AFFYMETRIX® PRODUCT FAMILY >

RNA ARRAYS AND REAGENTS >

AFFYMETRIX®

Technical Note

Performance and Validation of the GeneChip® Human Genome U133 Set

The GeneChip® Human Genome U133 (HG-U133) Set incorporates numerous major advances in array design with regard to sequence and probe selection methods[1]. Additionally, advanced statistical algorithms[2] for monitoring gene expression were recently introduced with the latest release of Affymetrix® analysis software, Microarray Suite, version 5.0 (MAS 5.0). As in previous product design updates, perhaps more significantly in this instance, these product refinements result in data outputs from the HG-U133 Set that differ from previous human array designs.

This document summarizes the verification and validation testing done to demonstrate the performance of the HG-U133 Set. These studies employed a number of different testing strategies over a variety of tissue and sample types. The specific performance parameters presented in this document include array sensitivity, data concordance with the HG-U95 Set, array performance as it relates to sequence selection, and tissue-specific and cell-cycle expression data. In addition, there is a discussion and demonstration of the new normalization control genes, and a performance comparison of bacterial control sequences represented by either 20 or 11 probe pairs per sequence. Collectively, these data sets exemplify the performance of the HG-U133 Set.

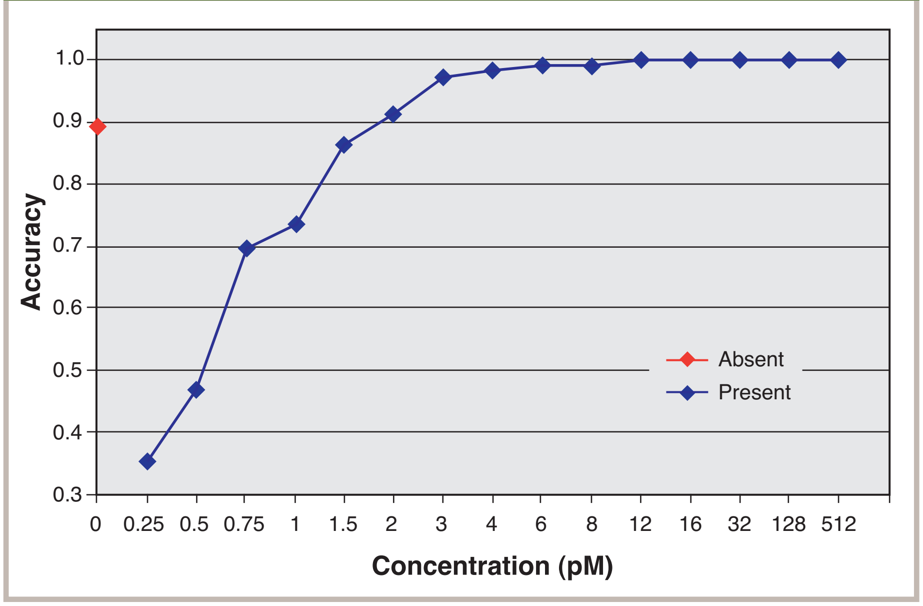
ARRAY SENSITIVITY WITH HUMAN CLONE SPIKES

One key to increasing sequence content per array on the HG-U133 Set was reducing the number of probe pairs per sequence from 16 to 11. Preliminary development work showed no significant change in array sensitivity when the 11 probe pairs per sequence, generated with the new probe selection rules, were compared to the previous 16-probe pair per sequence sets. To test the new probe selection rules as part of the HG-U133 development process, a set of human control clones was tiled using the 11-probe pairs per sequence on a prototype of the new array design. The control set contained over 50 clones that were spiked into a complex sample at known concentrations and hybridized to the prototype array to evaluate array sensitivity.

The sensitivity study was modeled after the Latin Square experiments described in the algorithm development technical note[2]. Briefly, in order to remove endogenous messages in the sample homologous to the control clones and prevent the generation of confounded data, total RNA was

Figure 1. Fifty-one labeled transcripts derived from human clones were spiked into a "subtracted" background derived from total human heart RNA. Spikes were added at the concentrations shown above (picomolar). Accuracy is defined as the ratio of spikes correctly called Present or Absent after data analysis using the MAS 5.0 detection algorithm.

Latin Square experiment demonstrating HG-U133 Array sensitivity in Detection calls.

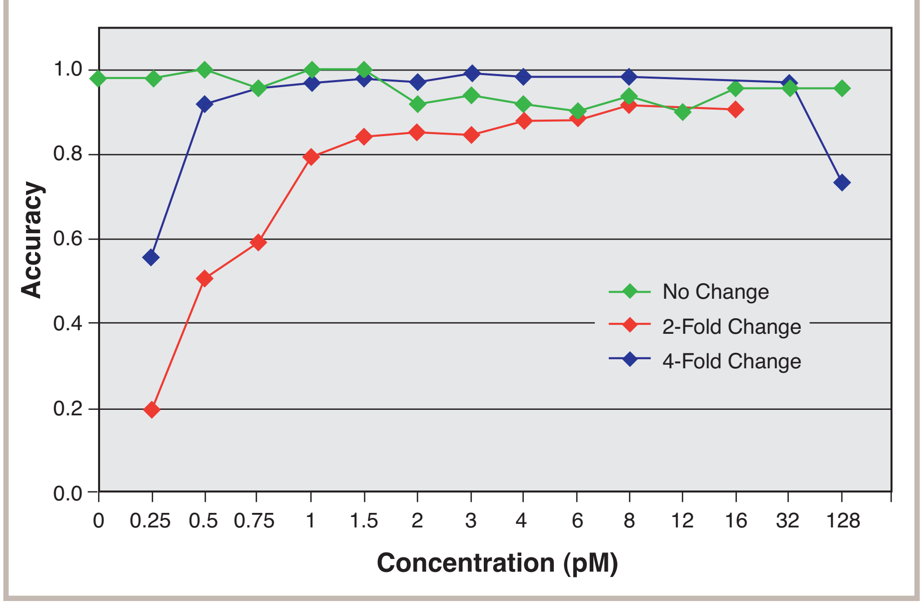


annealed to oligonucleotides hybridized to the endogenous clone sequences adjacent the poly-A tail of the mRNA. After treatment with RNase H, resultant cleaved mRNA molecules were effectively removed from the target preparation. Labeled cRNA spikes were added to the “subtracted” background RNA as groups in different concentrations, creating a Latin Square experiment. The overall sensitivity of the HG-U133 Set was determined by analyzing detection and comparison calls at the different concentrations.

