AFFYMETRIX®

Technical Note

GeneChip® Exon Array Design

The primary objective for the design of this first-generation GeneChip® Exon Array is to interrogate each potential exon with one probe set over the entire genome on a single array. With this array, researchers are able to cost-effectively analyze gene expression at the level of transcript diversity on a whole-genome scale for the first time.

This Technical Note describes in detail the implementation of the exon array design concept and provides a summary of the array content and probe selection results using the GeneChip® Human Exon 1.0 ST Array (Sense Target) as an example. In particular, changes in this design are compared and contrasted with existing arrays, such as the GeneChip® Human Genome U133 Plus 2.0 Array, in order to highlight the implications and advantages of using the Human Exon 1.0 ST Array for expression analysis.

Introduction

On the GeneChip® Human Exon 1.0 ST Array, 5,362,207 features are used to interrogate one million exon clusters (collections of overlapping exons) with over 1.4 million probe sets. Probe sequences were selected using the high-quality human genome assembly (July 2003, hg16, build 34) and a variety of genome annotations including those inferred from human, mouse, and rat cDNAs (Appendix 1, Appendix 2, and Appendix 3). A number of publicly available gene prediction sets were also used (Appendix 4 and Appendix 5) including Ensembl, GENSCAN, and Vega.

Implementation of the Exon Array Design Concept

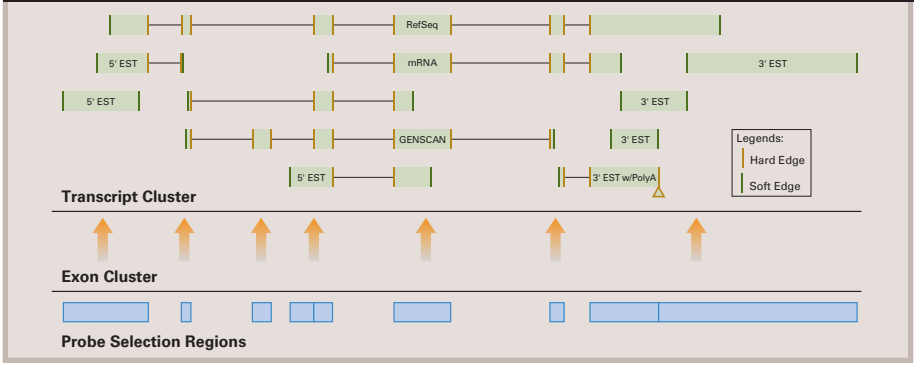
PROBE SELECTION REGIONS, EXON CLUSTERS, TRANSCRIPT CLUSTERS, AND GENES

One significant feature of this exon array design is that the various input sequences and annotations were consolidated onto the genome and then divided into probe selection regions (PSRs). The PSRs are contiguous and do not overlap in genomic space. An example of PSRs resulting from the consolidation process for the GeneChip® Human Exon 1.0 ST Array is shown in Figure 1.

The consolidation process was carried out by projecting all the input annotations onto the genome to infer the exonic vs. intronic and intergenic regions. The transcribed blocks (exon clusters) were further fragmented into PSRs when hard edges were observed as described in the next section Hard Edges and Soft Edges for Defining PSRs.

To delineate the relationship of individual exons, exon clusters were grouped into transcript clusters as a post-design annotation activity. Those exon clusters that shared splice sites, or were derived from overlapping exonic sequences, or were single-exon clusters bounded on the genome by spliced content, were annotated to belong to the same transcript cluster. A transcript cluster roughly corresponds to a gene.

Figure 1. Consolidation of input content into probe selection regions (PSRs). In the array design input sequence consolidation process, all the input annotations were projected onto the genome to infer transcribed regions. Internal splice sites, polyadenylation sites (indicated by a triangle), and CDS start/stop positions were typically used to infer "hard edges." Hard edges may result in the fragmentation of a contiguous piece of transcribed sequence (an exon cluster) into multiple probe selection regions (PSRs).



HARD EDGES AND SOFT EDGES FOR DEFINING PSRs

The consolidation process leveraged several tunable metrics. The exact settings were adjusted for each annotation source (see Appendix 6) to achieve a balance between sufficient fragmentation of exon clusters into PSRs and avoid erroneous fragmentation due to incorrect information and noise in the input annotation sources. Two concepts were used in the content consolidation process to describe the different treatments of various input sequences:

- Hard Edges: the end of the sequence that defines the boundary of a PSR and cannot be extended beyond the border by other evidence.
- Soft Edges: the end of the sequence that can be extended by another available source of evidence that may not result in the boundary of a PSR.

Hard edges were inferred from either 3' and 5' splice sites, CDS start and stop positions, polyadenylation sites, or internal splice sites. Notable exceptions were the syntenic cDNA content from mouse and rat as well as ESTs which contained non-consensus splice sites. The syntenic content from mouse and rat has a tendency to be fragmented, particularly within the untranslated regions (UTRs); thus it was used only to infer transcribed regions (i.e., exon clusters), and not splice events (i.e., PSR boundaries within an exon cluster). Similarly, ESTs with non-consensus splice sites were used only to infer transcribed regions, and not splice events. This significantly reduced over-fragmentation in highly expressed genes which have thousands and even tens of thousands of ESTs.

For Ensembl and putative full-length mRNA-based annotations, the CDS starts and stops were treated as hard edges. This helped ensure good probe coverage for the coding portion of the terminal exons and also for single-exon protein-coding genes. It also mitigated possible issues associated with overextension of the terminal exons by other annotation sources.

Figure 2. Representation of the TPM2 gene on the GeneChip® Human Exon 1.0 ST Array. The probes and probe sets for TPM2 (2nd tier below the coordinates) are shown in this figure in relationship to a subset of the annotations used to design probe sets for the TPM2 locus. Note that there are several probe sets on the opposite strand based on exoniphy, sgp, and other annotation sources.

