



Operon Biotechnologies, Inc.
2211 Seminole Drive
Huntsville, AL 35805
www.operon.com

***Canis familiaris* (dog) Genome Array Ready Oligo Set™ (Version 1.0)**

Canis familiaris (dog) V1.0.1 AROS is designed using state-of-the-art methodology and proprietary software for detection of alternative splice variants from Ensembl *C. familiaris* release 44.2b genebuild. This version of the Ensembl gene and transcript predictions and annotations were based on from the CanFam 2.0 whole genome shotgun assembly (May 2006) provided by the Broad Institute. CanFam 2.0 genebuild predicts 27,310 gene transcripts and 1,926 RNA genes. In this AROS set, 27,039 transcripts are covered by at least one AROS oligo probe providing essentially full coverage of the canine nuclear and mitochondrial genomes.

Probe design targets each oligo to be fully contained within a single exon and makes this probe set also applicable for CGH (comparative genome hybridization) experiments. The *C. familiaris* AROS V1.0 contains 25,383 probes including 134 Stratagene SpotReport™ Alien™ control oligos, 48 Negative controls, 288 Positive and Positive Stringency control oligos. Additionally, 267 Production control oligos used by Operon for internal checks and also useful to researchers for plate layout checks are included free of charge. These AROS control probes supplement with the set to facilitate experimental validation and quality control.

Release date: July, 2007

I. The sequence sources

Canis familiaris genome sequence and its associated annotations are from Ensembl *C. familiaris* release 44.2b genebuild. Additional sequences used in the AROS design are accessible at the National Center for Biotechnology Information (www.ncbi.nih.gov).

II. Design criteria and selection rules:

The following selection rules were executed in the design.

- 1) Target length of oligo probe sequences was 70 bases. Exceptions for genes and regions without qualified oligo candidates at this length are routinely made in AROS probe sets in order to maintain tight control of hybridization characteristics to allow

for highly stringent experimental conditions. These exceptions are based on the inherent characteristics of the *C. familiaris* genome and not a limitation of the design algorithm.

- 2) The melting temperatures (T_m) of $73 \pm 5^\circ\text{C}$ was used for first round probe candidate screening. T_m was calculated using the nearest neighbor algorithm.
- 3) The oligo candidates were pre-selected with 40 bases away from both 5'- and 3'-end of transcripts as well as a maximal distance of less than 1000 bases from the 3'-end.
- 4) The oligo candidates were screened against the low complexity of sequences: (a) The contiguous single nucleotide base repeat or poly (N) tract (≤ 8 bases); (b) the pre-set normalized simple repeat score.
- 5) The oligo candidates were selected against the potential hairpin structure with stem length of over 9 bases.
- 6) The oligo candidates were screened against the cross-hybridization of the non-self transcripts. The non-self transcript is the transcript not represented by a specific oligo. The selection criteria were used in the selection: (a) the cross-hybridization identity score $\leq 70\%$; (b) contiguous match length to any non-self transcript ≤ 20 bases.
- 7) Limited exceptions were made if no qualified oligos were available from the selection. Information covering these exceptions is detailed in the *C. familiaris* (dog) AROS V1.0 genelists.
- 8) Probe design targets each oligo to be fully contained within a single exon and makes this probe set also applicable for CGH (comparative genome hybridization) experiments.

III. Annotation notes:

Oligos were identified by using NCBI BLAST to map each oligo to the sequences in the databases described above. Identifiers are retained, added, or deleted according to the BLAST score (score $>97\%$ identity).

Ensembl transcripts are further updated by using NCBI BLAST to map each oligo to the transcripts and the exons for the Ensembl release as described below. The Ensembl transcript ID(s) from the gene list were compared to the archive list in the core database to obtain the latest version number to check the match. It is possible that an oligo maps to more than one Ensembl transcript and/or gene and in this case the transcript IDs are separated by 'comma' symbols (,) in the same order as the genes appear.

Oligos based on Ensembl data include the following ancillary data from the core, compara, and external databases: gene symbol and description; human and rat homologies; gene ontology terms; and gene family information. The latter is useful for determining the impact of oligos that map to more than one gene.

BLAST hits for Ensembl transcripts and exons are compared to determine the extent of the coverage. An oligo may map to more than one transcript and it may spread across multiple exons. The definitions for transcript coverage are shown below.

Oligo Type Definition

I Oligo represents one transcript of an Ensembl gene with multiple transcripts

CI Oligo represents the only transcript of an Ensembl gene

C Oligo represents all transcripts of an Ensembl gene

P Oligo represents a group of transcripts of an Ensembl gene(s)

Oligos are further described according to whether they span more than one exon as described below.

Exon Type Definition

E Oligo is fully contained in one exon

T Oligo spans more than one exon

X No exon hits, indicating that the transcript is not represented.