Figure 1 illustrates detection call sensitivity. In this representative experiment, over 80% of the spikes are called Present at a concentration of 1.5 pM. This concentration corresponds to approximately one transcript in 100,000 or 3.5 copies per cell. Within the same experiment the false positive rate of making a Present call was roughly 10%, as noted by 90% of the clones being called Absent when not spiked into the sample (0 pM concentration). The same data were evaluated in a comparison analysis (Figure 2). Two-fold changes are detected for greater than 80% of the spikes at a baseline concentration of 1.5 pM compared to 3.0 pM. Four-fold changes are detected for greater than 80% of the spikes at a baseline concentration of 0.5 pM compared to 2.0 pM. The percentage of spikes correctly called No Change remains consistently near 100% over the concentration range tested. Detection and comparison call sensitivity remained comparable to previous designs, though the number of probe pairs per sequence was reduced from 16 to 11.

Figure 2. Same data as in Figure 1 analyzed using the MAS 5.0 comparison call algorithm. Fold changes were determined using the lower concentration as the baseline file. For 2-fold changes 0.25 pM was compared to 0.5 pM, 0.5 pM was compared to 1 pM, 0.75 pM was compared to 1.5 pM, etc. The same strategy was used to generate 4-fold change data. No change data were generated by evaluating replicate samples. Data points represent the spike concentration of the baseline file. Accuracy is defined as the (number of spikes called Increase/total number of spikes) for 2- and 4-fold changes. For the No Change data, accuracy is defined as the (number of spikes called No Change/total number of spikes).

Latin Square experiment demonstrating HG-U133 Array sensitivity in comparison calls.



CONCORDANCE TO THE HG-U95 ARRAY SET

While sequence selection methods differed between HG-U95 and HG-U133 designs, a group of 32,126 probe sets were identified as significantly related between the two designs (representing a similar region of the same transcript) based on the sequence identity and overlap of their respective probe selection regions. These probe sets are listed in the table “HG-U95 Set to HG-U133 Set, Best Match” on the Affymetrix web site, www.affymetrix.com. These related probe sets were used to examine call concordance and signal log ratios on samples run on both array sets.

Six tissues were hybridized to all arrays in the HG-U95 and HG-U133 Sets. Detection calls were obtained for the HG-U133 hybridizations using MAS 5.0 with its new statistical algorithms[2]. The HG-U95 hybridizations were analyzed using the new MAS 5.0 expression algorithms and the MAS 4.0 algorithms. Table 1 shows the results for fetal brain total RNA, which is representative of the samples analyzed. When comparing MAS 5.0 analyzed data from the HG-U95 and HG-U133 arrays, the percentage of concordant calls (Present or Absent in both samples) was 80%. When HG-U95 data analyzed in MAS 4.0 were compared to the HG-U133 data analyzed in MAS 5.0, the concordance dropped slightly to 78%. Discordant calls were more likely to be called absent in HG-U133 and present in HG-U95 (regardless of algorithm).

*The overall result users should expect to see is a lower percentage of Present calls when making comparisons between HG-U95 and HG-U133 data**.

*The HG-U95 probe sets included on the HG-U133 Set make up roughly 1% of the current design. The analysis of probe sets representing UniGene clusters with greater than 50 members identified a set of discordant probe sets over a set of ten tissues. In these cases the HG-U95 probe set was called Present in at least one tissue while the corresponding HG-U133 probe set was called Absent in all of the tissues. The discordant probe sets follow the general trend of having high signal values in the HG-U95 design with low signal in the HG-U133 design. A conclusive resolution to the apparent discrepancy is currently under investigation. To mitigate any potential loss of valid data in the HG-U133 design, both probe sets are included in the current design.

Table 1. The data were derived from total fetal brain RNA. HG-U133 calls were derived using Affymetrix® Microarray Suite (MAS) 5.0 software. HG-U95 calls were derived using both MAS 4.0 and 5.0 software. Over the six tissue set, concordant calls (Present on both arrays or Absent on both arrays) ranged from 79-84% (HG-U95 analyzed with MAS 5.0) and 76-82% (HG-U95 analyzed with MAS 4.0). To create this table, Marginal calls were considered Absent.

Call concordance between HG-U95 and HG-U133 probe sets.

		HG-U133 (MAS 5.0)	
		Present	Absent
HG-U95 (MAS 5.0)	Present	6,062 (19%)	3,616 (11%)
	Absent	2,678 (8%)	19,770 (62%)
HG-U95 (MAS 4.0)	Present	6,392 (20%)	4,563 (14%)
	Absent	2,348 (7%)	18,823 (59%)

To analyze comparisons of signal log ratio, the metric used to measure change in transcript concentration between two samples was a subset of the Best Match table. Probe sets found on the HG-U95Av2 array with a corresponding probe set on HG-U133A array (10,507 total) were used for the analysis. In Figure 3, brain and kidney total RNA samples were hybridized to HG-U95Av2 and HG-U133A arrays. Signal log ratios were generated from a tissue-to-tissue comparison analysis in MAS 5.0. Signal log ratios from probe sets identified as matches in Table 1 were then plotted. The first graph (data points in green) shows the correlation for all probe sets while the data in the second graph (data points in red) show the correlation for probe sets called Present in both tissues. The correlation greatly improves when the probe sets were restricted to only those called Present (r-squared value of 0.89 vs. 0.54).

Figure 3. Brain and kidney total RNA samples were hybridized to HG-U95Av2 and HG-U133A arrays. A comparison analysis (MAS 5.0) was performed on data from the same array using the brain sample as the baseline. Signal log ratios generated from the previous analysis were plotted as scatter graphs aligning HG-U95Av2 probe sets with their respective best match on the HG-U133A array. The top graph shows the correlation for all probe sets (r-squared value (RSQ) = 0.54, n = 10,507). The bottom graph shows the correlation for probe sets called Present in both tissues (RSQ = 0.89, n = 2,910).

Correlation of signal log ratio values between HG-U133A and HG-U95Av2 probe sets.

