

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

Transcript Assignment for NetAffx™ Annotations

The NetAffx™ Transcript Assignment Pipeline (Figure 1) creates a relationship between GeneChip® probe sets and the current transcript record. The number of mRNA transcripts and Expressed Sequence Tags (ESTs) available in public databases continues to evolve from the original time of design. The NetAffx website maintains a current view of transcripts that GeneChip probe sets interrogate.¹

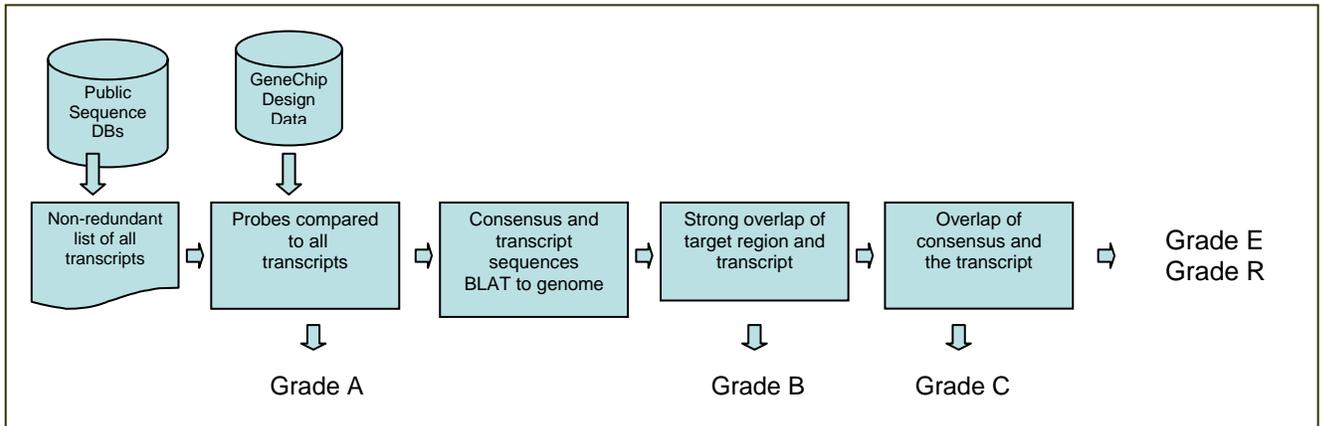


Figure 1. The NetAffx™ Transcript Assignment Pipeline.

In contrast to a spotted cDNA array, the relationship of a GeneChip probe set to a gene transcript is constantly evolving and dynamic because Affymetrix GeneChip Array designs incorporate all available nucleotide data, including Expressed Sequence Tags (ESTs) and mRNA records.² Since ESTs are single reads from fragments of the mRNA, they are useful because they contain untranslated region (UTR) sequence data, which is omitted in many mRNA records. In Vitro Transcription (IVT) GeneChip probe sets, such as U133 and MOE420, detect the ends of the transcripts, such that the target and consensus sequences often consist mostly or entirely of UTR.

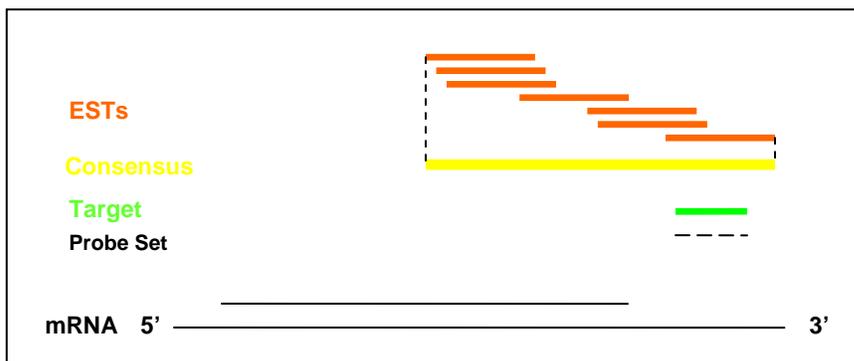


Figure 2. Evolution of a GeneChip® probe set designed exclusively from ESTs.

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

The consensus sequence shown in Figure 2 above is derived from a cluster of EST sequences by a base-calling algorithm. Target sequences are chosen near the 3' end of the consensus/exemplar sequence with boundaries defined as the 5' end of the first probe and the 3' end of the last probe. Subsequent to design, one or more mRNA transcripts become available through further sequencing. Updated information about these transcripts must be derived using bioinformatic data.

