

Array-Ready Oligo Set™ for the Zebrafish Genome
Version 1.0

We are pleased to announce Version 1.0 of the Zebrafish (*Danio rerio*) Array-Ready Oligo Set (AROS) containing 3479 70mer probes representing 3479 *Danio rerio* genes. The set of 3479 probes represents cloned Zebrafish genes and some expressed sequence tags (ESTs). For our probe design we use state-of-the-art methodology and proprietary software. An amino linker is attached to the 5' end of each oligo.

Gene Sequence Source and Selection

All probes are designed from the UniGene Database Build Dr 41, and the Zebrafish Reference Sequence (RefSeq) database, developed and maintained at the National Center of Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov>). For AROS design, we selected 1206 sequences, representing well-annotated genes, and 2273 ESTs with high or moderate homology to known genes.

Advantages of Using Gene Sequences from UniGene and RefSeq

UniGene is an open source database freely available to everyone. UniGene automatically clusters GenBank sequences into a non-redundant set of gene-oriented clusters. It has become one of the most widely used *de facto* standards in the public domain for cataloguing genes. Each UniGene cluster contains cloned genes and/or ESTs sequences that represent a unique gene. UniGene sequences are filtered for contaminant sequences, genomic repetitive regions, and low-complexity sequences using NCBI's Dust program. All 70mer oligos are designed from the representative sequence. This chosen representative sequence is the sequence with the longest region of high-quality sequence in each cluster. For more information, refer to the Zebrafish gene list.

The NCBI Reference Sequence project is an effort to standardize gene sequence references by providing a NCBI-staff curated gene sequence and avoiding gene sequence redundancy. Each RefSeq is linked through a LocusID number to NCBI's LocusLink database. RefSeq genes are quickly becoming the *de facto* standard and are used by a large research community.

The following are advantages of designing a probe directly from a gene sequence from the RefSeq collection. Each of these advantages are provided by the NCBI LocusLink interface (<http://www.ncbi.nlm.nih.gov/LocusLink/>)

Features provided for each RefSeq through LocusLink	Example
Gene	Beta-catenin
Alternate gene symbols	ctnnb
Locus ID	30265
LocusLink Report	http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=30265
Chromosome	16
Official gene symbol and description (ZFIN) ZFIN.ZDB-GENE-980526-362 http://zfin.org	ctnnb: beta-catenin is involved in cadherin-mediated intercellular adhesion, and in mediating signal transduction by some Wnt and Frizzled homologs. It is expressed throughout development in most if not all cell types.

Sequence source	Number of oligos
Simultaneously represented in UniGene and RefSeq	712
Represented in UniGene only	2767
Total number oligos in Zebrafish Genome Oligo Set	3479

Probe Design and Selection Rules

Once a gene has been selected to be included in the set, a probe is selected with an optimal set of parameters. Sufficient numbers of 70mer candidate probes for each gene are selected using the following criteria for the Zebrafish Genome Oligo Set.

1) All oligos are within $78^{\circ}\text{C} \pm 5^{\circ}\text{C}$ using the following formula:

$$T_m = 81.5 + 16.6 \times \log[\text{Na}^+] + 41 \times (\#\text{G} + \#\text{C})/\text{length} - 500/\text{length}$$

where $[\text{Na}^+] = 0.1 \text{ M}$ and $\text{length} = \#\text{A} + \#\text{C} + \#\text{G} + \#\text{T}$

2) Each oligo is within 750 bases from the 3' end of the available gene sequence. For 161 (4.6%) oligos location from the 3' end is greater than 750 and less than 970 bases.

3) An oligo cannot have a contiguous single nucleotide base repeat or poly (N) tract longer than 10 bases.

4) An oligo cannot have a potential hairpin structure with a stem length longer than 9 bases.

5) A normalized score is assigned to each oligo based on the number of repeats. Oligos with more repeats having a normalized score greater than a certain threshold are filtered out.

6) Each oligo has less than or equal to 70% identity to all other genes. For all oligos in the Zebrafish Genome Oligo Set, using BLAST, each oligo is aligned against all 15,956 representative sequences in Zebrafish UniGene Build Dr 41. Using the alignment with the candidate oligo versus the highest scoring non-self gene, a BLAST percent identity score is computed. The highest scoring non-self gene is defined as the sequence that yields the most matched bases in an alignment. This BLAST percent identity is also referred to as cross-hybridization identity of the oligo.

