AFFYMETRIX®

Technical Note

GeneChip® Gene 1.0 ST Array Design

This Technical Note describes in detail the implementation of the GeneChip® Gene 1.0 ST Array design concept, and provides a summary of the array content and probe selection results using the Human Gene 1.0 ST Array as an example. Additional support materials and more detailed information, such as quality assessment of whole transcript-based microarray data and whole transcript-based Gene 1.0 ST Array performance, can be found at www.affymetrix.com.

Introduction

The GeneChip® Gene 1.0 ST Array System is designed to measure the gene expression of well-annotated genes, using a single probe set per gene comprised of multiple probes that are distributed along the entire length of the genomic locus. This design strategy yields a more robust and accurate analysis of the cumulative transcription activities of the gene of interest. Affymetrix provides a complete system solution for this new-generation gene expression profiling tool. The Gene 1.0 ST Array System is compatible with the existing Whole Transcript (WT) Sense Target Labeling and Control Reagents, the Fluidics Station 450 and GeneChip Scanner 3000 7G.

Key Human Gene 1.0 ST Array Properties

- Interrogates 28,869 well-annotated genes with 764,885 distinct probes
- Based on the March 2006 (UCSC hg18, NCBI Build 36) human genome sequence assembly
- Content derived from curated and provisional RefSeq mRNAs (November 3, 2006; 24,188 sequences), the full EMBL/EBI Ensembl data set (October 2006, version 41.36c; 56,364 transcripts), and GenBank mRNAs annotated as having a complete coding sequence (CDS) (November 3, 2006; 56,249 sequences)
- 25-mer probes designed to be distributed across the transcribed regions of each gene with a median of 26 probes per gene
- Compatible with WT Sense Target Labeling and Control Reagent Kits for maximum coverage of the entire gene
- Perfect match-only (PM-only) array design with probes that hybridize to sense target
- Various control probe sets including:
    - o Hybridization control probe sets for BioB, BioC, BioD and CreX pre-labeled spikes
    - o Assay control probe sets for the Dap, Phe, Lys and Thr polyA unlabeled spikes
    - o Putative exon and intron control probe sets from putative constitutive genes for use as pseudo positive/negative controls
    - o Generic background probes

Evolution of Expression Array Designs and Design Strategies

3′ BIASED EXPRESSION ARRAY DESIGNS

Initial expression array designs such as the HuGeneFL array were based on known transcript sequences. Probe sets were selected directly against the mRNA sequence using a set of heuristics for probe performance. This design strategy was replaced with more comprehensive bioinformatics and probe modeling for the Human Genome U133 family of array designs (i.e., the GeneChip® HG-U133 Plus 2.0 Array). For the Human Genome U133 family of arrays, multiple transcript sequences, including expressed sequence tags (ESTs), were used to generate consensus transcript sequences, orient consensus sequences and identify polyadenylation sites. Significant improvements were made in modeling probe performance and selecting probes. For each polyadenylation site, a probe set consisting of 11 perfect match and mismatch probe pairs were selected within the 5' region of the polyadenylation site (for more details, see the Technical Note, *Array Design for the Human Genome U133 Set*, available at http://www.affymetrix.com/support).

Figure 1 shows the HG-U133 Plus 2.0 consensus sequence (in yellow) for the Mitochondrial Ribosome Recycling Factor (MRRF) gene aligned to the genome along with an 11-probe probe set (small red box overlays) at the 3' end of this gene. Also shown is a smaller EST-based consensus sequence within the same genomic region.