Figure 2 also illustrates the discovery process for a probe set designed only with EST data and shows the relationship between the consensus, target and probe sequences. The advantage of including EST data as opposed to only complete mRNA records is that GeneChip probe array data has substantially greater longevity. As the mRNA record matures, data taken in previous experiments remains interpretable for later investigation.

Initially, designs from organisms that have had little mRNA sequencing performed may have a small number of transcript assignments. Over time, mRNA records associated with unassigned probe sets demonstrate that Expressed Sequence Tag (EST) clusters effectively predict the later discovery of new mRNA. Figure 3 below shows how human, mouse and rat EST-only probe set annotations have improved steadily since these arrays were released.

When Human U133 array design was released, 36.2 percent of the 44,199 probe sets were EST-only probe sets. Two years later, only 20.7 percent remain unassigned to a specific mRNA. Both mouse and rat arrays have had less intensive sequencing support efforts than human arrays, and each one currently has approximately eight and ten times more non-EST mRNA entries in GenBank, respectively. Accordingly, these chip arrays have seen a greater improvement since their initial releases (Figure 3). Currently, only 17 percent of EST-only probe set assignments remain unassigned for the Mouse MOE430s, versus 44.2 percent (19,900) unassigned at its release in March 2003. For the Rat 230s, 51.1 percent of the probe sets were EST-based at the end of 2004, versus 81.5 percent (25,301) at its release in March 2003. This trend demonstrates that as sequencing catches up to the EST record, the Rat chip array data will include transcript assignments and cover a substantial majority of rat genes.

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

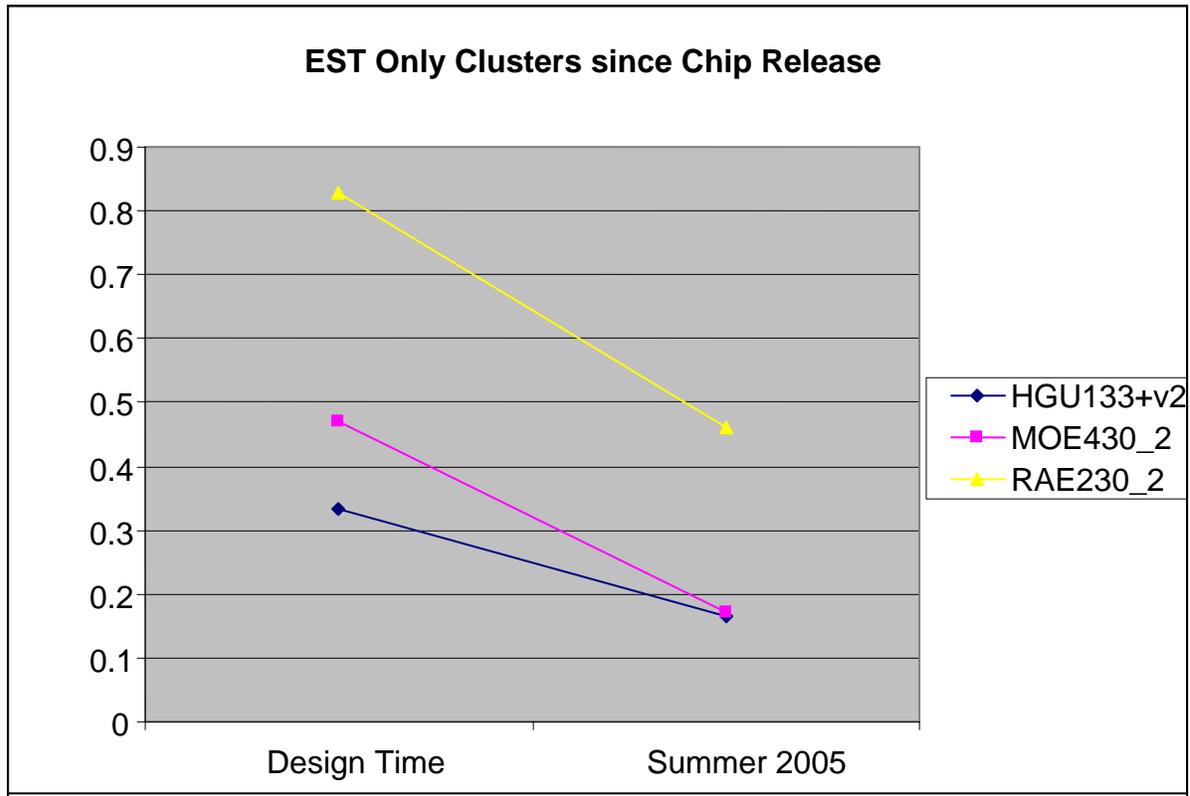


Figure 3. **EST-only Assignments (including both E and R Grades) for the U133_2.0, MOE 420 and RAE 230 arrays as a percentage of total probe sets from the first annotation runs to the most current. The x axis indicates time since release. The release dates for the GeneChip® arrays are January 2002 for U1332.0 and March 2003 for mouse and rat, respectively.**

