Annotation of Codelink probes with transcript information.

Summary:

The Codelink platform has 33k probes. Codelink probe annotations are available from sources like the rwgcod package in BioConductor. However, the method by which these annotations are maintained is not very clear. Most annotations methods incrementally update an annotation by replacing obsolete Entrez Ids by their replacements and do not carry out a full sequence based annotation. Since the data being release by J&J might be used for model building purposes, we decided to re-create the probe annotations from scratch using available high-confidence transcript and gene annotations. The end result of our analysis is a partition of the overall set of Codelink probes into the following 5 confidence categories:

1. probes mapping to known full-length reviewed transcripts,
2. probes mapping to full-length unreviewed transcripts,
3. probes mapping to intronic and near-gene regions,
4. probes mapping to the intergenic regions, and
5. probes that do not map to either the genome or the transcriptome.

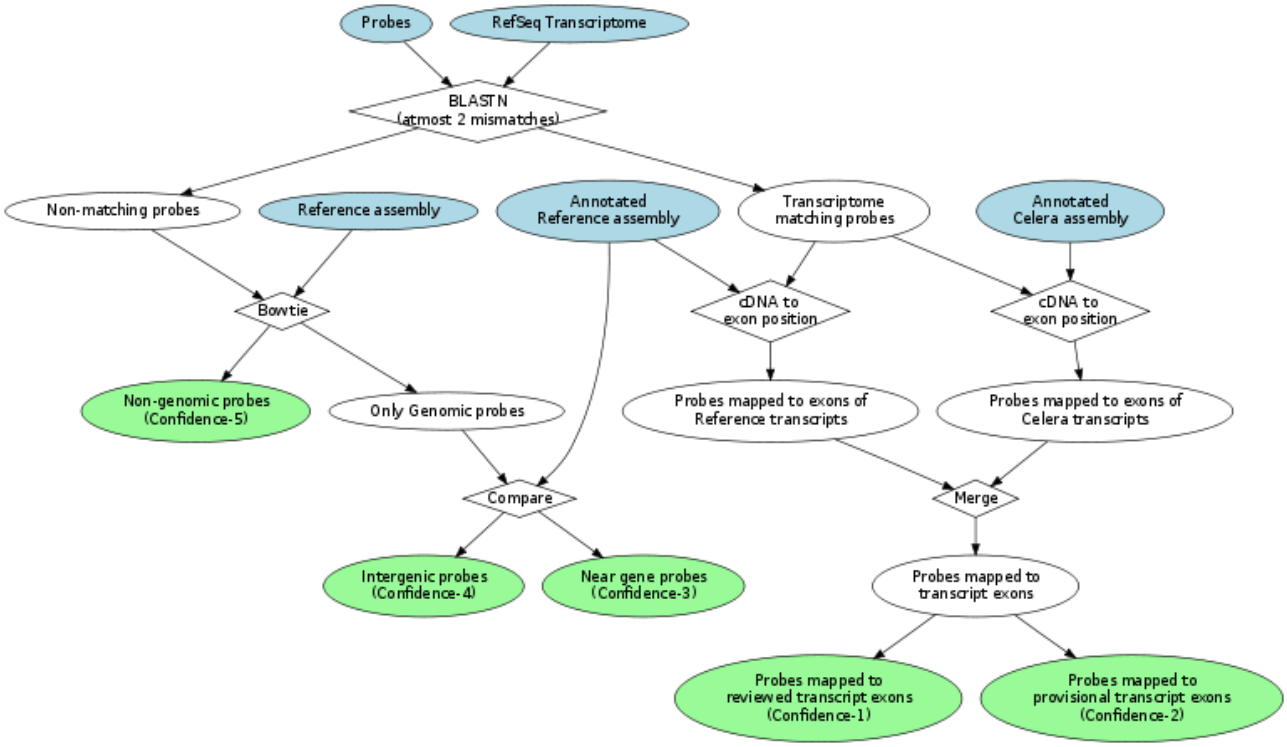
This classification can help end-users in restricting their analysis to probes they deem to be of high-confidence. The exon level information provided by this annotation may also be useful in ordering probes from ABI for PCR validation purposes.

Data used for analysis:

- Transcriptome: the collection of all 30,423 RefSeq transcripts were downloaded from GenBank in .gbk format. These files contain the mRNA sequences and coding region information, but do not contain the positions of the exons in the transcript.
- Transcript annotations on the assembly: Annotated chromosomes in .gbs format were downloaded from the NCBI genomes ftp site. The annotation is with respect to a particular assembly. There are two alternate assemblies available for rat – the reference assembly, and the Celera assembly. The transcripts annotated on these assemblies are different. Even when the same transcript is present on both assemblies, its exon-intron structure can be differ in the two assemblies. Both assembly annotations were downloaded.
- Reference assembly chromosome files: Indexed versions of the reference assembly sequences which allow quick sequence mapping were downloaded from the Bowtie site on sourceforge.
- Codelink probes: A fasta file containing 33,961 thirty-length probe sequences was downloaded from the Hydra site at J&J.
- Reference Entrez genes: The Entrez Ids of all rat genes present in the reference genome assembly was obtained by parsing files from the NCBI Gene ftp site.

