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. 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.

Human Array Designs

SEQUENCE SELECTION FOR THE HG-U133 FL DESIGN

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. In the consensus sequence, a base was 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.

Sources and numbers of sequences used in the HG-U133 design:

<table>
<thead>
<tr>
<th>Source</th>
<th>Release Date</th>
<th>Human Sequences Used in Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniGene</td>
<td>April 20, 2001 (#133)</td>
<td>2,688,626</td>
</tr>
<tr>
<td>dbEST</td>
<td>April 28, 2001</td>
<td>3,471,886</td>
</tr>
<tr>
<td>WUSTL</td>
<td>Feb 2001</td>
<td>1,430,516</td>
</tr>
<tr>
<td>GenBank</td>
<td>April 25, 2001 (#123.0)</td>
<td>61,523</td>
</tr>
<tr>
<td>RefSeq</td>
<td>April 30, 2001</td>
<td>12,716</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>7,665,267</td>
</tr>
</tbody>
</table>

Table 1. 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.
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 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)3 (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.

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 Manual4 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.

---

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

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence Collection</strong></td>
<td>Sequences and annotations collected from a variety of sources.</td>
</tr>
</tbody>
</table>
| **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 re-subclustered 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 chosen due to alternative polyadenylation sites. |
| **Prioritizing Probe Selection Regions** | - Regions for probe selection were prioritized based on the quality of sequence and annotation. |
VECTOR CONTAMINATION AND REPEATS
Each sequence was assessed for repeats using RepeatMasker software and for vector contamination using BLASTN and the UniVector database.3

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. 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.

Multiple 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).

Figure 2. 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 signals), if n ≥ 8 or if n = 5 and 4 ≤ n, then probes are also picked from region 2. Otherwise, if n < 5 or if n = 6, then region 2 is skipped. If the exemplar has no 3' UTR sequence then region 1 is not picked. Consequentially, if n ≥ 2, region 2 is picked, otherwise region 3 is picked.
Hybridization characteristics allowing multiple linear regression (MLR) model selection of probes based on predicted standing of probe uniqueness and behavior. The new array design uses a complete coding region and some 3’ untranslated sequence. EST-only clusters are prioritized according to strong evidence from genomic mapping and orientation information. The actual prioritization categories are summarized in the bottom section of Table 2.

### 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. 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 unifies 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”. 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 50 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 intensities (LnI) are highly correlated with observed Log (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 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. These probes are usually GC-rich, and tend to cross-hybridize with nontarget species, and hence no longer exhibit a concentration response to their specific target.

### Appendix

<table>
<thead>
<tr>
<th>U95 and HuGene FL</th>
<th>U133</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence Sources</strong></td>
<td>UniGene, GenBank</td>
<td>UniGene, RefSeq, Genbank, dbEST, WUSTL, Golden Path Draft Assembly</td>
</tr>
<tr>
<td><strong>Sequence Curation</strong></td>
<td>Filtered for repeats, vector</td>
<td>Repeats and vector screening, EST quality trimming</td>
</tr>
<tr>
<td><strong>Sequence Subclustering</strong></td>
<td>Subcluster by similarity and orientation</td>
<td>Similarity, orientation, and genomic position</td>
</tr>
<tr>
<td><strong>Sequence Orientation</strong></td>
<td>According to CDS annotation and EST read direction</td>
<td>Genomic sequence, poly-A prediction, CDS, and EST read direction</td>
</tr>
<tr>
<td><strong>Sequence Region</strong></td>
<td>600 base region from end of consensus</td>
<td>600 base region from end of exemplary full length mRNA or consensus. Multiple poly-A sites selected.</td>
</tr>
<tr>
<td><strong>Probe Quality</strong></td>
<td>Heuristic rules (e.g. not more than 10 A’s in a probe). Probe quality is assessed as a binary (yes/no) function.</td>
<td>Thermodynamic multiple linear regression model predicts intensity of probes. Probe quality assessed on a continuous scale.</td>
</tr>
<tr>
<td><strong>Probe Uniqueness</strong></td>
<td>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.</td>
<td>Probes which have two 8-mer matches, including at least one 12-mer match will be avoided. Unintended sequences.</td>
</tr>
<tr>
<td><strong>Probe Spacing</strong></td>
<td>Approximately equally spaced.</td>
<td>Spacing weighted to favor high quality and independent probes.</td>
</tr>
<tr>
<td><strong>Number of Probes</strong></td>
<td>16-20</td>
<td>11</td>
</tr>
<tr>
<td><strong>Probe Set Annotation</strong></td>
<td><em>s</em>, <em>g</em>, <em>f</em>, <em>n</em>, <em>r</em>, <em>i</em></td>
<td>Discontinued: <em>r</em>, <em>i</em>, <em>n</em></td>
</tr>
<tr>
<td><strong>Feature Size</strong></td>
<td>20 microns</td>
<td>18 microns</td>
</tr>
</tbody>
</table>
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 59 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 μM represents absent call accuracy for spike probe sets. The graphs represent 11 probe pairs, selected by the previous heuristic method (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). A. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). B. Comparison of change call accuracy for 11 new versus 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). C. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). A. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). B. Comparison of change call accuracy for 11 new versus 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). C. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). A. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). B. Comparison of change call accuracy for 11 new versus 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). C. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). A. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). B. Comparison of change call accuracy for 11 new versus 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves). C. Signal profiles are shown for probe sets comprised of 16 probe pairs, selected by the previous model (blue curves), and 11 probe pairs, selected by the new model method (red curves). 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 μM. Results are shown for probe sets comprised of 16 probe pairs, selected by the previous model (red curves), and 11 probe pairs, selected by the new model method (red curves).
**Figure 4.** A. Predicted and observed probe intensities. Profiles are for predicted (red) and observed (blue) intensity values for all 35-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.