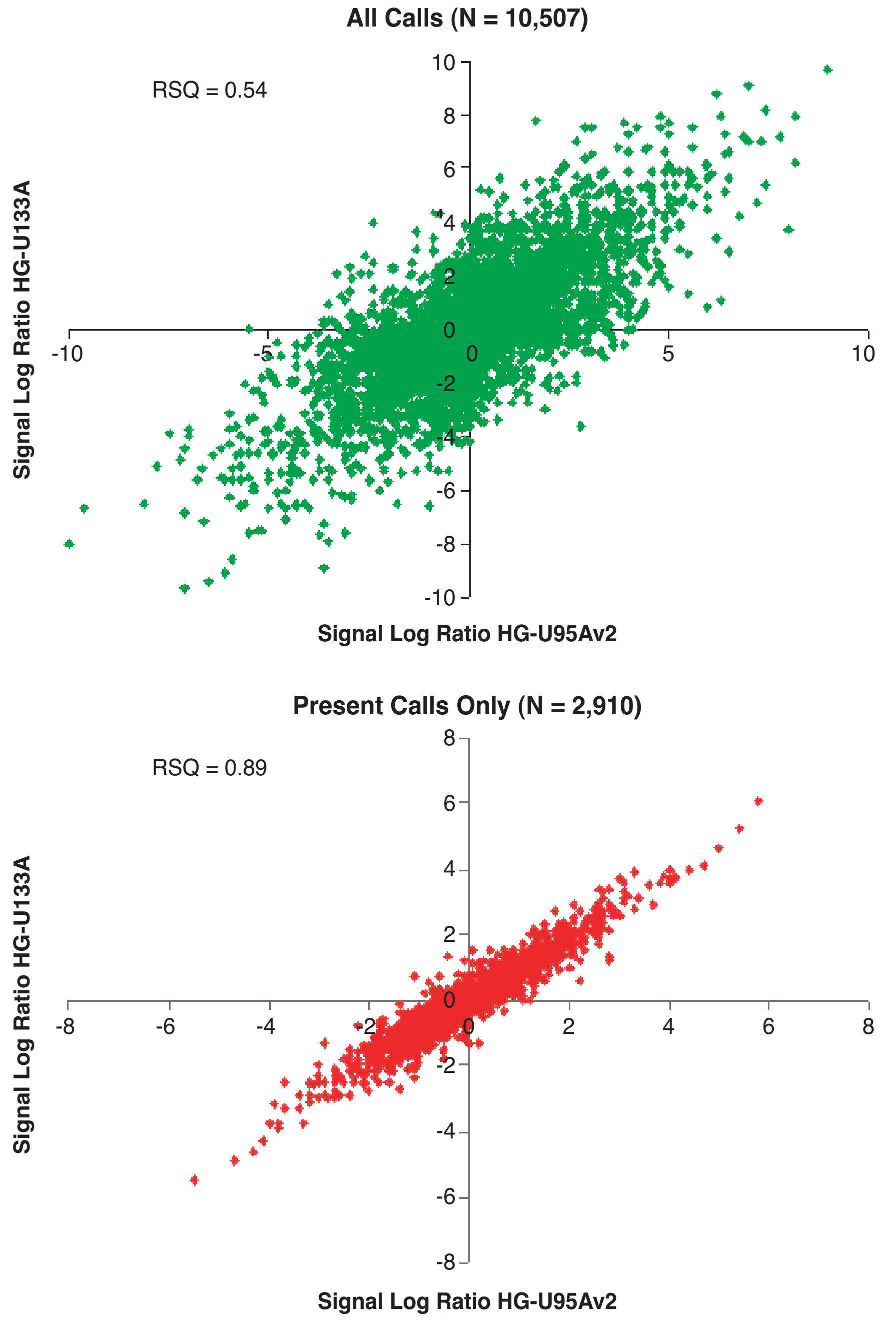


Table 2. A probe set is counted as Present if it was called Present in at least one of the tissue or cell line samples. The probe set classifications are as follows: potential full length sequences with 3′ UTR (Full Length Including UTR); consensus sequences where there is strong evidence for polyadenylation within 400 bases of an mRNA containing a complete CDS but lacking a 3′ UTR (Extended Full Length); consensus sequences with strong evidence for polyadenylation (Strongest Evidence for Polyadenylation); consensus sequence ends from subclusters containing a complete CDS mRNA (Complete CDS Consensus End); consensus sequence ends from subclusters containing a non-EST sequence (Non-EST Consensus End); consensus sequences with evidence for polyadenylation (Evidence for Polyadenylation). Probe sets from EST-only subclusters are grouped by cluster annotation quality and include subclusters with a minimum cluster assembly depth of 6 sequences or 3 sequences supporting the consensus end as the 3′ end of a transcript. Cluster annotations include whether the cluster is oriented (Oriented), whether the cluster maps to the draft assembly of the human genome (Mapped), and whether the cluster contains at least one 3′ EST or a sequence containing a polyadenylation site (3′). For completeness, probe sets to the opposite strand of an unknown or problematic subcluster are selected (Opposite Consensus End) and consensus ends which are over 1.2 kb from a probe set on the same strand are also selected (Distant Consensus End).

Probe set classification and detection over ten tissues and six cell lines.

	Array A			Array B			Array Set		
Classification	**Present**	**% Present**	**Total**	**Present**	**% Present**	**Total**	**Present**	**% Present**	**Total**
Full Length Including UTR	9,083	70%	13,049	969	62%	1,556	9,976	69%	14,529
Extended Full Length	89	52%	171	29	50%	58	118	52%	229
Strongest Evidence for Polyadenylation	2,488	77%	3,228	4,541	66%	6,929	7,012	69%	10,140
Complete CDS Consensus End	208	36%	570	25	34%	74	232	36%	643
Non-EST Consensus End	1,202	48%	2,526	979	36%	2,755	2,178	41%	5,278
Evidence for Polyadenylation	818	82%	993	473	79%	595	1,289	81%	1,586
EST-Only Subclusters									
Oriented, Mapped, and 3'	114	41%	279	3,572	39%	9,153	3,686	39%	9,432
Oriented and 3'	17	52%	33	204	35%	590	221	35%	623
Mapped and 3'	3	21%	14	27	36%	76	30	33%	90
3'	0	0%	0	4	18%	22	4	18%	22
Opposite Consensus End	210	31%	683	176	28%	619	385	30%	1,301
Distant Consensus End	103	59%	176	83	55%	150	186	57%	326
Total	**14,335**	**66%**	**21,722**	**11,082**	**49%**	**22,577**	**25,317**	**57%**	**44,199**

The data demonstrated that related probe sets perform similarly in a comparative sense. Relative degrees of increase and decrease are roughly the same between designs. Probe sets with Present calls typically generate signal values above background, giving more robust measurements of transcript abundance. By definition, signal values from probe sets with absent calls are sufficiently noisy so that they cannot be distinguished from 0. Therefore, signal log ratios from probe sets with Present calls in the baseline and experimental files will be more stable than signal log ratios derived from probe sets that are Absent in one or both of the baseline and experimental files. This is consistent with the data shown in Figure 3.