Methods:

The following picture presents a high-level overview of the pipeline that was used.



The starting point for our analysis is a mapping of probes to the transcript sequences in the RefSeq transcriptome using BLASTN. The fundamental difference between RefSeq transcripts and other mRNA accessions at GenBank is that RefSeq transcripts are meant to represent a non-redudant set of full-length transcript models. A large number of accessions are used as supporting evidence for constructing a RefSeq transcript.

Ideally, every probe present should align without any mismatches against some transcript in the RefSeq transcriptome. However, the BLASTN output consisted only of 20,291 matches involving 15,776 distinct probes matching to 15,964 distinct transcripts. Since the total number of probes is 33,961, this is rather surprising. One possibility is that the RefSeq transcriptome is not comprehensive – i.e. there are genes with transcripts that are not covered by the RefSeq transcriptome . To check this possibility we compared the genes covered by the transcriptome with the list of all reference assembly genes available at Entrez Gene. The results shown in the below table indicate that the transcriptome has very good coverage (93%) of the protein-coding genes at Entrez Gene.

Gene Type	All Entrez Genes	Entrez Genes not represented in the transcriptome
Protein-coding	22,385	1,540
Pseudo	8,261	7,826
Other	292	153
Unknown	91	6
MiscRNA	33	31
rRNA	2	2
snRNA	1	1

Total	31,087	9,581

The transcriptome data contains the sequence for each transcript but does not contain exon information. So the probe matches to the transcriptome are only in cDNA co-ordinates but do not directly indicate the exon number(s) that the probe overlaps. The exon-intron structure of a transcript is only available when the transcript has been aligned against a genomic assembly. Two different assemblies are available for rat – the Reference assembly and the Celera assembly. In addition to the fasta sequences of the assembled chromosomes, NCBI also provides RefSeq transcript annotations with respect to these assemblies.

RefSeq transcripts can be classified along two orthogonal directions:

- manually reviewed transcripts start with the letter 'N', while those created by computational pipelines start with the letter 'X'. The latter are called provisional RefSeqs and these accessions are typically short-lived. As and when they are reviewed, they are either discarded because they do not meet the standards or promoted to curated RefSeq status and assigned a new accession starting with 'N'.
- transcripts of protein-coding genes have the second letter as 'M', while non-coding transcripts have the second letter as 'R'. These could be miRNA, snoRNA, ribosomal RNAs etc.

Taken together we have 4 possibilities – NM, NR, XM, XR.

The following table shows the distribution of the different types of RefSeq transcripts in the two assemblies and in the overall RefSeq transcriptome.

	Reference	Celera	Transcriptome
NM	15518	15739	16754
NR	23	24	0
XM	6412	6149	12600
XR	512	511	1069
Total	22465	22423	0

Most of the NM accessions are common between the two assemblies, while XM and XR transcripts are usually specific to an assembly.

We use the transcriptome mapping results along with the assembly annotations to infer the exon to which each probe maps. This needs to be done once w.r.t the Celera annotations, and once with the reference assembly transcript annotations. Most of the time the exon number inferred from the Celera annotation agrees with the exon number inferred from the reference assembly annotations – but in 7% of the cases there is a disagreement. For instance, probe GE22218 matches to the 41-70 position of NM_019346. However, the first exon of NM_019346 has length 67 in the reference assembly and length 103 in the Celera assembly. So the probe will be annotated as spanning exons 1,2 w.r.t the reference assembly and exon 1 w.r.t the Celera assembly.

6% of the probes are exon-spanning probes while the remaining 94% fall entirely within a single exon. Overall these 15,535 probes that get mapped to exons should be considered as high quality reliable probes. Within these we can again make a distinction between the 13,219 that map to

curated transcripts (NM_) and treat them as the most reliable probes.

So, overall, we can create 5 lists:

1. probes mapping to curated transcripts (13,219)
2. probes mapping to provisional transcripts (2,316)
3. probes mapping to intronic, or within 500-bp of a RefSeq transcript (this can be inferred only by mapping to the whole genomic assembly). We have done this only for the reference assembly and not for the celera assembly (6,756)
4. probes mapping to other genomic locations on the reference assembly (9587)
5. probes not mapping anywhere (2083)

Validation:

To validate this classification, we checked which of these probes had valid Entrez Ids in the latest rwgcod annotation package in BioConductor. Intuitively, a large percentage of our highest confidence probes (List-1) should have a valid Entrez ID, while only a small percentage of our lowest confidence probes (List-5) should have valid Entrez Ids. The following table shows that this in indeed the case.

	Size	**w/o rwgcod Entrez ID**	**% without Entrez**
List-1	13219	276	2.08
List-2	2316	438	18.91
List-3	6756	3803	56.29
List-4	9587	6843	71.37
List-5	2083	1294	62.12