

Array-Ready Oligo Set™ for the *Candida albicans* Genome
Version 1.1

We are pleased to announce the release of our *Candida albicans* Genome Oligo Set Version 1.1 containing 6266 70mer probes. The majority of the 70mers, 5948 in total, are designed from *C. albicans*-predicted Open Reading Frames (ORFs). The remaining set of 318 oligos is designed from *C. albicans* gene sequences from GenBank. For our probe design methodology we use state-of-the-art methods and proprietary software. An amino linker is attached to the 5' end of each oligo.

Gene Sequence Source and Selection

Number of oligos designed from a predicted ORF from <i>C. albicans</i> ORF set Assembly 6 from Stanford University	5948
Number of oligos designed from a <i>C. albicans</i> gene sequence from GenBank	318
Total number of oligos	6266

A total of 5948 oligos are designed from the set of *C. albicans*-predicted ORF Set Assembly 6, which is developed, predicted, and maintained at Stanford University. Assembly 6 uses 10.4x coverage from whole genome shotgun sequencing of *C. albicans* strain SC5314. Sequence data for *C. albicans* were obtained from the Stanford Genome Technology Center website (<http://www.sequence.stanford.edu/group/candida>). Sequencing of *C. albicans* was accomplished with the support of the National Institute of Dental and Craniofacial Research (NIDCR) and the Burroughs Wellcome Fund.

The remaining set of 318 oligos is designed from gene sequences obtained from GenBank maintained by the National Center of Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov>). If the ORF is marked in the GenBank entry, then the oligo is designed within the ORF. Most of these gene sequences represent cloned genes.

Probe Design and Selection Rules

Once an ORF or 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 ORF or gene are selected using the following criteria for the *C. albicans* Genome Oligo Set Version 1.1.

Probe design and selection rules

- 1) All oligos are within 73°C ± 5°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 ORF or gene sequence.
- 3) An oligo cannot have a contiguous single nucleotide 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 *C. albicans* Genome

Oligo Set Version 1.1, using BLAST, each oligo is aligned against all 9168 predicted ORFs in the *C. albicans* ORF Set Assembly 6. Using the alignment with the candidate oligo versus the highest scoring non-self ORF, 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 the cross-hybridization identity 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 more 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 identity. Please note that for 525 ORFs and/or 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
- Greater than 70% cross-hybridization percent identity
- Contiguous base match length greater than 20 bases to another sequence

SUMMARY

Oligo selection criteria	Value	Number of oligos
Length	70mer	5741
Melting temperature	73°C ± 5°C	
Location from 3' end	≤ 1000	
Poly(N)tract length	≤ 10	
Hairpin stem length	≤ 9	
Cross-hybridization to all other sequences	≤ 70%	
Contiguous base match to another sequence	≤ 20	
Total number of oligos not satisfying one or more of the above criteria		525
Cross-hybridization to all other sequences	> 70%	334*
Contiguous base match to another sequence	> 20	443*
Hairpin stem length	9 < x ≤ 14	8*
Total		6266*

* Out of 525 oligos.

As mentioned in the summary table above, 334 oligos have a greater than 70% cross-hybridization identity. This is due to the following:

1) Some of the 318 Genbank sequences have large overlapping regions with some of the sequences in the set of 9168 *C. albicans*-predicted ORF Set Assembly 6, indicating that the predicted ORF for the GenBank sequence might be present. Using BLAST, these 318 genes sequences were aligned using the sense strand against the set of 9168 *C. albicans*-predicted ORFs in the sense strand. The table below summarizes the significant results.

Length of overlapping region between a GenBank gene sequence and an ORF	Number of GenBank gene sequences in this set
> 500 bases at > 98% identity	238*
> 100 bases at > 98% identity	258*
> 50 bases at > 98% identity	259*
Total	318

* Out of 318 gene sequences.

