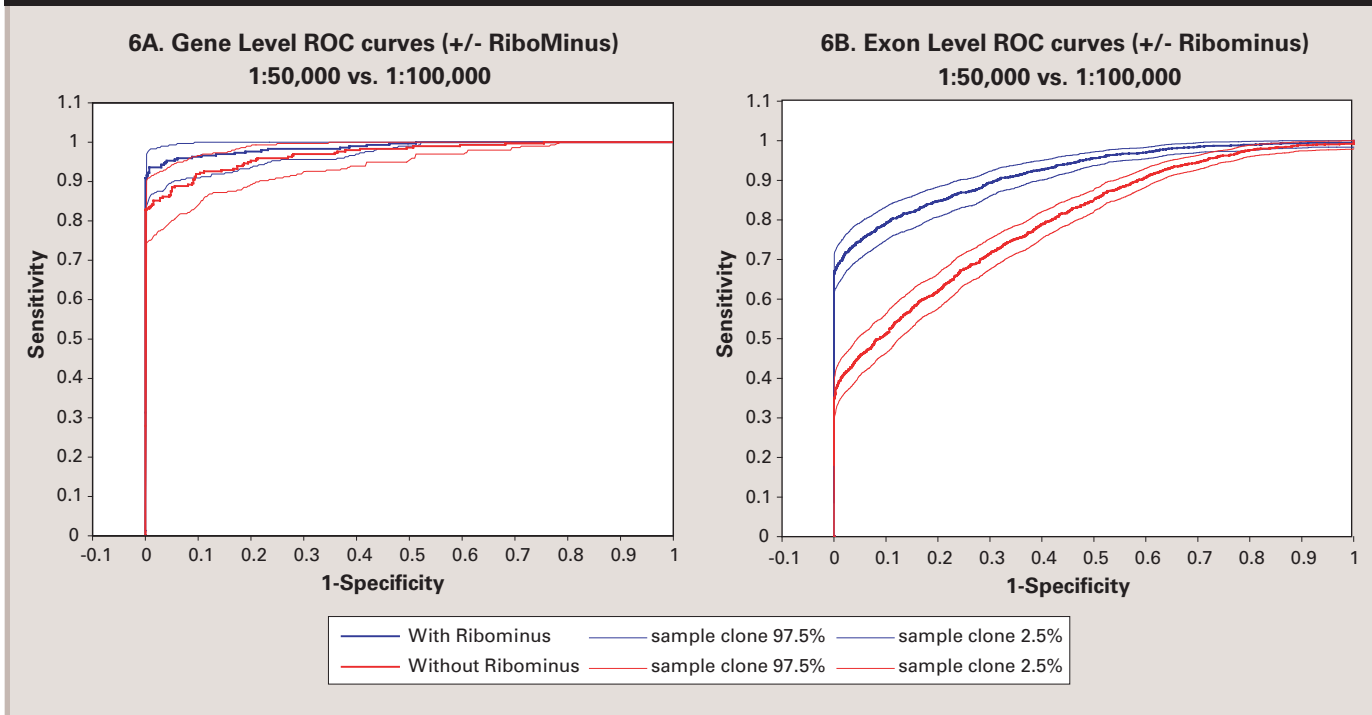
Introduction to New Array Quality Assessment Metrics

An additional feature with this new platform is a Quality Report generated by the ExACT software. With the introduction of a completely new array design, a new assay, and a new detection algorithm in ExACT, Detection Above Background (DABG), it is important to establish quality assessment metrics in order to have a better understanding of the overall array performance, and eliminate outlier arrays, prior to in-depth, higher level analysis of an experiment. Some of the metrics that are generated by ExACT in the Quality Report are described below. For a more detailed description of the calculation of these metrics, refer to the Affymetrix

Figure 5. Representation across the entire transcript for gene PDE6B. Full-length human transcripts from cDNA clones were spiked into the human cervical cell line HeLa total RNA at 1:100,000 (ratio of copy number relative to mRNA), where these clones were previously determined not to be expressed in the HeLa complex background by quantitative RT-PCR. The samples were then labeled according to the recommended protocol. PLIER signal estimates were generated using ExACT. One of the clones, PDE6B, was visualized in the Integrated Genome Browser (IGB). The bottom track in green shows the RefSeq transcript structure. The yellow boxes in the second track from the bottom show probe selection regions represented by the clone insert sequences, with the 5' end on the left-hand side of the graph, and the 3' end on the right-hand side of the graph. The bars in the top two tracks (red and green) represent relative PLIER signal estimates for each probe set matching the cloned sequence from two different target preparations. The transcript is approximately 3 kb in length.



