Accurate Error Rates: Calculating Reproducibility by Minor Allele Frequency

J. Romm, P. Zhang, H. Ling, E. Pugh, K. Doheny

Center for Inherited Disease Research (CIDR), IGM, JHU-SOM, Baltimore, MD

Introduction

The Center for Inherited Disease Research (CIDR) provides high quality next-generation sequencing (NGS), genotyping and statistical genetics consultation to investigators working to discover genes that contribute to diseases. Many commercially available genotyping arrays (HumanExome, HumanOmni2.5, HumanOmni5 and all Illumina "plus exome" arrays) contain SNP content that enables users to assay low minor allele frequency (MAF) variants (<1%). Although overall sample reproducibility rates have not changed with the use of arrays containing rare variants, there are concerns about the accuracy of low MAF SNPs given the nature of the genotype cluster and calling algorithms. In order to get a better picture of the entire project, minor allele reproducibility rates from study duplicate pairs were evaluated in different MAF bins (0-1%, 1%-5%, 5%-10%, 10%-50%) and with different statistics.

Classic Reproducibility (all data)

$$Classic\ Reproducibility = \frac{n}{N}$$

For each MAF bin, *N* is the number of sites that both have genotypes in one duplicate pair, *n* is the number of the pair of sites that have the same genotypes out of *N* sites.

With classic reproducibility all MAF bins have rates > 99.6% and counterintuitively, the lowest MAF (0, 0.01] bin shows the highest reproducibility. This is because for low MAF SNPs almost all duplicate pairs are the major homozygote, which is the easiest of the alleles to call for a rare SNP.

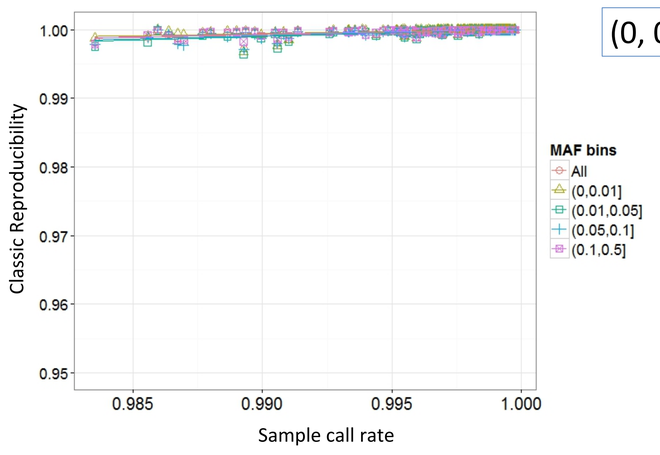

Figure 1: Scatterplot of classic reproducibility estimates (all data) versus sample call rate (minimum of each duplicate pair).

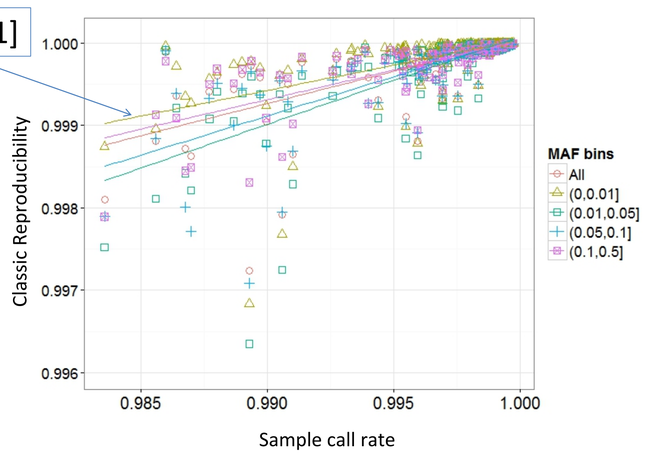

Figure 2: Scatterplot of classic reproducibility estimates (all data) but with zoomed in scale to focus on differences.

Results

The example project has 11,639 total samples, including 228 blind duplicate pairs. Samples were run on the Illumina HumanOmni5Exome-4v1 array (4,511,703 SNPs). Prior to release, a cluster file was created using the project's samples and a technical filter was applied to the SNPs in order to eliminate assay failures.

Reproducibility excluding major allele homozygotes

$$Reproducibility\ excluding\ major\ allele\ homozygotes = \frac{n-m}{N-m}$$

For each MAF bin, *N* is the number of sites that both have genotypes in one duplicate pair and out of *N* sites, *m* sites are both major allele homozygotes and *n* sites have the same genotypes in the duplicated pairs.

Kappa

$$k = \frac{f_0 - f_e}{N - f_e}$$

Kappa statistic is a measure of agreement excluding agreement by chance. Here N is the number of sites that both genotyped in one duplicate pair. f_0 is the number of observed pairs that have the same genotypes out of N pairs. f_e is the expected number of pairs that have the same genotypes out of N pairs (Jacob Cohen 1960).