Figure 5. Observed intensity vs. concentration for five representative probes. The black line is the best fit line relating Ln I and Ln Conc. Brown and green represent low affinity probes 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.

**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.

**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.

Chimeric 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.

**REFERENCES**

1. A description of the data analysis algorithm that accompanies this design is available in **AFFYMETRIX TECHNICAL NOTE**: New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays. Product No. 701097

2. CATT, Double-Tech™, Oakland, CA


Good Quality of a represented human gene. 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.

**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.

The blue represents an extremely high affinity, saturated probe that will changes in target concentration. These are typical of the probes that are selected. Also be discarded. Pink and red represent good quality probes with a linear response 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.

**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. 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.

**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).

**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.

**REFERENCES**

1. A description of the data analysis algorithm that accompanies this design is available in Affymetrix Technical Note: “New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays” Product No. 701097.
2. CAT, DoubleTree™, Oakland, CA.
Comparison of Feature Size.

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 59 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 probes. The graphs represent 11 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 sets 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 18 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.

Figure 8. 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.

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 EST clusters, while the second tier indicates the number of probe sets representing these clusters with regard to source annotation and annotation. The sequences are assigned to unique classifications as follows: potential full-length sequences with 3’ UTR (Full Length Including UTR); consensus evidence for polyadenylation (Evidence for Polyadenylation); consensus sequence ends from substrates containing a complete CDS (Complete CDS Consensus End); consensus sequence ends from substrates containing an EST (EST Consensus End); consensus sequence ends with evidence for polyadenylation (Evidence for Polyadenylation). Probe sets from EST-only probe sets are grouped by cluster annotation quality and include substrates with 3 or more sequences or containing a maximum assembly cluster 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), whether the cluster contains at least one EST or a sequence containing a polyadenylation site (3’). For completeness, probe sets to the opposite strand of an EST or problematic substrates are selected (Opposite Consensus End). C. Signal profiles are shown for probe sets comprised of 18 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.

Table 2.

