

### Array-Ready Oligo Set™ for the *Arabidopsis thaliana* Genome Version 1.0

We are pleased to announce the release of our *Arabidopsis thaliana* Genome Oligo Set Version 1.0 containing 26,090 70mer probes representing 26,029 *A. thaliana* genes. The 26,090 probes are designed from experimentally determined or cloned genes, expressed sequence tag (EST) sequences, and predicted transcripts. For our probe design methodology we use sophisticated design methods, which were developed based on our experience. An amino linker is attached to the 5' end of each oligo.

#### Gene sequence source and selection

The 26,086 probes are designed from the UniGene Database Build At 4 developed and maintained at the National Center of Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov>). Additionally 4 probes are designed from GenBank sequences not in the UniGene Build At 4.

#### Introduction to Ensembl

Ensembl (<http://www.ensembl.org/>) is a joint project between the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL–EBI) and the Sanger Institute to develop an automated genome annotation database and browser that includes human, zebrafish, rat, and several other eukaryotic genomes. The success of array oligo design hinges on the quality and annotation of sequence information.

#### Advantages of using gene sequences from UniGene

UniGene is an open source database. UniGene automatically clusters GenBank sequences into a nonredundant 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 *A. thaliana* UniGene cluster contains experimentally verified or cloned genes, EST sequences, and/or predicted transcripts 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 oligos are designed from the representative sequence of each of the 26,029 clusters. This chosen representative sequence is the sequence with the longest region of high-quality sequence in each cluster.

#### Number of EST and mRNA-aligned oligos

All 133,330 *A. thaliana* EST sequences available in the AtEST database, built on 2/20/02 from The *A. thaliana* Information Resource (TAIR) web site (<http://www.arabidopsis.org>), were downloaded. Using BLAST, 26,090 oligos were aligned with all ESTs in the AtEST database. Oligos with a hit to an EST at greater than 93% identity over the length of the oligo are categorized as EST aligned. For a 70mer oligo, a greater than 93% identity translates to greater than 65 matched bases. Percent identity is the number of matches in the BLAST alignment divided by the length of the oligo. A total of 9698 oligos have greater than 93% identity to an *A. thaliana* EST over the length of the oligo.

To obtain a set of mRNA-aligned or EST-aligned oligos, 26,090 oligos were aligned using BLAST to sequences from three different sources. *A. thaliana* UniGene component sequences with the word cDNA or mRNA were extracted. Also included were all ESTs from the AtEST database and 5000 Ceres-TIGR collaboration full-length cDNAs available at the TIGR web site (<http://www.tigr.org>). The number of oligos having greater than 93% identity to a full-length cDNA, mRNA, or EST is 14,331 oligos.

The result of this analysis shows that 14,331 probes have significant similarity to some transcript by having a high percent identity to an EST, mRNA, or full-length cDNA.

Number of EST-aligned oligos	9698
Number of mRNA-aligned or EST-aligned oligos	14,331
Total number of oligos in <i>A. thaliana</i> Genome Oligo Set	26,090

#### 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. Large numbers of 70mer candidate probes for each gene are selected using the following criteria for the *A. thaliana* Genome Oligo Set Version 1.0.

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

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

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

2) Each oligo is within 1000 bases from the 3' end of the available gene sequence.

3) An oligo cannot have a contiguous single nucleotide base repeat or poly (N) tract longer than 8 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 *A. thaliana* Genome Oligo Set, using BLAST, each oligo is aligned against all 26,062 representative sequences in *A. thaliana* UniGene Build At 4. 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 homology or similarity of the oligo.

This calculated percent identity score is dependent on the size of the sequence database used to BLAST against, oligo sequence, and the 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 cross-hybridization similarity.

Please note that for 3898 genes that did not yield oligos satisfying all the above criteria, certain rules were relaxed. For those genes, one or more of the following criteria may apply:

- Oligo has a longer predicted hairpin stem length
- $T_m$  less than  $73^{\circ}\text{C}$
- Probe location greater than 1000 bases from 3' end of the gene or ORF
- Greater than 70% cross-hybridization percent identity
- Contiguous base match greater than 20 bases to another gene

