



Technical Note

■ Array Design and Performance of the GeneChip® Mouse Expression Set 430

The mouse is a popular model system for the study of human biology and disease processes. Gene expression profiling is an important tool in understanding the genetic interactions underlying these processes. This document is an overview of the approach and parameters used in the design of the GeneChip® Mouse Expression Set 430.

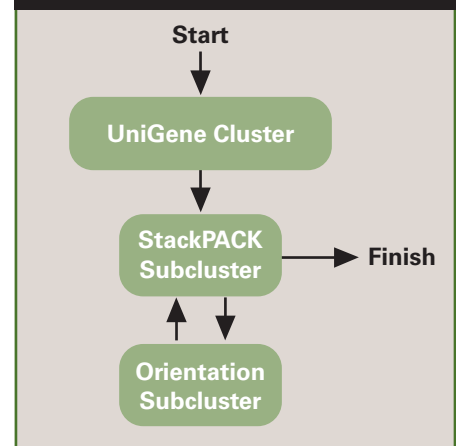
GeneChip® Mouse Expression Set 430 Design

Since the success of an array design is highly dependent on the quality and curation of sequence information, we built tools to address these issues during the design of the GeneChip® Human Genome U133 Set (HG-U133).^{*} We then leveraged the knowledge gained from this process and applied a similar method to the GeneChip Mouse Expression Set 430 (Mouse 430) design. The design modifications for the Mouse 430 are detailed below.

Various public data sources were used for the Mouse 430 design (Table 1). Sequence data were obtained from dbEST (NCBI, June 2002), GenBank (NCBI, Release 129, April 2002), and RefSeq (NCBI, June 2002). Additionally, the draft assembly of the mouse genome (Whitehead Institute Center for Genome Research, April 2002) was used to assess sequence orientation and quality. The initial sequence curation process involved:

- Collection of sequences and annotations from various public sources
- Identification and removal of vector sequences
- Sequence alignment to the mouse draft assembly
- Detection of polyadenylation sites
- Orientation of sequences, using consensus splice sites from genome alignments, detected polyadenylation sites, and CDS and EST read direction annotations

Figure 1: Sequence cluster information from UniGene was used to create initial seed clusters. Seed clusters were subclassified into one or more StackPACK subclusters with assemblies using StackPACK (Electric Genetics). Subclusters with orientation problems were further subclassified into orientation subclusters, which were then processed by StackPACK.

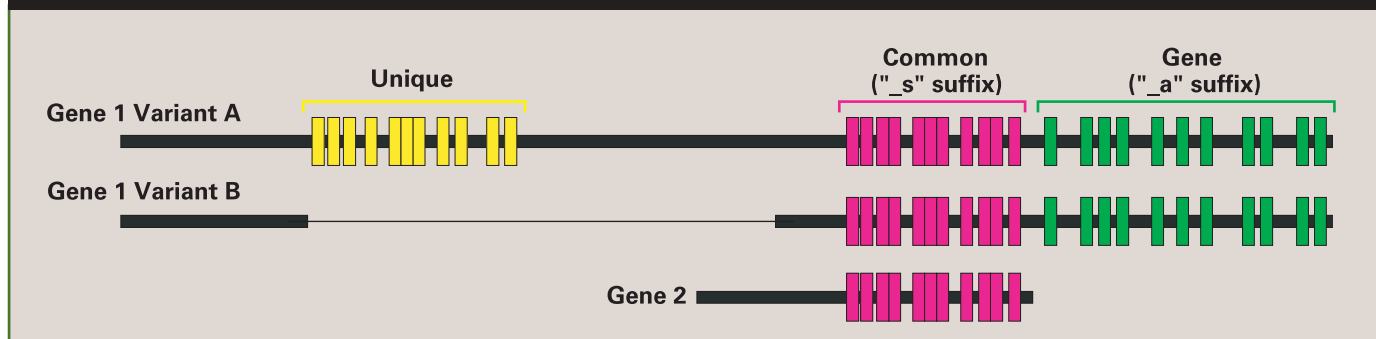


- Identification and removal of low-quality regions of EST sequences

UniGene (NCBI, Build 107) was then used to create initial clusters of cDNA sequences (Figure 1). To prevent overly complicated probe set annotations, genome-based subclustering was not performed. Sequence based subclustering was accomplished using StackPACK software (Electric Genetics). This step partitions alternative transcript isoforms into separate clusters and helps to remove problematic sequences. In some cases, sequence-based subclusters were further subgrouped due to conflicting orientation

^{*}For more information on the design process for the HG-U133 Set, please refer to the Technical Note: "Array Design for the GeneChip Human Genome U133 Set."

Figure 2: Different probe set types are indicated by suffixes to the probe set name. Unique probe sets are predicted to perfectly match only a single transcript. Gene probe sets, with an “_a” suffix, are predicted to only perfectly match transcripts from the same gene. Common probe sets, with a “_s” suffix, are predicted to perfectly match multiple transcripts, which may be from different genes. Probe sets that have a “_x” suffix are not shown here but are described in the text.



calls within the subcluster assembly. To be conservative when selecting probes, at least 75% identity in all of the member sequences was required when calling a consensus sequence.