<table>
<thead>
<tr>
<th>Classification</th>
<th>HG-U133A</th>
<th>HG-U133B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniGene Clusters</td>
<td>14,593</td>
<td>19,318</td>
<td>31,928</td>
</tr>
<tr>
<td>Additional Potential Full Lengths</td>
<td>513</td>
<td>198</td>
<td>711</td>
</tr>
<tr>
<td>Subclusters</td>
<td>18,462</td>
<td>21,070</td>
<td>39,532</td>
</tr>
<tr>
<td>Full Length Including UTR</td>
<td>13,049</td>
<td>1,556</td>
<td>14,605</td>
</tr>
<tr>
<td>Extended Full Length</td>
<td>171</td>
<td>58</td>
<td>229</td>
</tr>
<tr>
<td>Strongest evidence for polyadenylation</td>
<td>3,228</td>
<td>6,929</td>
<td>10,140</td>
</tr>
<tr>
<td>Complete CDS Consensus End</td>
<td>570</td>
<td>74</td>
<td>643</td>
</tr>
<tr>
<td>Non-EST Consensus End</td>
<td>2,526</td>
<td>2,765</td>
<td>5,291</td>
</tr>
<tr>
<td>Evidence for polyadenylation</td>
<td>993</td>
<td>5,156</td>
<td>5,149</td>
</tr>
<tr>
<td>ESTOnly clusters</td>
<td>683</td>
<td>619</td>
<td>1,302</td>
</tr>
<tr>
<td>Oriented, Mapped, and 3’</td>
<td>176</td>
<td>150</td>
<td>326</td>
</tr>
<tr>
<td>Oriented and 3’</td>
<td>279</td>
<td>9,153</td>
<td>9,432</td>
</tr>
<tr>
<td>Mapped and 3’</td>
<td>33</td>
<td>590</td>
<td>623</td>
</tr>
<tr>
<td>3’</td>
<td>14</td>
<td>76</td>
<td>90</td>
</tr>
<tr>
<td>Opposite Consensus End</td>
<td>22</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Distant Consensus End</td>
<td>683</td>
<td>619</td>
<td>1,302</td>
</tr>
</tbody>
</table>

Identification of Cross-hybridizing Probes

Microarrays with 794 different probes and 533 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 separate array which had 2604 probes, with 100 random mismatch probes to each to perfect match probe.

Probes were designed to each target a unique probe set to a single transcript or common among a small set of transcript variants. A probe set in which all the probes exactly match multiple transcripts is annotated with an ‘_x’ 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 ‘_x’ 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 ‘_x’ probe set when transcript variation is evident. In most cases, ‘_x’ 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 such 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 ‘_s’ appended to the
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. 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 unifies 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.” 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 Lat (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 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.

**Multiple sequences and probe sets.**

<table>
<thead>
<tr>
<th>Hs.54829.2 Hs.40855060.0</th>
<th>0 500 1000 1500 2000 2800</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.54829.1 Hs.40855060.0</td>
<td>0 500 1000 1500 2000 2800</td>
</tr>
<tr>
<td>Hs.54829.3 Hs.40855060.0</td>
<td>0 500 1000 1500 2000 2800</td>
</tr>
<tr>
<td>Hs.900486.556 Hs.13661193.1</td>
<td>0 500 1000 1500 2000 2800</td>
</tr>
</tbody>
</table>

**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.

- **Sequence Sources**: UniGene, GenBank
- **Sequence Curation**: Filtered for repeats, vector
- **Subclustering**: Subcluster by similarity and orientation
- **Probe Quality**: Thermodynamic multiple linear regression model predicts intensity of probes. Probe quality assessed on a continuous scale.
- **Probe Spacing**: Approximately equally spaced.
- **Number of Probes**: 16-20

**Justification**

- Improved annotation, classification, and sequence quality
- Avoid low quality EST sequence regions, thereby improving consensus sequence quality
- Reduces false clusters of homologs
- Improves orientation calls by using sequence-based methods in addition to annotations
- Full-length exemplars may be of higher sequence quality than consensus. Multiple poly-A sites improve sensitivity for alternative transcripts
- Minimize specific cross hybridization to similar targets from unintended sequences
- Ensure multiple probes give independent measurements of the target
- Combined with algorithm and probe quality improvements, allows greater information density without reduction in information quality
- Non-unique probe set types were simplified and adjusted to account for improvements in probe selection rules
- Allow greater information density without reduction in information quality

**Figure 3.** This 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. Probes with the highest alignment to the seed sequence are aligned to the right side of the probe. The red bars at the top indicate the span of each probe set that detects the seed sequence. Due to the high similarity of the two regions, they are considered to be one probe set. The light blue regions are divergent, unaligned sequences. Variance bars indicate either unaligned sequences due to low complexity or divergence of aligned sequences. For UniGene cluster Hs.79732 representing the fibulin1 gene, there are four subclusters represented by four exemplars (blue bars) and another by a single consensus sequence (gray bar). There is also a potential full-length sequence shown in this alignment (Hs.500458.560) that is not listed in UniGene, but is also a transcript for fibulin1. 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).

