A test of genotyping platform bias for multiethnic case/control association studies merging external controls

Tiffany Tu [1] Rasheed Gbadegesin [2] Adebowale Adeyemo [3] Alejandro Ochoa [1]

[1]Computational Biology and Bioinformatics Program, Duke Ctr. for Statistical Genetics and Genomics, and Dept. of Biostatistics and Bioinformatics, Duke Univ., Durham, NC [2]Dept. of Pediatrics, Div of Nephrology, Duke Univ. Med. Ctr., Durham. NC [3]NIH, Bethesda, MD

Background

Every genotyping platform, including array, sequencing, and imputation, is known to have biased errors that vary between platforms. Merging data from different genotype platforms is a common step prior to conducting GWAS, especially in case-control studies where control samples from various sources are often combined to increase the sample size while reducing costs by reusing existing data. There is an increased risk of type I error when these platform-differential biases are correlated with a variable of interest, confounding the analysis.

Data	African	European	South Asian	Total
Array	301	274	248	823
WGS:TGP	661	503	489	1653
WGS:gnomAD	20740	34026	2419	57185

Table 1. Sample size from three major ancestry groups in array controls (from Duke collaborators) and Whole-Genome Sequencing (WGS) controls (from 1000 Genomes Project and gnomAD)

Merging controls from different platforms

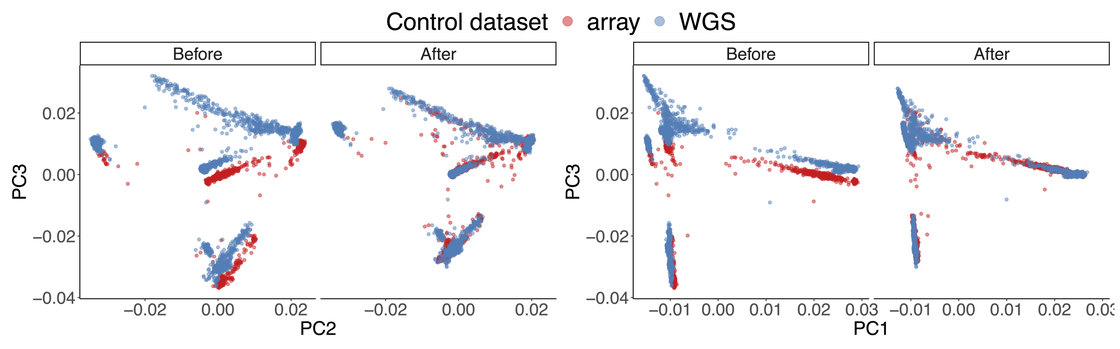

Figure 1. PCA is commonly used in population genetics to identify population substructure. PCA plots here show the genetic distribution of array control and WGS control before and after removing biased SNPs

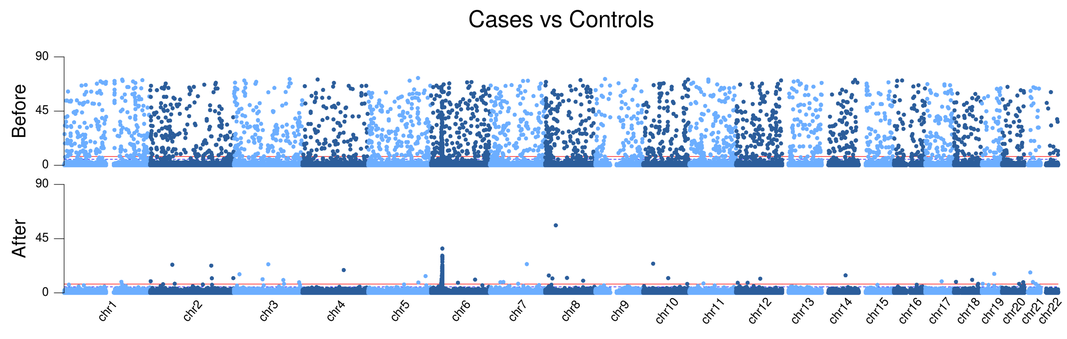

Figure 2. Stacked Manhattan plots of GMMAT association test results using array cases and controls merged with TGP WGS controls. Top plot shows before processing biased SNPs from merged controls and bottom plot shows post processing of biased SNPs.

SNP Classification

We performed a likelihood ratio test for each SNP, which models allele counts per ancestry and platform as Binomial with some group-specific allele frequency, and tested the **null hypothesis that allele frequencies for each ancestry are equal between array and WGS**. Each ancestry is treated as independent data, and the resulting log-likelihood ratio statistic has a Chi-squared distribution with 3 degrees of freedom (number of ancestries), which is used to calculate p-values.

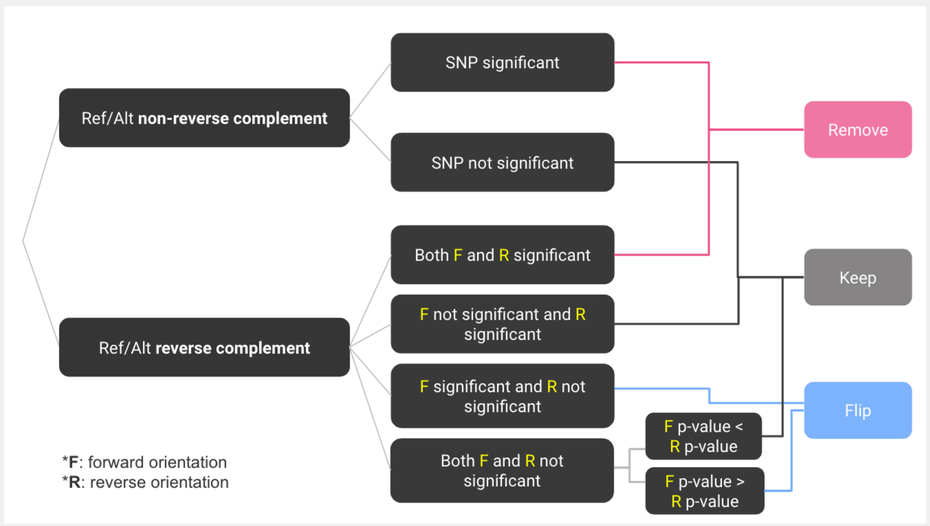

Figure 3. Decision Table for SNP classification

Likelihood Ratio Test

For each ancestry i in k total ancestries:

- x_{ij}: number of reference alleles of one ancestry in platform j
- n_{ij}: total number of ref and alt alleles for one ancestry in platform j
- p_{ij}: allele frequency for each SNP for one ancestry in platform j

$$x_{ij} \sim Binomial(n_{ij}, p_{ij})$$

$$H_0 : p_{i1} = p_{i2}, \;\; H_1 : p_{i1} \neq p_{i2}$$

The test statistics is derived from the total log probability mass function (pmf) of the null and alternative hypothesis for k ancestries and 2 platforms:

$$\hat{p}_{i,0} = \frac{x_{i1} + x_{i2}}{n_{i1} + n_{i2}}, \quad \hat{p}_{ij} = \frac{x_{ij}}{n_{ij}}$$

$$\lambda = -2 \sum_{i=1}^{k} \sum_{j=1}^{2} [\ell(x_{ij}, n_{ij}, \hat{p}_{i0}) - \ell(x_{ij}, n_{ij}, \hat{p}_{ij})] \;\sim \chi^2_k$$

Allele Frequency Test on array vs WGS

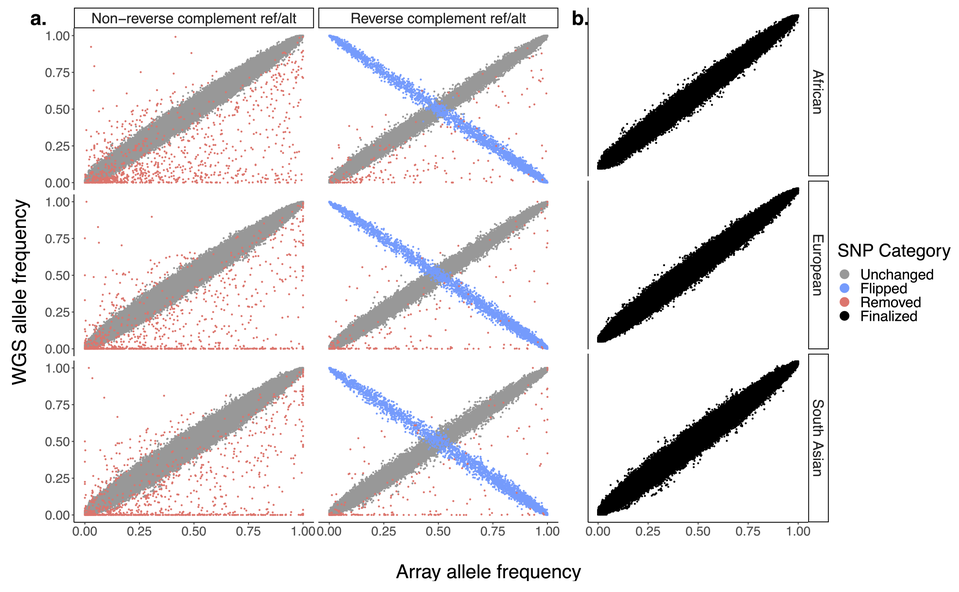

Figure 4. Allele frequency (AF) distribution of array controls vs TGP WGS controls. Each SNP is categorized into (1) removed, (2) flipped, or (3) unchanged according to Fig3 using a p-value threshold of 1e-10. (a) Non-reverse complement and reverse complement AF distribution before removing biased SNPs (b) AF distribution after processing biased SNPs according to its category.

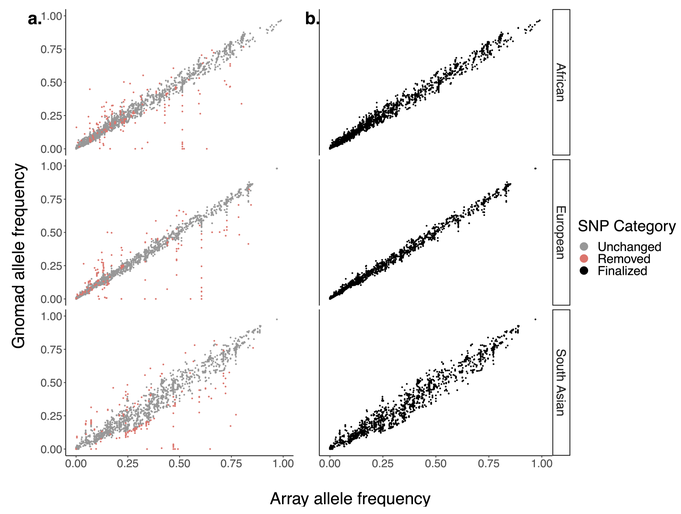

Figure 5. Categorization of SNPs in array controls vs gnomAD WGS controls using the same approach as a replication analysis. (a) AF distribution before processing biased SNPs (b) AF distribution post processing.

Summary

- Highlight challenges with merging datasets from various genotyping platforms
- Identify SNPs with platform-differential bias using novel statistical methods
- Critique differential biases that confound statistical analyses due to correlation with variables of interest

Funded by NIH NIAID grant 1U01AI152585-01