It should be noted that there were exceptions when the CDS start or stop was not treated as a hard edge if the CDS start or stop corresponded to the transcript start or stop. This exception was due to greater ambiguity of the full-length transcript when only the CDS was annotated.

Small gaps in the annotations relative to the genomic coordinates were ignored and the ends of some of the cDNA-based annotations were trimmed to prevent misalignment of extra bases in the cDNA sequence resulting in aberrant hard edges for small, dubious terminal exons.

Additionally, splice sites were not treated as hard edges when the cDNA alignment(s) that inferred the splice site contained unaligned cDNA sequence bases. In most cases, the outer bounds of the transcript annotation were also treated only as soft edges due to the fact that many annotations are 5' and/or 3' incomplete.

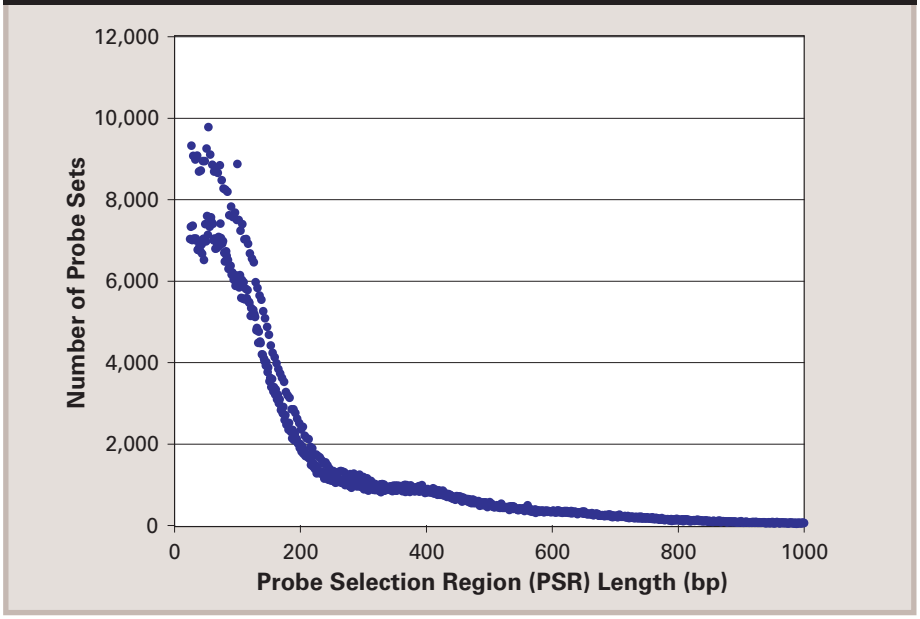
An example of the result of the consolidation process is shown for TPM2 (see Figure 2). Due to the space limitations, most of the input data are not shown here, hence there are PSRs outside of the exonic regions inferred by RefSeq and the other input annotations shown. Notice there is a large amount of coverage by PSRs for the entire RefSeq sequence for TPM2 including the known canonical cassette exons for TPM2.

Another observation to note in the TPM2 example is the presence of PSRs and probe sets on the opposite strand (these are illustrated in the top half of Figure 2). A certain percentage of transcript clusters also have evidence of antisense transcription. In reviewing these occurrences, the opposite strand probe sets frequently fall into one of the following categories: misoriented cDNA sequences, correct or incorrect *ab initio* gene predictions, or legitimate antisense transcripts inferred from cDNA input data.

CONTENT OVERVIEW AND IMPLICATIONS

The exon array design implementation for the human genome resulted in 1,796,124 PSRs from 1,084,639 exon clusters (overlapping sets of exon variants). A large number of very small PSRs (less than 10 bp) were obtained primarily from noise in the cDNA sequence set, particularly ESTs. Only those PSRs 25 bp in length or longer

Figure 3. Distribution of PSR lengths on GeneChip® Human Exon 1.0 ST Array. For probe sets represented in the design, the median PSR length is 123 bp. The number of PSRs with lengths that are a multiple of three are generally higher than the number of PSRs whose lengths are not multiples of three due to *ab initio* gene predictions that are based on protein coding potentials. Hence, the basis for what appears to be two separate distributions in this graph. There are 7,072 probe sets with a PSR length of 2 kb or greater that are not plotted here.



are represented on the array. In the final design of the Human Exon 1.0 ST Array, the median PSR length is 123 bp (Figure 3). There is a slight bias in the PSR length (in bp) to be multiples of three since this is driven by protein coding single exon predictions (i.e., GENSCAN suboptimal exon predictions).

Following probe selection, over 1.4 million PSRs representing about 1 million exon clusters are represented on the Human Exon 1.0 ST Array by over 1.4 million probe sets and over 5 million probes. The probe sets are grouped into over 300,000 transcript clusters with over 90,000 transcript clusters containing more than one probe set.

The type of annotation supporting each probe set can be an important parameter for downstream data analysis. About half of the probe sets on the Human Exon 1.0 ST Array are based on a single type of annotation as shown in Figure 4. Most of these single-annotation type probe sets are derived from ESTs and GENSCAN predictions (Figure 5 and Figure 6). Collectively, about a quarter of the array content is based solely on ESTs and another quarter is based solely on GENSCAN predictions. The other half of the array consists primarily of content based on a combination of annotation sources (Figure 4, Figure 5, and Figure 6).

In terms of the depth of the evidence supporting the design, roughly a sixth of the probe sets derived from EST-only content are supported by more than one EST.