The oligos are also mapped with NCBI BLAST to the genome represented by the Ensembl reference chromosome sequences to determine their potential for comparative genome hybridization (CGH) experiments. An oligo that is unique is indicated as 'Predicted_CGH_Oligo' and an oligo whose maximum cross-hybridization match is less than 85% identical is indicated by 'Predicted_CGH_Oligo_Potential'. The maximum cross-hybridization score is provided.

Several microarray data analysis platforms utilize Genbank identifiers to incorporate gene annotation data. All BLAST hits greater than 97% identity are included.

IV. Characteristics of *Canis familiaris* (dog) AROS V1.0 Control Probes

1) Control Oligos:

- Negative Control Oligo Probes: Randomly-generated oligo sequences which were selected after filtering against the transcripts.
- Positive Control Oligo Probes: 24 oligo probes were randomly selected from this AROS as the positive controls including both known and predicted transcripts. They and the corresponding Hybridization Stringency Control Oligos (see below) are randomly scattered throughout the AROS set to be used to check a variety of potential issues including wash stringency and variation of hybridization stringency in different areas of the array.
- Hybridization Stringency Control Oligos: Generated based on several positive control oligo sequences with sequence homology of 50% 60% 70%, 80%, 90% and antisense identity to the original oligo probe sequences. They're intended as the standards to calibrate the microarray hybridization conditions and stringency.
- Stratagene SpotReport™ Alien™ Spike Control Oligos: Licensed from Stratagene (www.stratagene.com). Coupled with the Alien mRNA spikes in Stratagene SpotReport Alien Oligo Array Validation System, they are intended as the internal controls for the normalization and standardization of dye incorporation, microarray hybridization and data analyses.
- Production Tracking Oligo opHsV04NC000001: It is a randomly-generated oligo sequence with a length of 30 bases. The tracking oligo is randomly positioned in the 384-well and 96-well plates so that each 384-well plate has

four, and each 96-well plate has one for the purpose of quality control in production. The CyDye labelled antisense probe for this oligo can be used in a QC assay to ensure the correct placement/order of the probe printing plates on the spotting robot.

