Exon Probeset Annotations and Transcript Cluster Groupings

I. Introduction

This whitepaper covers the procedure used to group and annotate probesets. Appropriate grouping of probesets into transcript clusters and subsequent filtering of probesets within the transcript cluster plays a critical role in generating gene-level signal estimates. (See the Gene Signal Estimates from Exon Arrays whitepaper for more information on the implications of probeset groupings on gene level signal estimates.) The annotations and groupings provided are intended to be a baseline of information for each exon array. How the information is applied will depend on the specific goals for each experiment.

Before reading further, note that this document is intended as a supplement to the appropriate exon array design technote (i.e. GeneChip Human Exon 1.0 ST Array Design Technote). It is assumed that the reader is familiar with the key design points and terminology described in that paper.

The GeneChip Human Exon 1.0 ST Array (and other exon arrays) are more exploratory than predecessor expression arrays such as the HG-U133 2.0 Plus. Previous array designs used a cDNA assembly consensus or exemplar approach where one or a few probeset were associated with a specific gene. For the exon arrays the design was exon focused rather than gene or transcript focused. As a result there are no intrinsic transcript or gene entities in the exon array designs. Instead there are only probesets associated with exons or contiguous parts of exons. Groupings of exon probesets into transcripts and genes is now a dynamic post-design process.

As new genome assemblies and annotations are released there is an opportunity to generated improved transcript cluster groupings. The meta-probeset lists available in the support files for the exon arrays are one such example of these groupings. Meta-probeset lists from different versions of the genome are likely to have different groupings of exon probesets into transcript cluster. Furthermore these different groupings may in turn have subtle (or even substantial) impacts on gene level analysis.

The basic approach used to generate transcript cluster groupings for a particular genome is to (1) construct gene annotations on the genome using a variety of annotation sources merged using a set of rules, (2) map exon probesets to gene annotations using the genome, and (3) use the target genome to group the probesets that mapped to the gene annotations. A fall out from this is that gene annotations are easily created from the transcript annotations used to group the probesets. The gene grouping process and probeset annotation process are described in more detail below.

The result of the entire process is a collection of “Design Annotation Files”. These design annotation files are available from the appropriate exon array support page on http://www.affymetrix.com. Specifically, for each genome version a set of master GFF files are generated. These files are used to populate parts of the NetAffx content and are used to generate the other files including:

- design level probeset annotation CSV file
- meta-probeset list files
- IGB binary annotation (BGN) files

II. Defining Gene Annotations

The following five steps outline our approach to grouping probesets.

Cluster transcript annotations on the same strand of the target genome using a set of rules involving exon overlap and splice site sharing.
Label each transcript cluster as a different gene.
Join exons of clustered transcript annotations on the target genome to determine gene structure.
Map each probeset to a single gene, based on whether it falls within the gene’s annotated exon boundaries on the target genome.
Group probesets together that map to the same gene as a transcript cluster.

There are situations where simply clustering transcripts based on exon-exon overlap can produce gene annotations that actually represent multiple genes. Sometimes this is probably due to two genes actually sharing some transcribed region on the genome (i.e. 3’ UTR of one gene overlapping the 5’ UTR of a downstream gene on the same strand). In other cases this is due to erroneous cDNA sequences, alignment algorithms, or gene predictions. There was an effort to avoid such over-clustering of transcripts, even at the expense of fragmenting some gene annotations. This decision was motivated by the fact that falsely joined exons (from more than one gene) would generate results which at first pass would appear to be alternative splicing.