The results of the exon array design strategy have several characteristics and implications that should be taken into consideration when interpreting results obtained from the Human Exon 1.0 ST Array:

1. The design captures moderate to large variations in transcript diversity. PSRs smaller than 25 bp, i.e., a 3 bp shift in splice site, are not represented on the array. Although it is possible to interrogate such splice events with junction probe sets including probes spanning multiple exons which are not contiguous on the genome, junction probe sets are not included as part of this exon array design (see below).
2. There are no junction probe sets on the array. Given that we have extremely limited understanding and knowledge of which splice forms are present, this

Figure 4. Distribution of probe sets based on the number of different supporting annotation sources. Probe sets can be supported by annotations from multiple sources. The number of annotation sources supporting a PSR is plotted here. For example, a PSR supported by several ESTs, an Ensembl transcript, and two RefSeq transcripts would have a tier count of 3. There is no supporting annotation sources based on the build 34 annotations for 26,022 probe sets. For annotation purposes a probe set must be contained by the individual annotation, whereas the design of the array can take into account multiple annotations which are merged into a single probe selection region.

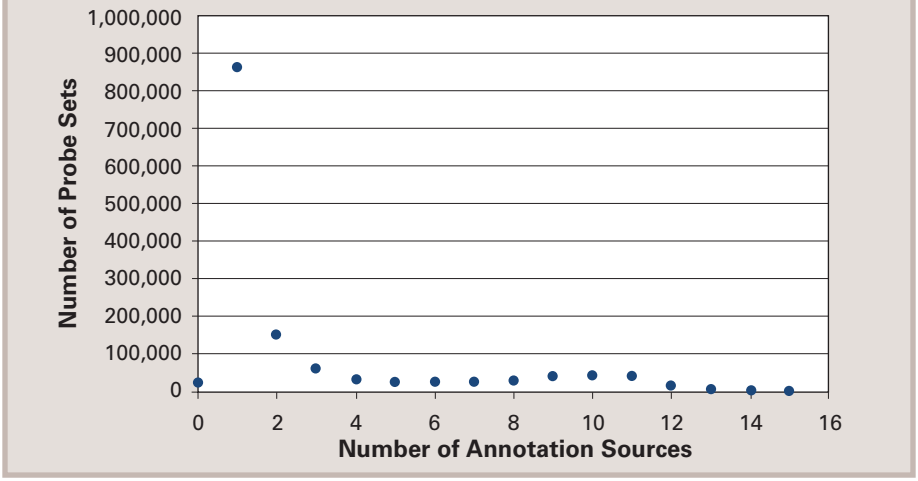
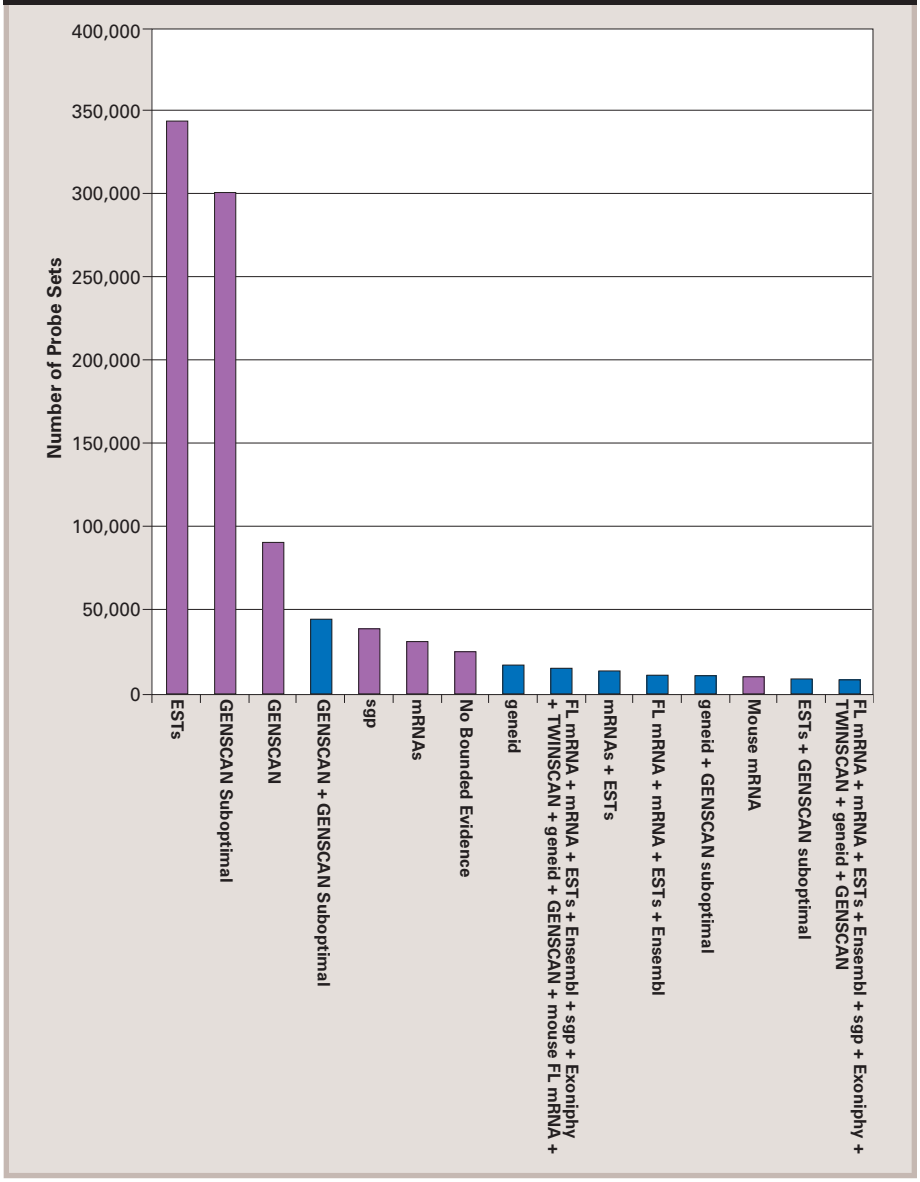


Figure 5. Distribution of probe sets based on their supporting evidence sources. Only the top 15 collections are shown. Note that a large part of the array is supported only by ESTs or GENSCAN predictions. Collections with more than one annotation type are shown in blue.



first-generation, whole-genome, transcript-variation array design focuses on interrogating transcribed regions (i.e., exons) rather than splice events. For researchers interested in focusing on a subset of well-characterized genes and transcripts, splice junction probe array designs are possible through Affymetrix' custom design program.