**Appendix**

<table>
<thead>
<tr>
<th>U95 and HuGene FL</th>
<th>U133</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence Sources:</td>
<td>UniGene, GenBank</td>
</tr>
<tr>
<td>Sequence Curation</td>
<td>Filtered for repeats, vector</td>
</tr>
<tr>
<td>Subclustering</td>
<td>Subcluster by similarity and orientation</td>
</tr>
<tr>
<td>Sequence Orientation</td>
<td>According to CDS annotation and EST read direction</td>
</tr>
<tr>
<td>Sequence Region</td>
<td>600 base region from end of consensus</td>
</tr>
<tr>
<td>Probe Quality</td>
<td>Thermodynamic multiple linear regression model predicts intensity of probes. Probe quality assessed on a continuous scale.</td>
</tr>
<tr>
<td>Probe Spacing</td>
<td>Approximately equally spaced.</td>
</tr>
<tr>
<td>Number of Probes</td>
<td>16-20</td>
</tr>
<tr>
<td>Justification</td>
<td>Improved annotation, classification, and sequence quality</td>
</tr>
</tbody>
</table>

**Probe Quality Metric**

- **Probe Quality**: Thermodynamic multiple linear regression model predicts intensity of probes. Probe quality assessed on a continuous scale.
- **Probe Spacing**: Approximately equally spaced.
- **Number of Probes**: 16-20

**Probe Selection**

- **Sequence Sources**: UniGene, GenBank
- **Sequence Curation**: Filtered for repeats, vector
- **Subclustering**: Subcluster by similarity and orientation
- **Probe Quality**: Thermodynamic multiple linear regression model predicts intensity of probes. Probe quality assessed on a continuous scale.
- **Probe Spacing**: Approximately equally spaced.
- **Number of Probes**: 16-20

**Justification**

- Improved annotation, classification, and sequence quality
- Avoid low quality EST sequence regions, thereby improving consensus sequence quality
- Reduces false clusters of homologs
- Improves orientation calls by using sequence-based methods in addition to annotations
- Full-length exemplars may be of higher sequence quality than consensus. Multiple poly-A sites improve sensitivity for alternative transcripts
- Minimize specific cross hybridization to similar targets from unintended sequences
- Ensure multiple probes give independent measurements of the target
- Combined with algorithm and probe quality improvements, allows greater information density without reduction in information quality
- Non-unique probe set types were simplified and adjusted to account for improvements in probe selection rules
- Allow greater information density without reduction in information quality
VECTOR CONTAMINATION AND REPEATS
Each sequence was assessed for repeats using RepeatMasker software and for vector contamination using BLASTN and the UniVector database.5

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.6 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 aambiguous 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).

Figure 2. If the full-length exemplar has 3’ UTR, then probes are picked from the 600 base region ending at 1. For the stack of ESTs (annotated 3’, or possessing polyadenylation signals), if n≥2 and x<400, then probes are also picked from region 2. Otherwise, if n<8 or if n≥2 and x<400, then region 2 is skipped. If the exemplar has no 3’ UTR sequence then region 1 is not picked. Consequently, if n≥2, region 2 is picked, otherwise region 3 is picked.

Gene expression monitoring
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 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) (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.

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 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.

---

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

Human Array Designs

**SEQUENCE SELECTION FOR THE HG-U133 FL SET**

In the HG-U133 FL 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 HG-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). 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

---

**Table 1.**

<table>
<thead>
<tr>
<th>Source</th>
<th>Release Date</th>
<th>Human Sequences Used in Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniGene</td>
<td>April 20, 2001 (#133)</td>
<td>2,688,626</td>
</tr>
<tr>
<td>dbEST</td>
<td>April 28, 2001</td>
<td>3,471,886</td>
</tr>
<tr>
<td>WUPLST</td>
<td>Feb 2001</td>
<td>1,430,516</td>
</tr>
<tr>
<td>GenBank</td>
<td>April 25, 2001 (#123.0)</td>
<td>61,523</td>
</tr>
<tr>
<td>RefSeq</td>
<td>April 30, 2001</td>
<td>12,716</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>7,665,267</td>
</tr>
</tbody>
</table>

**Sources and numbers of sequences used in the HG-U133 design.**