The Development of a High-Density Canine Microarray


From Nestlé Research Center Lausanne, Nestec Ltd., Vers-chez-les-Blanc, 1000 Lausanne 26, Switzerland (Holzwarth, Mansourian, and Raymond); and Nestlé Purina Research, Checkerboard Square, St. Louis, MO 63164 (Middleton, Roberts, and Hannah).

Address correspondence to James A. Holzwarth at the address above, or e-mail: james.holzwarth@rdls.nestle.com.

Abstract

DNA microarrays can give global transcriptional views of cellular responses to disease, development, nutrition, and other biological states. They can be used to elucidate biological networks, develop diagnostics, and identify genetic targets and molecular mechanisms. The technology is widely used and can be a valuable complement to more “disease-centric” focused arrays. For these reasons, Nestlé designed a custom canine Affymetrix microarray representing transcripts from multiple tissues for use in areas where a more focused microarray had not already been developed. Sufficient numbers of sequences representing messenger RNAs (mRNAs) or expressed sequence tags (ESTs) is integral for the design of a global microarray chip. This chip was designed using public domain sequences (GenBank) and sequences from a proprietary canine EST database. In order to enrich the chip with annotated transcripts, both of these sequence sets were BLASTed against the nonredundant protein database. The sequences on the microarray were isolated from more than 48 different tissues. The final compliment of sequences had sequences unique to GenBank (3160), unique to the proprietary EST database (17,620), and present in both sources (1996). In comparison with human sequences (RefSeq), 74% of the canine sequences matched a human sequence.

Array-based transcript profiling experiments enable the detection of thousands of gene expression patterns in a single experiment (Duggan et al. 1999). In the Gene Expression Omnibus (GEO) database, there are currently 30,000 such experiments performed on more than 100 organisms (Barret et al. 2005). Having such resources available makes the ability to compare experiments much more valuable. The ability to compare past and future measurements depends on the equivalence of probe sequences (Kuo et al. 2002). The Affymetrix platform is a good candidate as a stable platform, since it has been shown to have good reproducibility (Ishii et al. 2000; Novak et al. 2002) and, once made, a microarray will have an unchanging compliment of probes. The annotations of the sequences change rapidly, which means that it is possible to improve the accuracy of the analysis (Gautier et al. 2004).

Here we describe the design of a canine array using messenger RNA (mRNA) and expressed sequence tag (EST) sequences. The quality of the sequences was the most important criteria for inclusion in the target sequence set. For the sequences assembled out of several sequences, the consensus was recalculated. Comparison with a nonredundant protein database also allowed the exclusion of sequences that did not correspond to a protein in order to reduce genomic sequence contamination of the target sequence set.

Materials and Methods

DNA Sequences

The sequences used were the canine mRNA sequences present in GenBank (Benson et al. 2004; July 2003 version) and a proprietary EST sequence database containing approximately 350,000 ESTs. The EST sequences had associated phred (Ewing and Green 1998) quality values. The singletons were included when the average phred quality was greater than 25. The nonsingletons (sequence clusters with more than one representative sequence) were accepted if the assembly sequence was composed of sequences with an average quality value of more than 25. The GenBank sequences were used directly from the database, whereas the sequences from the proprietary EST database were assembled using phrap (Green 2005) and had the consensus sequence recalculated based on the quality values. The sequences were BLASTed (Altschul et al. 1990) against the nonredundant protein database generated by LION Biosciences.
Germany). A set of selection criteria were established based on these results: fraction of protein sequence coverage more than half (exclusion of potential chimaeras, misassemblies), sufficient 3' coverage (comparison with protein C-terminus), exclusion of sequences with a change in direction of match between protein and DNA (misassembly), at least 50% similarity between the DNA and protein sequence where the length of the match corresponded to at least 35 amino acids.