3. Gene-level probe coverage may vary. This is dependent on the number of exons, or more specifically, the number of PSRs, present in a gene. Most known protein coding transcripts are covered by many probes; it is not uncommon to have 50 or more probes against a RefSeq mRNA sequence (Figure 7) spanning the entire transcript. In very few cases, a single-exon transcript may have fairly low probe coverage.
4. A very small percentage of RefSeq exons may have no coverage. Although represented in very low frequency, some exons larger than 25 bp may be fragmented by other evidence, resulting in PSRs under 25 bp; therefore, with no probe coverage for that exon. One such example is one of the exons in the CLTA gene (Figure 8). Approximately 0.2 percent of exons inferred from RefSeq mRNAs are fragmented into sets of PSRs less than 25 bp in length.

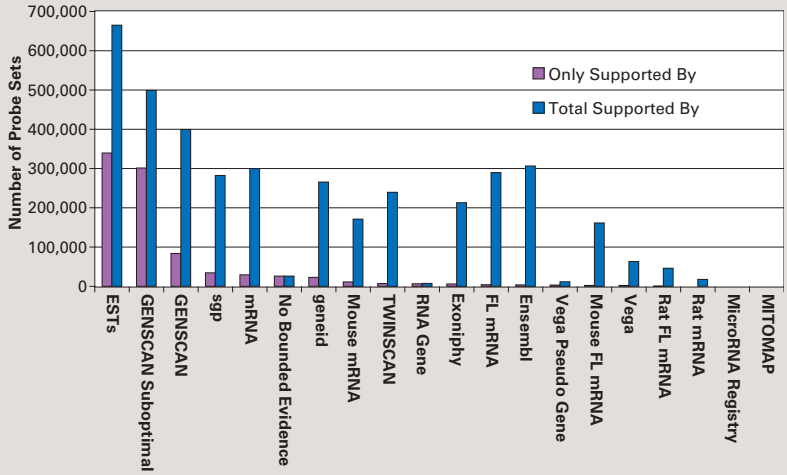
It is important to note that much of the design information described above for each probe set, including the type of annotation, the number of supporting evidence sequences, the PSR sequence, the genomic coordinates, and sequence of each probe, are provided in annotation files and through the NetAffx™ Analysis Center. Providing such detailed information should help the researchers in quickly narrowing down their analysis results to a smaller subset of genomic loci as well as to assist in better interpretation of the array results.

PROBE SELECTION

Probe sets were selected against each of the PSRs. The same probe model and probe selection process used for currently existing expression arrays, such as the GeneChip Human Genome U133 Plus 2.0 Array (HG-U133 Plus 2.0 Array), were utilized here. However, some key modifications were made specifically for the Human Exon 1.0 ST Array in order to provide a complete solution suitable for the array's unique application as well as the new Whole Transcript (WT) Sense Target Labeling Assay.

- Four perfect match probes were selected for each probe set. The number of probes representing each PSR is primarily constrained by the length of the PSR. Due to the smaller PSR size with a median of 123 bp, up to four probes were chosen for each probe set. It has been observed that using multiple probes for each probe set generates more robust quantitative information and that reducing the number of probes results in lower sensitivity and specificity. Thus at the PSR level, performance will be compromised rel-

Figure 6. Distribution of probe sets supported by a particular evidence source. ESTs and GENSCAN predictions alone account for about 50 percent of the array design. Of the ~340,000 EST-only supported probe sets, about 65,000 of them are supported by more than 1 EST.



ative to the 11-probe-pair HG-U133 Plus 2.0 Array probe set. At the gene level, however, performance is expected to be recovered because of the multiple PSRs representing each gene with the gene-level performance comparable to that obtained on the HG-U133 Plus 2.0 Array.

- A common set of probes is used for background correction for all probes on the array. This is in contrast to a probe-specific background subtraction approach where a specific mismatch is selected for each perfect match probe. Two collections of background probes have been tiled in this design:
  - **Antigenomic Background Probes.** This is a collection of probes that were selected because they are not present in the human genome (or seven other genomes including mouse, rat, Drosophila, *C. elegans*, *S. cerevisiae*, Arabidopsis, and *E. coli*) and are not expected to cross-hybridize to transcribed human sequences. For each of the 26 bins of varying GC count (zero Gs or Cs out of a 25-mer sequence, to all 25 bases are Gs or Cs), approximately 1,000 25-mer probes were chosen and represented on the array.
  - **Genomic Background Probes.** This is an alternative collection of probes to be used in place of a specific mismatch. Unlike the Antigenomic Background Probes, these Genomic Background Probes were mismatch probes of which their perfect match counterparts do match the genome, but in regions less likely to be expressed. For each of the 26 bins of varying GC count (0-25), approximately 1,000 25-mer probes were chosen.

These two collections of background probes are used to estimate the probe-specific background by comparing a perfect match probe intensity to the median intensity of all the background probes with a matching GC content. When used in combination with the GeneChip® Whole Transcript (WT) Sense Target Labeling Assay targets, this background subtraction strategy has been shown to be an effective replacement for the probe-specific mismatch probe approach, which has been used in 3' designs such as the HG-U133Plus 2.0 Array.

- The probe model parameters were tuned for DNA target. This is because the new WT Sense Target Labeling Assay generates DNA targets possessing similar, but distinct, hybridization properties from that of the cRNA target generated by the GeneChip 3' IVT Amplification Assay.
- All probes on the array were evaluated for potential cross hybridization to other PSRs in the Human Exon 1.0 ST Array design as well as for splice junctions observed in the input data set. It should be noted that probes were not evaluated for cross hybridization against the entire human genome, which would have unnecessarily removed good-performing probes because they share similarity

Figure 7. RefSeq mRNA probe coverage on the GeneChip® Human Exon 1.0 ST Array. RefSeq sequences are more likely to be represented by multiples of four probes due to the fact that in general four probes are selected for each probe set. There are 23 mRNAs with more than 400 probes covering the sequence that are not plotted in this graph.

