Array-Ready Oligo Set™ for the Rat Genome
Version 3.0

We are pleased to announce Version 3.0 of the Rat Genome Oligo Set containing 26,962 longmer probes representing 22,012 genes and 27,044 gene transcripts. Pseudogenes are excluded from oligo design. A total of 2,413 oligos from Rat Genome Oligo Set Version 1.1 are included in the Rat Genome Oligo Set Version 3.0. The design is fully based on the Ensembl Rat Database Version v19.3b.2 (http://www.ensembl.org/) and Rat Genome Project, an international collaboration to sequence the genome of the brown rat (*Rattus norvegicus*). This approach allows detection of alternative splicing variants using common, partial common, or individual transcript oligos. For 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 and Transcript Sequence Source and Selection

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, mouse, rat, and several other eukaryotic genomes. The success of array oligo design hinges on the quality and annotation of sequence information. The Human, Mouse, Rat, etc. Genome Sequencing Projects produced a highly accurate and contiguous sequence (1). Ensembl can be used as either an interactive website complete with sequence visualization tools or downloaded as flat files or tables for local installation.

Advantages of Using Ensembl

Excluding pseudogenes, the Ensembl Rat Database Version v19.3b.2 contains 22,159 genes, 28,545 transcripts, and 183,320 exons and is fully based on the Rat Genome Sequencing Project. The following are the advantages of using the Ensembl database:

- *Alternative splicing variants* – There are 3,723 genes in the Ensembl Rat Database Version v19.3b.2 that have more than one transcript. The Ensembl database provides the best non-redundant set of genes with exon and intron coordinates.
- *Comparative genome analysis* – Complete ortholog information between human, mouse, and rat genomes is provided. In addition, large-scale synteny and crossspecies genome analysis data are accessible through the Ensembl database.
- *Additional functional annotation* – Additional functional annotation from the InterPro database and Ensembl protein family annotation are accessible through Ensembl.
- *Portability* – Ability to locally install Ensembl, including all software, source code, and sequence data. All components of Ensembl are open source and freely available.
- Provides mapped external identifiers, such as the National Center for Biotechnical Information (NCBI) RefSeqs, GenBank/EMBL accessions, LocusLink, Swiss-Prot, TrEMBL, and identifiers of other microarray platforms for many genes and transcripts.
- The Ensembl genome browser includes MapView, ContigView, GeneView, TransView, and ProteinView for zooming in on any region of the genome.

- EnsMart (http://www.ensembl.org/EnsMart/) permits fast retrieval of integrated gene/protein annotation, disease information, expression data, single nucleotide polymorphism (SNP), and cross-species analyses by external identifiers.

Additional information can be found at the Ensembl website (http://www.ensembl.org) and in the articles by Hubbard et al. (2) and M. Clamp et al. (3).

Differences between UniGene and Ensembl

The previous Rat Genome Oligo Set Version 1.1 is based on the UniGene database. UniGene automatically clusters GenBank sequences into a non-redundant set of geneoriented clusters. Each UniGene cluster contains cloned genes and expressed sequence tag (EST) sequences that represent a unique gene. This chosen representative sequence is the longest sequence in each cluster. UniGene design is based on the one cluster, one gene scheme. The following table outlines certain differences between the two platforms.

Database Features	UniGene	Ensembl
Transcriptome information	Yes	Yes
Genomic gene structure (exon/intron)	No	Yes
Alternative splice variants information	No	Yes
Comparative genomic synteny view	No	Yes
SNP information	No	Yes

The main disadvantage for design based on UniGene is that exon/intron gene structure and splicing variants information are not present to be taken into account in the oligo design.

Probe Design and Selection Rules

The concept of an "exon oligo," which is an oligo fully contained in an exon, was previously used to experimentally identify alternative splice variants in the human genome in a study by Shoemaker et al. (4). A "transcript oligo" is an oligo contained in multiple exons. These two oligo types are further classified below as three different types depending on number of transcripts represented: "common oligo", "partial common oligo" and "individual transcript oligo". These three oligo classifications are essential for differentiating alternative splice variants and maximizing the number of represented transcripts. The common oligo type is used for representing all transcripts of one gene. The partial common oligo classification is present for those genes when a common oligo or an individual transcript oligo could not be found. The design platform makes use of these oligo type classifications.