Probes are selected from the 600 bases most proximal to the 3' end of each transcript. Probe selection regions were defined using any of the following criteria:

- 3' ends of RefSeq and complete CDS mRNA sequences (Full Length End)
- Eight or more 3' EST reads terminating at the same position (Strong Evidence for Polyadenylation)
- 3' end of the assembly (Consensus End)

This approach identifies alternative polyadenylation sites internal to the assembly end. In contrast to the HG-U133 design, the probe selection region is always selected from the consensus sequence to simplify the bioinformatics associated with data analysis. When alternative polyadenylation sites are less than 600 bases apart, only the probe selection region based on the upstream polyadenylation site is used.

In summary, sequence content for the Mouse 430 design (Table 2) was selected and prioritized based on the following rules. Probe selection regions were selected for subclusters containing:

1. Non-EST sequences
2. Only EST sequences where a transcript end is confirmed by eight or more 3' EST reads
3. Only EST sequences where the cluster consensus end coincides with two or more 3' EST reads, the cluster is oriented, and sequences from more than one cDNA library are included in the cluster

In general, probe selection regions matching rule three were only tiled when less than three probe selection regions for that UniGene cluster matched rules one and two. One other special class of sequence content was also included if they did not already meet the rules above:

- Best matching probe set for a GeneChip Murine Genome U74Av2 probe set

Probe and probe set selection were performed as described by the “Array Design for the GeneChip Human Genome U133 Set” technical note. In short, a thermodynamic multiple linear regression model was used to predict probe performance. Eleven probe pair probe sets were then selected based on predicted probe characteristics, such as performance, uniqueness metrics, and spacing rules. A new non-unique probe set type, “_a”, was added to indicate those probe sets that recognize multiple alternative transcripts from the same gene (Figure 2). Probe sets with

Source	Release Date	Sequences	Used in Design
UniGene	June 2002 (#107)	84,459	34,323
dbEST	June 2002	2,590,400	1,471,886
GenBank	April 2002 (#129)	54,154	30,887
RefSeq	June 2002	9,527	7,968
Total		2,738,540	1,545,064

Table 1. Sources and numbers of sequences used in the Mouse 430 Design. UniGene clusters were used as a starting point for the design process but were not used as the main source of sequence information. The use of primary sequence sources provided better control over the regions used and access to additional annotation information, such as sequence quality parameters from dbEST. A draft assembly of the mouse genome from the Whitehead Institute Center for Genome Research (April 2002) was used to improve cDNA sequence orientation and annotation.

Classification	Mouse Set 430	Mouse 430A	Mouse 430B
Probe Sets	45,037	22,626	22,511
UniGene Clusters	34,323	14,109	20,214
Additional Potential Full Lengths	25	25	0
Subclusters	39,015	18,116	20,899
Full Lengths	14,484	14,484	0
Full Length End and Strong Evidence for Polyadenylation	6,839	6,839	0
Strong Evidence for Polyadenylation	2,408	2,408	0
Full Length End	4,452	4,452	0
Consensus End	785	785	0
Non-ESTs (excluding Full Lengths)	9,450	3,771	5,679
Strong Evidence for Polyadenylation	3,390	1,880	1,510
Consensus End	6,060	1,891	4,169
ESTs	21,103	4,371	16,732
Strong Evidence for Polyadenylation	8,341	3,362	4,979
Library Coverage >1			
Evidence for Polyadenylation >1	11,769	260	11,509
Single Evidence for Polyadenylation	101	50	51
No Direct Evidence for Polyadenylation	145	108	37
Single Library Coverage >1			
Evidence for Polyadenylation >1	35	28	7
Single Evidence for Polyadenylation	382	297	85
No Direct Evidence for Polyadenylation	330	266	64

Table 2. Classification and number of probe sets placed on the Mouse 430. It is estimated that this set interrogates approximately 39,000 transcripts from approximately 34,000 genes. The first tier provides a summary of content with regard to the listed metrics. The second tier provides a summary of probe set content based on annotation quality. The probe sets are assigned to the classifications based on the sequence quality of the subcluster (Full Lengths, Non-ESTs, ESTs) and the justification for the region from which probes were selected (Strong Evidence for Polyadenylation, Full Length End, Consensus End). Probe sets based on EST-only subclusters were also grouped based on cDNA library coverage for the subcluster.

common probes among multiple transcripts from separate genes are annotated with a “_s” suffix. Occasionally, it is not possible to select a unique probe set or a probe set with identical probes among multiple transcripts. In this case, similarity criteria are suspended and the resulting probe set is annotated with a “_x” suffix. Such probe sets will contain some probes that are identical or highly similar to other sequences. The probe set may cross-hybridize in an unpredictable manner with other sequences, but should hybridize correctly to the main target. Data generated from these probe sets should be interpreted with caution due to the likelihood that some of the Signal measure-

ments for a subset of the probes in the probe set are from transcripts other than the one being intentionally measured.

GeneChip® Mouse Expression Set 430 Performance Improvements

COMPARISONS OF TISSUE PANEL SIGNALS FOR PROBE SET PAIRS

In order to show that the new Mouse 430 design, with reduced probe set size, produced equivalent or more informative data compared to the previous design, GeneChip Murine Genome U74v2 Set (MG-U74v2), we compared the Signal output from Affymetrix® Microarray Suite v.5.0 from both. The comparison began

with the identification of probe set pairs, one probe set from each design, that were identified as being the best representatives of overlapping probe selection regions (PSR) from the MG-U74v2 and the Mouse 430 designs. For a detailed comparison of the Mouse 430 and the MG-U74v2 designs see the Appendix. More specifically, probe sets were paired by requiring that all sixteen probes of each MG-U74v2 probe set align within the probe selection region (PSR) of its matched Mouse 430 probe set. All possible pairings were made between probe sets from the Mouse 430 Array Set (A and B) and the MG-U74v2 Array Set (A, B, and C).