The NetAffx™ Transcript Assignment Pipeline

In order to provide the most complete transcript assignments with the highest reliability, the NetAffx Transcript Assignment Pipeline (Figure 1) employs a battery of techniques to produce probe set transcript assignments which are graded by quality (Figure 4). Transcript assignment grades fall into five categories that describe the quality of the direct evidence, which supports the assignment. For a given probe set, only the highest grade assignments are kept (i.e., if there is an A assignment, then B, C, E and R assignment results are not presented).

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

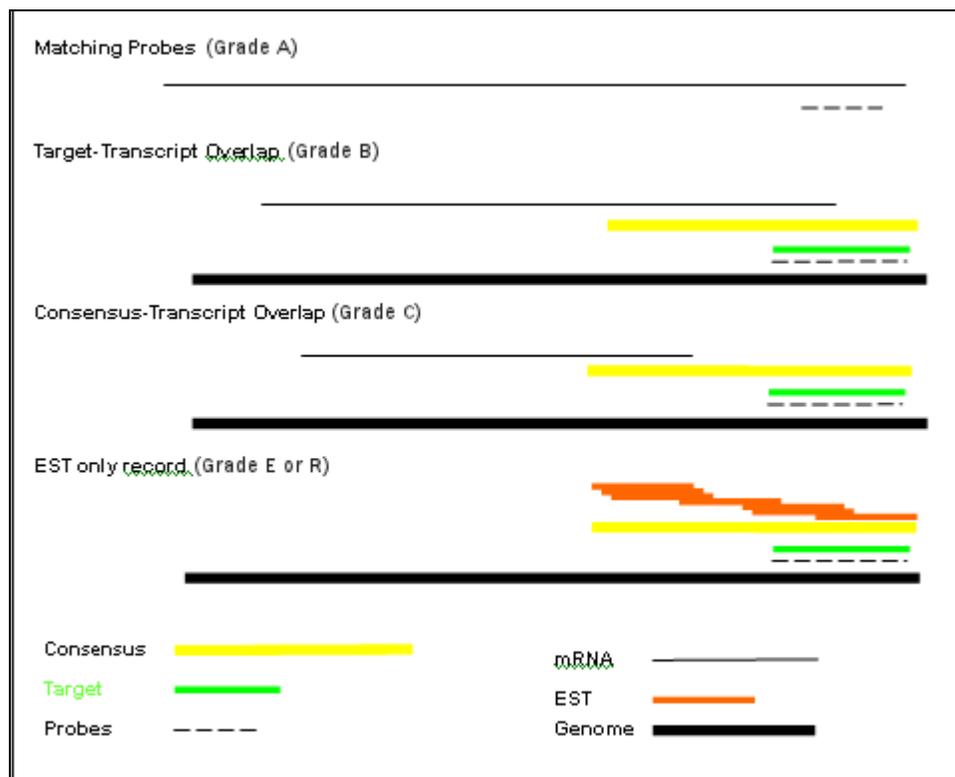


Figure 4. Transcript class assignments reported by NetAffx™.

Matching Probe (Grade A) probe sets have nine or more probes matching a transcript mRNA or gene model sequence. *Genome Target-Transcript Overlap* (Grade B) transcript assignments have a partial overlap between the transcript and the target sequence. *Genome Consensus-Transcript Overlap* (Grade C) transcript assignments result when the transcript sequence overlaps the consensus, but not a significant portion of the target. Overlap transcript assignments must be substantiated by a correspondence to the genome. If no transcript is found, then EST-only data described through a UniGene EST cluster are given a Grade E (EST record) assignment.⁵ If no UniGene EST cluster is built for that probe set, then the representative sequence IDs from the original design record are designated with a Grade R (Representative Sequence) annotation.

Prior to the October 2004 annotation release, transcript-to-LocusLink and transcript-to-UniGene-to-LocusLink mappings, as generated by LocusLink, were used to create many probe set assignments. The resulting associations were confusing in some instances. For example, the *1439963_x_at* probe set from the mouse *MOE430B* array was assigned to the *AK031518* mRNA by the NetAffx™ annotation pipeline, which in turn was associated with the *Ptch1* locus in the LocusLink dataset. However, the transcript sequence primarily spanned the intronic region of the *Ptch1* transcript, as defined by the RefSeq *NM_008957*.

