3-15-2021

Robust, flexible, and scalable tests for Hardy-Weinberg Equilibrium across diverse ancestries

Alan M. Kwong

Thomas W. Blackwell

Jonathon LeFaive

Mariza de Andrade

John Barnard

See next page for additional authors

Recommended Citation

Alan M Kwong, Thomas W Blackwell, Jonathon LeFaive, Mariza de Andrade, John Barnard, Kathleen C Barnes, John Blangero, Eric Boerwinkle, Esteban G Burchard, Brian E Cade, Daniel I Chasman, Han Chen, Matthew P Conomos, L Adrienne Cupples, Patrick T Ellinor, Celeste Eng, Yan Gao, Xiuqing Guo, Marguerite Ryan Irvin, Tanika N Kelly, Wonji Kim, Charles Kooperberg, Steven A Lubitz, Angel C Y Mak, Ani W Manichaikul, Rasika A Mathias, May E Montasser, Courtney G Montgomery, Solomon Musani, Nicholette D Palmer, Gina M Peloso, Dandi Qiao, Alexander P Reiner, Dan M Roden, M Benjamin Shoemaker, Jennifer A Smith, Nicholas L Smith, Jessica Lasky Su, Hemant K Tiwari, Daniel E Weeks, Scott T Weiss, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Analysis Working Group, Laura J Scott, Albert V Smith, Gonçalo R Abecasis, Michael Boehnke, Hyun Min Kang, Robust, flexible, and scalable tests for Hardy-Weinberg Equilibrium across diverse ancestries, Genetics, 2021;, iyab044, https://doi.org/10.1093/genetics/iyab044

Robust, flexible, and scalable tests for Hardy-Weinberg Equilibrium across diverse ancestries

Alan M. Kwong[1], Thomas W. Blackwell[1], Jonathon LeFaive[1], Mariza de Andrade[2], John Barnard[3], Kathleen C. Barnes[4], John Blangero[5], Eric Boerwinkle[6,7], Esteban G. Burchard[8,9], Brian E. Cade[10,11], Daniel I. Chasman[12], Han Chen[6,13], Matthew P. Conomos[14], L. Adrienne Cupples[15,16], Patrick T. Ellinor[17,18], Celeste Eng[9], Yan Gao[19], Xiuqing Guo[20], Marguerite Ryan Irvin[21], Tanika N. Kelly[22], Wonji Kim[23], Charles Kooperberg[24], Steven A. Lubitz[17,18], Angel C. Y. Mak[9], Ani W. Manichaikul[25], Rasika A. Mathias[26], May E. Montasser[27], Courtney G. Montgomery[28], Solomon Musani[29], Nicholette D. Palmer[30], Gina M. Peloso[15], Dandi Qiao[23], Alexander P. Reiner[24], Dan M. Roden[31], M. Benjamin Shoemaker[32], Jennifer A. Smith[33], Nicholas L. Smith[34,35,36], Jessica Lasky Su[23], Hemant K. Tiwari[37], Daniel E. Weeks[38], Scott T. Weiss[23], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Analysis Working Group, Laura J. Scott[1], Albert V. Smith[1], Gonçalo R. Abecasis[1], Michael Boehnke[1], Hyun Min Kang[1,*]

1 - Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109; 2 - Mayo Clinic, Rochester, MN 55905; 3 - Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44106; 4 - Department of Medicine, Anschultz Medical Campus, University of Colorado, Aurora, CO 80045; 5 - Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX 78520; 6 - Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030; 7 - Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030; 8 - Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94143; 9 - Department of Medicine, University of California San Francisco, San Francisco, CA 94143; 10 - Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA 02115; 11 - Division of Sleep Medicine, Harvard Medical School, Boston, MA 02115; 12 - Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA 02215; 13 - Center for Precision Health, School of Public Health and School of Biomedical

Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030; 14 - Department of Biostatistics, University of Washington, Seattle, WA 98195; 15 - Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118; 16 - Framingham Heart Study, Framingham, MA 01702; 17 - Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114; 18 - Cardiovascular Disease Initiative, The Broad Institute of MIT and Harvard, Cambridge, MA 02124; 19 - Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216; 20 - The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute at Harbor-UCLA Medical Center, Torrance, CA, 90502; 21 - Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294; 22 - Department of Epidemiology, Tulane University, New Orleans, LA 70112; 23 - Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115; 24 - Fred Hutchinson Cancer Research Center, Seattle, WA 98109; 25 - Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908; 26 - GeneSTAR Research Program and Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins University, Baltimore, MD 21205; 27 - Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; 28 - Sarcoidosis Research Unit, Genes and Human Disease Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104; 29 - Jackson Heart Study, University of Mississippi Medical Center, Jackson, MS 39216; 30 - Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC 27157; 31 - Departments of Medicine, Pharmacology, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232; 32 - Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232; 33 - Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48109; 34 - Department of Epidemiology, University of Washington, Seattle WA 98195; 35 - Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle WA 98101; 36 - Seattle Epidemiologic Research and Information Center, Office of Research and Development, Department of Veterans Affairs, Seattle WA 98108; 37 - Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294; 38 - Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, PA 15261

*Corresponding author: Center for Statistical Genetics and the Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109. E-mail: hmkang@umich.edu

KEYWORDS

population structure; principal components analysis; next-generation sequencing; genotype likelihoods

CORRESPONDING AUTHOR

Hyun Min Kang

Department of Biostatistics

University of Michigan School of Public Health

1415 Washington Heights

Ann Arbor, MI 48109

Phone: 734-647-1980

E-mail: hmkang@umich.edu

ABSTRACT

Traditional Hardy-Weinberg equilibrium (HWE) tests (the χ^2 test and the exact test) have long been used as a metric for evaluating genotype quality, as technical artifacts leading to incorrect genotype calls often can be identified as deviations from HWE. However, in datasets comprised of individuals from diverse ancestries, HWE can be violated even without genotyping error, complicating the use of HWE testing to assess genotype data quality. In this manuscript, we present the Robust Unified Test for HWE (RUTH) to test for HWE while accounting for population structure and genotype uncertainty, and evaluate the impact of population heterogeneity and genotype uncertainty on the standard HWE tests and alternative methods using simulated and real sequence datasets. Our results demonstrate that ignoring population structure or genotype uncertainty in HWE tests can inflate false positive rates by many orders of magnitude. Our evaluations demonstrate different tradeoffs between false positives and statistical power across the methods, with RUTH consistently amongst the best across all evaluations. RUTH is implemented as a practical and scalable software tool to rapidly perform HWE tests across millions of markers and hundreds of thousands of individuals while supporting standard VCF/BCF formats. RUTH is publicly available at https://www.github.com/statgen/ruth.

INTRODUCTION

Hardy-Weinberg equilibrium (HWE) is a fundamental theorem of population genetics and has been one of the key mathematical principles to understand the characteristics of genetic variation in a population for more than a century (HARDY 1908; WEINBERG 1908). Genetic variants in a homogeneous population typically follow HWE except for unusual deviations due to very strong case-control association and enrichment (NIELSEN *et al.* 1998), sex linkage, or non-random sampling (WAPLES 2015).

HWE tests are often used to assess the quality of microsatellite (VAN OOSTERHOUT *et al.* 2004), SNP-array (WIGGINTON *et al.* 2005), and sequence-based (DANECEK *et al.* 2011) genotypes. Testing for HWE may reveal technical artifacts in sequence or genotype data, such as high rates of genotyping error and/or missingness, or sequencing/alignment errors (NIELSEN *et al.* 2011). It can also identify hemizygotes in structural variants which are incorrectly called as homozygotes (MCCARROLL *et al.* 2006). Quality control for array- or sequence-based genotypes typically includes a HWE test to detect and filter out artifactual or poorly genotyped variants (LAURIE *et al.* 2010; NIELSEN *et al.* 2011).

