AFFYMETRIX®

Technical Note

Array Design and Performance of the GeneChip® Rat Expression Set 230

The rat is a common model system for the study of a number of biological processes, in particular toxicology, neurobiology, and immunology. Gene expression profiling is an important tool in understanding the genetic interactions underlying these basic processes. This document is an overview of the approach and parameters used in the design of the GeneChip® Rat Expression Set 230.

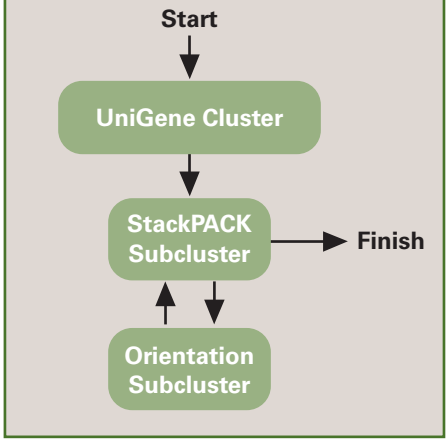
GeneChip® Rat Expression Set 230 Design

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 Rat Expression Set 230 (Rat 230) design. The design modifications for the Rat 230 are detailed below.

Various public data sources were used for the Rat 230 design (Table 1). Sequence data were obtained from dbEST (NCBI, June 2002), GenBank (NCBI, Release 129, April 2002), and RefSeq (NCBI, June 2002). Additionally, a preliminary draft assembly of the rat genome (Baylor College of Medicine Human Genome Sequencing Center, June 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 rat 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
- Identification and removal of low-quality regions of EST sequences

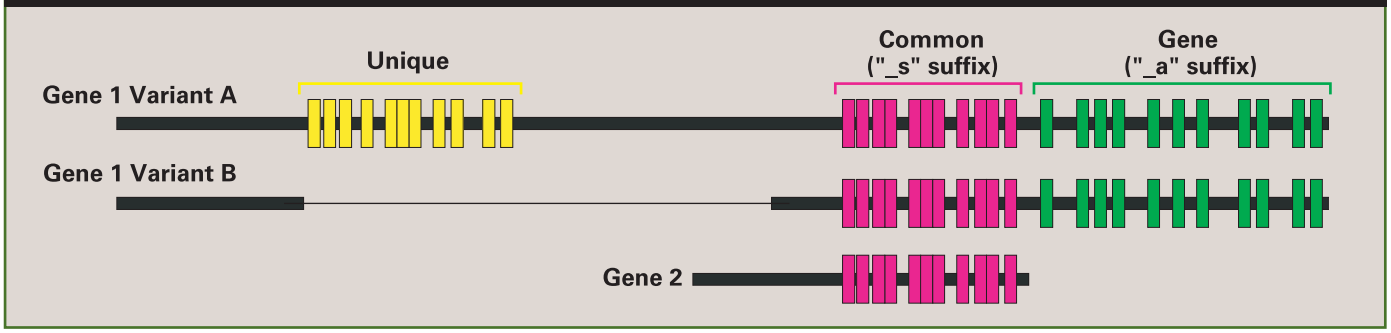
Figure 1: Sequence cluster information from UniGene was used to create initial seed clusters. Seed clusters were subclassified into one or more StackPACK subclusters with assemblies using StackPACK (Electric Genetics). Subclusters with orientation problems were further subclustered into orientation subclusters,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



UniGene (NCBI, Build 99) 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 subclusters 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)
- Six 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 on the upstream polyadenylation site is used.

In summary, sequence content for the Rat 230 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 six 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 the same UniGene cluster matched rules one and two. Other special classes of sequence content were included if they did not already meet the rules above:

- Best matching probe set for a GeneChip Rat Genome U34A probe set
- Best matching probe set to a high-quality mouse probe selection region

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 common probes among multiple transcripts from separate genes are annotated

Source	Release Date	Sequences	Used in Design
UniGene	June 2002 (#99)	59,904	28,757
dbEST	June 2002	347,955	243,272
GenBank	April 2002 (#129)	9,934	6,948
RefSeq	June 2002	3,785	3,783
Total		**421,578**	**282,760**

Table 1. Sources and numbers of sequences used in the Rat 230 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 rat genome from the Baylor College of Medicine Human Genome Sequencing Center (June 2002) improved cDNA sequence orientation and annotation.