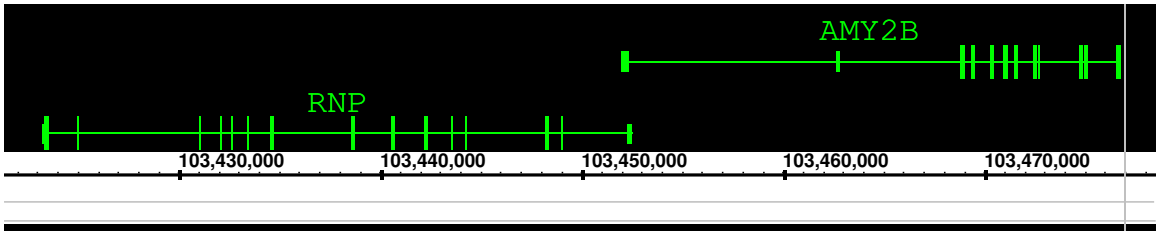

Figure 1: RefSeq transcript annotations for two different genes on chr1 (NCBI build 34) have overlapping exons.

Gene annotations were constructed from clusters of transcript annotations. Because of the wide variety of transcript annotations sources, it was desirable to find a way to qualitatively categorize the different types of sources. For example, RefSeq sequences are manually curated and generally trusted as accurate

isoforms of genes. While ESTs and partial mRNAs are not manually curated, they cover greater portions of the target genomes and are supported by the fact that they are derived from cloned sequence. And although *ab-initio* gene predictions are not inherently supported by clone sequence, they may be reasonable predictors of transcript structures. Acknowledging that different annotation sources have different levels of confidence, an iterative approach was implemented to ensure that high confident annotations were not merged by less confident annotations.

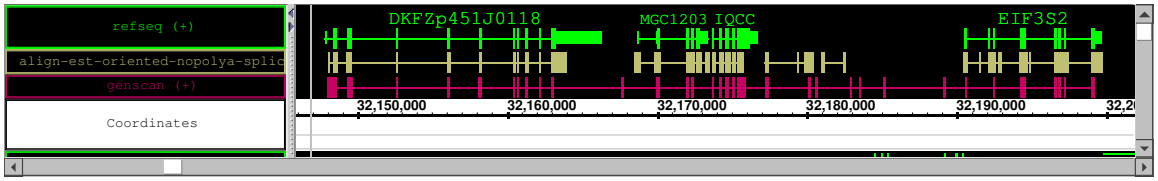

Figure 2: RefSeq transcript annotations on chr1 (NCBI build 34) are shown on top of spliced EST alignment annotations and a Genscan prediction.

For the purposes of establishing a hierarchy of gene confidence levels, we partitioned the sources of input transcript annotations into three types. From highest to lowest confidence, the types were labeled *core*, *extended*, and *full.* Broadly defined, the core type consisted of (BLAT) alignments of mRNA with annotated full-length CDS regions, the extended type consisted of cDNA alignments and annotations based on cDNA alignments, and the full type consisted of sets of *ab-initio* gene predictions.

For the GeneChip Human Exon 1.0 ST Array, the annotation sources for each level are:

- Core Gene Annotation sources
- RefSeq alignments
- Genbank alignments of 'complete CDS' transcripts
- Extended Gene Annotation sources
- cDNA alignments
- Ensembl annotations (Hubbard, T. et al.)
- Mapped syntenic mRNA from rat and mouse
- microRNA annotations
- Mitomap annotations
- Vegagene (The HAVANA group, Hillier et al., Heilig et al.)
- VegaPseudogene (The HAVANA group, Hillier et al., Heilig et al.)
- Full Gene Annotations
- Geneid (Grup de Recerca en Informàtica Biomèdica)
- Genscan (Burge, C. et al.)
- GENSCAN Suboptimal (Burge, C. et al.)
- Exoniphy (Siepel et al.)
- RNAgene (Sean Eddy Lab)
- SgpGene (Grup de Recerca en Informàtica Biomèdica)

- TWINSCAN (Korf, I. et al.)

The core type was so named because the annotations in this type were intended to be the foundation from which we built our gene annotations. The extended type derived its name from the sense that these annotations would extend the boundaries of the core genes. The idea behind the name of the full type was that it would signify all possible content.

