Technical Note

Roche NimbleGen Probe Design Fundamentals

The ultimate success of an experiment often depends on the effort expended during the initial design. This is especially true for *in silico* microarray probe selection. This document describes probe selection strategies used for the majority of NimbleGen microarray tiling and transcript-based designs and describes the following criteria:

- Synthesis cycles
- Melting temperature
- Repetitive elements
- Uniqueness measures

General Concepts

Roche NimbleGen generates two basic types of designs:

- Tiling – Genome-centric designs with all probes mapping to the genome. Tiling designs are generally used for CGH, ChIP-chip, DNA methylation, and expression tiling applications.
- Transcript-based – Gene-centric designs with all probes mapping to specific transcripts. Transcript-based designs are used for gene-based expression applications.

The approaches used to generate these designs share several fundamental methods, described below.

Synthesis Cycles

Oligonucleotide synthesis on NimbleGen arrays is limited to a finite number of synthesis cycles based on the length of the probe. In general, synthesis cycles are limited to three times the probe length, rounded down to nearest multiple of 4. Most long oligo designs are limited to the synthesis cycle associated with a 50mer design.

Probe Length	Max. Synthesis Cycles
24	72
36	108
50	148
60	148

Probes are synthesized from 3′ → 5′. The 3′ end of all probes is adjacent to the array surface. Synthesis order is always A, C, G, then T, with each base in the order constituting a synthesis cycle. Therefore, the least expensive 4mer probe to synthesize is 5′-TGCA-3′ and the most expensive is 5′-TTTT-3′ with the former requiring 4 cycles and the latter requiring 16 cycles.

Melting Temperature

Roche NimbleGen uses several different methods of calculating oligonucleotide melting temperature (T_m). Methods used depend on the length of oligo, the type of application, and the conditions used for the hybridization.

Repetitive Elements

To avoid non-specific binding, highly repetitive elements of the genome or transcript sets are often excluded from microarray designs. As an alternative to the RepeatMasker program (1), Roche NimbleGen has developed a new method that utilizes a strategy similar to the WindowMasker program, developed by Margolis (2006), to identify these regions and exclude them from probe selection (2).

Roche NimbleGen's method is based on the frequency of 15mers within the target genome. (*Note: Smaller, less complex genomes may use shorter nmers.*) The process compares the set of probes against a pre-computed frequency histogram of all possible 15mer probes in the target genome. For each probe, the frequencies of the 15mers comprising the probe are then used to calculate the average 15mer frequency of the probe.

The higher the average 15mer frequency, the more likely the probe is to lie within a repetitive region of the genome. For large eukaryotic genomes, only probes with an average 15mer frequency of less than 100 are used. For prokaryotic genomes and smaller eukaryotic genomes, this threshold is reduced. This method results in better coverage of the genome, as compared to the conventionally used RepeatMasker program, while still avoiding highly repetitive regions.

Uniqueness Measures

Roche NimbleGen defines two terms to describe the cross-hybridization potential of any given probe: frequency and careful uniqueness. There is a subtle but significant difference between *frequency* and *careful uniqueness*. For example, an oligo may appear as an exact duplicate hundreds of times within a genome but be significantly distant from all other oligos. An example of this might be an oligo that lies in a conserved domain in a large gene family. That oligo may uniquely identify members of that gene family but will not cross-hybridize with any other genes in the transcriptome. Conversely, an oligo can be carefully unique in the genome but be within just a few mismatches of many other oligos. Again, using the example of a large gene family, an oligo may exactly match only one member of a gene family, but contain only one or two nucleotide differences from a number of other genes in the same family. Therefore, both metrics are calculated using a proprietary application and used in the final probe selection process.

Frequency. Frequency, also referred to as count or perfect matches, is defined as the number of times an oligo of a given size appears exactly within a set of comparison sequences. For tiling designs, this comparison is usually against the genomic sequence for the target organism. Probes generated for transcript-based designs are generally compared to the transcriptome.

Careful Uniqueness. Like frequency, careful uniqueness is based on a comparison of an oligo against the source transcriptome or genome. Careful uniqueness, however, is evaluated based on the existence of close matches rather than exact matches. In addition, the method for calculating careful uniqueness depends on the type of design being generated.