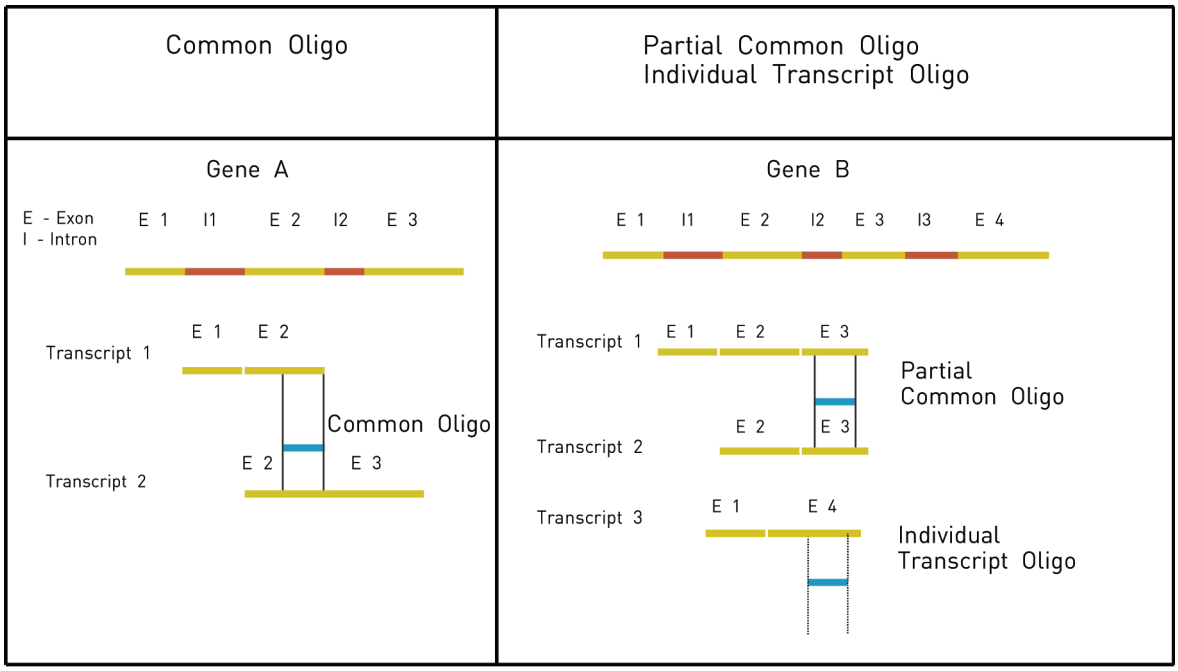
Using Operon's array-ready oligo design platform, optimal oligos were selected within each exon as exon oligos. For certain transcripts, if exon oligos could not be identified, optimal oligos that cover multiple exons were chosen as transcript oligos.

Oligo Type	Definition	Number of Oligos
Exon oligo	Probe fully contained in one exon	22,167
Transcript oligo	Probe contained in multiple exons	4,795
Total		26,962

As mentioned before, exon and transcript oligos are further characterized as one of three oligo types depending on the number of transcripts represented: common oligos, partial common oligos, or individual transcript oligos. A common oligo is defined as an oligo that represents all transcripts of one gene. A partial common oligo is defined as an oligo that represents a subset of transcripts of one gene. An individual transcript oligo is defined as an oligo that represents only one transcript of one gene.

Oligo Type	Oligo Type Symbol	Definition	Number of Transcript Oligos	Number of Exon Oligos
Common Oligo	C	The oligo represents all transcripts of one gene	375	2,224
Partial Common Oligo	P	The oligo represents a subset of transcripts of one gene	31	733
Individual Transcript Oligo	I	The oligo represents only one transcript of one gene	4,389	19,210

Figure 1. Illustration of common, partial common, individual transcript oligos for two different genes



Non-self transcripts for common oligos are all transcripts of other genes. Non-self transcripts for partial common oligos are all transcripts not represented by the oligo. Non-self exons are all exons other than the exon used for oligo design. Non-self transcripts for individual transcript oligos are simply all transcripts other than the transcript used for oligo design. These classifications of self and non-self transcripts/exons are used below for computing certain design criteria.

Sufficient numbers of 70mer candidate probes for each exon and each gene transcript are selected using the following criteria:

1) All oligos are within 78°C ± 5°C using the following formula:
T_m = 81.5 + 16.6 x log[Na+] + 41 x (#G + #C)/length – 500/length where [Na+] = 0.1 M and length = #A + #C + #G + #T

2) Each oligo is within 2000 bases from the 3' end of the available transcript 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 exon oligo has less than or equal to 70% identity to all other non-self exons. Exon oligos, using BLAST, are

aligned against all 183,320 exon sequences in the Ensembl Rat Database Version v19.3b.2. Using the alignment with the candidate oligo versus the highest scoring non-self exon, a cross-hybridization identity score is computed. The highest scoring non-self exon is defined as the exon sequence that yields the most matched bases in an alignment. A non-self exon is defined above.