Figure 6. Sensitivity in detecting different amounts of spiked-in transcripts. Thirty human and 10 Arabidopsis cloned transcripts that are not endogenously expressed in the HeLa cell line were spiked into HeLa total RNA at 1:100,000 or 1:50,000 (ratio copy number relative to mRNA), or approximately 2 to 3 or 4 to 6 copies per cell respectively. Labeled DNA targets were then prepared according to the recommended protocol. The ROC plot shows sensitivity and specificity for samples with transcripts spiked in at 1:50,000 compared to samples with transcripts spiked in at 1:100,000, and represents triplicate sample preparations for each sample type. The blue traces represent samples that have gone through rRNA removal using the RiboMinus Kit prior to target preparation. The red traces represent samples where 100 ng of total RNA was used without first depleting rRNA with the RiboMinus Kit.



white paper “Quality Assessment of Exon Arrays” available on www.affymetrix.com.

The Quality Report includes summary metrics, such as the number of probes and probe sets analyzed and the percentage of probes and probe sets detected by the DABG algorithm. These and other summary metrics in the report file illustrate the performance of a single array relative to all arrays in the experiment, thereby enabling the identification of outlier array results within a given experiment.

For example, the mean of the absolute Relative Log Expression (RLE) over various collections of probe sets is generated by calculating the difference in PLIER values for a probe set on one array compared to the median value of that probe set on all arrays within a set of experiments. The mean absolute value per array is computed from the RLE for all probe sets used in the

analysis of that array.

Other metrics include the mean probe set PLIER target response that represents the mean of the PLIER signal estimates from all probe sets used in the analysis. The mean probe set absolute PLIER residuals is a summary of the residuals from the PLIER model fit for all probe sets analyzed on a given array.

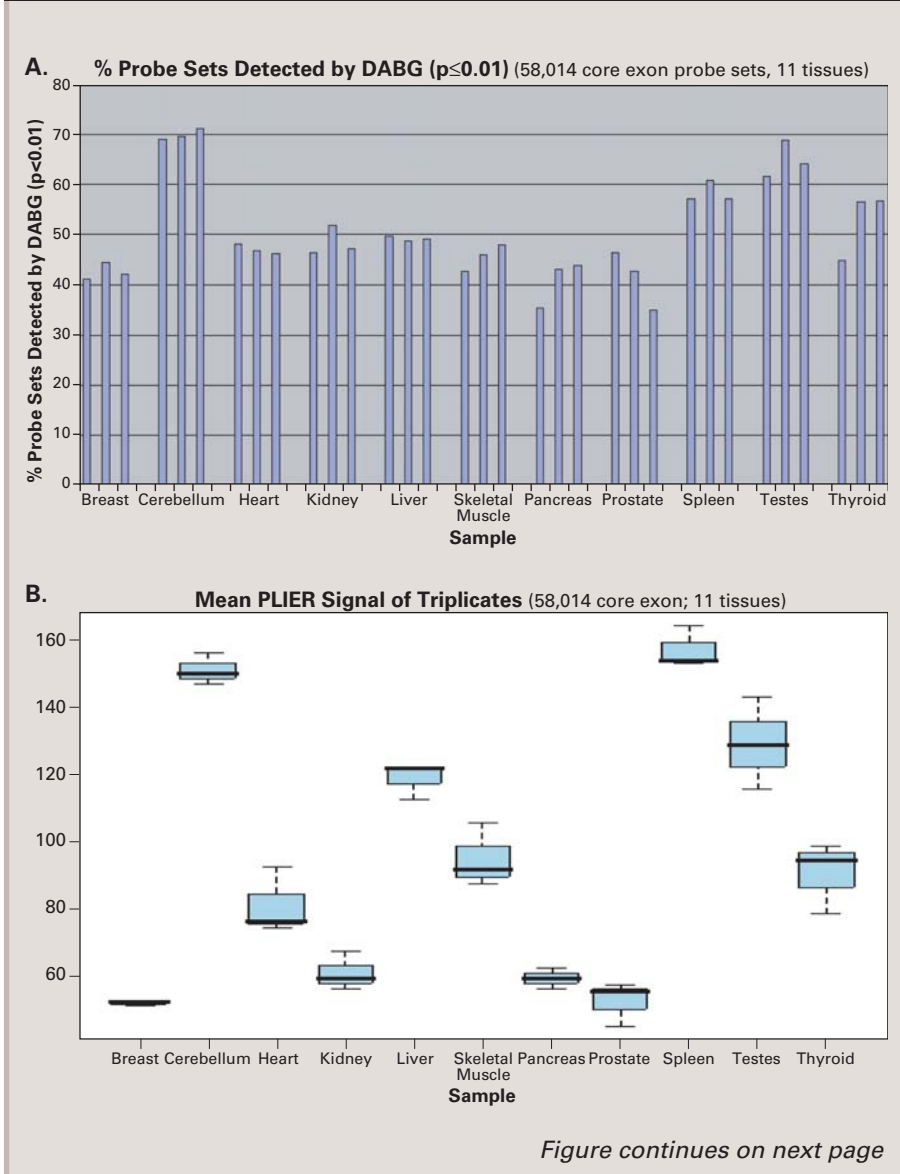
A unique design element of the GeneChip Human Exon 1.0 ST Array is the inclusion of probe sets for both intron and exon sequences from the 100 housekeeping genes, previously used as alternative normalization controls on the GeneChip® Human Genome U133 Arrays.

