Array Design for the GeneChip® Human Genome U133 Set

Human Array Designs

Microarray gene expression analysis is an integrated system comprised of sequence selection, probe selection and array design, as well as the algorithms used to analyze the data[I]. The new Human Genome U133 Set (HG-U133) represents a culmination of Affymetrix array design experience. This document is an overview of the parameters used in the design of the HG-U133 set. A brief review of previous array designs is presented for comparison. The differences between designs are summarized in the Appendix.

Sequence Selection for the HuGeneFL Array

In the HuGeneFL design an exemplar method was used to select sequences. An exemplar is a single sequence that represents a cluster of sequences. This method consisted of selecting GenBank exemplar sequences from preliminary Unigene clusters, and tiling probes to these sequences. This method is simple, but limited in the extent of the transcribed genome represented as well as the total number and quality of sequences tiled.

Sequence Selection for the Human Genome U95 Set

The next step in array design evolution was the more complex consensus calling method used in the design of the Human Genome U95 Set (HG-U95). In this method, a consensus sequence was created from each sequence cluster. The consensus sequence was then used for probe selection.

Briefly, Unigene build 95 data were pruned, primarily by removing expressed sequence tags (ESTs) from excessively large clusters with an identified protein-coding sequence (CDS). The remaining sequences were aligned and subclustered by a Cluster and Alignment Tool (CAT)[II]. For a given base to be called within the consensus sequence, that position had to agree for at least 75% of the aligned sequences. Otherwise, the base was identified with an "N" and not used for probe selection. A weighted majority algorithm was used to identify a direction for each sequence. One or two consensus sequences per cluster were tiled to create the array.

Sequence Selection for the Human Genome U133 Set

A major advance in the HG-U133 design is the use of genomic sequence to verify sequence selection, sequence orientation, and the quality of sequence clustering. The genomic information also anchors EST sequences with a concurrent increase in annotation and orientation information. Additional enhancements result from an emphasis on combining primary sequence and annotation information from a large variety of public databases (Table 1), with initial clustering information from the UniGene Human Database, Hs.data, release 133 (U133). The strategies used for sequence alignment and sequence

Source	Release Date	Human Sequences	Used In Design
UniGene	April 20, 2001 (#133)	2,688,626	7,907
dbEST	April 28, 2001	3,471,886	2,619,747
WUSTL	Feb 2001	1,430,516	1,195,490
GenBank	April 25, 2001 (#123.0)	61,523	38,168
RefSeq	April 30, 2001	12,716	12,461
Total		**7,665,267**	**3,873,773**

Table 1. Sources and numbers of sequences used in the HG-U133 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. Raw base call information, which enables better polyadenylation identification, was obtained for a substantial number of EST sequences from Washington University (WUSTL). The draft assembly of the human genome from the University of California, Santa Cruz (Golden Path) was used to improve cDNA sequence annotations.

annotation have been significantly expanded (Figure 1). Over six million sequences were considered for inclusion in this design.

Sequence Collection and Analysis

The success of an array design is highly dependent on the quality of sequence information used. To provide the most complete starting information, cDNA sequence data were obtained from primary sequence sources: GenBank, RefSeq, dbEST and Washington University (WUSTL)[III] (Table 1). Sequence meta information such as descriptions and definitions, clone identifiers, library identifiers, read directions, CDS annotations, low quality base annotations, gene names, and gene products were extracted from the external data files in addition to the actual sequence.

All input sequences were aligned to the draft assembly of the human genome (April 2001 release). Only high quality regions of genome alignment were used to annotate and analyze the input sequences. The genomic alignments confirmed and implied sequence orientation through the identification of consensus splice sites.

In an effort to improve consensus sequence quality, low-quality EST sequence regions were identified and removed according to the following mutually exclusive rules:

1) Sequences were trimmed if the primary sequence annotation indicated poor quality regions.

2) If EST sequences aligned to the genomic sequence, the unaligned bases were removed.

3) The 3' ends were trimmed in cases where the sequence read was abnormally long.

Sequence Collection

- Sequences and annotations collected from a variety of sources.

Sequence Analysis

- Aligned to the draft assembly of the genome.
- Low quality bases trimmed from ESTs.
- Polyadenylation sites detected.

Seed Cluster Creation

- UniGene Clusters were used as seed clusters.
- Additional seed clusters were created with potential full-length mRNAs not in UniGene.

Genomic Based Subclustering

- Seed clusters were subclustered based on the contig.

Sequence Based Subclustering

- Sequences were subclustered in transcriptome space.

Orientation Based Subclustering

- EST read directions, CDS annotations, consensus splice sites, and polyadenylation sites were used to predict the orientation of the subcluster. The subcluster was resubclustered if substantial conflicts exist.

Picking Probe Selection Regions

- For each subcluster, regions for probe selection were selected from either the consensus or exemplar sequences.
- Multiple regions may be choosen due to alternative polyadenylation sites.