While HWE tests are commonly and reliably used for variant quality control in samples from homogeneous populations, applying them to more diverse samples remains challenging. When analyzing individuals from a heterogeneous population, the standard HWE tests may falsely flag real, well-genotyped variants, unnecessarily filtering them out for downstream analyses (HAO AND STOREY 2019). This problem is important since genetic studies increasingly collect genetic data from heterogeneous populations. In principle, HWE tests in these structured populations can be performed on smaller cohorts with homogenous backgrounds

(BYCROFT *et al.* 2018), and the test statistics combined using Fisher's or Stouffer's method (MOSTELLER AND FISHER 1948; STOUFFER 1949). However, such a procedure requires much more effort than using a single HWE test across all samples. In addition, this approach cannot account for any heterogeneity within each of the smaller cohorts.

Here, we describe RUTH (Robust Unified Test for Hardy-Weinberg Equilibrium) which tests for HWE under heterogeneous population structure. Our primary motivation for developing RUTH is to robustly filter out artifactual or poorly genotyped variants using HWE test statistics. RUTH is (1) computationally efficient, (2) robust against various degrees of population structure, and (3) flexible in accepting key representations of sequence-based genotypes including best-guess genotypes and genotype likelihoods. We perform systematic evaluations of RUTH and alternative methods for HWE testing using simulated and real data to explore the advantages and disadvantages of these methods for samples of diverse ancestries.

MATERIALS AND METHODS

Unadjusted HWE tests

Consider a study of n participants with true (unobserved) genotypes $g_1, g_2, \cdots, g_n$ at a bi-allelic variant coded as 0 (reference homozygote), 1 (heterozygote), or 2 (alternate homozygote). Represent the best-guess/hard-call (observed) genotypes as $\hat{g}_1, \hat{g}_2, \cdots, \hat{g}_n$. A simple HWE test uses the chi-squared statistic to compare the expected and observed genotype counts assuming no population structure and no genotype uncertainty. The chi-squared HWE test statistic is defined as $T_{\chi^2} = \sum_{k=0}^{2} \frac{(c_k - \hat{c}_k)^2}{\hat{c}_k}$ where $c_j = \sum_{i=0}^{n} I(\hat{g}_i = j)$ (ignoring missing

genotypes), $\hat{p} = \frac{c_1+2c_2}{2n}, \hat{q} = 1 - \hat{p}\,, \hat{c}_0 = n\hat{q}^2, \hat{c}_1 = 2n\hat{p}\hat{q}$, and $\hat{c}_2 = n\hat{p}^2$. Under HWE, the asymptotic distribution of T_{χ^2} is assumed to follow χ_1^2 (ROHLFS AND WEIR 2008). An exact test is known to be more accurate for finite samples, particularly for rare variants (WIGGINTON *et al.* 2005), and using mid-p values instead of exact p-values will lead to slightly less conservative estimates (GRAFFELMAN AND MORENO 2013). HWE tests stratified by case-control status are known to prevent an inflation of Type I errors for disease-associated variants (LI AND LI 2008). Widely used software tools such as PLINK (PURCELL *et al.* 2007) and VCFTools (DANECEK *et al.* 2011) implement an exact HWE test based on best-guess genotypes. We will refer to the exact test as the unadjusted test.

Existing HWE tests accounting for structured populations

The unadjusted HWE test assumes a homogeneous population. If a study is comprised of a set of discrete structured subpopulations, a straightforward extension of the unadjusted test is to (1) stratify each study participant into exactly one of the subpopulations, (2) perform the unadjusted HWE test for each subpopulation separately, and (3) meta-analyze test statistics across subpopulations to obtain a combined p-value using Stouffer's method (STOUFFER *et al.* 1949). More specifically, let $z_1, z_2, \cdots, z_s$ be the z-scores from HWE test statistics for *s* distinct subpopulations with sample sizes $n_1, n_2, \cdots, n_s$. A combined meta-analysis HWE test statistic across the subpopulations is $T_{meta} = \frac{\sum_{i=1}^{s} z_i\sqrt{n_i}}{\sqrt{\sum_{i=1}^{s} n_i}}$, which asymptotically follows a standard normal distribution when each subpopulation follows HWE.

