

# Robust Chi-square Test to Distinguish Population Stratification from Other Reasons of Hardy-Weinberg Disequilibrium

Yumi Jin<sup>1\*</sup> and Yimin Bao<sup>1\*</sup>

<sup>1</sup>SMART National Biobank of China, Shenzhen Medical Academy of Research and Translation, Shenzhen, China

\*Corresponding authors. yumijin@hotmail.com and baoyiming@smart.org.cn

Keywords: PopHWE, Hardy-Weinberg Equilibrium, chi-square test, population stratification, Wahlund effect

## Abstract

The Hardy-Weinberg Equilibrium (HWE) test is a standard component of quality control pipelines for genomic studies and databases. However, applying the HWE test to large, multi-ethnic datasets raises substantial challenges: natural departures from HWE due to population stratification and admixture can mask technical errors. We extended the classical chi-square goodness-of-fit test to account for the presence of population stratification. Assuming genotype datasets can be partitioned into relatively homogeneous strata, we changed the null hypothesis of the classical HWE test to assume that all strata are in HWE, and created a test statistic similar to the classical chi-square but applicable to stratified populations. This test statistic, together with a few others, can distinguish HWE deviations caused by population stratification from those resulting from other reasons. We have implemented this method in the software package PopHWE. Test results using simulated data and 1000 Genomes data showed that PopHWE outperforms the classical HWE meta-analysis using Stouffer's method, while being much faster, easier to implement, and more robust to small sample sizes, large numbers of subpopulations, and

the presence of ancestry grouping errors. Using PopHWE, HWE deviations caused by population stratification and those caused by genotyping errors can be displayed in different clusters on scatterplots. Furthermore, we demonstrated that the combined z-value of meta-analysis is biased toward excess heterozygosity for genotype datasets with up to moderately large sample sizes, and showed that replacing the combined z-value with its squared value could significantly improve the performance.

## Introduction

The rapid proliferation of high-throughput sequencing technologies has led to an explosive accumulation of genomic data. Large-scale repositories such as NCBI's database of Genotypes and Phenotypes (dbGaP), the UK Biobank (UKBB), and SMART-National Biobank of China (SMART-NBC) have been built to archive and distribute petabytes of sequence information. This deluge of data has made robust genetic data quality control (QC) increasingly critical but profoundly more challenging, as QC must now deal not only with per-sample technical artifacts, but also with cross-study heterogeneity, population stratification, and the computational burden of scaling QC pipelines to millions of samples. Hardy-Weinberg equilibrium (HWE) test is being used in almost all genomic studies and databases as a QC tool to filter out genotype data with potential errors (Hao et al., 2016; Marees et al., 2018; Kwong et al., 2021).

The HWE test calculates the p-value of obtaining the observed genotype counts for a variant in a dataset, under the null hypothesis that the data are sampled from a population in HWE. A deviation from HWE indicates potential data errors for that variant, which is filtered out if the p-value falls below a preselected threshold. However, the null hypothesis of HWE test also requires that the dataset be drawn from a randomly mating, homogeneous population (Graffelman & Weir, 2016; Meisner et al., 2021; Schaid et al., 2006; Abramovs et al., 2020; Hsu, 2026). When population stratification exists, as it does in almost all genomic dataset (Hao & Storey, 2019; Novembre & Peter, 2016; Pearman et al., 2022), a deviation from HWE is expected even in the absence of data errors. This phenomenon, known as the Wahlund effect, has become increasingly important as modern sequencing consortia include millions of individuals from multiple ancestries and admixed populations (Bycroft et al., 2018; Gaziano et al., 2016). In the presence of population stratification, the HWE p-value decreases as sample size increases. To avoid removing too many variants that are free of data errors, researchers adopt lower p-value thresholds, with  $P < 1 \times 10^{-6}$  commonly used in large studies (Fritsche et al., 2016).

Extremely low thresholds such as  $1 \times 10^{-50}$  or even  $1 \times 10^{-300}$  (near the underflow limit of the machine) are also frequently employed (Neamatzadeh et al., 2024; Sha & Zhang, 2011; Shyr et al., 2024). However, without accounting for population stratification, classical HWE test will produce inflated type I and/or type II errors regardless of the p-value threshold chosen.

The simplest and most widely used method to account for population stratification is the stratified HWE test, which divides samples into subsets of relatively homogeneous ancestry groups, performs HWE test within each subgroup and then either combines the results into one test statistic (Hao & Storey, 2019; Hemstrom et al., 2024; Kwong et al., 2021) or filters variants based on p-values from separated ancestry groups (Pearman et al., 2022). In the past decade, several methods (Hao et al., 2016; Hao & Storey, 2019; Kwong et al., 2021; Meisner & Albrechtsen, 2019; Meisner et al., 2021) have been developed to use individual-specific allele frequencies (IAF) to account for population stratification, without partitioning samples into discrete ancestry groups. IAFs are usually estimated by assuming a linear relationship between them and some continuous ancestry scores, such as principal components (PC). P-values are calculated based on the null hypothesis that the population is in HWE after the genetic ancestry of each individual are accounted for and IAFs correctly reflect the genetic ancestry. IAF-based approaches avoid the need for ancestry grouping but are considerably more computationally demanding than stratified HWE test. Currently, stratified HWE test remains the most widely adopted strategy for addressing population stratification. Nonetheless, many researchers continue to apply classical HWE test for dataset QC, in part because it is implemented as the default in widely used tools such as PLINK (Chang et al., 2015; Purcell et al., 2007) and VCFtools (Danecek et al., 2011). In a literature review, Pearman et al. (Pearman et al., 2022) estimated that the classical pooled method, referred to as “Out Combo” in their study, was used in 22% or more of 219 scientific publications involving HWE filtering of RADseq data. The authors also noted that this filtering approach can substantially affect downstream analyses, as loci with strong signals of population structure are more likely to be excluded. Given concerns that HWE filtering may adversely influence population structure inference, many researchers have opted to omit this step from their data quality control pipelines (Adams et al., 2026; Atsawawaranunt et al., 2026; Faure et al., 2025; Mon et al., 2025; Zeng et al., 2026; Zhou et al., 2026).

The low computational complexity of stratified HWE test makes it feasible for genome-wide analyses involving millions of individuals and variants. It also provides a

straightforward, interpretable metric that directly measures deviation from expected genotype frequencies, and it serves as a critical QC filter for identifying genotyping errors, sample contamination, and other data quality problems. It is currently used in most large genomic databases, including UKBB, (Bycroft et al., 2018), EGA (Freeberg et al. 2021), All of Us (Ramirez et al., 2022). As noted above, however, this method has several problems. Test results from all strata need to be combined to avoid substantial power loss, yet there is no easy, widely accepted method for doing so. Kwong et al (2021) described a meta-analysis approach that combines p-values using Stouffer's method ([Stouffer 1949](#)), but to our knowledge this approach is not commonly used. As Pearman et al. (2022) pointed out, many researchers instead perform HWE filtering using separate p-values from different ancestry groups, which inevitably leads to substantial type I and/or type II errors. A rapid, accurate, and easily automatable HWE test method is urgently needed for genomic studies, especially for large-scale genomic databases.

Here we describe how the classical chi-square goodness-of-fit test can be extended to account for the presence of population stratification, and report a method that separates HWE deviations caused by population stratification from those caused by other reasons. This method has been implemented in a publicly available software tool called PopHWE.

## Methods

### The classical chi-square goodness-of-fit HWE test

The classical chi-square test assesses whether observed genotype counts in a dataset deviate from expected values under random mating. Suppose there is a dataset with  $N$  individuals. Select one variant with reference and alternative alleles  $a$  and  $b$ , respectively, and denote  $N_{aa}$ ,  $N_{ab}$ ,  $N_{bb}$  as the observed counts (traditionally  $O$  is used for the observed numbers) of genotypes  $g \in \{aa, ab, bb\}$ , and denote the frequencies of the two alleles as  $p = N_a/N$  and  $q = 1 - p$ , where  $N_a$  is the count of the  $a$  allele. Under the null hypothesis that the dataset is sampled from a large population in HWE, denoted as  $H_{0C}$  in this manuscript, using  $p$  to estimate the population allele frequency, the expected genotype counts are  $E_{aa} = p^2N$ ,  $E_{ab} = 2pqN$ , and  $E_{bb} = q^2N$ . The classical chi-square test compares the observed genotype counts with the expected ones and calculates  $\chi^2$  test statistic as:

$$\chi^2 = \frac{(N_{aa} - E_{aa})^2}{E_{aa}} + \frac{(N_{ab} - E_{ab})^2}{E_{ab}} + \frac{(N_{bb} - E_{bb})^2}{E_{bb}}$$

(1)

Since the allele frequencies used to compute expected counts are calculated from the observed genotype counts,  $\chi^2$  is a function, denoted as *Chi2*, of the observed genotype counts:

$$\chi^2 = \text{Chi2}(N_{aa}, N_{ab}, N_{bb})$$

Under  $H_{0C}$  and provided that all expected counts are sufficiently large, the test statistic asymptotically follows a  $\chi^2$  distribution with 1 degree of freedom (*df*), denoted as  $\chi^2_{(1)}$  in this manuscript, i.e.,  $\chi^2 \sim \chi^2_{(1)}$ . One of the reasons why it has *df*=1 is that allele frequencies are estimated from the observed allele counts. Suppose the count of *aa* is changed by  $\delta$ , the *ab* count needs to be changed by  $-2\delta$  to keep the count of allele *a* unchanged, and hence the *bb* count also needs to be changed by  $\delta$ . If the sample size is small or variants are rare, the HWE exact test is more accurate, as it enumerates all possible genotype configurations compatible with the observed allele counts and sums the exact multinomial probabilities of configurations no more likely than the observed data, although this accuracy comes at the cost of increased computational complexity (Graffelman & Weir, 2016; Wigginton et al., 2005).

The HWE test is widely used in genomic studies as a QC tool. Variants showing Hardy-Weinberg disequilibrium (HWD), as indicated by a statistically significant p-values, are often flagged as potentially containing genotyping errors or low-quality data and are subsequently filtered out. However, HWD can arise from multiple causes beyond data artifacts. Among these, population stratification is one of the most important factors that cannot be overlooked in modern genomic studies.

## The effect of population stratification on HWE test

Suppose a dataset with size  $N$  includes two subsets with same size, but randomly sampled from two different homogeneous populations, whose reference allele frequencies (RAFs) are  $p_1$  and  $p_2$  for a variant, which are the unbiased estimations of the population RAFs. For the whole dataset, the overall RAF  $p$  will be the average of the two RAFs:  $p = (p_1 + p_2)/2$ , and the expected genotype counts calculated based on the assumption that the whole dataset being sampled from one homogenous population would be  $E_{w\_aa} =$

$p^2N$ ,  $E_{w_{ab}} = 2p(1-p)N$ , and  $E_{w_{bb}} = (1-p)^2N$ . However, these counts are incorrect due to the Wahlund effect.

The actual genotype counts can be calculated from the original assumption, i.e., the two subsets are sampled from two homogeneous populations. The total observed genotype counts are expected to be:

$$\begin{aligned} E_{s_{aa}} &= \frac{N}{2}(p_1^2 + p_2^2) \\ E_{s_{ab}} &= N[p_1(1-p_1) + p_2(1-p_2)] \\ E_{s_{bb}} &= \frac{N}{2}(q_1^2 + q_2^2) \end{aligned}$$

The difference between  $E_{s_{aa}}$  and  $E_{w_{aa}}$ , denoted as  $\delta_{aa}$ , will be:

$$\delta_{aa} = E_{s_{aa}} - E_{w_{aa}} = \frac{N}{2}(p_1^2 + p_2^2) - \frac{N}{4}(p_1 + p_2)^2 = \frac{N}{4}(p_2 - p_1)^2$$

Let  $\delta_p = p_2 - p_1$ , we have  $\delta_{bb} = \delta_{aa} = \delta_p^2 N/4$  and  $\delta_{ab} = -\delta_p^2 N/2$ .

When  $\delta_p = 0$ , the expected genotype counts assuming datasets are sampled from two homogenous populations are the same as those obtained with one population. If  $p_1 \neq p_2$ ,  $E(N_{ab})$  is fewer than  $E_{ab}$ , which means that population stratification results in heterogeneity deficiency (HetDef).

If the whole dataset is treated as from one population, the  $\chi^2$  value calculated using the classical chi-square test, denoted as  $\chi_c^2$ , will be:

$$\chi_c^2 = \frac{(N_{aa} - E_{w_{aa}})^2}{E_{w_{aa}}} + \frac{(N_{ab} - E_{w_{ab}})^2}{E_{w_{ab}}} + \frac{(N_{bb} - E_{w_{bb}})^2}{E_{w_{bb}}}$$

If  $p_1 = p_2$ , the case is the same as all samples in the dataset being sampled from one homogenous population, and we have  $\chi_c^2 \sim \chi_{(1)}^2$ . However, when  $p_1 \neq p_2$ ,  $\chi_c^2$  doesn't follow a  $\chi^2$  distribution, since the genotype count differences are not expected to be zero. The  $\chi_c^2$  value is expected to be:

$$E(\chi_c^2) = \frac{N\delta_p^4}{4(p_1 + p_2)^2} + \frac{N\delta_p^4}{2(p_1 + p_2)(2 - p_1 - p_2)} + \frac{N\delta_p^4}{4(2 - p_1 - p_2)^2}$$

(2)

These results indicate that classical HWE test fails to work when population stratification exists. Also note that even in the presence of population structure, for a particular variant, as long as its RAFs in all ancestry groups are almost the same, classical HWE test, either  $\chi^2$  goodness-of-fit test or HWE exact test, can still be used as a QC tool.

Note that  $E(\chi_c^2)$  is directly proportional to the sample size, which explains why lower p-value thresholds are needed when  $N$  increases. In an extreme case when  $p_1 \approx 0$  and  $p_2 \approx 1$ , we have  $E(\chi_c^2) \approx N$ .

If multiple subpopulations with different sample sizes are included in the population, it can be shown that  $E(N_{ab}) < E_{w_{ab}}$  based on Jensen's inequality, and  $E(N_{aa}) > E_{w_{aa}}$ ,  $E(N_{bb}) > E_{w_{bb}}$ , which result in  $E(\chi_c^2) > 0$ . Also in all cases, we have  $\chi_c^2 \leq N$  (Attia et al., 2010).

## HWE test with admixed populations

Now suppose all individuals in the dataset are sampled from one population recently admixed from two ancestry groups. Each individual  $i$  has a certain ancestry proportion, reflected by an ancestry score  $S$ , which ranges from 0 to 1 and uniformly distributed among individuals. Let's separate the individuals into  $k$  groups by binning them based on  $S$ , and for a locus assume RAF  $p_j$  in group  $j$  is determined by the mean  $S$  score  $S_j$  as  $p_j = p_0 + (p_1 - p_0)S_j$ , where  $p_0$  and  $p_1$  are the minimal and maximal RAFs in the two ancestries. Let  $q = 1 - p$ , and define  $q_i = 1 - p_i$ , for each individual  $i$ , genotype counts can be estimated using the following equations:

$$E(N_{aa}) = \frac{N}{k} \sum_{j=1}^k p_j^2, \quad E(N_{ab}) = \frac{N}{k} \sum_{j=1}^k 2p_j q_j, \quad E(N_{bb}) = \frac{N}{k} \sum_{j=1}^k q_j^2$$

When  $N, k$  are sufficiently large, we have:

$$E(N_{aa}) \approx \int_0^1 N[p_0 + (p_1 - p_0)x]^2 dx = \frac{N}{3}(p_0^2 + p_0 p_1 + p_1^2) \quad (3)$$

The counts of the other two genotypes can be calculated based on similarity and the constraint  $N_{ab} = N - N_{aa} - N_{bb}$ :

$$E(N_{bb}) \approx \frac{N}{3} (q_0^2 + q_0 q_1 + q_1^2) \quad (4)$$

$$E(N_{ab}) \approx \frac{N}{3} (1 + 2p_0 q_0 + 2p_1 q_1 - p_0 p_1 - q_0 q_1) \quad (5)$$

The classical chi-square can be calculated using the *Chi2* function:  $\chi_c^2 = \text{Chi2}(N_{aa}, N_{ab}, N_{bb})$ . If  $p_0 = p_1 = p$ , the three genotype counts will be  $p^2 N$ ,  $2pqN$ ,  $q^2 N$ , indicating that the population is in HWE and  $\chi_c^2 \sim \chi_{(1)}^2$ . When  $p_0 \neq p_1$ ,  $\chi_c^2$  doesn't follow chi-square distribution, which means classical  $\chi^2$  test cannot be applied to admixed populations. In the extreme case when  $p_0 = 0$  and  $p_1 = 1$ , we have  $N_{aa} = N_{ab} = N_{bb} = N/3$ , and the expected counts  $E_{aa} = E_{bb} = N/4$ , and  $E_{ab} = N/2$ , so:

$$\frac{E(\chi_c^2)}{N} \approx \frac{\left(\frac{1}{3} - \frac{1}{4}\right)^2}{\frac{1}{4}} + \frac{\left(\frac{1}{3} - \frac{1}{2}\right)^2}{\frac{1}{2}} + \frac{\left(\frac{1}{3} - \frac{1}{4}\right)^2}{\frac{1}{4}} = \frac{1}{9} \quad (6)$$

Note that  $\chi_c^2$  is directly proportional to sample size  $N$ . When  $N$  increases, the p-value calculated based on  $\chi_{(1)}^2$  distribution will decrease quickly, even if the degree of population stratification is small. In fact, p-value reflects the confidence level of HWD, not the degree of HWD, which should be measured by statistical metrics like inbreeding coefficient ( $F$ ) (Weir 1996), Disequilibrium parameter ( $D$ ) (Hernandez and Weir 1989, and Alpha ( $\alpha$ ) (Lindley 1998). The above calculations show that  $\chi_c^2/N$ , which is equal to  $F^2$ , can also be used to determine the degree of HWD. The above two examples represent two of the extreme cases. When individuals in a dataset have only two ancestries, if no individuals are admixed, the maximum  $\chi_c^2/N$  value is 1. If all individuals are randomly admixed, we have  $\max\left(\frac{\chi_c^2}{N}\right) \approx \frac{1}{9}$ , corresponding to  $\max(F) \approx \frac{1}{3}$ .

## Stratified HWE-test and meta-analysis

According to Kwong et al (2021), HWE meta-analysis across stratified populations is a commonly used approach to perform HWE test in the presence of population structure. Each stratum is analyzed separately for HWE using chi-square goodness-of-fit test or HWE-exact test. However, since many ancestry groups can be included in the test

dataset, it would be common to have strata with small sample sizes. Chi-square test requires rather large sample sizes and high minor allele frequencies (MAF). The results are valid only when all the expected genotype counts are 5 or more (Weir 1996, Grover et al 2010). HWE exact test, although more computationally intensive, is valid for small sample sizes and low allele frequencies. For a population with  $N$  individuals, under the hypothesis of HWE, the probability of observing genotype counts  $G = (N_{aa}, N_{ab}, N_{bb})$  given allele counts  $N_a$  and  $N_b$  is

$$P(G|N, N_a) = \frac{2^{N_{ab}} N!}{N_{aa}! N_{ab}! N_{bb}!} \times \frac{N_a! N_b!}{(2N)!}$$

Summing up the probabilities of all possible configurations that are at least as extreme as the observed data, a two-tailed p-value can be obtained.

We implemented the algorithm as described by Chang et al (2015) in python and verified the results with PLINK. To do meta-analysis, we converted the two-tailed p-values into z-values for each stratum, assigning a negative sign if the observed heterozygote count is more than the expected value (HetExc). The z-values from all the  $k$  strata with sample

size  $N_i$  for each stratum  $i$  are combined using Stouffer's method:  $Z_m = \frac{\sum_{i=1}^k (z_i \sqrt{N_i})}{\sqrt{\sum_{i=1}^k N_i}}$

(Kwong et al, 2021).

The combined  $Z_m$  is expected to follow a standard normal distribution, denoted as  $N(0,1)$ , when all subpopulations are in HWE, plus that each subpopulation has sufficiently large sample size. However, when sample size is small,  $Z_m$  does not follow an  $N(0,1)$  distribution, especially when MAFs are low. Assuming HWE and let  $p$  be the MAF in the population or its best estimation, randomly draw  $N$  individuals, the probability to observe  $N_a$  minor alleles is:

$$P(X = N_a) = \binom{N_a}{2N} p^{N_a} (1-p)^{2N-N_a}$$

When  $N$  is small, we can enumerate all the genotype configurations  $G$  and calculate the HWE exact p-values for all these configurations, and hence predict the expected z-value:

$$E(Z_m) = \sum_{g \in G} P(X = N_a) P(g|N_a) Z_g$$

where  $Z_g$  is the  $Z_m$  value when genotype configuration is  $g$ . The expected z-value is a function of sample size and MAF. Table 1 shows a simple example when  $N = 3$  and  $p =$

0.1. Note that  $E(Z_m) = 0.0206 > 0$ , indicating that the combined z-values don't follow  $N(0,1)$  distribution.

More  $E(Z_m)$  are displayed in Table 2. When  $N = 50$  and  $p = 0.1$ ,  $E(Z_m)$  is about 0.3. Also note that when  $N > 2000$ ,  $E(Z_m) \approx 0$  for  $p$  between 0.1 and 0.5. These results indicate that HWE exact test should work well for datasets with a few thousands or more samples. However, if the dataset needs to be partitioned into ancestry groups with hundreds or fewer samples, the p-values calculated by HWE exact test, as well as the meta-analysis relied on HWE exact test, might not be accurate. Of course, in this case, the classical  $\chi^2$  test will perform much worse than the HWE exact test.