Each transcript oligo has less than or equal to 70% identity to all other transcripts. Using BLAST, each transcript oligo is aligned against all 28,545 transcripts in Ensembl. Similarly, a cross-hybridization identity score is computed versus the top non-self transcript. A non-self transcript is defined above.

Certain genes and transcripts are highly homologous to each other. In order to represent a highly homologous gene or transcript in the genome oligo set, the cross-hybridization identity criterion was gradually relaxed. Oligos that meet the other selection criteria with the lowest possible cross-hybridization identity were included. For details, please see the Summary table and Figure 7, below.

7) Each exon oligo cannot have greater than 20 contiguous bases common to any non-self exons. Each transcript oligo cannot have greater than 20 contiguous bases common to any non-self transcripts.

For a number of exons and transcripts that did not yield oligos satisfying all the above criteria, certain rules were relaxed. Certain selection rules such as oligo length, T_m, location, cross-hybridization identity, and contiguous bases were relaxed.

Once oligo candidates have been selected satisfying the selection rules mentioned above, each oligo is ranked based on cross-hybridization identity.

A final exon oligo is selected based on the above criteria and the exon location. Each final exon oligo is preferentially selected first from the 3' most exon. If the 3' most exon fails to yield a successful exon oligo, then a successful exon oligo is selected scanning the various exons from the 3' end towards the 5' end of gene.

For each final exon oligo, a cross-hybridization identity and contiguous bases score versus non-self transcripts is also computed. Non-self transcripts for common oligos are all transcripts of other genes. Non-self transcripts for partial common oligos are all transcripts not represented by the oligo. This cross-hybridization identity and contiguous bases score is reflected in the tables and graphs below.

By selecting a combination of common oligos, partial common oligos, and individual transcript oligos from both exon and transcript oligos based on minimized cross-hybridization identity scores, this set was designed to differentiate alternative splicing variants.

A summary of the selection criteria is shown in the table below. Complete data for the 3' end location criteria is shown in Figure 5.

SUMMARY

Oligo Selection Criteria	Criteria Values	Number of Oligos in Rat Genome Oligo Set Version 3.0
Oligo Length‡ (See Figure 2)	50, 60, 70mer	26,962
Melting Temperature	78°C ± 5°C	26,962
Poly(N)tract Length	≤ 8	26,945
Stem Length in Potential Hairpin	≤ 9	26,955
Cross-hybridization identity to all other transcripts§	≤70%	24,193
Contiguous base match to any other transcript§	≤20	24,278
Poly(N)tract Length‡ (See Figure 2)	›8	17*
Stem Length in Potential Hairpin§	›9	7*
Cross-hybridization identity to all other transcripts§	›70%	2,769*
Contiguous base match to any other transcript§	›20	2,684*
Number of oligos not satisfying one or more of the above criteria		3,681
Total		26,962

**Out of 3,681 probes.*
§For common oligos, the top non-self transcript is always a transcript of another gene. For a partial common oligo, the top non-self transcript can be any transcript other than the transcripts represented by the partial common oligo.
‡Amino linker is not counted in oligo length.

The following illustrations show the distribution of all 26,962 oligos representing the Rat Genome Oligo Set Version 3.0 for oligo length, melting temperature, GC content, location from 3' end, longest stem length, and cross-hybridization identity.