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

Classification	Rat Set 230	Rat 230A	Rat 230B
Probe Sets	31,042	15,866	15,276
UniGene Clusters	28,757	14,280	14,477
Additional Potential Full Lengths	65	65	0
Subclusters	30,248	15,256	14,992
Full Lengths	**4,699**	**4,699**	**0**
Full Length End and Strong Evidence for Polyadenylation	1,497	1,497	0
Strong Evidence for Polyadenylation	307	307	0
Full Length End	2,811	2,811	0
Consensus End	84	84	0
Non-ESTs (excluding Full Lengths)	**700**	**700**	**0**
Strong Evidence for Polyadenylation	187	187	0
Consensus End	513	513	0
ESTs	**25,643**	**10,467**	**15,176**
Strong Evidence for Polyadenylation	6,759	6,759	0
Library Coverage >1			
Evidence for Polyadenylation >1			
Hit to Good Mouse Probe Selection Region	2,940	1,384	1,556
Does Not Hit a Good Mouse Probe Selection Region	5,701	1,393	4,308
Single Evidence for Polyadenylation			
Hit to Good Mouse Probe Selection Region	824	73	751
Does Not Hit a Good Mouse Probe Selection Region	2,113	114	1,999
No Direct Evidence for Polyadenylation			
Hit to Good Mouse Probe Selection Region	1,412	146	1,266
Does Not Hit a Good Mouse Probe Selection Region	2,467	139	2,328
Single Library Coverage >1			
Evidence for Polyadenylation >1			
Hit to Good Mouse Probe Selection Region	165	6	159
Does Not Hit a Good Mouse Probe Selection Region	1,152	25	1,127
Single Evidence for Polyadenylation			
Hit to Good Mouse Probe Selection Region	979	29	950
Does Not Hit a Good Mouse Probe Selection Region	387	252	135
No Direct Evidence for Polyadenylation			
Hit to Good Mouse Probe Selection Region	596	65	531
Does Not Hit a Good Mouse Probe Selection Region	148	82	66

Table 2. Classification and number of probe sets placed on the Rat 230. It is estimated that this set interrogates approximately 30,000 transcripts from approximately 29,000 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 the number of distinct cDNA libraries from which the subcluster sequences were originally sequenced and the probe set hits to high-quality mouse probe selection regions. The mouse probe selection regions were from an internal design based on mouse UniGene Build 107.