In some cases, multiple probe sets from one or both array types represented the same Mouse 430 PSR. Such multiple probe sets may represent splice variant regions or alternative 3' ends. In order to prevent pairing of a probe set from one array type that represents a highly expressed region to a probe set from the other array type, that represents a rarely expressed region, we select only the *most responsive* probe set from each array type to represent the PSR. Specifically, we select the Mouse 430 probe set with the highest $NLP_T_{Mouse\ 430}$ value (see Methods) and the MG-U74v2 probe set with the highest $NLP_T_{MG-U74v2}$ value (see Methods).

We compared the relative Signal levels and relative Signal responsiveness to tissue diversity for these pairs of probe sets. Responsiveness to tissue diversity is a desired trait of the probe sets in a design because it indicates that probe set Signal

values change in response to varying levels of transcript. More intense Signals may be indicative of a better design, in terms of improved probe selection and sequence representation. Relative Signal levels are measured by counting the number of cases where a probe set from one array design produces significantly higher Signal values than a probe set from the other array design. We evaluated 12,613 PSRs represented by probe sets from both the Mouse 430 and the MG-U74v2 Sets (see Methods).

We ran the samples in triplicate across eleven tissue types on both array designs (11 tissues X 3 replicates X 2 array types). Single array analysis was performed on each experiment and Signal values were used to generate scatter plots. These histograms allowed a visual comparison of the Signal values produced between the paired Mouse 430 and MG-U74v2 probe sets. In

Figure 3, each point is a $\log(\text{Signal})$ value (y axis) for one of the eleven tissue types (x axis) produced by a MG-U74v2 probe set (red) or a Mouse 430 probe set (blue).

EFFECT OF PROBE SET DESIGN ON SIGNAL LEVELS

This analysis indicates that the majority of Mouse 430 probe sets tend to produce higher Signal values relative to corresponding probe sets on the MG-U74v2 Set. The metric for relative magnitude of Signals, NLP_PrbSet , is the negative log of the probability that the probe set design has no effect on the magnitude of $\log(\text{Signal})$ values for a probe set pair (see Methods). NLP_PrbSet values increase as the mean of $\log(\text{Signal})$ values produced by the one probe set increasingly differs from the mean of $\log(\text{Signal})$ values produced by the second probe set over the tissue panel. In other words, as NLP_PrbSet increases so does the probability that the array design

Figure 3. Log(Signal) profiles for three probe set pairs. The x-axis represents the tissue types, while the y axis represents the $\log(\text{Signal})$. Each point is the $\log(\text{Signal})$ value produced by a probe set for one of eleven tissues: 1=brain, 2=embryo, 3=heart, 4=kidney, 5=liver, 6=lung, 7=muscle, 8=ovary, 9=spleen, 10=testicle, 11=thymus. Three replicate experiments for each tissue type and array type produce three profiles for the Mouse 430 Set probe set (blue) and three profiles for the paired MG-U74v2 Set probe set (red). Definitions of NLP_PrbSet , $NLP_T_{Mouse\ 430}$ and $NLP_T_{MG-U74v2}$, are given in the text and in the Methods. **A.** Probe Set Pair with $NLP_PrbSet=1.5$; Mouse 430 probe set is 1426773_at ($NLP_T_{Mouse\ 430}=14.3$); MG-U74v2 probe set is 95594_at ($NLP_T_{MG-U74v2}=16$). **B.** Probe Set Pair with $NLP_PrbSet=14$; Mouse 430 probe set is 1420967_at ($NLP_T_{Mouse\ 430}=16$); MG-U74v2 probe set is 104007_at ($NLP_T_{MG-U74v2}=16$). **C.** Probe Set Pair with $NLP_PrbSet=16$; Mouse 430 probe set is 1422476_at ($NLP_T_{Mouse\ 430}=16$); MG-U74v2probe set is 97444_at ($NLP_T_{MG-U74v2}=16$).