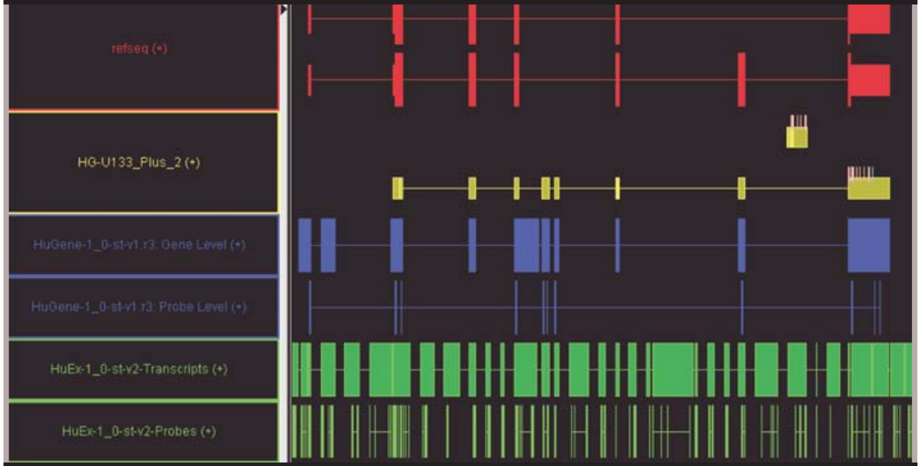
Over the course of all these designs, array density increased, enabling more transcripts to be interrogated on fewer arrays. Additionally, public sequence data, including the full human genome sequence, became more complete. The end result was a mature platform for genome-wide, gene-level expression analysis on a single array, the Human Genome U133 Plus 2.0 Array.

WHOLE TRANSCRIPT-BASED ARRAY DESIGNS

All of these earlier designs relied upon the use of a 3' biased target preparation protocol. As a result, the probe sets are biased toward the 3' ends with multiple probe sets per gene in cases of known alternative polyadenylation and alternative terminal exon usage. With the launch of the Human Exon 1.0 ST Array came a new target preparation protocol, the Whole Transcript (WT) Assay, which generates target across the entire transcript, not just the 3' end. For more detailed information regarding the WT Assay and performance, refer to the Technical Note, *Human Exon 1.0 ST Array and WT Sense Target Labeling Assay for Genome-Wide, Exon-Level Expression Analysis*, available at http://www.affymetrix.com/support.

This innovation removed the need to anchor probes to the 3' end of genes, and opened the door to improved gene-level estimates with probes covering the entire gene, as well as the opportunity for exon-level analysis of splice variants on the Human Exon 1.0 ST Array. For more information on the Human Exon 1.0 ST Array, refer to the Technical Note, *Array Design for the Human Exon 1.0 ST Array*, available at http://www.affymetrix.com/support.

Figure 1: MRRF coverage on the HG-U133 Plus 2.0 Array, Human Gene 1.0 ST Array and Human Exon 1.0 ST Array.


refseq (+)
HG-U133_Plus_2 (+)
HuGene-1_0-st-v1.r3: Gene Level (+)
HuGene-1_0-st-v1.r3: Probe Level (+)
HuEx-1_0-st-v2-Transcripts (+)
HuEx-1_0-st-v2-Probes (+)

This Integrated Genome Browser (IGB) screen shot shows the Mitochondrial Ribosome Recycling Factor (MRRF) locus coverage for three different Affymetrix expression arrays. The top track (red) shows two RefSeq splice variants for this gene. The second track (yellow) shows two HG-U133 Plus 2.0 consensus sequences with the 3' biased probe sets (shown as red ticks on the yellow consensus sequence). One HG-U133 Plus 2.0 probe set represents the known MRRF gene structure while the other HG-U133 Plus 2.0 consensus sequence reflects a more speculative EST cluster. The Human Gene 1.0 ST Array gene bounds for this gene (first blue track) have broader probe coverage (second blue track) than the HG-U133 Plus 2.0 Array. The Human Exon 1.0 ST Array design (in green) shows higher per-exon probe coverage (second gene-level track showing probes) and additional probe sets covering more speculative parts of this locus (first green track showing the full gene bounds for this locus on the exon array).

In addition to assay improvements, the Human Exon 1.0 ST Array also represented another large step in increased density, with over 1 million exons. Exons are interrogated by four probes in most cases, resulting in over 5 million probes on a single array. The Human Exon 1.0 ST Array was designed from the ground up around the completed human genome assembly. As a result, the well-known exons, as well as the more speculative and predictive parts of a gene, are represented. An example of the strength of this design strategy is seen by looking at the MRRF locus shown in Figure 1. Both known parts of the MRRF gene structure as well as more speculative and predicted parts of the MRRF gene are covered. A convenient implication of the new assay and array is that a single gene-level expression estimate can be generated from the exon array.