The signal values generated from ExACT for the exon and intron probe sets effectively serve as qualitative positive and negative controls for a set of experiments. Pseudo-ROC curves can also be generated and the

area under the curve (AUC) calculation for these control genes is provided as a metric in the Quality Report from ExACT.

Figure 7 illustrates the various QC metrics for the tissue panel hybridized to GeneChip® Human Exon 1.0 ST Arrays, including RLE, mean probe set PLIER target response, mean probe set absolute PLIER residuals, percent probe sets detected above background (DABG), and AUC from the exon-intron pseudo-ROC curve. Although metrics vary considerably between tissues, the within-tissue type technical replicates are expected to demonstrate consistent results, in general.

Figure 7. Overall array quality assessment metrics. Sense DNA target was prepared according to the protocol using 1 μ g each from human breast, cerebellum, heart, kidney, liver, skeletal muscle, pancreas, prostate, spleen, testes, and thyroid total RNA. Three sample preparation replicates were carried out for each starting total RNA. Analysis was carried out in ExACT to generate PLIER signal estimates. CEL files were quantile normalized within each tissue, then median normalized across all tissues prior to carrying out the probe summarization step. QC metrics reported by ExACT are shown, including percent probe sets detected (DABG), mean probe set PLIER target response, mean probe set absolute PLIER residuals, mean probe set absolute PLIER RLE, and Positive vs. Negative (exon-intron) ROC AUC.