This calculated cross-hybridization identity is dependent on the size of the sequence database used to BLAST against, oligo sequence, and use of either gapped or no-gap alignment method.

7) Each oligo of any length cannot have greater than 20 contiguous bases common to any other gene.

Once oligo candidates have been selected satisfying all the selection rules mentioned above, each oligo is ranked based on BLAST percent identity as computed in Step 6. One final oligo for each gene is selected with the minimum percent identity or crosshybridization similarity.

SUMMARY

Oligo selection criteria	Criteria values	Number of oligos in genome set satisfying these criteria
Length Melting temperature Poly(N)tract length Stem length in potential hairpin Cross-hybridization identity to all other genes Contiguous base match to any other gene	70mer 78°C ± 5°C ≤ 10 ≤ 9 ≤ 70% ≤ 20	3479
Location from 3' end	≤ 750	3318
Location from 3' end	750 < x < 970	161
Total		3479

The following illustrations show the distribution of all 3479 oligos representing the Zebrafish AROS for melting temperature, GC content, location from 3' end of gene sequence, length of maximum stem length, and BLAST percent identity or cross-hybridization similarity.

Figure 1. Melting Temperature

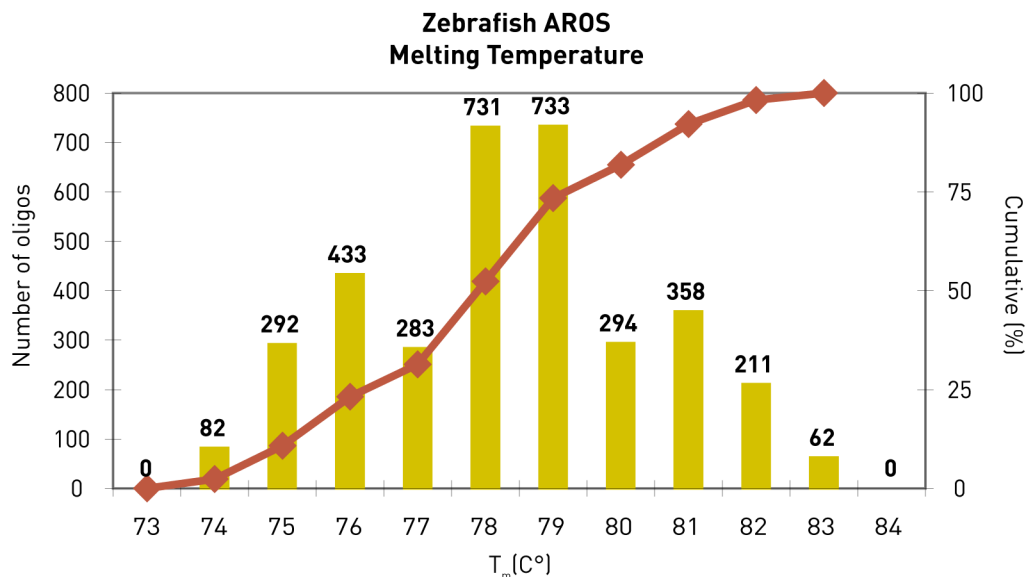


Figure 2. GC Content

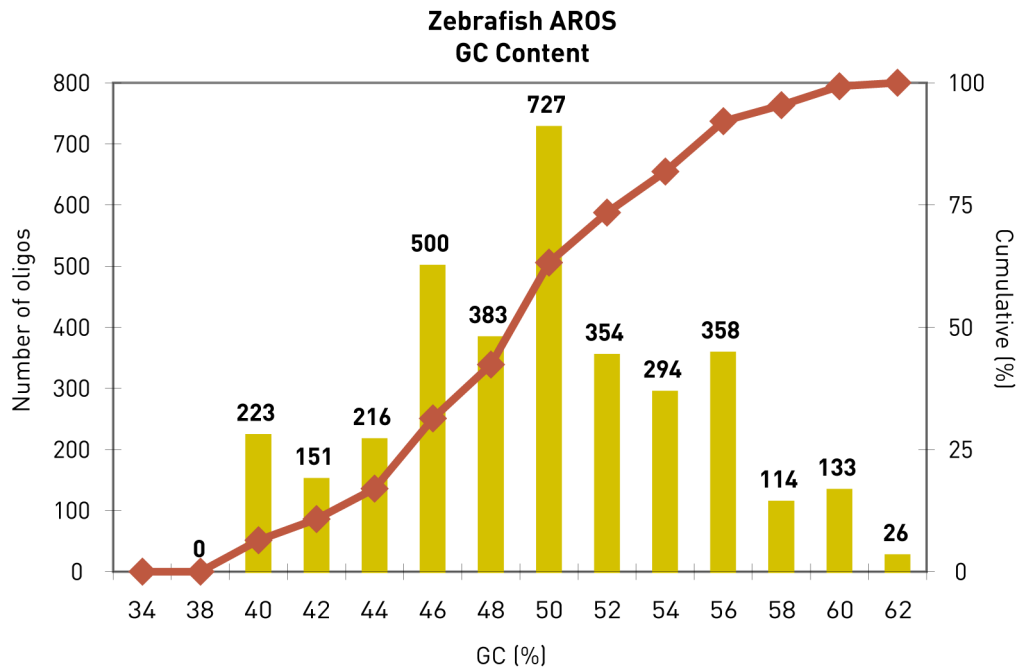


Figure 3. Location from 3' End

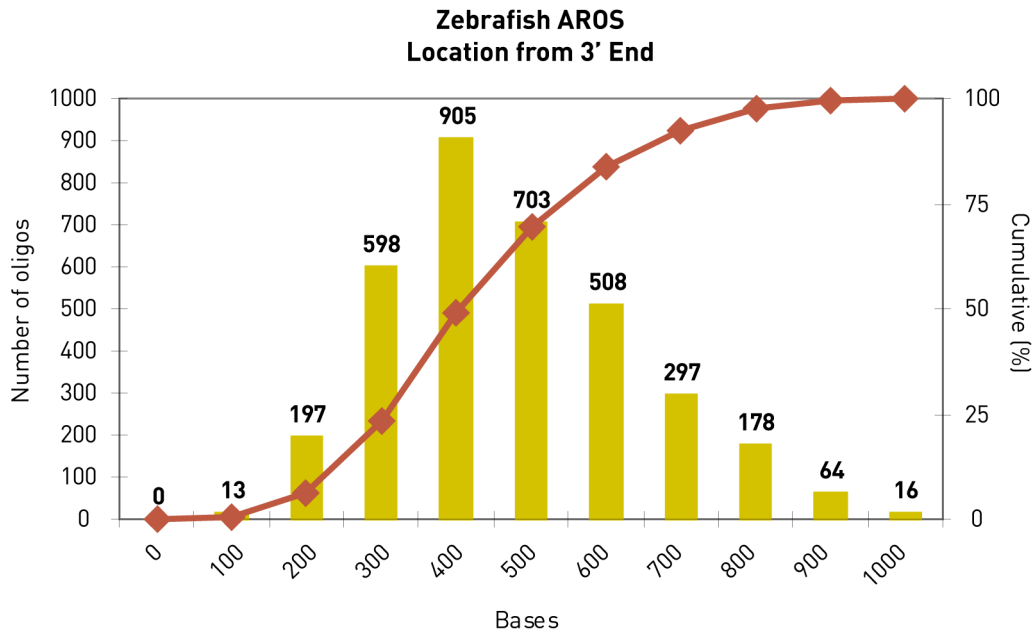


Figure 4. Length of the Longest Hairpin Stem