ARRAY PERFORMANCE COMPARED TO SEQUENCE SELECTION

As in the HG-U95 Set, the A array of the HG-U133 Set was designed to contain probe sets for the well-annotated genes. The vast majority of full-length mRNA sequences are contained on the HG-U133A array. In contrast, the majority of the EST-only clusters are represented by probe sets found on the HG-U133B array. To examine relative performance of the two arrays within the set, Detection calls were examined over a set of ten tissues and six cell lines.

For the HG-U133A array, 66% of the probe sets were called Present at least once in the sample set (see Table 2). Probe sets within the largest classification group, “Full Length Including UTR” (13,049 probe sets), were called Present 70% of the time in at least one tissue. This set contains the majority of RefSeq genes and the group was expected to contain one of the highest percentages of Present calls. While not confined to one of the arrays within the set, two groups performed equivalently to this largest group in terms of percentage Present calls. The two groups, “Strongest Evidence” and “Evidence of Polyadenylation,” were based on subclusters containing relatively large polyadenylation stacks.

In the case of the HG-U133B array, 49% of the probe sets were called Present. Probe sets within the group, EST-only subclusters found primarily on HG-U133B, were called Present 39% of the time in at least one tissue. EST-only subclusters were prioritized during sequence selection using several criteria including cluster size, polyadenylation evidence, mapping to the chromosome and orientation. The relatively high percentage of Present calls strongly suggests that the prioritization strategy selected the highest quality EST-only subclusters. Although representing sequences expressed in fewer tissues and at lower abundance levels than HG-U133A, HG-U133B contains qualified probe sets that will prove valuable in extracting the greatest amount of expression data from a single sample.

TISSUE-SPECIFIC EXPRESSION STUDIES

Most tissues are characterized by producing a set of specific mRNA transcripts. To further evaluate the performance of the HG-U133 Array Set, multiple tissues were tested to confirm tissue-specific expression patterns. HG-U133 data from different tissue sources were analyzed into self-organizing maps (SOM) using Affymetrix® Data Mining Tool 3.0. Tissue-specific genes were identified as clusters created by the SOM analysis.

Total RNA from ten human tissues was used to prepare the cRNA target, labeled and then hybridized to HG-U133A and B arrays. The tissue types examined were from the adrenal gland, adult brain, fetal brain, adult heart, fetal heart, kidney, pancreas, placenta, stomach and thyroid. Clusters representing tissue-specific expression patterns were clearly identified in the analysis as having a single peak of expression for a given tissue. Twelve clusters, averaging 970 members each, were created in the analysis of the ten tissues. Figure 4 shows four selected clusters exhibiting specific expression patterns for adult and fetal brain, adult heart, pancreas or placenta, respectively. Genes within the identified clusters were checked to confirm the tissue-specific expression patterns. For example, the top 20 probe sets (based on signal) for the pancreas cluster are shown in the table below the cluster maps (Figure 4). The list contains a number of genes whose expression is annotated as specific to the pancreas. We observed a number of pancreatic enzymes among the list of highest abundance transcripts in the pancreas tissue sample.

Fifteen genes exhibiting tissue-specific expression in one of four tissues — heart, fetal brain, pancreas and placenta — were analyzed and confirmed using real-time RT-PCR (TaqMan®). TaqMan primers and probes were selected using a 3’ bias. The real-time RT-PCR data were plotted against log signals to determine the degree of correlation.

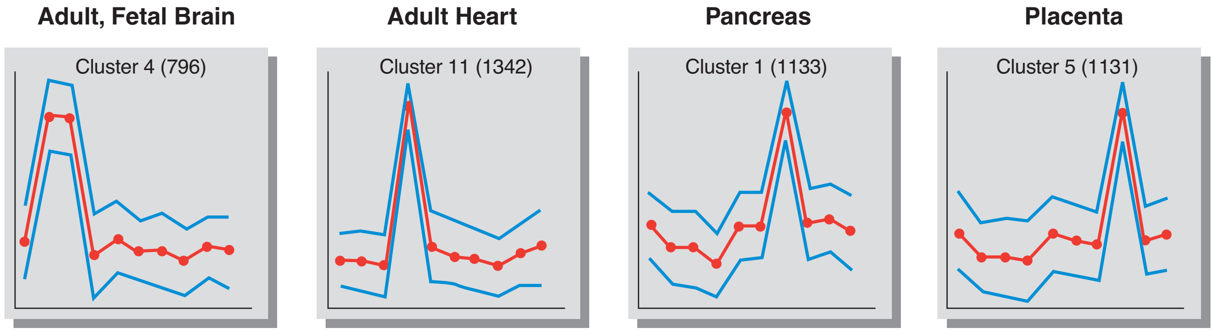
Figure 5 presents data obtained from the heart sample. Five transcripts shown to have high expression levels only in heart (myoglobin, cardiac LIM protein, troponin I, muscle creatine kinase and myosin light chain 2A) were examined along with transcripts showing tissue-specific expression patterns in fetal brain, pancreas and placenta. Change threshold (CT) values were plotted against log signal data and the expected inverse relationship was observed. Heart-specific genes (red) produced a high log signal value and a corresponding low CT value in the graph. Conversely, genes identified as tissue-specific for fetal brain (blue), pancreas (green) and placenta (gray) showed lower log signal values and higher CT values. The same fifteen genes were analyzed in either fetal brain, pancreas or placenta RNA preparations (data not shown). In these tissues the heart-specific genes showed a decrease in signal and higher CT value while the appropriate genes (depending on the tissue) increased in signal and decreased in CT value. Tissue-specific expression patterns identified by HG-U133 array analysis were consistent with gene annotations and confirmed by real-time RT-PCR in the subset tested.