The transcript annotation clustering procedure utilized the confidence rankings of the transcripts in such a way that non-overlapping transcripts from a higher rank could not be joined together by a transcript annotation from a lower rank. For example, annotations of EST alignments would not be used as evidence to join two separate RefSeq alignment annotations – however, a third RefSeq alignment annotation could. Lower ranking annotations could only be added to the transcript clusters established at a higher ranking, or they would establish new clusters without any higher ranking content. Here is the algorithm for clustering the transcript annotations.

1) Core annotations are merged first.
    a. Spliced core annotations are merged first.
        i. If two core annotations share a splice site, they belong to the same cluster.
        ii. If two core annotations overlap with two different exons in each transcript, they belong to the same cluster.
        iii. If a core annotation lies within the boundaries of exactly one other core exon, the two belong to the same cluster.
    b. Single exon core annotations are added next.
        i. If a single exon core annotation overlaps an exon of exactly one spliced core cluster, it is added to that cluster. (However, the single exon is not used to determine exon overlap for subsequent single exon core annotations – i.e. prevent “chaining” single exons).
        ii. If a single exon core annotation lies within the boundaries of exactly one other core exon, the two belong to the same cluster.
        iii. Single exon core annotations that don’t overlap any spliced core annotations are merged with each other based on overlap.
        iv. If a cluster of single exon core annotations formed in step (iii) overlaps with exactly one other spliced core cluster (with the single exons from step (i) now included), it is added to the spliced cluster (i.e. now allow chaining if it is completely unambiguous)
2) Extended annotations are added next, in a similar manner as the single exon core annotations.
    c. If an extended annotation exon overlaps the exon of exactly one core cluster or shares a splice site with exactly one core cluster, it is added to that cluster. (However, the added extended annotation is not used to determine exon overlap for subsequent extended annotations.)
    d. If an extended annotation falls entirely within the bounds of exactly one core exon, the extended annotation is added to that core cluster.

e. If a spliced extended annotation overlaps the exons of more than one core cluster and does not share a splice site with exactly one of them, the extended annotation is broken up into its underlying exons. These underlying exons are then treated as separate single exon transcripts for the purposes of merging.
f. Extended annotations that don't overlap any core clusters are merged with each other.
g. If a cluster of extended annotations formed in step (d) overlaps exactly one other core cluster (with the extended annotations from step (a) now included), it is added to that core cluster.

3) Full annotations are added last, in exactly the same manner as the extended annotations, except that extended clusters are treated the same way as core clusters.

The following figures illustrate hypothetical examples of how some of the rules for clustering transcript annotations are applied.

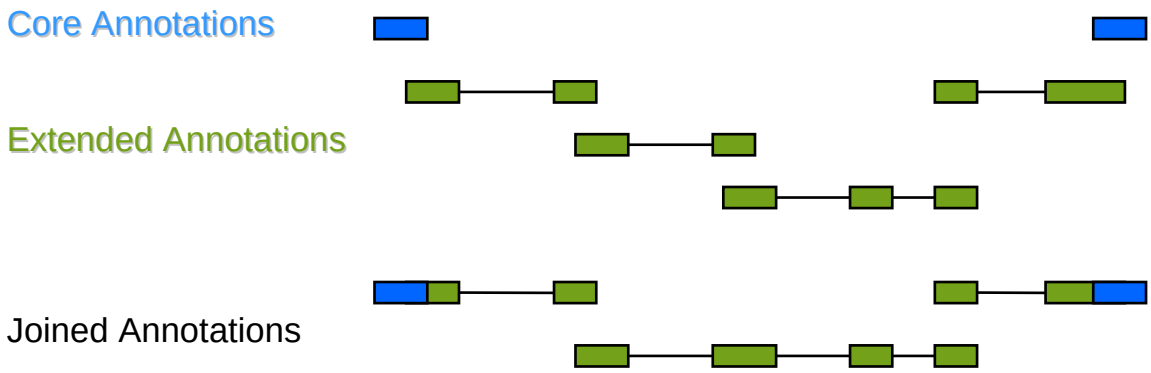

Figure 3: (Rule 2e) Extended annotations that do not directly overlap any core annotations are not added to those associated clusters. Instead they are merged to form a cluster of their own.