Detection of Alternatively Utilized Exons

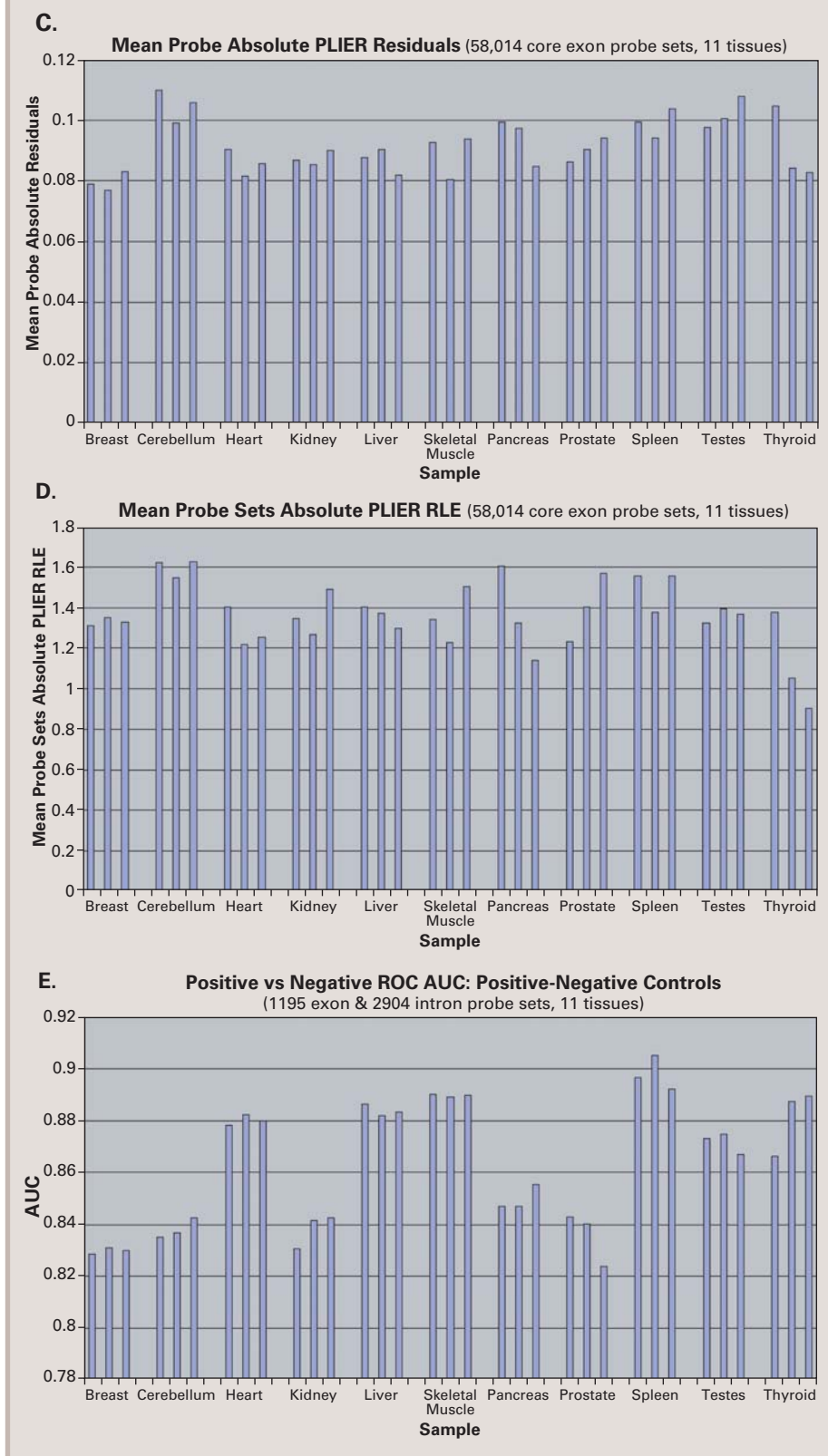
One of the advantages of using the exon arrays is to carry out expression analysis at the exon level, therefore revealing critical information on alternative utilization of exons, or alternative splicing events. As a first rudimentary proof-of-principle check from the tissue panel data discussed previously, it was presumed that different splice isoforms may be present in different tissue types.

The TPM2 gene has two well-documented predominant isoforms known to be expressed in smooth or skeletal muscles. The structures of the two transcripts from TPM2 are visualized in IGB, as shown in Figures 8A and 8B, in the track labeled "RefSeq."

Data from heart (red) and skeletal muscle (blue) on the TPM2 gene are shown in the tracks above and below the RefSeq track. The expression level is notably higher in muscle than in heart sample. It is evident that preferential utilization of the documented cassette exons in the literature can be reproduced from the array data, as shown in the highlighted regions. In addition, the array results may also reveal novel intronic retention expression in certain tissues that have not been previously identified (i.e., Expression within the 3' intron relative to the first RefSeq transcript in the red track in Figure 8A is not supported by RefSeq, but is supported by Vega annotations [data not shown]).

For details about additional algorithms that are being assessed for their utility in identifying alternative splicing events, refer to the Affymetrix white paper "Alternative Transcript Analysis Methods for Exon Arrays" available on www.affymetrix.com.