## Modifying chi-square test to account for population stratification

Suppose a dataset includes  $k$  ancestry groups. Directly applying the classical HWE test to this dataset, the null hypothesis will be that the whole dataset is sampled from one population in HWE, which is neither accurate nor useful. Checking any group  $i$  with size  $N_i$ , denoting  $N_{iaa}$ ,  $N_{iab}$ , and  $N_{ibb}$  as the genotype counts, and  $p_i$ ,  $q_i$  as the frequencies of alleles  $a$  and  $b$ , under the null hypothesis that each group is sampled from a population in HWE, the expected genotype counts in this group will be:  $E_{iaa} = p_i^2 N_i$ ,  $E_{iab} = 2p_i q_i N_i$ , and  $E_{ibb} = q_i^2 N_i$ . For the whole dataset, the total genotype counts observed will be:  $N_{aa} = \sum_{i=1}^k N_{iaa}$ ,  $N_{ab} = \sum_{i=1}^k N_{iab}$ , and  $N_{bb} = \sum_{i=1}^k N_{ibb}$ . Summing up the expected counts in all groups, we get:  $S_{eaa} = \sum_{i=1}^k E_{iaa}$ ,  $S_{eab} = \sum_{i=1}^k E_{iab}$ , and  $S_{ebb} = \sum_{i=1}^k E_{ibb}$ , where  $S_{eaa}$ ,  $S_{eab}$ , and  $S_{ebb}$  are the sums of the expected genotype counts in all groups under null that all groups are in HWE, denoted as  $H_{0P}$  in this manuscript.

Classical  $\chi^2$  test uses the expected genotype counts calculated under the null that the whole dataset is sampled from a population in HWE, or  $H_{0C}$ , i.e.,

$$\chi_c^2 = \frac{(N_{aa} - E_{aa})^2}{E_{aa}} + \frac{(N_{ab} - E_{ab})^2}{E_{ab}} + \frac{(N_{bb} - E_{bb})^2}{S_{bb}}$$

However, the more useful null is that the individual ancestry groups, not the whole dataset, are sampled from populations in HWE. Under  $H_{0P}$ , the expected genotype count  $E(N_{aa})$  will be:

$$E(N_{aa}) = E\left(\sum_{i=1}^k N_{iaa}\right) = \sum_{i=1}^k E(N_{iaa}) = \sum_{i=1}^k E_{iaa} = S_{eaa}$$

The expected counts of the other two genotypes can be calculated similarly. Using the expected genotype counts estimated under  $H_{0P}$ , we can calculate a new statistic metric, denoted as  $\chi_s^2$ :

$$\chi_s^2 = \frac{(N_{aa} - S_{eaa})^2}{S_{eaa}} + \frac{(N_{ab} - S_{eab})^2}{S_{eab}} + \frac{(N_{bb} - S_{ebb})^2}{S_{ebb}} \quad (7)$$

Under this new null, the difference between the observed and expected genotype counts, such as  $N_{aa} - S_{eaa}$ , is expected to be zero, i.e.,  $E(\chi_s^2) = 0$ . However, below we will show that  $\chi_s^2$  does not follow a chi square distribution, especially for the variants with large allele frequency differences among subpopulations.

We can define another statistic metric, the expected chi square with population stratification as:

$$\begin{aligned} \chi_e^2 &= E(\chi_c^2) = \text{Chi2}(S_{eaa}, S_{eab}, S_{ebb}) \\ &= \frac{(S_{eaa} - E_{aa})^2}{E_{aa}} + \frac{(S_{eab} - E_{ab})^2}{E_{ab}} + \frac{(S_{ebb} - E_{bb})^2}{E_{bb}} \end{aligned} \quad (8)$$

We also define another test statistic  $\chi_d^2$  as:

$$\chi_d^2 = \chi_c^2 - \chi_e^2$$

Based on Jensen's inequality, we have  $S_{eab} \leq E_{ab}$ , which means that population stratification results in HetDef, which is known as Wahlund effect.

Note that the expected genotype count used for calculating  $\chi_e^2$  are the same as those used for  $\chi_c^2$ , since the estimated allele frequencies are the same.  $\chi_e^2$  doesn't follow  $\chi^2$  distribution under  $H_{0P}$  either. It indicates the expected HWD caused by population stratification, which will be shown below to be useful in data QC.

## Statistic metrics for testing HWE in the presence of population stratification

Equation (1) is the classical Pearson chi-square form. However,  $\chi^2$  can also be expressed in other forms. Using the disequilibrium parameter ( $D$ ) (Hernandez and Weir 1989, Schaid et al., 2006),  $\chi^2$  can be calculated as:

$$\begin{aligned}
 \chi^2 &= \frac{(ND)^2}{Np^2} + \frac{(2ND)^2}{2Npq} + \frac{(ND)^2}{Nq^2} \\
 &= ND^2 \left( \frac{1}{p^2} + \frac{2}{pq} + \frac{1}{q^2} \right) \\
 &= ND^2 \left( \frac{q^2 + 2pq + p^2}{p^2q^2} \right) = \frac{ND^2}{p^2q^2} \\
 &= \frac{(N_{aa} - Np^2)^2}{Np^2q^2} = \frac{(N_{aa} - E_{aa})^2}{Np^2q^2}
 \end{aligned}
 \tag{9}$$

Denoting  $Z = \frac{N_{aa} - E_{aa}}{\sqrt{Npq}}$ ,  $Z$  is the z-score of a normal distribution and expected to follow  $N(0,1)$  under null of HWE. Since  $Z \sim N(0,1)$ , the denominator corresponds to the standard deviation of an independent random variable. Equation (9) is the same as that of the test statistic  $T_D$  used by Garg et al (2026). Note that both  $Z$  and  $\chi^2$  can be used to test for HWE, but  $Z$  as calculated above is more informative than  $\chi^2$  since it is signed and hence can be used for one-tailed test.

In the presence of population structure, suppose the dataset can be stratified to  $k$  subgroups as above, the expected  $aa$  genotype count is  $S_{aaa} = \sum_{i=1}^k E_{iaa}$ , under  $H_{0P}$ . Denoting  $\sigma_i = \sqrt{N_i p_i q_i}$  for each subgroup, the overall standard deviation will be  $\sigma = \sqrt{\sum_{i=1}^k \sigma_i^2}$ . We can create a test statistic  $Z_p$  as:

$$Z_p = \frac{N_{aa} - S_{aaa}}{\sqrt{\sum_{i=1}^k (N_i p_i^2 q_i^2)}} = \frac{N_{aa} - \sum_{i=1}^k (N_i p_i^2)}{\sqrt{\sum_{i=1}^k (N_i p_i^2 q_i^2)}}
 \tag{10}$$

and another one, denoted as  $\chi_p^2$ :

$$\chi_p^2 = Z_p^2 = \frac{(N_{aa} - S_{aaa})^2}{\sum_{i=1}^k (N_i p_i^2 q_i^2)}
 \tag{11}$$

Comparing  $Z_p$  with  $Z$ , the latter is the z-score of a normal distribution, which can be expressed as  $\frac{x-\mu}{\sigma}$ , for a random variable  $x$  with mean  $\mu$ , and standard deviation  $\sigma$ . For chi-square test with population allele frequencies being estimated from the observed genotype counts,  $N_{aa}$ ,  $E_{aa}$ , and  $\sqrt{N}pq$  correspond to  $x$ ,  $\mu$ , and  $\sigma$ , respectively, which is why  $Z \sim N(0, 1)$  under  $H_{0C}$ . Under the new null  $H_{0P}$ , for the same random variable, the correct mean and standard deviation are shown in Equation (10), and hence  $Z_p$  is also expected to asymptotically follow  $N(0,1)$ , which also implies  $\chi_p^2 \sim \chi_{(1)}^2$ . Therefore, the classical chi-square test has been extended to deal with the existence of population stratification. If the null to be tested is  $H_{0C}$ , i.e.,  $k = 1$ , or  $\forall i, j \in \{1, \dots, k\}: p_i = p_j$ , Equation (10) is reduced to the classical Pearson chi-square form.

Note that whether a dataset is heterozygosity excess or deficient depends on which hypothesis is used. The sign of the  $Z_p$  value calculated in Equation (10) reflects whether the observed heterozygotes are more than the count expected under  $H_{0P}$ . However, HetExc or HetDef is determined based on the classical null hypothesis  $H_{0C}$ , which is also adopted by us in this manuscript.

### Comparing $\chi_s^2$ , $\chi_p^2$ and $\chi_c^2$ in excess and deficiency of heterozygosity

We have shown that existence of population stratification results in HetDef. HWD can also arise from HetExc, which might be the more common reason in sequencing data (Abramovs et al. 2020). The two test statistics  $\chi_s^2$  and  $\chi_c^2$  perform differently when in HetDef and in HetExc. If a variant has genotype counts of HetExc, then  $\chi_s^2$  is greater than  $\chi_c^2$ . Below is the proof.

Based on Jensen's inequality, we have  $E_{ab} \geq S_{eab}$ , and hence  $E_{aa} \leq S_{eaa}$  and  $E_{bb} \leq S_{ebb}$ . The genotype counts expected under the null that whole population is in HWE are:  $E_{aa} = p^2N$ ,  $E_{ab} = 2pqN$ , and  $E_{bb} = q^2N$ . Assuming HetExc genotype counts, let  $N_{ab} = (2pq + 2\delta)N$ , and hence  $N_{aa} = (p^2 - \delta)N$ , and  $N_{bb} = (q^2 - \delta)N$ , we will have  $pq \leq \delta \leq \min(p^2, q^2)$ . Note that  $\delta = -D$ , where  $D$  is the disequilibrium parameter (Shoemaker et. al., 1989). Let  $S_{eab} = E_{ab} - 2\varepsilon N$ ,  $S_{eaa} = E_{aa} + \varepsilon N$ , and  $S_{ebb} = E_{bb} + \varepsilon N$ , where  $0 \leq \varepsilon \leq pq$ .  $\chi_s^2$  can be calculated as:

$$\chi_s^2 = \frac{N(p^2 - \delta - p^2 - \varepsilon)^2}{p^2 + \varepsilon} + \frac{N(2pq + 2\delta - 2pq + 2\varepsilon)^2}{2pq - 2\varepsilon} + \frac{N(q^2 - \delta - q^2 - \varepsilon)^2}{q^2 + \varepsilon}$$

$$\frac{\chi_s^2}{N} = \frac{(\delta + \varepsilon)^2}{(p^2 + \varepsilon)} + \frac{2(\delta + \varepsilon)^2}{pq - \varepsilon} + \frac{(\delta + \varepsilon)^2}{(q^2 + \varepsilon)}$$

For the middle term, it is obvious that the value increases when  $\varepsilon$  increases. For the first term, we calculate the derivative:

$$\begin{aligned} \frac{\partial (\delta + \varepsilon)^2}{\partial \varepsilon (p^2 + \varepsilon)} &= \frac{2(\delta + \varepsilon)(p^2 + \varepsilon) - (\delta + \varepsilon)^2}{(p^2 + \varepsilon)^2} \\ &= \frac{\delta(2p^2 - \delta) + 2p^2\varepsilon + \varepsilon^2}{(p^2 + \varepsilon)^2} \end{aligned}$$

Since  $\delta, \varepsilon, p$  are all positive and  $2p^2 - \delta > p^2 - \delta \geq 0$ , the derivative is positive for all  $\varepsilon$  values. Therefore, we have  $\chi_s^2 \geq \chi_c^2$  when HetExc. Note that  $\chi_c^2$  is a special case of  $\chi_s^2$  when  $\varepsilon = 0$ . When HetDef,  $\chi_s^2$  can be greater or smaller than  $\chi_c^2$ .

For  $\chi_p^2$ , our test results (shown below) with simulated and real data also imply that  $\chi_p^2 > \chi_c^2$  when HetExc is possible true, but it is a conjecture needing to be proved.

## Effects of unaccounted population stratification on HWE test results

We have shown that the expected chi-square value  $\chi_e^2$  under  $H_{0p}$  can be calculated using Equation (8). Here we are going to show that if population stratification exists in at least one of the subpopulations,  $\chi_c^2$  will be greater than  $\chi_e^2$  under the null that no population stratification exists in any of the subpopulations.

Assume there is a genotype dataset  $A$  with sample size  $N$  and the frequencies of the two alleles ( $a, b$ ) being  $p$  and  $q$ . This dataset is partitioned into  $k$  distinct ancestry groups, i.e.,  $A = A_1 \cup A_2 \dots \cup A_k$ . For any group  $A_i, i \in \{1, \dots, k\}$ , denote  $N_i$  as the sample size, and  $p_i$  and  $q_i$  as the frequencies of the two alleles and assume it is in HWE. The genotype counts in  $A$  will be  $N_{iaa} = N_i p_i^2, N_{iab} = 2N_i p_i q_i, N_{ibb} = N_i q_i^2$ , and the total genotype counts in the dataset will be  $N_{aa} = \sum_{i=1}^k N_{iaa}, N_{ab} = \sum_{i=1}^k N_{iab}$ , and  $N_{bb} = \sum_{i=1}^k N_{ibb}$ ,

and the classical chi square value is calculated as  $\chi_c^2 = Chi2(N_{aa}, N_{ab}, N_{bb})$ . Due to Wahlund effect, we will have HetDef, i.e.,  $N_{ab} < E_{ab}$ .

Now assume there is a similar dataset  $B$ , with same sample size  $N$ . Dataset  $B$  includes  $k$  ancestry groups, of which the first  $k - 1$  groups are the same as in  $A$ , i.e.,  $B = A_1 \cup A_2 \dots \cup A_{k-1} \cup B_k$ , where  $B_k$  has same size  $N_k$  and same allele frequencies  $p_k$  and  $q_k$  as in  $A_k$ . Therefore, the overall allele frequencies in  $B$  are  $p$  and  $q$ , same as those in  $A$ , and hence the expected genotype counts used to calculate  $\chi_c^2$  remain the same, i.e.,  $E_{aa} = Np^2$ ,  $E_{ab} = 2Npq$ ,  $E_{bb} = Nq^2$ .

Suppose  $B_k$  is *not* in HWE. Instead, it includes two ancestry groups in HWE,  $B_{k1}$  and  $B_{k2}$ , with sizes  $N_{k1}$  and  $N_{k2}$ , and RAFs  $p_{k1}$  and  $p_{k2}$ . When  $p_{k1} \neq p_{k2}$ , based on Jensen's inequality, we have  $N_{k1ab} + N_{k2ab} < N_{kab}$ , where  $N_{kab}$  is the  $ab$  genotype count in subpopulation  $A_k$  and  $N_{k1ab}$  and  $N_{k2ab}$  are the  $ab$  counts in  $B_{k1}$  and  $B_{k2}$ . Let  $\Delta N_k = 2(N_{kab} - N_{k1ab} - N_{k2ab})$ , the differences of the  $aa$  or  $bb$  genotype counts will be  $-\Delta N_k$ .

For dataset  $A$ , the classical  $\chi^2$  is:

$$\chi_c^2 = \frac{(N_{aa} - E_{aa})^2}{E_{aa}} + \frac{(N_{ab} - E_{ab})^2}{E_{ab}} + \frac{(N_{bb} - E_{bb})^2}{E_{bb}}$$

For dataset  $B$ , the classical  $\chi^2$  becomes:

$$\chi_c'^2 = \frac{(N_{aa} - E_{aa} + \Delta N_k)^2}{E_{aa}} + \frac{(N_{ab} - E_{ab} - 2\Delta N_k)^2}{E_{ab}} + \frac{(N_{bb} - E_{bb} + \Delta N_k)^2}{E_{bb}}$$

Since  $E(N_{ab} - E_{ab}) < 0$  and  $E(N_{aa} - E_{aa}) = E(N_{bb} - E_{bb}) > 0$ , due to Wahlund effect, we have  $E(\chi_c'^2) > E(\chi_c^2)$ . Therefore, we have proved that for a dataset including any number of subgroups, if any subgroup is further split into two groups with different RAFs,  $\chi_e^2$  will increase. The increase in  $\chi_e^2$  value accumulates every time when one more subgroup is split. In general, for a genotype dataset randomly sampled from one or more populations, under the null hypothesis that each population is in HWE,  $\chi_e^2$  is the unbiased estimation of  $\chi_c^2$ . Alternatively, when one or more populations are not in HWE,  $\chi_e^2$  is a biased estimation that underestimates  $\chi_c^2$ . The difference between the mean  $\chi_c^2$  and  $\chi_e^2$  reflects the degree of unaccounted population stratification that remains after the ancestry grouping.

As for  $\chi_s^2$  and  $\chi_p^2$ , when population stratification is not fully accounted for, both are expected to shift to the side of HetDef. In the extreme case when population stratification

in the dataset is completely unaccounted for, we have  $\chi_p^2 = \chi_s^2 = \chi_c^2$  as shown above. In addition, we will have  $\chi_e^2 = 0$  in this case.

## The PopHWE software

We have developed a software tool called PopHWE to perform stratified HWE test using the improved and extended chi square goodness-of-fit method to account for population stratification. For human genotype data, PopHWE uses GrafAnc (Jin et al., 2026) to separate individuals into ancestry groups, then calculates genotype counts for each group and performs HWE test using statistics including  $\chi_p^2$ ,  $\chi_e^2$  and  $\chi_c^2$ . It allows the users to select and/or merge ancestry groups before HWE test. PopHWE also accepts genotype counts pre-aggregated using any ancestry grouping methods, for *homo sapiens* or any other diploid species. In addition, the package includes some visualization tools for users to plot and check the results. PopHWE is publicly available for downloading from <https://github.com/jimmy-penn/PopHWE>.

## Testing PopHWE using simulated data

We simulated genotype datasets including different numbers of variants and different sample sizes. Each dataset included different numbers ( $k$ ) of subsets, and each subset was randomly sampled from a population with infinitely large sample size in HWE. In each dataset, most of the variants had correct genotypes and could be treated as negative controls. A certain percentage of variants were simulated to include errors, with randomly generated genotype counts without HWE restriction and unrelated to the allele frequencies in the populations. All these variants were treated as with errors (positives), even if they happened to be in HWE by chance. Same method was used to simulate a dataset without population structure, by simply setting  $k = 1$ .

PopHWE was tested using the simulated data and compared to the meta-analysis method performed using HWE exact test as describe above. The z-value of each stratum was calculated from the p-value based on  $N(0,1)$  distribution, and assigned a negative sign if the stratum had HetDef. Z-values from all strata were combined using Stouffer's method and denoted as  $Z_m$ . In order to compare  $Z_m$  to the  $\chi^2$  values generated by PopHWE, we calculated the square of  $Z_m$  and denoted it as  $\chi_m^2 = Z_m^2$ , and plotted histograms showing its distribution alongside with that of  $\chi_p^2$  and  $\chi_s^2$ . On the other hand, since the normal

distribution curves are easier to be visually inspected than the  $\chi^2$  curves, we also converted  $\chi^2$  values to z-values by taking square root of them. Note that  $Z_p$  can be calculated using Equation (10), and it is already signed. The z-values corresponding to the other  $\chi^2$  values are calculated as:  $Z_s = \sqrt{\chi_s^2}$ ,  $Z_c = \sqrt{\chi_c^2}$ , and  $Z_d = \sqrt{|\chi_d^2|}$ , with  $Z_s$  set to negative if  $N_{ab} > S_{eab}$ , and  $Z_c$  and  $Z_d$  assigned negative signs for HetDef variants. In addition, for visualization convenience we created  $Z_{rm}$ ,  $Z_{rp}$ ,  $Z_{rs}$ ,  $Z_{rc}$ ,  $Z_{rd}$  by randomly assigning signs to the corresponding z-values. These z-values with random signs faithfully reflect the corresponding  $\chi^2$  values when sample sizes are sufficiently large.

## Application of HWE test methods to 1000 Genomes Project data

Two datasets from 1000 Genomes Project were used to test PopHWE. First, we extracted the genotypes of the ~282K GrafAnc “Ancestry SNPs” from the 1KGP WGS data. These SNPs are independent, ancestry informative SNPs preselected by GrafAnc (Jin et al., 2026) for ancestry inference. We also used the VCF file including genotypes of chromosome 22 for testing. Only the “best guess” GT values were used.

## Results

### HWE test on a large dataset without population structure

We simulated a dataset of 10,000 individuals randomly sampled from a population of infinite size under HWE. The dataset contained 100,000 independent biallelic SNPs, 90% of which had no genotyping errors in any sample, with MAFs uniformly distributed between 0.1 and 0.5 in the population. The remaining 10% of SNPs contained genotyping errors, with genotypes randomly drawn from a uniform distribution over the three genotype classes, independent of the population allele frequencies.

Fig. 1 compares the test results of PopHWE with those of the meta-analysis. When the dataset is not stratified, we have  $\chi_p^2 = \chi_s^2 = \chi_c^2$ ,  $\chi_e^2 = 0$ , and  $\chi_m^2$  and  $Z_m$  are the results of one HWE exact test without meta-analysis. The top row shows that both  $\chi_p^2$  and  $\chi_m^2$  follow the  $\chi^2_{(1)}$  distribution closely. The distribution curves in the 2<sup>nd</sup> row show that the corresponding z-values follow the  $N(0,1)$  distribution closely. The discontinuities observed in the distribution curve of  $Z_m$  values (Fig. 1D) can be explained by the discrete nature of the HWE exact test p-values.

The bottom two rows in Fig. 1 show the distributions of the test statistics for SNPs with genotyping errors. Fig. 1E confirms that the maximum  $\chi_c^2$  equals to  $N$ , as predicted in Equation (2), and hence the maximum z-value is  $\sqrt{N}$  (Fig. 1G). The limits in  $\chi_m^2$  and  $Z_m$  values (Figs. 1F and 1H) result from underflow when p-values are calculated using HWE exact test.

## HWE test on a large dataset with two subpopulations

To test the methods in the presence of population stratification, we simulated another dataset with same size ( $N = 10,000$ ), but including two subgroups, each containing 5,000 individuals randomly sampled from a large but independent population in HWE. For all SNPs, the RAFs were randomly distributed between 0.1 and 0.9, and 10% of the SNPs were simulated as with errors as described above.

