applied biosystems

Rare Heterozygous Adjusted Genotyping

Introduction

Applied Biosystems™ Axiom™ custom and catalog genotyping arrays offer highly accurate genotyping for common and rare variants. The use of genotyping arrays for extremely rare variants, defined as those with a Minor Allele Frequency below 0.01%, can be challenging. We describe a novel genotyping algorithm (Rare Heterozygote (Het) Adjusted Genotyping), tailored to very rare variants that achieves excellent accuracy for these variants.

This technical note provides the background information on the rare het adjustment algorithm, including results of analyses on several training sets with samples with known genotypes. In addition, it describes verification results on the UK Biobank Data, a health resource that follows the health and well-being of 500,000 participants, all genotyped on Thermo Fisher Scientific's Applied Biosystems Axiom UKBiobank and UKBileve microarrays. We include a comparison of the results of the new algorithm compared to Exome sequencing data, available for ~50,000 of the UKBiobank cohort.

Rare Het Adjusted Genotyping can be used with Applied Biosystems™ Array Power Tools (APT 2.11.3)* or Axiom Analysis Suite Software (AxAS) v5.1.

Basics of Rare Het Adjusted Genotyping

The Axiom Rare Het Adjusted Genotyping algorithm optimizes Axiom GT1 genotyping for very rare variants. While the location and shape of the heterozygous (het) cluster provides powerful evidence for het calls for common variants, rare variants often have a single sample in the het cluster, making the call less robust. An in-depth analysis of the distribution of replicate probe signal on Axiom microarrays revealed that this distribution can be used to significantly improve the accuracy of heterozygous calls.

The algorithm examines heterozygous calls from probesets** which have two replicates on the array and which have three or less het calls in the batch. The hets are selectively adjusted to “No Calls,” when the distribution of replicate probe signals for the sample and probeset in question suggest that the call is a false positive.

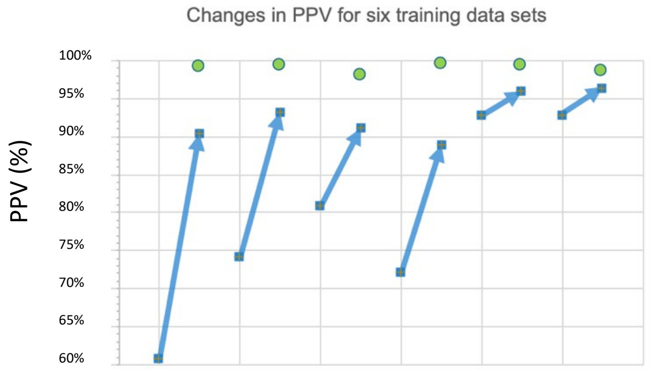
The number of het calls that are examined and set to No Call depends on the array and more specifically on the number of rare variants on the array. Typically less than a few hundred calls are adjusted in a batch of 96 samples genotyped on a high density array. If the percentage of rare markers on an array is around 1% of all markers, we expect the rare het adjustment to affect (on average) less than one het per sample.

Our Rare Het Adjusted Genotyping algorithm was designed and trained on many independent datasets. One important metric for rare het calls is the Positive Predictive Value (PPV) of such calls. PPV is the fraction of correct calls out of all het calls. Figure 1 shows the positive predictive value (PPV) for six representative data sets before and after the application of Rare Het Adjusted Genotyping to clusters with a single het, all genotyped with 1000 Genome samples with published genotypes. Another important metric is the number of true positives calls retained by the algorithm. Figure 1 shows that on average over 99% of the true positives calls are retained, while significantly improving the PPV to over 90%. Similar results were obtained when we applied the algorithm to the UKBiobank+ data.

* Array Power Tools was previously called Affymetrix Power Tools.

** A probeset is a group of one or more probe sequences that interrogates a specific known polymorphic or nonpolymorphic location in the genome.

+ Analysis of UK Biobank data is conducted using UK Biobank Resource under Application Number 55681.



Blue squares & arrows: change in PPV after rare het adjustment.