Figure 2. Oligo Length - Rat Genome Oligo Set Version 3.0

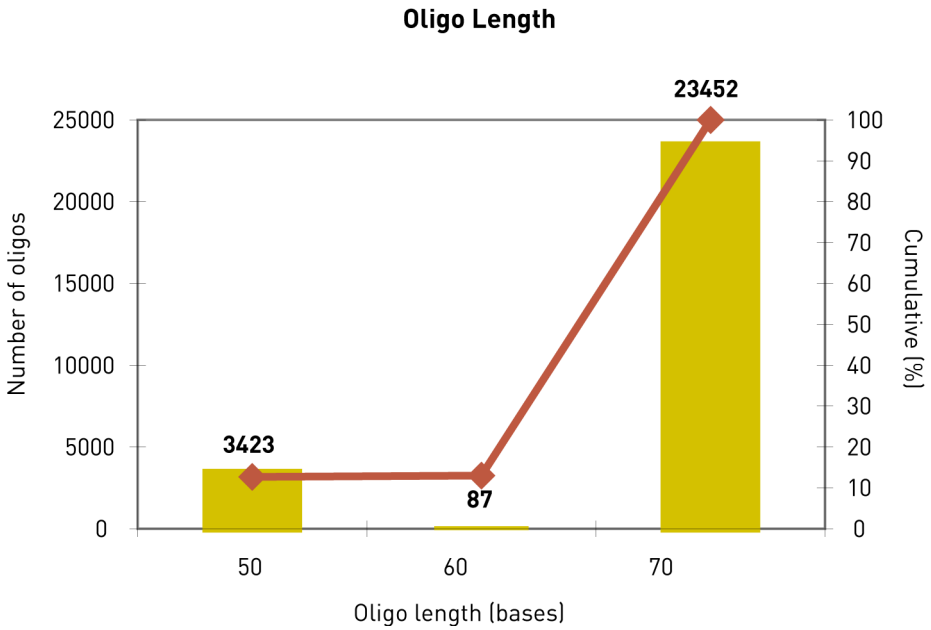


Figure 3. Melting Temperature - Rat Genome Oligo Set Version 3.0

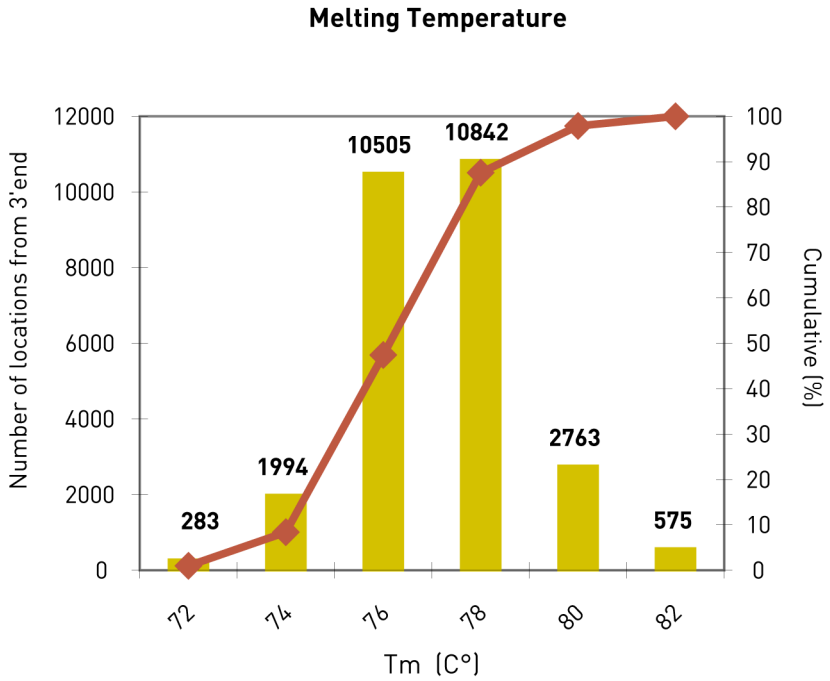


Figure 4. GC Content - Rat Genome Oligo Set Version 3.0

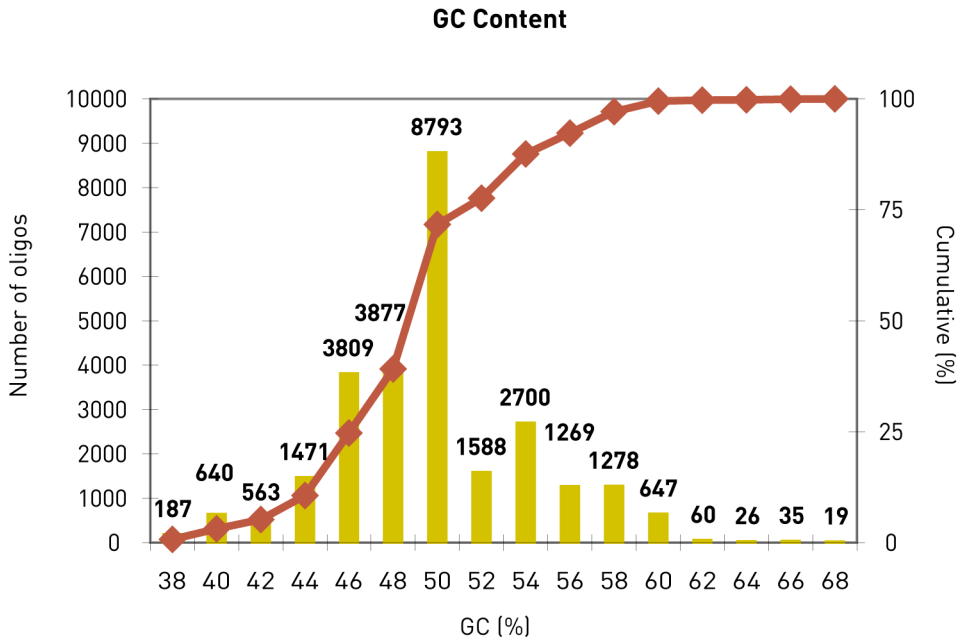