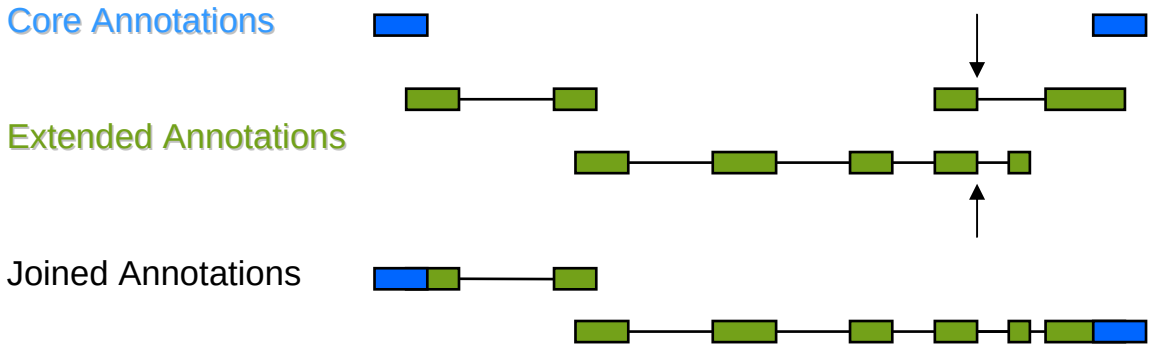

Figure 4: (Rule 2a) The arrows indicate a shared splice site between the two extended annotations. The bottommost extended annotation is joined with the cluster that it shares a splice site with.

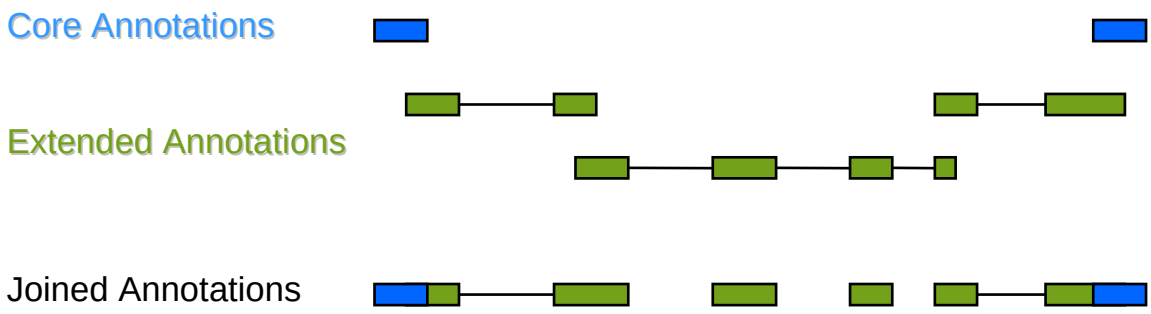

Figure 5: (Rule 2c) The extended annotation is broken up into underlying exons because it overlaps two gene clusters from a higher confidence level.

The resulting transcript clusters define a set of exon boundaries, which in turn determine the gene definitions used for probeset clustering. Each gene annotation is constructed from transcript annotations from one or more confidence levels. Some parts of a gene annotation may derive from high confidence core annotations, while other parts derive from the lower confidence extended or full annotations. Therefore, different parts of the gene annotation can be labeled according to the highest confidence level of transcript annotation that supports that part. This labeling of the resulting gene annotations according to confidence level was used to further annotate the probesets that mapped to the gene.