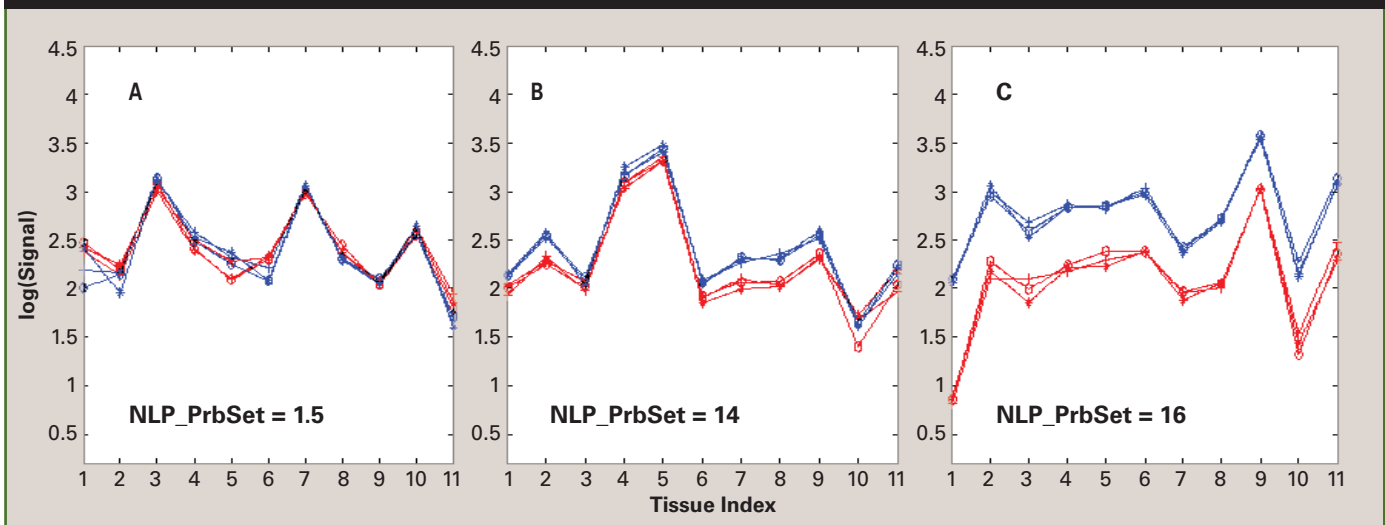
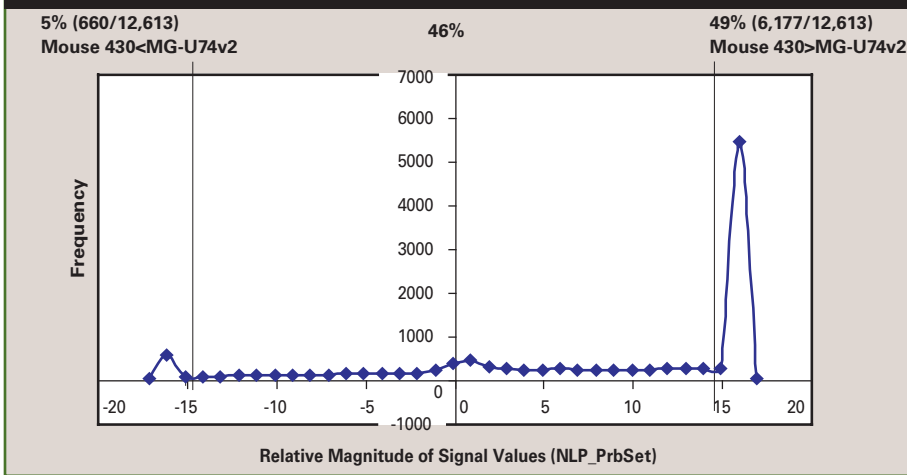


Figure 4: Relative Magnitude of Signal values (NLP_PrSet). Bars indicate the boundaries of significance cutoffs.



affects the magnitude of the Signals. For clarity in viewing the data we set the NLP_PrSet to a negative value if the MG-U74v2 probe set Signals are greater than the Mouse 430 probe set Signals on average for that probe set pair.

Figures 3, panels A-C show how the absolute NLP_PrSet values increase as the MG-U74v2 and Mouse 430 $\log(\text{Signal})$ profiles resolve (note that the blue curves and red curves become separated). A probe set pair, whose NLP_PrSet value equals the maximum value of sixteen (Figure 3C, corresponding to p -value=0) produces blue curves that are well above the red curves. The purpose of the relative Signal analysis is to detect cases where the probe set design causes most Signal values to clearly resolve. As a result, we have selected a stringent cutoff for significance, which requires absolute NLP_PrSet values to be greater than fifteen. These cutoffs were used in generating the results shown in Figure 4.

Figure 4 shows the distribution of the NLP_PrSet values for the 12,613 probe set pairs. The bars separate the cases, where a probe set from one array type produces significantly higher Signal values (absolute NLP_PrSet values are greater

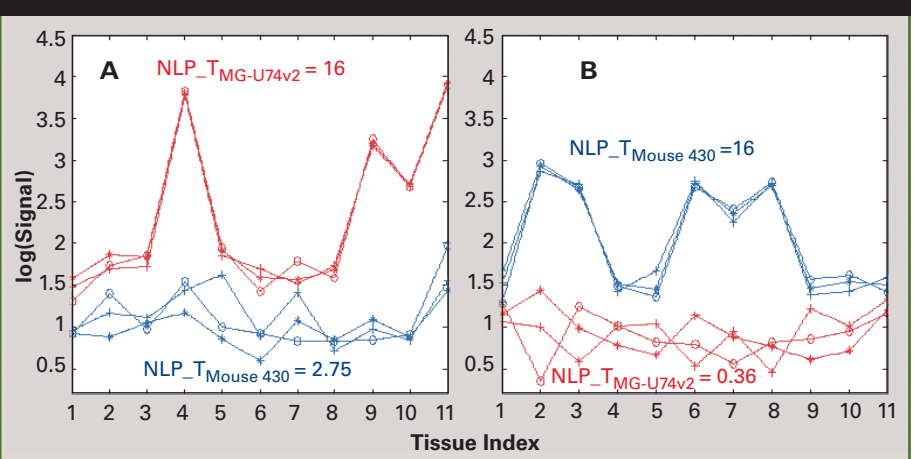
than fifteen) than a probe set from the other array type. Given this separation, 49% (6,177/12,613) of the probe set pairs have significantly higher Mouse 430 probe set Signals. Only 5% (660/12,613) of the probe set pairs have significantly higher MG-U74v2 probe set Signals. When Signal values are evaluated along with responsive-

ness to tissue diversity, however, only a small fraction, 99 of the 660 MG-U74v2 probe set pairs, were found to be potentially discordant and more informative for expression profiling out of the total number probe set pairs evaluated. This concept will be discussed further in the *Discordant Probe Set Pairs* section.