We plotted histograms to compare the HWE test results of PopHWE and those of meta-analysis (Fig. 2). Note that the classical  $\chi^2$  values don't follow  $\chi^2$  distribution in the presence of population stratification (Fig. 2J), indicated by the long tail in the histogram. On the contrast, the PopHWE  $\chi_p^2$ ,  $\chi_s^2$  (Figs 2A, 2D) and the meta-analysis  $\chi_m^2$  values (Fig. 2G) all appeared to follow  $\chi_{(1)}^2$  distribution.  $\chi_d^2$  values can be both positive and negative (Fig. 2M), since they were calculated as the difference between  $\chi_c^2$  and  $\chi_e^2$ .

Since  $\chi_{(1)}^2$  distribution is highly skewed and difficult to be plotted, we converted the  $\chi^2$  values to z-values and plotted histograms for them. Fig. 2C, Fig. 2F, and Fig. 2I show that  $Z_{rp}$ ,  $Z_{rs}$ , and  $Z_{rm}$  follow  $N(0, 1)$  distribution closely, confirming that the corresponding  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$  follow  $\chi_{(1)}^2$  distribution. On Fig. 2L and Fig. 2O, we can see that  $Z_{rc}$  and  $Z_{rd}$  do not follow normal distribution, reflecting the facts that  $\chi_c^2$  and  $\chi_d^2$  do not follow  $\chi_{(1)}^2$  distribution (Fig. 2J and Fig. 2M).

Fig. 3 displays the scatterplots of  $\chi_c^2$  versus  $\chi_e^2$ ,  $\chi_p^2$ ,  $\chi_s^2$ ,  $\chi_m^2$ , and absolute  $\chi_d^2$  values, in the five rows from top to bottom, respectively. Variants without genotyping errors are colored black, and those with errors are red. The second and third columns show HetDef and HetExc variants, respectively. The panels in the last column show the  $\chi^2$  values between 0 and 100. From the first row we can see that there is a strong correlation between  $\chi_c^2$  and  $\chi_e^2$ , which also implies that more ancestry-informative variants would be filtered out if p-value threshold were set based on  $\chi_c^2$  values, as predicted above. These plots show that both population stratification and data errors can cause HWE deviations, as indicated by

$\chi_c^2$  values. By plotting  $\chi_e^2$  against  $\chi_c^2$ , we can roughly see which HWE deviations were caused by population stratification and which are from genotyping errors.

The middle three rows show that variants without errors had much lower  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$  values in general, compared to those with errors, indicating that these scores can be used for QC of genotype data in the presence of population stratification. The second and third rows show that results of  $\chi_p^2$  and  $\chi_s^2$  values are very similar.

$\chi_e^2$  values alone cannot be used for QC filtering, but  $\chi_d^2$ , the difference between  $\chi_c^2$  and  $\chi_e^2$ , can. Plots in the bottom row show that most of the variants with errors could be separated from those without errors in this test case by  $\chi_d^2$ , although it seems to have lower discriminative power than  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$ .

The plots in the 4<sup>th</sup> row show that  $\chi_m^2$  values had an upper limit around 3000, corresponding to  $p \approx 10^{-300}$  in each population, the underflow limit of the PLINK HWE exact test.  $\chi^2$  values from PopHWE don't have similar limits since no extremely small or large values are involved in the calculations.

The middle two columns show that the  $\chi^2$  values of the HetDef (Column 2) and HetExc (Column 3) variant had different distribution patterns. Fig. 3K verifies the theorem that  $\chi_s^2 > \chi_c^2$  under HetExc, as proved in the “Method” section. Fig. 3J shows that for most of the variants, especially those with high  $\chi_s^2$  values, it is true that  $\chi_s^2 < \chi_c^2$  under HetDef. For  $\chi_p^2$  values, it also seems true that  $\chi_p^2 > \chi_c^2$  when HetExc.

Fig. 3. shows that  $\chi_c^2$ ,  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$  can be used to separate variants with genotyping errors from those without errors. As for  $\chi_e^2$ , although it would not be a good test statistic alone, it can help us see which HWE deviations were caused by population stratification and which were from other reasons, after being plotted against  $\chi_c^2$ . The difference between  $\chi_c^2$  and  $\chi_e^2$ , denoted as  $\chi_d^2$ , reflects the HWE deviations excluding the effect of population stratification, and hence can also be used to as a statistical metric for genotype data QC.

The discriminative powers of these test statistics are compared in Fig. 4. The results of  $\chi_s^2$  is not plotted since they are very similar to those of  $\chi_p^2$ . Since the cases with genotyping errors were completely randomly generated, most of them had high degrees of HWD, and hence easily to be detected. Even the classical HWE test statistic  $\chi_c^2$  has high discriminative powers. In the last two columns, we filtered out some easy-to-detect genotyping errors and compared statistical metrics with the more difficult cases, in which

only the data points with  $\chi^2 < 1000$  (column 2) and  $\chi^2 < 100$  (column 3) were included. In all these three cases,  $\chi_p^2$  had the highest AUC values.

The ROC curves for  $\chi_m^2$  exhibited irregularities, with the true positive rate remaining low until the false positive rate approached nearly 1. As with the upper bound of  $\chi_c^2$  shown in Fig. 3, this issue stemmed from numerical underflow during computation. When sample sizes are large, p-values from HWE tests can become extremely small. Most existing HWE test software, including PLINK, only computes p-values down to approximately  $10^{-300}$ ; any p-value below this threshold is truncated to 0, and the corresponding z-value is set to the value associated with this minimum p-value. In the context of HWE meta-analysis, underflow is not uncommon when there are large subpopulations. When there are only two subpopulations with similar sizes, if the directions of heterozygosity differ between the two subpopulations, the z-values may cancel each other out, yielding a combined z-value close to 0. For instance, consider genotype counts of 4000/0/6000 and 0/10000/0 in two populations. Due to underflow, both p-values are set to 0, and the z-values are assigned the same magnitude (corresponding to the minimum p-value) but with opposite signs. Given equal sample sizes, the combined z-value becomes 0, despite the clear deviation from HWE. Had the p-values been computed accurately, the combined z-value would have been substantial. Thus, this represents a false negative, where a genuine HWD is incorrectly classified as being in equilibrium.

This elevated false negative rate is mitigated when more subpopulations are included or when subpopulation sample sizes vary. Fig. S1 presents similar ROC curves for a dataset comprising 100 subpopulations, each with a sample size of 100. In this scenario,  $\chi_m^2$  had discriminative powers higher than  $\chi_d^2$ , but  $\chi_p^2$  remained the most powerful test statistic.

## HWE test on simulated datasets including different numbers of subpopulations

To test how number of subpopulations affect the HWE test results, we simulated 6 datasets with total sample size 10,000, but including 10, 20, 50, 100, 200, and 500 subpopulations, respectively. Each subpopulation was randomly sampled from a population in HWE, as described above. Fig. 5 plots the test results of PopHWE and meta-analysis, showing the histogram of  $Z_p$ ,  $Z_m$ ,  $\chi_p^2$  and  $\chi_m^2$  values, from left to right columns. When there were 10 subpopulations,  $Z_p$  and  $Z_m$  values still seem to follow  $N(0, 1)$ , similar to those with 2 subpopulations (Fig. 2). However, when more subpopulations

were included in the dataset, the  $Z_m$  and  $Z_p$  histograms shifted to left, i.e., the HetExc side, and hence no longer followed  $N(0, 1)$  (Fig. 5, first two columns). However, the  $\chi_p^2$  and  $\chi_m^2$  values (the last two column) seemed to still follow  $\chi_{(1)}^2$ , when there were 100 or fewer subpopulations

Results in Fig. 5 imply that  $Z_m$  is not a good test statistic metric for HWE meta-analysis, especially when the datasets include many ancestry groups. Instead of using  $Z_m$  values to do one-tailed or two-tailed tests, a better way is to convert them to the unsigned  $\chi_m^2$  values and perform one-tailed test, which is equivalent to using the absolute  $Z_m$  values for one-tailed test with  $N(0, 1)$  distribution.

Fig. 5 also shows that both  $\chi_m^2$  and  $\chi_p^2$  can be used to calculate p-values with the  $\chi_{(1)}^2$  distribution model, even if the data is partitioned into 100 strata, each of which containing only 100 individuals. However, ROC curves in Fig. 6 show that  $\chi_p^2$  might have greater discriminative power than  $\chi_m^2$ . When there were 10 subpopulations, the AUC values (when only test statistic value  $> 100$  were considered) of  $\chi_p^2$  and  $\chi_m^2$  were very close, with AUC difference of  $\chi_p^2$  and  $\chi_m^2$  being about 0.001. However, when number of subpopulations increased to 500, the AUC difference increased to 0.024, indicating  $\chi_p^2$  is more robust to dataset partitioning than the combined test statistic score of the HWE meta-analysis.

## HWE test on simulated datasets including fixed allele frequencies in each subpopulation

In order to check the effects of allele frequency differences on the HWE test results, we simulated datasets with total sample size 10,000 and two subpopulations with the same sample size, each of which was randomly sampled from a population with same allele frequencies for all the 100,000 variants. These datasets were tested with PopHWE and meta-analysis. Fig. 7 includes the results of 6 datasets with the following RAF differences between the two populations: 0.1, 0.2, 0.3, 0.8, 0.94, and 0.98, from the top to the bottom row. The three columns plot the histograms of PopHWE  $Z_{rs}$ ,  $Z_{rp}$  and meta-analysis  $Z_{rm}$  values, respectively, reflecting the corresponding  $\chi_s^2$ ,  $\chi_p^2$ , and  $\chi_m^2$  values. The first column shows that  $Z_{rs}$  closely follows a  $N(0, \sigma)$  distribution, where  $\sigma < 1$ , in other words,  $\sigma Z_{rs} \sim N(0, 1)$ , which indicates that the corresponding  $\chi_s^2$  values will follow  $\chi_{(1)}^2$  distribution after multiplied by  $\sigma^2$ , i.e.,  $\sigma^2 \chi_s^2 \sim \chi_{(1)}^2$ . Note that  $\sigma$  is close to 1 when the RAF difference is 0.1, and decreases when RAF difference increases. The second

column verifies that  $Z_{rp}$  follows a  $N(0, 1)$  distribution, as predicted above, which also means that  $\chi_p^2 \sim \chi_{(1)}^2$ .  $Z_{rm}$  values were calculated using meta-analysis based on Stouffer's method, which is expected to follow  $N(0, 1)$  distribution when sample size is sufficiently large. Our test results (last column) verify that when sample size is 10,000 and there are two subpopulations with RAF difference 0.3 or less, this is approximately true (with  $\sigma$  between 0.97 and 0.99). Comparing the three test statistics,  $Z_{rp}$  is the best one since it follows a  $N(0, 1)$  distribution under null and tolerates RAF difference, indicating that  $\chi_p^2$  is a good test statistic for HWE test in the presence of population stratification, and it can replace both  $\chi_s^2$  and  $\chi_m^2$ .

## HWE test on simulated datasets including admixed populations

PopHWE can also be applied to admixed populations. We simulated datasets randomly sampled from a recently admixed population, in which all individuals had two ancestries A and B, with proportions of A ancestry uniformly distributed from 0 to 1. For each of the 100,000 variants, ancestry groups A and B had different RAFs, and for each individual in the admixed population, the RAF was only determined by the ancestry proportion.

We first tested PopHWE using a dataset with sample size 10,000 and simulated as above, without partitioning it into separated subsets. When the dataset is not stratified,  $\chi_e^2$  values can be calculated using the total genotype counts  $N_{aa}$ ,  $N_{ab}$ ,  $N_{bb}$ . However,  $\chi_p^2$  and  $\chi_e^2$  need to be computed slightly differently. In this admixed model, samples are not aggregated to ancestry groups, and hence no allele frequencies in groups can be calculated. Instead, for each sample, an individual-specific allele frequency (IAF) can be estimated from its ancestry proportion, and expected genotype counts can be calculated from IAF. For each sample  $i$ , let the ancestry proportion be  $r_i$ , for a certain locus, the IAF  $p_i$  was estimated as  $p_i = p_b + r_i p_t$ , where  $p_b$  and  $p_t$  were the RAFs in the two ancestry groups for this variant. We calculated the expected genotype counts as  $E_{iaa} = \sum_{i=1}^N 2p_i^2$ ,  $E_{iab} = \sum_{i=1}^N 2p_i(1 - p_i)$  and  $E_{ibb} = \sum_{i=1}^N (1 - p_i)^2$ . Alternatively, these counts can also be estimated using Equations (3), (4), (5). The sum of the expected genotype count was calculated as  $S_{eaa} = \sum_{i=1}^N E_{iaa}$ , and  $\chi_s^2$ ,  $\chi_e^2$ , and  $\chi_p^2$  values were calculated using Equations (7), (8) and (11).

The test results are presented in Fig. 8. Fig. 8A shows the distribution of  $\chi_c^2$  values, exhibiting a long visible tail extending to  $\sim 1,111$ , as expected from Equation (6). The scatterplot in Fig. 8B reveals a tight correlation between  $\chi_c^2$  and  $\chi_e^2$ , a slope of 0.98 and

an  $r^2$  of 0.99 for the linear regression. Fig. 8C, 8E and 8D show that  $\chi_s^2$ ,  $\chi_p^2$  and  $\chi_d^2$  values had much lower standard deviations than  $\chi_c^2$  values, although they did not follow  $\chi_{(1)}^2$  distribution very closely. Fig. 8F plots the histogram of  $Z_p$ , which doesn't follow normal distribution.

We then aggregated the samples in the above dataset into difference numbers of groups such that each group contained only individuals with ancestry proportions within a certain range. These datasets were used to test PopHWE and meta-analysis for stratified HWE tests. Fig. 9 presents the test results when the dataset was partitioned into 2, 4, 6, 8, 10, 20, and 50 subgroups, from the top row to the bottom row, respectively. Histograms of  $Z_p$ ,  $Z_m$ ,  $\chi_p^2$ , and  $\chi_m^2$  values are plotted on columns 1 to 4, and scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$  are plotted on the last column.

When the 10,000 samples were aggregated into only 2 groups, neither  $Z_p$  nor  $Z_m$  values followed the  $N(0, 1)$  distribution, with  $Z_p$  values shifted to the HetDef side and  $Z_m$  to the HetExc side, and  $\sigma$  values much higher than 1 (2.4 and 2.8, respectively). In agreement with the z-values, the corresponding  $\chi_p^2$  and  $\chi_m^2$  values did not follow  $\chi_{(1)}^2$  distribution either. The scatterplot shows that  $\chi_e^2$  values, the expected  $\chi^2$  values resulting from population stratification calculated using this model, were much lower than the observed  $\chi_c^2$  values. Both the mean value shifts and the low slope of the  $\chi_c^2$  vs  $\chi_e^2$  scatterplot indicate the existence of unaccounted population stratification, as described in "Method". When the dataset was partitioned into more groups based on ancestry proportions, the mean values of  $Z_p$  and  $Z_m$  became closer to 0, and the slope on the scatterplots increased to near 1, indicating agreements between expected and observed  $\chi^2$  values. When number of subgroups was 8 or 10, the distributions of  $Z_p$  and  $\chi_p^2$  were very close to  $N(0, 1)$  and  $\chi_{(1)}^2$ , respectively, and the slopes of the scatterplot were near 1 ( $> 0.96$ ), but the  $Z_m$  values were still shifted to the HetExc side. When the dataset was divided into 50 subgroups, although the slope on the that  $\chi_c^2$  vs.  $\chi_e^2$  plot remained close to 1 (Fig 9i), the distributions of  $Z_p$  and  $Z_m$  values shifted away from  $N(0, 1)$  (Figs. 9e and 9f). It seems that both too few and too many subgroups are not good for HWE test of admixed populations. Population stratification of an admixed population can be fully accounted for by partitioned it into 8-10 strata based on ancestry backgrounds and using  $\chi_p^2$  to estimate the null distribution. Comparing the results in Fig. 9 to those in Fig. 8, for admixed populations, when appropriately aggregated, the stratified PopHWE test is possibly better than the test using IAFs.

## Applying PopHWE and meta-analysis to 1000 Genomes Project data

We applied PopHWE and the meta-analysis method to the 1KGP data. We first tested the method using the genotypes of the ~282K GrafAnc Ancestry SNPs extracted from 1KGP data. GrafAnc was used to group 1KGP individuals in the following three ways: 1. With 5 superpopulations including AFR (n=648), EUR (n=511), SAS (n=489), EAS (n=504), and AMR (n=320). 2. Seven continental ancestry groups by splitting the 156 African Americans (*AncGroupID* = 107) from the AFR group, and 212 Latin American 2 (*AncGroupID* = 602) samples from the AMR group. 3. With all the 29 subcontinental ancestry groups assigned by GrafAnc, with sample sizes ranging from 1 to 199. Fig. 10 shows the histograms of  $Z_p$ ,  $Z_s$ ,  $Z_m$ ,  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$  values obtained using the above 3 grouping methods. The top row shows the distributions of the  $Z_p$  values. We can see that  $Z_p$  shifted to the HetDef side when the dataset was partitioned into 5 and 7 superpopulations. When it was assigned to 29 ancestry groups by GrafAnc, there was an almost perfect match of  $Z_p$  and  $N(0, 1)$  distributions (Fig. 10C). These results are also verified by the histograms of  $\chi_p^2$  (the 4<sup>th</sup> row), indicating (1) Ancestry grouping made by GrafAnc correctly reflects the genetic ancestry of 1KGP data. (2) Most of the selected SNPs do not have genotyping errors in the dataset. (3) PopHWE tolerates ancestry grouping at fine scale. Results of  $Z_s$  (2<sup>nd</sup> row) and  $\chi_s^2$  (4<sup>th</sup> row) are very close to those of  $Z_p$  and  $\chi_p^2$ , but the former looks slightly better than the latter. For meta-analysis,  $Z_m$  (3<sup>rd</sup> row) always shifted to the HetExc side, no matter which ancestry grouping method was used. As pointed out above, the undirected  $\chi_m^2$  would be a much better test statistic than  $Z_m$ . However, even if  $\chi_m^2$  is used, meta-analysis is not as good as PopHWE, since  $\chi_p^2$  follows  $\chi_{(1)}^2$  more closely than  $\chi_m^2$ , especially when the dataset is separated into ancestry groups at subcontinental level.

Fig. 11 plots the scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$ ,  $\chi_d^2$ ,  $\chi_p^2$ , and  $\chi_e^2$ , from the top row to the bottom row, respectively. As in Fig. 10, the three different ways to partition the dataset are shown in different columns. The  $\chi_c^2$  vs.  $\chi_e^2$  scatterplots show that there were tight correlations between  $\chi_c^2$  and  $\chi_e^2$  values, with slopes close to 1, confirming that the  $\chi_e^2$  was a roughly unbiased estimation of  $\chi_c^2$  for 1KGP dataset, and the GrafAnc ancestry grouping correctly accounted for the population stratification. However, when the dataset was coarsely partitioned into only 5 continental populations, the slope was rather low (0.88, Fig 11A), indicating that some population stratification was not accounted for, as described above. The slope increased when the dataset was split into more ancestry groups.

The  $\chi_c^2$  vs.  $\chi_e^2$  scatterplots also separate the HWE deviations caused by population stratification from those caused by other reasons, mostly likely by data errors. SNPs with high  $\chi_c^2$  values have high observed HWD. The scatterplots show that these SNPs were separated into two clusters: one cluster with high  $\chi_e^2$  values and the other with low  $\chi_e^2$ , likely reflecting population stratification and data errors, respectively. These scatterplots are potentially useful for genotype data QC. For example, as a method in addition to automatic filtering using  $\chi_p^2$ , after visual inspection, one can draw cutoff lines on  $\chi_c^2$  vs.  $\chi_e^2$  scatterplot to filter out SNPs with potential data errors.

Scatterplots in the first row also indicates that  $\chi_d^2$ , the difference between  $\chi_c^2$  and  $\chi_e^2$  might also be useful in genotype data QC. The second row of Fig. 11 shows the relationship between  $\chi_c^2$  and  $\chi_d^2$ . Based on the graph, a threshold of  $\chi_d^2$  could be set to filter out SNPs with potential data errors. However, the setting of  $\chi_d^2$  cutoff would be useful only if the population stratification had been fully accounted for, i.e., the slope was closed to 1. When the slope is much lower than 1,  $\chi_e^2$  values should be estimated using the regression equation, which might not be easy to obtain since SNPs with potential errors should be excluded before linear regression analysis.

The bottom three rows show the scatterplots of  $\chi_c^2$  vs.  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$ , all of which had much lower average values than  $\chi_d^2$ . Our simulation results in Fig. 3 already showed that the variants with high  $\chi_p^2$ ,  $\chi_s^2$ , or  $\chi_m^2$  values were likely to have genotyping errors, and that these test statistics probably had more discrimination powers than  $\chi_d^2$ . However, through visual inspection comparing the scatterplots in the first row to those in the last three rows, we can see that for the SNPs with high  $\chi_c^2$  values, it seems that  $\chi_c^2$  vs.  $\chi_e^2$  scatterplot can separate HWD caused by population stratification from HWD caused by data errors more clearly than other plots. These results indicated that  $\chi_e^2$  and  $\chi_d^2$  might also be useful for genotype data QC. The scatterplots of  $\chi_c^2$  vs.  $\chi_s^2$  also verify that for SNPs with HetExc (the red dots),  $\chi_s^2$  is always greater than  $\chi_c^2$ , as proved above. On the other hand, almost all the SNPs with HetDef had  $\chi_s^2 < \chi_c^2$ . For  $\chi_p^2$ , it seems also true that  $\chi_p^2 > \chi_c^2$  for HetExc SNPs (Figs. G-I).