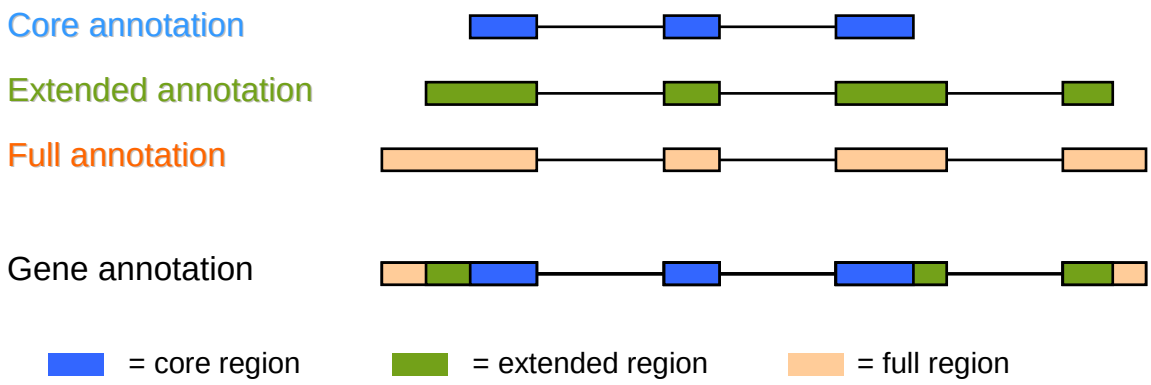

Figure 6: Transcript annotations from different confidence levels are merged to form a gene annotation. The regions of the gene annotation can be labeled according to the highest level confidence transcript that supports that region.

III. Probeset Gene Mapping and Probeset Grouping

Probesets were grouped together if they mapped to the same gene annotation. Generally, a probeset was said to map to a gene annotation if it fell inside the

exon of that gene. However there were cases that were not so straight-forward, and so a set of specific rules was established.

4) If a Probe Selection Region (PSR) overlaps exactly one gene, then the associated probeset is mapped to that gene (even if the probeset does not overlap any genes).
5) If a probeset falls entirely within the exon bounds of exactly one gene, then it is mapped to that gene.
6) If a probeset falls entirely within the exons of multiple genes, the probeset is mapped to its own unique transcript.
    - Exception: if the probeset falls within the exons of multiple genes, but within the core region of only one gene, then it is mapped to the gene with the core region.
7) If a PSR or probeset does not overlap any genes, the probeset is mapped to its own unique transcript.

There were some cases where probesets were selected from Probe Selection Regions (PSRs) that were based on content that was removed after the design. Rule 1 attempted to associate the probesets from these PSRs with the genes that they were proximal to.

Cases where probesets fell within exons from multiple genes were covered by Rule 3, however there were special cases involving core genes. If a probeset fell within the exons of multiple genes, but only one of them was a core gene, then the probeset was mapped to the core gene. The motivation behind this rule was that since core genes were regarded with the highest confidence, a probeset that fell within core exon should be mapped to that gene, despite the lower confidence content.

IV. Probeset Annotation

There were three types of annotation that were applied after the probesets had been grouped:

1. confidence ranking
2. whether the probeset was ‘bounded’ to a transcript cluster (less confident association) or actually contained (high confident association)
3. whether the probeset was within a high quality CDS region.

While all probesets were given a confidence ranking, the ‘bounded’ and ‘CDS’ flags were applied to only some probesets.

IV.A. Probeset Confidence Ranking

The annotations used to generate the gene bound definitions were further used to rank the probesets with respect to their confidence level. As mentioned above, the gene annotations used to group the probesets had core, extended, or full regions according to the confidence level of the transcript annotations used to

compose the gene. A probeset with a position in the gene that was within a core region of the gene was given a 'core' level ranking. Likewise for probesets that fell within the extended and full regions of their associated genes. In the cases where a probeset crossed level boundaries, the probeset was labeled with the lower confidence rating. If a probeset did not overlap any genes, it was labeled as 'free'.