GeneChip® Rat Expression Set 230 Performance Improvements

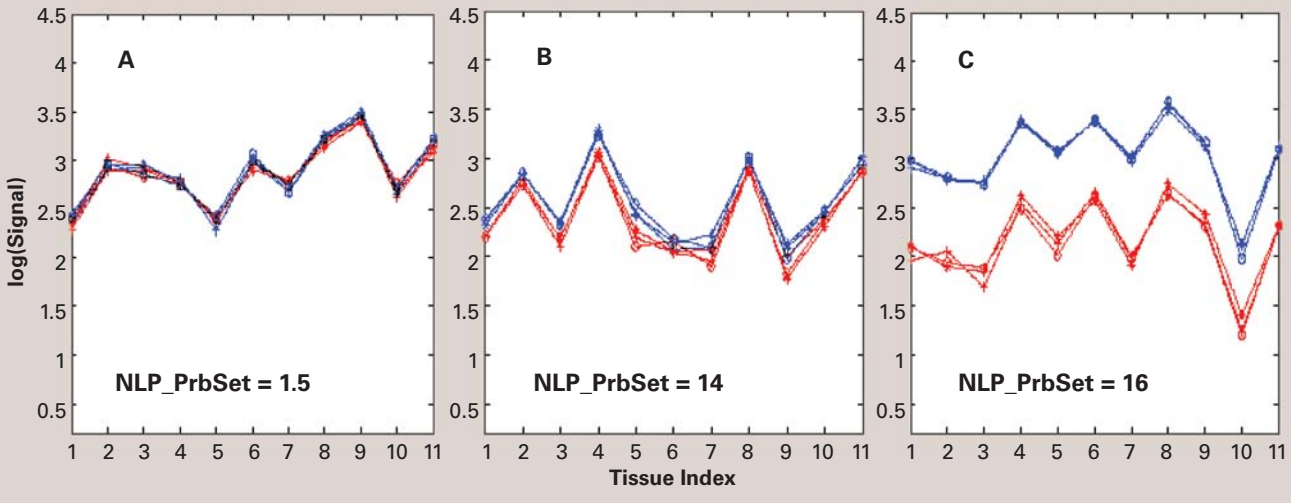
COMPARISONS OF TISSUE PANEL SIGNALS FOR PROBE SET PAIRS

In order to show that the new Rat 230 design, with reduced probe set size, produced equivalent or more informative data compared to the previous design, GeneChip® Rat Genome U34 Set (RG-U34), 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 RG-U34 and the Rat 230 sets. For a detailed comparison of the Rat 230 and the RG-U34 designs see the Appendix. More specifically, probe sets were paired by requiring that all sixteen probes of each RG-U34 probe set align within the PSR of its matched Rat 230 probe set. All possible pairings were made between probe sets from the Rat 230 Array Set (A and B) and the RG-U34 Array Set (A, B, and C). In some cases, multiple probe sets from one or both array types represented the same Rat 230 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 design (representing a highly expressed region) to a probe set from the other array design (representing a rarely expressed region), we select only the probe set from each array design that is *most responsive* to tissue diversity to represent the PSR (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 11,879 PSRs represented by probe sets from both the Rat 230 Set and the RG-U34 Set (see Methods).

We ran the samples in triplicate across eleven tissue types on both array designs (11 tissues X 3 replicates X 2 array designs). The tissue types were brain, embryo, heart, kidney, liver, lung, skeletal muscle, ovary, spleen, testicle, and thymus. Single array analysis was performed on each experiment and Signal values were used to generate histograms. These plots allowed a visual comparison of the Signal values produced between the paired Rat 230 and RG-U34 probe sets. In Figure 3, each point is a log(Signal) value (y axis) for one of the eleven tissue types (x axis) produced by a RG-U34 probe set (red) ora Rat 230 probe set (blue).

Figure 3. Log(Signal) profiles for three probe set pairs. The x axis represents the tissue types, while the y axis represents the log(Signal). Each point is the log(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 hybridizations for each tissue type and array type produce three profiles for the Rat 230 Set probe set (blue) and three profiles for the paired RG-U34 Set probe set (red). Definitions of NLP_PrbSet, NLP_$T_{Rat\ 230}$, and NLP_$T_{RG\text{-}U34}$ are given in the text and in Methods. **A.** Probe Set Pair with NLP_PrbSet=1.5; Rat 230 probe set is 1372564_at (NLP_$T_{Rat\ 230}$=16); RG-U34 probe set is rc_AA851618_at (NLP_$T_{RG\text{-}U34}$=16). **B.** Probe Set Pair with NLP_PrbSet=14; Rat 230 probe set is 1370282_at (NLP_$T_{Rat\ 230}$=16); RG-U34 probe set is U44948_at (NLP_$T_{RG\text{-}U34}$=16). **C.** Probe Set Pair with NLP_PrbSet=16; Rat 230 probe set is 1371774_at (NLP_$T_{Rat\ 230}$=16); RG-U34 probe set is rc_AI236332_at (NLP_$T_{RG\text{-}U34}$=14.8).