Adopting the approach of partitioning the same 1KGP dataset into 29 continental populations, we tested how SNPs with potential errors were filtered out using  $\chi_d^2$ ,  $\chi_p^2$  and  $\chi_m^2$ . We set the thresholds of these test statistics so that exactly 0.1% of the SNPs exceeded the cutoffs and would be filtered out. Fig. 12 shows the test results. From top to bottom, each row presents scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$ ,  $\chi_d^2$ ,  $\chi_p^2$  and  $\chi_m^2$ . From left to right, each column highlights the SNPs filtered out using  $\chi_d^2$ ,  $\chi_p^2$  and  $\chi_m^2$ , colored in red, purple,

and chocolate color, respectively. To filter out 282 (0.1%) SNPs, cutoff values of  $\chi_d^2$ ,  $\chi_p^2$  and  $\chi_m^2$  needed to be set to 116.9, 45.8, and 32.7, respectively. Since we do not know which 1KGP variants have genotyping errors, we do not know which are true positives and which are false. However, histograms in Fig. 10 imply that most of the Ancestry SNPs don't have data errors, and the  $\chi_p^2$  and  $\chi_m^2$  values approximately follow  $\chi_{(1)}^2$  distribution. Cutoff values 45.8 and 32.7 correspond to p-values  $1.3 \times 10^{-11}$  and  $1.1 \times 10^{-8}$ , based on  $\chi_{(1)}^2$  distribution, meaning that we are extremely confident that the variants to be filtered out did have data errors, and the confidence level was much higher using  $\chi_p^2$  than using  $\chi_m^2$ . For  $\chi_m^2$ , when same cutoff (45.8) as for  $\chi_p^2$  was used, only 231 (0.08%) SNPs were filtered out, implying that  $\chi_p^2$  might have lower false negative rates than  $\chi_m^2$  when same p-value threshold was used. In addition, Fig. 10 also showed that  $\chi_p^2$  values had lower standard deviations than  $\chi_m^2$  values, as reflected by the low  $\sigma$  values of  $Z_p$  than those of  $Z_m$  (Fig. 10), which implies that  $\chi_p^2$  probably has lower type I errors than  $\chi_m^2$ , when same p-value threshold is used.

## Testing PopHWE using all variants in Chromosome 22 of 1KGP data

We then applied PopHWE and meta-analysis methods to the 1KGP Chromosome 22 dataset, using all the 1.9M variants. In the Ancestry SNP dataset used above, the variants were preselected by GrafAnc to keep ancestry informative SNPs, and hence very few SNPs have low MAFs. Since many variants with low MAFs are included in the Chromosome 22 dataset, and  $\chi^2$  test requires a minimum expected genotype count, usually 5, a higher MAF threshold needs to be set. The 1KGP dataset has a rather small sample size 2,504. For variants with  $RAF = 0.01$ , the minimum expected genotype count will be 0.25, far fewer than the required value. So, we set MAF threshold to 0.05, such that the minimum expected genotype count will be 6.3, a little larger than the required value.

The dataset was partitioned into 5, 7, and 29 ancestry groups, same as what we did with the 1KGP Ancestry SNP dataset, and histograms and scatterplots (Figs 13-15) similar to Figs 10-12 were plotted to present the results.

Results of HWE test with all variants in chromosome 22 of the 1KGP data are similar to those of the Ancestry SNPs, except that the former had many more variants with potential errors. Comparing Fig. 13 with Fig. 10, we can see that  $Z_p$  and  $Z_s$  values shifted to the HetDef side when the dataset was partitioned into 5 or 7 continental ancestry groups,

while  $Z_m$  shifted to the HetExc side with all ancestry grouping methods. It seems the histograms don't follow  $N(0, 1)$  or  $\chi^2_{(1)}$  distribution (the red curves), but this is expected because many variants might have genotyping errors. Assuming 38% variants had genotyping errors, we plotted curves (purple) of  $N(0, 1)$  or  $\chi^2_{(1)}$  distribution for the rest 62% variants. These curves fit the distributions of  $Z_p$ ,  $Z_s$ ,  $\chi_p^2$ , and  $\chi_s^2$  very well, when the dataset was split into 29 ancestry groups and  $\chi_p^2$  and absolute z-values were small (last column, 1<sup>st</sup>, 2<sup>nd</sup>, 4<sup>th</sup>, and 5<sup>th</sup> columns), which indicates that the estimate of the genotyping error rate (38%) is reliable, and the purple curves can be used to estimate the true positive and false positive rates when variants are filtered using any  $Z_p$  or  $\chi_p^2$  cutoffs.

Fig. 14 plots the scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$ ,  $\chi_d^2$ ,  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$ , showing HetDef variants in blue and HetExc in red. It appears there were no large differences between the scatterplots obtained with different ancestry grouping methods. The 1<sup>st</sup> row shows that variants with high HWD (as reported by  $\chi_c^2$ ) were separated into two clusters: the top one caused by population stratification, and the bottom one by genotyping errors. These scatterplots also show that all HetExc variants (colored red), except a few with very low  $\chi_c^2$  values, had  $\chi_e^2$  values near 0, which is expectable since population stratification would result in HetDef. In other words, for HetExc variants, the effects of population stratification can be ignored, and the dataset can be treated as one without population structure so that the classical  $\chi^2$  test is applicable. For HetDef variants, those with HWD were separated into two clusters on the  $\chi_c^2$  vs.  $\chi_e^2$  scatterplot. We believe that for most of the variants close to the diagonal line, the HWE deviations are caused only by population stratification, and for those far away from the diagonal line, the HWE deviations may be caused by data errors, or both data errors and population stratification. Fig. 14 also show that most of the variants with extremely large  $\chi_c^2$  values were the HetExc ones, and for all HetExc variants,  $\chi_s^2 > \chi_c^2$  (4<sup>th</sup> row) as proved above, and it seems we also have  $\chi_p^2 > \chi_c^2$  (3<sup>th</sup> row).

Fig. 15 presents how  $\chi_d^2$ ,  $\chi_p^2$ , and  $\chi_m^2$  could be used to find variants with potential data errors, both for HetExc and HetDef variants. Out of the 1.87M variants in Chromosome 22, 185K variants had  $MAF > 0.05$ . Setting p-value cutoff to  $1 \times 10^{-5}$ , corresponding to  $\chi^2 = 19.5$  based on  $\chi^2_{(1)}$  distribution, 71K (39%), 56K (30%), and 53K (28%) variants would be filtered out by using  $\chi_d^2$ ,  $\chi_p^2$ , and  $\chi_m^2$ , respectively. On the other hand, in order to filter out 25% (46K) of the SNPs, the cutoff values should be set at 59, 38, and 45, for using  $\chi_d^2$ ,  $\chi_p^2$ , and  $\chi_m^2$ , respectively. Fig. 15 shows the distributions of the 25% variants removed using different metrics. Since we don't know which SNPs really had genotyping errors, it

is difficult to compare the precisions of different statistic metrics in finding data errors. However, in Fig. 15B, we can see that almost all variants filtered out by  $\chi_p^2$  (purple dots) were also in the cluster away from the diagonal, implying that the HWE deviations of these variants might be caused by reasons other than population stratification. Similar results were observed on Fig. 15C, for variants filtered out using  $\chi_m^2$  (chocolate color), except that the removed and kept variants were not separated as well as in Fig. 15B, implying that  $\chi_p^2$  might be more accurate than  $\chi_m^2$  in finding variants with genotyping errors. This was confirmed by the plots on the bottom two rows. Comparing Fig. 15I and Fig. 15L, we can see that some variants identified as without data errors at  $\chi_m^2$  threshold 38 had  $\chi_p^2$  values as high as 200 and above. However, Figs. 15H and 15K show that variants with low  $\chi_p^2$  (colored black) also have low  $\chi_m^2$  values, implying that filtering with  $\chi_p^2$  might be more reliable.

Note that the areas of clusters appearing on the plots do not reflect the size. Actually 75% of the variants were colored in black.

## Discussion

The HWE test has become one of the most widely used methods for detecting genotyping errors in genetic studies, based on the observation that severe departures from HWE expectations often signal such errors. However, applying the HWE test to large genomic datasets has introduced substantial challenges. In modern genomic studies, population structure is present in nearly all genotype datasets, especially large ones. Such structure causes deviations from HWE even in the absence of data errors. When the  $\chi^2$  test is used to assess HWE, its values are directly proportional to sample size if population stratification exists. The p-values calculated using the  $\chi_{(1)}^2$  distribution therefore decrease rapidly and can easily reach the underflow limit when sample sizes are large, even if the degree of population stratification is low. A p-value measures the confidence that the dataset was sampled from a population in HWE, not the degree of HWD (Attia et al 2010, Rothman 2002). Nevertheless, p-values are often used with the intention of filtering out variants with substantial HWD. In addition, a p-value is meaningful only when the model is correct. In fact, in most studies using large genotype datasets, the investigators already know that the datasets contain some degree of population structure. When the classical HWE test is used to identify genotyping errors, an incorrect null hypothesis that the population is in HWE has been assumed, and it should therefore not

be expected that the  $\chi_c^2$  values of variants without genotyping errors follow a  $\chi_{(1)}^2$  distribution. Because the model is wrong, the p-values may be useless or even misleading. Extremely low p-values, such as  $10^{-300}$ , do not mean that it is extremely unlikely that variants have no genotyping errors. These p-values would be much higher if a correct model were used. In this case, the  $\chi_c^2$  values might still be useful, but p-values calculated using the  $\chi_{(1)}^2$  distribution are misleading.

When population structure exists in a dataset, a simple and straightforward way is partitioning it into ancestry groups, hoping each group is roughly genetically homogenous, and hence the null hypothesis should be each group being in HWE. We have shown that expected genotype counts can be easily calculated under this hypothesis, and so can the  $\chi_e^2$  expected due to population stratification. Plotting  $\chi_e^2$  against  $\chi_c^2$ , we can separate HWE deviations caused by population stratification from those resulting from genotyping errors. After visual inspection on the scatterplots, we can draw a cutoff line (or curve) to roughly identify the variants with potential genotyping errors. This is useful as an extra method in addition to automatic variant filtering. Although no p-values can be calculated from the  $\chi_e^2$  and  $\chi_c^2$  values since no distribution models are available, we can still directly use  $\chi_e^2$ ,  $\chi_c^2$ , and/or  $\chi_d^2$  values to filter out variants with potential genotyping errors. Since the variants are clustered on the scatterplots, it is not required to use p-values to accept or reject the null hypothesis. In addition, even if the null hypothesis is false, i.e., population stratification exists in one or more subpopulations, as long as the ancestry grouping captures some degree of population stratification, many variants with potential genotyping errors can still be separated from those without errors. If the population stratification is fully captured by the ancestry grouping, the slope of the  $\chi_e^2$  vs.  $\chi_c^2$  regression on the scatterplot will be 1. Lower slope indicates population stratification exists in one or more subpopulations. With more careful examination on these subpopulations, it is possible to regroup the dataset to capture more population stratification and hence improve the accuracy of genotype data QC using HWE test.

By using the genotype counts and expected counts in all subpopulations, we have extended the classical chi-square goodness-of-fit test for datasets with population stratification. Under  $H_{0P}$ , we created a new test statistic  $\chi_p^2$  to replace the classical  $\chi^2$  metric, denoted as  $\chi_c^2$  in this manuscript.  $\chi_p^2$  is applicable to testing for the  $H_{0P}$  that all subpopulations are in HWE, as well as for the classical  $H_{0C}$  that the whole dataset is in HWE. The  $\chi_p^2$  metric is expected to asymptotically follow a  $\chi_{(1)}^2$  distribution, and hence meaningful p-values can be calculated using  $\chi_p^2$  when population stratification exists.

Due to the Wahlund effect, population stratification results in heterozygote deficiency (HetDef). Therefore, the HWD of HetExc variants are most likely caused by factors other than population stratification, which is consistent with the observation that HetExc variants often reflect genotyping errors (Royo 2020). We have shown that  $\chi_p^2$  was greater than  $\chi_c^2$  for all HetExc variants with all the data we tested, which implies that it might be safe to use the classical HWE test for HetExc variants. For these variants, the presence of population structure does not increase the type I error of the classical HWE test. In addition, even if population structure exists in the datasets, for some variants, e.g., the allele frequencies between different subpopulations may be very low. We have shown above, e.g., Equation (2) and Jensen's inequality, that HWD is caused by the allele frequency differences between subpopulations. Strictly speaking, population stratification, not population structure, is the reason that causes HWE deviation.

In Equations (7) and (8),  $\chi_s^2$  and  $\chi_e^2$  values are calculated with the requirement that the dataset be partitioned into distinct ancestry group. However, for many species, including *homo sapiens*, there are no sharp, discontinuous boundaries between populations. Especially for some recently admixed populations, such as African Americans and Latin Americans, the samples appear as continuous clines, instead of distinct clusters. For these populations, if the individual-level allele frequencies (IAF) can be estimated, using GrafAnc or the method used in RUTH, we can use the same equations to calculate  $\chi_p^2$  and  $\chi_e^2$  values, by treating each sample as a subpopulation with sample size being one (or one unit). Simulation results with an admixed population with RAFs ranging from 0 to 1 (Fig. 8B) show that the samples were distributed near the  $y=x$  line on the  $\chi_c^2$  vs.  $\chi_e^2$  scatterplot, indicating  $\chi_e^2$  is the unbiased estimation of  $\chi_c^2$ . However,  $\chi_p^2$  did not follow  $\chi_{(1)}^2$  distribution, with standard deviations greater than  $\chi_{(1)}^2$ , although it looked much better than  $\chi_c^2$  (Fig. 8A) for HWE-test.

Treating one individual as one subpopulation may introduce too much noise, which may explain the large standard deviations of  $\chi_p^2$  and  $Z_p$  values as observed in Figs. 8E and 8F. Admixed populations lacking distinct clusters can also be partitioned into separate, genetically similar ancestry groups for HWE testing. Our simulation results demonstrate that even in a rather extreme scenario, where each individual carries two ancestries and ancestry proportions are uniformly distributed across the full range from 0 to 1, population stratification can be almost fully accounted for by dividing the dataset into 8 to 10 subpopulations, such that the maximum ancestry proportion difference within each subpopulation is 0.1 to 0.125. Extending these findings to human genotype data, for recently admixed populations such as African Americans, one can first estimate ancestry

proportions using methods like PCA or GrafAnc, then partition the population into groups in which the maximum ancestry proportion difference is about 0.1, and perform a stratified HWE test using PopHWE. When sample size is sufficiently large, one can plot scatterplots and histograms as in Fig. 8 to check if population stratification is captured by the ancestry grouping.

Stratified HWE tests can be performed using classical per-stratum methods, after which the results across strata should be combined to avoid loss of power (Schaid et al., 2006). Several statistical methods are available for combining results, of which Stouffer's method is the most commonly used. Nevertheless, researchers still tend to treat subpopulations separately, probably because meta-analysis methods are not implemented in commonly used software tools. Moreover, although Kwong et al. (2021) describe meta-analysis using Stouffer's method as "an unpublished but commonly used approach," the performance of different HWE meta-analysis methods has not been well evaluated or compared. Other methods, such as the omnibus stratified framework of Troendle and Yu (1994), which calculates  $\chi^2$  values for all strata and combines them by summing the  $\chi^2$  values, are preferred by some investigators (Garg et al., 2026), out of concern that cancellation of z-value, when different groups deviate from HWE in opposite directions, may result in power loss. However, cancellation should not be a problem, since the combined z-value follows a  $N(0,1)$  distribution under the null even after cancellation, which is also confirmed by our simulation results. Summing the squares of z-values avoids cancellation but may not increase power, for two reasons: (1) a one-tailed test cannot be performed, and (2) the degrees of freedom increase. In any case, two large z-values pointing in different directions are more likely to arise from random fluctuation, compared with those in same directions, so allowing them to cancel each other is reasonable.

Stouffer's method can be applied to datasets with small sample sizes by using the HWE exact test to calculate p-values and deriving signed z-values from those p-values. The problem with this method is the case in which identical large z-values differ only in sign, which in many cases are artifacts caused by numerical underflow. Another problem is that the combined z-value shifts toward the HetExc side, which can be avoided as discussed below. However, with the omnibus stratified framework of Troendle and Yu,  $\chi^2$  (or  $Z^2$  as in Garg et al. 2026) need to be calculated for each stratum. These  $\chi^2$  values might not be accurate, since some strata may have much smaller sample sizes than the whole dataset. For example, Garg et al. (2026) divided 1KGP data into 10 strata, of which the smallest had only 170 individuals. For variants with  $MAF = 0.1$ , the minimum expected genotype

count would be 1.7, far below the number required for a valid  $\chi^2$  test. In addition, the combined z-values produced by Stouffer's method are comparable across studies, whereas the combined  $\chi^2$  values obtained with the omnibus stratified framework are not, owing to their differing degrees of freedom.

In this manuscript, we elucidate some usage details of meta-analysis with Stouffer's method and evaluate its performance using simulated and 1KGP data. As described above, the HWE exact test is usually needed for analyzing each stratum. Theoretically, after the z-values from all strata are combined using Stouffer's method, the overall z-value follows a  $N(0, 1)$  distribution under the null hypothesis that each stratum is in HWE. However, this holds only if the sample sizes are infinite or "sufficiently large". We show that for a moderately large dataset, e.g., with thousands of samples spanning roughly a dozen subpopulations, the combined z-values depart significantly from the expected  $N(0, 1)$  null distribution, with the distribution shifted toward HetExc. This shift will inflate the type I error rate of the one-sided test for HetExc (the left tail), while reducing the power of the one-sided test for HetDef (the right tail) and thus increasing its type II error rate. The type I error rate of the two-sided test will also be inflated. We attribute this shift in part to the discreteness of HWE exact test p-values and to the need to assign signs to the resulting z-values, both of which can distort the null distribution. However, our results show that when the signs of the z-values are ignored, these values follow  $N(0, 1)$  distribution on one side very closely when sample sizes are moderately large. Since the square of a random variable following  $N(0, 1)$  follows  $\chi^2_{(1)}$  distribution, we can convert the signed z-values ( $Z_m$ ) to the unsigned  $\chi^2$  values, called  $\chi^2_m$  for meta-analysis in this manuscript. As shown above,  $\chi^2_m$  is a much better test statistic metric than  $Z_m$ , indicating that HWE meta-analysis should be modified to convert the combined z-values to  $\chi^2_m$  values and calculate p-values based on  $\chi^2_{(1)}$  null distribution.

For chi-square goodness-of-fit test with biallelic variants, the  $\chi^2$  values are expected to follow  $\chi^2_{(1)}$  null distribution, and hence  $\chi^2_m$  are comparable to the  $\chi^2$  values obtained from chi-square tests. Similarly, the square roots of  $\chi^2_c$  and  $\chi^2_p$  values obtained from chi-square test are also comparable to the meta-analysis  $Z_m$  values. Randomly assigning signs to the converted z-values will make them easily comparable to  $Z_m$  values on graphs. The extra noises introduced by the random sampling are minimal when there are thousands or more samples.

Our test results with simulated and 1KGP data demonstrate that HWE test using the PopHWE  $\chi^2_p$  performs better than the modified HWE meta-analysis using  $\chi^2_m$ . In

addition,  $\chi_p^2$  is more robust to the small sample sizes and large numbers of subpopulations. Even when datasets are split into hundreds of subpopulations containing only a few samples, HWE test using  $\chi_p^2$  can still distinguish most variants with genotyping errors from those without errors, with discrimination powers much higher than those obtained using the meta-analysis  $\chi_m^2$ . The robustness of PopHWE is probably related to its simple algorithm. For each subpopulation, PopHWE only uses the observed allele counts to calculate the expected genotype counts, and HWD is estimated only after the expected genotype counts in all subpopulations are calculated and combined. No matter how small the sample size of a subgroup is, the allele frequency calculated using the observed allele counts is always the unbiased and best estimation of the allele frequency in the population where the subgroup was sampled from. However, for HWE meta-analysis, no matter whether chi-square test or HWE exact test is used, HWD is estimated for each subgroup, which might introduce some extra random fluctuations.

Another advantage of PopHWE over the HWE meta-analysis is that the former can distinguish HWD caused by genotyping errors from those caused by other reasons, and display them in different clusters on scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$  for visual inspection. Same as the classical chi-square test, the alternative hypothesis of PopHWE test is ambiguous. When  $H_{0P}$  is rejected with a high  $\chi_p^2$  value, one needs to determine if it is caused by genotyping errors or existence of population stratification in some subpopulations. However, in this case one can examine the scatterplot of  $\chi_c^2$  vs.  $\chi_e^2$  and find the correlation between them and calculate the slope, as well as estimate the existence and degree of population stratification that is not accounted for by the ancestry grouping. After that, one can choose to either manually draw a line on the scatterplot to filter variants with potential errors, or improve the ancestry grouping and rerun HWE test. Although the existence of unhandled population stratification can also be found by meta-analysis by checking the distribution of the  $\chi_m^2$  values, PopHWE provides an extra way to resolve this problem. As we see in test results with 1KGP data,  $\chi_e^2$  separates variants with potential genotyping errors from those without errors on the scatterplot. Especially for those with high degrees of HWD, it seems  $\chi_e^2$  has higher discrimination power than  $\chi_p^2$  and  $\chi_m^2$ , which needs to be verified using more simulated and real genotype data.

The algorithms for calculating the PopHWE  $\chi_p^2$  and  $\chi_e^2$  values are very simple, without the numerical underflow problem encountered by HWE exact test. Numerical underflow in HWE meta-analysis not only affects the precisions of the results, but also increase false negative rates in some cases, especially when the dataset includes few large subpopulations with similar sizes. PopHWE has very low computation complexity.

Adding genotype counts to total takes constant time, needing only three additions. Given genotype counts of  $n$  samples and  $k$  subpopulations, the time complexity to calculate  $\chi_p^2$  and  $\chi_e^2$  values is  $O(k)$ . However, calculating the meta-analysis  $Z_m$  or  $\chi_m^2$  values usually takes  $O(kn)$  time. Using the optimized algorithm implemented in PLINK 1.9 and above, the time complexity can be reduced to  $O(k\sqrt{n})$ , still much higher than that of PopHWE. In modern genomic studies, it is not uncommon to have datasets including hundreds of thousands of samples with genotypes of multi-million variants. For these large genotype datasets, it is impractical and unnecessary to enumerate all the genotype configurations and calculate the exact p-values based on a binomial distribution, and PopHWE is a much simpler, faster, and more robust method for HWE test.