Figure 1. Melting Temperature Distribution

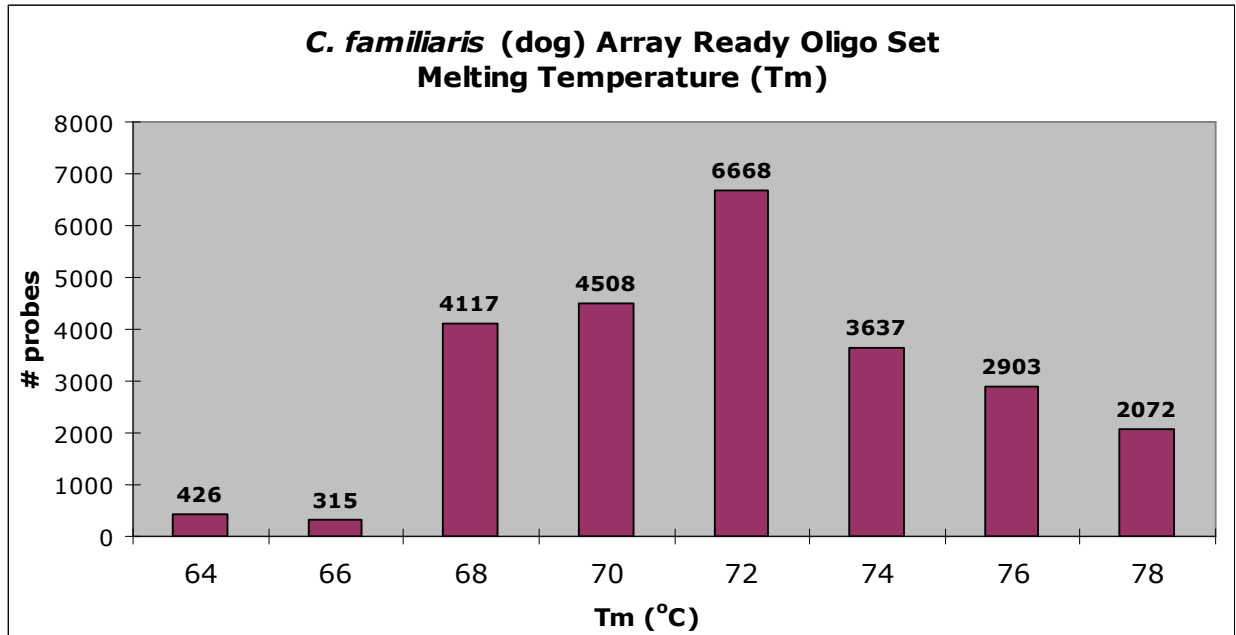


Figure 2. Distribution of Oligo Probe GC Content

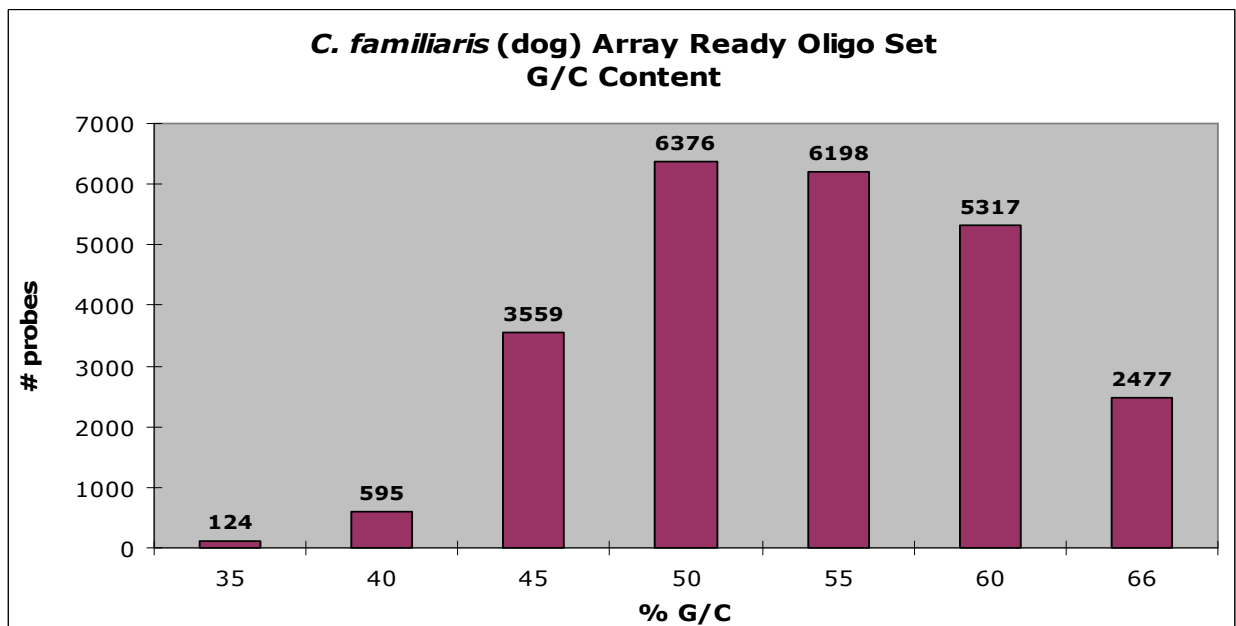


Figure 3. Distance to the 3'-end of Transcripts

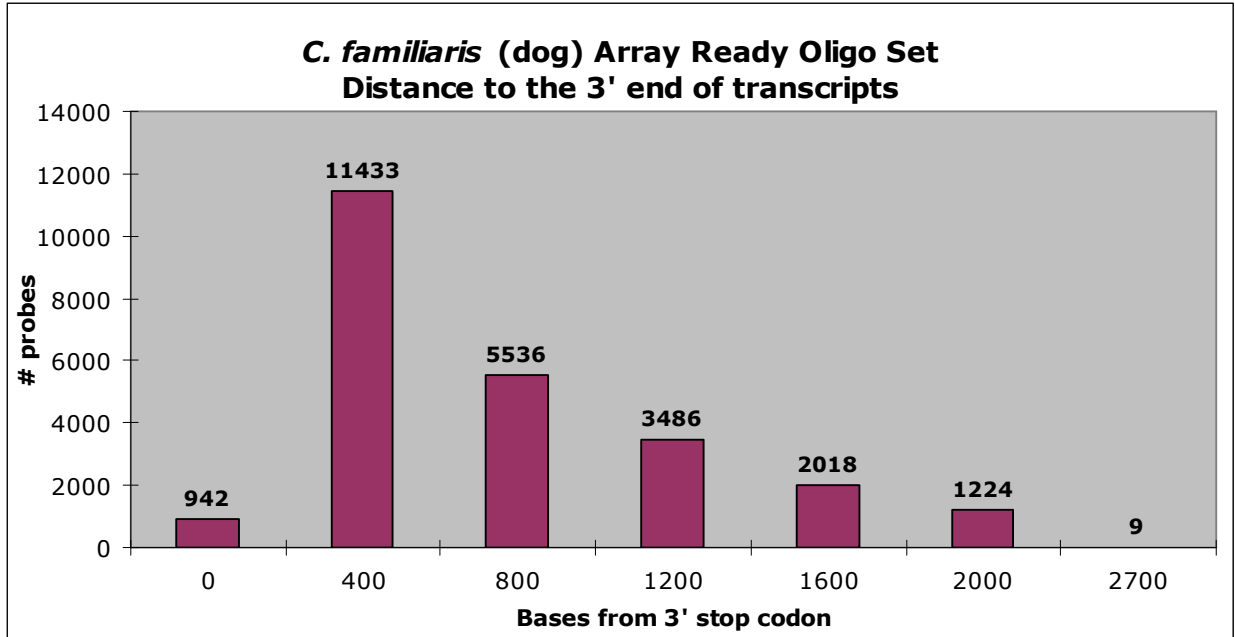


Figure 4. Hairpin Stem Length