When the population cannot be easily stratified into distinct subpopulations (e.g. intra-continental diversity or an admixed population), a quantitative representation of genetic

ancestry, such as principal component (PC) coordinates or fractional mixture over subpopulations, can be more useful for representing genetic diversity (ROSENBERG *et al.* 2002; PRICE *et al.* 2006). HWES takes PCs as additional input to perform HWE tests under population structure with logistic regression (SHA AND ZHANG 2011), and a similar idea was suggested by Hao and colleagues (2016). However, existing implementations do not support sequence-based genotypes (where genotype uncertainty may remain at low or moderate sequencing depth) or other commonly used formats for genetic array data. A recent method PCAngsd estimates PCs from uncertain genotypes represented as genotype likelihoods (MEISNER AND ALBRECHTSEN 2019) and uses these estimates to perform a likelihood ratio test (LRT) for HWE, similar to the LRT version of RUTH with differences in computational performance (see below).

Robust HWE testing with RUTH

Here we describe RUTH (Robust and Unified Test for Hardy-Weinberg equilibrium) to enable HWE testing under structured populations, which is especially useful for large sequencing studies. We developed RUTH to produce HWE test statistics to allow quality control of sequence-based variant callsets from increasingly diverse samples. RUTH models the uncertainty encoded in sequence-based genotypes to robustly distinguish true and artifactual variants in the presence of population structure, and seamlessly scales to millions of individuals and genetic variants.

We assume the observed genotype for individual i can be represented as a genotype likelihood (GL) $L_i^{(G)} = \Pr(Data_i | g_i = G)$, where $Data_i$ represents observed data (e.g. sequence or array), and $g_i \in \{0,1,2\}$ the true (unobserved) genotype. For example, GLs for

sequence-based genotypes can be represented as $L_i^{(G)} = \prod_{j=1}^{d_i} \Pr(r_{ij}|g_i = G; q_{ij})$ where d_i is the sequencing depth, r_{ij} is the observed read, and q_{ij} is the corresponding quality score (EWING AND GREEN 1998; JUN *et al.* 2012). We model GLs for best-guess genotypes $\hat{g}_i$ from SNP arrays as $L_i^{(G)} = (1-e_i)^2,\ 2e_i(1-e_i),\ e_i^2$ for $\hat{g}_i = 2, 1, 0$ where e_i is the assumed per-allele error rate. Imputed genotypes may also be approximately modeled using this framework, but the current implementation requires creating a pseudo-genotype likelihood to describe this uncertainty (see Discussion).

Accounting for Population Structure with Individual-Specific Allele Frequencies

We account for population structure by modeling individual-specific allele frequencies from quantitative coordinates of genetic ancestry such as PCs, similar to HAO *et al.* (2016). For any given variant, instead of assuming that genotypes follow HWE with a single universal allele frequency across all individuals, we assume that genotypes follow HWE with heterogeneous allele frequencies specific to each individual, modeled as a function of genetic ancestry. Let $\boldsymbol{x_i} \in \mathbb{R}^k$ represent the genetic ancestry of individual i, where k is the number of PCs used. We estimate individual-specific allele frequency p as a bounded linear function of genetic ancestry

$$p(\boldsymbol{x}_i; \boldsymbol{\beta}) = \begin{cases} \boldsymbol{\beta}^T \boldsymbol{x}_i & \varepsilon \leq \boldsymbol{\beta}^T \boldsymbol{x}_i \leq 1-\varepsilon \\ \varepsilon & \boldsymbol{\beta}^T \boldsymbol{x}_i < \varepsilon \\ 1-\varepsilon & \boldsymbol{\beta}^T \boldsymbol{x}_i > 1-\varepsilon \end{cases},$$

where ε is the minimum frequency threshold. We estimate $\widehat{\boldsymbol{\beta}}$ with an E-M algorithm. We used $\varepsilon = \frac{1}{4n}$ in our evaluation. Even though we used a linear model for $p(\boldsymbol{x}_i; \boldsymbol{\beta})$ for computational efficiency, it is straightforward to apply a logistic model, which is arguably better (YANG *et al.* 2012; HAO *et al.* 2016).