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

There were also situations where a probe set would map to more than one UniGene or LocusLink. In such cases, an attempt was made to resolve the ambiguity by using the mRNA with the maximum number of matching probes.

The current transcript assignment pipeline calculates all transcript assignments directly and makes a record of the evidence used to make the assignment, classifying it by grade (Figure 4).

Non-Redundant Transcript Database

To reduce the number of identical entries found in GenBank, non-redundant data sets are used for all transcript assignments for NetAffx. mRNA sequences are obtained from the appropriate public databases (GenBank, RefSeq, Ensembl, Saccharomyces Genome Database, TIGR, and etc). The mRNA sequences for each organism are clustered, eliminating redundant entries whose BLAT alignment³ overlaps more than 97 percent of their entire length with a sequence identity of 97 percent or more. The longest sequence in each cluster is then used as the representative of that cluster. Preference is given to RefSeq sequences, which are never removed as redundant transcripts, even if two RefSeq sequences appear to be identical to our pipeline. The peptide translation record for each transcript is then kept for protein functional annotation.

The removal of redundant sequences is important because it helps to clean the database of fragmentary and redundant transcripts and provides a more realistic idea of the transcript record. The advantages of this database cleaning are substantial. For example, for the Q2 2005 Annotation Update, GenBank release 145 consisted of 226,514 *Homo sapiens* mRNAs. Clustering at 97 percent sequence identity produced 96,742 clusters containing 100,101 sequences.

Annotation Methods and Classifications

Probe Matched Transcript Assignment (Grade A)

Matching Probe or Grade A assignments represent the best quality transcript assignments. Pairwise alignment of the probe sequences with gene transcripts is the most accurate method to precisely determine the transcript sequences detected by probe sets.⁴ The 25-mer probe sequences are aligned with the non-redundant mRNA set. mRNA sequences that match perfectly with at least 9 probes in a probe set are identified as Grade A assignments. For example, consider probe set *200018_at* from the Human U133 Plus array. This probe set has 11 out of 11 probes matching *BC00672* and 1 cross-hybridization probe against *X04297*.

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

Genome Based Transcript Assignment (Grade B)

If there are no adequate Matching Probe assignments for a probe set, then genomic alignments of the target consensus or exemplar sequence are used. The consensus/exemplar sequences and the non-redundant mRNA sequences for each organism are aligned with the genomic sequence.

If the target region of the consensus/exemplar sequence aligns with the genome and overlaps with the genomic alignment of an mRNA, then the transcript assignment is annotated as *Target-Transcript Overlap (Grade B)*. The rationale here is that several mRNA sequences with incomplete 3' UTR regions may not overlap significantly with the 5' end of a consensus sequence based on the EST data. However, if a consensus and a transcript can be shown to be associated by overlapping alignments to the genome, then the assignment has some significant evidence. Since proximity to the probe sequences defines a Grade B assignment, no overlap thresholds are imposed; even one nucleotide overlap is recorded.

Genome Consensus-Transcript Overlap Assignment (Grade C)

Grade C assignments are also based on genomic alignments, but the target region either does not align with the genomic region or does not overlap with the mRNA-to-genome alignment. This could also indicate a potentially erroneous EST-based extension of the 3' region of the transcript.

Together, Grade A, B, and C annotations relate probe sets to transcript models. The list of transcripts provided is as complete as possible, excluding duplicates, and includes gene predictions and alternative transcripts. True splicing products can be distinguished as related to the probe by viewing them in a genome browser such as the Integrated Genome Browser (IGB)⁶ or the UCSC Genome Browser.⁷

EST-Only Probe Set Annotations (Grades E and R)

If no current mRNA sequence or other gene model can be attributed to a probe set, NetAffx users are provided with the EST-related information used to define the probe set at the time of design. NetAffx annotation pipelines do not update EST data assignments; they extract them from the design files. Two grades of EST-only probe sets are currently supported.

The *EST-Only Record (Grade E)* assignments occur when a UniGene EST-only cluster is known to relate to the probe set. UniGene BLASTX functional annotations are provided with these clusters. Annotations based upon nucleotide-to-amino acid sequence comparisons should be used cautiously.

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

Sometimes mRNA sequences may be included in the UniGene cluster for an E or R class assignment if the mRNA sequence has been retired or retracted from the public databases.

The *Representative Sequences Only (Grade R)* assignments are given to probe sets which only have raw EST evidence associated with them. For arrays with mature UniGene EST sequencing efforts, there will be few or no R probe sets. For convenience, links to the individual ESTs (Group Members) that produced this probe set at date of design are given in the annotation description and through a link on the NetAffx detail pages for Expression Array probe sets.