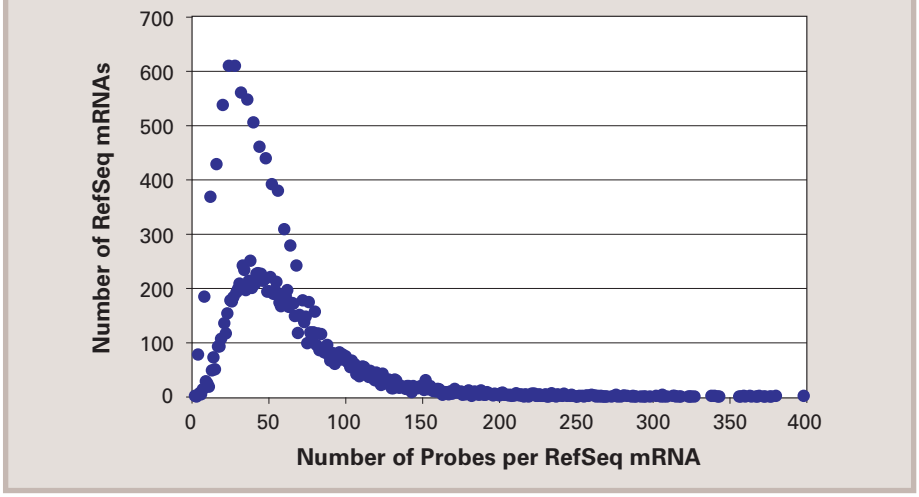
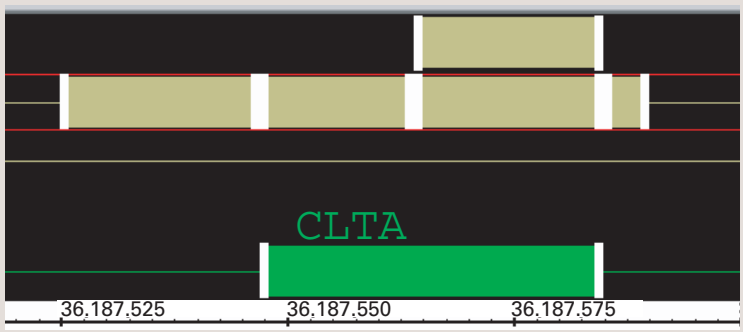


Figure 8. Over-fragmentation of a small exon in the CLTA gene. Due to a hard edge inferred from a GENSCAN suboptimal exon prediction, the 38 bp exon in the CLTA gene was fragmented into two PSRs, each of which is less than 25 bp. The top brown block is the GENSCAN suboptimal exon prediction. The bottom green block is the 38 bp CLTA exon with RefSeq support. The set of brown blocks separated by white lines in the middle is the resulting PSRs, of which no probe sets were selected for the GeneChip® Human Exon 1.0 ST Array design.



with the untranscribed region in the genome resulting in poorer performing probe sets. Most of the resulting probe sets are unique to a single location in the putative transcriptome (Figure 9).

- Probes were selected for hybridization to the sense-strand target, not the antisense target as is done for 3' array designs.

One implication of an exon-based array design relative to the 3' array designs is that the number of candidate probes for any given probe set is much smaller (see Figure 3) compared with the typical 3' 600 bp used for 3'-array designs. The result is that more of the probes in the probe set are not independent with respect to the region of the transcript they are interrogating (Figure 10). The number of independent interrogation positions for each probe set is also provided in the annotation library files and can be used for filtering and prioritizing the array results.

Processing of cDNA Sequences into Genomic Annotations

While the public genome annotations provide a simple source of input for the probe selection region enumeration, the cDNA sequences do not. Using an internally developed cDNA sequence analysis pipeline, genome annotations were generated from cDNA sequences using the following approach:

1. When possible, EST read direction and CDS annotations were determined.
2. The cDNA sequences (and WUSTL EST trace, when available) were checked for the presence of polyadenylation sites and signals.
3. The cDNA sequences were aligned with their respective genomes using blat.
4. The cDNA-genome alignments were evaluated for consensus splice sites.
5. The cDNA sequence orientation was determined using a probabilistic model combining information about:
   a) EST read direction
   b) CDS orientation
   c) Polyadenylation sites and signals
   d) Consensus splice site usage
6. A genome transcript annotation was inferred for each cDNA sequence using the orientation and alignment information.

The result is that all the cDNA sequences that could be oriented (Appendix 7) and aligned were used as a source of content for the design of the Human Exon 1.0 ST Array. Only those cDNA sequences that aligned with at least 80 percent of their bases to a given genomic locus are included in the design. In addition, only the locus with the best alignment to cDNA in the genome was used; in cases where there were multiple best alignments, all were used.

The mouse and rat genome annotations were further processed by "mapping" the annotations from their respective genome to the human genome (Figure 11). This was achieved by using the whole-genome alignments from the Genome Browser group at the University of California, Santa Cruz to translate coordinates in mouse and rat genomes to syntenic coordinates in human. The final coverage of the human genome is summarized in Figure 12.

Other Control Probes on Human Exon 1.0 ST Array

In addition to the main content, additional probe sets including critical controls are represented on the Human Exon 1.0 ST Array and they are divided into several

Figure 9. Probe set cross-hybridization types. As part of the probe selection process, probes were evaluated for potential cross-hybridization against other putative transcribed sequences in the design (but not the entire genome). Most of the probe sets on the array are predicted to uniquely hybridize to a single target ("unique" or "cross hybridization type 1"). A small number of probe sets have probes that cross-hybridize to sequences elsewhere in the genome, but all probes in the probe set still hybridize to a common transcribed genomic region ("similar" or "cross hybridization type 2"). For the exon array design strategy, the similar probe sets reflect probe sets which interrogate more than one location in the genome. A number of probe sets have inconsistent cross-hybridization properties amongst the probes in the same probe set. These mixed ("mixed" or "cross hybridization type 3") probe sets may be measuring target from multiple sources.