Tiling Designs – Uniqueness for tiling designs is not a Boolean measure as for transcript-based designs but rather a count of the number of close matches to the oligo in the genome. The method for determining uniqueness for transcript-based designs is, in general, too time-intensive for longer oligos. Therefore, the uniqueness information is generated using the SSAHA program (3), a software tool for rapid matching and alignment of DNA sequences. Each oligo is mapped to the genome and a count is kept of the number of exact matches (perfect matches) and close matches in the genome. Roche NimbleGen uses a word length of 12 and a step size of 1 so that single nucleotide differences can be found, even in short oligos. Generally the minimum match size is set to the minimum oligo length minus the word length, so for a 50mer oligo, the minimum match size allowed would be 38 nucleotides. Up to 5 insertions/deletions are generally allowed, as well as a maximum gap size of 5 nucleotides.

Transcript-Based Designs – Uniqueness for a transcript-based design is a Boolean measure that classifies an oligo as either similar or dissimilar to other oligos of the same size in the transcriptome. Each oligo of a given size is compared to all other oligos of that size in the transcriptome, and a weighted score is calculated based on the number and position of mismatches between the two oligos. The resulting score is compared to a set threshold, and if the score exceeds the threshold, the oligo is classified as unique (or carefully unique).

The default weighting scheme is trapezoidal in shape, giving greater weight to mismatches in the center of the oligo than to mismatches at either end. The default threshold for being considered carefully unique is a weighted mismatch score greater than 10. If in any comparison the score is less than 10, the oligo fails and is declared to be non-unique.

Probe Selection for Tiling Designs

Tiling designs are generally the simplest designs to create, as they are normally an attempt to evenly space probes over a genomic region of interest. The most basic tiling design consists of tiling through a region of interest at a fixed interval, generating probes of specific length, T_m, or synthesis cycle characteristics. Probes falling into repetitive regions are eliminated and the resulting set of probes can be used to create a design.

In most cases, however, filtering the probe set is recommended. The filtering process requires generating a pool of candidate probes and then applying filters based on repetitiveness, uniqueness, T_m, and base pair composition. The initial pool of probes is generated by tiling through the genome as described above.

This tiling interval, *I*, the start-to-start distance from the 5′ end of one probe to the 5′ end of the next probe, must be determined beforehand. The upper bound on *I* is calculated as

$$I = \#\_Bases\_Covered / \#\_Features$$

but the final value of *I* will likely be smaller, indicating a denser tiling.

This difference is due to characteristics of the genomic sequence, like repetitive elements, ambiguous sequence, and non-unique probes. The recommended tiling interval ranges are shown in the table (right).

Product	Lower Bound	Upper Bound
ChIP-chip	10	100
DNA Methylation	10	100
CGH	10	6,000

Tiling intervals less than 10bp are not recommended. If selection criteria (15mer frequency masking, uniqueness, T_m, base pair composition) are to be applied during the selection process, the candidate probe set must be over-selected by a factor of 10 to support filtering on these criteria down to an appropriately sized probe set.

As outlined above, the average 15mer frequency for each probe is calculated by comparison with the source genome. For large eukaryotic genomes, any probe with a frequency greater than 100 is automatically excluded at this step. Following the frequency masking step, the uniqueness of the probes is calculated using SSAHA.

The final probe selection for tiling designs is done using window-based rank selection. For each probe within the window, a score is calculated from the frequency, T_m, and uniqueness information. The best probe, or probes, within the window are then selected for inclusion in the final probe set. After a probe is selected within a window, the window is moved forward *I* bases and the ranked selection process is repeated.

How the tiling interval and window size are apportioned depends on the type of design and user requirements. Maximizing the window size allows for the most over-sampling and the selection of the best probe out of the maximum size candidate pool. This comes at the expense of regular spacing.

Selecting a large interval and a small window size ensures a more regular probe spacing at the cost of reducing the size of the candidate probe pool for each window.