The Human Gene 1.0 ST Array is a focused gene-level expression array that represents 28,869 well annotated full-length genes from RefSeq, Ensembl and putative complete CDS GenBank transcripts. The Human Gene 1.0 ST Array design, wherever possible, uses a subset of the same probes on the Human Exon 1.0 ST Array to interrogate the more focused, better-annotated content at the gene level. Probes are designed across the whole transcript to provide a more accurate representation of total transcription activity for the gene locus. Predicted and discovery-oriented content from the Human Exon 1.0 ST Array has been dropped and in most cases there are fewer probes per exon; together this permits the use of a smaller, more affordable chip format for the Human Gene 1.0 ST Array. Figure 1 shows the reduced probe coverage in known parts of the MRRF gene as well as a loss of probes within the more speculative parts of the MRRF gene (blue track).

Robust and Accurate Gene-level Expression Analysis with the Gene 1.0 ST Array

In order to target each "gene" for the desired level of probe coverage on the Gene 1.0 ST and Exon 1.0 ST Arrays, the boundaries of each gene were defined by calculating "gene bounds". Conceptually, gene bounds are the projection of all exons for a given gene onto the genome. See the gene-level gene bounds for the ATPase, class I, type 8B, member 2 (ATP8B2) locus in Figure 2A (first track). The gene bounds were calculated in a hierarchical manner, using better-annotated evidence first (i.e., RefSeq) and more speculative content later (i.e., Ensembl predictions). For more details on the construction of gene bounds, please see the white paper, *Exon Probe Set Annotations and Transcript Cluster Groupings,* available at http://www.affymetrix.com/support. For the Human Gene 1.0 ST Array, only RefSeq, Ensembl and putative complete CDS mRNA from GenBank were used to generate gene-bound annotations.

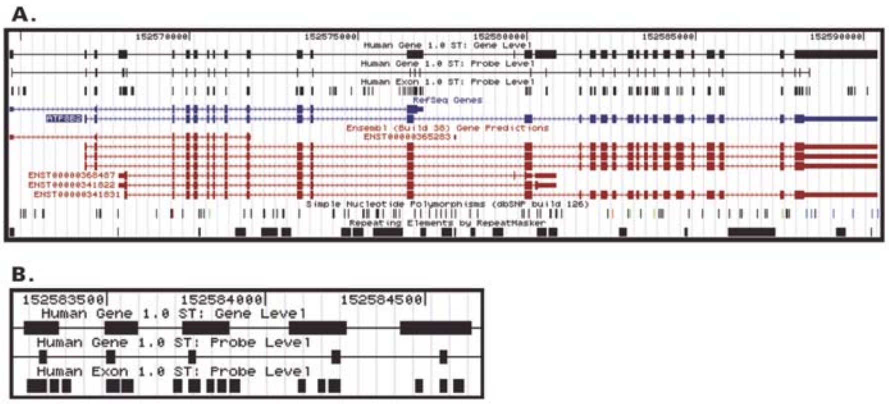
To ensure robust gene-level estimates of expressed RNA, a target of 25 probes per gene was selected for each gene bound. Preference was given to probes already on the Human Exon 1.0 ST Array. In short, probes were selected uniformly over the gene structure (Figure 2A, Probe-level Track). In many cases only a subset of the probes on the exon array were selected (Figure 2B).

Some of the gene bounds were covered by less than 25 probes from the Human Exon 1.0 ST Array, so additional probes were selected from the genome sequence for the gene bounds. In these cases new probes were picked uniformly over the gene structure. Around 80 percent of the probes on the Human Gene 1.0 ST Array are also present on the Human Exon 1.0 ST Array. The Human Exon 1.0 ST Array has a substantial number of additional probes for more speculative genes and exons as well as more probes in each exon for many known genes (Figure 2B).