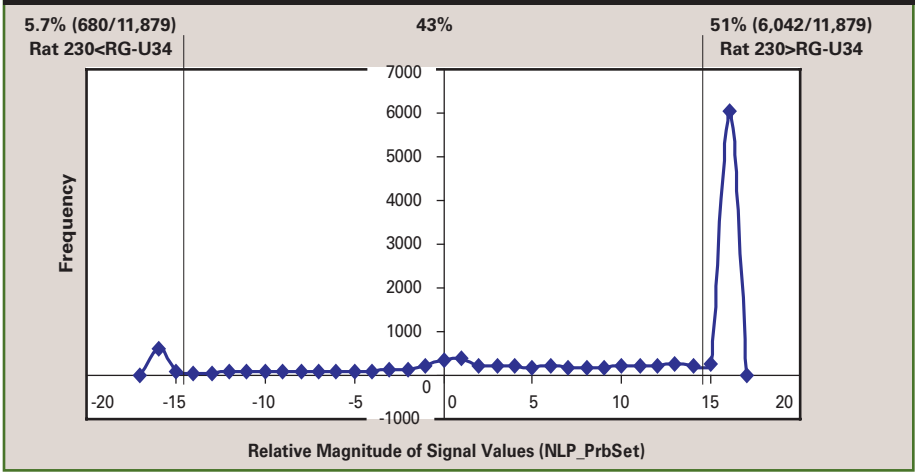
EFFECT OF PROBE SET DESIGN ON SIGNAL LEVELS

This analysis indicates that the majority of Rat 230 probe sets tend to produce higher Signal values relative to corresponding probe sets on the RG-U34 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(Signal) values for a probe set pair (see Methods). NLP_PrbSet values increase as the mean of log(Signal) values produced by one probe set increasingly differs from the mean of log(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 affects the magnitude of the Signals. For clarity in viewing the data we set the NLP_PrbSet to a negative value if the RG-U34 probe set Signals are greater than the Rat 230 probe set Signals on average for that probe set pair.

Figure 3, panels A-C show how the absolute NLP_PrbSet values increase as the RG-U34 and Rat 230 log(Signal) profiles resolve in three different probe set pairs (note that the blue curves and red curves become separated). A probe set pair whose NLP_PrbSet 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_PrbSet 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_PrbSet values for the 11,879 probe set pairs. The bars separate the cases where a probe set from one array design produces significantly higher Signal values (absolute NLP_PrbSet values are greater than fifteen) than a probe set from the other array design. Given this separation, 51% (6,042/11,879) of the probe set pairs have significantly higher Rat 230 probe set Signals. Only 5.7% (680/11,879) of the probe set pairs have significantly higher RG-U34 probe set Signals. When Signal values are evaluated along with responsiveness to tissue diversity, however, only a small fraction, 59 of the 680 RG-U34 probe set pairs, were found to be potentially discordant and more informative for expression profiling out of the total number of probe set pairs evaluated. This concept will be discussed further in the *Discordant Probe Set Pairs* section.

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



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 Rat 230 design producing more responsive probe sets. The values of $NLP\_T_{Rat\ 230}$ and $NLP\_T_{RG\text{-}U34}$ 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(Signal) values for the given probe set (see Methods). NLP_T values increase as the mean of log(Signal) values produced by each tissue increasingly differs from the means of log(Signal) values produced by the other tissues. In other words, the higher the value of NLP_T, the more variable the log(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(Signal) values of the upper curves vary significantly with regard to at least one tissue, producing NLP_T values of 16 (Figure 5A) and 11.2 (Figure 5B). However, the log(Signal) values for lower curves are essentially flat with regard to the variation across the different tissues, producing low NLP_T values of 1.4 (Figure 5A) and 0.52 (Figure 5B).

We analyzed the distribution of relative responsiveness to the tissue panel 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: $(NLP\_T_{Rat\ 230})$ - $(NLP\_T_{RG\text{-}U34})$, and generated the distribution of relative responsiveness values for the 11,879 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 11,879 probe set pairs (blue curve) is centered about zero, with a slight bias towards Rat 230 probe sets being more responsive. 19% of the probe set pairs have equally responsive RG-U34 and Rat 230 probe sets, 47% have more responsive Rat 230 probe sets, and 34% have more responsive RG-U34 probe sets. Random experimental variation is expected to produce non-zero values in both directions. The fact that percentages are not equal (47% vs. 34%) suggests that differences between the Rat 230 design and RG-U34 design may also contribute.

