GeneChip® Targeted Genotyping Algorithm Enhancements for Release 1.5

January 2006

Introduction

With the launch of the GeneChip® Targeted Genotyping System 1.5, algorithm enhancements have been made to the Molecular Inversion Probe (MIP) data analysis algorithms. These improvements do not affect the overall framework of MIP data analysis described in [1]. Their effect on the data is to improve the quality of genotyping as well as the speed of clustering. Improvements vary according to the assay panel and experiments analyzed. We have tested the new algorithm on a wide variety of assay panels and found increases of 0-0.2% in trio accuracy, 0-0.3% in completeness and 0.3-1% in conversion (passed assays).

This document describes in detail the algorithm changes between releases 1.0 and 1.5 and should be read in conjunction with [1].

Sample Normalization

As described in [1], data from each experiment are normalized on a per experiment basis by applying three transformations to the data, in the following order: 1) background subtraction, 2) spectral overlap correction, and 3) allele balancing. We have made some changes to the background subtraction and spectral overlap correction steps, whereas we have not changed the allele balancing step.

Background Subtraction

1. Tiles of neighboring features in local background estimation are circles with 9 features per diameter in release 1.5, whereas they were 7-feature-wide squares in release 1.0. The circular tiles are (slightly) more closely matched to the natural shape of local background patterns.
2. In release 1.0, a given percentile of the signal ranked features was used to estimate background in each tile. In contrast, in release 1.5 we use an average of the features between the 2nd and 15th percentiles as ranked by signal. Averaging over features provides a smoother estimate, whereas picking a given percentile can result in picking the central feature itself, which then subtracts to exactly zero. This self-subtraction leads to a distortion of the histogram described below. It is avoided by averaging over features from the 2nd to the 15th percentile (in signal).
3. After local background subtraction we perform an overall background subtraction (a single offset applied to all features). In release 1.0 this offset was determined from the peak of the signal histogram (over all features). Since there is always a

significant number of essentially ‘zero’ signal features in any MIP experiment, the peak of the signal histogram provides a good estimate of the true zero point (average array background). In release 1.5 we use the same histogram, but instead of its peak we use the leading edge of the histogram (signals below peak) and extrapolate the slope of the leading edge to the same height as the peak. This is (slightly) preferable to using the peak itself since the peak is asymmetric (it has a longer tail on the high side) and the extent of the asymmetry on the high side (dominated by assay background and spectral overlap magnitude) is more variable than on the low side (dominated by array noise).

Spectral Overlap Correction

As in release 1.0, we measured the spectral overlap between each pair of colors (alleles) and then use these values to construct a 4x4 spectral overlap matrix which we invert and apply to the background subtracted data. We also still use the homozygous genotypes to measure each pair of spectral overlaps. In both release 1.0 and release 1.5, two data cuts are used to isolate these homozygous genotypes and these two cuts are controlled by two parameters: 1) HomFraction, the fraction of genotypes that are considered homozygous (as opposed to heterozygous) for the given sample, and 2) LowSignalFraction, the fraction of data that has too low a signal to use for spectral overlap estimation. In release 1.0 these parameters were set at conservative values of HomFraction=20% and LowSignalFraction=50%. Typical (human) samples have a real HomFraction in the range 60-80%, depending on the average allele frequency of the assay panel. Using a lower estimate of the HomFraction than is actually the case for any given sample causes extremely slight underestimation of spectral overlap values. This has no practical effect on genotype quality. Overestimating HomFraction can have more serious consequences for genotype quality and, hence, in release 1.0 we used a conservative underestimate (20%) that allowed the accurate genotyping of samples with real HomFractions as low as 10% (e.g., cross breed mouse samples).