To improve the accuracy of the gene-level estimates, an effort was made to minimize the number of potentially cross-hybridizing probes. Candidate probes with a large number of 16-mer matches to repetitive portions of the human genome were excluded. In addition, during the selection of the final set of probes for each gene bounds, preference was given to probes with a unique 25-mer match to the human genome. As a result, some parts of the gene structure may lack probe coverage. In the end, 90 percent of the gene-level probe sets contain only probes that match uniquely to the genome. For the remaining 10 percent, mixed probe sets were selected, allowing for both unique and non-unique probes. In most cases, these mixed gene-level probe sets reflect highly conserved gene families or genes that have high homology to pseudogene(s). Unique versus mixed hybridization designations are available from NetAffx™ Analysis Center (http://www.affymetrix.com/analysis/) and in the NetAffx CSV annotation file, located at http://www.affymetrix.com/support.

An additional 227 transcript-based probe sets were included to increase the coverage of RefSeq transcripts not covered by the main design; these include transcripts which did not align to the genome. A single transcript-based probe set was generated for each of the 227 RefSeq mRNA sequences by selecting probes from the entire length of the transcript sequence.

Figure 2: Human Gene 1.0 ST Array content for ATP8B2.


A.
152570000
152575000
152580000
152585000
152590000
Human Gene 1.0 ST: Gene Level
Human Gene 1.0 ST: Probe Level
Human Exon 1.0 ST: Probe Level
RefSeq Genes
ATP8B2
Ensembl (Build 38) Gene Predictions
ENST00000365283
ENST00000368487
ENST00000341822
ENST00000341831
Simple Nucleotide Polymorphisms (dbSNP build 126)
Repeating Elements by RepeatMasker
B.
152583500
152584000
152584500
Human Gene 1.0 ST: Gene Level
Human Gene 1.0 ST: Probe Level
Human Exon 1.0 ST: Probe Level

This figure shows two screenshots from the UCSC Genome Browser (http://genome.ucsc.edu/) showing the ATPase, class I, type 8B, member 2 (ATP8B2) gene locus and the bed files for the Human Gene and Exon 1.0 ST Arrays.

(A) The full transcribed region of this locus. The first track, "Human Gene 1.0 ST: Gene Level," shows the Affymetrix gene bounds annotation for this locus based on RefSeq, Ensembl and GenBank transcript annotations. The second track, "Human Gene 1.0 ST: Probe Level," shows the individual probes for this gene; probes in the same gene-level probe set are reflected by the horizontal line through the probes. This track shows the relatively uniform probe coverage over all the exons for this gene. The third track, "Human Exon 1.0 ST: Probe Level," shows all the probes on the Human Exon 1.0 ST Array. This track shows that there are several probes on the Human Exon 1.0 ST Array that map between the exons of the RefSeq and Ensembl transcripts representing the more speculative content on the Human Exon 1.0 ST array. The RefSeq and Ensembl track reflect some of the input transcript annotations used to generate the Affymetrix gene bounds annotation.

(B) A close-up of several exons, illustrating that the Human Exon 1.0 ST Array has increased probe coverage for some of the exons in the gene compared to the Human Gene 1.0 ST Array.

Figure 3: Distribution of probes per gene on the Human Gene 1.0 ST Array. The array design target was 25 probes per gene with probes being relatively uniform over all of the exons for each gene. As such, the number of probes per probe set can vary depending on how many exons the gene has and the sequence composition of the gene. The mean number of probes is 28 per gene and the median is 26.