Figure 7. Overall array quality assessment metrics. (continued)



Conclusions and Discussion

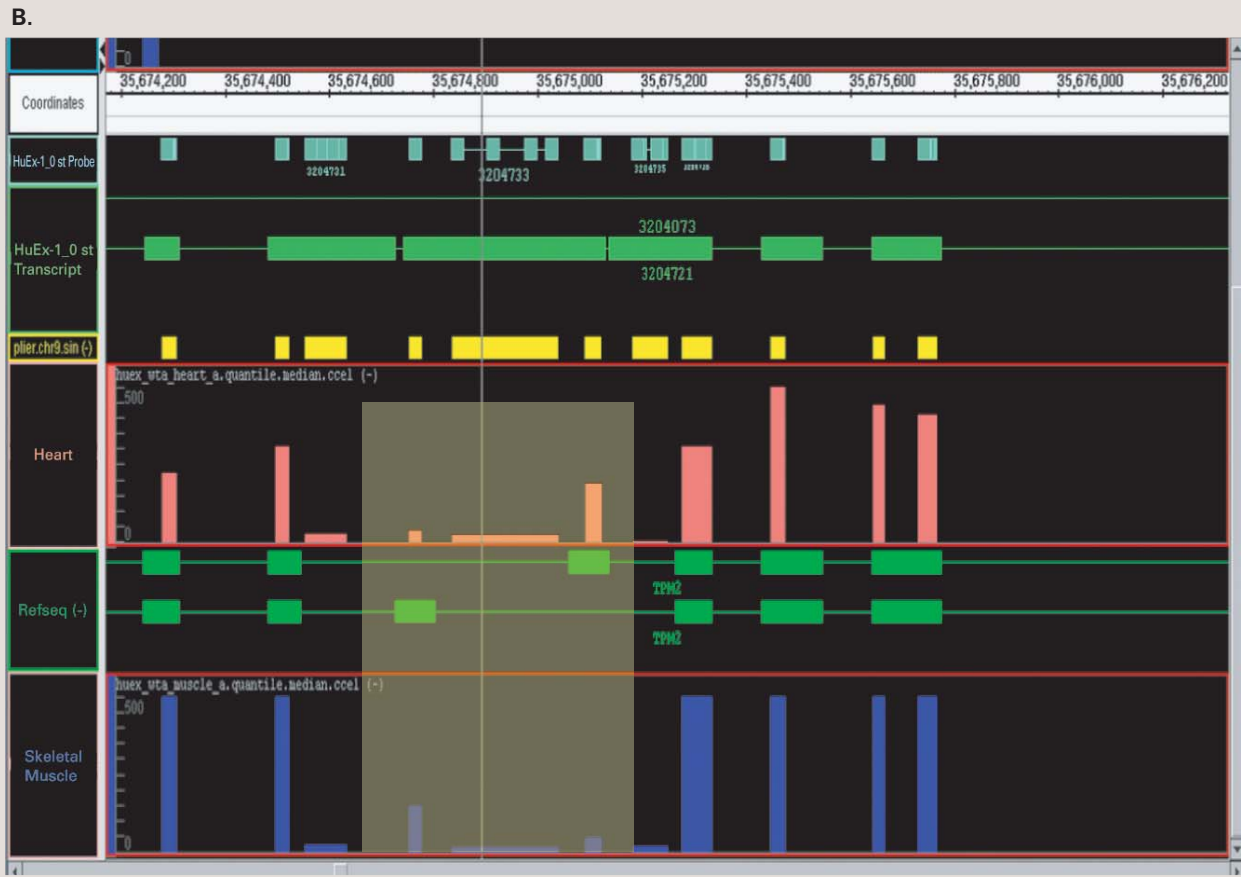
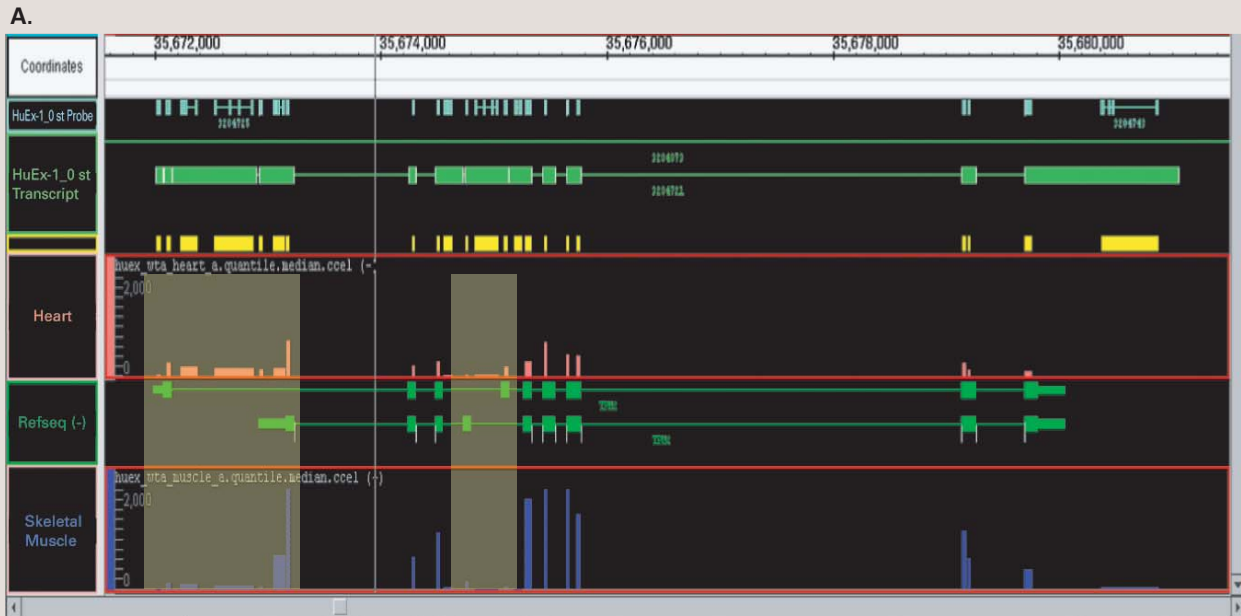
This Technical Note describes the basic performance characteristics of the new system, including the GeneChip WT Sense Target Labeling and Control Reagents and the GeneChip Human Exon 1.0 ST Array, for genome-wide, exon-level expression and alternative splicing analysis. It has been demonstrated that when used together, the overall assay is robust, reproducible, and sensitive for detecting small changes at the low end of expression at the gene level, and to a more moderate degree at the exon level. New array quality assessment metrics and typical results are also shown.

