Application Brief

Statistical methods for off-target variant genotyping on Affymetrix Axiom® Arrays

T. A. Webster, A. Pirani, M. Shen, L. Bellon, and H. Gao

Abstract

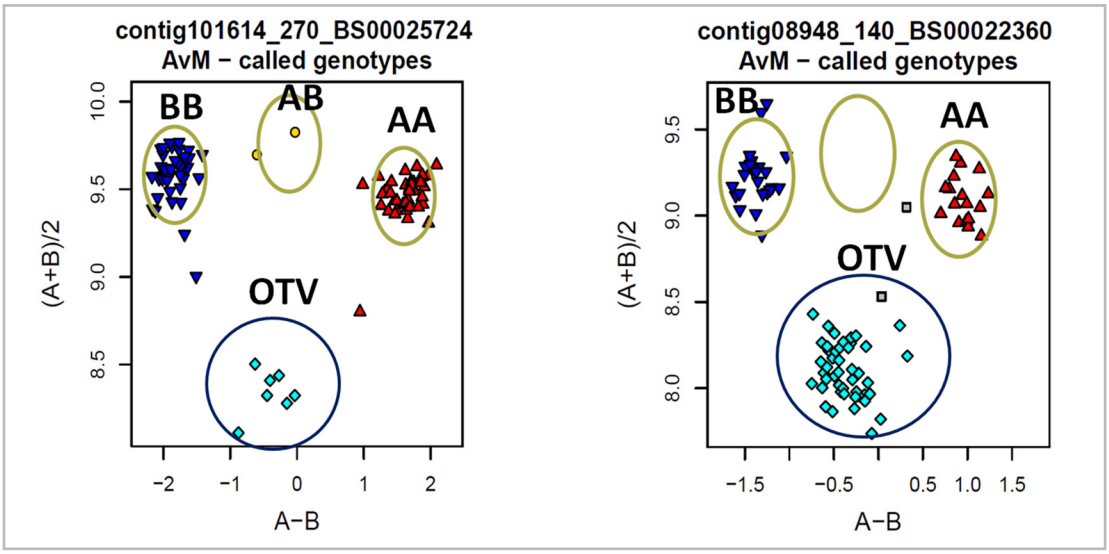
Off-target variants (OTVs) (Didion, *et al.*, *BMC Genomics* **13**:34, 2012) are genomic markers with sequences that are significantly different from the sequences of hybridization probes used in microarray-based genotyping. This dissimilarity between probe and target can lower hybridization intensities such that genotypes of the subpopulation are often incorrectly called as heterozygotes by genotyping algorithms that cluster intensities into three expected genotypes, namely AA, AB, and BB. OTVs tend to group members of the divergent subpopulation into a fourth cluster denoted as the OTV cluster. Therefore, correct identification of this OTV cluster, a process described as OTV genotyping, is a critical step.

We have developed a statistical method called "OTVCaller" to genotype all OTVs based on the posterior genotyping information generated by the automated clustering algorithm AxiomGT1, which is used by Axiom® Genotyping Console™ Software. OTVCaller uses posterior genotypes as the prior information to initiate the Baum-Welch algorithm to iteratively search for the optimal locations of the four clusters AA, AB, BB, and OTV. It then derives genotypes and OTV types when convergence is reached. We applied OTVCaller to genotype data from multiple species and demonstrated that OTVCaller can successfully identify the OTV cluster in the presence of all three genotypes (AA, AB, and BB) or in the presence of only two genotypes (AA and BB). OTVCaller is available from Affymetrix in a SNP data analysis post-processing "SNPolisher™" R package and is applicable to Axiom® genotyping arrays.

Background

Off-target variant (OTV) probes (Didion, *et al.*, 2012)[1] usually demonstrate substantially low hybridization intensity and center at zero in contrast dimension due to double deletion or sequence non-homology. Thus, they tend to confound the genotype calling of heterozygotes. Therefore, there is a critical need for statistical methods to distinguish OTVs from true heterozygotes.

Figure 1: Two examples of OTV genotypes. Left has four genotypes: AA, AB, BB, and OTV. Right has three genotypes: AA, BB, and OTV. Diamonds in cyan represent off-target variants.