CELL CYCLE SPECIFIC EXPRESSION STUDIES

In addition to examining tissue-specific expression patterns, we evaluated temporal expression patterns in a human diploid fi-

Figure 4. This figure illustrates four clusters derived from SOM analysis of ten tissue samples. The data represent samples hybridized on HG-U133A arrays and analyzed using Affymetrix® Data Mining Tool 3.0 software. The ten tissues represented by dots in the cluster graphs are from left to right adrenal, adult brain, fetal brain, adult heart, fetal heart, kidney, pancreas, placenta, stomach and thyroid. Clusters are arbitrarily numbered and the number of probe sets in each cluster is found within the parenthesis. The table lists the top 20 probe sets from the pancreas cluster based on signal.

Representative tissue-specific clusters from a Self-Organizing Map (SOM) analysis and examples of genes identified in the pancreas cluster.



Adult, Fetal Brain

Adult Heart

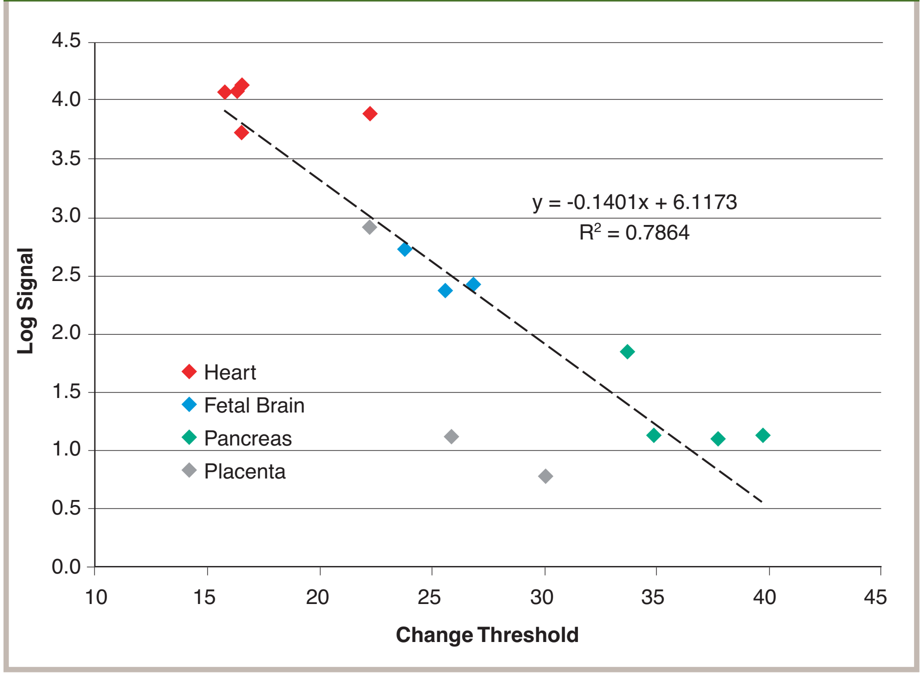
Pancreas

Placenta

Probe Set	UniGene Title	UniGene Cluster	Pancreas Signal
207412_x_at	carboxyl ester lipase-like (bile salt-stimulated lipase-like)	Hs.169271	33,378
206311_s_at	phospholipase A2, group IB (pancreas)	Hs.992	32,786
206212_at	carboxypeptidase A2 (pancreatic)	Hs.89717	32,674
205912_at	pancreatic lipase	Hs.102876	31,997
206131_at	colipase, pancreatic	Hs.1340	31,505
206297_at	chymotrypsin C (caldecrin)	Hs.8709	31,473
208473_s_at	pancreatic zymogen granule membrane associated protein GP2 beta form	Hs.274493	30,050
213421_x_at	protease, serine, 4 (trypsin 4, brain)	Hs.58247	29,796
205509_at	carboxypeptidase B1 (tissue)	Hs.180884	29,757
209752_at	regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	Hs.1032	29,590
202376_at	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	Hs.234726	29,504
205971_s_at	chymotrypsinogen B1	Hs.74502	29,093
216470_x_at	T cell receptor beta locus	Hs.303157	28,846
205615_at	carboxypeptidase A1 (pancreatic)	Hs.2879	28,798
206681_x_at	glycoprotein 2 (zymogen granule membrane)	Hs.53985	28,411
211738_x_at	Similar to elastase 3, pancreatic (protease E), clone MGC:14514, mRNA	Hs.181289	27,796
214411_x_at	chymotrypsinogen B1	Hs.74502	27,707
206239_s_at	serine protease inhibitor, Kazal type 1	Hs.181286	27,014
207463_x_at	protease, serine, 3 (trypsin 3)	Hs.278310	26,615
205910_s_at	carboxyl ester lipase (bile salt-stimulated lipase)	Hs.99918	26,104

Figure 5. Fifteen probe sets were identified as tissue-specific in array hybridizations. Corresponding TaqMan primers and probes were designed for these probe sets. The graph plots the $\log_{10}$ signal against the change threshold values derived from TaqMan analysis. The tissue-specific genes and probe sets include: Heart (red), troponin I, cardiac, 205742_at, cysteine and glycine-rich protein 3 (cardiac LIM protein), 205553_s_at, myoglobin, 204179_at, myosin light chain 2a, 219942_at, creatine kinase, muscle, 204810_s_at; Fetal Brain (blue), collapsin response mediator protein 1, 202517_at, ectodermal-neural cortex (with BTB-like domain), 201341_at, P311 protein, 201310_s_at; Pancreas (green), colipase, pancreatic, 206131_at, carboxypeptidase A1 (pancreatic), 205615_at, glucagon, 206422_at, pancreatic lipase, 205912_at; Placenta (gray), annexin A1, 201012_at, keratin 19, 201650_at, tissue factor pathway inhibitor 2, 209277_at.