Although some of the metrics are still rudimentary, it is anticipated that they will continue to evolve as they are applied to more biological systems. Affymetrix has now provided a basic set of tools for a new level of genomic research that was not previously possible.

Related References

- GeneChip® WT Sense Target Labeling Assay Manual
- ExACT User Manual
- GeneChip® Exon Array Design Technical Note
- White Paper “Gene Signal Estimates from Exon Arrays”
- White Paper “Alternative Transcript Analysis Methods for Exon Arrays”
- White Paper “Quality Assessment of Exon Arrays”
- White Paper “Exon Probe Set Annotations and Transcript Cluster Grouping”
- White Paper “Exon Array Background Correction”

Figure 8. Alternative isoforms of gene TPM2 showing differential expression of exons in different tissues. 8A. Screen shot from IGB illustrating two alternative RefSeq transcript isoforms from the TPM2 gene. Top three tracks show probe sets, transcript clusters, and probe selection regions. The two tracks above and below the RefSeq transcript structure show signal values from heart (red) and muscle (blue). Highlighted regions show differences in expression between heart and muscle for the exons included. 8B. Zoomed in screen shot from the same figure.





AFFYMETRIX, INC.

3420 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-DNA-CHIP (1-888-362-2447)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd

Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom
UK and Others Tel: +44 (0) 1628 552550
France Tel: 0800919505
Germany Tel: 01803001334
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

AFFYMETRIX JAPAN K.K.

Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com Please visit our web site for international distributor contact information.

For research use only. Not for use in diagnostic procedures.

Part No. 702197 Rev. 1
©2006 Affymetrix, Inc. All rights reserved. Affymetrix®,  GeneChip®, HuSNP®, GenFlex®, Flying Objective™, CustomExpress®, CustomSeq®, NetAffix™, Tools To Take You As Far As Your Vision®, and The Way Ahead™, Powered by Affymetrix™, and GeneChip-compatible™, are trademarks of Affymetrix, Inc. All other trademarks are the property of their respective owners. Array products may be covered by one or more of the following patents and/or sold under license from Oxford Gene Technology: U.S. Patent Nos. 5,445,934; 5,700,637; 5,744,305; 5,945,334; 6,054,270; 6,140,044; 6,261,776; 6,291,183; 6,346,413; 6,399,365; 6,420,169; 6,551,817; 6,610,482; 6,733,977; and EP 619 321; 373 203 and other U.S. or foreign patents.