Figure 5: Log(Signal) profiles for two discordant probe set pairs. Each point is the log(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 hybridizations for each tissue type and array design produce three profiles for the Rat 230 probe set (*blue*) and three profiles for the paired RG-U34 probe set (*red*). *NLP_PrbSet*, *NLP_T$_{Rat\ 230}$*, and *NLP_T$_{RG\text{-}U34}$* are described in the text and in the Methods. **A.** Probe Set Pair with NLP_PrbSet= -16. The Rat 230 probe set is 1388247_at (NLP_T$_{Rat\ 230}$=1.4). The RG-U34 probe set is U48828_g_at (NLP_T$_{RG\text{-}U34}$=16). **B.** Probe Set Pair with NLP_PrbSet= 16. Rat 230 probe set is 1392826_at (NLP_T$_{Rat\ 230}$=11.2). The RG-U34 probe set is rc_AI014100_s_at (NLP_T$_{RG\text{-}U34}$= 0.52).

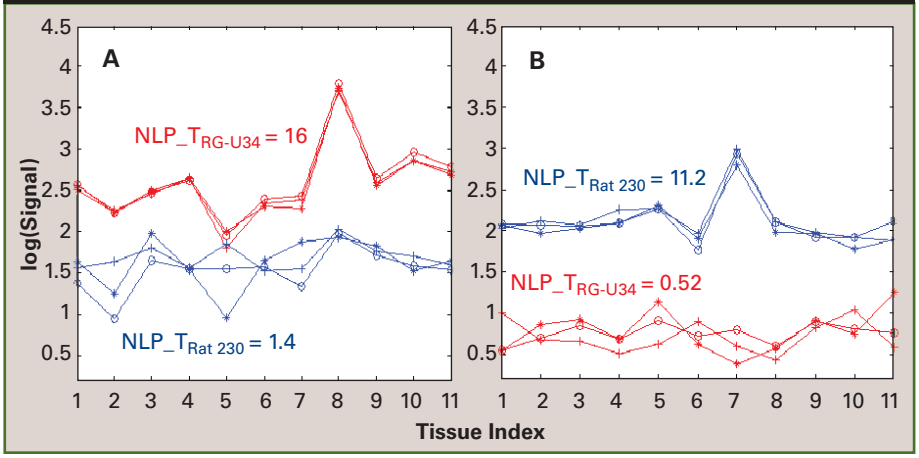
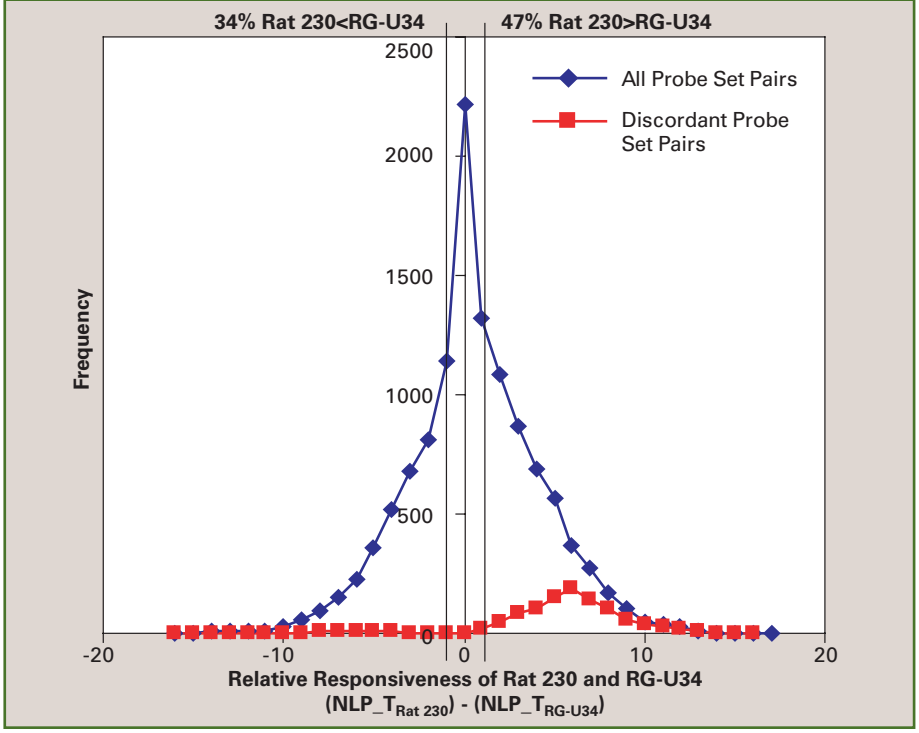