In comparison with the IAF-based methods like RUTH, the advantages of PopHWE in terms of time complexity and explainability of the test results are much more prominent. Besides, for modern genomic studies, it is often necessary to partition datasets into ancestry groups and compare allele frequencies between these groups. PopHWE, as well as HWE meta-analysis, provide ways for researchers to further check HWE deviation in each ancestry group, while with IAF-based methods one can only blindly filter out variants using a cutoff p-value. IAF-based methods have an advantage: it avoids the rather arbitrary ancestry grouping. However, to estimate the IAFs, genetic ancestry inference still needs to be done and parameters such as number of PCs needs to be set. It is true that ancestry grouping is not trivial. For human genotype data, one can use GrafAnc (Jin et al., 2026), for ancestry grouping. GrafAnc assigns samples into 38 distinct subcontinental ancestry groups, which are relatively homogeneous and can be directly used for PopHWE, or merged or filtered by the users before HWE test, depending on users' requirements such as minimum sample sizes of these groups. Since PopHWE tolerates ancestry partitioning, subpopulations can be further split using GrafAnc scores like ancestry proportions.

In summary, we have developed PopHWE, a statistical method and software tool that performs the HWE test in the presence of population stratification by extending the classical chi-square goodness-of-fit test. We also point out that HWE meta-analysis using Stouffer's method can be improved by employing the square of the combined z-value as the test statistic and calculating p-values based on the  $\chi_{(1)}^2$  distribution. Compared with meta-analysis, PopHWE is not only more accurate but also much simpler, faster, and more robust. We expect that PopHWE will be widely used for genotype data QC, especially for large genotype datasets or databases.

## Code availability

PopHWE is publicly available at <https://github.com/jimmy-penn/PopHWE>.

GrafAnc is available at: <https://github.com/jimmy-penn/grafanc>

1000 Genomes Project high-coverage sequence data:

[http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/1000G\\_2504\\_high\\_coverage/working/20190425\\_NYGC\\_GATK/](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/)

## Conflicts of interest

The authors declare no competing interests.

## References

- Abramovs, N., Brass, A., & Tassabehji, M. (2020). Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Front Genet*, 11, 210. <https://doi.org/10.3389/fgene.2020.00210>
- Adams, M. W. D., Ewart, K. M., Carlile, N., Rose, H. A., Walker, J. A., Hutton, I., Ho, S. Y. W., & Lo, N. (2026). Pleistocene Sea-Level Fluctuation Shapes Archipelago-Wide Population Structure in the Critically Endangered Lord Howe Island Cockroach *Panesthia lata*. *Ecol Evol*, 16(3), e72760. <https://doi.org/10.1002/ece3.72760>
- Atsawawaranunt, K., Stuart, K. C., Whibley, A., Ewart, K. M., Rollins, L. A., Johnson, R. N., & Santure, A. W. (2026). Missing or Mis-Telling the Story? Trade-Offs for Restriction-Site Associated Compared to Whole Genome Sequencing. *Mol Ecol*, 35(2), e70211. <https://doi.org/10.1111/mec.70211>
- Attia J, Thakkinstian A, McElduff P, Milne E, Dawson S, Scott RJ, Klerk Nd, Armstrong B, Thompson J. Detecting genotyping error using measures of degree of Hardy-Weinberg disequilibrium. *Stat Appl Genet Mol Biol*. 2010;9:Article 5. doi: 10.2202/1544-6115.1463. Epub 2010 Jan 6. PMID: 20196755.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203-209. <https://doi.org/10.1038/s41586-018-0579-z>

- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4, 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Freeberg MA, Fromont LA, D'Altri T, et al. The European Genome-phenome Archive in 2021. *Nucleic Acids Research*. 2022;50(D1):D980-D987. doi:10.1093/nar/gkab1059
- Faure, N., Vilcot, M., Pichot, F., Riutort, J. J., Barroil, A., Holon, F., Tomasi, N., Mouillot, D., Deter, J., & Manel, S. (2025). A Fragile Stronghold: Genomics Reveal Angelshark Population Vulnerability in Corsica, a Key Mediterranean Refuge. *Ecol Evol*, 15(10), e72275. <https://doi.org/10.1002/ece3.72275>
- Fritsche LG, Igl W, Bailey JN, Grassmann F, Sengupta S, et al. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*. 48:134–143.
- Garg E, Romain J, Sun L, Paterson AD (2026) Assessing Hardy-Weinberg equilibrium in T2T-aligned 1000 genomes project. *PLoS Genet* 22(9): e1012280. <https://doi.org/10.1371/journal.pgen.1012280>
- Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., . . . O'Leary, T. J. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*, 70, 214-223. <https://doi.org/10.1016/j.jclinepi.2015.09.016>
- Graffelman, J., & Weir, B. S. (2016). Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. *Heredity (Edinb)*, 116(6), 558-568. <https://doi.org/10.1038/hdy.2016.20>
- Grover, V.K., Cole, D.E.C. and Hamilton, D.C. (2010), Attributing Hardy-Weinberg Disequilibrium to Population Stratification and Genetic Association in Case-Control Studies. *Annals of Human Genetics*, 74: 77-87. <https://doi.org/10.1111/j.1469-1809.2009.00552.x>
- Hao, W., Song, M., & Storey, J. D. (2016). Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5), 713-721. <https://doi.org/10.1093/bioinformatics/btv641>
- Hao, W., & Storey, J. D. (2019). Extending Tests of Hardy-Weinberg Equilibrium to Structured Populations. *Genetics*, 213(3), 759-770. <https://doi.org/10.1534/genetics.119.302370>
- Hemstrom, W., Grummer, J. A., Luikart, G., & Christie, M. R. (2024). Next-generation data filtering in the genomics era. *Nat Rev Genet*, 25(11), 750-767. <https://doi.org/10.1038/s41576-024-00738-6>

- Hernández J. L., Weir B. S. A disequilibrium coefficient approach to Hardy-Weinberg testing. *Biometrics*. 1989 Mar;45(1):53–70.
- Hsu, Y. H. (2026). Hardy-Weinberg Equilibrium Filtering in Population Genomics: Empirical Review and Decision Framework for Improved Practice. *Ecol Evol*, 16(1), e72688. <https://doi.org/10.1002/ece3.72688>
- Jin, Y., Wang, H., Naj, A. C., Wang, L. S., & Lee, W. P. (2026). GrafAnc: Reliable and reproducible inference of continental and regional population structure. *HGG Adv*, 7(1), 100530. <https://doi.org/10.1016/j.xhgg.2025.100530>
- Kwong, A. M., Blackwell, T. W., LeFaive, J., de Andrade, M., Barnard, J., Barnes, K. C., Blangero, J., Boerwinkle, E., Burchard, E. G., Cade, B. E., Chasman, D. I., Chen, H., Conomos, M. P., Cupples, L. A., Ellinor, P. T., Eng, C., Gao, Y., Guo, X., Irvin, M. R., . . . Kang, H. M. (2021). Robust, flexible, and scalable tests for Hardy-Weinberg equilibrium across diverse ancestries. *Genetics*, 218(1). <https://doi.org/10.1093/genetics/iyab044>
- Lindley D. (1998). Statistical inference concerning Hardy-Weinberg equilibrium. In: Bernardo JM BJ, Dawid AP, and Smith AFM, editor. *Bayesian Statistics*. 3 ed. Oxford: Oxford University Press. p 307-26.
- Marees A.T., de Kluiver H., Stringer S., Vorspan F., Curis E., Marie-Claire C., Derks E.M. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res*. 2018; 27:e1608.
- Meisner, J., & Albrechtsen, A. (2019). Testing for Hardy-Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. *Mol Ecol Resour*, 19(5), 1144-1152. <https://doi.org/10.1111/1755-0998.13019>
- Meisner, J., Albrechtsen, A., & Hanghoj, K. (2021). Detecting selection in low-coverage high-throughput sequencing data using principal component analysis. *BMC Bioinformatics*, 22(1), 470. <https://doi.org/10.1186/s12859-021-04375-2>
- Mon, S. L. Y., Lwin, M., Maw, A. A., Htun, L. L., Bawm, S., Nagano, Y., Nagano, A. J., Kawabe, K., Wada, Y., Okamoto, S., & Shimogiri, T. (2025). Exploring genetic diversity and population structure of Myanmar indigenous chickens using double digest restriction site-associated DNA sequencing. *Anim Genet*, 56(4), e70038. <https://doi.org/10.1111/age.70038>
- Neamatzadeh, H., Dastgheib, S. A., Mazaheri, M., Masoudi, A., Shiri, A., Omid, A., Rahmani, A., Golshan-Tafti, A., Aghasipour, M., Yeganegi, M., Bahrami, M., Aghili, K., Khajehnoori, S., & Mosavi Jarrahi, A. (2024). Hardy-Weinberg Equilibrium in Meta-Analysis Studies and Large-Scale Genomic Sequencing Era. *Asian Pac J Cancer Prev*, 25(7), 2229-2235. <https://doi.org/10.31557/APJCP.2024.25.7.2229>
- Novembre, J., & Peter, B. M. (2016). Recent advances in the study of fine-scale population structure in humans. *Curr Opin Genet Dev*, 41, 98-105. <https://doi.org/10.1016/j.gde.2016.08.007>

- Pearman, W. S., Urban, L., & Alexander, A. (2022). Commonly used Hardy-Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Mol Ecol Resour*, 22(7), 2599-2613. <https://doi.org/10.1111/1755-0998.13646>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3), 559-575. <https://doi.org/10.1086/519795>
- Ramirez, A. H., Sulieman, L., Schlueter, D. J., Halvorson, A., Qian, J., Ratsimbazafy, F., Loperena, R., Mayo, K., Basford, M., Deflaux, N., Muthuraman, K. N., Natarajan, K., Kho, A., Xu, H., Wilkins, C., Anton-Culver, H., Boerwinkle, E., Cicek, M., Clark, C. R., . . . All of Us Research, P. (2022). The All of Us Research Program: Data quality, utility, and diversity. *Patterns (N Y)*, 3(8), 100570. <https://doi.org/10.1016/j.patter.2022.100570>
- Rothman K. (2002). Random error and the role of statistics. Epidemiology: An introduction. Oxford: Oxford University press.
- Royo, J. L. (2020). Hardy Weinberg Equilibrium Disturbances in Case-Control Studies Lead to Non-Conclusive Results. *Cell J*, 22(4), 572-574. <https://doi.org/10.22074/cellj.2021.7195>
- Schaid, D. J., Batzler, A. J., Jenkins, G. D., & Hildebrandt, M. A. (2006). Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. *Am J Hum Genet*, 79(6), 1071-1080. <https://doi.org/10.1086/510257>
- Sha, Q., & Zhang, S. (2011). A test of Hardy-Weinberg equilibrium in structured populations. *Genet Epidemiol*, 35(7), 671-678. <https://doi.org/10.1002/gepi.20617>
- Shoemaker J, Painter I, Weir BS. A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics*. 1998 Aug;149(4):2079-88. doi: 10.1093/genetics/149.4.2079
- Shyr, D., Dey, R., Li, X., Zhou, H., Boerwinkle, E., Buyske, S., Daly, M., Gibbs, R. A., Hall, I., Matise, T., Reeves, C., Stitzel, N. O., Zody, M., Neale, B. M., & Lin, X. (2024). Semi-supervised machine learning method for predicting homogeneous ancestry groups to assess Hardy-Weinberg equilibrium in diverse whole-genome sequencing studies. *Am J Hum Genet*, 111(10), 2129-2138. <https://doi.org/10.1016/j.ajhg.2024.08.018>
- Stouffer SA. 1949. The American Soldier. Princeton: Princeton University Press.
- Troendle JF, Yu KF. A note on testing the Hardy-Weinberg law across strata. *Ann Hum Genet*. 1994 Oct;58(4):397-402. doi: 10.1111/j.1469-1809.1994.tb00735.x. PMID: 7864594
- Wahlund, S. (1928). Composition of population and correlation phenomenon from the viewpoint of heredity. *Hereditas*, 11, 65-106.
- Weir BS . (1996) Genetic Data Analysis II. Sinauer Associates: Massachusetts.
- Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*, 76(5), 887-893. <https://doi.org/10.1086/429864>

Zeng, W., Ye, Z., Liu, X., Lin, H., & Wu, J. (2026). Genotyping-by-Sequencing Reveals Low Genetic Diversity and Pronounced Geographic Structuring in the Endangered Medicinal Plant *Coptis chinensis* var. *brevisepala*. *Plants (Basel)*, 15(3). <https://doi.org/10.3390/plants15030371>

Zhou, W., Qin, H., Wang, H., Zhi, H., Huang, R., Hou, S., Wang, J., Chen, L., Tian, X., Qiao, Z., Mu, Z., & Diao, X. (2026). Genome-wide association analysis and candidate gene identification for plant height in Shanxi local foxtail millet varieties. *Theor Appl Genet*, 139(7). <https://doi.org/10.1007/s00122-026-05303-9>

Tables and Figures

Table 1. Calculation of expected z-value when three individuals are randomly drawn from a population in HWE with frequency of allele a being 0.1

| $N_a$  | $P(N_a)$ | $N_{aa}$ | $N_{ab}$ | $N_{bb}$ | $P(a)$ | $E_{ab}$ | $P(g N_a)$ | $Z_g$   | $W(z)$        |
|--------|----------|----------|----------|----------|--------|----------|------------|---------|---------------|
| 0      | 0.5314   | 0        | 0        | 3        | 0.000  | 0.000    | 1.0        | 0.0000  | 0.0000        |
| 1      | 0.3543   | 0        | 1        | 2        | 0.167  | 0.833    | 1.0        | 0.0000  | 0.0000        |
| 2      | 0.0984   | 1        | 0        | 2        | 0.333  | 1.333    | 0.2        | 1.2816  | 0.0252        |
| 2      | 0.0984   | 0        | 2        | 1        | 0.333  | 1.333    | 1.0        | 0.0000  | 0.0000        |
| 3      | 0.0146   | 1        | 1        | 1        | 0.500  | 1.500    | 1.0        | 0.0000  | 0.0000        |
| 3      | 0.0146   | 0        | 3        | 0        | 0.500  | 1.500    | 0.4        | -0.8416 | -0.0049       |
| 4      | 0.0012   | 2        | 0        | 1        | 0.667  | 1.333    | 0.2        | 1.2816  | 0.0031        |
| 4      | 0.0012   | 1        | 2        | 0        | 0.667  | 1.333    | 1.0        | 0.0000  | 0.0000        |
| 5      | 0.0001   | 2        | 1        | 0        | 0.833  | 0.833    | 1.0        | 0.0000  | 0.0000        |
| 6      | 0.0000   | 3        | 0        | 0        | 1.000  | 0.000    | 1.0        | 0.0000  | 0.0000        |
| $E(z)$ |          |          |          |          |        |          |            |         | <b>0.0206</b> |

$N_a$ : number of *a* allele;  $P(N_a)$ : probability to have  $N_a$  *a* allele;  $N_{aa}$ ,  $N_{ab}$ ,  $N_{bb}$ : counts of the three genotypes;  $P(a)$ : Frequency of *a* allele in the observed genotypes;  $E_{ab}$ : expected ab genotypes calculated using observed genotype counts assuming HWE;  $P(g|N_a)$ : probability of observing these genotype counts given  $N_a$ ;  $Z_g$ : z-value calculated for this genotype count configuration;  $W(z)$ : z-value weighted by  $P(N_a)$ .

Table 2. Expected z-values of HWE exact test given different sample sizes and minor allele frequencies.

| #Individuals | Minor allele frequency |              |              |              |              |
|--------------|------------------------|--------------|--------------|--------------|--------------|
|              | 0.1                    | 0.2          | 0.3          | 0.4          | 0.5          |
| 2            | 0.016                  | 0.050        | 0.085        | 0.111        | 0.121        |
| 3            | 0.021                  | 0.039        | 0.036        | 0.022        | 0.015        |
| 4            | 0.042                  | 0.111        | 0.155        | <b>0.170</b> | <b>0.172</b> |
| 5            | 0.051                  | 0.100        | 0.084        | 0.036        | 0.012        |
| 10           | 0.113                  | <b>0.202</b> | <b>0.160</b> | 0.117        | 0.108        |
| 20           | 0.194                  | 0.170        | 0.066        | 0.054        | 0.071        |
| 50           | <b>0.298</b>           | 0.120        | 0.048        | 0.023        | 0.044        |
| 100          | 0.269                  | 0.073        | 0.024        | 0.005        | 0.031        |
| 200          | 0.178                  | 0.050        | 0.018        | 0.008        | 0.022        |
| 500          | 0.105                  | 0.032        | 0.011        | 0.004        | 0.012        |
| 1000         | 0.076                  | 0.019        | 0.007        | -0.003       | 0.010        |
| 2000         | 0.050                  | 0.021        | 0.016        | 0.007        | -0.009       |
| 5000         | 0.004                  | 0.006        | -0.001       | 0.004        | 0.003        |
| 10000        | 0.003                  | 0.020        | 0.003        | -0.013       | 0.004        |

Note: Expected z-values for number of individuals > 500 were obtained through simulations.

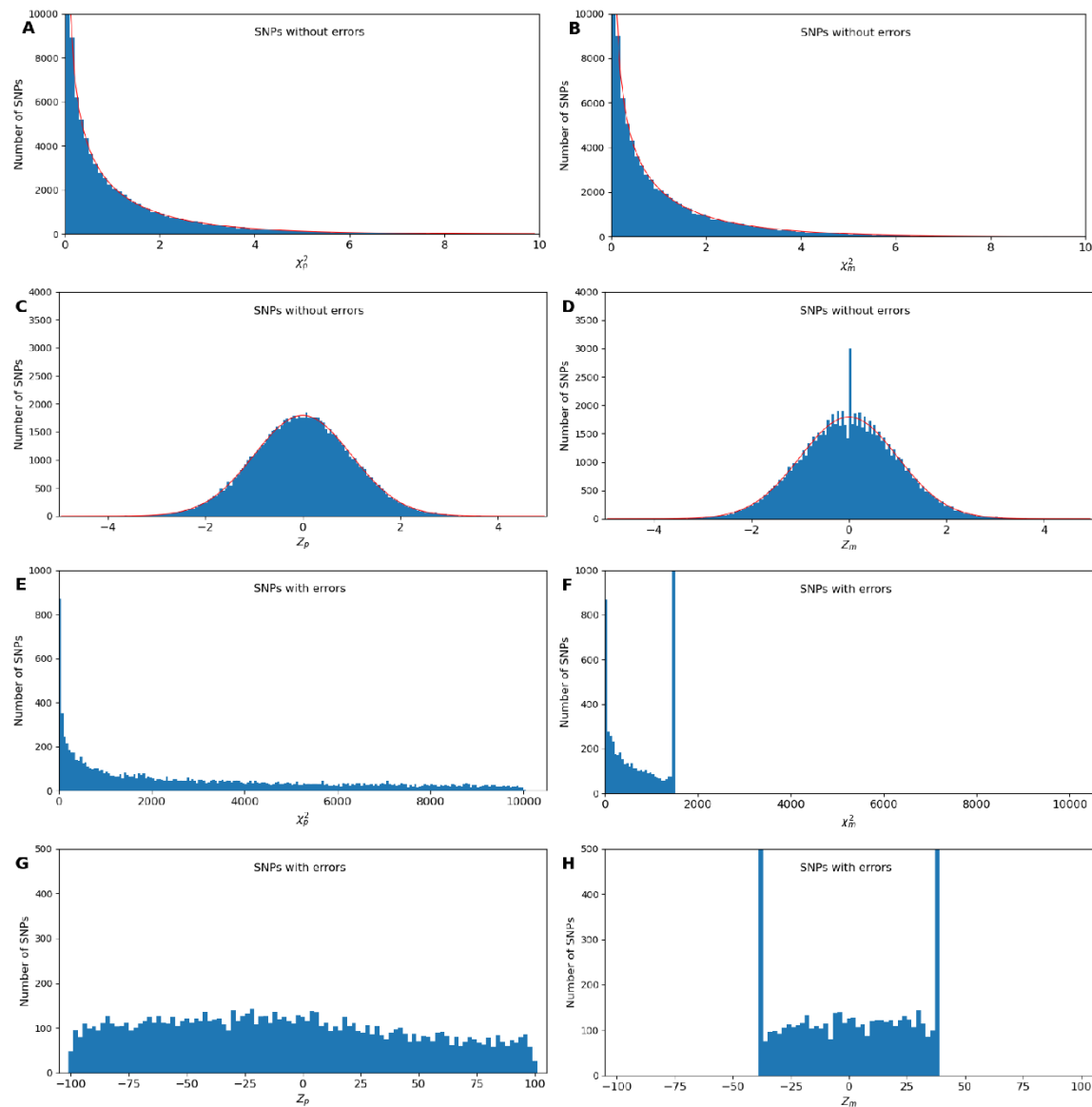


Fig. 1. Test results of PopHWE (left column) and HWE-exact test (right column) with simulated homogeneous dataset. 10,000 individuals were randomly sampled from an infinitely large population in HWE. Total 100K biallelic loci were checked, of which 90K loci don't have data errors (top two rows), with allele frequencies in the population uniformly distributed between 0.1 and 0.9. 10K loci were simulated with errors (bottom two rows), whose genotype counts were randomly generated numbers under the only constraint that they summed up to 10,000. Since there are no subpopulations,  $\chi^2_p$  and  $Z_p$  values are same as the classical  $\chi^2_c$  and  $Z_c$  values, and  $\chi^2_m$  and  $Z_m$  are obtained using HWE-exact test without meta-analysis.