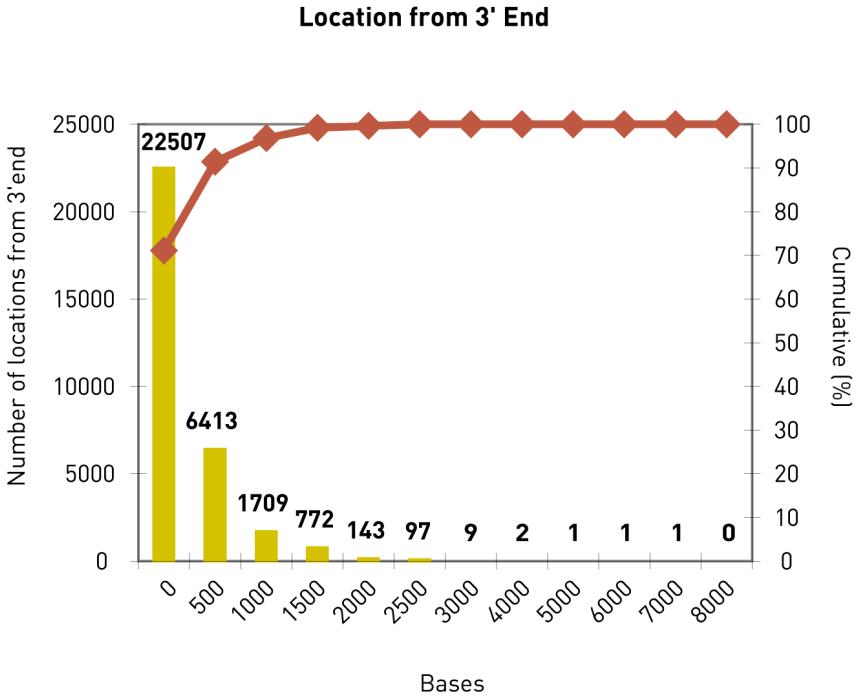


Figure 5. Location from 3' End - Rat Genome Oligo Set Version 3.0



Common and partial common oligos have multiple locations from 3' end shown as they represent multiple transcripts. Individual transcript oligos have only one location from 3' end shown. Total number of 3' end locations shown is 31,656.