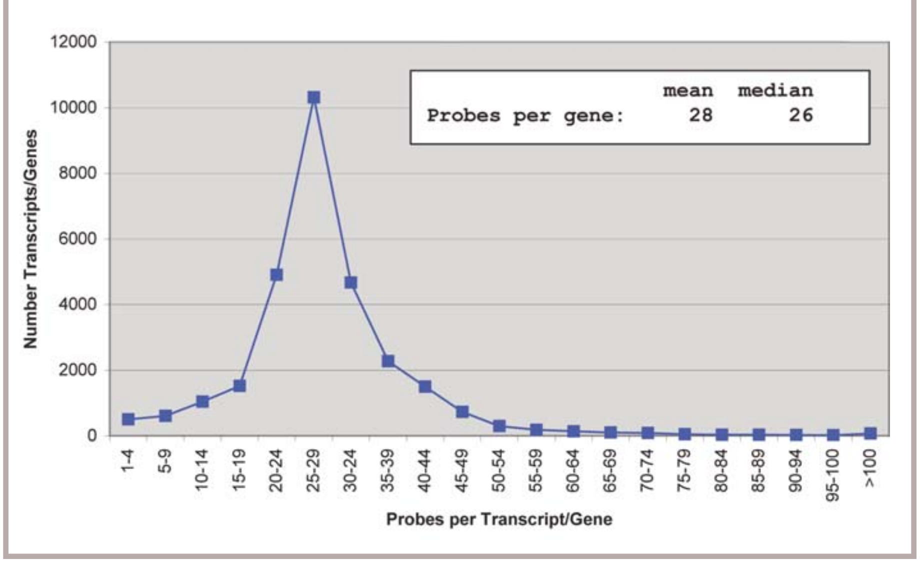
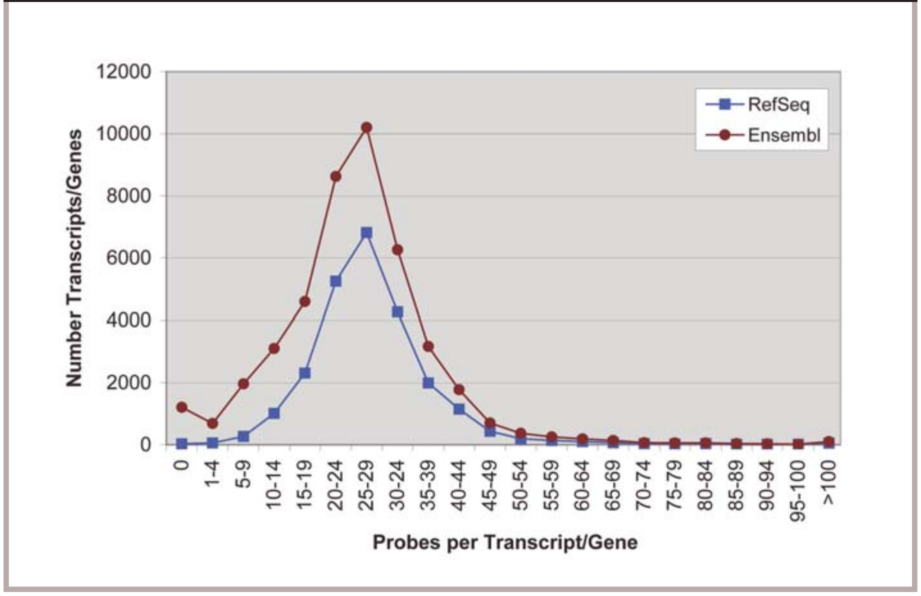


Figure 4: Distribution of probes per transcript on the Human Gene 1.0 ST Array. While the Human Gene 1.0 ST Array design focuses on having one probe set per gene, one way to evaluate gene coverage is to look at transcript coverage instead. Here we evaluate RefSeq and Ensembl transcript coverage by showing how many probes on the Human Gene 1.0 ST Array match each of the transcripts. More than 98 percent of the RefSeq transcripts is covered by 10 or more probes; more than 90 percent of the Ensembl protein-coding transcript set is covered by 10 or more probes.



The end result is a single array with 28,869 gene-level probe sets composed of 764,885 distinct probes, with a mean of 28 probes per gene and a median of 26 probes per gene (Figure 3).

Excellent Coverage of Known Genes

As of January 4, 2007, 24,198 of the 24,259 (99.7 percent) sequences present in the RefSeq database are covered by four or more probes on the array (Figure 4). More than 98 percent of RefSeq and more than 90 percent of the Ensembl protein-coding transcripts are covered by 10 or more probes.