**Probe Sequence Selection**

The Affymetrix GeneChip Custom Expression Array (Affymetrix, Santa Clara, CA) was used to generate the microarrays. The parameters for the microarray chosen were as follows: 18 μm feature size, standard 12.8 mm array format, eukaryotic antisense target type, probe selection region 600 bases from the 3' end, and eight as the minimum acceptable probes per sequence. The criteria for selection of the probes were only unique probes (do not match other sequences within the list of target sequences), probe sets with 11 probe pairs, the number of independent probes being two or more, and a probeset score greater than 2.1021 (chosen empirically).

**Probe Genome and Probe RefSeq Comparison**

Probe sequences were BLASTed (Altschul et al. 1990) against the CanFam 1.0 assembly (Lindblad-Toh et al. 2005). A probe was considered to match if it was 100% identical for the entire length. A probe set was identified as matching if more than half of the probes in the probe set matched with 100% identity; similarly the categories for no match or more than one match were also identified either based on the lack of match or more than one match, respectively, for the majority of the probes within a probe set. For the dog–human comparison, canine target sequences were BLASTed (Altschul et al. 1990) against human RefSeq (Pruitt et al. 2005) sequences. A match was considered where the expect value was less than 1e-10 and a 200 bp match. For the human–dog comparison, RefSeq human sequences were BLASTed against canine target sequences. A match was considered where the expect value was 1e-10 or less and there was a 200 bp match.

**Results**

Canine-specific sequences from two sources, GenBank mRNAs and custom ESTs, were gathered and subjected to different types of filtering criteria (Figure 1). When EST sequences were used, associated quality data was present; these data were used to ensure that low quality sequences were not included in the subsequent stages of selection. The GenBank DNA database sequences used corresponded to approximately 24,000 mRNA sequences. The proprietary canine database used for this design contained approximately 350,000 clones (average length 616 nucleotides) or ESTs coming from different organ systems, with the majority (approximately 300,000) being 3' ESTs. The assemblies (contigs) were performed using phrap (Green 2005) and the consensus sequence recalculated. There were approximately 100,000 clusters of sequences with about 70,000 clusters with one sequence and about 30,000 clusters with more than one sequence in the proprietary canine database.

In order to allow the determination of the degree of annotation for each of the sequences and to exclude potentially erroneous sequences from the selection process, the sequences were compared with the nonredundant protein database generated by LION Biosciences containing 1,185,971 protein sequences. Sequences identified as mRNA sequences or EST sequences with similarity to independently identified protein sequences have increased reliability. Sequences were also identified that correspond to full-length transcripts (as indicated by the corresponding protein sequence coverage). This is useful for functional annotation purposes. Sequences matching the C-terminal end of the protein sequence were also identified. Affymetrix recommends selecting sequences within 600 nucleotides of the 3’ end.

The selection criteria reduced the input to 24,550 sequences (5376 from GenBank and 19,173 proprietary), with an average length of 1408 bases. These sequences were submitted to Affymetrix for probe selection.

The Affymetrix selection file contained several parameters for selecting the probe sets. A given target sequence is represented on the chip by a set of probe pairs. The
Affymetrix high-density microarray we chose could accommodate 250,661 probes. The probe pairs were selected to be unique in order to avoid ambiguous results. A combination of empirically chosen probeset scores and a further elimination of 50 sequences that corresponded to protein sequences with uninformative annotations were used in order to arrive at the final selection of probe sets representing 22,786 target sequences. In the final selection, 3160 were from GenBank, 17,620 were from the proprietary sequences, and 1996 were present in both. Of the GenBank sequences on the chip, 2075 were not found in UniGene; 242 were found in one tissue and 1860 were found in multiple tissues. Of the proprietary sequences, 5554 were found in only one tissue and 13,045 were found in many tissues. Table 1 shows the various tissues represented on the array and the number of sequences from each tissue.

The probe sequences were compared with the CanFam 1.0 assembly. Some of the probesets matched once to the CanFam 1.0 assembly (18,735) (Figure 2), others matched more than once (598), and the rest did not match at all (3036). Seventy-three percent of the probes matched perfectly, while 27% did not match the CanFam 1.0 assembly. A very few of the probesets had more than half of the probes that matched more than once, which would indicate that they are part of gene families. Those that did not match at all on the genome (Figure 2) could be either due to errors, single nucleotide polymorphisms (SNPs), or probes that cross exon boundaries.