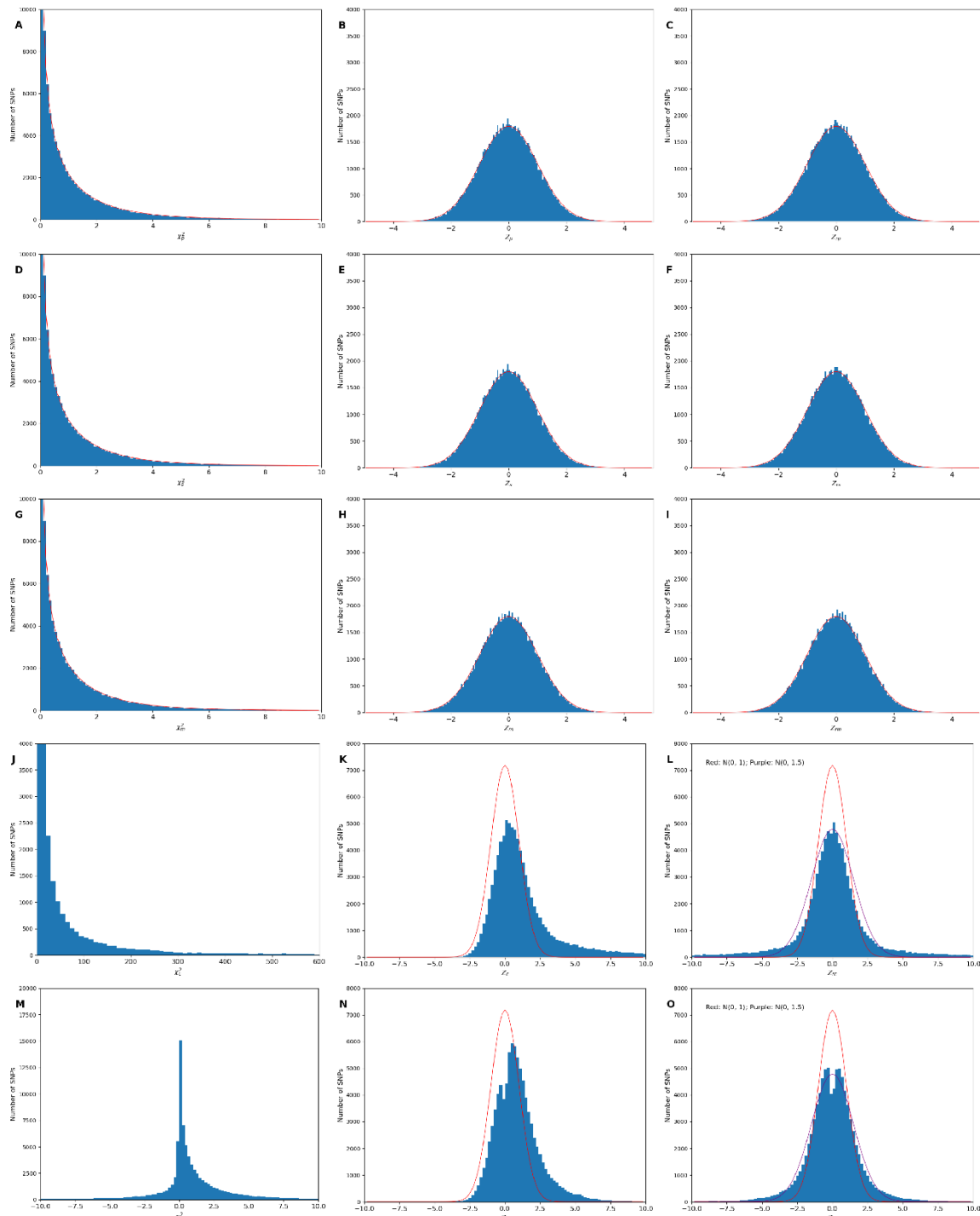


**Fig.2.** Histograms of PopHWE and meta-analysis test statistic values obtained with simulated dataset including 5,000 samples, 2 subpopulations and 100K SNPs. Only 90K SNPs without genotyping errors are displayed. The first column plots chi-square values  $\chi_p^2$ ,  $\chi_s^2$ ,  $\chi_m^2$ ,  $\chi_c^2$ , and  $\chi_d^2$ , from top to bottom. The second column shows the corresponding z-value, with HetDef SNPs assigned negative values. The last column shows the z-values

with randomly assigned signs. The red lines are the distribution curves of  $\chi^2_{(1)}$ ,  $N(0, 1)$ , and the purple lines are  $N(0, 1.5)$  curves (the bottom two rows).

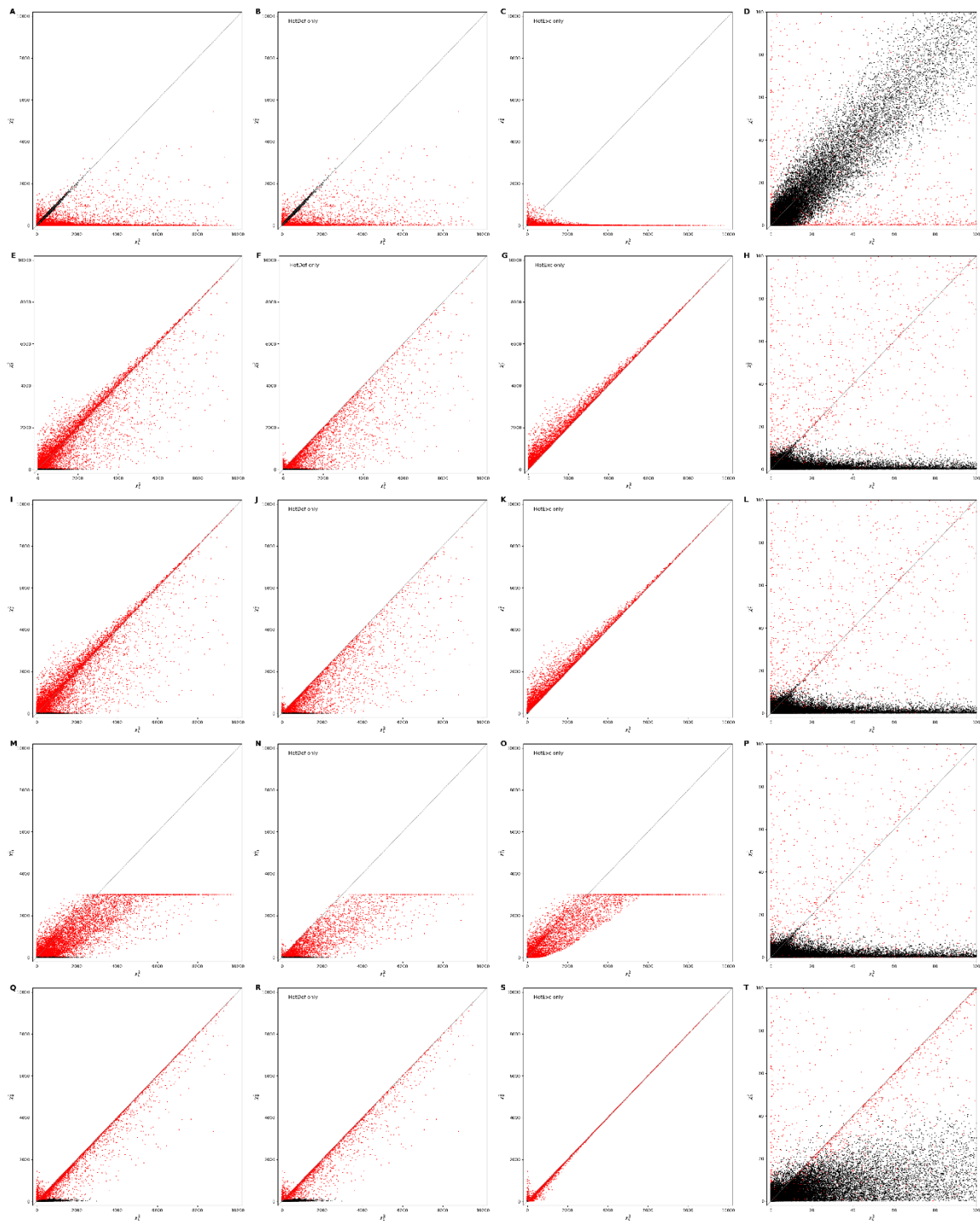


Fig.3. Scatterplots showing distributions of different the  $\chi^2$  values of PopHWE and meta-analysis against the classical  $\chi^2$  values. The x-axis shows  $\chi_c^2$  values, and the y-axis in each row from top to bottom shows  $\chi_e^2$ ,  $\chi_p^2$ ,  $\chi_s^2$ ,  $\chi_m^2$ , and absolute  $\chi_d^2$  values, respectively. Variants without genotype errors are colored black, and those with errors are red. The second column only shows HetDef variants, and the third column has HetExc ones only. The plots in the last column show the  $\chi^2$  values between 0 and 100.

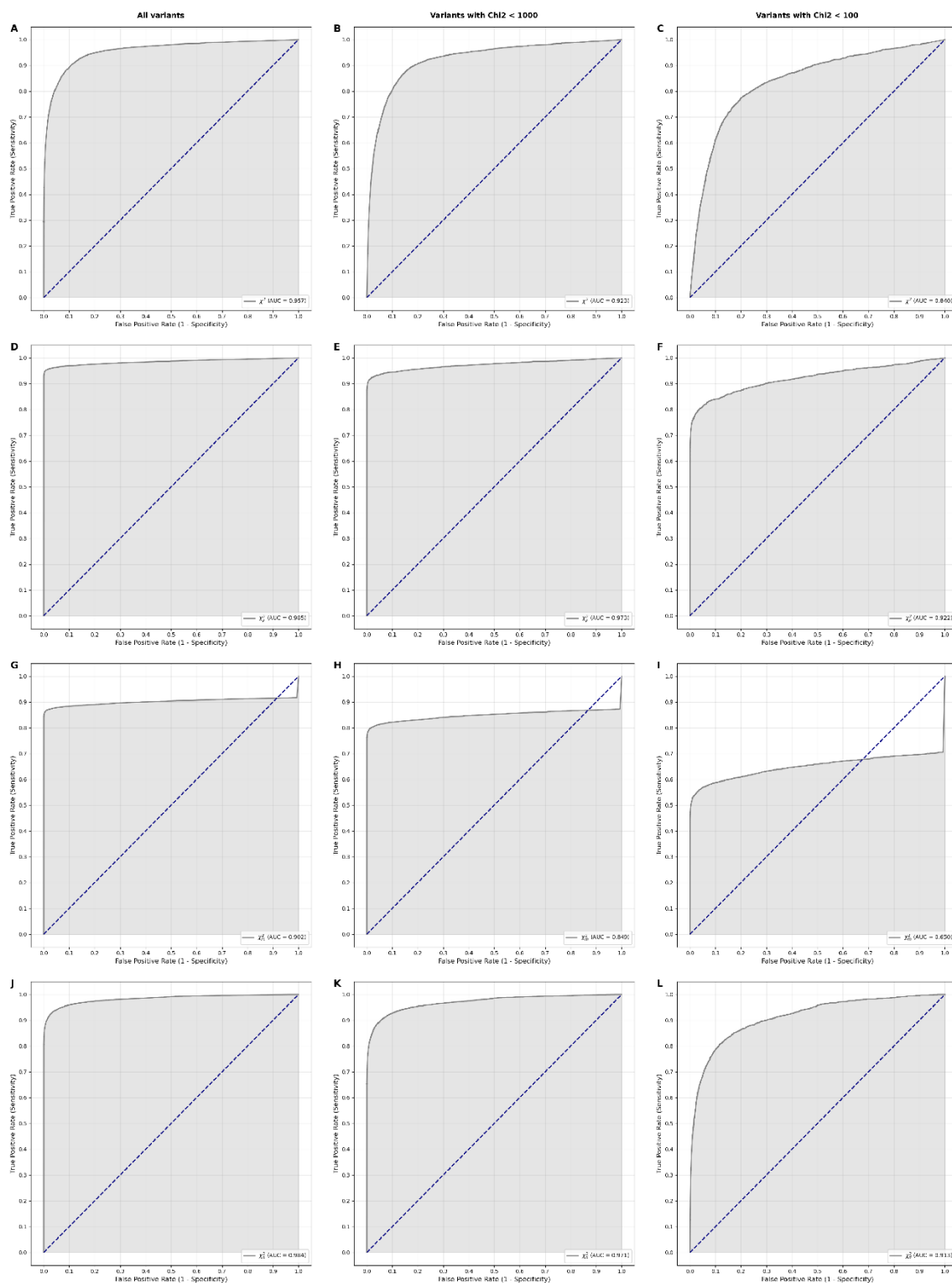


**Fig. 4. Receiver Operating Characteristic (ROC) curves comparing the discrimination powers of PopHWE  $\chi^2$  scores and meta-analysis  $\chi_m^2$  scores. HWE methods were tested with the same dataset as used in Figs. 2 and 3, i.e., 10K samples in two same-size**

subpopulations. ROC curves were plotted for the following test statistics:  $\chi_e^2$ ,  $\chi_p^2$ ,  $\chi_m^2$ , and absolute  $\chi_d^2$ , from top to the bottom row. An AUC (Area Under the Curve) value was calculated for each ROC curve and used to represent the discrimination power. Since cases with data errors were randomly generated, and most of them were very well separated from those without errors for all statistical metrics, we also plotted ROC curves after removing some easy to detect errors. Panels in the middle column plot ROC curves after excluding the variants with statistic scores of the test statistic greater than 1000, and in the last column those with scores greater than 100 were excluded.

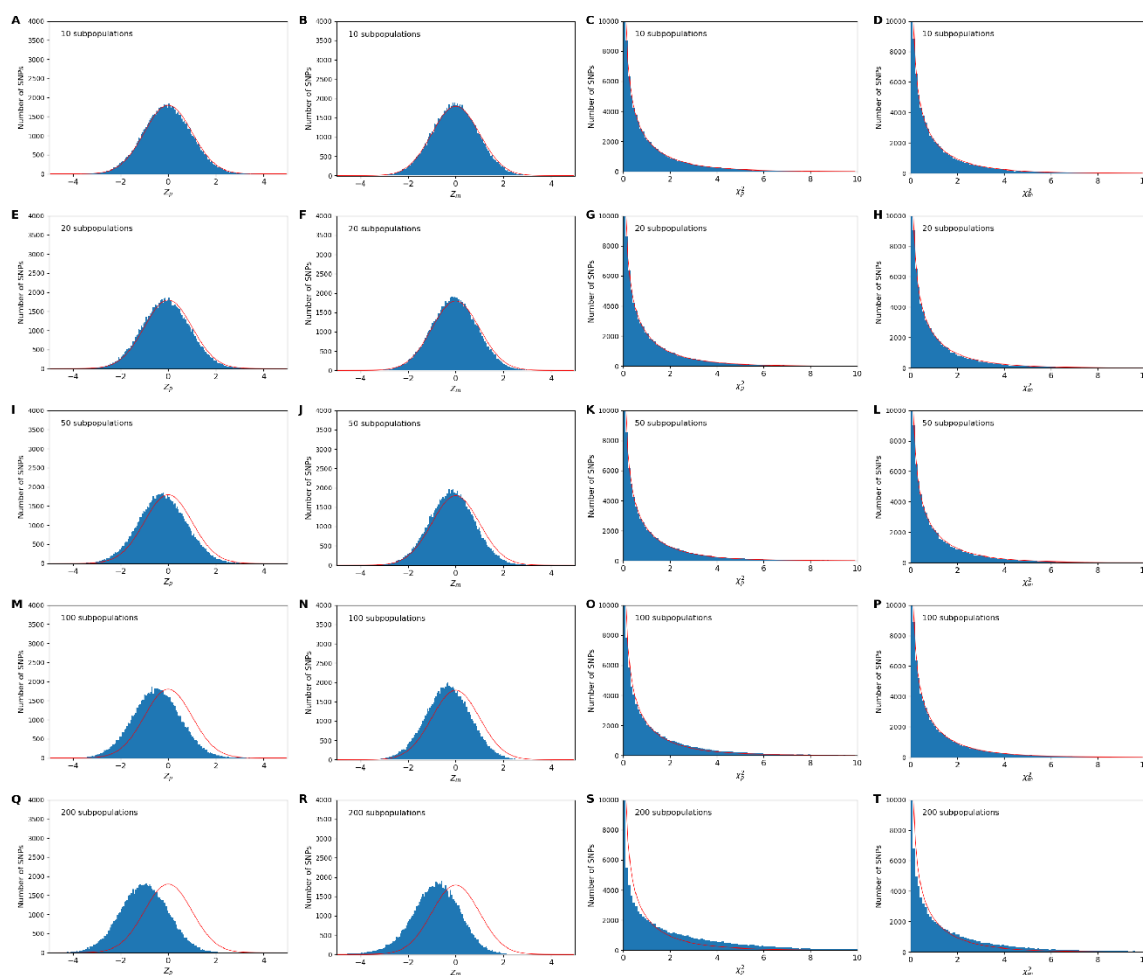


Fig. 5. Comparison of test statistic metrics of PopHWE and meta-analysis when the datasets include different numbers of subpopulations. Datasets were simulated to have same sample size 10,000 but different numbers of same-size subpopulations, each of which was randomly sampled from a population in HWE, same as those in Figs 2, 3, 4. From top to bottom, numbers of subpopulations in each row were 10, 20, 50, 100, 200, and the corresponding subpopulation sample sizes were 1000, 500, 200, 100, 50. From

left to right, the variables are  $Z_p$ ,  $Z_m$ ,  $\chi_p^2$  and  $\chi_m^2$ . Red lines are curves of  $N(0, 1)$  and  $\chi_{(1)}^2$  distributions.

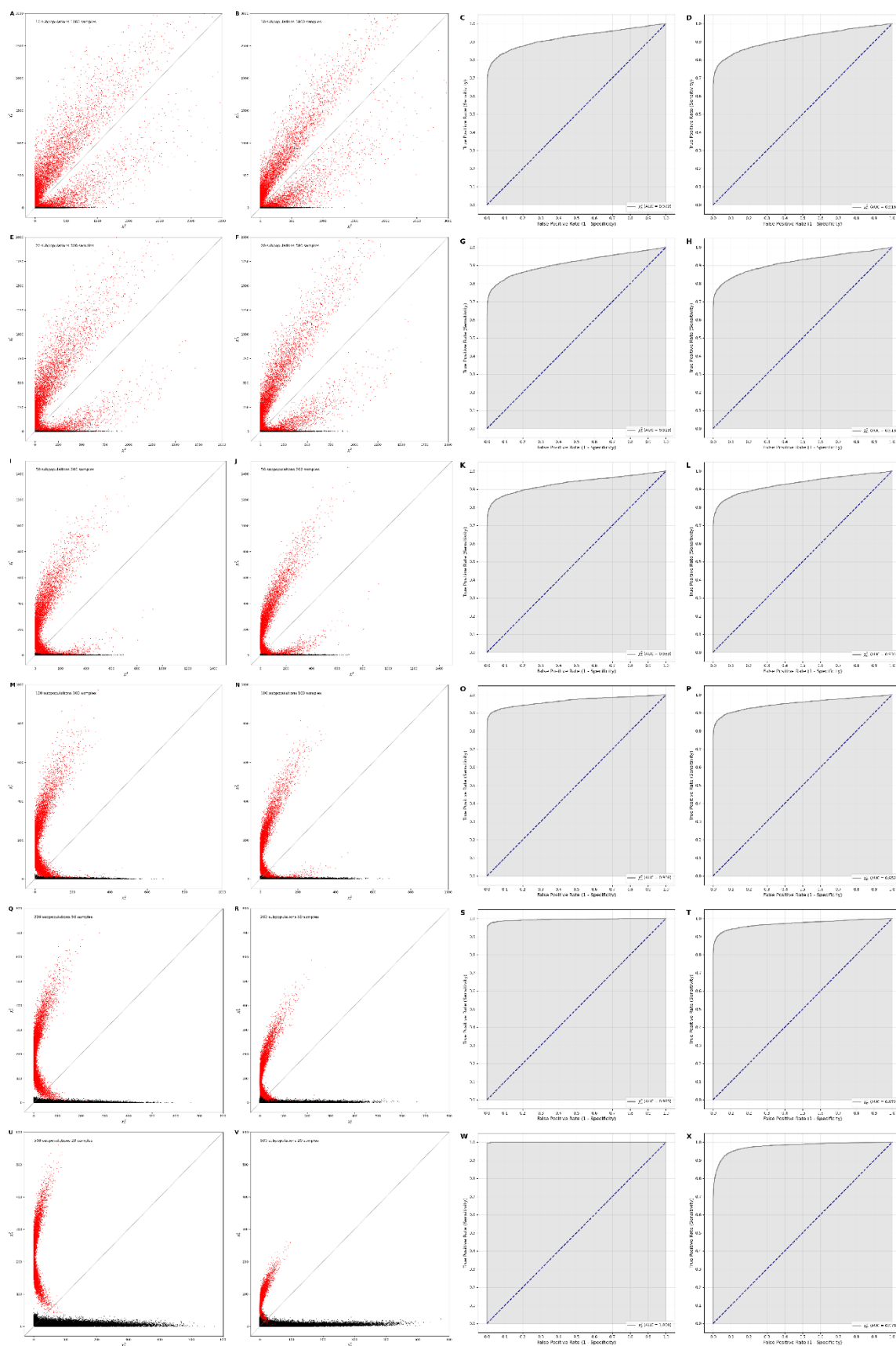


Fig. 6. Change of discrimination powers of PopHWE and meta-analysis test statistics when the datasets include different numbers of subpopulations. Same datasets as in Fig. 5 were used for this figure, except that an extra dataset with 500 subpopulations (each included 20 samples) was used for the panels in the bottom row. From left to the right, each column shows scatterplot of  $\chi_c^2$  vs.  $\chi_p^2$ , ROC curve of  $\chi_p^2$ ,  $\chi_c^2$  vs.  $\chi_m^2$ , ROC curve of  $\chi_p^2$ . The AUC values were calculated only considering the variants with test statistic ( $\chi_p^2$  or  $\chi_m^2$ ) value smaller than 100.