Prioritizing Probe Selection Regions

- Regions for probe selection were prioritized based on the quality of sequence and annotation.

Figure 1. Sequence selection steps. A number of analyses were involved in generating the HG-U133 design. See the corresponding text for more details.

These approaches reduced the presence of low quality bases, which disrupt the clustering process and potentially contaminate the sequence content on the array.

Polyadenylation Sites

The portion of mRNA sequence adjacent to a poly-A site is most efficiently converted into labeled target (see the Expression Analysis Technical Manual[IV] for details of the labeling reaction). Great care was therefore taken to identify polyadenylation sites since optimal probes are generally located within 600 bp upstream of the site. The use of untrimmed, primary sequence information helped significantly in this regard because poly-A or poly-T tracts are often removed prior to submission to public databases. Polyadenylation sites were identified and a site score was calculated using a heuristic that accounts for the length of the poly-A (or poly-T, 5' read), the amount of 5' (or 3') extraneous sequence, and the degree of interruption within the poly-A tract (or poly-T). For those sequences with a polyadenylation site, the presence of a polyadenylation signal was determined using a probabilistic model.

Vector Contamination and Repeats

Each sequence was assessed for repeats using RepeatMasker software and for vector contamination using BLASTN and the UniVector database.[V]

Cluster Creation

The initial cluster information was derived from UniGene build U133. Additional potential full-length sequences not in UniGene were used to create an additional 1,144 singleton clusters.

Genome Based Subclustering

In a number of cases a UniGene cluster represents several genes within a gene family. Genome based subclustering was applied using the alignment information for each member sequence to the genomic sequence. Sequences that aligned to different contigs were assigned to separate subclusters. Those sequences that did not align to the genomic sequence were added to the largest subcluster.

Sequence Based Subclustering

At this time, the human genome assembly remains incomplete and the quality is highly variable. It was therefore still necessary to refine seed clusters using a transcriptome-based clustering approach. This was accomplished using the CAT^{II}. To be conservative in selecting probes, 75% identity in all of the member sequences is required when a consensus is called. This eliminates problems with ambiguous and polymorphic bases.

Orientation Based Subclustering

Subcluster orientation is determined using information from the following:

1) Sequence-label information is used, such as CDS annotations and read directions.

2) In cases where introns are clearly delineated, consensus splice-site flanking sequences are used where an intron flanking sequence GT-AG indicates the sense orientation while CT-AC implies an anti-sense orientation.

3) Polyadenylation signals and sites (5’ stretches of T's or 3’ stretches of A’s) also provide orientation information.

A combination of the above information is used to make an orientation call of sense, anti-sense, or unknown for each member sequence used. Clusters with a ambiguous orientation were re-subclustered by placing all the sequence members with evidence of a sense orientation into one subcluster and all the members with evidence of an anti-sense orientation into another subcluster. Sequences with an unknown orientation were placed into the larger of these.

Probe Selection Regions

A given subcluster, while typically representing one transcript variant, may represent several alternative polyadenylation sites that may be sufficiently spaced to warrant more than one probe selection region. Based on the orientation call, the 3’ end of the cluster was identified. For clusters of unknown or ambiguous orientation, probes were picked against both ends of the sequence. Potential transcript ends are identified by the 3’ end of a potential full length member sequence, by a set of 8 or more EST ends (5’ end of a 3’ EST or a polyadenylated EST), or by the end of the consensus sequence (Figure 2). A 600 base region upstream of the end is chosen for probe selection. For putative transcript ends based on a potential full-length mRNA, the corresponding mRNA sequence is used as an exemplar when picking probes. For all other transcript ends, the consensus sequence is used. A consequence of this strategy is that there can be multiple probe sets representing a particular sequence (Figure 3).

Prioritization

The number of potential regions to represent the human transcriptome and to pick probes from is in the hundreds of thousands. A number of these regions or clusters are speculative, consisting of EST singletons. Some are aberrant minority subclusters resulting from cloning or sequencing errors. At the other extreme are transcripts that have been documented in the

