

Array-Ready Oligo Set™ for the Human Genome Version 2.1

We are pleased to announce Version 2.1 of the Human Genome Oligo Set containing 21,329 70mer probes representing 21,329 genes. Using refined probe design methods, we have formulated a new set of probes for most known human genes. This set represents many new genes not covered by our Human Genome Oligo Set Version 1.1, and incorporates more sophisticated design methods along with more gene sequence information. We also provide a map from the Human Genome Oligo Set Version 1.1 that corresponds to this new set.

Mapping Human Version 2.0 oligos to Ensembl 17.33 annotation

In order to provide customers with an integrated solution for the numerous versions of the Human Genome Oligo Set (Version 2.0, Version 2.0 Upgrade, and Version 2.1), we mapped the oligos from the Human Genome Oligo Set Version 2.0 to the current Ensemble annotation.

The original design of the Human Genome Oligo Set Version 2.0 is based on the Human UniGene Build 147. Mapping to the current Ensembl annotation is based on the following:

1. Basic Local Alignment Search Tool (BLAST) oligo sequences from the Human Genome Oligo Set Version 2.0 to various databases.

- a. Human Ensembl 17.33 database main gene transcript build — <http://www.ensembl.org>.
- b. Human Ensembl 17.33 database main gene transcript build with 3' end exon extended with 2000 base genomic sequence.
- c. UniGene Human 162 Build — <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>.
- d. Acembly.ncbi_33.transcripts - <http://www.ncbi.nih.gov/IEB/Research/Acembly/>.

2. If the oligo has greater than 97% identity to a transcript or gene sequence from three databases above, the oligo represents that transcript or gene. If the oligo matches to multiple transcripts or genes, only one ID is randomly chosen. Preference is given to a 100% match.

3. For oligos that map to Ensembl genes or transcripts, the annotation is based on Ensembl_mart_17_1. Oligos are further characterized as C, P, I (for details, see the datasheet for Human Genome Oligo Set Version 3.0) oligo_type based on the number of transcripts matched. For oligos that map only to Human UniGene Build 162, the annotation is based on Human UniGene Build 162. Only 987 oligos from the Human Genome Oligo Set Version 2.0 cannot be mapped to any current annotation (Ensembl, UniGene Build 162, Acembly transcript).

For details, please download the new genelist for the Human Genome Oligo Set Version 2.1 from the Oligo Microarray Database (OMAD) Download Center (<http://omad.qiagen.com/download/>).

Gene Sequence Source and Selection

The majority of the probes (21,322) are designed from the UniGene Database Build Hs 147 (February 2002) and the Human Reference Sequence (RefSeq) Database with the exception of 7 probes designed from the UniGene Database Build Hs 152. These databases are developed and maintained at the National Center of Biotechnology Information

(NCBI) (<http://www.ncbi.nlm.nih.gov>).

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 nonredundant set of gene-oriented clusters. It has become one of the most widely used de facto standards in the public domain for cataloguing human genes. Each UniGene cluster contains cloned genes and expressed sequence tag (EST) sequences that represent a unique gene. All oligos are designed from the representative sequence of each of the 21,329 clusters. This chosen representative sequence is the longest sequence in each cluster. UniGene representative sequences are filtered for contaminant sequences, genomic repetitive regions, and low complexity sequences using NCBI's Dust program.

The NCBI Reference Sequence (RefSeq) project is an effort to standardize gene sequence references by providing an 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 supported by a large research community.

The advantages of designing a probe directly from a gene sequence from the RefSeq collection are given below. Each of these advantages is provided by the NCBI LocusLink Interface.

| Features provided for each RefSeq | Example |
|---|--|
| Alternate gene symbols | VEGFA |
| Full-length coding regions known | Present |
| Gene Ontology™ (GO) terms | 1) Molecular function: soluble fraction, vascular endothelial growth factor receptor ligand 2) Biological process: stress response, homophilic cell adhesion, positive control of cell proliferation 3) Cellular component: extracellular, stress response |
| Chromosome/Cytogenetic map information | 6p12 |
| Official gene symbol and name by Human Gene Nomenclature Committee (HGNC) | Vascular endothelial growth factor (VEGF) |

Furthermore, curated gene sequences undergo manual review by NCBI staff and offer additional advantages. The following corrections are made to curated gene sequences.

