Transcript Assignment for Affymetrix GeneChip® probe arrays

The NetAffx™ Analysis Center maintains a current view of transcripts that GeneChip® microarray probe sets interrogate.[1] The challenges of maintaining this evolving sequence and annotation database include immature and fragmented mRNA records, inherited sequence errors from data used to design the arrays, and errors in mRNA and EST sequences. The NetAffx transcript assignment pipeline employs state-of-the-art bioinformatics and assures the highest quality data possible.

To understand the methods described herein, it is necessary to lay out some of the inherent caveats of how GeneChip® expression probe arrays are designed (fully described in reference 2).

A typical problematic probe set is illustrated in Figure 1 below. Evidence of transcription (*exemplar sequences*) for a gene is exclusively EST data. This happens often when sequencing projects are new for an organism. The exemplars are clustered together by similarity, combined by a clustering algorithm, and abstracted into a single *consensus sequence*. From a subsequence of the consensus, the *target sequence* is extracted from which probes of the *probe set* are selected.

Since ESTs have lower sequence accuracy, resulting probe sequences will often show as erroneous when compared to the genome sequence and transcript record over time.

The transcript record, mRNA and EST evidence, is constantly changing in databases around the world. mRNA sequences, which may appear later, are usually truncated in the 5' end so that EST-based consensus sequences will overlap poorly with subsequent mRNA sequences that they correspond to *in vivo* (top of Figure 1). This can lead to omissions in transcript assignment when probe set matching to mRNA is used exclusively.

To this end, the NetAffx Analysis Center employs a tiered assignment protocol. The NetAffx transcript assignment pipeline delivers the broadest assignment of mRNA transcript to probe set matching with the best reliability available.

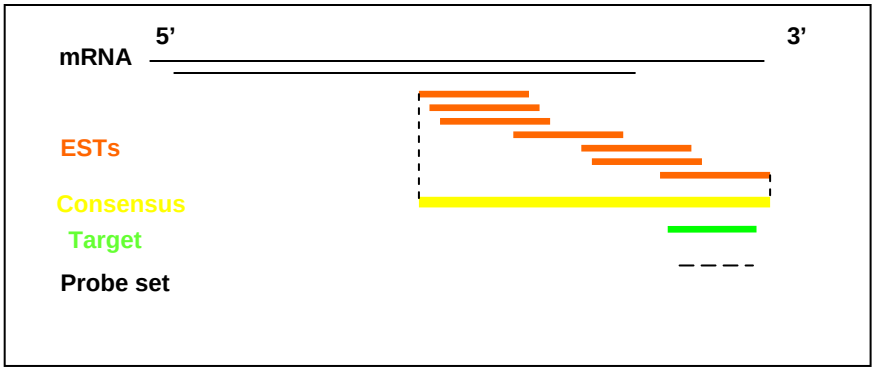

Figure 1: **A problematic GeneChip® microarray probe set.** *Exemplar* sequences are EST and mRNA that are directly represented on the GeneChip array. *Consensus* sequences are built by the base calling algorithm derived base-by-base from a cluster of public mRNA and/or EST sequences. In this case only ESTs define this consensus. *Target* sequences are chosen near the 3' boundary of the consensus/exemplar sequence with boundaries from the 5' end of the first probe to the 3' end of the last probe. Each transcript is measured by 11 to 20 probe sequences of 25-nt length collectively referred to as *probe set*.