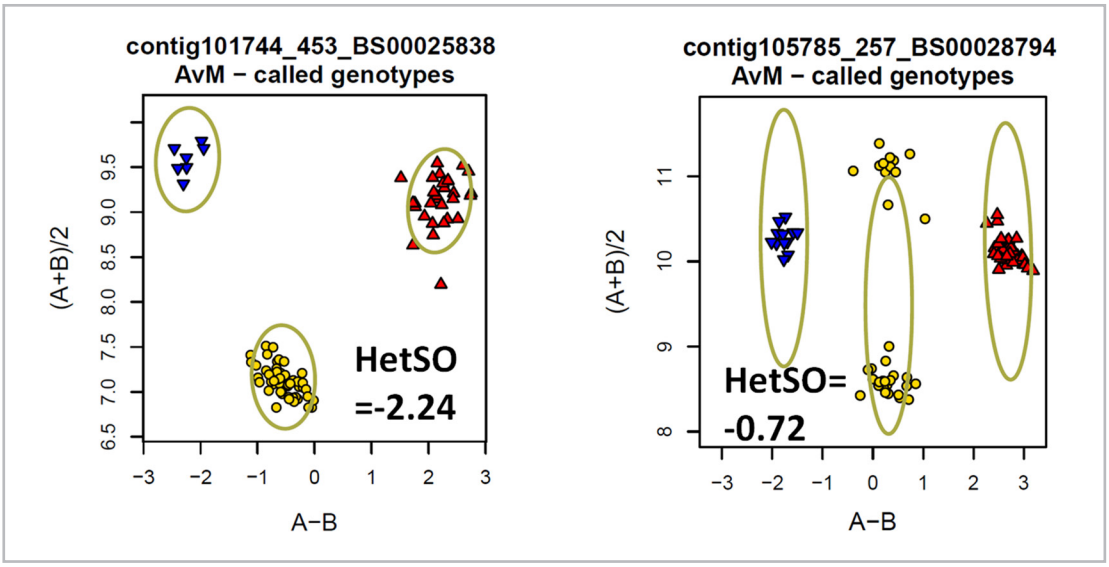


Identification of OTV genotypes

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

Large negative values are usually associated with OTV clusters. Therefore, we define OTVs as genotypes with good cluster resolution but with HetSO <-0.3, which is a customizable threshold. Identification of OTV genotypes is executed automatically by SNPolisher™ R package.

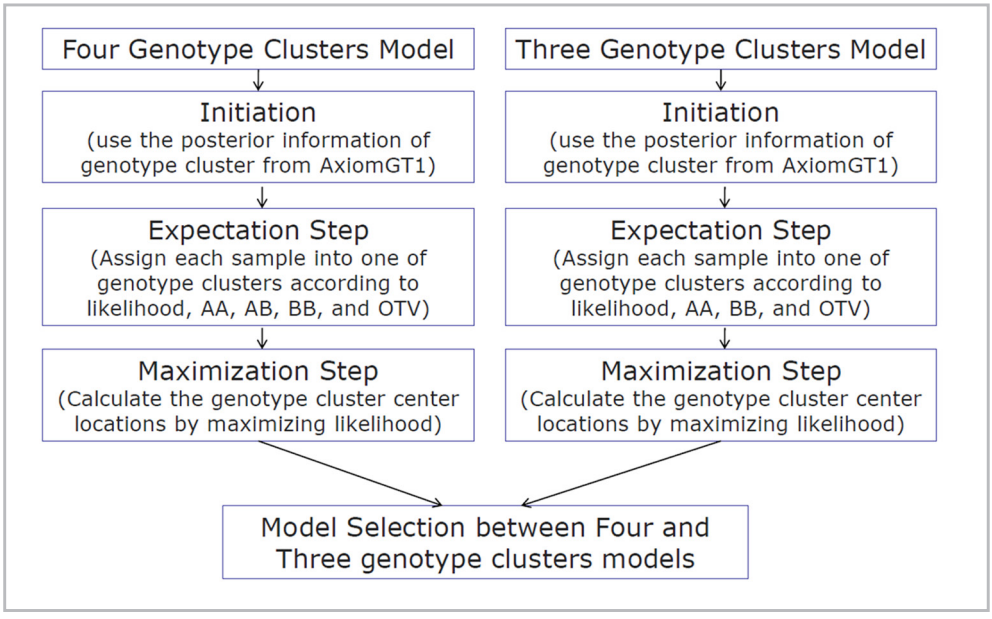
Figure 2: Two examples of genotype clusters with low HetSO values. They are identified as OTV genotypes as their HetSO values are below -0.3.



Algorithm for OTV genotyping

We developed a statistical approach called "OTVCaller" to genotype the OTV SNPs. This algorithm uses the posterior genotyping information generated by the automated clustering algorithm AxiomGT1 to initiate the 2D Baum-Welch algorithm in order to search for the optimal clustering in both intensity strength and contrast dimensions. It then iteratively updates sample assignments and cluster centers until convergence is reached. The detailed workflow of the OTVCaller algorithm is shown in Figure 3 and is implemented in SNPolisher R package.