EFFECT OF PROBE SET DESIGN ON RESPONSE TO TISSUE DIVERSITY

Responsiveness to tissue diversity is a desired outcome of an array design because it indicates that probe set Signals change in response to real variation of transcript levels. This analysis indicates that there is a slight skewing in favor of the Mouse 430 design producing more responsive probe sets. The values of $NLP_T_{\text{Mouse 430}}$ and $NLP_T_{\text{MG-U74v2}}$ represent the metrics (we will refer to these values generically as NLP_T) for responsiveness of each probe set design. Each is the negative log of the probability that the tissue type has no effect on the magnitude of $\log(\text{Signal})$ values for the given probe set type (see

Figure 5: $\log(\text{Signal})$ profiles for two discordant probe set pairs. Each point is the $\log(\text{Signal})$ value produced by a probe set for one of eleven tissues: 1=brain, 2=embryo, 3=heart, 4=kidney, 5=liver, 6=lung, 7=muscle, 8=ovary, 9=spleen, 10=testicle, 11=thymus. Three replicate experiments for each tissue type and array type produce three profiles for the Mouse 430 probe set (blue) and three profiles for the paired MG-U74v2 probe set (red). NLP_PrSet, $NLP_T_{\text{Mouse 430}}$ and $NLP_T_{\text{MG-U74v2}}$, is described in the text. **A.** Probe Set Pair with NLP_PrSet=-16. Mouse 430 probe set is 1420701_at ($NLP_T_{\text{Mouse 430}}=2.75$). MG-U74v2 probe set is 94716_f_at ($NLP_T_{\text{MG-U74v2}}=16$). **B.** Probe Set Pair with NLP_PrSet=16. Mouse 430 probe set is 1424807_at ($NLP_T_{\text{Mouse 430}}=16$). MG-U74v2 probe set is 161793_at ($NLP_T_{\text{MG-U74v2}}=0.36$).



Methods). NLP_T values increase as the mean of $\log(\text{Signal})$ values produced by each tissue type increasingly differs from the means of $\log(\text{Signal})$ values produced by the other tissue types. In other words, the higher the value of NLP_T, the more variable the $\log(\text{Signal})$ values are across the tissue panel, or the more responsive the probe set is to tissue diversity. Figure 5 provides the NLP_T values for two probe set pairs, where the probe sets from the two array designs have different degrees of responsiveness. The $\log(\text{Signal})$ values of the upper curves vary significantly with regard to at least one tissue, producing NLP_T values of 16. However, the $\log(\text{Signal})$ values for lower curves are essentially flat with regard to the variation across the replicate experiments, producing low NLP_T values of 2.75 (Figure 5A) and 0.36 (Figure 5B).

We analyzed the distribution of relative responsiveness to determine if there is an overall trend towards one array design or the other producing more responsive probe sets. We set relative responsiveness to be the difference between NLP_T values of the probe sets in a pair: $(\text{NLP_T}_{\text{Mouse 430}}) - (\text{NLP_T}_{\text{MG-U74v2}})$, and generated the distribution of relative responsiveness values for the 12,613 probe set pairs. The shape of the resulting distribution indicates the degree to which the array designs have an effect on this NLP_T metric, or bias it in one direction or the other. If the only source of differences in response to tissue diversity is random experimental variation, then the shape of the distribution will be normal, or bell shaped, and centered about zero. Skewing in either direction suggests that the array designs contribute to the differences in responsiveness.

Figure 6 shows that the distribution of relative responsiveness for 12,613 probe set pairs (blue curve) is centered about