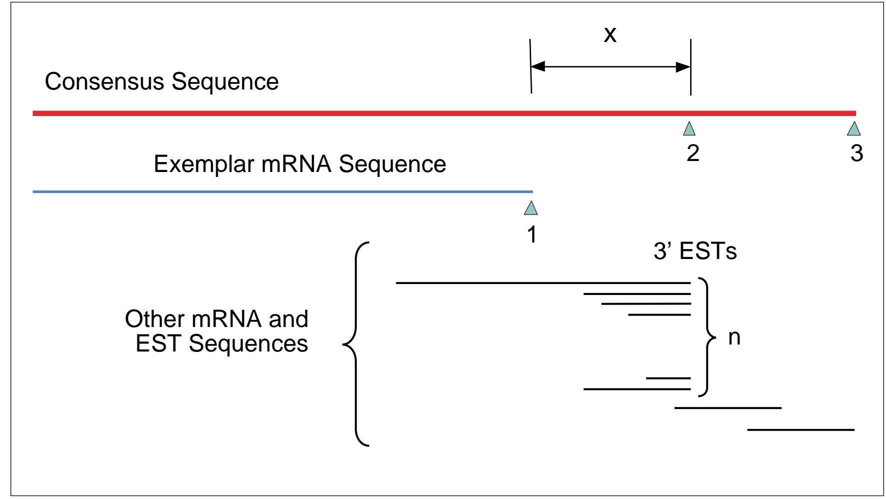


Figure 2. Multiple probe selection regions. If the full-length exemplar has 3′ UTR, then probes are picked from the 600 bp region ending at 1. For the stack of 3′ ESTs (annotated 3′, or possessing polyadenylation site/signal), if $n \geq 8$, and $x > 400$, then probes are also picked from region 2. Otherwise, if $n < 8$ or if $x \leq 400$, then region 2 is skipped. If the exemplar has no 3′ UTR sequence then region 1 is not picked. Consequently, if $n \geq 8$, region 2 is picked, otherwise region 3 is picked.

databases over 1,000 times. Thus the set of potential regions for probes must be prioritized so that well-annotated and strongly supported regions are given the highest likelihood of being represented on the arrays in the set. In general, the highest priority was given to regions representing mRNAs annotated as containing the complete coding region and some 3' untranslated sequence. EST-only clusters are prioritized according to strong evidence for polyadenylation, the size of the cluster, maximal depth of the cluster assembly, genomic mapping and orientation information. The actual prioritization categories are summarized in the bottom section of Table 2, below.

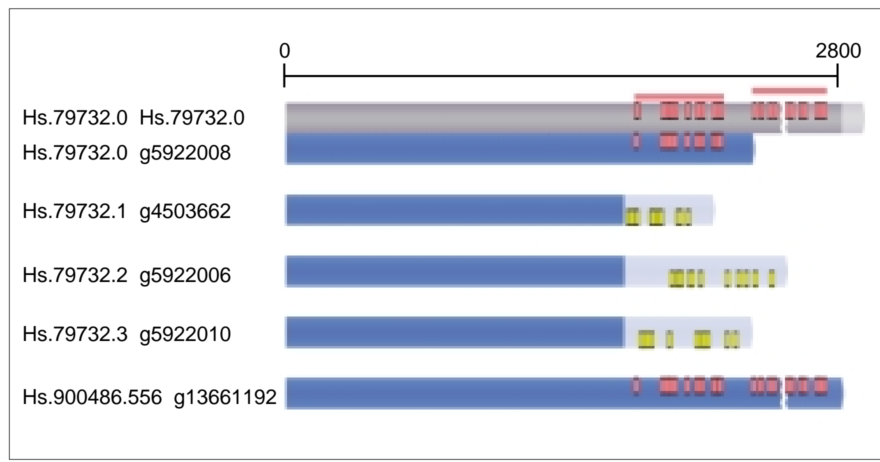


Figure 3. Multiple sequences and probe sets. The figure represents multiple pair-wise alignments to the seed sequence, Hs79732.0. The first portion of each label indicates the subcluster and the second portion indicates the sequence. Consensus sequences start with 'Hs' while exemplar sequences start with 'g'. Red boxes on the transcript bar indicate the regions that are probed. The two red bars at the top indicate the span of each probe set that detects the seed sequence. Dark blue regions are strongly conserved while the light blue regions are divergent, unaligned sequence. Wavy line gaps indicate either unaligned sequence due to low complexity filtering or divergent sequence.

For UniGene cluster Hs.79732 representing the fibulin 1 gene, there are four subclusters represented by four exemplars (blue bars) and again by a single consensus sequence (grey bar). There is also a potential full-length sequence shown in this alignment (Hs.900486.556) that is not listed in UniGene, but is also a transcript for fibulin 1. Two probe sets represent possible alternative polyadenylation sites (red probe sets) while another three probe sets represent possible alternative 3′ transcript ends (yellow probe sets).

Classification	HG-U133A	HG-U133B	Total
UniGene Clusters	**14,593**	**19,318**	**31,728**
Additional Potential Full Lengths	**513**	**198**	**707**
Subclusters	**18,462**	**21,070**	**38,903**
Full Length Including UTR	**13,049**	**1,556**	**14,529**
Extended Full Length	**171**	**58**	**229**
Strongest evidence for polyadenylation	**3,228**	**6,929**	**10,140**
Complete CDS Consensus End	**570**	**74**	**643**
Non-EST Consensus End	**2,526**	**2,755**	**5,278**
Evidence for polyadenylation	**993**	**595**	**1,586**
EST-only clusters			
Oriented, Mapped, and 3′	**279**	**9,153**	**9,432**
Oriented and 3′	**33**	**590**	**623**
Mapped and 3′	**14**	**76**	**90**
3′	**22**	**0**	**22**
Opposite Consensus End	**683**	**619**	**1,301**
Distant Consensus End	**176**	**150**	**326**

Table 2. Classifications and counts of sequences placed on the HG-U133 Set. It is estimated that the HG-U133 Set interrogates approximately 39,000 transcripts.