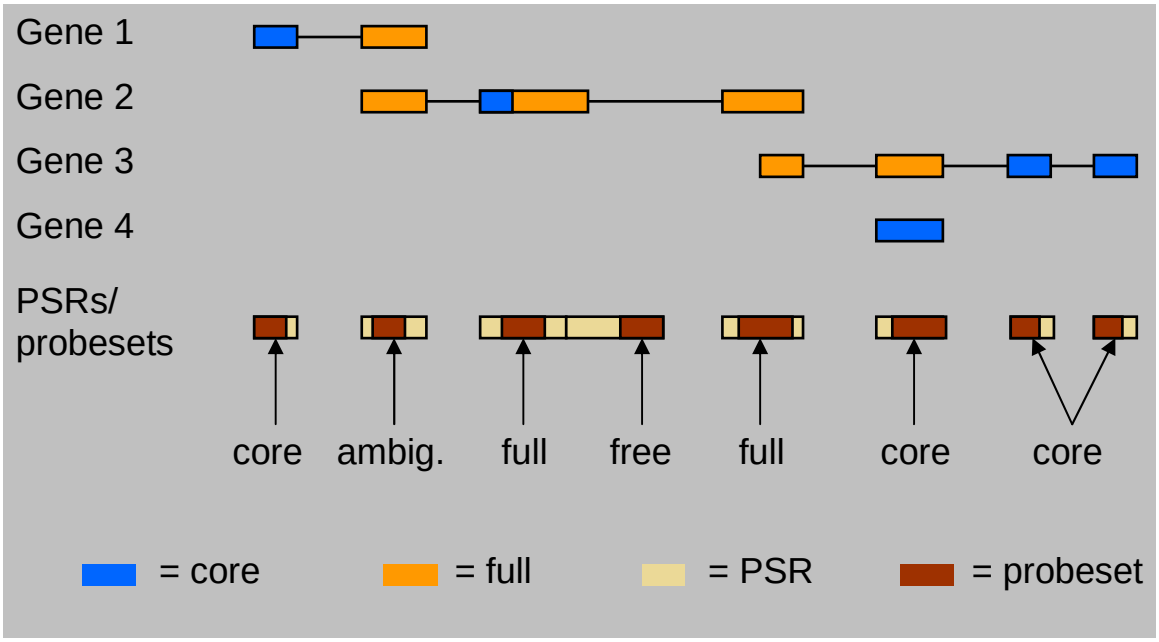

Figure 7: **Probesets are labeled with a confidence ranking according to the confidence level region of the overlapping gene. Probesets that fall within multiple genes are labeled 'ambiguous', unless the probeset falls within exactly one core region of a gene; then it is labeled 'core'. Probesets that overlap confidence region boundaries are labeled with the lower confidence level. Probesets that do not fall within any genes are labeled 'free'.**

There were other instances where a probeset fit inside the exon of more than one gene annotation, and no determination could be made as to which gene annotation it belonged to. In these cases, these probeset was labeled 'ambiguous' and placed singly in its own gene annotation. The exception to this rule was if a probeset mapped to the exons of multiple genes, but fell within the core region of only one gene. Then the probeset was given the core level annotation instead of 'ambiguous'.