Correlation of array log signal and TaqMan change thresholds for various tissue-specific transcripts in heart RNA.



broblast cell line as the cells progressed through the early phases of the cell cycle (Iyer, *et al.*, *Science* 283:83-87 (1999))[3], allowing us to verify expression of genes known to be involved in cell cycle arrest and proliferation. In order to synchronize cells, they were grown to 50-60% confluence and then serum-starved (by growing in media plus 0.1% FBS) for 48 hours. Serum was reintroduced (media + 10% FBS) and cells were harvested at 30 minutes, 2 hours and 6 hours following addition of serum containing media. Total RNA was also isolated from quiescent cells just prior to serum addition (time 0). Total RNA was used to generate cRNA target and samples were hybridized to HG-U133A arrays. The time points chosen focus on the transition of cells from an arrested state to proliferation.

Hierarchical clusters were generated using signal data (GeneMaths software 1.5, Pearson correlation, neighbor joining). Transcripts with similar temporal expression patterns were grouped together in the resulting dendrogram. Figure 6 illustrates the progressive expression patterns observed. In the absence of growth factors in serum, cells were in a low metabolic state. As cells began to proliferate, the abundance level of Group 1 transcripts decreased over time. Members of this group, including the cyclin-dependent kinase inhibitor 1B (p27, kip1), cyclin-dependent kinase inhibitor 2C, growth arrest-specific 1 (GAS1) and cyclin G2, are reported to be highly expressed during growth inhibition. Group 2 contained previously identified immediate early response transcripts, and showed transient increases in expression as early as 30 minutes after addition of serum. Transcription factors included c-FOS, FOSB, JUN, JUNB, activating transcription factor 3, early growth response 1, 2 and 3, immediate early response 3 and TGFB inducible early growth response (TIEG). We also observed increased expression for growth factors and genes involved in signal transduction. Transcripts clustered in Group 3 were those with low abundance in arrested cells, and in the early times post-serum addition, with higher abundance by 2-6 hours. This group contained additional transcription factors and signal transduction genes, but also included genes involved in inflammation and coagulation, such as IL-6, IL-8, IL1-beta and tissue factor pathway inhibitor 2. Observed expression patterns were consistent with genes known to have roles in cell cycle regulation, as well as the reported expression of genes involved in wound healing[3].

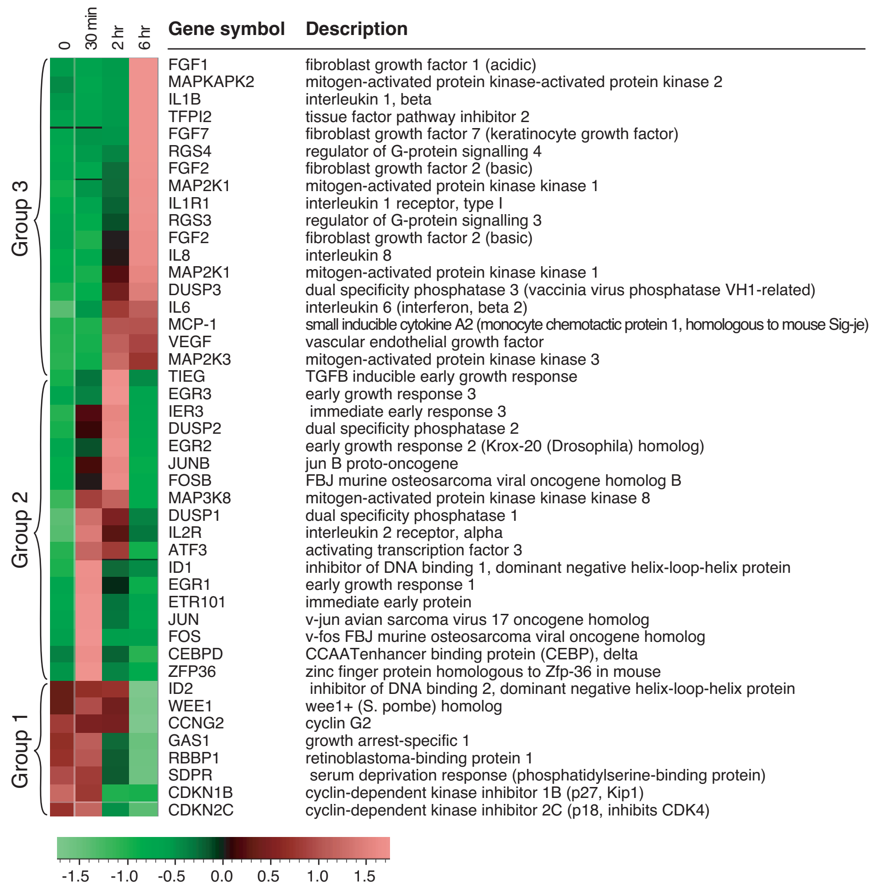
NORMALIZATION CONTROL PROBE SETS

As mentioned earlier, gene sequences are not randomly distributed between the HG-U133A and B arrays. Probe sets representing well-annotated genes are found primarily on the HG-U133A array, and as such, tend to produce higher signal values on average when compared to the HG-U133B array. Strategies to normalize array data, such as global scaling, are not always appropriate. In some instances these methods may artificially increase the actual signal values of probe sets if a common global scaling value is used, especially if overall intensities of the arrays being normalized are quite different. As an alternative means to relate signal values between arrays, a set of 100 normalization control genes were represented on both the HG-U133A and B arrays.

These normalization controls were originally identified from a data set of HG-

Figure 6. Cluster image showing expression profiles of transcripts with at least 2-fold changes relative to the 0 time point as measured by MAS 5.0 analysis. The transcripts shown were clustered hierarchically using GeneMaths 1.5 software matrix-based cluster analysis. The neighbor joining clustering method was used with Pearson correlation as the similarity coefficient. Expression levels are represented on a continuum from green (low abundance) to red (high abundance) as denoted on the scale below. Group 1: decreasing abundance over time. Group 2: increasing abundance 30 minutes and 2 hours post-serum addition. Group 3: transcripts with low abundance at early time points, increasing at 2-6 hours after serum addition.

Cell cycle expression profile.