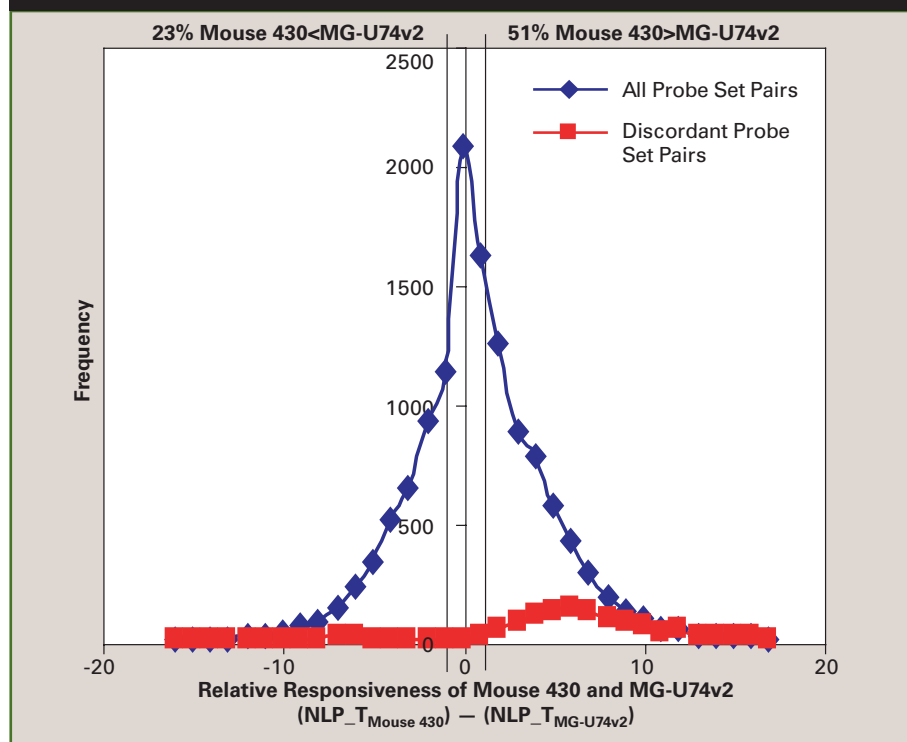
zero, with a bias towards Mouse 430 probe sets being more responsive. 16% of the probe set pairs have equally responsive MG-U74v2 and Mouse 430 probe sets, 51% have more responsive Mouse 430 probe sets, and 23% have more responsive MG-U74v2 probe sets. Random experimental variation is expected to produce non-zero values in both directions. The fact that percentages are not equal (51% vs. 23%) suggests that differences between the Mouse 430 design and MG-U74v2 design may also contribute.

DISCORDANT PROBE SET PAIRS

In this section, we compare counts of discordant cases (see Methods) where the probe set from one design appears to produce significantly more information for expression profiling relative to the paired probe set from the other array design. We

define these discordant cases as those where a probe set from one array design not only produces significantly higher $\log(\text{Signal})$ values ($\text{NLP_PrbSet} > 15$) but is considered to be responsive to tissue diversity ($\text{NLP_T} > 11$), while the probe set from the second array design not only produces significantly lower $\log(\text{Signal})$ values but also produces a response to tissue diversity that falls below the threshold ($\text{NLP_T} < 11$). The probe set producing the higher and more responsive Signals is counted as being more informative and, therefore, better for expression profiling. The probe set pairs in Figure 5 (discussed previously) are examples of cases that are counted as discordant. In contrast, the probe set pairs in Figure 3C are not counted in the discordant category, despite the significant difference in $\log(\text{Signal})$ levels, because both probe sets are responsive to

Figure 6: Relative Responsiveness: difference $(\text{NLP_T}_{\text{Mouse 430}} - \text{NLP_T}_{\text{MG-U74v2}})$ between Mouse 430 and MG-U74v2 response to tissue diversity. Distributions are generated for all probe set pairs evaluated (12,613 blue curve) and for 1,144 discordant (defined in the text and in the Methods) probe set pairs (red curve). The bars bracket the cases for which Relative Responsiveness is zero (i.e., the response for both designs is equivalent).



tissue diversity and, therefore, should be informative for expression profiling.

Only 9.1% (1,144/12,613) of probe set pairs fall into the discordant category. For the probe set pairs within this category, the Mouse 430 design is 11 times (1,045/1,144 vs. 99/1,144) more likely to exhibit characteristics of a superior, more informative probe set and only rarely produces an inferior one. The distribution of relative responsiveness of the 1,144 discordant cases is shown in Figure 6 (red curve). Since the Mouse 430 design is more likely to produce a responsive probe set among the discordant cases, the bulk of the discordant cases fall on the right side of the distribution.

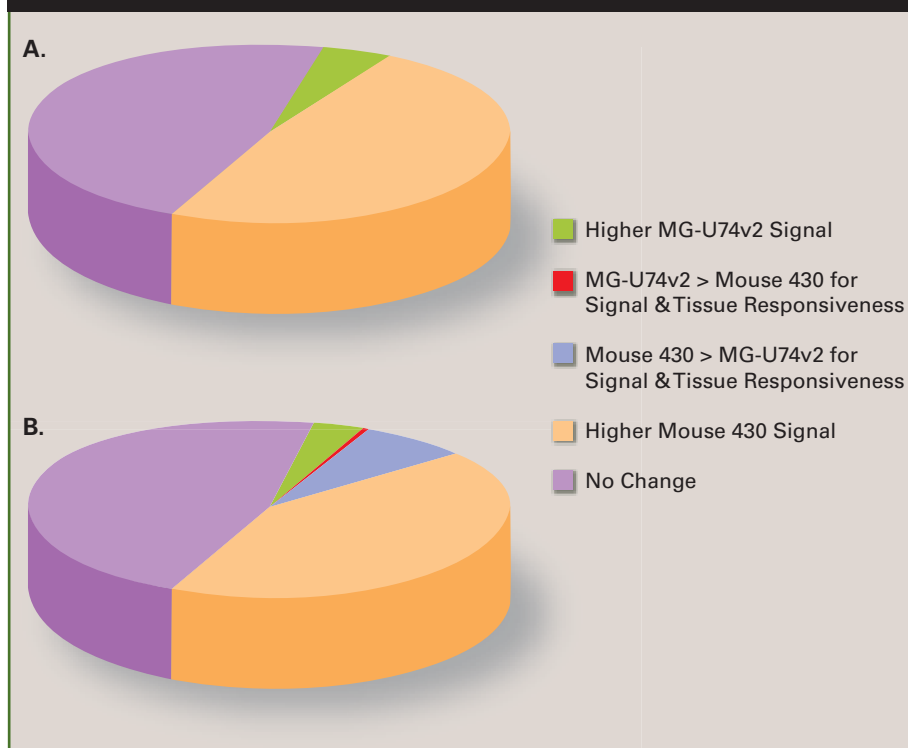
In summary, Figure 6 illustrates that both array designs perform in an equivalent manner for the majority of probe set pairs. In cases of discordant probe set pairs, where we examined both the magnitude of Signal values combined with tissue responsiveness, the Mouse 430 demonstrated superior performance compared to the MG-U74v2.