Figure 6: Relative Responsiveness: difference (NLP_T$_{Rat\ 230}$-NLP_T$_{RG\text{-}U34}$) between Rat 230 and RG-U34 response to tissue diversity. Distributions are generated for all probe set pairs evaluated (11,879 blue curve) and for 1,057 discordant (defined in the text and in the Methods) probe set pairs (red curve). The bars bracket the cases for which Relative Responsiveness is zero (i.e., the response for both designs is equivalent).



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 discordant cases as those where a probe set from one array design not only produces significantly higher log(Signal) values (NLP_PrbSet >15) but is considered to be responsive to tissue diversity (NLP_T > 11), while the probe set from the second array design not only produces significantly lower log(Signal) values but also produces a response to tissue diversity that falls below the threshold used in this study (NLP_T <11). The probe set producing 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(Signal) levels, because both probe sets are responsive to tissue diversity and, therefore, should be informative for expression profiling.

Only 8.9% (1,057/11,879) of probe set pairs fall into the discordant category. For the probe set pairs within this category, the Rat 230 design is 17 times (998/1,057 vs. 59/1,057) 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 1,057 discordant cases is shown in Figure 6 (red curve). Since the Rat 230 design is more likely to produce a responsive probe set among the discordant cases, the bulk of the discordant cases (998/1,057) 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 Rat 230 demonstrated superior performance compared to the RG-U34.

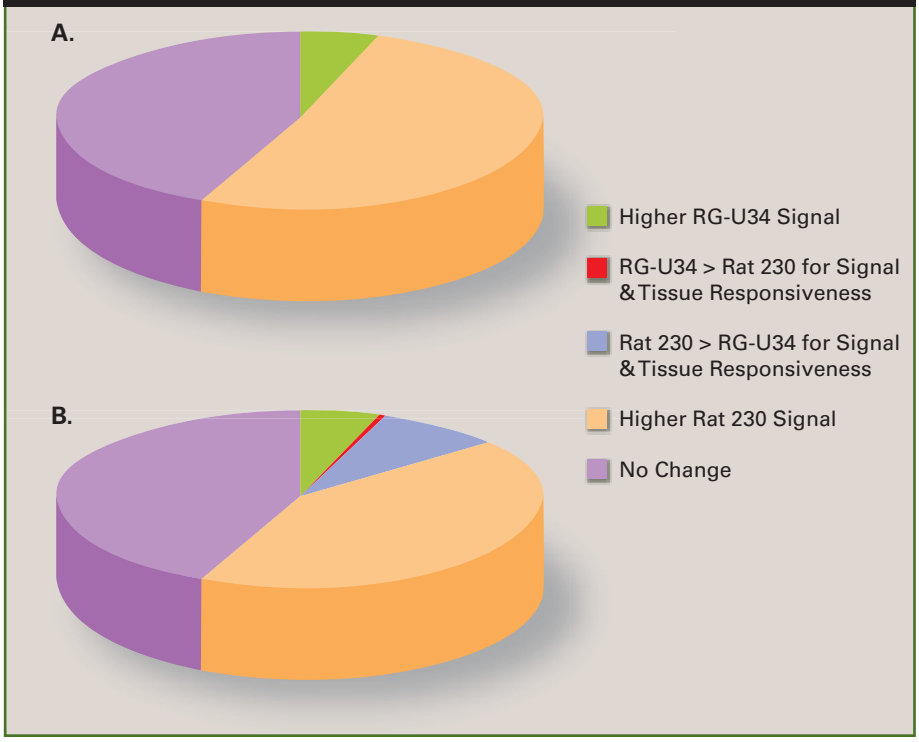
SUMMARY OF GENECHIP® RAT EXPRESSION SET 230 PERFORMANCE IMPROVEMENTS

Signal values tend to be higher for the majority of Rat 230 probe sets relative to the corresponding probe sets on RG-U34. 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 Rat 230 probe sets, on average, exhibit greater responsiveness to diverse tissue types. In addition, Rat 230 probe sets are more likely to outperform a RG-U34 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 two array types. The percentage of cases where the Rat 230 Signals are higher (51%) or, in contrast, the RG-U34 Signals are higher (5.7%) is small. 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 design compared to the other. In these cases, Rat 230 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 RG-U34 design.