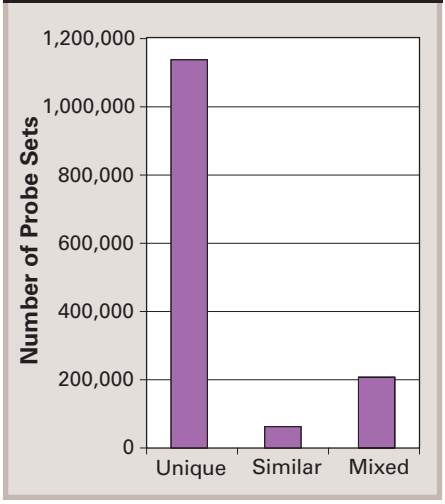
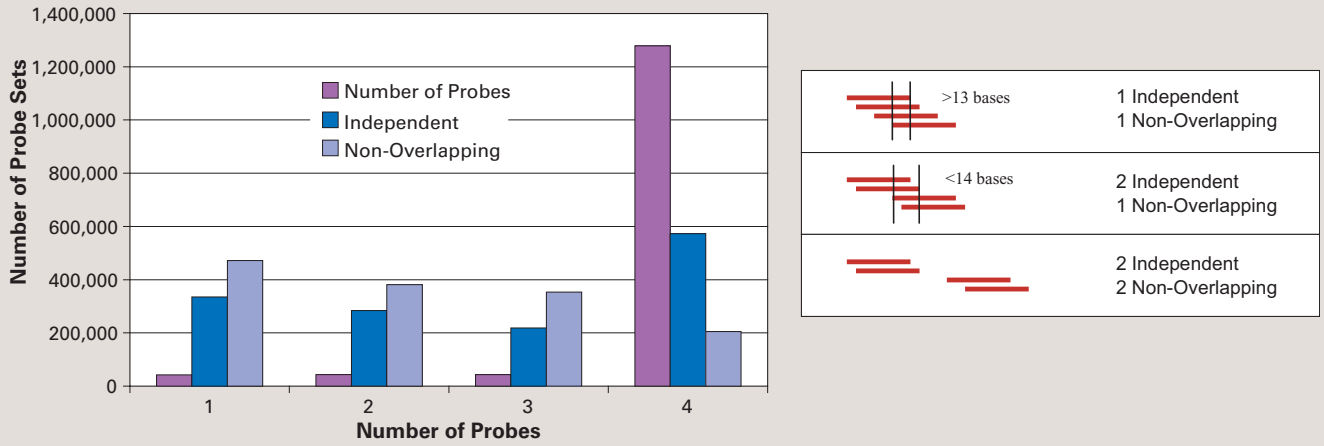


Figure 10. Distribution of non-overlapping and independent probes within the probe sets on the GeneChip® Human Exon 1.0 ST Array. Two distinct, but related, concepts are used to define the relationship among the probes within a probe set: a) The number of non-overlapping probes. A probe is considered non-overlapping if it does not share any sequence with another probe in the probe set. All probes that overlap with each other are counted as only one "non-overlapping" probe. b) The number of independent probes. Probes which overlap by more than 13 bases are counted as only one "independent probe." Several examples of this are shown by red lines representing probes positioned relative to the genome. As probe sets are chosen for small exons it is not uncommon to have probe sets with probes that overlap or are not independent. The annotation for these two concepts can be used to prioritize the probe sets from the array results to facilitate interpretation of results.



different classes:

- **Antigenomic Background Probes**
  Described previously.
- **Genomic Background Probes**
  Described previously.
- **Affymetrix Controls**
  A number of the standard Affymetrix control probe sets are tiled on the Human Exon 1.0 ST Array. These include the *bioB*, *bioC*, *bioD*, and *cre* probe sets for the bacterial spikes; antisense target perfect match and mismatch probes are tiled. Sense target probe sets are tiled for the standard poly-A spikes (*dap*, *lys*, *phe*, and *thr*).
- **Intron-Exon Controls**
  Probe sets were generated for the consensus sequences of the HG-U133 Plus 2.0 Array normalization control genes. All of the exons as well as 100 bp-long intronic regions were used as PSRs to select perfect match probes as controls.
- **Unmapped Human mRNAs**
  Putative full-length human mRNA sequences which did not align to the human genome were tiled with one perfect match sense target probe set every 5 bp for the last 600 bp of the transcript.

Figure 11. Result of synteny mapping from mouse to human. (A) An example where a mouse cDNA sequence is mapped and nearly identical to a human RefSeq annotation. (B) An example of a mouse RefSeq mapped onto a region of human with no corresponding annotation.