2) It is known that in the complete set of 9168 *C. albicans*-predicted ORF Set Assembly 6 has cases of predicted ORFs fully contained in another ORF or an ORF that has a large overlapping region with another ORF. Using BLAST, each ORF was aligned in the sense strand against all other ORFs in the sense strand in the set of 9168 Assembly 6 ORFs. The table below summarizes the significant results.

Length of overlapping region between two different ORFs	Number of ORFs
> 500 bases at 100% identity	128
> 100 bases at 100% identity	993
> 50 bases at 100% identity	1062
Number of total Assembly 6 <i>C. albicans</i> ORFs	9168

Therefore, it is possible that for certain cases in which we have reported a high cross-hybridization percent identity, the oligo could be "hitting" to a fragment of itself. Currently we are unaware of an existing non-redundant database for *C. albicans* that includes both cloned gene and predicted ORF sequences. Instead, BLASTing against this set of 9168 predicted ORFs currently presents the best solution.

The following illustrations show the distribution of all 6266 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 identity.

Figure 1. Melting Temperature

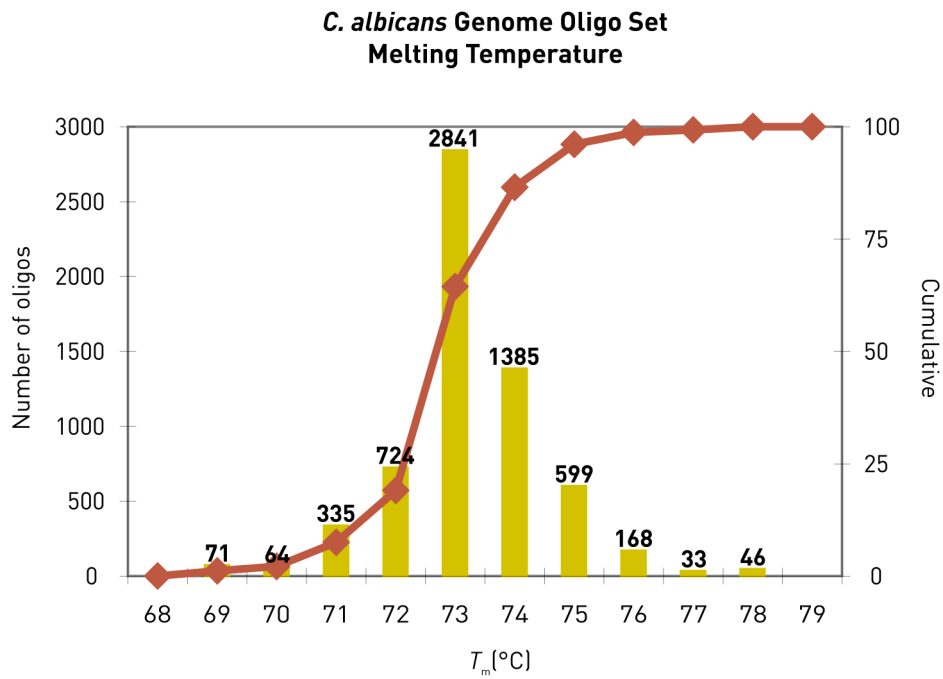


Figure 2. GC Content

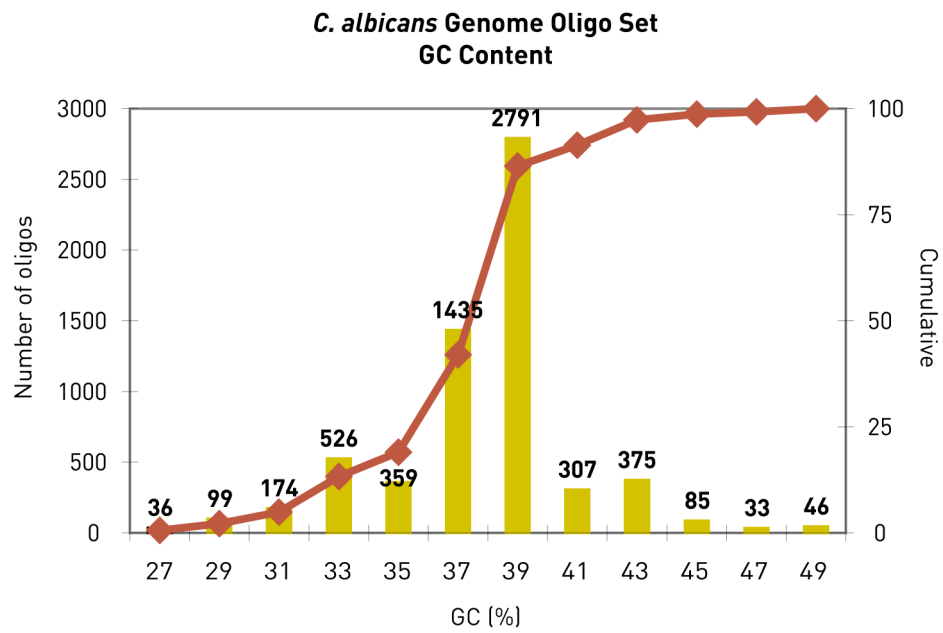


Figure 3. Location from 3' End

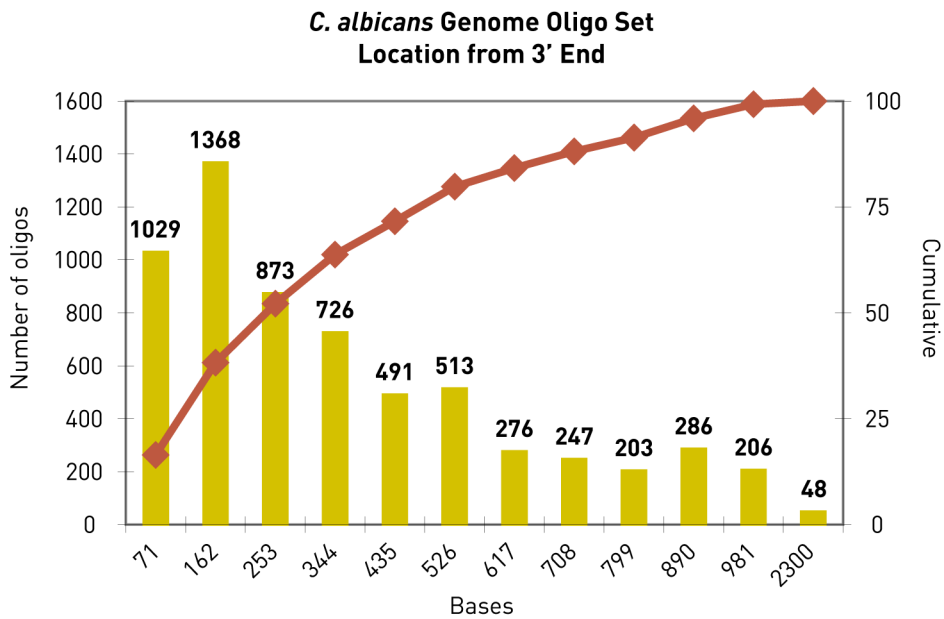


Figure 4. Length of the Longest Hairpin Stem