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 Rat 230 probe sets (orange) compared to the percentage of probe sets with higher Signal values in RG-U34 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 Rat 230 probe set pairs are 17 times more likely to show characteristics of superior probe sets (blue) and only rarely produce an inferior one (red).



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_$T_{Rat\ 230}$ and NLP_$T_{RG\text{-}U34}$.

NLP_PrbSet

We produce a *p*-value, p_{PrbSet}, for the probability that the array design has no effect on the magnitudes of the 66 (11 tissues X 3 replicates X 2 array designs) log(Signal) values for a probe set pair. Specifically, p_{PrbSet} 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: $mean_{Rat\ 230} = mean_{RG\text{-}U34}$

against the alternate hypothesis

H_1: $mean_{Rat\ 230} \neq mean_{RG\text{-}U34}$

where

$mean_{RG\text{-}U34}$=mean(tissue panel log(Signals) produced by the RG-U34 probe set)

and where

$mean_{Rat\ 230}$=mean(tissue panel log(Signals) produced by the Rat 230 probe set)

then

$NLP\_PrbSet = -\log(p_{PrbSet})$

and

NLP_PrbSet is set to a negative value if the $mean_{RG\text{-}U34} > mean_{Rat\ 230}$.

$NLP\_T_{Rat\ 230}$ and $NLP\_T_{RG\text{-}U34}$

We run a One-Way ANOVA on the tissue panel data for each probe set design to obtain two *p*-values, $p\_T_{Rat\ 230}$ and $p\_T_{RG\text{-}U34}$ for the null hypothesis

H_0: means of log(Signals) are the same for all tissue types

against the alternate hypothesis

H_1: means of log(Signals) are different for all tissue types

then

$NLP\_T_{Rat\ 230} = -\log(p\_T_{Rat\ 230})$

and

$NLP\_T_{RG\text{-}U34} = -\log(p\_T_{RG\text{-}U34})$.

DETECTION OF DISCORDANT PROBE SET PAIR

A discordant probe set pair has the following properties:

(1) Absolute NLP_PrbSet value is greater than 15
(2) The probe set producing the higher average Signal values has an NLP_T value (as described above) 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(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) NLP_PrbSet = 16 is greater than 15
(2) $NLP\_T_{Rat\ 230}$ = 11.2 is greater than 11, and
(3) $NLP\_T_{RG\text{-}U34}$ = 0.52 is less than eleven and the Rat 230 probe set produces a higher average log(Signal) value than the RG-U34 probe set

SUMMARY

The GeneChip® Rat Expression Set 230 design incorporates the same expertise that was used 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 99 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 Rat 230 probe sets relative to the corresponding RG-U34 probe sets.
- Rat 230 probe set pairs exhibit greater responsiveness to diverse tissue types.
- Rat 230 probe sets outperform RG-U34 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 rat genome.

GENECHIP® RAT GENOME U34 SET DESIGN

The first-generation GeneChip® rat genome set, the GeneChip Rat Genome U34 Set (RG-U34), was designed using exemplar sequences selected from UniGene Build 34 and GenBank. The longest member or exemplar in a cluster of sequences was used as the representative for a sequence. Orientation of exemplar sequences was based on coding sequence (CDS) and EST read direction annotations. Probe sets consisting of 16 probe pairs were selected against the 3' ends using a set of heuristic rules.

Appendix: Differences in the design characteristics of the RG-U34 and the Rat 230 are discussed in detail in the following table.

	RG-U34	Rat 230	Justification
Sequence Sources	UniGene, GenBank	UniGene, RefSeq, GenBank, dbEST, Rat 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	Selected UniGene and GenBank exemplar sequences	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 exemplar end	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	24 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. 701406 Rev. 1