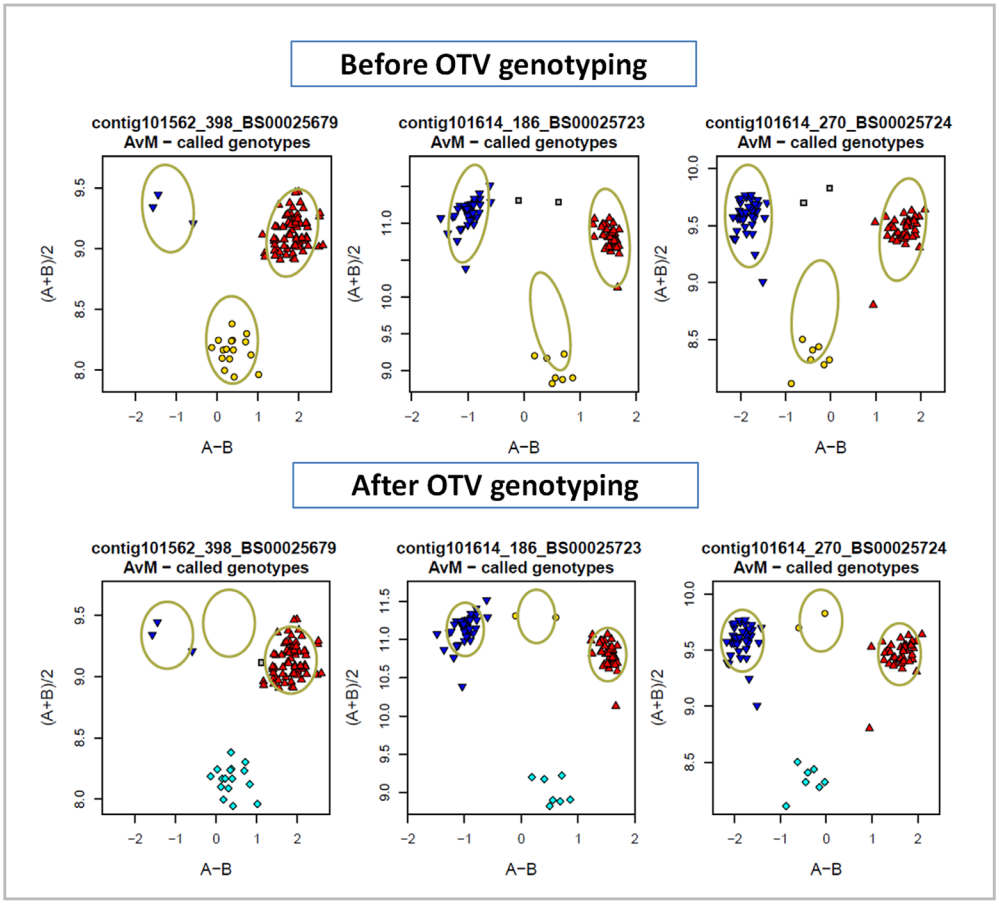
Figure 3: The workflow of the OTVCaller algorithm. Two models are considered, including a three-genotype model (AA, BB, and OTV) and a four genotype model (AA, AB, BB, and OTV). The last step is to select the best-fit model.



Application to wheat OTV genotypes

We applied the OTVCaller algorithm implemented in SNPolisher™ R package to bread wheat genotyping data generated with Affymetrix' Axiom® custom array. OTVCaller accurately corrected miscalled heterozygotes, as shown in Figure 4, where OTVs had been miscalled as heterozygotes and true heterozygotes had been miscalled as null calls before OTV genotyping. They are all corrected after OTV genotyping.

Figure 4: The workflow of the OTVCaller algorithm. Two models are considered, including a three-genotype model (AA, BB, and OTV) and a four genotype model (AA, AB, BB, and OTV). The last step is to select the best-fit model.



References

1. Didion J. P., Yang H., Sheppard K., Fu C., McMillan L., Villena F. P., Churchill G. A. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* **13**:34 (2012).

Affymetrix, Inc. Tel: +1-888-362-2447 ■ **Affymetrix UK Ltd.** Tel: +44-(0)-1628-552550 ■ **Affymetrix Japan K.K.** Tel: +81-(0)3-6430-4020
Panomics Solutions Tel: +1-877-726-6642 panomics.affymetrix.com ■ **USB Products** Tel: +1-800-321-9322 usb.affymetrix.com

www.affymetrix.com Please visit our website for international distributor contact information.

"For Research Use Only. Not for use in diagnostic procedures."

P/N DNA01919 Rev. 1