- Vector and linker sequence removed
- Chimeric regions removed
- Untranslated regions extended if possible
- Ambiguous bases corrected

In the Human Genome Oligo Set Version 2.1, 11,530 oligos are designed from a gene sequence from RefSeq and 3002 oligos are designed from curated or reviewed RefSeqs.

| | |
|--|--------|
| Number of oligos using information and provisional or predicted sequence from UniGene and RefSeq | 8528 |
| Number of oligos using information and curated sequence from UniGene and RefSeq | 3002 |
| Subtotal | 11,530 |
| Number of oligos designed from UniGene | 9799 |
| Total number oligos in Human Genome Oligo Set Version 2.1 | 21,329 |

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 Human Genome Oligo Set.

1) All oligos are within $76 \pm 5^\circ\text{C}$ 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 repeat or poly (N) tract longer than 7 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 Human Genome Oligo Set Version 2.1, using BLAST, each oligo is aligned against all 96,073 representative sequences in Human UniGene Build Hs 147. 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 use of either gapped or nogap 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.

For a small number of genes that did not yield oligos satisfying all the above criteria, certain rules were relaxed. For those genes, the oligo is selected anywhere in its sequence or is designed to be less than 70 bases long.

Figure 4. Length of the Longest Hairpin Stem

| 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 ≤ 7 ≤ 9 ≤ 70% ≤ 20 | 21,013 |
| 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 | 50 to 60mer 78°C ± 5°C Any ≤ 7 ≤ 9 ≤ 70% ≤ 20 | 316 |
| Total | | 21,329 |

The following illustrations show the distribution of all 21,329 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