Green dots: %TP hets retained after rare het adjustment

	Data Set	1	2	3	4	5	6
PPV	Initial PPV	60.9%	74.3%	81.0%	72.3%	92.9%	93.0%
	PPV after rare het adjustment	90.5%	93.2%	91.2%	89.0%	96.0%	96.5%
%TP retained	% TP retained after rare het adjustment	99.2%	99.4%	98.1%	99.6%	99.5%	98.6%
#TP	Initial #TP	2,477	332	3,178	2,328	8,045	5,862
	#TP after rare het adjustment	2,456	330	3,119	2,319	8,007	5,779
#FP	Initial #FP	1,593	115	746	894	612	442
	#FP after rare het adjustment	258	24	300	287	333	211
Size	# samples in set	266	374	280	275	95	279

Figure 1. Bottom: #TP: Number of Axiom concordant het calls among those examined by the algorithm; #FP: Number of Axiom non-concordant het calls among those examined by the algorithm. The table shows #TP and #FP before and after Rare Het Adjustment. PPV = TP /(TP+FP). %TP retained is the %TPs unchanged by rare het adjustment. Graph (top) shows the change in PPV pre and post Rare Het adjustment and %TP retained post rare het adjustment.

Preamble on UKBiobank Data

Before describing the verification results on UKBiobank Data[3] (UKBB), we define Minor Allele Frequency (MAF) bins. Such MAF bins were described by Weedon et al[1]. The authors binned UKBB probesets into Axiom computed MAF (AcMAF) bins, based on Axiom array genotypes in UKBB. Bin 1 has extremely low AcMAF with at most 9 of the nearly 500,000 individuals genotyped on the UKBB array having the minor allele.

Bin	1	2	3	4	5
AcMAF	0% –0.001%	0.001% -0.005%	0.005% - 0.01%	0.01% - 1%	1%-50%

Table 1. Five AcMAF bins as defined by Weedon et al. Note that Bins 1-3 have extremely low allele frequency. A variant in Bin 3 is expected to have less than 5-10 individuals with a het call among 50,000 individuals, so even Bin 3 has extremely low MAF.

The variants in Bin 1 through Bin 3 represent < 5% of all variants on the UKBB array.

Because of their low MAF, the large size of the UKBB cohort, and the availability of exome sequencing data on 10% of the cohort, it is a great resource to verify the new Rare Het Adjusted genotyping algorithm.

Note regarding AcMAF and MAF. AcMAF is the computed MAF based on UKBB data. This computed MAF is highly accurate for almost all variants, but enriched for errors in the lowest MAF bins. These lowest MAF bins contain true low MAF markers along with nonresponsive probesets, easily detected by comparison to population allele frequencies and routinely removed from more recent Axiom platform data. Figure 2 (below) shows a density plot comparing Axiom Computed MAF with GnomAD (non-Finnish European), which is a great match to the UK population.