Probe set annotation classes: The first tier indicates the number of UniGene clusters, additional potential full length sequences, and subclusters, while the second tier indicates the quality of the probe sets representing these clusters with regard to source sequence and annotations.
The sequences are assigned to unique classifications 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 probe sets are grouped by cluster annotation quality and include subclusters with 3 or more sequences or containing a maximum cluster assembly depth of at least 6 sequences. Cluster annotations include the orientation of the cluster (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 Selection

The probe selection method used for the HG-U133 reflects an advanced understanding of probe uniqueness and hybridization characteristics allowing selection of probes based on predicted behavior. The new array design uses a multiple linear regression (MLR) model that was derived from a thermodynamic model of nucleic acid duplex formation. This model predicts probe binding affinity and linearity of signal changes in response to varying target concentrations. In contrast, previous probe selection strategies were based on a set of heuristic rules[VI]. The heuristic rules were not predictive, but rather acted as filters to remove sequence features known to degrade probe performance. The heuristic rules did not meet the performance requirements for increased array density.

Model-Based Probe Selection Used for the Human Genome U133 Design

An advantage of the new model-based probe selection system is that it provides a physical and mathematical foundation for systematic and large-scale probe selection. It utilizes both sequence and empirical information to predict optimal probes for array-based gene expression analysis. A second advantage is that the system allows simultaneous optimization of probe selection for a number of parameters, such as linear response to target concentration, independence of probes within a set, and probe sequence uniqueness.

Probe Quality Metric

The probe quality metric was developed on training sets of Latin Square experiments as described in Affymetrix® technical note "New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays"[I]. In these experiments, labeled transcripts were spiked into a complex hybridization mixture at a series of concentrations. The mixture was hybridized to specially designed microarrays that represent 104 yeast and 90 human transcripts. These arrays contained all possible 25-mer probes from the complete expressed sequence regions of the genes. Multiple linear regression analysis was used to model the behavior of each probe. The predicted natural logarithms of hybridization intensity (LnI) are highly correlated with observed LnI (Figure 4A). The high correlation was demonstrated for concentrations from 0.25 to 1024 pM.

An essential criterion of probe selection for quantitative expression analysis is that hybridization intensities of the selected probes should be linearly related to target