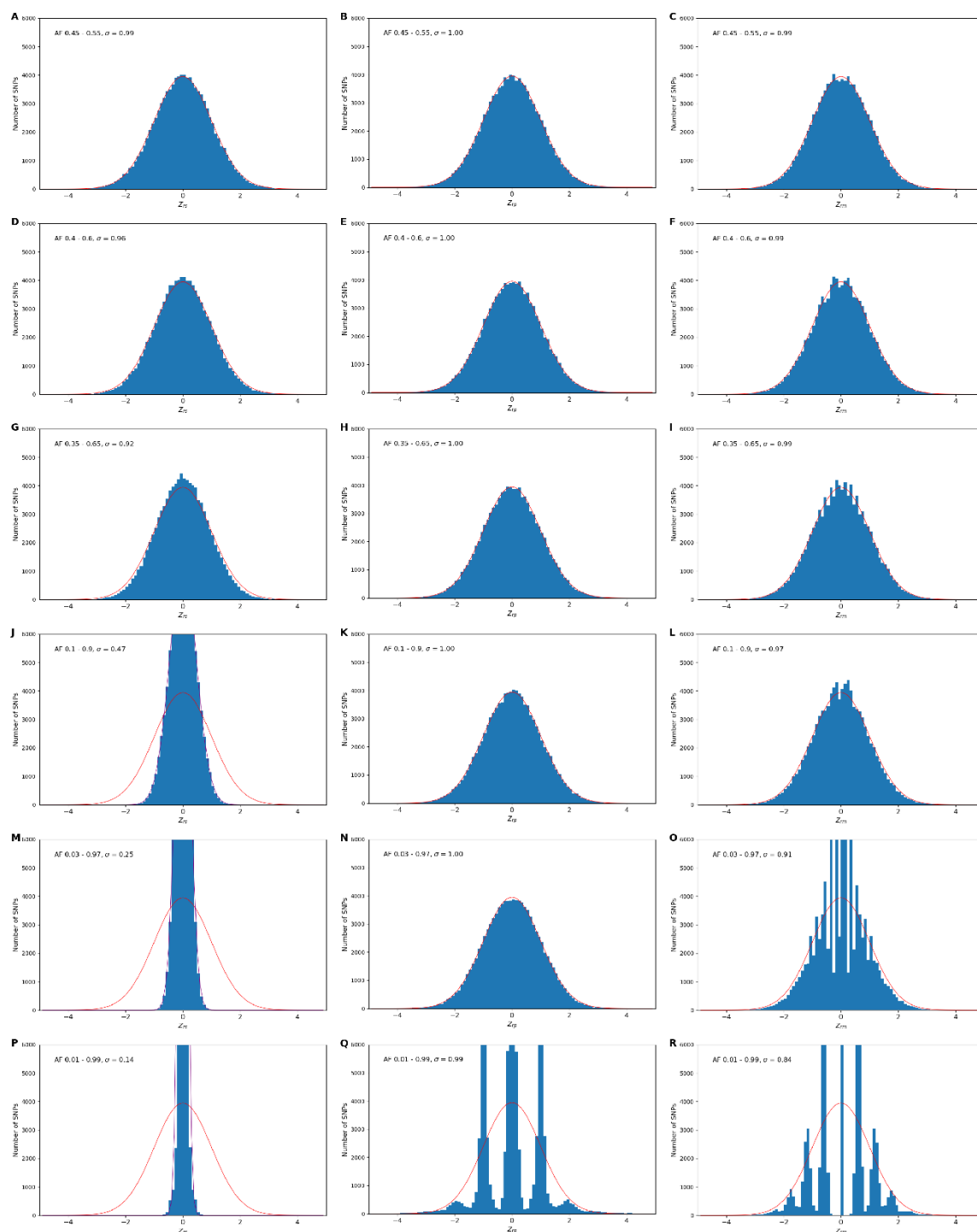


Fig. 7. Distributions of  $Z_{rs}$ ,  $Z_{rp}$ , and  $Z_{rm}$  values for variants with fixed allele frequencies in subpopulations. Test datasets were simulated to have sample size 10,000 and include 2 same-size subgroups, each of which was randomly sampled from a large population in HWE and with a fixed allele frequency for all the variants. Datasets sampled from populations with different allele frequencies were tested using PopHWE and meta-

analysis. Each row shows results of one dataset, and the variant RAFs of its two populations are displayed on each panel. Note that from top to the bottom, the differences of the population RAFs were 0.1, 0.2, 0.3, 0.8, 0.04, and 0.02. The three columns from left to right plot histograms of  $Z_{rs}$ ,  $Z_{rp}$ , and  $Z_{rm}$ , which are the  $Z_s$ ,  $Z_p$ , and  $Z_m$  values with randomly assigned signs, reflecting the corresponding  $\chi_s^2$ ,  $\chi_p^2$ , and  $\chi_m^2$  values. The red curves show  $N(0, 1)$  distribution, and the purple ones in Panels J, M, P are  $N(0, \sigma)$  curves, where  $\sigma$  is displayed on each panel.

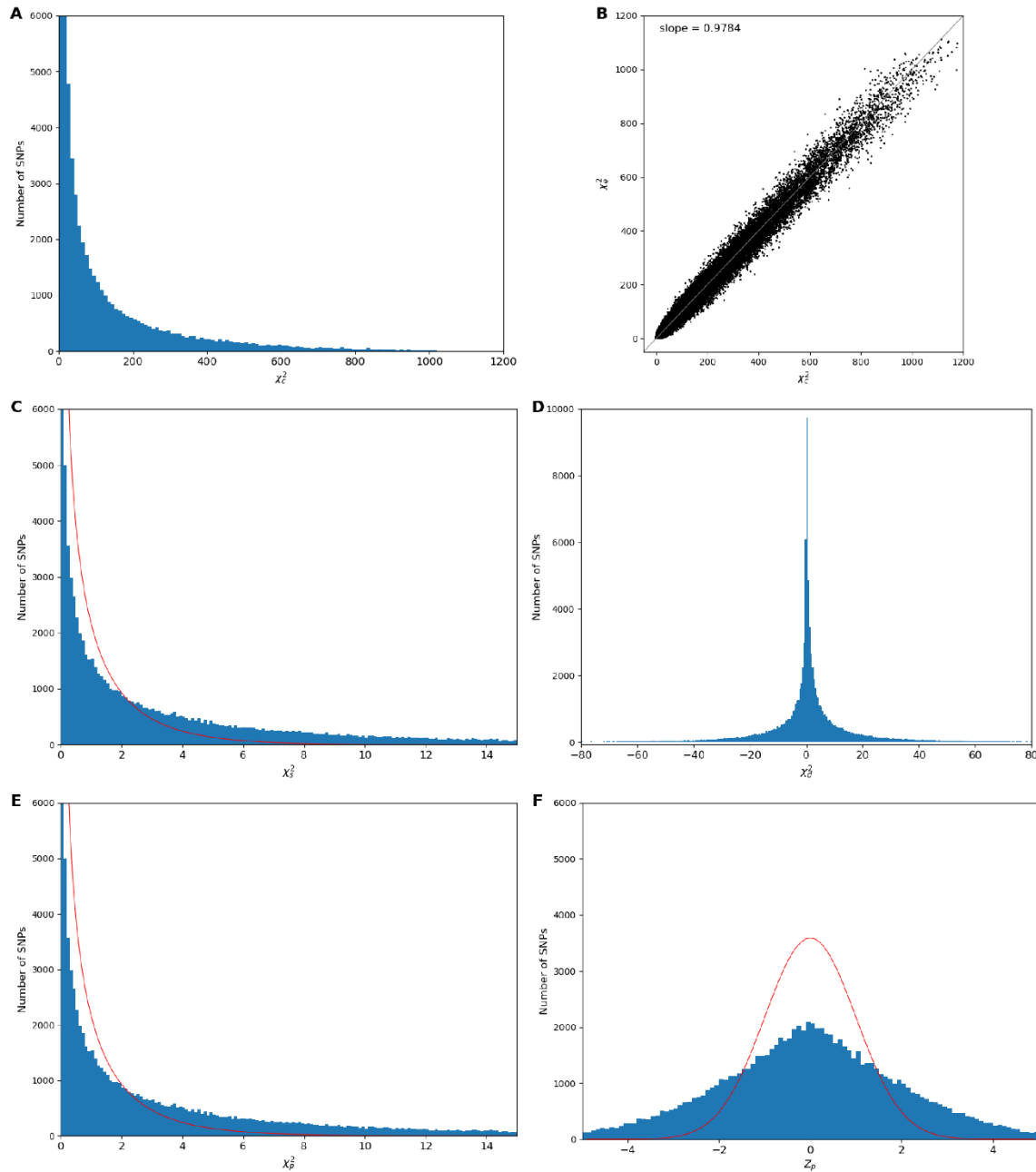


**Fig. 8.** PopHWE results obtained using a simulated dataset including an admixed population. A dataset with 10,000 samples were simulated to be randomly sampled from an admixed population, in which each individual was admixed from two ancestries. The ancestry proportions of the individuals were evenly distributed between 0 and 1. On each of the 100K variants, the two ancestry groups had two different AFs. The individual allele frequency of a sample was determined by its ancestry proportion, and the AFs in its two ancestries. The curves in red are  $\chi^2_{(1)}$  and  $N(0, 1)$  distributions.

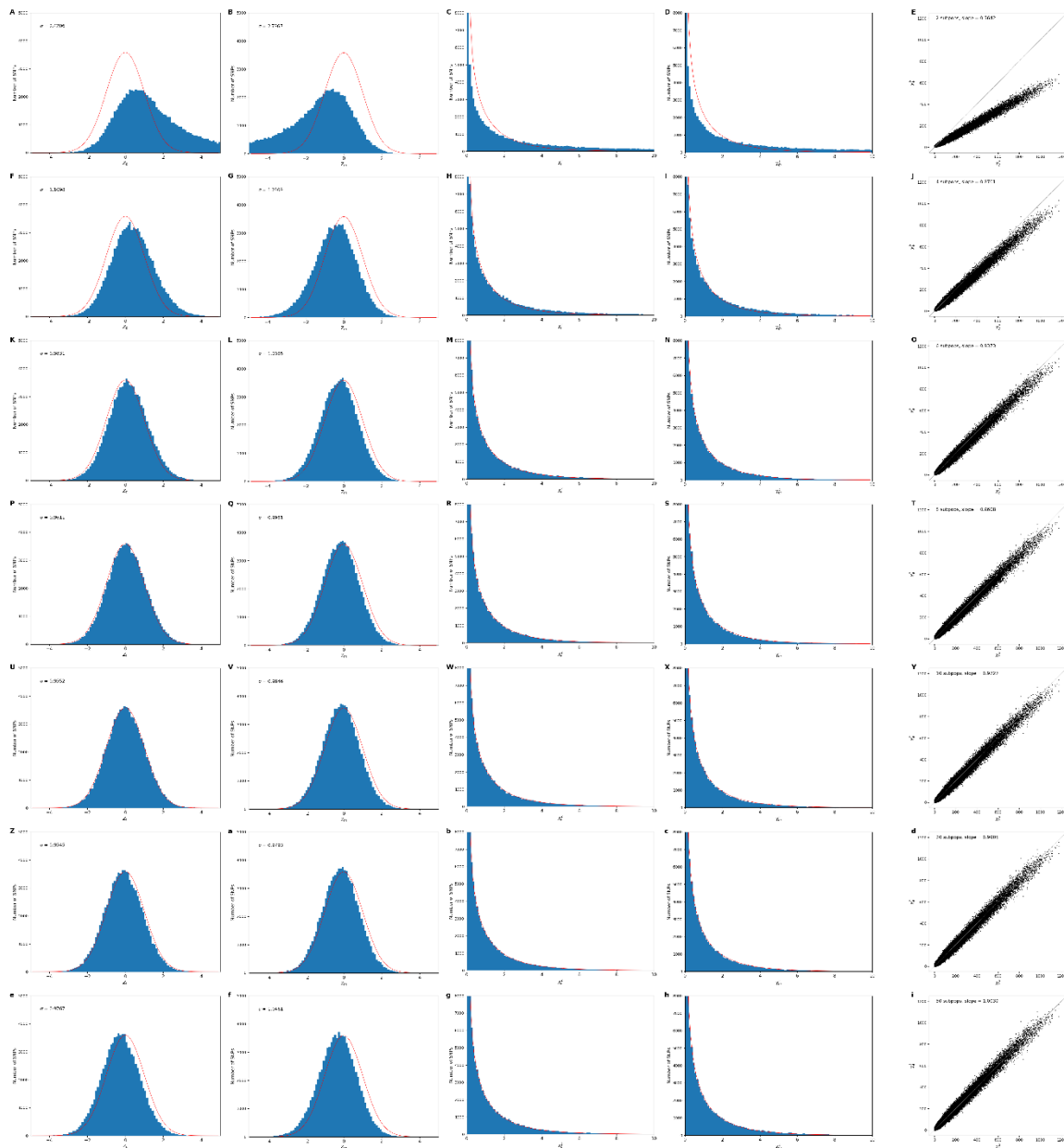


Fig. 9. Results of PopHWE and meta-analysis when the simulated dataset used in Fig. 8 was stratified into 2, 4, 6, 8, 10, 20, 50 subpopulations, from top to bottom row. The left 4 columns are histograms of  $Z_p$ ,  $Z_m$ ,  $\chi_p^2$ , and  $\chi_m^2$ , respectively. The last column shows the scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$ . The slope of linear regression equation is displayed on each panel. The red curves are the  $N(0, 1)$  and  $\chi^2(1)$  distributions.

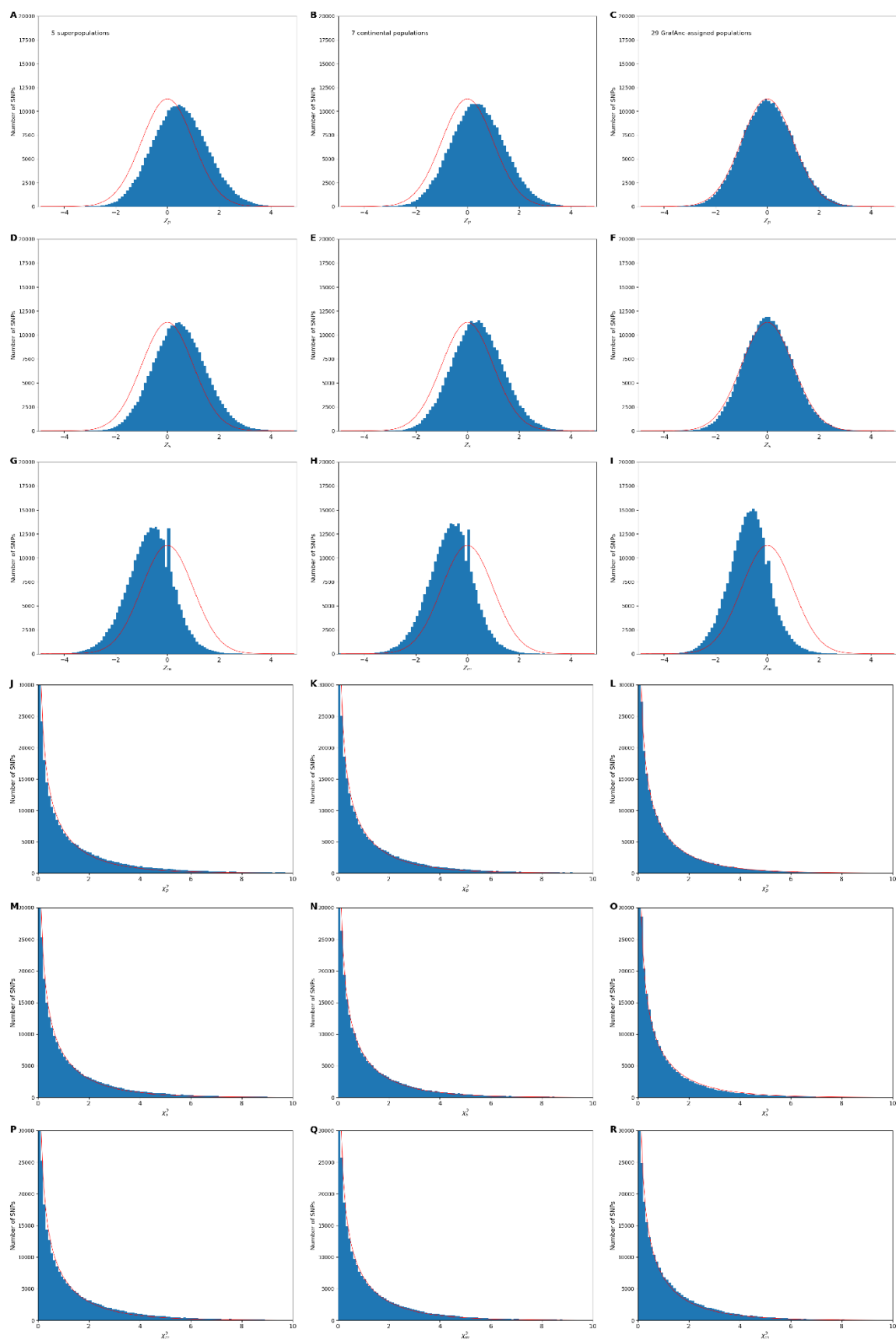


Fig. 10. Histograms of PopHWE and meta-analysis test statistics obtained using 1KGP dataset with GrafAnc ancestry SNPs. Genotypes of ~282K SNPs used by GrafAnc for ancestry inference were extracted from 1KGP WGS dataset. 2,504 individuals in the dataset were aggregated based on the ancestry groups assigned by GrafAnc. Three different ways were used to group the samples: 1. With 5 superpopulations including AFR (n=648), EUR (n=511), SAS (n=489), EAS (n=504), and AMR (n=320). 2. Seven continental ancestry groups by splitting the 156 African Americans (AncGroupID = 107) from the AFR group, and 212 Latin American 2 (AncGroupID = 602) samples from the AMR group. 3. Using all the 29 subcontinental ancestry groups assigned by GrafAnc, with sample sizes ranging from 1 to 199. PopHWE and meta-analysis were tested using these datasets and the results were plotted in histograms. From the top to the bottom row,  $Z_p$ ,  $Z_s$ ,  $Z_m$ ,  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$  values were plotted, and each column from the left to the right presents the results of the three grouping methods. The  $N(0, 1)$  and  $\chi_{(1)}^2$  distribution curves were plotted in red color. Variants with the minor allele frequencies (in the whole dataset) less than 0.01 were excluded.

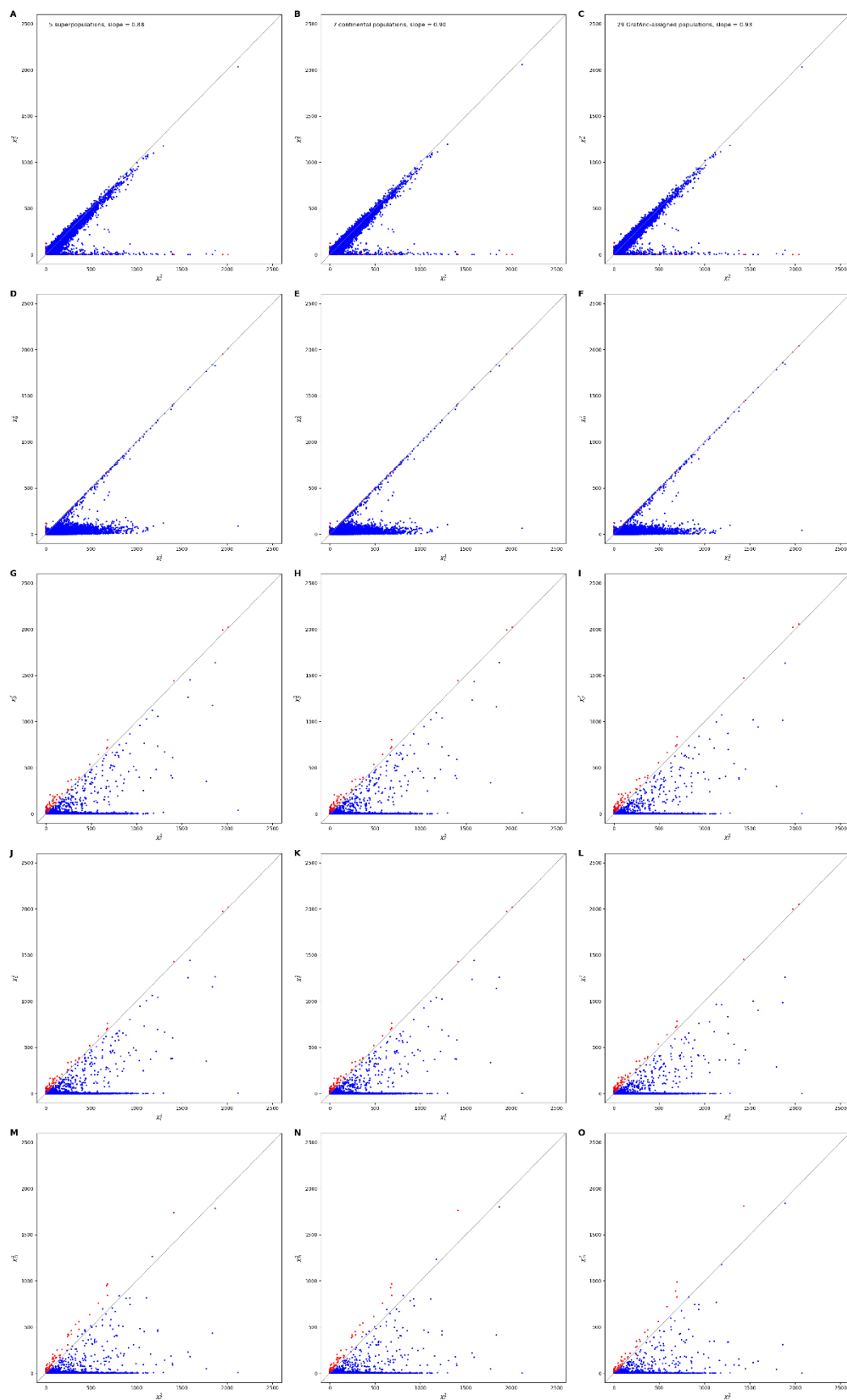


Fig. 11. Scatterplots of PopHWE and meta-analysis test statistics obtained using 1KGP dataset with GrafAnc ancestry SNPs. Same data and methods were used as in Fig. 10. Scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$ ,  $\chi_d^2$ ,  $\chi_p^2$ , and  $\chi_m^2$  were plotted in rows from the top to the bottom, respectively. The HetDef SNPs were plotted in blue, and the HetExc ones were plotted in red. In the top row, slopes of the linear regression equations of  $\chi_c^2$  vs.  $\chi_e^2$  values, obtained after SNPs with  $\chi_p^2 > 20$  were filtered, are displayed.