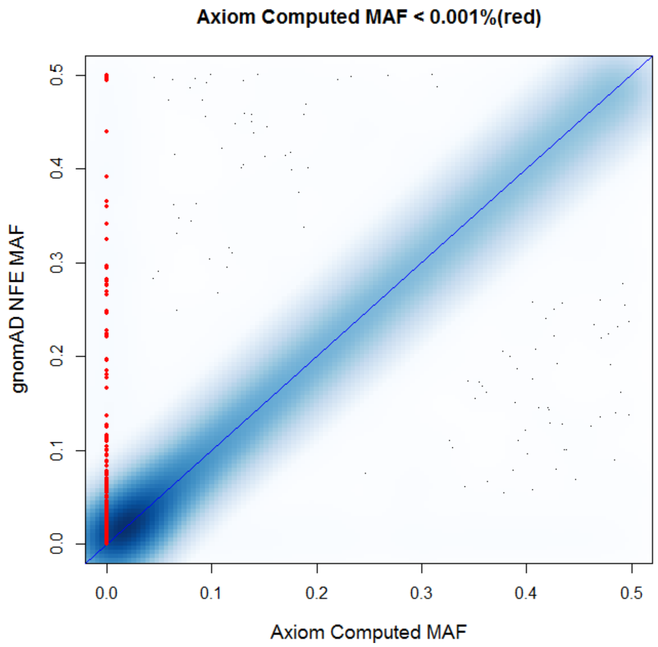


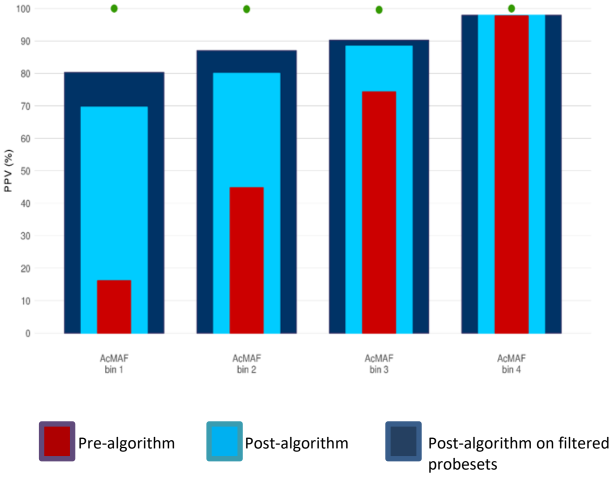
Figure 2. Density plot comparing Axiom Computed MAF with GnomAD NFE MAF (non-Finnish European), showing extremely high correlation overall.

Bin 1 probesets are colored in red, showing Bin 1 contains some common markers with nonresponsive probesets, easily detected by comparison to population allele frequencies and routinely removed from more recent Axiom platform data. About 30% of markers in bin1 have an expected allele frequency greater than 0.005% (5x greater than the boundary for the bin), and 20% have an expected allele frequency greater than 0.01% (10x greater).

Results on UKBiobank Data

As shown in Figure 3 below, Rare het Adjusted genotyping significantly improves PPV in all MAF ranges. Bin 4 and Bin 5 both have excellent performance.

These bins do not have any rare variants and are therefore not affected by the Rare Het Adjusted Genotyping. Bin 4 is shown in the graph below for completeness, while Bin 5, which looks identical to Bin 4 has been omitted from graph. In addition to applying Rare Het Adjusted Genotyping, the graph also shows that careful filtering of probesets significantly improves overall performance. As shown in Figure 2, a small number of nonresponsive probesets can significantly affect overall performance and these probesets are easily detected by comparing to population allele frequencies (a standard practice for more recent array designs).



● Percentage of True positives retained when applying the algorithm to filtered probesets

Note: Bin 5 with MAF >1% not shown as all values are virtually at 100% pre and post rare het adjustment

Figure 3: PPV and sensitivity were calculated based on the exome sequencing data for ~50,000 UKBB samples. Statistics were calculated on all heterozygous genotypes from probesets in the various computed MAF groups, applied on two-replicate probesets with a het cluster size of up to 4 within the respective batch. Probesets were filtered based on the rare het algorithm prediction, as well as GnomAD MAF completely out of range for the bin. Filtering did not use any exome sequencing concordance data.

Conclusions

Improved algorithms for genotype calls for Axiom microarrays achieve excellent PPV for very rare variants, removing false het calls with high accuracy, while keeping true calls virtually intact.

General observations on very rare variants (below 0.01% MAF, e.g. less than 1 expected het in 5,000 individuals)

Axiom Array genotypes give excellent performance on these markers, especially when curated with samples carrying the rare het to confirm performance.

We observed that when we restrict the variants to those probesets that produced at least one true positive het in the exome sequencing data, we achieve a PPV of 98%, along with an average sensitivity of 96%, (data not shown). This superior performance was observed on the AcMAF bin 1, the most challenging bin.

References

1. Weedon et al., (2019) Assessing the analytical validity of SNP-chips for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing, https://www.biorxiv.org/content/10.1101/696799v2
2. UK Biobank Axiom Array. Data sheet: https://assets.thermofisher.com/TFS-Assets/LSG/brochures/uk_axiom_biobank_genotyping_arrays_datasheet.pdf
3. Bycroft C, Freeman C, Petkova D et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203–209

Find out more at **thermofisher.com/microarrays**

For Research Use Only. Not for use in diagnostic procedures. **COL32811 0319**