Let $p_i = p(\boldsymbol{x}_i; \boldsymbol{\beta})$ and $q_i = 1 - p_i$ be the individual specific allele frequencies of the non-reference and reference alleles for individual i. Under the null hypothesis of HWE, the frequencies of genotypes (0, 1, 2) are $[q_i^2,\ 2p_iq_i,\ p_i^2]$. Under the alternative hypothesis, we assume these frequencies are $[q_i^2 + \theta p_i q_i,\ 2p_iq_i(1-\theta),\ p_i^2 + \theta p_i q_i]$ where θ is the inbreeding coefficient. This model is a straightforward extension of a fully general model where p_i, q_i is identical across all samples. Then the log-likelihood across all study participants is

$$l(\boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \log\left[L_i^{(0)}(q_i^2 + \theta p_i q_i) + L_i^{(1)}\, 2p_iq_i(1-\theta) + L_i^{(2)}(p_i^2 + \theta p_i q_i)\right]$$

Under both the null ($\theta = 0$) and alternative ($\theta \neq 0$) hypotheses, we maximize the log-likelihood using an Expectation-Maximization (E-M) algorithm (DEMPSTER *et al.* 1977). As we empirically observed quick convergence within several iterations in most cases, we used a fixed (n=20) number of iterations in our implementation (Figure S2).

RUTH Score Test

The score function of the log-likelihood is the derivative of the log-likelihood with respect to θ:

$$U(\theta) = \sum_{i=1}^{n} \frac{p_iq_i\left[L_i^{(0)} - 2L_i^{(1)} + L_i^{(2)}\right]}{L_i^{(0)}(q_i^2 + \theta p_i q_i) + L_i^{(1)}\ 2p_iq_i(1-\theta) + L_i^{(2)}(p_i^2 + \theta p_i q_i)} = \sum_{i=1}^{n} u_i(\theta)$$

Since $u_i'(\theta) = -u_i^2(\theta)$, we construct a score test statistic of H_0: $\theta = 0$ vs H_1: $\theta \neq 0$ as:

$$T_{score} = \frac{[U(0)]^2}{I(0)} = \frac{[\sum_{i=1}^{n} u_i(0)]^2}{\sum_{i=1}^{n} u_i^2(0)}$$

where $I(0)$ is the Fisher information under the null hypothesis. Under the null, T_{score} has an asymptotic chi-squared distribution with one degree of freedom, i.e. $T_{score} \sim \chi_1^2$. A detailed algorithm can be found in Figure S1.

RUTH Likelihood Ratio Test

The log-likelihood function $l(\boldsymbol{\beta}, \theta)$ can also be used to calculate a likelihood ratio test statistic:

$$T_{LRT} = 2\left[\max_{\boldsymbol{\beta},\theta} l(\boldsymbol{\beta}, \theta) - \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, 0)\right].$$

Like the score test, we estimate MLE parameters $\boldsymbol{\beta}, \theta$ iteratively using an E-M algorithm to test $H_0: \theta = 0$ vs $H_1: \theta \neq 0$. Under the null hypothesis, the asymptotic distribution of T_{LRT} is expected to follow χ_1^2. This test is very similar to the likelihood-ratio test proposed by PCAngsd (MEISNER AND ALBRECHTSEN 2019), except PCAngsd does not re-estimate $\boldsymbol{\beta}$ under the alternative hypothesis. In principle, the RUTH LRT should be slightly more powerful due to this difference; we expect the practical difference in power to be small, as deviations from HWE usually do not change the estimates of $\boldsymbol{\beta}$ substantially.

Simulation of genotypes and sequence reads under population structure

We simulated sequence-based genotypes under population structure using the following procedure. First, for each variant, we simulated an ancestral allele frequency and population-specific allele frequencies. Second, we sampled unobserved (true) genotypes based on these allele frequencies. Third, we sampled sequence reads based on the unobserved genotypes. Fourth, we generated genotype likelihoods and best-guess genotypes based on sequence reads.

Our goal was to simulate variants such that each subpopulation will have different average allele frequencies from other subpopulations.

