Statistical methods for off-target variant genotyping with Axiom arrays

Abstract

Off-target variants (OTVs) are genomic markers with sequences that differ significantly from those of hybridization probes that are used in microarray-based genotyping [1]. Mismatches between probes and targets can reduce hybridization intensity, so the genotypes of subpopulations are often incorrectly identified as heterozygous by genotyping algorithms that cluster intensities into the expected AA, AB, and BB genotypes. OTVs tend to group members of a divergent subpopulation into a fourth cluster denoted as the OTV cluster. Correctly identifying the OTV cluster, a process known as OTV genotyping, is thus a critical step.

We have developed a statistical method called OTVCaller to genotype OTVs based on the posterior genotyping information generated by the automated clustering algorithm AxiomGT1 developed for Applied Biosystems™ Axiom™ Analysis Suite software. OTVCaller uses information about posterior genotypes to initiate the Baum-Welch algorithm, which iteratively searches for the optimal locations of the AA, AB, BB, and OTV clusters. It then derives genotypes and OTV types when convergence is reached. We applied OTVCaller to genotype data for multiple species and demonstrated that it could successfully identify the OTV cluster in the presence of all three AA, AB, and BB genotypes or just AA and BB genotypes. OTVCaller is available with the Applied Biosystems™ Axiom™ SNPolisher™ R post-processing package for SNP data, and it can be applied with Applied Biosystems™ Axiom™ genotyping arrays.

Background

OTV probes usually display quite low hybridization intensity and center at zero in the contrast dimension due to double deletion or sequence nonhomology [1]. Consequently, they tend to confound genotype calling for heterozygotes. There is thus a critical need for statistical methods to distinguish OTVs from true heterozygotes.

Identification of OTV genotypes

We internally developed a metric for SNP quality control called heterozygote strength offset (HetSO). It is defined as the vertical distance from the center of the heterozygote cluster to the line connecting the centers of the homozygote clusters, as measured by (A+B)/2 or signal size (Figure 1).

Large negative values are usually associated with OTV clusters. We define OTVs as genotypes with good cluster resolution but HetSO values below –0.3, which is a customizable threshold (Figure 2). Identification of OTV genotypes is executed automatically by the SNPolisher R package.

Algorithm for OTV genotyping

We used the OTVCaller approach to genotype OTV SNPs. This algorithm takes the posterior genotyping information generated by the automated clustering algorithm AxiomGT1 and initiates the 2D Baum-Welch algorithm, which searches for optimal clustering in both the intensity and contrast dimensions. It then iteratively updates sample assignments and cluster centers until convergence is reached. Detailed workflows for implementing the OTVCaller algorithm in the SNPolisher R package are shown in Figure 3.

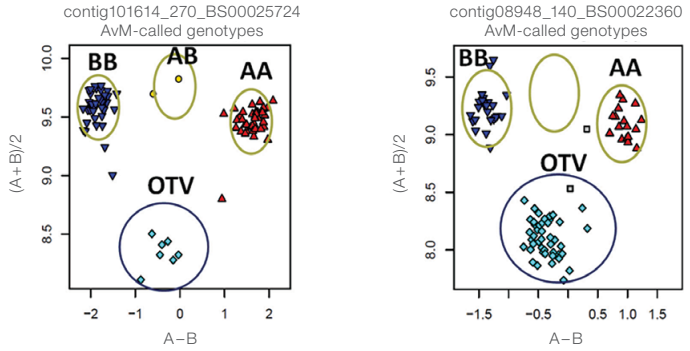

Figure 1. Two examples of OTV genotypes. At left are four genotype clusters: AA, AB, BB, and OTV. At right are three genotype clusters: AA, BB, and OTV. The cyan diamonds represent off-target variants.

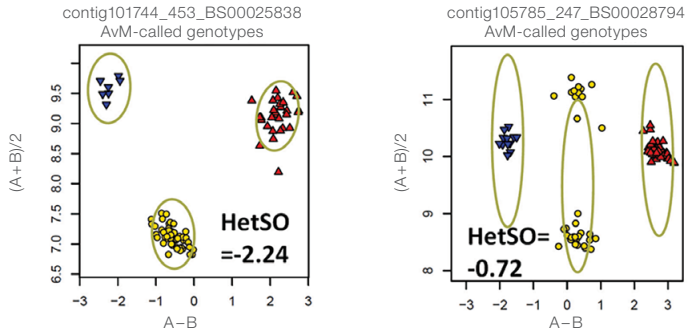

Figure 2. Examples of genotype clusters with low HetSO values. They are identified as OTV genotypes since their HetSO values are below –0.3.