Probe Selection for Transcript-Based Designs

Transcript-based designs begin as tiling designs with one major difference: oligos are generated from the transcripts rather than the genomic sequence.

Regardless of the final oligo length, the full set of 24mer oligos tiling a transcript is generated at a 1bp interval. For each 24mer oligo, scores are calculated for the average 15mer frequency in the genome, the oligo (24mer) frequency in the transcriptome, and uniqueness. The 15-mer frequency score is used to avoid probes in repetitive regions while the oligo (24mer) frequency score is used to identify the number of exact matches within the transcriptome. The uniqueness information, as described above, is reduced to a Boolean value based on the weighted mismatch score. To be considered unique, a 24mer oligo must have a weighted mismatch score greater than 10.

The T_m, base pair composition, and self-complementarity of the 60mer oligo are also part of the oligo selection parameters. The 60mer oligos and their associated parameter values are subjected to a rank-based

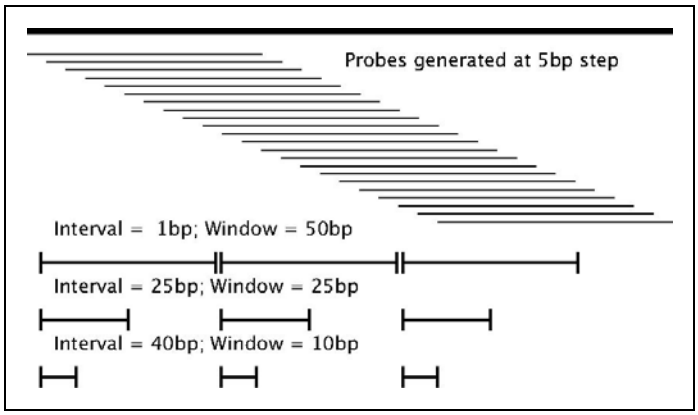

Figure 1. Window-based rank selection

selection process. Each parameter is assigned a different weight (which maybe adjusted by the designer), a score is calculated by summing the weighted values, and the oligo with the highest score is selected. After selection of the first oligo, the rank selection algorithm also accounts for distance to other selected probes within the same transcript. This distance weight helps to space probes over the usable length of the transcript. (*Note: When end-labeling sample preparation methods are used, oligos are limited to the 3′ 1,500bp of the transcript.*)

Unlike tiling where the number of probes is determined by probe spacing and sequence composition, expression designs attempt to have the same number of probes per transcript. Therefore, the rank selection process attempts to select up to *N* probes per transcript where

$$N = \#\_Features / \#\_Transcripts / \#\_Replicates$$

and $6 \leq N \leq 20$. The downside to this process is some truly undesirable probes, in the global sense, may be selected because they are the best choices for a particular transcript.

Removal of Redundant Probe Sets

The selection process attempts to select the best oligos for a given transcript: those that are the most unique, have the lowest oligo frequency and 15mer frequency, and are as close as possible to the target T_m for the design. If there is a conserved domain within a transcript, the selection process will automatically select oligos that fall within the non-conserved region(s). Sometimes, however, there is redundancy in transcript sets, either due to annotation mistakes, legitimate isoforms, or very closely related gene family members. To maximize the number of useful oligos for a given expression design, oligos are selected for all transcripts and then a filtering process is performed on the resulting probe sets. For a given N, the probes for each transcript are compared to all others. When completely duplicate probe sets are found, one transcript is selected as an exemplar for the others. This exemplar is used in the design, while the other transcripts that are also represented by that probe set are noted in the NimbleGen gene description file created for the design.

References

1. www.repeatmasker.org
2. *Bioinformatics*. 2006 Jan 15;22 (2):134-41.
3. www.sanger.ac.uk/Software/analysis/SSAHA/

For More Information

Toll-free in US: (877) NimbleGen / (877) 646-2534
(608) 218-7600
ngsales@nimblegen.com
www.nimblegen.com

NimbleGen

HIGH - DEFINITION GENOMICS℠

Roche

05338301001 • Reprinted 06/08 • Original Publication 11/07

Roche NimbleGen, Inc.
500 S. Rosa Road
Madison, WI 53719 USA