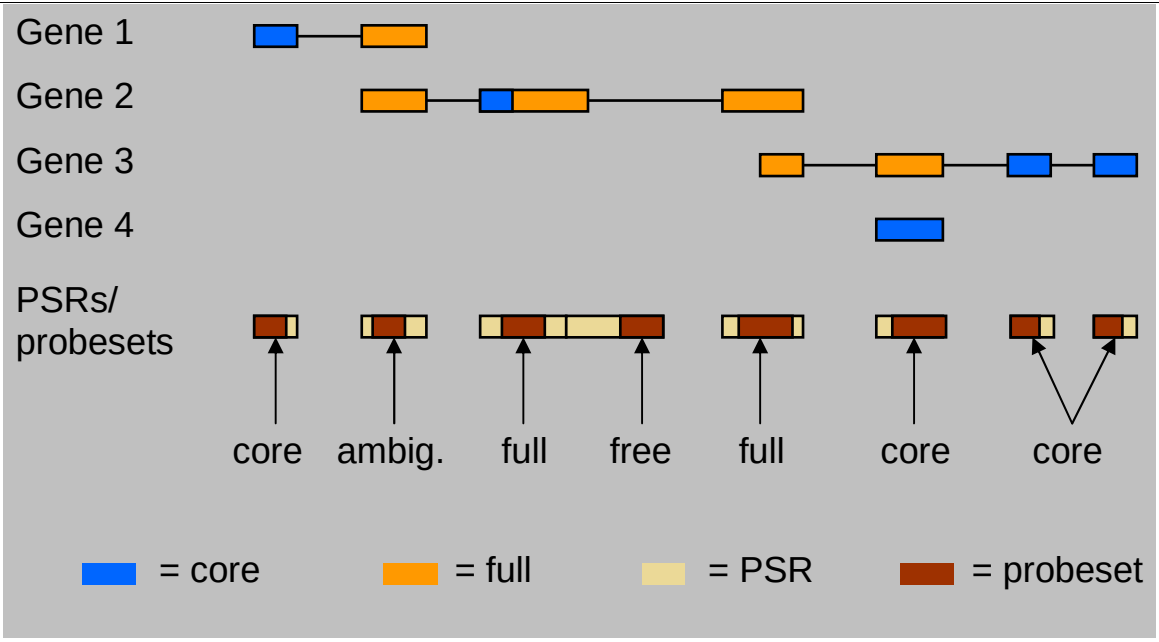

Figure 8: **Probesets are labeled with a confidence ranking according to the confidence level region of the overlapping gene. Probesets that fall within multiple genes are labeled 'ambiguous', unless the probeset falls within exactly one core region of a gene; then it is labeled 'core'. Probesets that overlap confidence region boundaries are labeled with the lower confidence level. Probesets that do not fall within any genes are labeled 'free'.**

IV.B. Bounded Probesets

Many of the genes generated by the transcript clustering procedure were actually single exon genes defined by either a solo GENSCAN Suboptimal exon annotation or a single exon EST alignment. These annotations occurred at relatively high frequency throughout the greater transcribed regions of the target genome. Many of these annotations occurred within the introns of larger, spliced, higher confidence genes. It was decided to include these probesets with the transcript clusters they fall in, but to provide annotations to the effect that the grouping confidence is lower. For example, this allows researchers to easily include or exclude these single exon probesets for alternative transcript analysis with the same locus.

Therefore, if a probeset mapped to a single exon extended or full gene that lay within the intron of exactly one other gene, then the probeset would be "regrouped" as part of the larger, spliced gene and given the additional label, 'bounded'. The intuition behind the word 'bounded' is that the probeset is bounded by the larger spliced gene.

