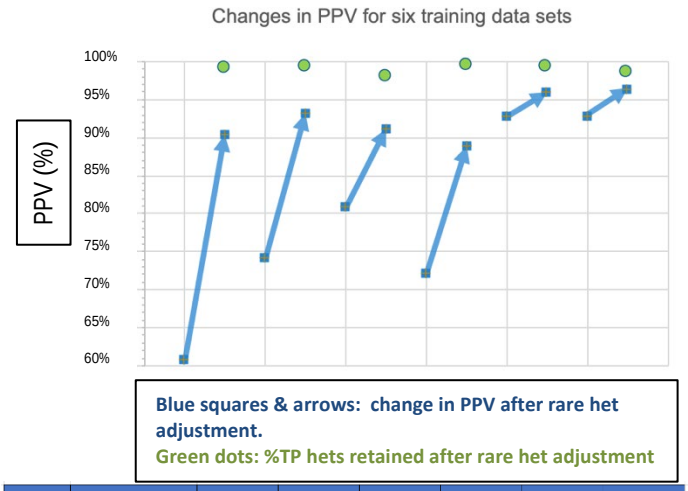

	Data Set	1	2	3	4	5	6
PPV	Initial PPV	60.9%	74.3%	81.0%	72.3%	92.9%	93.0%
	PPV after rare het adjustment	90.5%	93.2%	91.2%	89.0%	96.0%	96.5%
%TP retained	% TP retained after rare het adjustment	99.2%	99.4%	98.1%	99.6%	99.5%	98.6%
#TP	Initial #TP	2,477	332	3,178	2,328	8,045	5,862
	#TP after rare het adjustment	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% F0.001%	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.

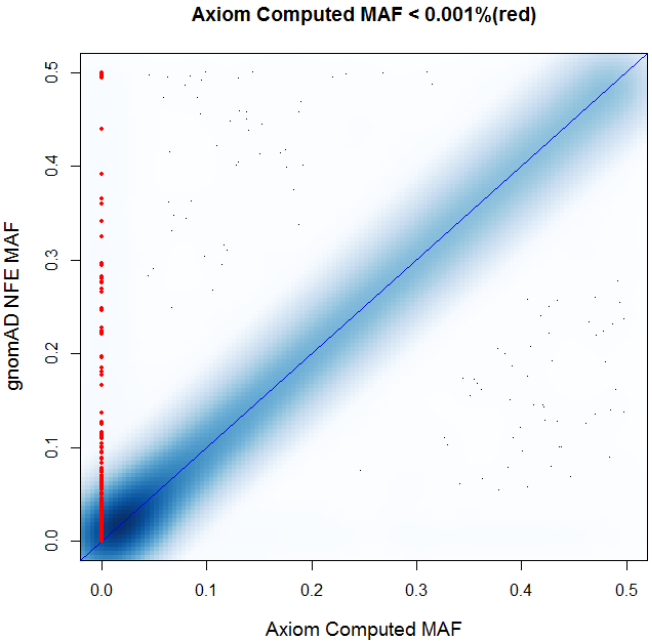

Figure 2. Density plot comparing Axiom Computed MAF with GnomAD NFE MAF (non-Finnish European), showing extremely high correlation overall.

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 foU 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).

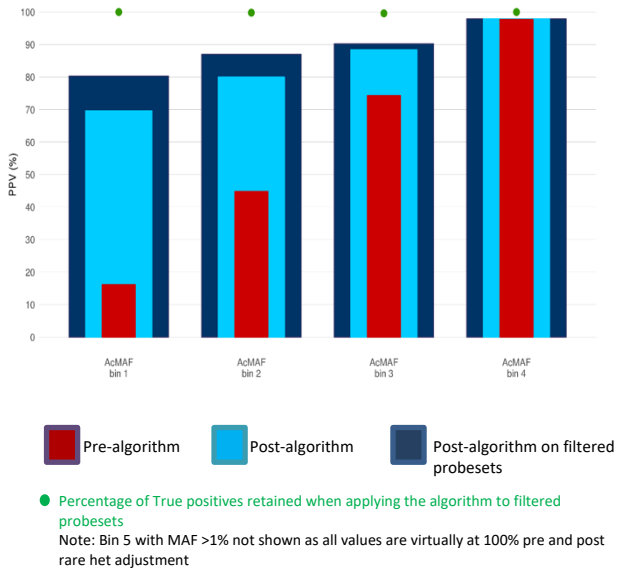

Figure 3: PPV and sensitivity were calculated based on the exome sequencing data for ~50,000 UKBB samples. Statistics were calculated on all heterozygous genotypes from probesets in the various computed MAF groups, applied on two-replicate probesets with a het cluster size of up to 4 within the respective 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**