Supplemental Material

CGR algorithms:

I. Algorithm to select comparative genomic hybridization probes for resequencing.

The custom software used to select CGH probes for resequencing performs the following tests on the reference (Ref) and test (Test) genome data:

1) Load Reference and test probe intensities. Create a list of forward and reverse data. Ensure that for each position in the Ref Fwd, there is one in Ref Rev, Test Fwd and Test Rev.
2) Calculate Ratio of PM for Ref/Test for Fwd and Rev. Calculate Average of these 2 numbers.
3) Get 50% Global median of Ratio Fwd, Ratio Rev, Ratio Ave as Median Global Fwd, Median Global Rev, Median Global Ave.
4) Calculate Global Threshold as described below and Std Dev under threshold of Ratio Fwd, Ratio Rev, Ratio Ave as Thresh Global Fwd, Thresh Global Rev, Thresh Global Ave.
5) Iterate through list of Data to calculate local thresholds for CGR Calls. The threshold is calculated using the data that is within 1800 BP of the current position as described below. In those cases where all data is above the cutoff, the global Threshold is used. If the Ratio Ave is greater than Threshold Ave then its position is added to a list of interesting positions.
7) A new list is created. For every interesting position, Pos, the set [Pos-14,Pos+14] is added to the new list.
8) The new list is sorted and made unique. This is the list of 5' positions for the start of probes to be resequenced. The location of the test SNP is in the middle of each of these probes.
9) The probes are generated from the Reference Genome.

Threshold calculation:
Input: First Cutoff. Percentile. Std Dev Multiplier.
1) Any value above the first cutoff is removed from the list.
2) The 80^{th} percentile of the list is calculated as baseline1
3) The Std Dev is calculated as StdDev1
4) Any value above (baseline1 + Std Dev Multiplier × StdDev1) is removed from the list.
5) The 80^{th} percentile of the list is calculated as baseline2
6) The Std Dev is calculated as StdDev2
7) The resulting Threshold is (baseline2 + Std Dev Multiplier × StdDev2).

The Std Dev Multiplier = 3.5

II. Algorithm for selection of oligos for design of resequencing oligos.

An algorithm was designed to select probes for resequencing arrays that optimizes the oligo length, mismatch position, and melting temperature. The algorithm accepts input from the user to specify the minimum and maximum probe lengths, and the target Tm for all the probes on the array. The target T_m of the probe is then divided in half, and each portion of the probe on either side of the mismatch is varied in length, to both stay within the specified length parameters, and to reach half of the specified target Tm. The following equation was used to calculate the probe T_m: Probe $T_m=5*(G_n+C_n)+1*(A_n+T_n)$, where G_n is

the number of Gs present in the probe, and so-on for the other three bases. This is a modification of the Wallace rule[1] empirically modified to better reflect surface probe behavior. The result of the algorithm is to produce array probes that all have a similar T_m, and the mismatch position is held at the approximate thermodynamic center of each oligo, where it is theoretically the most destabilizing. For the current study, oligos were selected between 29 and 39 oligos in length, and the T_m was set at 72°C.

III. Algorithms for SNP uniqueness testing.

The central bases of probes that called a SNP in the genome are checked for uniqueness in two ways. The first is a weighted frequency count, where a subset of the probe representing an increasing number of central bases is checked for a perfect match within the genome. This subset ranges from 19-mer up to 29mers, with shorter oligos contributing less to the N-mer frequency score than longer oligos. Because the probe that is calling a SNP is tested, these probes should be unique in the reference genome, and not effected by the position being called a SNP. The following equation is used to calculate the N-mer frequency score:

$$\text{N-mer frequency Score} = [\text{N-mer frequency}] \times [0.75^{(29\text{mer-Nmer length})/2}]$$

The sum of the frequence scores for central N-mers from 19, 21, 23, 25, 27, and 29mers is calculated for each Probe calling a SNP. The higher the N-mer frequency score, the more likely it is that a position in the genome other than the SNP position is being called a SNP, and thus is likely to represent a false positive. Scores >0 are considered likely false positives.

For the second uniqueness test the central 29-mer spanning the SNP site is compared to the target genome and given a pass/fail score for uniqueness. Since we know that mismatches at the end of a probe are less important than mismatches in the center, we place a vector of weights onto the length of the probe, giving full weight to mismatches in the core 14 bp around the SNP site and less weight to mismatches at the end. The weight vectors used in are listed here:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	1	1	2	2	2	2	4	4	4	4	4	4	4	0	4	4	4	4	4	4	4	2	2	2	2	1	1	1

To pass the uniqueness threshold, the sum of the mismatch weights must be greater than 10. Three mismatches between positions 8 and 22 would be considered unique. Six mismatches in positions 1 to 6 would not meet the uniqueness criterion (1+1+1+2+2+2 = 9). SNP sites that pass this criterion (a score < 10) and given a passing scores of 1, while SNP sites that fail (a score >10) are given a failing score of 0. SNP sites that fail this criterion (have a score of 0) are considered likely false positives.

---

[1] Wallace, R.B.; Shaffer, J.; Murphy, R.F.; Bonner, J.; Hirose, T.; Itakura, K. Nucleic Acids Res. 6, 3543 (1979)