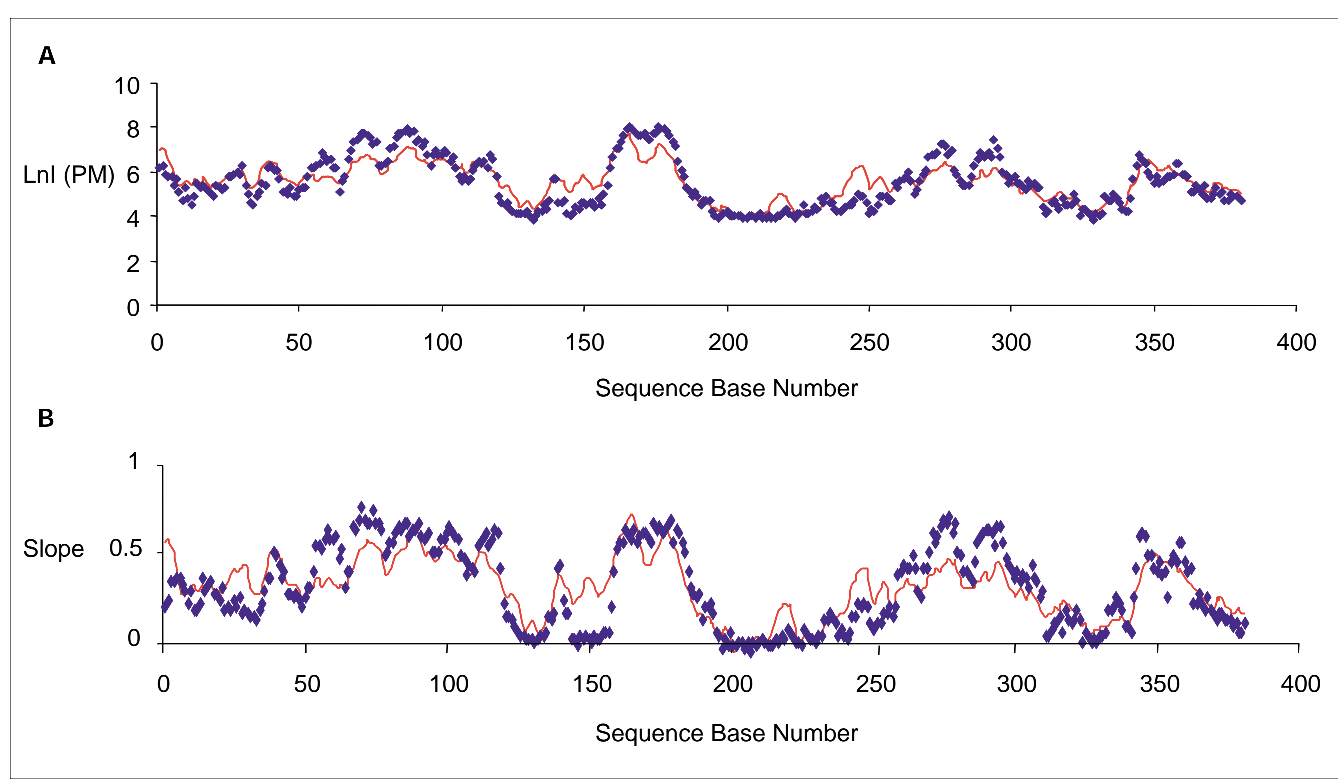


Figure 4. A. Predicted and observed probe intensities. Profiles are for predicted (red) and observed (blue) intensity values for all 25-mer perfect match probes of an expressed sequence region of a human gene, given an 8 pM target concentration. **B.** Predicted and observed slope of intensity vs. concentration. Profiles are for predicted (red) and observed (blue) slope values for all possible probes of a represented human gene.

concentrations. The quality metric is defined as the slope of the line that relates natural logarithms of intensities and of target concentrations for each probe (Figure 4B and Figure 5). The predicted slopes are highly correlated with observed slopes (Figure 4B).

Probes with low binding affinities typically correspond to small slope values because the hybridization affinity is too low for effective hybridization (Figure 5, brown and green lines). As affinity increases, the slope increases and probes are considered to be good quality (Figure 5, pink and red lines). However, when the affinity becomes extremely high (Figure 5, blue line) the slope drops to a small value. Those probes are usually GC-rich, and tend to cross-hybridize with nonspecific targets, and hence no longer exhibit a concentration response to their specific target.

Identification of Cross-hybridizing Probes

Microarrays with 794 different probes and 333 specific mismatches to each of those probes were used in a Latin Square experiment to develop a cross-hybridization model. Transcripts were spiked into a complex hybridization mixture at a series of spike concentrations. Multiple cross-hybridization rules were examined to determine which one best differentiated between probes which show significant cross-hybridization signal, and those which do not.
The rule which was chosen is that a probe is not expected to show cross-hybridization if it does not have at least two 8-mer perfect matches, including at least 12 consecutive matching bases to anything else in the transcriptome. These results were then validated on a different array which had 2694 probes, with 100 random mismatch probes to each to each perfect match probe.

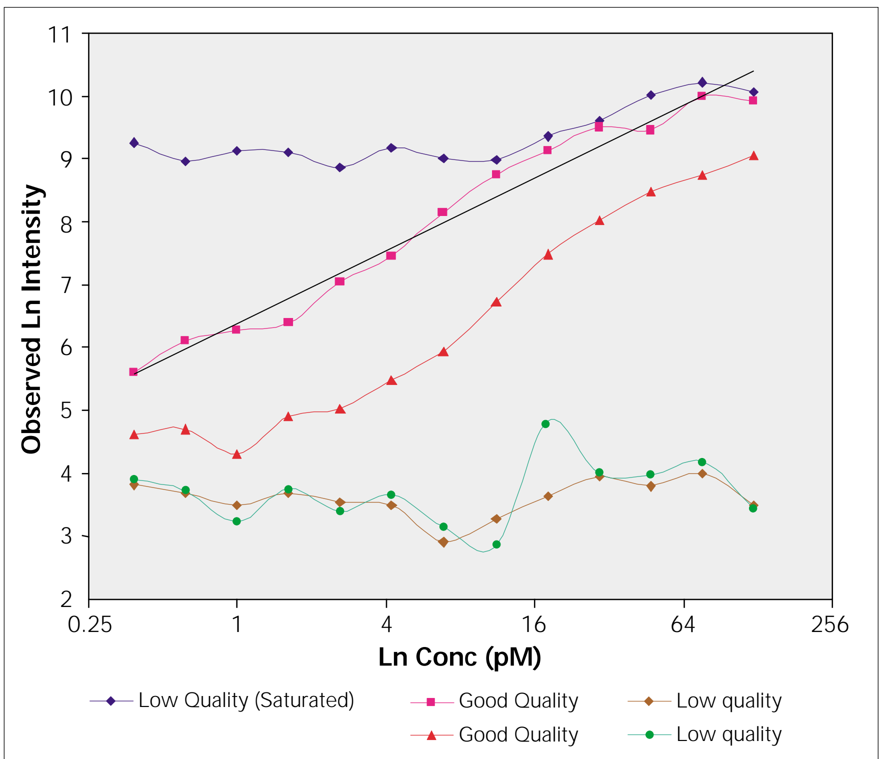


Figure 5. Observed intensity vs. concentration for five representative probes. The black line is the best fit line relating LnI and Ln Conc. Brown and green represent low affinity probes that are discarded. The blue represents an extremely high affinity, saturated probe that will also be discarded. Pink and red represent good quality probes with a linear response to changes in target concentration. These are typical of the probes that are selected.