SUMMARY OF GENECHIP® MOUSE EXPRESSION SET 430 PERFORMANCE IMPROVEMENTS

Signal values tend to be higher for the majority of Mouse 430 probe sets relative to the corresponding probe sets on MG-U74v2. Although differences were observed in the magnitude of Signal values between arrays, we expanded our investigation to explore how these probe set pairs perform across a biologically diverse set of tissues. Our results show that the Mouse 430 probe sets, on average, exhibit greater responsiveness to diverse tissue types. In addition, Mouse 430 probe sets are more likely to outperform a MG-U74v2 probe set when both magnitude of Signal and responsiveness to a biologically diverse tissue panel are evaluated. This conclusion is illustrated in Figure 7. In panel A, it is evident that just less than half of the probe set pairs have equivalent Signal values between the array types. There are cases where the Mouse 430 Signals are higher (49%) or, in

Figure 7: Percentage of discordant probe set pairs and effect of probe set type on Signal levels.

- A.** Displays the percentage of probe set pairs that have higher Signal values for Mouse 430 probe sets (orange) compared to the percentage of probe sets with higher Signal values in MG-U74v2 probe sets (green). Purple shows the percentage of probe set pairs where Signal values are not significantly different between the two array types.
- B.** Illustrates the percentage of discordant probe set pairs, as a fraction of the Signal values. Discordant probe set pairs are cases where a probe set pair in one array outperforms the matched probe set pair in the other array, with respect to both magnitude of Signal and tissue responsiveness. For the probe pairs within this category, the Mouse 430 probe set pairs are 11 times more likely to show characteristics of a superior probe set (blue) and only rarely produce an inferior one (red).



contrast, the MG-U74v2 Signals are higher (5%). In panel B, we see the overlap in the number of cases where both Signal and tissue responsiveness are considered. This highlights cases where Signal plus tissue responsiveness are better in one array compared to the other. In these cases, Mouse 430 is more likely to have a superior outcome for both the quantitative measurement and the biological effect compared to its partner probe set pair on the MG-U74v2 design.

Methods

ANALYSIS OF VARIANCE (ANOVA)

We use Two-Way ANOVA to compare the magnitude of Signals produced by two probe sets using the entire tissue panel, and compute NLP_{PrbSet} . We use One-Way ANOVA to analyze Signals produced by each probe set independently and compute the *Responsiveness* of each probe set type to tissue diversity: $NLP_{T_{Mouse\ 430}}$ and $NLP_{T_{MG-U74v2}}$.

NLP_PrSet

We produce a p -value, p_{PrSet} , for the probability that the array design of the probe sets has no effect on the magnitudes of the 66 (11 tissues X 3 replicates X 2 array designs) $\log(\text{Signal})$ values for a probe set pair. Specifically, p_{PrSet} is the p -value produced by a Two-Way ANOVA (factor one is probe set design, and factor two is tissue type) for the null hypothesis

$$H_0: \text{mean}_{\text{Mouse 430}} = \text{mean}_{\text{MG-U74v2}}$$

against the alternate hypothesis

$$H_1: \text{mean}_{\text{Mouse 430}} \neq \text{mean}_{\text{MG-U74v2}}$$

where

$$\text{mean}_{\text{MG-U74v2}} = \text{mean}(\text{tissue panel } \log(\text{Signals}) \text{ produced by the MG-U74v2 probe set})$$

and where

$$\text{mean}_{\text{Mouse 430}} = \text{mean}(\text{tissue panel } \log(\text{Signals}) \text{ produced by the Mouse 430 probe set})$$

then

$$\text{NLP}_{PrSet} = -\log(p_{PrSet})$$

and

NLP_{PrSet} is set to a negative value if the $\text{mean}_{\text{MG-U74v2}} > \text{mean}_{\text{Mouse 430}}$.