Figure 6. Hairpin Stem Length - Rat Genome Oligo Set Version 3.0

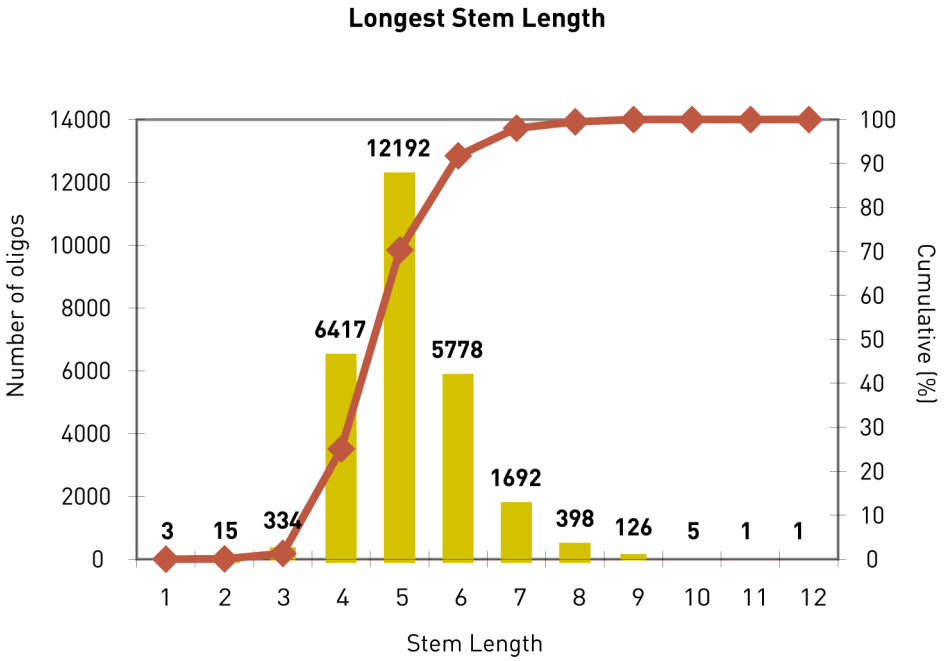
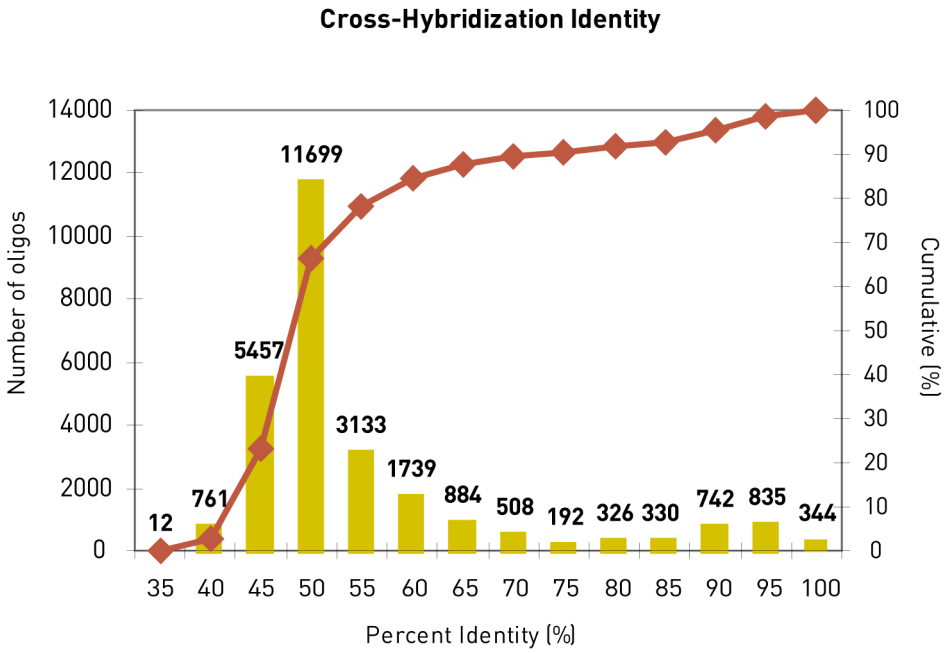


Figure 7. Cross-Hybridization Identity - Rat Genome Oligo Set Version 3.0



Quality Check of Probe Design Specifications

Once the final oligo set has been selected to represent a gene transcript, 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 26,962 probes.

Probe Design Specification	Expected Value	Verified Range	Number of Oligos in Rat Genome Oligo Set Version 3.0
Melting Temperature (C°)	78°C ± 5°C	72.0-85.0	26,962
GC Content (%)	35-70	39.0-68.0	26,962
Poly(N)tract length (base pairs)	≤8	1-10	26,945
Poly(N)tract length (base pairs)	›8	9-10	17
Hairpin stem length (base pairs)	≤9	1-12	26,955
Hairpin stem length (base pairs)	›9	10-12	7
Cross-hybridization identity (%)	≤70	35-100	24,193
Cross-hybridization identity (%)	›70	71-100	2,769
Contiguous bases (base pairs)	≤20	1-70	24,278
Contiguous bases (base pairs)	›20	21-70	2,684

References

1. *Rat Genome Sequencing Consortium assembles rat genome, HOUSTON (Nov. 25, 2002) - http://public.bcm.tmc.edu/pa/rgsc-genome.htm.*
2. *Hubbard, T. et al. (2002) The Ensembl genome database project. Nucleic Acids Res. 30, No. 1, 38–41.*
3. *Clamp, M. et al. (2002) Ensembl 2002: Accommodating comparative genomics. Nucleic Acids Res. 31, No. 1, 38-42.*
4. *Shoemaker, D.D. et al. (2001) Experimental annotation of the human genome using microarray technology. Nature. 409, 922–7.*