SUMMARY

Oligo Selection Criteria	Value	Number of Oligos in Genome Set Satisfying These Criteria
Length Melting temperature Location from 3' end Poly(N)tract length Stem length in potential hairpin Cross-hybridization to all other genes Contiguous base match to any other gene	70mer 78°C ± 5°C ≤ 1000 ≤ 8 ≤ 9 ≤ 70% ≤ 20	22,192
Length Melting temperature Location from 3' end Poly(N)tract length Stem length in potential hairpin Cross-hybridization to all other genes Contiguous base match to any other gene	70mer 78°C ± 5°C ≤ 1000 ≤ 8 ≤ 9 > 70% ≤ 20	2793
Length Melting temperature Location from 3' end Poly(N)tract length Stem length in potential hairpin Cross-hybridization to all other genes Contiguous base match to any other gene	70mer 78°C ± 5°C ≤ 1000 ≤ 8 ≤ 9 ≤ 70% > 20	189
Length Melting temperature Location from 3' end Poly(N)tract length Stem length in potential hairpin Cross-hybridization to all other genes Contiguous base match to any other gene	70mer 78°C ± 5°C ≤ 1000 ≤ 8 ≤ 9 > 70% > 20	69
Length	< 70	26
Stem length in potential hairpin	9 < x ≤ 15	31
Location from 3' end	Any	742
Contiguous base match to any other gene	> 20	291
Cross-hybridization to all other genes	> 70%	3067
Melting temperature	68°C < x < 73°C	48
<b>Total</b>		<b>26,090</b>

As mentioned in the Summary Table, 3067 oligos have greater than 70% crosshybridization identity to another gene. This is because a large number of genes in the *A. thaliana* genome have overlapping regions with each other. Using BLAST, all 26,029 genes were aligned against each other. The table below summarizes the number of overlaps and the percent identity.

Length of overlapping region between two <i>A. thaliana</i> genes	Percent identity of overlap	Number of genes
> 500 bases	> 95%	785
> 100 bases	> 95%	1628
> 50 bases	> 95%	1896
> 500 bases	> 90%	1434
> 500 bases	> 80%	4726

Furthermore, using BLAST, all 26,090 oligos were aligned against one another to obtain a percent identity of each probe versus all other non-self probes. The following table summarizes oligo versus oligo percent identity. Again the percent identity is the number of matched bases divided by the length of the oligo.

Percent identity of oligo versus another oligo	Cumulative number of oligos	Cumulative percent of oligos
≤ 30%	938	3.6
≤ 40%	18,821	72.1
≤ 70%	25,845	99.1
≤ 90%	26,078	99.95
100%	4	0.01

The following illustrations show the distribution of all 16,463 oligos 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