Gene symbol	Description
FGF1	fibroblast growth factor 1 (acidic)
MAPKAPK2	mitogen-activated protein kinase-activated protein kinase 2
IL1B	interleukin 1, beta
TFPI2	tissue factor pathway inhibitor 2
FGF7	fibroblast growth factor 7 (keratinocyte growth factor)
RGS4	regulator of G-protein signalling 4
FGF2	fibroblast growth factor 2 (basic)
MAP2K1	mitogen-activated protein kinase kinase 1
IL1R1	interleukin 1 receptor, type I
RGS3	regulator of G-protein signalling 3
FGF2	fibroblast growth factor 2 (basic)
IL8	interleukin 8
MAP2K1	mitogen-activated protein kinase kinase 1
DUSP3	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)
IL6	interleukin 6 (interferon, beta 2)
MCP-1	small inducible cytokine A2 (monocyte chemotactic protein 1, homologous to mouse Sig-je)
VEGF	vascular endothelial growth factor
MAP2K3	mitogen-activated protein kinase kinase 3
TIEG	TGFB inducible early growth response
EGR3	early growth response 3
IER3	immediate early response 3
DUSP2	dual specificity phosphatase 2
EGR2	early growth response 2 (Krox-20 (Drosophila) homolog)
JUNB	jun B proto-oncogene
FOSB	FBJ murine osteosarcoma viral oncogene homolog B
MAP3K8	mitogen-activated protein kinase kinase kinase 8
DUSP1	dual specificity phosphatase 1
IL2R	interleukin 2 receptor, alpha
ATF3	activating transcription factor 3
ID1	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein
EGR1	early growth response 1
ETR101	immediate early protein
JUN	v-jun avian sarcoma virus 17 oncogene homolog
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog
CEBPD	CCAATenhancer binding protein (CEBP), delta
ZFP36	zinc finger protein homologous to Zfp-36 in mouse
ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
WEE1	wee1+ (S. pombe) homolog
CCNG2	cyclin G2
GAS1	growth arrest-specific 1
RBBP1	retinoblastoma-binding protein 1
SDPR	serum deprivation response (phosphatidylserine-binding protein)
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)
CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)

U95Av2 hybridizations representing a large number of different tissues and cell lines. The data from these probe sets shared the common characteristic of consistently being called Present while exhibiting relatively low signal variation over different sample types. Overall, the normalization controls represent a wide range of expression levels. In order to implement these genes as a normalization tool on the HG-U133 Set, corresponding probe sets in the new design were identified through BLAST analysis and potential probe sets were further screened for low signal variation over a ten-tissue sample set. By using a combination of these and additional analyses, a final group of 100 normalization probe sets was selected, which are represented on both arrays and given probe set ID numbers 200000-200099.

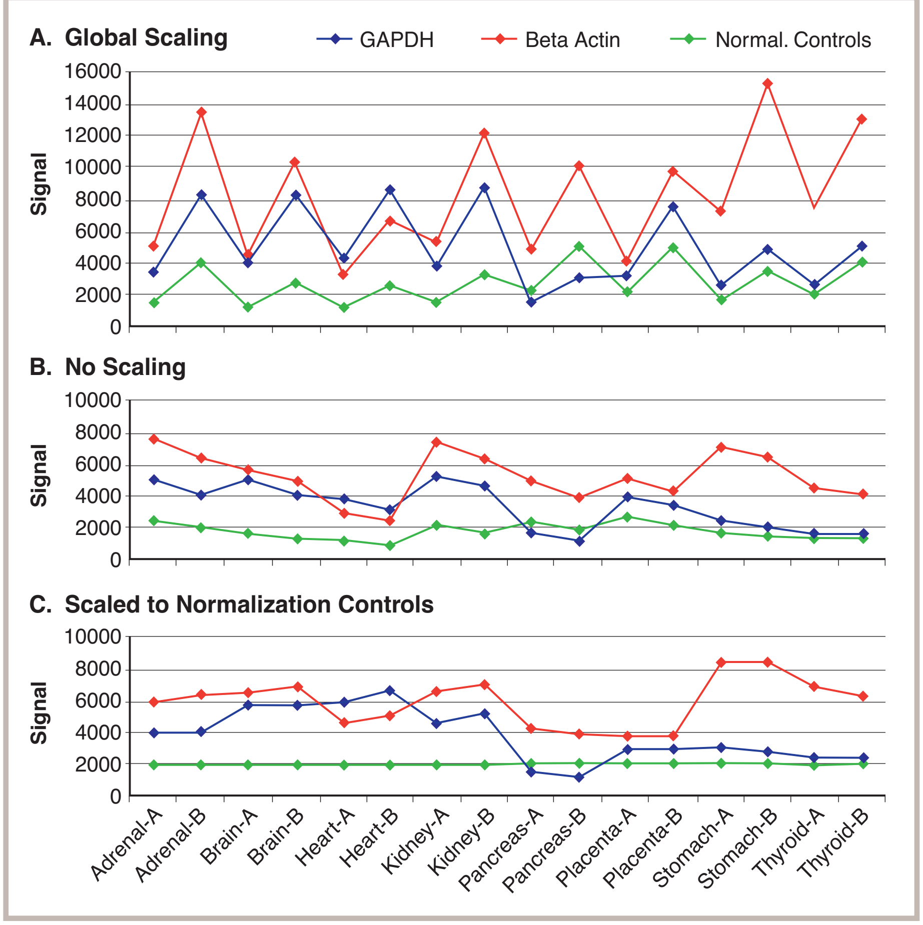
Figure 7 illustrates the use of normalization controls for scaling between the HG-U133A and B arrays. The graphs plot signals of glyceraldehyde-3-phosphate dehydrogenase (GAPDH), beta actin and the average of the 100 normalization controls over eight tissues and the two array types (GAPDH and beta actin are not part of the 100 normalization controls). After normalization, in an ideal case, the level of signal from these three measurements should be relatively consistent between the two arrays. Graph A presents the data after global scaling to a value of 150. All three signals measured are elevated in the output from the HG-U133B array illustrating that the higher overall raw signal from the HG-U133A array has skewed the normalized data on the HG-U133B array. Graph B shows the same data with no scaling applied. In this case, GAPDH, beta actin and the normalization control average signals are much closer in value for the same sample analyzed on either the HG-U133A or B array, as expected. Graph C shows the same data after scaling against the 100 normalization controls (average value set at 1,800). In this case, signal values for the HG-U133A and B arrays are more in line with each other as compared to Graph B, demonstrating that this method of normalization provides the expected scaling of the data. More importantly, scaling against normalization controls avoids the data skewing that is shown in Graph A, providing an improved alternative tool to global scaling when scaling data between the HG-U133A and B arrays. It is interesting to note that in all cases, GAPDH and beta actin signals vary between the different tissues, illustrating why these genes did not end up as part of the normalization control set.