$\text{NLP}_{T_{\text{Mouse 430}}}$ and $\text{NLP}_{T_{\text{MG-U74v2}}}$

We run a One-Way ANOVA on the tissue panel data for each probe set design to obtain two p -values, $p_{T_{\text{Mouse 430}}}$ and $p_{T_{\text{MG-U74v2}}}$ for the null hypothesis

$$H_0: \text{means of } \log(\text{Signals}) \text{ are the same for all tissue types}$$

against the alternate hypothesis

$$H_1: \text{means of } \log(\text{Signals}) \text{ are different for all tissue types}$$

then

$$\text{NLP}_{T_{\text{Mouse 430}}} = -\log(p_{T_{\text{Mouse 430}}})$$

and

$$\text{NLP}_{T_{\text{MG-U74v2}}} = -\log(p_{T_{\text{MG-U74v2}}})$$

DETECTION OF DISCORDANT PROBE SET PAIR

A discordant probe set pair has the following properties:

- (1) Absolute NLP_{PrSet} value is greater than 15
- (2) The probe set producing the higher average Signal values has an NLP_T value greater than 11
- (3) The probe set producing the lower average Signal values has an NLP_T value that is less than eleven, where the average Signal is the average over 33 $\log(\text{Signals})$ produced by a probe set for the tissue panel (3 replicates X 11 tissues)

For example, the probe set pair in Figure 5B is considered to be discordant because:

- (1) $\text{NLP}_{PrSet} = 16$ is greater than 15
- (2) $\text{NLP}_{T_{\text{Mouse 430}}} = 16$ is greater than 11, and
- (3) $\text{NLP}_{T_{\text{MG-U74v2}}} = 0.36$ is less than eleven and the Mouse 430 probe set produces a higher average $\log(\text{Signal})$ value than the MG-U74v2 probe set

SUMMARY

The GeneChip® Mouse Expression Set 430 incorporates the same expertise that was utilized for the design and performance of the GeneChip Human Genome U133 Set.

- Genomic sequences were used to verify sequence selection, orientation, and the quality of sequence clustering.
- Clustering information from UniGene Build 107 was used with primary sequences and annotation information combined from a large variety of public databases to provide higher quality data.
- Signal values are higher for the majority of Mouse 430 probe sets relative to the corresponding MG-U74v2 probe sets.
- Mouse 430 probe set pairs exhibit greater responsiveness to diverse tissue types.
- Mouse 430 probe sets outperform MG-U74v2 probe sets when both magnitude of Signal and responsiveness to tissue diversity are evaluated.

The resulting two-array set design and performance makes it the premier array product for the analysis of the transcribed mouse genome in order to explore human biology and disease processes using this widely used model system.

GENECHIP® MURINE GENOME U74v2 SET DESIGN

In the GeneChip® Murine Genome U74v2 Set design, consensus sequences were built from subclusters of UniGene Build 74 (September 1999). We then selected one

or two sequences from each cluster as the sequences to tile on the chip. The sequences were selected on the basis of matching previous mouse designs, and for containing full length sequences. After

that, large subclusters were preferred over smaller subclusters. Probe sets of 16 probe pairs were selected against the 3' ends using a set of heuristic rules.

Appendix: Differences in the design characteristics of the MG-U74v2 and the Mouse 430 are discussed in detail in the following table.

	MG-U74v2	Mouse 430	Justification
Sequence Sources	UniGene	UniGene, RefSeq, GenBank, dbEST, Mouse Draft Assembly	Improved annotation, classification, and sequence quality
Sequence Curation	Filtered for repeats, vector	Repeats and vector screening, EST quality trimming	Avoid low-quality EST sequence regions, thereby improving consensus sequence quality
Sequence Subclustering	Pangea CAT tool	Similarity and orientation	Reduces chimeric clusters
Sequence Orientation	According to CDS annotation and EST read direction	Genomic sequence, poly-A prediction, CDS, and EST read direction	Improves orientation calls by using sequence-based methods in addition to annotations
Sequence Selection Region	600 base region from the 3' end of consensus sequences	600 base regions selected from the consensus with regions based on strong evidence for polyadenylation, a full-length 3' end, and consensus sequence ends	Comprehensive detection of true 3' transcript ends prevents selection of probe sets against aberrantly extended clusters and allows for detection of shorter form transcripts
Probe Quality	Heuristic rules and Neural Net model. Probe quality is assessed as a binary (yes/no) function	Thermodynamic multiple linear regression model predicts intensity of probes. Probe quality assessed on a continuous scale.	Improved selection of probes that hybridize well to the correct target and reduce non-specific cross hybridization
Probe Uniqueness	Probes unique if 20 or fewer bases match pruning sequences, with up to 5 base total gap.	Probes that have two 8-mer matches, including at least one 12-mer match will be avoided	Minimize specific cross hybridization to similar targets from unintended sequences
Probe Spacing	Not considered for probe selection	Spacing weighted to favor high-quality and independent probes	Ensure multiple probes give independent measurements of the target
Number of Probes	16	11	Combined with algorithm and probe quality improvements, allows greater information density without reduction in information quality
Probe Set Annotation	_s, _g, _f, _r, _i	_a, _s, _x Discontinued: _r, _i Transformed: _g → _s or _a, _f → _x	Probe set types were simplified and adjusted to account for improvements in probe selection rules
Feature Size	20 micron	18 micron	Allows greater information density without reduction in information quality

AFFYMETRIX, INC.

3380 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
Tel: +44 (0) 1628 552550
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

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

Part No. 701405 Rev. 1

©2003 Affymetrix, Inc. All rights reserved. Affymetrix®, GeneChip®, ®, ®, HuSNP®, Jaguar™, EASI™, MicroDB™, GenFlex®, CustomExpress™, CustomSeq™, NetAffx™, "Tools to take you as far as your vision™", and "The Way Ahead™" are trademarks owned or used by Affymetrix, Inc. 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; 5,744,305; 6,261,776; 6,291,183; 5,700,637; 5,945,334; 6,346,413; and 6,399,365; and EP 619 321; 373 203 and other U.S. or foreign patents.