To simulate ancestral and population-specific allele frequencies, we followed the BALDING AND NICHOLS (1995) procedure, except we sampled ancestral allele frequencies from $p \sim Uniform(0,1)$ instead of $p \sim \mathrm{Uniform}(0.1, 0.9)$ to include rare variants. For each of $K \in \{1, 2, 5, 10\}$ populations, we sampled population-specific allele frequencies from $p_k \sim Beta\left(\frac{p(1-F_{st})}{F_{st}}, \frac{(1-p)(1-F_{st})}{F_{st}}\right)$, where $k \in \{1, \cdots, K\}$, and $F_{st} \in \{.01, .02, .03, .05, .10\}$ was the fixation index to quantify the differentiation between populations, as suggested by Holsinger (HOLSINGER 1999) and implemented in previous studies (HOLSINGER *et al.* 2002; BALDING 2003). Because p_k no longer follows the uniform distribution, we used rejection sampling to ensure that $\bar{p} = \frac{1}{K}\sum_{k=1}^{K} p_k$ is uniformly distributed across 100 bins across simulations to avoid artifacts caused by systematic differences in allele frequencies.

The unobserved genotype $G_i \in \{0,1,2\}$ for individual $i \in \{1, \cdots, n_k\}$, belonging to population k with sample size n_k, was simulated from genotype frequencies $(q_k^2 + \theta\, p_k q_k, 2p_k q_k(1-\theta), p_k^2 + \theta\, p_k q_k)$, where $q_k = 1 - p_k$ and $\theta \in \left[-\min\left(\frac{q_k}{p_k}, \frac{p_k}{q_k}\right), 1\right]$ quantifies deviation from HWE; $\theta = 0$ represents HWE, while $\theta < 0$ and $\theta > 0$ represent excess heterozygosity and homozygosity compared to HWE expectation, respectively. In our experiments, we evaluated $\theta \in \{0, \pm.01, \pm.05, \pm.1, \pm.5\}$. When θ was smaller than the minimum possible value for a specific population, we replaced it with the minimum value.

We simulated sequence reads based on unobserved genotypes, sequence depths, and base call error rates. To reflect the variation of sequence depths between individuals, we

simulated the mean depth of each sequenced sample as $\mu_i \sim Uniform(1, 2D-1)$, where D is the expected depth and $D = 5$ and $D = 30$ representing low-coverage and deep sequencing, respectively. For each sequenced sample and variant site, we sampled the sequence depth from $d_i \sim Poisson(\mu_i)$. Each sequence read carried either of the possible unobserved (true) alleles $r_{ij} \in \{0,1\}$, where $j \in \{1, \cdots, d_i\}$. Given unobserved genotype G_i, we generated $r_{ij} \sim Bernoulli\left(\frac{G_i}{2}\right)$, with observed allele $o_{ij} = (1-e_{ij})r_{ij} + e_{ij}(1-r_{ij})$ flipping to the other allele when a sequencing error occurs with probability $e_{ij} \sim Bernoulli(\epsilon)$. We used $\epsilon = 0.01$ throughout our simulations (which corresponds to phred-scale base quality of 20) and assumed that all base calling errors switched between reference and alternate alleles.

We then generated genotype likelihoods and best-guess genotypes from the simulated alleles. Let $t_i = \sum_{j=1}^{d_i} o_{ij}$ be the observed alternate allele count. The GLs for the three possible genotypes are $L_i^{(0)} = (1-\epsilon)^{d_i - t_i}(\epsilon)^{t_i}$, $L_v^{(1)} = 0.5^{d_i}$, $L_i^{(2)} = (\epsilon)^{d_i - t_i}(1-\epsilon)^{t_i}$. We called best-guess genotypes by using the overall ancestral allele frequency $\bar{p}$ for a given variant as the prior, then calling the genotype corresponding to the highest posterior probability among $\left(L_i^{(0)}(1-\bar{p})^2,\ 2L_i^{(1)}\bar{p}(1-\bar{p})^2,\ L_i^{(2)}\bar{p}^2\right)$ for each individual. For each possible combination of F_{st}, K, and θ, we generated 50,000 independent variants across a set of $n = 5{,}000$ samples with per-ancestry samples sizes $n_k = \frac{n}{K}$.