A comparison between human and canine target sequences revealed that 74% of the canine sequences used matched human sequences. Similar values were obtained using human Ensembl sequences (data not shown). Using the human sequence to search against the canine sequence set showed that 69% of the human sequences were also found.

**Discussion**

Quality filtering was performed to ensure that only high quality sequences were chosen. The way this was performed may have excluded correct sequences that did not possess associated high quality data. However, the consequences of low quality (or incorrect) sequence inclusion were much greater than the potential incompleteness of the sequence set. Incorrect sequence inclusion (false positive) has a higher cost with respect to omission (false negative) because of the expense of interpretation, validation, and follow-up.

The selection based on comparisons with protein sequences was used to exclude any genomic contamination. While every possible strategy is used to ensure genomic
sequences are not included in such libraries, there remains some small potential for contamination. These criteria were seen as a way to further decrease the risk. It is possible that these criteria excluded some unknown sequences from being included in the sequence set. While this is possible, the fact that the protein database contained all known protein sequences from all organisms minimized this risk. This strategy also excludes sequences with a very long 3' untranslated region (UTR) or in cases where the EST sequence is not extended into the coding sequence. Of course, assembled sequences are often longer than their EST counterparts; in such cases, there could be a bias that would influence the selection toward sequence assemblies as compared to singleton sequences. The possibility of excluding unknown sequences or sequences not extending into the protein coding sequence remains, but this measure was considered a reasonable trade-off in an effort to exclude spurious false positives. And having a match to an independently identified protein sequence increases the quality of the chosen sequences, since they have corroborating data.

Selecting probesets with little cross-hybridization excludes probesets that would recognize the common elements of several splice variants of the same gene. In such a case, probesets for genes with many splice variants could be excluded for uniqueness criteria.

The comparison with the CanFam 1.0 assembly was also important for two principal quality control reasons. First, it was important as a validation of the data by comparison with data generated in a completely different manner. Second, it was important since the libraries from which the majority of the sequences were taken originated from a beagle and the CanFam 1.0 assembly was generated from a boxer. The 27% that do not map to the boxer assembly could indicate probes crossing exon boundaries as well as probes falling on SNPs and sequencing errors.

The comparison of canine sequences and human sequences (RefSeq) is important in order to indicate that the sequence set is relatively complete. Seventy-four percent of the canine genes on the microarray matched to human genes. A smaller percentage (69%) of human genes match a canine sequence present on the microarray. While this shows that the sequence set represented is not complete, it certainly represents a large proportion of the known human genes. It is also possible that dog simply has fewer genes than human.

The present chip was designed to have wide coverage and high quality sequences. The proprietary sequences added 17,620 sequences that were not available in the public domain, while 3160 were present in the public domain and not in the proprietary sequences. The design decisions were chosen in such a way as to have both sequences with high and low abundance represented. The majority of the sequences were found in multiple tissues (14,905), while 5796 were only found in one tissue. The array represents sequences taken from 48 different tissues. The probesets present on the array were chosen in order to decrease ambiguity and to have the clearest identification of transcripts.

The described canine microarray represents the majority of canine genes, with sequences selected from many different tissue sources, with a mix of sequences found in one tissue and multiple tissues. The availability of rare transcripts, the number of sequences not found in GenBank, and the quality of the included sequences make this array quite unique and a useful tool for general transcript profiling.

The availability of a canine-specific microarray will give veterinary research a valuable tool. Not only can such a tool be used to develop diagnostics, it can also be used for more basic research. The elucidation of biological networks and the identification of molecular mechanisms possible with such a resource can lead to a deeper understanding of canine health and disease.

Acknowledgment

This paper was delivered at the 2nd International Conference on the “Advances in Canine and Feline Genomics: Comparative Genome Anatomy and Genetic Disease,” Universiteit Utrecht, Utrecht, The Netherlands, October 14–16, 2004.

References


Corresponding Editor: Kerstin Lindblad-Toh