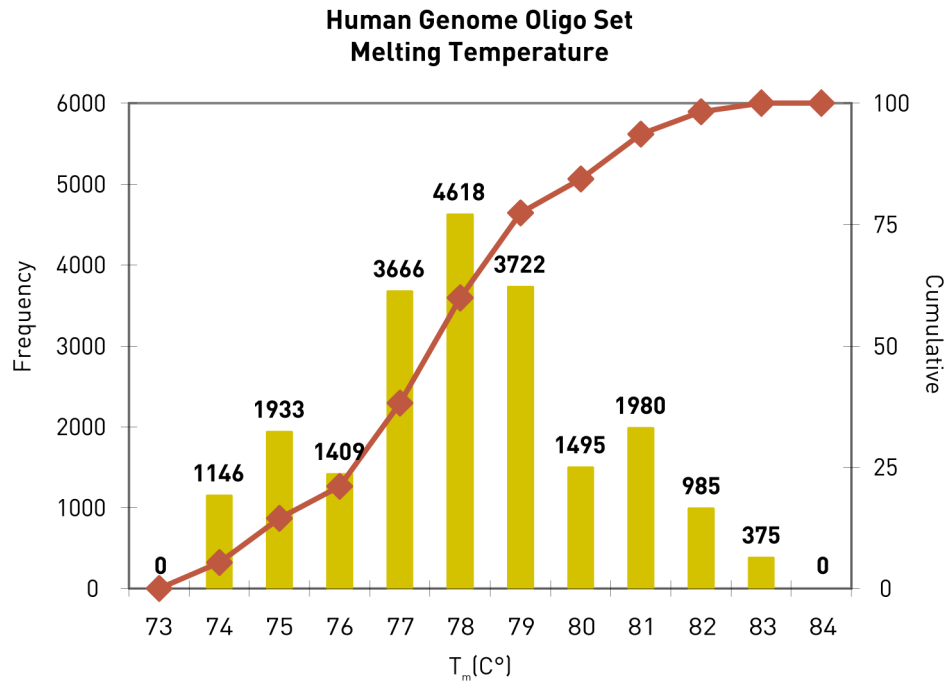


Figure 2. GC Content

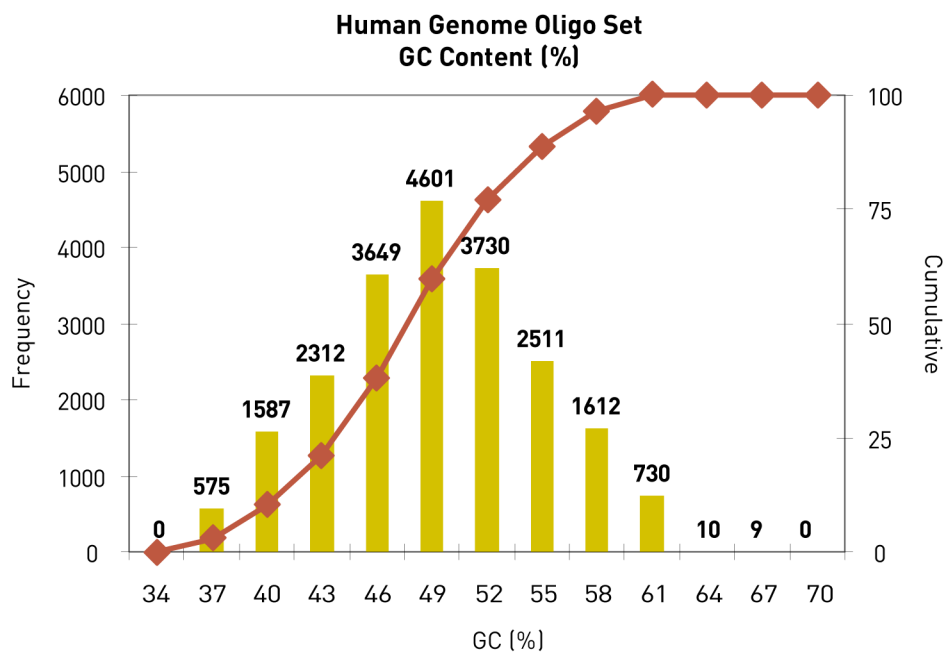


Figure 3. Location from 3' End

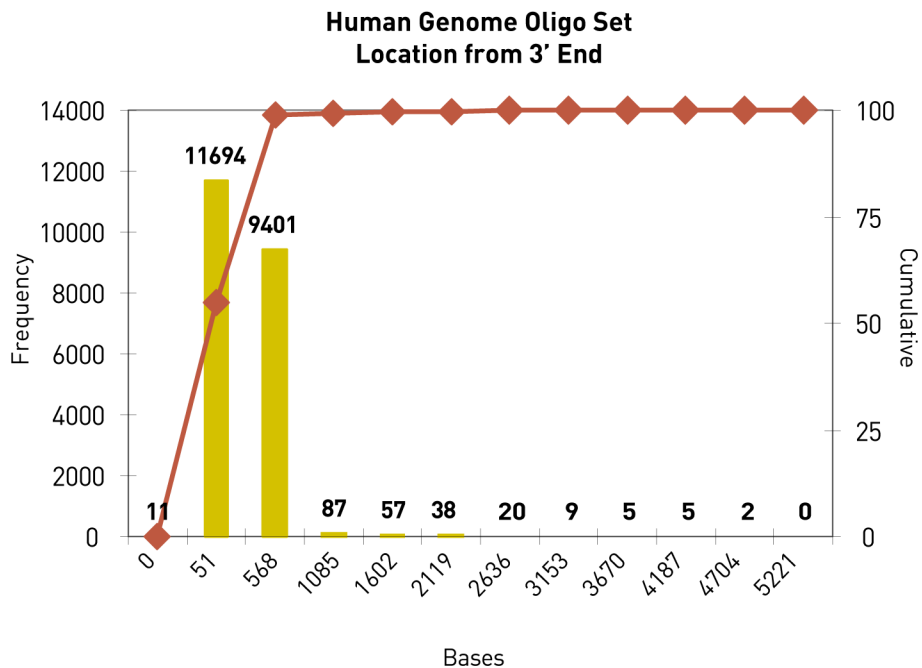


Figure 4. Length of the Longest Hairpin Stem

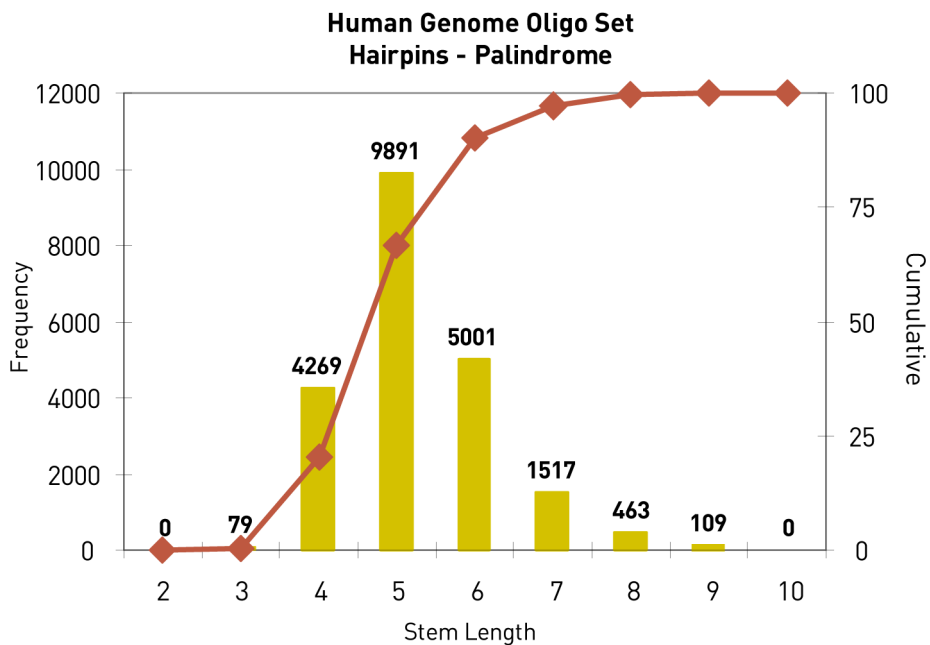
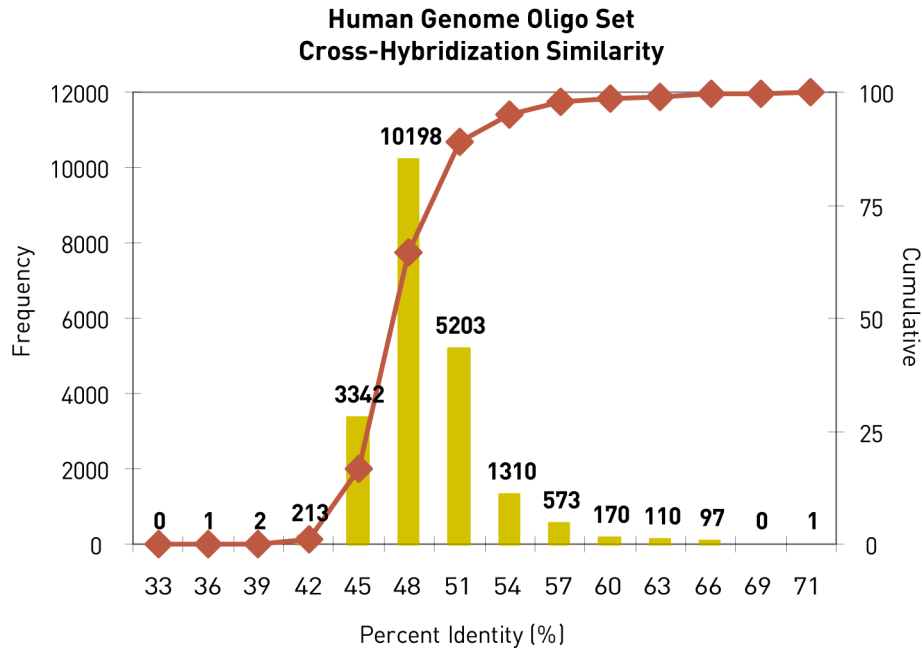


Figure 5. Cross-Hybridization Similarity



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 21,329 oligos in the set.

| Probe design specification | Expected value | Verified range |
|------------------------------------|----------------|----------------|
| Melting temperature [C°] | 78°C ± 5°C | 73.1–82.8 |
| GC content (%) | 35–70 | 37.7–68.0 |
| Hairpin stem length (base pairs) | ≤ 9 | 3–9 |
| Cross-hybridization similarity (%) | ≤ 70 | 36–70 |