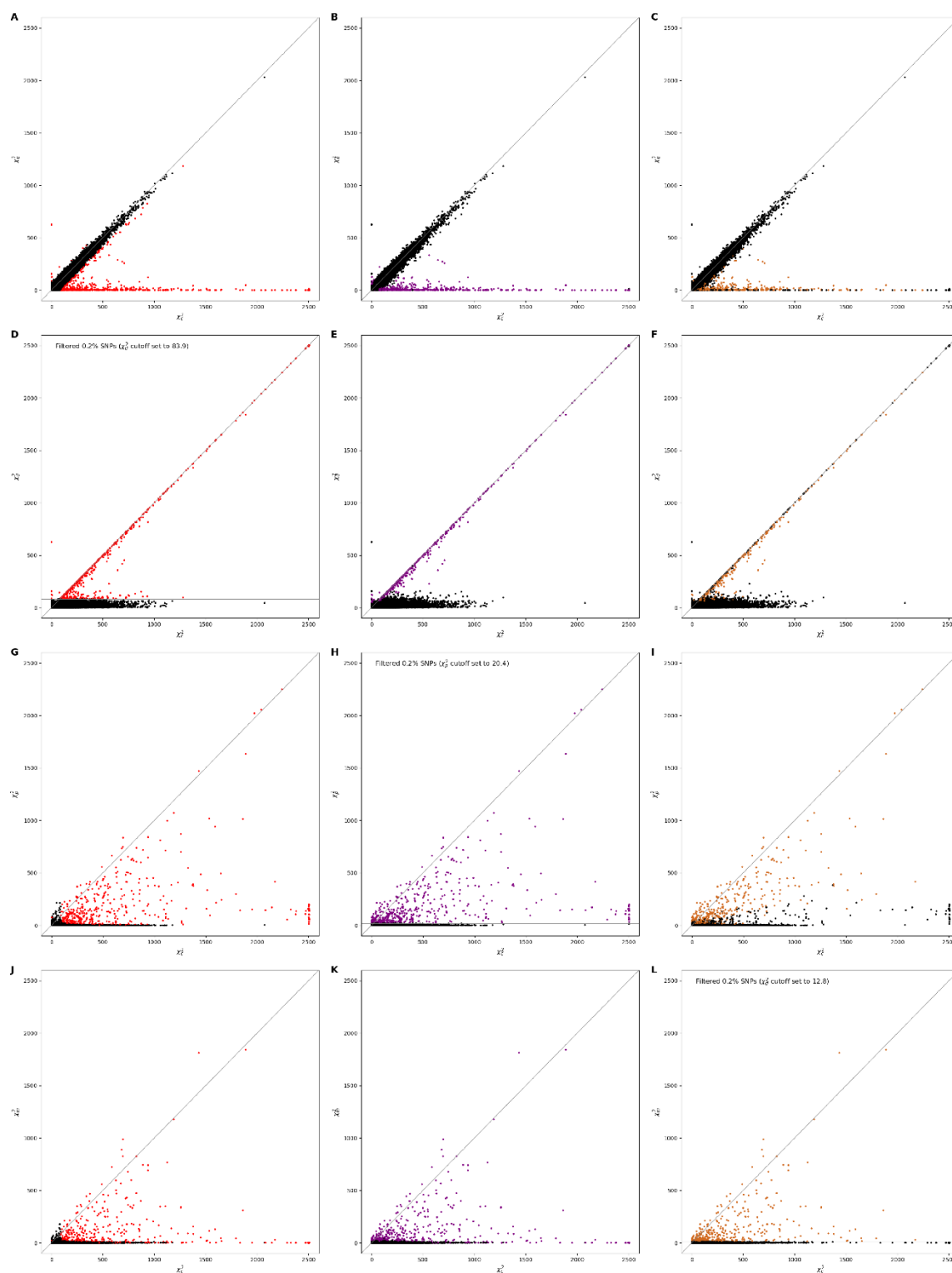


**Fig. 12.** Scatterplots of test results of 1KGP genotypes of GrafAnc ancestry SNPs showing SNPs filtered out using difference test statistic metrics. Same data and methods were used as in Fig. 10 and Fig. 11, except that the dataset was only partitioned into 29

ancestry groups. SNPs were filtered out so that the top 0.1% (282 out of 282K) with the highest test statistic scores were removed, corresponding to  $\chi_d^2 > 116.9$  (Column 1, in red),  $\chi_p^2 > 45.8$  (Column 2, in purple), and  $\chi_m^2 > 32.7$  (Column 3, in chocolate color). Scatterplots in each row shows the relationship between  $\chi_c^2$  vs.  $\chi_e^2$ ,  $\chi_d^2$ ,  $\chi_p^2$ , and  $\chi_m^2$ , in the top row to the bottom row, respectively.

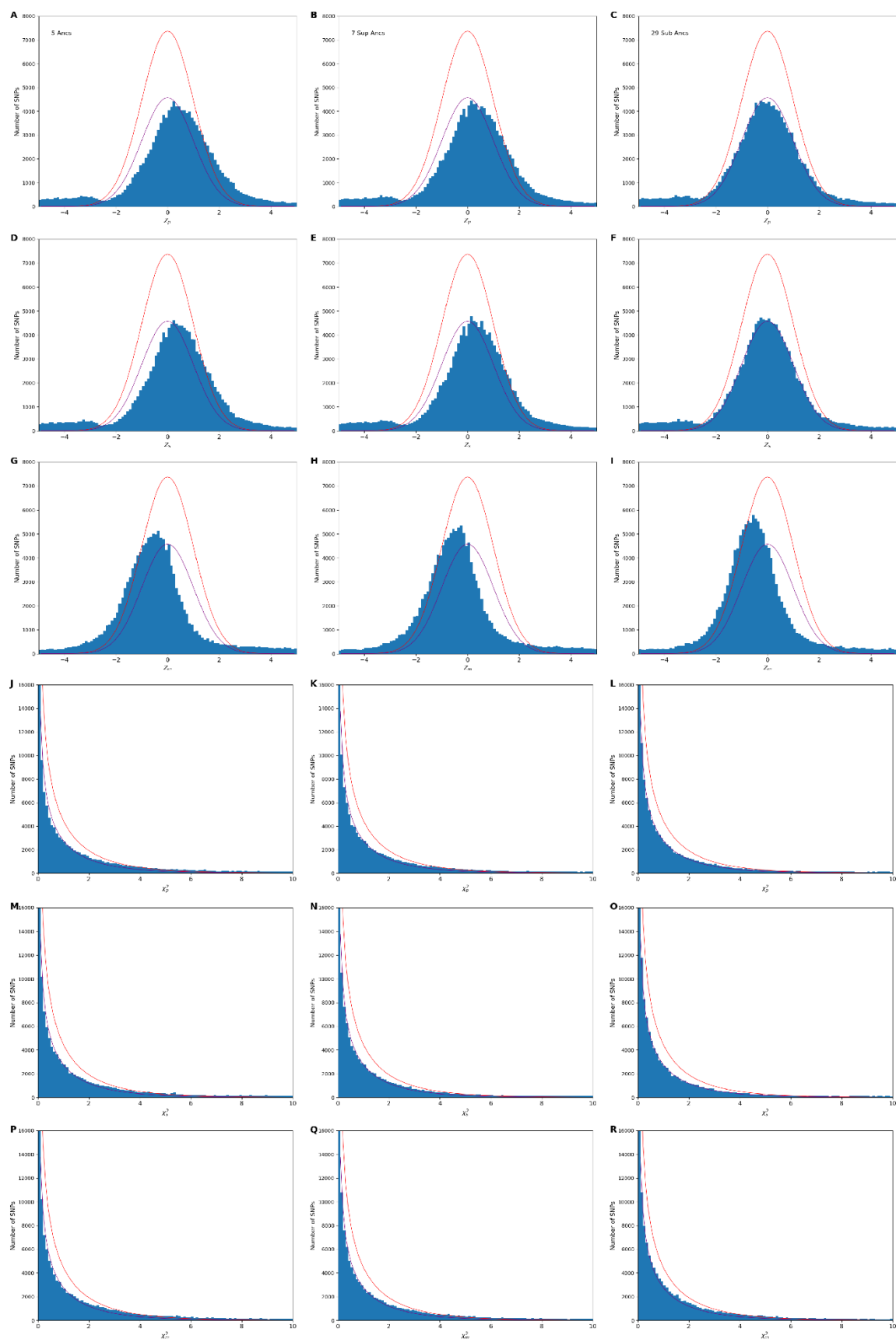


Fig. 13. Histograms of PopHWE and meta-analysis test statistics obtained using 1KGP Chromosome 22 dataset. After filtering variants with minor allele frequencies less than 0.01, 507K variants were plotted. Ancestry grouping was made the same way as that used in Fig.10, and the histograms were also plotted in the same way as in Fig. 10. The red curves show  $N(0, 1)$  and  $\chi^2_{(1)}$  distributions, with the area under the curve equaling to the total number of variants included. The purple curves also show  $N(0, 1)$  and  $\chi^2_{(1)}$  distributions, but the area under the curves equal to 62% of the variants, which is equivalent to plotting  $N(0, 1)$  and  $\chi^2_{(1)}$  for variants without errors, while assuming 38% of the variants have genotyping errors.

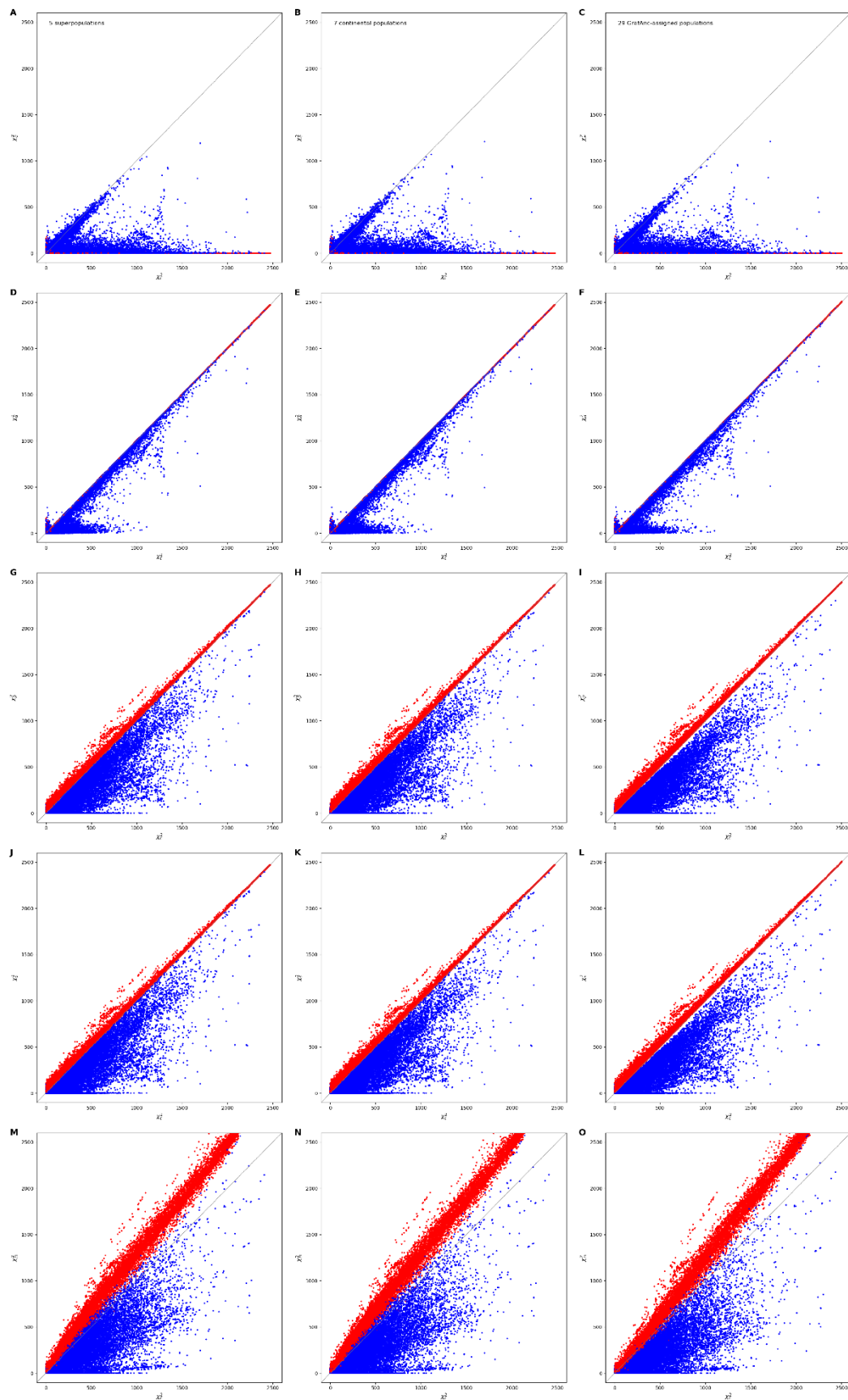


Fig. 14. Scatterplots of PopHWE and meta-analysis test statistics obtained using 1KGP Chromosome 22 dataset. Same data and methods were used as in Fig. 13. Scatterplots of  $\chi_c^2$  vs.  $\chi_e^2$ ,  $\chi_d^2$ ,  $\chi_p^2$ ,  $\chi_s^2$ , and  $\chi_m^2$  were plotted in rows from the top to the bottom, respectively. The HetDef SNPs were plotted in blue, and the HetExc ones were plotted in red.

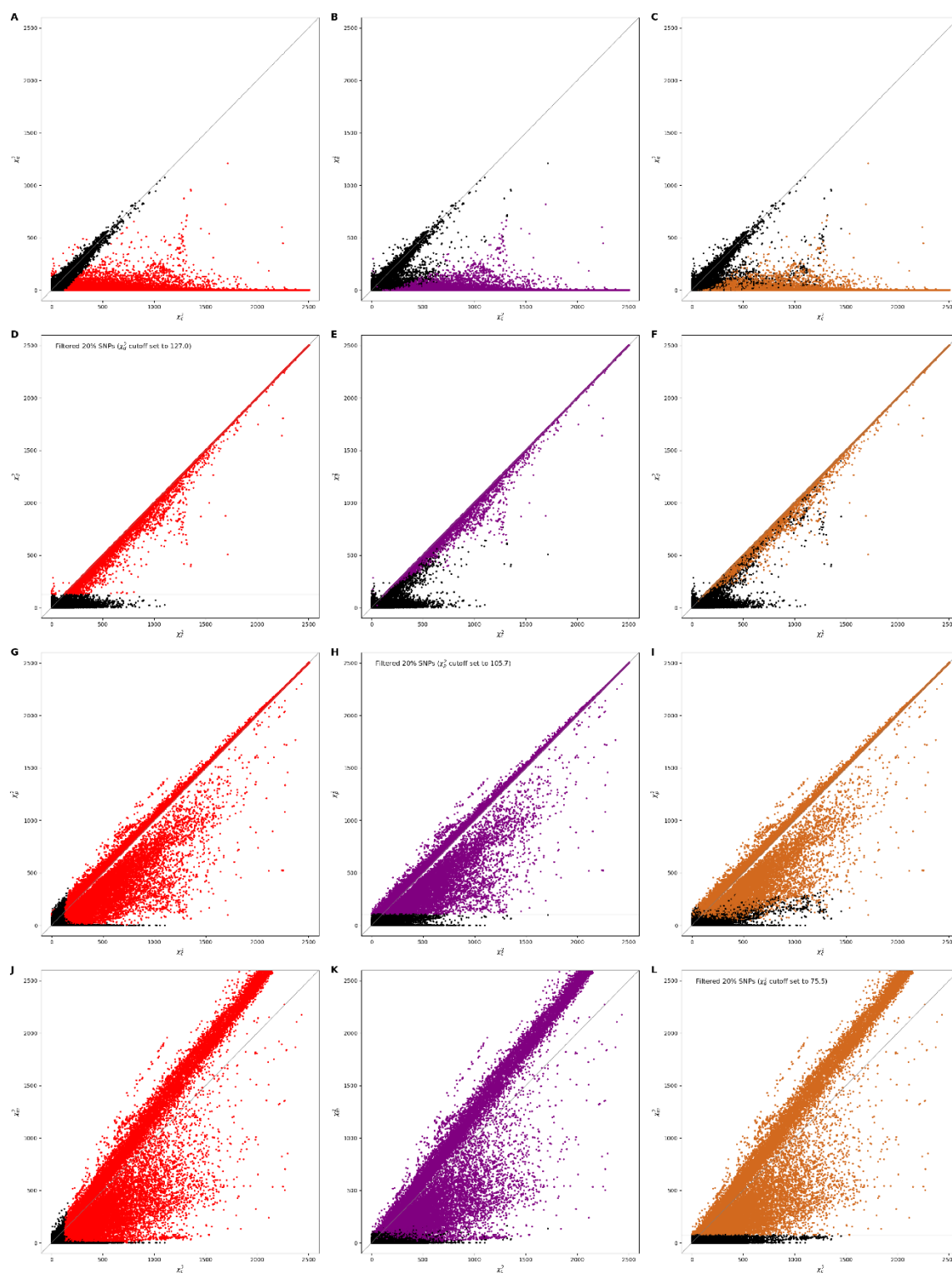


Fig. 15. Scatterplots of test results of 1KGP Chromosome 22 dataset showing SNPs filtered out using difference test statistic metrics. Same data and methods were used as in Fig. 13 and Fig. 14, except that the dataset was only partitioned into 29 continental

ancestry groups. SNPs were filtered out with thresholds  $\chi_d^2 > 116.9$  (Column 1, in red),  $\chi_p^2 > 45.8$  (Column 2, in purple), and  $\chi_m^2 > 32.7$  (Column 3, in chocolate color). Scatterplots in each row shows the relationship between  $\chi_c^2$  vs.  $\chi_e^2$ ,  $\chi_d^2$ ,  $\chi_p^2$ , and  $\chi_m^2$ , in the top row to the bottom row, respectively.

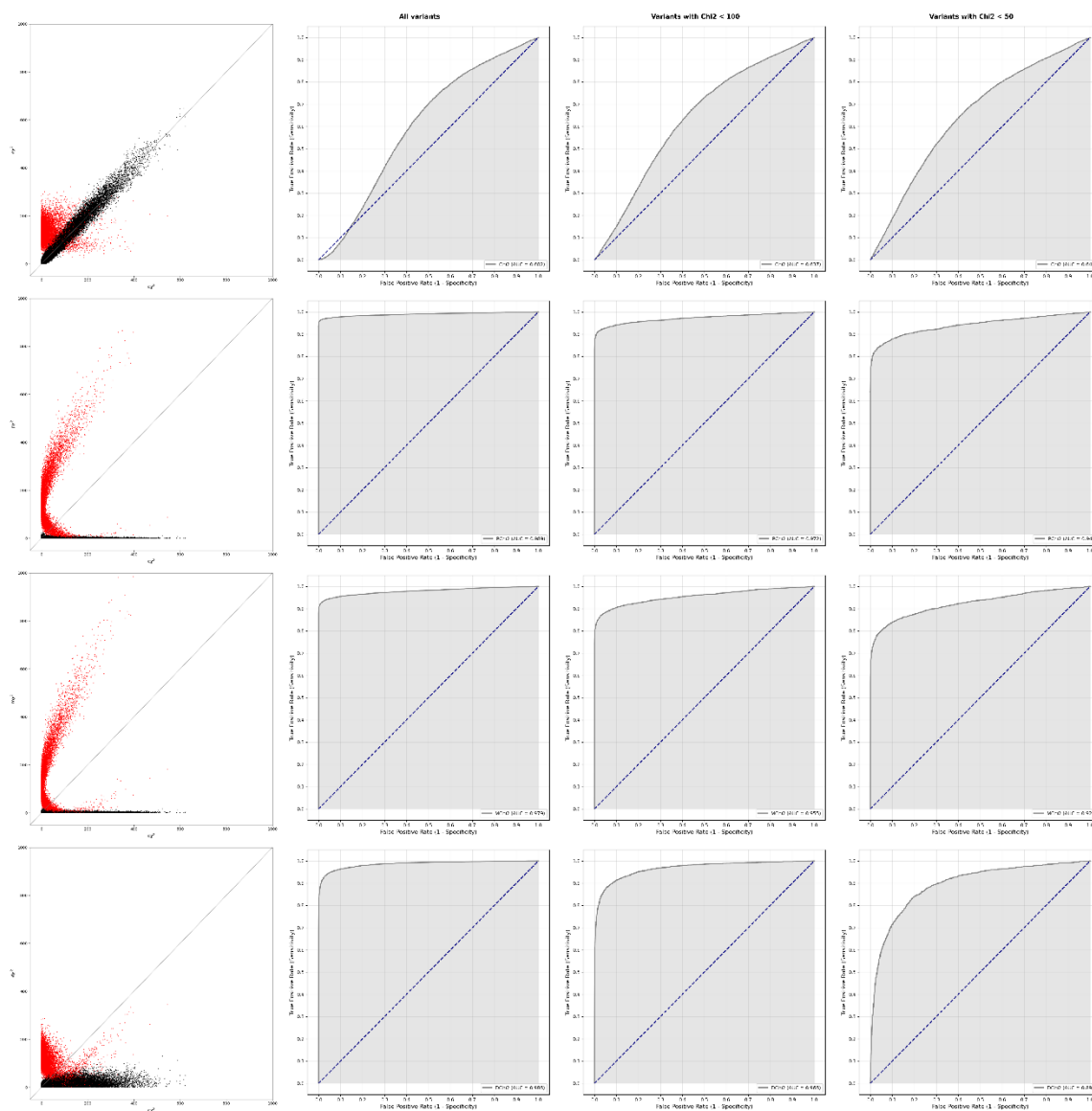


Fig. S1. Scatterplot of ROC curves of test dataset including 100 subpopulations with sample size 100 in each subpopulation.