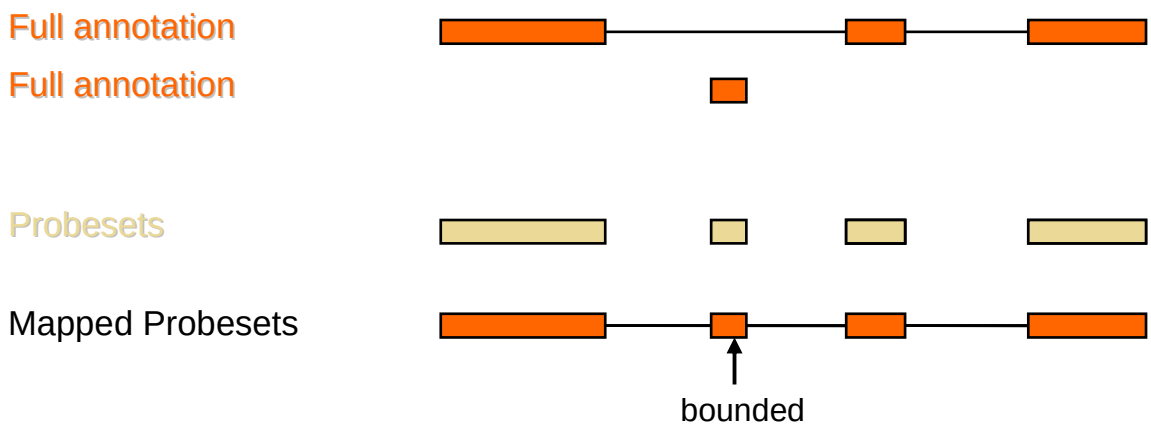

Figure 9: **Probesets mapping to non-core single exon annotations that fall within the intron of one other gene are grouped with that gene and given the label 'bounded'.**

IV.C. CDS Annotation

Probesets that fell completely within the CDS region of at least one transcript from one of the following sources were flagged as 'cds' probesets.

- RefSeq alignments
- Genbank alignments 'complete CDS' transcripts
- Ensembl annotations
- VegaGene annotations

This allows research to quickly focus down to a set of probesets against likely protein coding regions.

V. Probeset Evidence

The probesets were further characterized by listing the quantity and source of annotations that could be interrogated by the probeset. The same transcript annotations used as building blocks for gene definitions were also used to describe the evidence for the probesets (although, in general any annotation set can be used here). As with the other labeling procedures, the entire probeset had to be bounded by the exons of the annotation in order for it to be listed as evidence. This technique for listing evidence gave rise to some non-intuitive situations where a probeset could have been given a confidence level of 'core', 'extended' or 'full', but not have any evidence. This is due to the fact that the transcript annotations were merged into larger gene annotations when determining the confidence levels, but were not merged when determining evidence for the probesets. Therefore, a probeset could be mapped to an exon that was derived from a composition of several smaller annotations, yet the probeset did not fall within the bounds of any one of these smaller annotations.

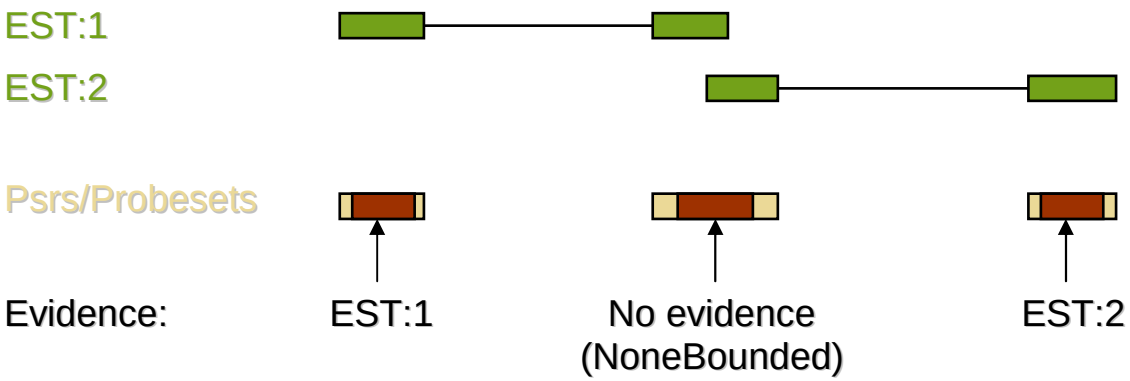

Figure 10: **An annotation is evidence for a probeset if the probeset lies completely within one of the exons of the annotation.**

The mapping of exon array probesets to specific annotations is contained within the GFF file as comments. These GFF comments can be easily parsed should this level of information be needed.

VI. Discussion

These probeset groupings and annotations are meant to provide a starting point for researchers and are not intended as the final word on how probesets should be clustered and analyzed. While the notion of transcript clusters loosely equals a gene, it should be noted that we make no attempt to merge non-overlapping transcript clusters for the same gene (based on paired EST reads for instance).