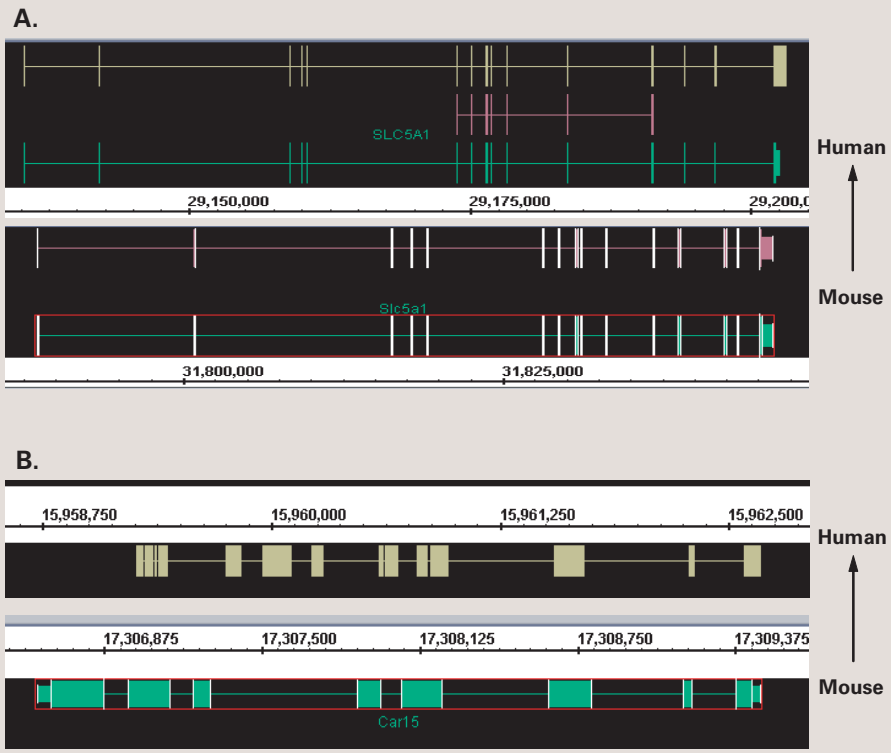
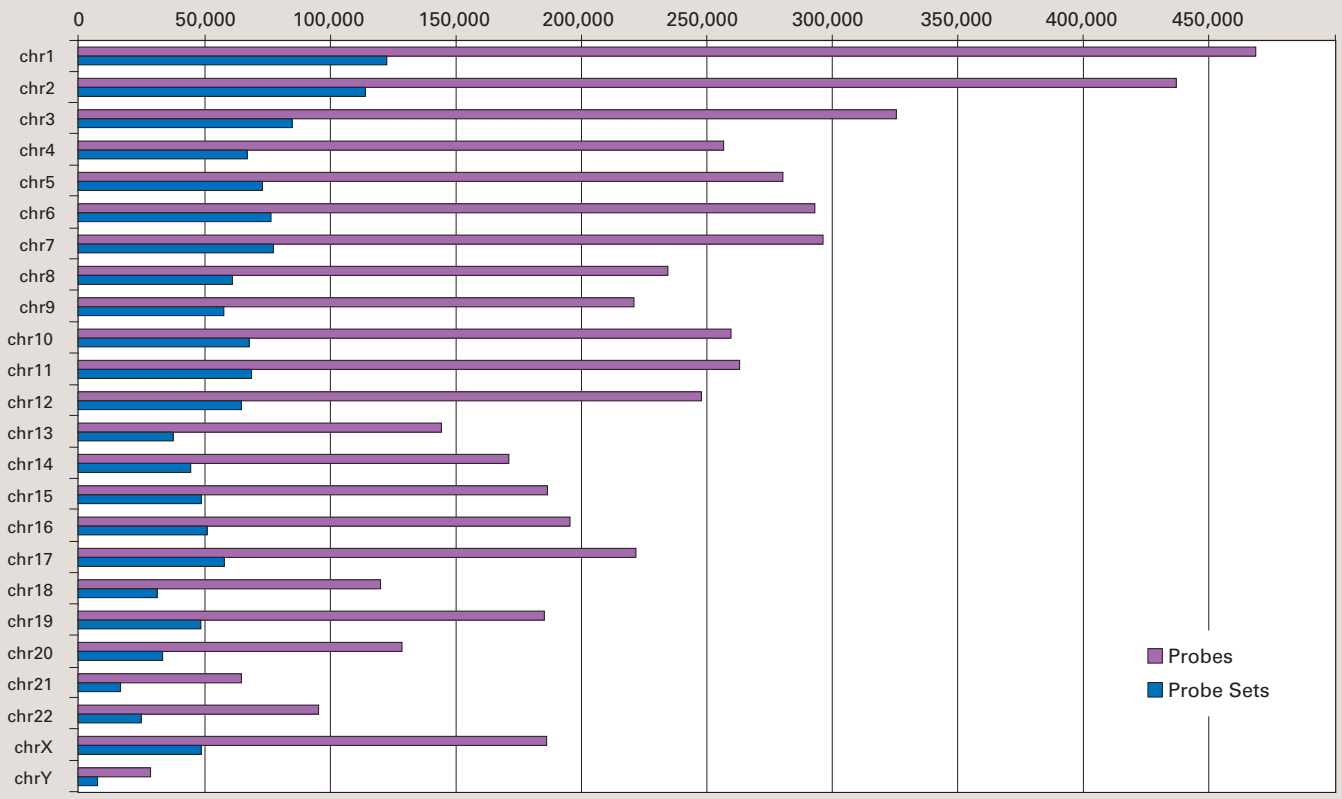


Figure 12. Probe and probe set genome coverage. Coverage for the mitochondria (chrM), unordered (random), and unplaced (chrUn) chromosomes are not shown.



Appendix

Appendix 1: Genome assemblies used to design the GeneChip® Human Exon 1.0 ST Array

Species	UCSC Version	Date	Source
Human	hg16 (build 34)	July 2003	genome.ucsc.edu
Mouse	mm4	Oct 2003	genome.ucsc.edu
Rat	rn3	June 2003	genome.ucsc.edu

Appendix 2: cDNA sequence sources used to design the GeneChip® Human Exon 1.0 ST Array

cDNA Source	Label Used	Date
GenBank Release 139	fl, mrna, mouse-fl, mouse-mrna, rat-fl, rat-mrna	Dec 15, 2003
RefSeq Cumulative Update	fl, mouse-fl, rat-fl	Feb 7, 2004
dbEST	est	Feb 5, 2004
WUSTL EST Traces		Jan 30, 2004
Entrez Query for recent human mRNA sequence Submissions	fl	Jun 7, 2004

"fl" is an abbreviation for full-length and includes RefSeq mRNA transcripts as well as putative full-length transcripts from GenBank® as indicated by a "Complete CDS" on the definition line.