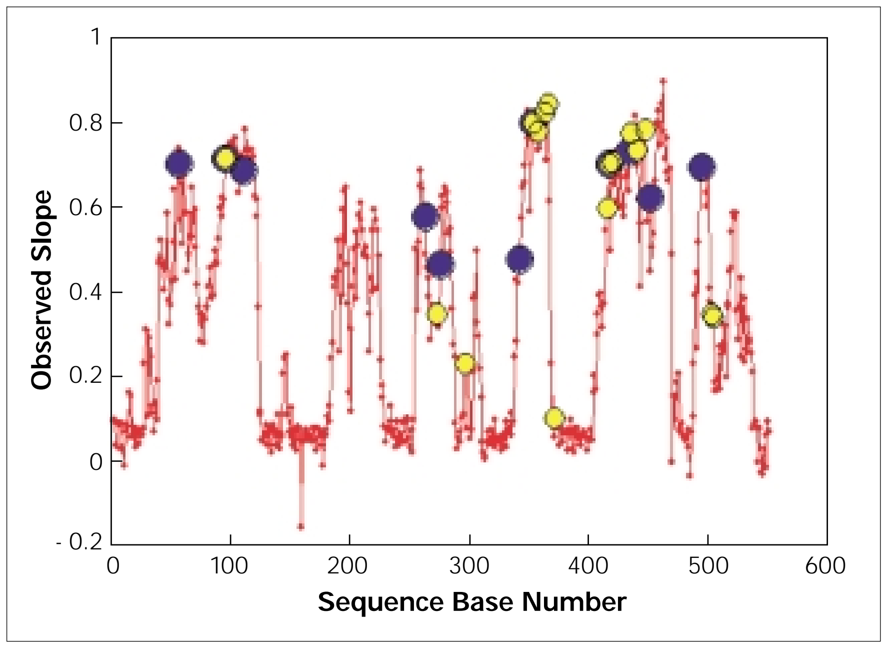


Figure 6. Example of probe selection. Observed slopes of intensity vs. concentration of all 25-mer perfect match probes are shown for one gene. Blue circles represent the eleven probes selected by the new model. Yellow circles represent the 16 probes selected by the previous heuristic method. The range of selected slopes is significantly reduced with the new model compared to the previous heuristic model. Probes selected by the new model are expected to behave more uniformly. The distribution rules also ensure, as can be seen in the figure, that the probes are more evenly spaced on the transcript.

Probe Spacing and Independence

Maximizing probe spacing is a strategy to ensure that multiple probes produce multiple, independent measurements of the target. The new probe selection system selects probe sets that are optimized for both spacing and quality as shown in Figure 6.

Number of Probes in a Probe Set

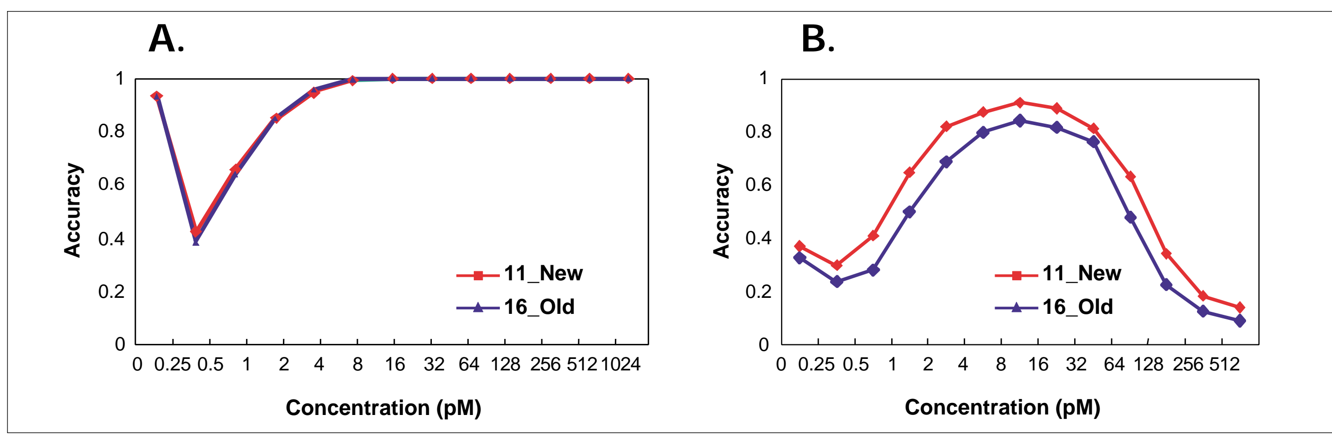
The optimized selection of high-quality probes enables the reduction of probe pairs per probe set. This allows an increased number of probe sets per array while maintaining performance. The HG-U133 set contains 11 probe pairs per probe set compared to 16-20 probe pairs per target on previous designs. Even with this decrease in number of probe pairs per probe set, the accuracy of both Detection and Change calls is the same or better for the new design (Figure 7).