	MAF bin				
	All	**(0, 0.01]**	**(0.01, 0.05]**	**(0.05, 0.1]**	**(0.1, 0.5]**
Reproducibility excluding major allele homozygotes	(0.9845 - 1) [687,264 -897,563]	(0.6046 - 0.9996) [5,703 - 45,028]	(0.9125 - 0.9998) [34,487 - 102,614)	(0.9771 - 1) [42,146 - 78,108]	(0.9955 - 1) [604,113 -683,661]
Kappa	(0.9899 - 1) [4,160,840 - 4,244,348]	(0.7516 - 0.9999) [1,441,278 -1,464,964]	(0.9518 - 0.9999) [967,620 - 986,799]	(0.9865 - 1) [357,754 -364,041]	(0.996 - 1) [1,374,416 - 1,428,618]
Classic Reproducibility (all data)	(0.9972 - 1) [4,160,840 - 4,244,348]	(0.9968 - 1) [1,441,278 - 1,464,964]	(0.9964 - 1) [967,620 - 986,799]	(0.9971 - 1) [357,754 - 364,041]	(0.9979 - 1) [1,374,416 - 1,428,618]

Table 1: Reproducibility excluding major allele homozygotes, kappa and classic reproducibility ranges (in parentheses) and SNP count ranges [in brackets] for each MAF bin for the 228 duplicate pairs.

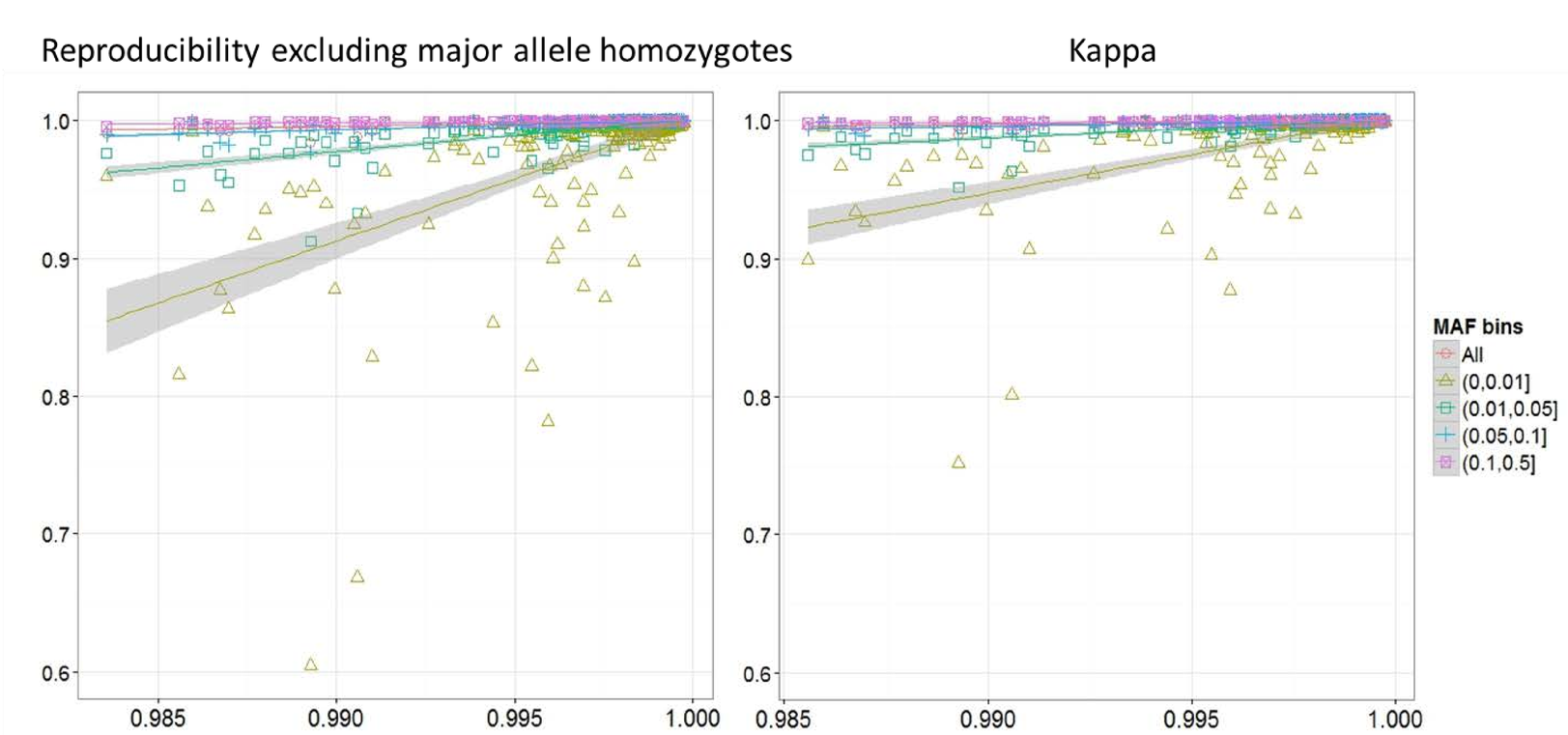

Figure 3: Scatter plots show reproducibility excluding major allele homozygotes (Y axis, left) and kappa statistic (Y axis, right) are correlated with sample call rate (X axis, the scale is the same as Figure 1)

Conclusions

1) The classic reproducibility suffers from the overwhelming effect of high concordance rate of major allele homozygotes on lower MAF bins and thus cannot accurately reflect the error rate for variants in lower MAF bins.

2) Both reproducibility excluding major allele homozygotes and kappa statistic can be used to assess data quality and are well correlated with sample call rates. As expected, SNPs in the lower MAF bins have lower reproducibility rates, especially when the sample call rate is low for one member of the pair. As the MAF of the SNPs increases, the reproducibility rate for that MAF bin increases and is less dependent on sample call rate.

3) Our data can be used to guide decisions on expected realistic error rates for variants of different allele frequencies. It can also be used to guide decisions on sample call rate cut-offs for data cleaning based on the type of analysis being done – i.e. rare versus common variant analyses may require different sample level call rate threshold cut-offs.