Appendix

Appendix 3: The number of cDNA sequences used to design the GeneChip® Human Exon 1.0 ST Array

cDNA Source	Label Used	# of Human Sequences	# of Mouse Sequences	# of Rat Sequences
GenBank mRNAs	fl	46,753	24,746	6,386
RefSeq mRNAs	fl	34,933	16,548	4,797
GenBank mRNAs	mrna	86,420	25,747	5,286
GenBank High-throughput mRNAs	mrna	7,732	64,106	344
dbESTs	est	5,471,625	4,056,277	583,838

Rat and mouse EST sequences were not used for the Human Exon 1.0 ST Array design. The numbers are shown for informational purposes only.

Appendix 4: Genome annotations used to design the GeneChip® Human Exon 1.0 ST Array

Annotation	Label Used	Date	Data Source	# of Transcripts	# of Exons
Ensembl V21.34d.1	ensGene	May 10, 2004	www.ensembl.org	35,685	325,353
Exoniphy	exoniphy	May 25, 2004	genome.ucsc.edu		184,616
geneid	geneid	Jun 15, 2004	genome.ucsc.edu	32,255	216,731
GENSCAN	genscan	Mar 12, 2004	genome.ucsc.edu	42,974	326,300
GENSCAN suboptimal exons	genscanSubopt	Mar 17, 2004	genome.ucsc.edu		518,038
miRBase	MicroRNAregistry	Apr 8, 2004	genome.ucsc.edu		187
MITOMAP	mitomap	Jun 16, 2004	www.mitomap.org		72
Non-Coding RNA Genes	maGene	Jun 16, 2004	genome.ucsc.edu		7,220
sgp	sgpGene	Mar 6, 2004	genome.ucsc.edu	42,880	236,382
TWINSCAN	twinscan	Jun 15, 2004	genome.ucsc.edu	21,369	193,454
Vega	vegaGene	Mar 12, 2004	genome.ucsc.edu	11,700	80,546
Vega Pseudo Genes	vegaPseudoGene	Mar 12, 2004	genome.ucsc.edu	3,071	5,125

Appendix 5: Genome annotation decriptions

Gene Annotation	Primary URL	Label Used	Description
geneid	www1.imim.es/software/geneid/	geneid	*ab initio* gene prediction
Ensembl	www.ensembl.org/	ensGene	confirmed gene prediction based on cDNA and protein sequences
sgp	genome.imim.es/software/sgp2/	sgpGene	combination of geneid *ab initio* predictions with tblastx comparisons of genomes between related organisms
GENSCAN	genes.mit.edu/GENSCANinfo.html	genscan, genscanSubopt	*ab initio* gene prediction
Exoniphy recomb2004.pdf	www.cse.ucsc.edu/~acs/	exoniphy	exon predictions based on both protein coding potential and exon conservation
Non-Coding RNA Genes	genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg16&g=rnaGene	rnaGene	various non-coding RNA annotations (tRNA, snoRNAs, ...) based on tRNAscan and Wublast results
MITOMAP	www.mitomap.org/	mitomap	curated mitochondria gene annotations
miRBase	microrna.sanger.ac.uk/sequences/index.shtml	microRNAregistry	curated precursor miRNA annotations
TWINSCAN	genes.cs.wustl.edu/	twinscan	gene prediction using combination of *ab initio* approaches (similar to GENSCAN) and genome conservation
Vega	vega.sanger.ac.uk/	vegaGene, vegaPseudoGene	curated transcript annotations

Appendix

Appendix 6: Tunable metrics applied to each annotation source when PSRs were generated

	Transcript Hard Edges	Splice Site Hard Edges	CDS Start/Stop Hard Edges	Merge Gaps	Trim Transcript Edges
ensGene		x	x	< 9 bp	0 bp
geneid		x		< 9 bp	0 bp
GENSCAN		x		< 9 bp	0 bp
GENSCANSuboptimal	x	x		< 9 bp	0 bp
Exoniphy	x	x		< 9 bp	0 bp
maGene	x	x		< 9 bp	0 bp
MITOMAP	x	x		< 9 bp	0 bp
microRNAregistry	x	x		< 9 bp	0 bp
sgpGene		x		< 9 bp	0 bp
TWINSCAN		x		< 9 bp	0 bp
vegaGene		x	x	< 9 bp	0 bp
vegaPseudoGene		x		< 9 bp	0 bp
mouse-fl				< 19 bp	10 bp
mouse-mrna				< 19 bp	10 bp
rat-fl				< 19 bp	10 bp
rat-mrna				< 19 bp	10 bp
fl		x	x	< 19 bp	10 bp
mrna		x		< 19 bp	10 bp
est		x		< 19 bp	20 bp

Hard edges are locations in the genome where one contiguous putative expressed region was split into two probe selection regions (PSRs). Hard edges were inferred from the splice sites of EST alignments only for those EST alignments where all the splice sites were one of the three consensus splice sites (gt-ag, at-ac, gc-ag). Syntenic content from mouse and rat is often fragmented, particularly in the UTRs; hence they were not used to set splice site hard edges to prevent over-fragmentation. Single exon annotation sources were defined as hard edges to prevent over extension of these single exon annotations. For annotations with putative-complete CDS annotations, the CDS start/stop positions were set as hard edges. Small gaps were merged and the ends of cDNA alignments were trimmed to prevent over-fragmentation of the PSRs.

Appendix 7: cDNA sequence orientation for human cDNA sequences

Sequence Class	Number Oriented	Number Unoriented	Percent Unoriented
fl	71,686	0	0%
mrna	66,834	27,318	29%
est	4,587,634	883,991	16%

Unoriented sequences were not used in the design.

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
UK and Others Tel: +44 (0) 1628 552550
France Tel: 0800919505
Germany Tel: 01803001334
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

AFFYMETRIX JAPAN K.K.

Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com Please visit our web site for international distributor contact information.

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

Part No. 702026 Rev. 1