METHODS

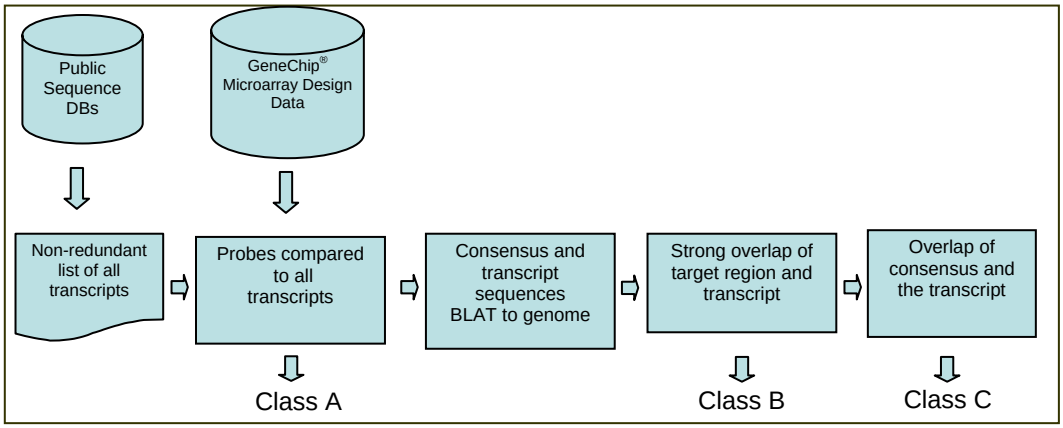

Figure 2: **Diagram of transcript assignment in the NetAffx™ Analysis Center.**

Non-redundant transcript database. mRNA sequences are obtained from the appropriate public databases (GenBank®, RefSeq, Ensembl, Saccharomyces Genome Database, TIGR, etc.). The mRNA sequences for each organism are clustered at 90 percent sequence identity using BLAT.[3] The longest sequence in each cluster is then used as the representative of that cluster, with preference given to RefSeq sequences. This non-redundant data set is the nucleotide record used for all transcript assignments for the NetAffx Analysis Center. The peptide translation record for each transcript is also kept for protein annotation.

E.g., GenBank release 142 has 135,632 mRNAs for *Homo sapiens*. Clustering at 90 percent sequence identity produced 61,950 clusters.

Probe matched transcript assignment. Pair-wise alignment of the probe sequences with gene transcripts is the most accurate method to precisely determine the transcript sequences detected by probe sets.[4] All 25-mer probe sequences are aligned with the non-redundant mRNA set. mRNA sequences that match perfectly with at least nine probes in a probe set are identified. These are referred to as "Matching Probe" or "Class A" assignments and represent the best quality assignments.

There are other relationships between a probe sequence and a transcript. If an mRNA sequence is found to match less than nine probes perfectly in a probe set, it is recorded as a "cross-hyb" probe set. If the proper orientation of the consensus sequence constructed from mRNA and/or EST data is unknown at the time of design, probes are tiled against both the strands of the consensus sequence to ensure that the true transcript is represented on the array. If probes align with the negative strand of the mRNA then the corresponding probe set is annotated as "negative strand" probe set.

For example, consider probe set 200018_at from the GeneChip Human Genome U133 Plus 2.0 Array (HG-U133 Plus 2.0). This probe set has 11/11 matching probes for BC00672 and 1 cross-hyb probe against X04297.

Probe set 1552279_a_at from the HG-U133 Plus 2.0 Array has 10/11 probes matching the sense strand of transcript AL832613 and 4/11 "negative strand" probes matching the antisense strand of RefSeq NM_015077.

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

"Genome Consensus/Exemplar Overlap" (Class C) assignment is also based on genomic alignments but the target region either does not align with the genomic region or does not overlap with the mRNA -> genome alignment. These may also indicate a potentially erroneous EST-based extension of the 3' region of the transcript.

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 "Genome Target Overlap" (Class B). There are no thresholds imposed in this process. Therefore, even if one nucleotide of the target sequence overlaps with the mRNA alignment it is recorded as "Genome Target Overlap." The rationale here is that several mRNA sequences with incomplete 3' UTR sequence may not overlap significantly with the 5' of a consensus sequence based on the EST data, but if their placement can be verified by placement on the genome, then the assignment has some significant evidence.