Cross-Hybridized Probe Sets

When NetAffx assigns a transcript to a probe set, the Annotation Description field on the details page lists other known transcripts to which a probe matches. The Annotation Notes field gives an estimate of the amount of cross-hybridization the probe set may see from other transcripts.

Antisense Probe Sets

When some consensus sequences are compiled and the orientation of the cluster sequences are ambiguous, probe sets are designed going in either direction off the same consensus. If one of these probe sets has no transcript assigned to it, it is designated as an Antisense Probe Set and linked to its sense partner in the annotation description.

Overview of NetAffx™ Transcript Assignments

A more current, detailed view of how transcript assignments mature on Human, Mouse, and Rat Arrays is given in Table 1 below. The majority of probe set transcript assignments are *Matching Probe (Grade A)* assignments, based upon a direct identity between the probe and transcript sequences. The transcript sequencing effort for all these organisms is maturing rapidly, even over a single quarter with an expansion of transcript assignments (Grades A, B or C) ranging from 1 to 7 percent. For human arrays, which are supported for a relatively mature transcript record, the probe sets, which have an actual transcript associated with them (Grades A, B, and C), increase 1 percent over a period of three months, from March to June 2005. See Table 1 below. The Rat array, whose transcript record is still maturing rapidly, can improve by 7 percent in a single quarter.

The genome-based Grade B and C assignments give an additional 7 to 15 percent of coverage. Table 1 below demonstrates that the quality of the transcript record evolves as the numbers of B and C assignments gradually diminishes and more transcript records with UTR are found in the public records. The remaining

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

probe sets, substantiated by EST sequence data alone (Grades E and R), diminish over time, and it appears that IVT GeneChip arrays will eventually migrate towards 85 percent or more mRNA transcript assignments.

Assignment Classification	HG133Plus Probe Sets (%)		Mouse430v2 Probe Sets (%)		Rat230v2 Probe Sets (%)	
	Q1 2005	Q2 2005	Q1 2005	Q2 2005	Q1 2005	Q2 2005
Matching Probes (A)	39,391 (72%)	41,797 (77%)	32,895 (73%)	33,363 (75%)	10,219 (33 %)	13,034 (42 %)
Target/Transcript Overlap (B)	2807 (5.2%)	2691 (5.0%)	3472 (7.7%)	2901 (6.4%)	3725 (12%)	2813 (9.1%)
Consensus/Transcript Overlap (C)	1298 (2.4%)	1105 (2.2%)	1011 (2.2%)	794 (1.8%)	1228 (4.0%)	934 (3.0%)
EST-Only Probe Set (E)	11,117 (20%)	8171 (15%)	7659 (17%)	6309 (14%)	15,870 (51%)	11,243 (36 %)
Representative Sequence Probe Set (R)	*	849 (1.6%)	*	1397 (3.1%)	*	3018 (9.7 %)

Table 1. Transcript assignment grade and coverage for three major array products (Dec 2004 release). The number of cross-hybridizing probes is a measure of the uniqueness of the probes on the array. * After Q1 2005, E-Grade assignments were subdivided into E- and R-Grade assignments.

Transcript Assignment for NetAffx™ Annotations

Revision Date: 2006-3-24

Revision Version: 2.3

References

1. Liu, G., Loraine, A.E., Shigeta, R., *et al.* (2003), 'NetAffx: Affymetrix probe sets and annotations', *Nucleic Acids Research* Vol. 31, pp. 82-86.
2. Affymetrix Technical Note, (2001), 'Array Design for the GeneChip Human Genome U133 Set', http://www.affymetrix.com/support/technical/technotes/hgu133_design_technote.pdf
3. Kent, W.J., (2002), 'BLAT-the BLAST-like alignment tool', *Genome Res.* Vol. 12(4), pp. 656-664.
4. Chalifa-Caspi, V., Yanai, I., Ophir, R., *et al.* (2004), 'GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes', *Bioinformatics* Vol. 20(9), pp. 1457-1458.
5. Pontius, J.U., Wagner, L., and Schuler, G.D., (2003), 'UniGene: A Unified View of the Transcriptome', In: *The NCBI Handbook*. Bethesda (MD): National Center for Biotechnology Information No. 21. See <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=UniGene>.
6. The Integrated Genome Browser homepage is: <http://genoviz.sourceforge.net/>. The IGB User's Guide is available at https://www.affymetrix.com/support/developer/tools/IGB_User_Guide.pdf
7. Kent, W.J., Sugnet, C.W., Furey, T.S., *et al.* (2002), 'The human genome browser at UCSC', *Genome Res.* Vol. 12(6), pp. 996-1006. See <http://genome.ucsc.edu>.