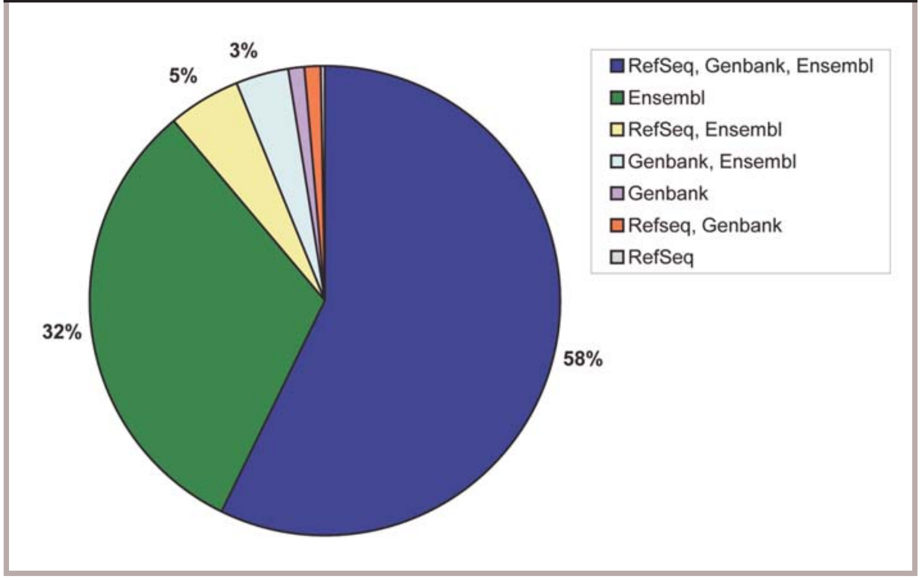
The majority (~60 percent) of gene bounds contained on the array are supported by evidence from all three sources—RefSeq, Ensembl and GenBank putative full-length mRNA's (Figure 5). Approximately one-third of the gene bounds are supported by Ensembl evidence alone. This fraction represents non-coding mRNAs and slightly more speculative gene content. The remaining probe sets are supported by one or two of the input sources, as illustrated in Figure 5.

Sample, Labeling and Hybridization Controls for Monitoring the Entire Microarray Experiment

BioB, BioC, BioD and CreX Prelabeled Spikes: Using the same probes as those on the Human Exon 1.0 ST Array, this panel of probe sets represents hybridization control genes consisting of *bioB*, *bioC*, *bioD* and *cre*. *BioB*, *bioC* and *bioD* are genes in the biotin synthesis pathway of *E. coli*, and *cre* is the recombinase gene from P1 bacteriophage. These controls are part of the Hybridization Control Kit and are useful in evaluating the hybridization and wash steps.

Dap, Phe, Lys and Thr PolyA Unlabeled Spikes: These probe sets are designed from several *B. subtilis* genes (*lys*, *phe*, *thr* and *dap*). These probe sets can be assayed using a Poly-A RNA Control Kit that contains *in vitro*-synthesized, polyadeny-

Figure 5: Breakdown of evidence source for gene-level probe sets. This figure shows the breakdown of supporting evidence for the gene-level probe sets on the Human Gene 1.0 ST Array. Around 60 percent of the probe sets are supported by RefSeq, Ensembl and GenBank. Another 32 percent are supported only by Ensembl.



lated transcripts that are pre-mixed at staggered concentrations to enable GeneChip® probe array users to assess the overall success of the target prep steps.

Putative Exon and Intron Control Probe sets: Using the same probes as those on the Human Exon 1.0 ST Array, these control probe sets are designed to interrogate putative exonic and intronic regions from a set of putative constitutively expressed genes. In short, four probe probe sets were selected for each putative exon. Four probes were selected for up to three 100bp segments within each putative intron. These control probe sets can be used to assess overall sample and data quality. (See the *Quality Assessment of Exon and Gene Arrays* white paper for more details.)

Generic Background Probes: Using the same probes as those on the Human Exon 1.0 ST Array (antigenomic background probes), these probes can be used to estimate probe-specific background. 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 (from zero G/Cs out of a 25-mer sequence, to all 25 bases being G/Cs), approximately 1,000 25-mer probes were chosen for each bin. (See the *Exon Array Background Correction* white paper for more information.)