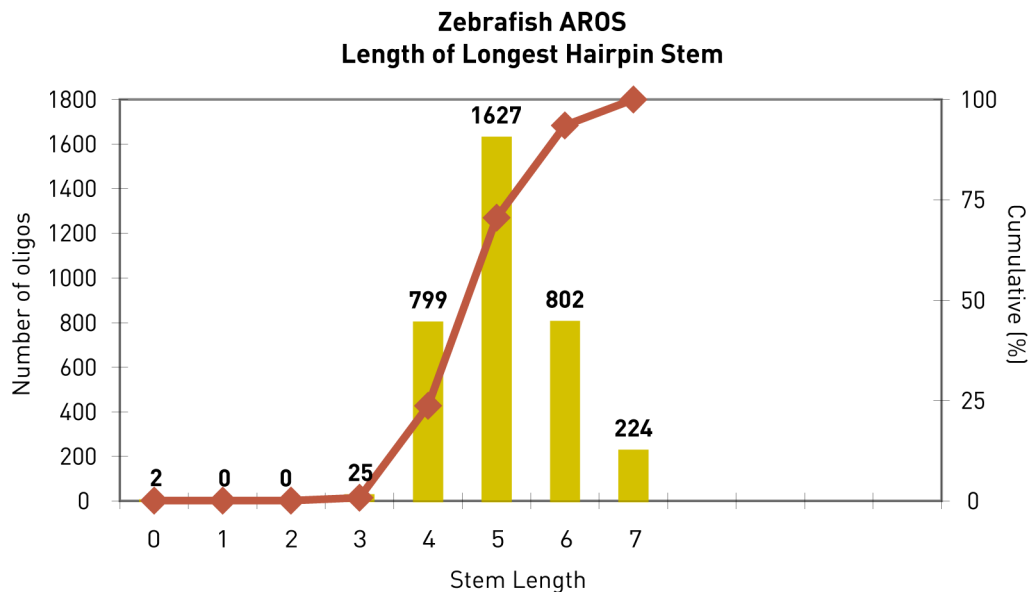
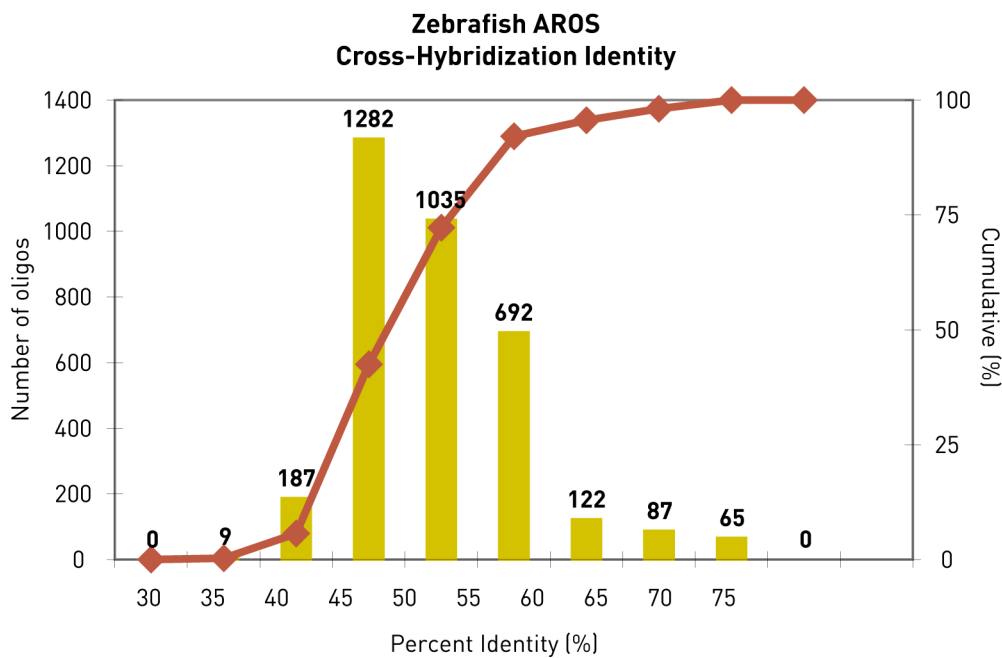


Figure 5. Cross-Hybridization Identity



Quality Check of Probe Design Specifications

Once the final oligo set has been selected to represent a gene, each oligo undergoes design specifications quality control where we use an independent method to confirm that all oligos have met the specified design specifications. The table below summarizes data from our quality check for probe design specifications for all 3479 probes in Version 1.0.

Probe design specification	Expected value	Verified range	Number of oligos
Melting temperature (C°)	78°C ± 5°C	73.6–82.9	3479
GC content (%)	38%–62%	38.6–61.4	3479
Cross-hybridization similarity (%)	≤ 70	34–70	3479