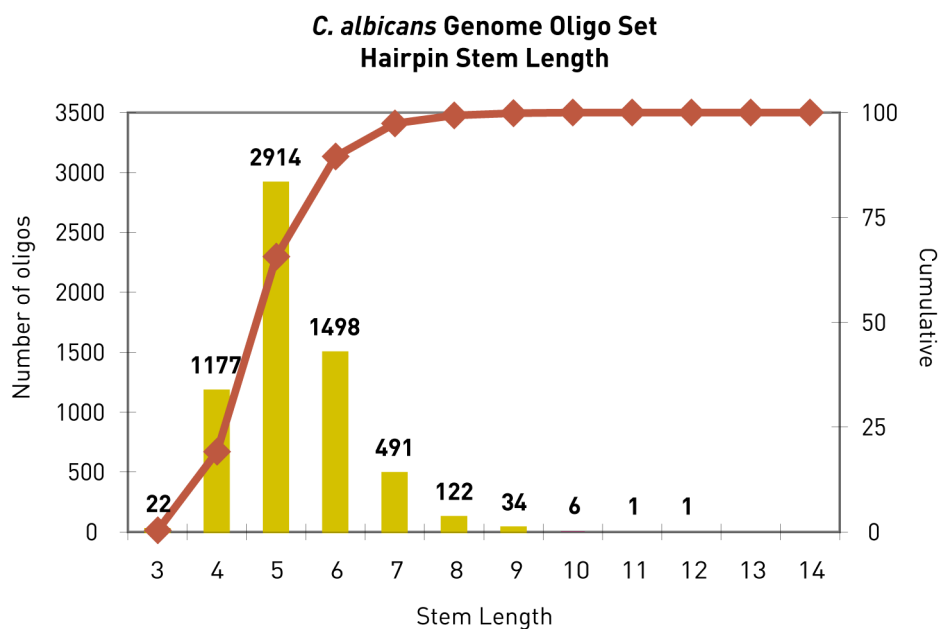
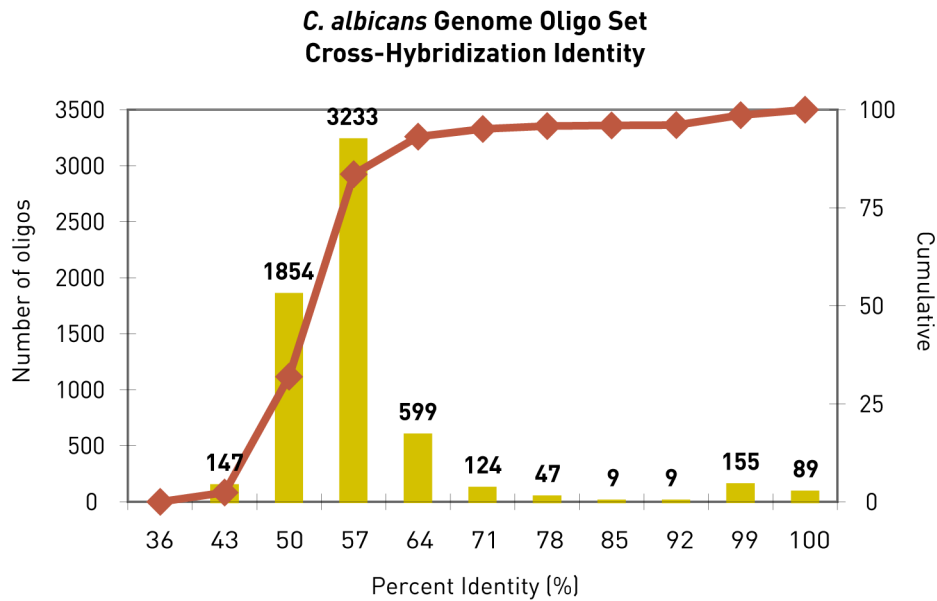


Figure 5. Cross-Hybridization Identity



Quality check of probe design specifications

Once the final oligo has been selected to represent an ORF or 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 6266 oligos in the set.

Table 3: Quality check probe design specifications

Probe Design Specification	Expected Value	Verified Range	Number of Oligos
Melting temperature [C°]	73° ± 5°C	68.3-77.7	6266
GC content (%)	25-50	25.71-48.57	6266
Location from 3' end (bases)	≤ 1000	70-993	6223
Location from 3' end (bases)	> 1000	1023-2230	43
Cross-hybridization identity (%)	≤ 70	37-70	5932
Cross-hybridization identity (%)	> 70	71-100	334