BACTERIAL CONTROL PROBE SETS

Bacterial control probe sets historically have been placed on Affymetrix expression array designs. They serve as indicators of array quality and proper hybridization and staining, especially when a sample includes the addition of the bacterial control spikes in the form of the quality-tested GeneChip® Eukaryotic Hybridization

Figure 7. GAPDH (blue), beta actin (red) and the normalization control average (green) signals are plotted for eight adult tissues (adrenal, brain, heart, kidney, pancreas, placenta, stomach and thyroid) hybridized onto either HG-U133A (-A) or HG-U133B (-B) arrays (x axis, below bottom graph). The scaling strategy used for each graph is as follows: **A.** Global Scaling. Hybridization data are scaled to a global value of 150. **B.** No scaling. Hybridization data are left as is. **C.** Scaled to Normalization Controls. Hybridization data are scaled to the 100 normalization controls using an average signal value of 1,800.

Signals of glyceraldehyde-3-phosphate dehydrogenase (GAPDH), beta actin and the average of 100 normalization controls over eight adult tissues using different scaling strategies.



Controls. In the HG-U133 design, bacterial controls are represented by a new 11-probe pair set as well as the historic 20-probe pair set for customers who continue to use these in their own quality tests. The one exception was the *B. subtilis trp* 11-probe pair control that performed as an outlier in terms of sensitivity, and was therefore eliminated from the final design.

The performance of the new 11-probe pair design was evaluated against the previous 20-probe pairs per sequence design in Latin Square experiments. As seen in Figure 8, the 11-probe pair set produced signal values similar to the 20-probe pair set for a given concentration. Additionally, the slopes of the trend lines are nearly identical. The 11-probe pair design produced less variation between the different transcripts at the same concentration, as demonstrated by their consistently smaller error bars. Therefore, users who currently use the bacterial control spikes as a check for overall quality of their experiment should be able to make a seamless transition to the 11-probe pairs per sequence controls.

Summary

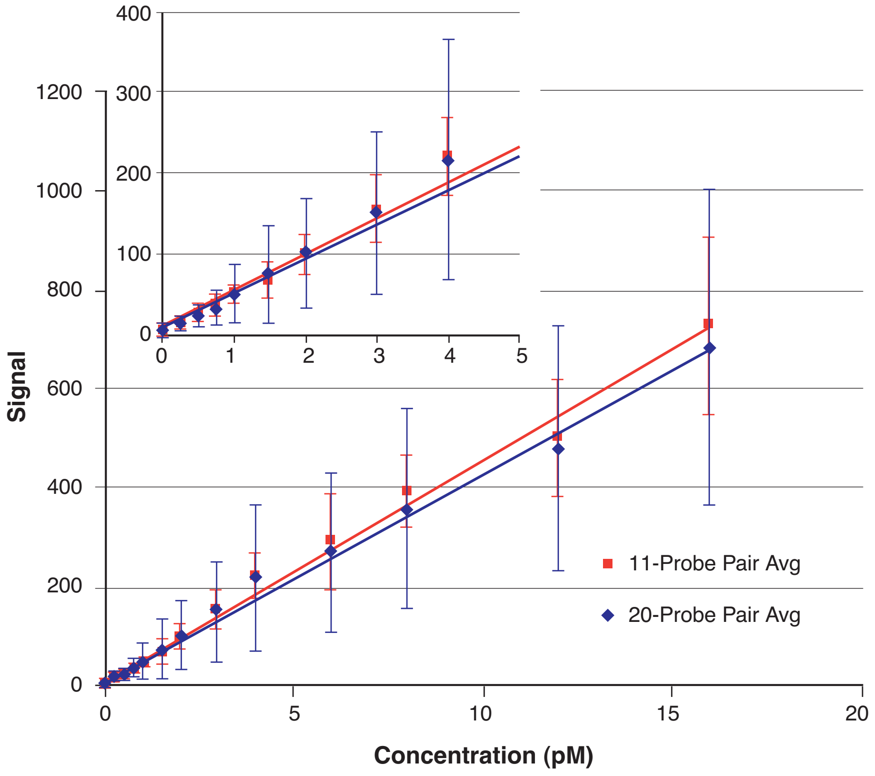
A number of biologically relevant assays were run to verify the performance of the HG-U133 Set array design. The results of this testing are summarized as follows:

- Using a set of human clone spikes, the sensitivity of the array was demonstrated at the 1.5 pM range.
- Analyzing array data over ten different tissues produced tissue-specific gene clusters as expected and served as a confirmation of array design and probe set annotations. These results were confirmed by TaqMan analysis.
- The new normalization controls are an improved alternative tool to global scaling when comparing data between arrays and experiments.
- The eleven probe-pair bacterial controls were shown to have similar performance characteristics to the previous 16-probe pair design, with reduced signal variation between the different transcripts.
- Despite changes in sequence and probe selection methods for the HG-U133 Set[1], there remains a relatively high level of concordance to its predecessor, the HG-U95 Set.

Overall, the HG-U133 Set is a powerful tool that provides the best view of transcription from the human genome.

Figure 8. The graph depicts the average signal of 18 bacterial control probe sets against concentration in picomolar (pM). Red identifies the 11-probe pair sets and blue identifies the 20-probe pair sets. Error bars represent the signal standard deviation.
The inset enlarges the view of the data between 0 and 5 pM.

Performance of 11-probe pair bacterial control probe sets relative to 20-probe pair designs.



FOOTNOTES

[1] Affymetrix Technical Note "Array Design for the GeneChip® Human Genome U133 Set."

[2] Affymetrix Technical Note "New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays."

[3] Iyer, V.R., *et al*. The transcriptional program in the response of human fibroblasts to serum. *Science* 283 (5398):83-7 (1999 Jan 1).

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 Number 701211 Rev. 2