Probe Set Selection

The primary goal in probe set selection is to select a probe set unique to a single transcript or common among a small set of similar transcript variants. A probe set in which all the probes exactly match multiple transcripts is annotated with an "_s" appended to the probe set name. The probe set selection process generally favors probe sets measuring fewer transcripts.

Probe sets with common probes among multiple transcripts, the _s probe sets, are frequent and to be expected due to alternative polyadenylation and alternative splicing. One transcript may be represented by both a unique and an _s probe set when transcript variation is evident. In most cases, _s probe sets represent transcripts from the same gene, but transcripts from homologous genes are sometimes also represented by the same probe set.

Occasionally, it is not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. In that case, similarity criteria are suspended and the resulting probe set is annotated with an "_x" appended to the probe set name. Such probe sets will contain some probes that are identical or highly similar to other sequences. The probes 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 is from transcripts other than the one being intentionally measured.



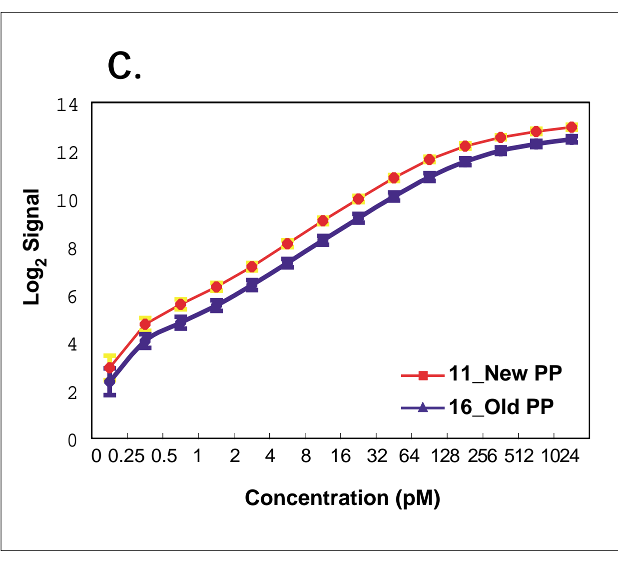


Figure 7. A. Comparison of detection call accuracy of 11 new probe pairs vs. 16 probe pairs selected using the previous rules. Accuracy values were determined for probe sets of 99 yeast genes in four replicate Latin Square experiments. At each point Accuracy is equal to the fraction of probe sets that detect the presence of target at the given concentration, using the detection call algorithm from MAS 5.0. The point for 0 pM represents absent call accuracy for spike probe sets. The graphs represent 16 probe pairs, selected by the previous heuristic method (blue curves), and 11 probe pairs, selected by the new model method (red curves).
B. Comparison of change call accuracy for 11 new probe pairs vs. 16 probe pairs selected using the previous rules. At each point Accuracy is equal to the fraction of probe sets that correctly call an increase for a two-fold change of target concentration from the base concentration in four replicate Latin Square experiments. The zero concentration point corresponds to a change from zero to 0.25 pM. Profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous heuristic method (blue curves), and 11 probe pairs, selected by the new model method (red curves). Eleven probe pairs selected by the new model outperform 16 probe pairs selected by the previous model.
C. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous heuristic method (blue curves), and 11 probe pairs, selected by the new model method (red curves). Eleven probe pairs selected by the new model outperform 16 probe pairs selected by the previous model.

Other Probe Sets

The sequence and probe selection rules described in this document were used in the creation of over 98% of the probe sets on the HG-U133 design. There exist a small group of probe sets based on the HG-U95 design that are retained on the current design. These probe sets and control probe sets are described in detail in a forthcoming technical note available in early 2002.

Feature Size

The feature size has been decreased from 20 microns to 18 microns to allow more probe sets per array. Comparative results from Latin square experiments (described above) showed that probe set performance for 18 micron features was equivalent to 20 micron features (Figure 8).