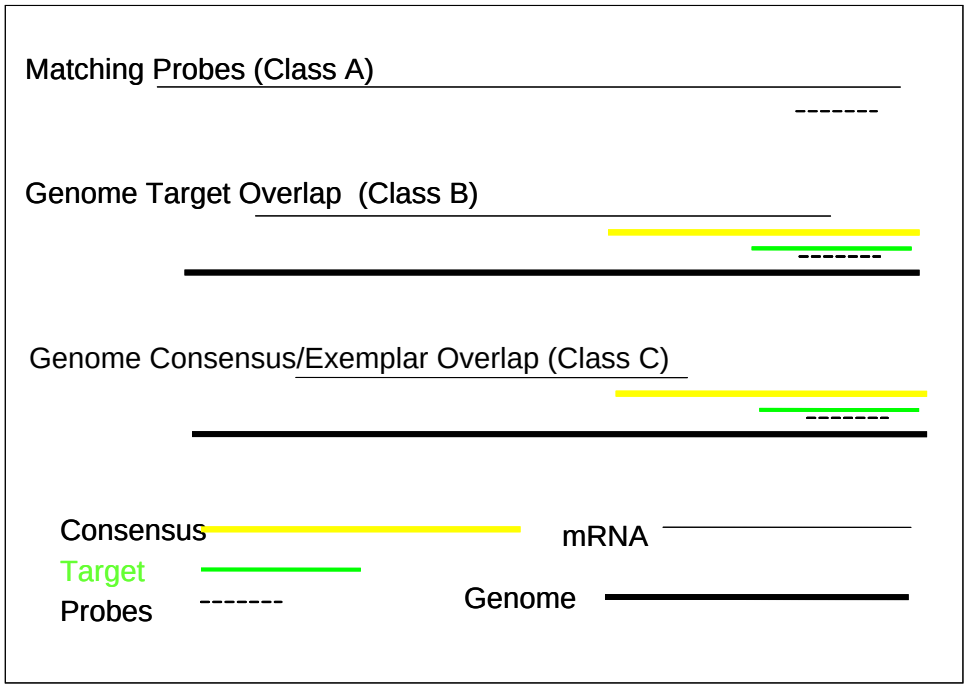

Figure 3: **Transcript assignments reported by the NetAffx™ Analysis Center**. *Matching Probe* (Class A) probe sets have nine or more probes matching the transcript sequence. *Genome Target Overlap* (Class B) transcript assignments have a partial overlap between the transcript and target sequence. *Genome Consensus/Exemplar Overlap* (Class 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.

OVERVIEW

As can be seen in Table 1, the more mature array designs have yielded very strong correlation to the public transcript record. Over 70 percent of probe sets are shown as "Matching Probes" for the latest Human and Mouse designs while the more recent Rat records show a lower number of Class A transcript assignments.

Negative strand errors on the arrays are consistently at about 12 percent of the total record across the designs.

It is important to note that the transcript-to-LocusLink and transcript-to-UniGene-to-LocusLink mappings are provided by the corresponding databases and are not computed by Affymetrix. These associations may be confusing in some instances. Consider the probe set 1439963_x_at from the Mouse 430B Array. This probe set is assigned to AK031518 mRNA by the NetAffx annotation pipeline. AK031518 is associated with the Ptch1 locus in the LocusLink data set. However, the transcript sequence mostly spans the intronic region of the Ptch1 transcript as defined by the RefSeq NM_008957. There are also situations where a probe set maps to more than one UniGene or LocusLink ID. In such cases, an attempt is made to resolve the ambiguity by using the mRNA with the maximum number of matching probes.
For users who only want the most substantiated probes, we recommend the use of the NetAffx Analysis Center's search facilities to include only probe sets which have Class A evidence.

Array name	Transcript assignment type	Probes assigned	% of total
HG-U133_Plus_2	Matching Probes (A)	40044	73.24
	Genome Target Overlap (B)	2727	4.99
	Genome Consensus/Exemplar Overlap(C)	1153	2.11
	Cross Hyb Matching Probes	16452	30.09
	Negative Strand Matching Probes	8521	15.58
	Total Number of Probes	54675	100.00
Mouse430_2	Matching Probes (A)	33182	73.57
	Genome Target Overlap (B)	2842	6.30
	Genome Consensus/Exemplar Overlap(C)	883	1.96
	Cross Hyb Matching Probes	14188	31.46
	Negative Strand Matching Probes	5074	11.25
	Total Number of Probes	45101	100.00
Rat230_2	Matching Probes (A)	10045	32.30
	Genome Target Overlap (B)	3796	12.21
	Genome Consensus/Exemplar Overlap (C)	1248	4.01
	Cross Hyb Matching Probes	8877	28.54
	Negative Strand Matching Probes	2134	6.86
	Total Number of Probes	31099	100.00

Table 1: Transcript assignment overview from Oct 2003 for three major arrays

REFERENCES

1. Liu G, *et al*. NetAffx: Affymetrix probe sets and annotations. *Nucleic Acids Research* **31,** 82-86 (2003).
2. Array Design for the GeneChip® Human Genome U133 Set. Affymetrix Technical Note. http://www.affymetrix.com/support/technical/technotes/hgu133_design_technote.pdf (2001)
3. Kent, W.J BLAT-the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
4. Chalifa-Caspi, *et al.* GeneAnnot: comprehensive two-way linking between oligonucleotide array probe sets and GeneCards genes. *Bioinformatics* **9**, 1457-1458 (2004).