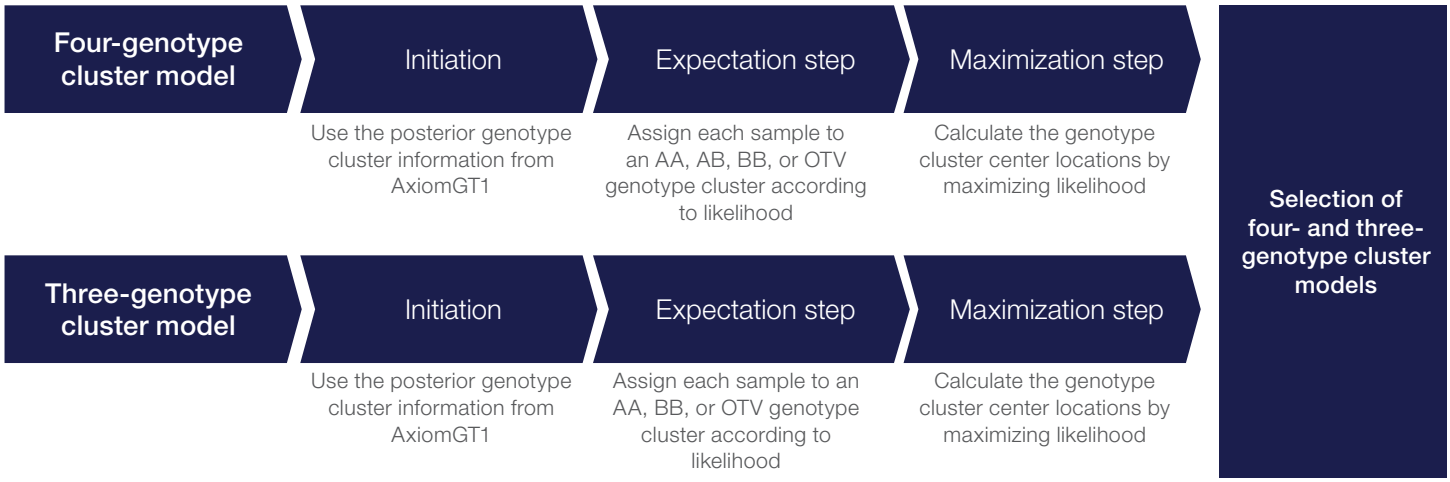

Figure 3. OTVCaller algorithm workflows. Two models are considered: a three-genotype model (AA, BB, and OTV) and a four-genotype model (AA, AB, BB, and OTV). The last step is selection of the best-fit model.

applied
biosystems

Application to wheat OTV genotyping

We implemented the OTVCaller algorithm in the SNPolisher R package to analyze bread wheat genotyping data generated with a custom Axiom™ array. OTVCaller accurately corrected miscalled heterozygotes, as shown in Figure 4. The OTVs had been miscalled as heterozygotes, and true heterozygotes had been miscalled as nulls before OTV genotyping. They were all correctly called after OTV genotyping.

References

1. Didion JP, Yang H, Sheppard K et al. (2012) Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* 13:34.

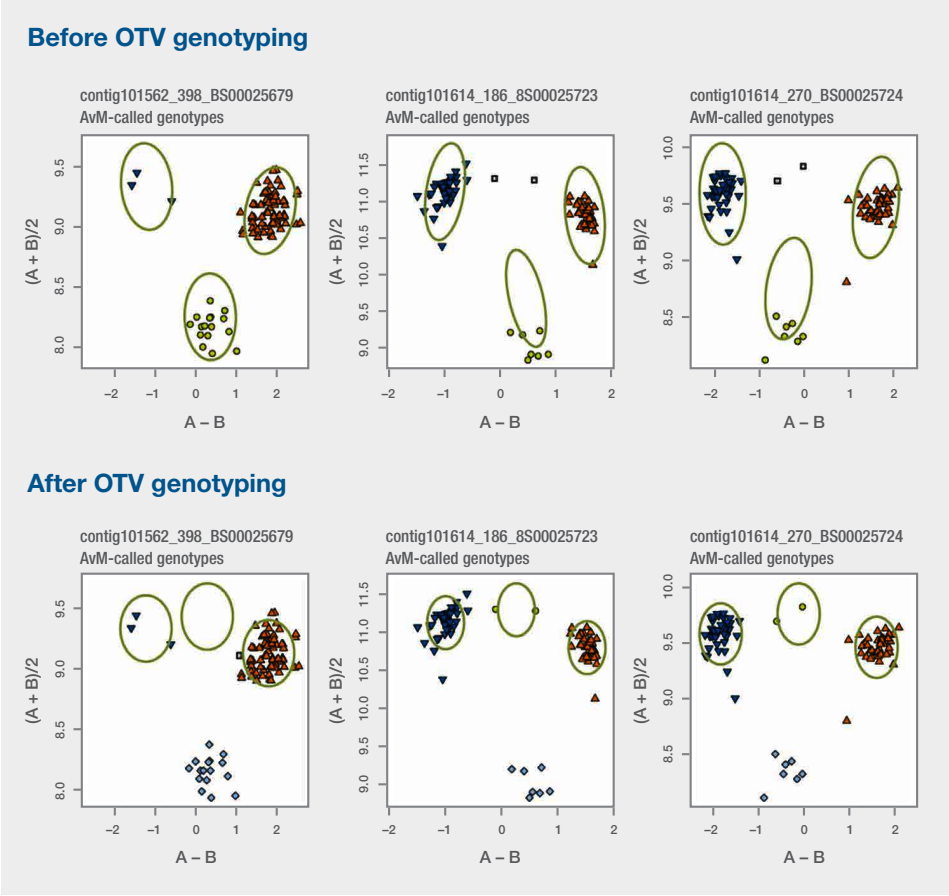

Figure 4. Bread wheat genotyping data illustrating the implementation of the OTVCaller algorithm.

Find out more at
thermofisher.com/microarrayagrigenomics

ThermoFisher SCIENTIFIC