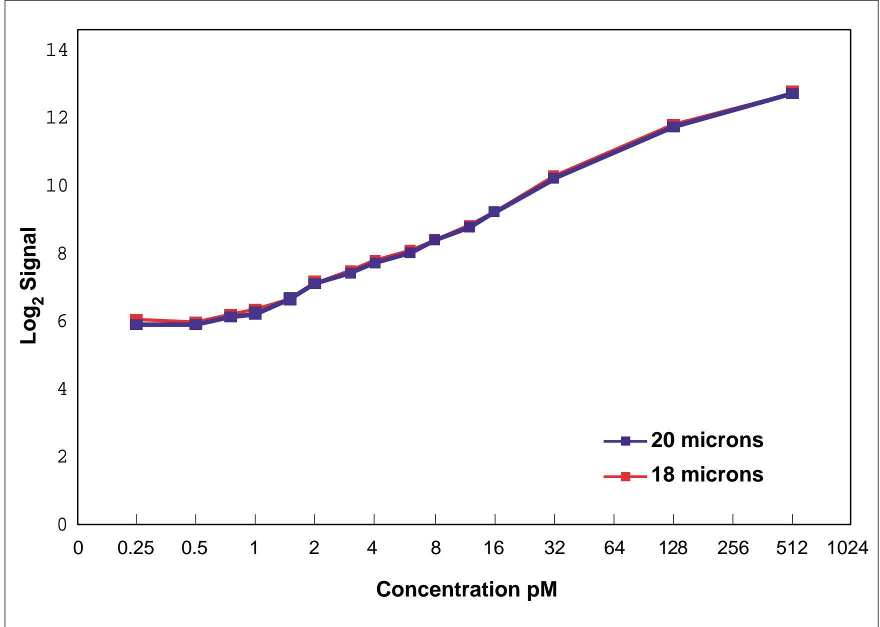


Figure 8. Comparison of Feature Size. Signal profile represents average signal of a set of 88 Human genes in response to increasing target concentration. Feature size of the probe sets was 20 microns (blue line) and 18 microns (red line). All sequences were represented by eleven probe pairs.

Summary

The Human Genome U133 Set incorporates significant advances in array design.

— Genomic sequences were used to verify sequence selection, orientation, and the quality of sequence clustering.

— Clustering information from UniGene, build 133 was used in conjunction with primary sequence and annotation information combined from a large variety of public databases to give a more complete and accurate starting sequence data set.

— Probe selection was improved by using a multiple linear regression model derived from a thermodynamic model of duplex formation, which predicts binding affinity and linear signal response to target concentration.

— Feature size reduction to 18 micron to allow for increased information density with no reduction in performance.

The resulting two-array set is a powerful tool that allows you the best view of transcription from the human genome.

Appendix

	U95 and HuGene FL	U133	Justification
Sequence Sources	UniGene, GenBank	Unigene, RefSeq, Genbank, dbEST, WUSTL, Golden Path 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	Subcluster by similarity and orientation	Similarity, orientation, and genomic position	Reduces false clusters of homologs
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 end of consensus	600 base region from end of exemplar full length mRNA or consensus. Multiple poly-A sites selected.	Full-length exemplars may be of higher sequence quality than consensus. Multiple poly-A sites improve sensitivity for alternative transcripts.
Probe Quality	Heuristic rules (e.g. not more than 10 A's in a probe). 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.	Improve selection of probes that hybridize well to the correct target and reduce non-specific cross hybridization
Probe Uniqueness	Probes which have 21 or more bases out of 25 matching targets expected to be in RNA samples are too similar, and will be avoided.	Probes which 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	Approximately equally spaced.	Spacing weighted to favor high quality and independent probes.	Ensure multiple probes give independent measurements of the target.
Number of Probes	16-20	11	Combined with algorithm and probe quality improvements, allows greater information density without reduction in information quality.
Probe Set Annotation	_s, _g, _f, _n, _r, _i	_s, _x Discontinued: _r, _i, _n Transformed: _g → _s _f → _x	Non-unique probe set types were simplified and adjusted to account for improvements in probe selection rules.
Feature Size	20 microns	18 microns	Allow greater information density without reduction in information quality

I. A description of the data analysis algorithm that accompanies this design is available in GENE EXPRESSION MONITORING TECHNICAL NOTE: New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays, Product No. 701097-rev 3.

II. CAT; DoubleTwist , Oakland, CA

III. L. Hillier, G. Lennon, M. Becker, M. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. Dubuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M. Soares, F. Tan, E. Trevaskis, K. Underwood, P. Wohldman, R. Waterston, R. Wilson and M. Marra. (1996). Generation and analysis of 280,000 human expressed sequence tags., Genome Research 6, 807-828.

IV. GeneChip® Expression Analysis: Technical Manual, 2001.

V. RepeatMasker: Smit, A.F.A. (1996) Origin of interspersed repeats in the human genome. Curr. Opin. Genet. Devel. 6 (6), 743-749.
BLASTN : Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs , Nucleic Acids Res. 25:3389-3402.
UniVector: Kitts, P.A., Madden, T.L., Sicotte, H. & Ostell, J.A. - Manuscript in preparation.

VI. Lockhart et.al. Nature Biotechnology, 14:1675-1680 (1996)

AFFYMETRIX, INC.

3380 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-362-2447 (1-888-DNA-CHIP)
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

www.affymetrix.com

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

Part No. 701133 Rev 1

©2001 Affymetrix, Inc. All rights reserved. Affymetrix®, GeneChip®, ™, HuSNP™, Jaguar™, EASI™, MicroDB™, GenFlex™, 417™, 418™, 427™, 428™, Pin-and-Ring™, Flying Objective™, ™, CustomExpress™, NetAffx™ and ® are trademarks owned or used by Affymetrix, Inc. 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, and 5,945,334; and EP 619 321; 373 203 and other U.S. or foreign patents.

AFFYMETRIX®