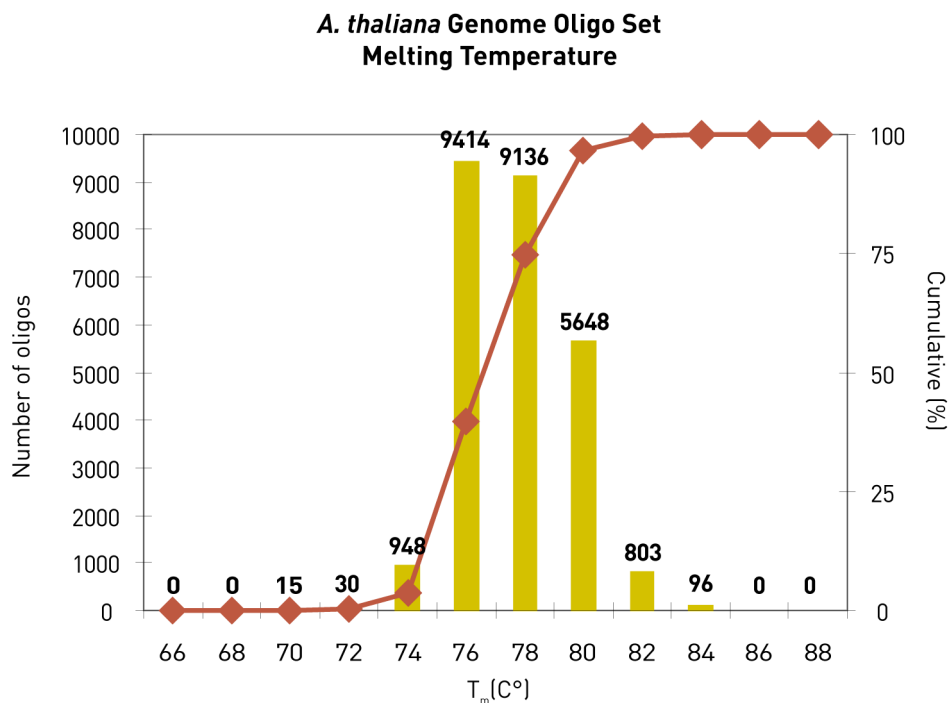


Figure 2. GC Content

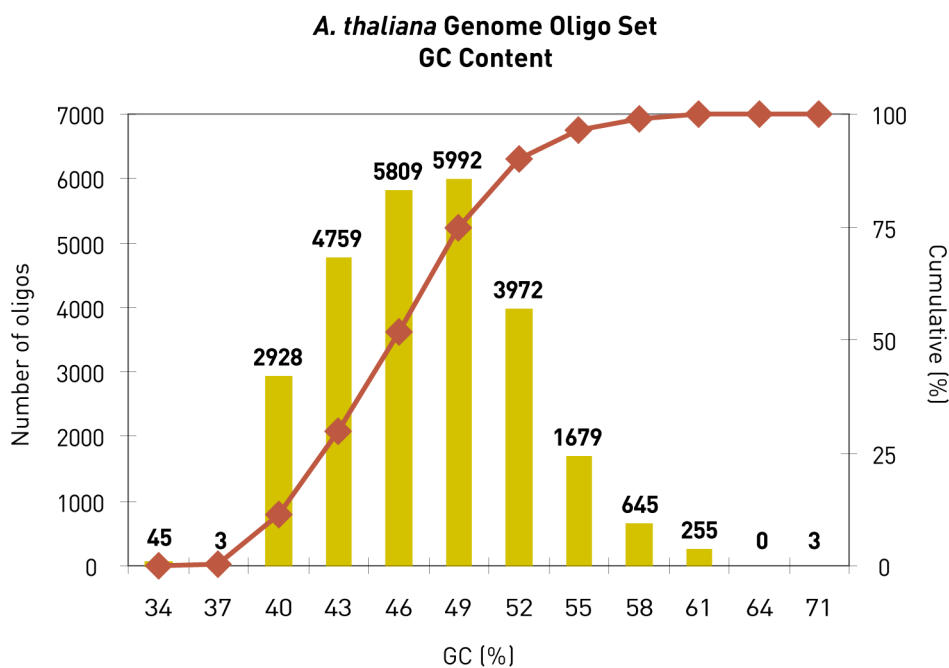


Figure 3. Location from 3' End

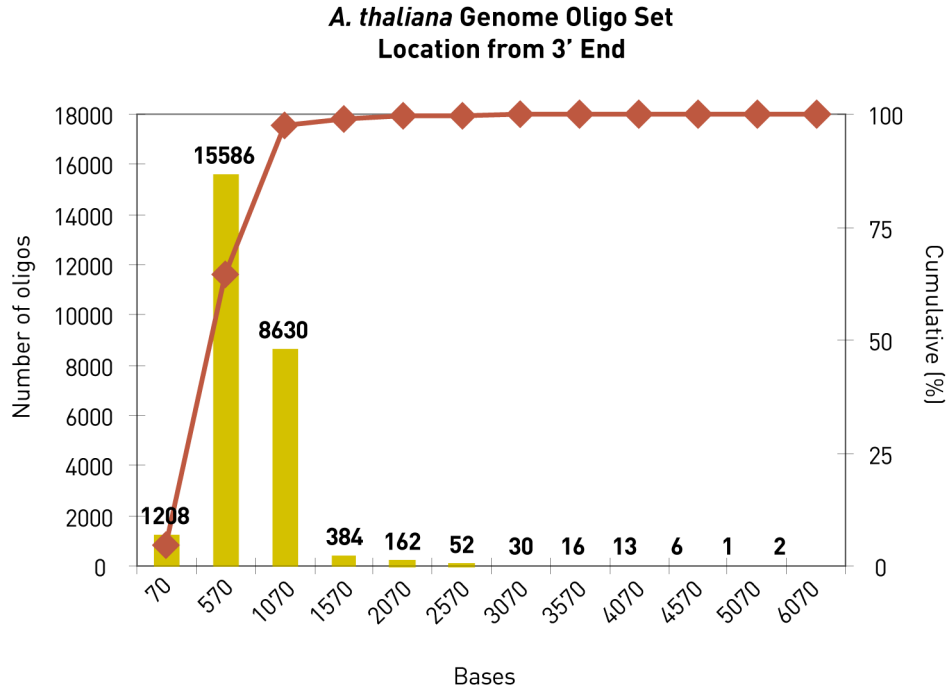


Figure 4. Length of the Longest Hairpin Stem

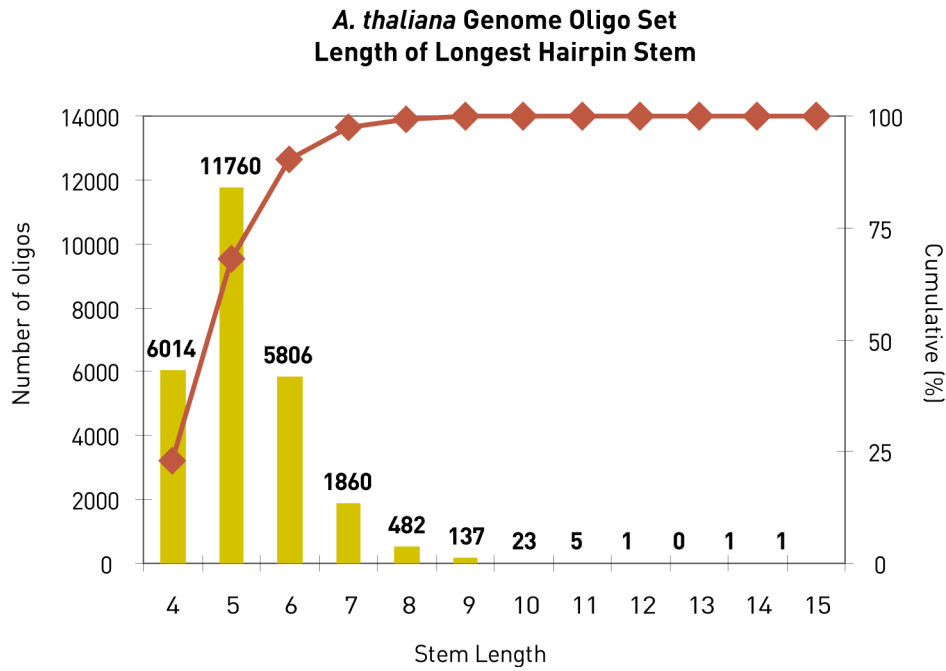
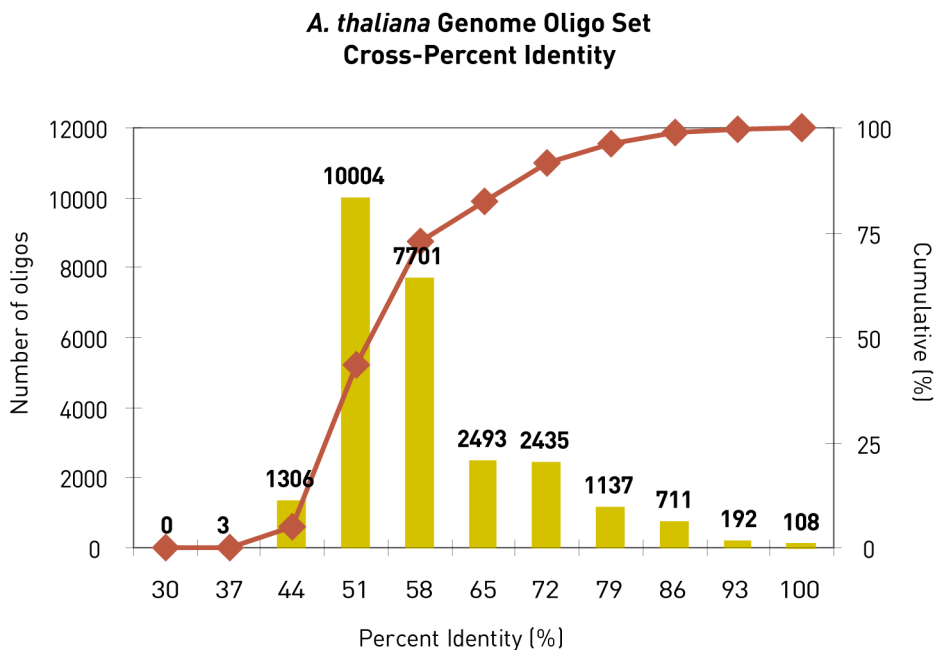


Figure 5. Cross-Hybridization Identity



Quality check of probe design specifications

Once the final oligo 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 26,090 oligos in the set.

Probe Design Specification	Expected Value	Verified Range	Number of Oligos
Melting Temperature (°C)	78°C ± 5°C	73.1 - 82.9	26,042
Melting Temperature (°C)	< 73°C	68 - 73	48
Hairpin Stem Length (base pairs)	≤ 9	4 - 9	26,051
Hairpin Stem Length (base pairs)	> 9	10 - 5	31
Cross-Hybridization Similarity (%)	≤ 70	37 - 69	23,023
Cross-Hybridization Similarity (%)	71 - 100	71 - 100	3,067