Complete Analysis Solution, Comprehensive and Current Annotations and Extensive Support Files

Using Affymetrix Expression Console™ software, available for download at www.affymetrix.com, analyzing gene-level expression on Human Exon 1.0 ST Arrays and Human Gene 1.0 ST Arrays is as easy as 3' biased expression analysis. Probe-level analysis of Human Gene 1.0 ST Array CEL files can be done within Expression Console including the ability to do quality assessment of array results based on the control probe sets mentioned above. From within Expression Console 1.1 or higher, users can download all necessary library files and NetAffx annotations from www.affymetrix.com, as well as export gene-level expression results along with NetAffx annotations for use in a variety of third-party tools.

Genomic and biological annotations for the Human Gene 1.0 ST Array are updated and accessible through the NetAffx™ Analysis Center, located at http://www.affymetrix.com/analysis. Genomic and biological annotations include:

- Genome location
- Sequence information
- Gene symbol/title
- Assignments to public transcripts
- Gene ontology assignments
- Pathway information

In addition, a number of supporting files are available through the array support page, located at http://www.affymetrix.com/support:

- **BED Files:** These files are used to visualize the Human Gene 1.0 ST Array content in the context of the human genome using either the Integrated Genome Browser (IGB) or the UCSC Genome Browser. The Human Gene 1.0 ST Array gene-level and probe-level tracks in Figure 1 and Figure 2 were created using these BED files.
- **GFF Files:** These files contain basic design time information about the array. These include probe sequence, probe location in the CEL file, probe location in the genome, and how probes are grouped into probe sets.
- **Analysis Library Files:** These files are required by Expression Console and the Affymetrix Power Tools (APT) for doing probe-level analysis. Expression Console can download the necessary files via the File -> Download Library Files menu option.
- **NetAffx CSV Annotation Files:** These files contain all of the genomic and biological annotations hosted in NetAffx. Expression Console can download these files via the File -> Download Annotation Files menu option.

- **Comparison Spreadsheets**: These files contain information on how content from various arrays relate to one another. Comparison spreadsheets are provided against the Human Genome U133 Plus 2.0 and the Human Exon 1.0 ST Arrays.

In addition to the support files, sample data is available from http://www.affymetrix.com/support/datasets.affx.

Summary

The Human, Mouse and Rat Gene 1.0 ST Arrays are the latest products in the family of Affymetrix expression arrays offering whole-transcript coverage. The 169-format Gene 1.0 ST Arrays contain approximately 25 probes designed across the full-length of 28,869, 28,853 and 27,342 well-annotated genes for Human, Mouse, and Rat, respectively, providing a more complete and more accurate picture of gene expression than 3' biased expression array designs.

The Gene 1.0 ST Array System uses sparser probe coverage than the Exon 1.0 ST Array System (the first in the WT-based array family) and covers only well-annotated content at the gene level. The Gene 1.0 ST Array System therefore provides the same advantages of the Exon 1.0 ST Array System for gene-level analysis in a more affordable format, but without the discovery content and exon-level coverage that enables the high-resolution study of known and predicted alternative splicing.

Notes:

AFFYMETRIX, INC.

3420 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. 702488 Rev. 2

©2007 Affymetrix, Inc. All rights reserved. Affymetrix®, ®, GeneChip®, HuSNP®, GenFlex®, Flying Objective™, CustomExpress®, CustomSeq®, NetAffx™, Tools To Take You As Far As Your Vision®, and The Way Ahead™, Powered by Affymetrix™, and GeneChip-compatible™, are trademarks of Affymetrix, Inc. All other trademarks are the property of their respective owners. Array products may be covered by one or more of the following patents and/or sold under license from Oxford Gene Technology: U.S. Patent Nos. 5,445,934; 5,700,637; 5,744,305; 5,945,334; 6,054,270; 6,140,044; 6,261,776; 6,291,183; 6,346,413; 6,399,365; 6,420,169; 6,551,817; 6,610,482; 6,733,977; and EP 619 321; 373 203 and other U.S. or foreign patents.