MYcroarray - MYbaits® Bait Design

Procedure and Data

Date: April 25th, 2016
Client: Lisa Komoroske
Project: TURTLE
Source genome: Customer-supplied genome sequence: *Cmyd.v1.1.fa.txt*
Designer: Alison Devault

Bait Design Summary

1) I first combined your 1st & 2nd priority sequences (in order) into a single file, and slightly edited the names
2) I then RepeatMasked all the sequences using default settings as "vertebrate" organism, soft-masking all repeats in lowercase sequence; ~13% of the overall bases were masked this way
3) Then I re-divided the sequences again by priority category, and each bait candidate was BLASTed against the provided TURTLE genome, and a hybridization melting temperature (T_m)* was estimated for each hit assuming standard MYbaits® buffers and conditions.

** T_m is defined as the temperature at which 50% of molecules are hybridized*

** T_m is defined as the temperature at which 50% of molecules are hybridized*

4) For each bait candidate, one BLAST hit with the highest T_m was first discarded from the results (allowing for 1 hit in the genome). Then, based on the distribution of remaining calculated T_m's, we then filtered out non-specific baits using the following criteria:

A) Stringent (only specific baits pass). Bait candidates pass if they satisfy *one* of these conditions:
- No hits with T_m above 60°C
- At most 2 hits 62.5 – 65°C
- At most 10 hits 62.5 – 65°C and at least 1 failing flanking bait
- At most 10 hits 62.5 – 65°C, 2 hits 65 – 67.5°C, and fewer than 2 passing flanking baits
- At most 2 hits 62.5 – 65°C, 1 hit 65 – 67.5°C, 1 hit 70°C or above, and < 2 passing flanking baits

B) Moderate (some non-specific baits pass)
Additional candidates pass if they have at most 10 hits 62.5 – 65°C and 2 hits above 65°C, and fewer than 2 passing baits on each flank.

C) Relaxed (more non-specific baits pass)
Additional candidates pass if they have at most 10 hits 62.5 – 65°C and 4 hits above 65°C, and fewer than 2 passing baits on each flank.

Files for Review

TURTLE-input-seq.fas: target sequences provided (separately by Priority 1 or 2)

TURTLE-probes-120.fas: probes file (same as above with slightly edited names)

TURTLE-probes-120-[stringency].bed: bait positions on each target locus (BED4 format).

TURTLE-probes-120-filtration.txt: for each bait (FIELD 1), the GC content (2), lowercase content (3), number of detected unintended reference hits* at various T_m bins (4-9), pass/fail call at a given filtration level (10-12), and the bait sequence (13) are tabulated.
** We record at most 101, 51 and 21 hits in the last three T_m bins, respectively.*

TURTLE-probes-120-targetcovg.txt: for each target locus (FIELD 1), the contig length (2), and percent baited at various filtration levels (3-5) are tabulated.

Design Metrics

When deciding which level of filtering to use, note that stringent filters might improve % on-target but reduce coverage breadth, while relaxed filters will likely achieve the opposite. Note especially that reads from non-unique regions can be difficult to properly locate in the genome and therefore may be useless.

Priority #1 Sequences				
Metric	**Unfiltered**	**Stringent**	**Moderate**	**Relaxed**
Baits pass	1,379	1,202	1,245	1,248
% baits pass	100.00%	87.16%	90.28%	90.50%
Target loci baited	1,379	1,202	1,245	1,248
Target loci unbaited	0	177	134	131
Total baited length*	165,480	144,240	149,400	149,760
% Target baited	100.00%	87.16%	90.28%	90.50%
Unbaited regions**	0	177	134	131

Priority #2 Sequences				
Metric	**Unfiltered**	**Stringent**	**Moderate**	**Relaxed**
Baits pass	1,400	1,233	1,272	1,274
% baits pass	100.00%	88.07%	90.86%	91.00%
Target loci baited	1,400	1,233	1,272	1,274
Target loci unbaited	0	167	128	126
Total baited length*	168,000	147,960	152,640	152,880
% Target baited	100.00%	88.07%	90.86%	91.00%
Unbaited regions**	0	167	128	126

* = Cumulative size in nucleotides of all baited regions (sequence covered by at least one bait).
** = Contiguous regions of at least one nucleotide that are not covered by at least one bait.