In release 1.5 we have increased the robustness of our spectral overlap correction by making the procedure iterative. The first iteration is exactly as described above except that the LowSignalFraction is set at 20% (instead of 50%). This allows a better estimate for cases where there are significant differences in average signal from the different alleles (unbalanced alleles). Having performed the first iteration of spectral overlap correction, we then apply a threshold basecalling algorithm to determine better estimates of the HomFraction and LowSignalFraction values that apply to this sample. We then use these values to estimate any additional spectral overlap correction that needs to be applied (e.g., if there was a slight under- or over-estimation in the first iteration). This procedure of threshold basecalling to estimate HomFraction and LowSignalFraction and then applying any remaining spectral overlap correction is repeated twice (respectively, the second and third iterations of spectral overlap correction). The procedure converges very rapidly. Typically 99% of the spectral overlap correction is performed in the first iteration and approximately 1% in the second iteration and <0.1% in the third iteration. The main advantage of the iterative procedure is that it is more robust to a wider range of true HomFraction values (down to 3%) and LowSignalFraction values (from 0 – 80%) in real samples. It also provides more accurate spectral overlap correction since a dynamically adjusted estimate of HomFraction and LowSignalFraction is used for each sample.

Clustering Algorithm for Genotyping

The expectation-maximization (E/M) clustering algorithm that is used to genotype each marker is the same as in release 1.0 and as described in [1]. However, a few changes have been made in the practical details of applying this algorithm. As discussed in [1], release 1.0 used 45 different seed conditions for the 9 parameters of the model (3 cluster centers, 3 cluster widths (sigma), and 3 cluster weights (sum to 1). Each separate set of 9 seed values was converged to its local maximum through the E/M process. Then a log-likelihood metric was computed for that fit. The highest log-likelihood of the 45 fits was deemed the 'global' maximum and used to determine genotypes. This procedure is good at avoiding local minima problems from inappropriate seed values, but it is also quite computationally intensive.

In release 1.5 we perform a pre-clustering evaluation of the marker so that more targeted choices of seed values are made for each marker (in release 1.0 the same set of 45 seed conditions are applied to each marker). This pre-clustering evaluation of the marker is essentially a threshold calling procedure where samples are 'called' according to their signal contrast values using fixed bins (in this 1-D space of signal contrast) where the homozygous and heterozygous calls are expected to be (on average). There are also 'no-call' zones between the expected bins. If the total number of samples in the 'no-call' zones is less than the number of samples in the smallest occupancy bin, then the marker is called 'well binned.' For well-binned markers we use the samples in each bin to determine the seed values for that cluster, e.g., average signal contrast defines the cluster center and likewise for the sigma and weight of the cluster associated with a given bin. If the marker is well binned then no further sets of seed values are tried for the given marker. About 2/3rds of the markers are typically well binned in any given assay panel so there is considerable speed improvement in the algorithm since we only use 1 set of seed values for these markers (as opposed to 45 sets in release 1.0).

Markers that are not well binned (typically 1/3rd) have significant numbers of samples in the no-call zones between the bins. For these markers, the bin boundaries are adjusted in a variety of ways that capture a wide spectrum of observed marker behavior, e.g., the heterozygous bin is expanded towards a homozygous bin and that homozygous bin is correspondingly pulled back. For each scenario of adjusted bin boundaries, we use the samples in each bin to determine seed values for the parameters of each cluster (just as in the well-binned case) and then perform an E/M clustering. The number of bin scenarios attempted depends on the marker but is never more than 22, so that even for these markers there is a significant improvement in algorithm speed with respect to release 1.0 where 45 E/M fits were performed for each marker.

In addition to the above changes in how cluster seed values are calculated, we have also made slight changes to overall constraints that are applied to the 9 parameters in each

E/M clustering fit, e.g., maximum and minimum allowed values for cluster sigmas. Also, in release 1.5 the ranges of cluster means do not overlap with each other, while they are allowed to overlap with each other in release 1.0. These constraints are empirically derived by testing over large data sets.

The net effect of the above clustering changes is to reduce the number of ‘mis-clustered’ markers. This has led to improvements in genotyping quality that depend somewhat on the assay panel and are in the following ranges: 0-0.2% increase in trio accuracy, 0-0.3% in completeness, and 0.3-1% in conversion (passed assays).

References

[1] Moorhead M, Hardenbol P, Siddiqui F, *et al*: Optimal genotype determination in highly multiplexed SNP data. *European Journal of Human Genetics* 2005 Nov 23; (in press).