Evaluation of Type I Error and Statistical Power

We used different p-value thresholds, F_{st} values, number of ancestry groups K, and average sequencing depth D to determine the number of variants significantly deviating from HWE. To evaluate Type I error, we simulated sequence reads under HWE ($\theta = 0$) and calculated the

proportion of significant variants at each p-value threshold. In RUTH tests, we assumed PCs were accurately estimated using true genotypes unless indicated otherwise. For real data, we summarized ancestral information by projecting PCs estimated from full genomes onto the reference PC space of the Human Genome Diversity Panel (HGDP) (LI *et al.* 2008) using verifyBamID2 (ZHANG *et al.* 2020), similar to the procedure for variant calling in the TOPMed Project, which has integrated RUTH as part of its quality control pipeline (https://github.com/statgen/topmed_variant_calling).

In all datasets, we evaluated the tradeoff between Type I Error and power for each method using precision-recall curves (PRCs) and receiver-operator characteristic curves (ROCs). In simulated data, we considered variants with $\theta = 0$ to be true negatives and variants with $\theta = -0.05$ to be true positives. For real data, we labeled HQ variants as negative and LQ variants as positive.

Data source

To evaluate our method, we used sequence-based genotype data from the 1000 Genomes Project (1000G) (THE 1000 GENOMES PROJECT CONSORTIUM *et al.* 2015) and the Trans-Omics Precision Medicine (TOPMed) Project (TALIUN *et al.* 2019). In both cases, we used subsets of variants from chromosome 20. For 1000G, we started with 1,812,841 variants in 2,504 individuals, with an average depth of $7.0\times$. For TOPMed, we started with 12,983,576 variants in 53,831 individuals, with an average depth of $37.2\times$.

Application to 1000 Genomes data

To test our method on 1000G data, we first needed to define two sets of variants: one set which is expected to follow HWE, and another set which is expected to deviate from HWE. Unlike simulated data, variants in 1000G are not clearly classified into "true" or "artifactual", so evaluation of false positives and power is less straightforward. We focused on two subsets of variants in chromosome 20 which serve as proxies for these two variant types. We selected non-monomorphic sites found in both the Illumina Infinium Omni2.5 genotyping array and in HapMap3 (THE INTERNATIONAL HAPMAP CONSORTIUM *et al.* 2010) as "high-quality" (HQ) variants that mostly follow HWE after controlling for ancestry, ending up with 17,740 variants. We selected variants that displayed high discordance between duplicates or Mendelian inconsistencies within family members in TOPMed as "low quality" (LQ) variants which should be enriched for deviations from HWE even after accounting for ancestry, ending up with 10,966 variants. Among 329,699 LQ variants from TOPMed in chromosome 20, we found that only 10,966 overlap with 1000 Genome samples. We suspect that a substantial fraction of these 10,966 LQ variants are true variants since they passed all of the 1000G Project's quality filters. Nevertheless, we still expect a much larger fraction of these LQ variants to deviate from HWE compared to HQ variants.

We evaluated multiple representations of sequence-based genotypes from 1000G. As 1000G samples were sequenced at relatively low-coverage of $7.0\times$ on average, best-guess genotypes inferred only from sequence reads (raw GT) tend to have poor accuracy. Therefore, the officially released best-guess genotypes in 1000G were estimated by combining genotype likelihoods (GL), calculated based on sequence reads, with haplotype information from nearby variants through linkage-disequilibrium (LD)-aware genotype refinement using SHAPEIT2 (DELANEAU *et al.* 2013). This procedure resulted in more accurate genotypes (LD-aware GT), but

it implicitly assumed HWE during refinement. As different representations of sequence genotypes may result in different performance in HWE tests, we evaluated all three representations - raw GT, LD-aware GT, and GL. In all tests of RUTH using hard genotype calls, we assumed the error rate for GT-based genotypes to be 0.5%, which is representative of a typical non-reference genotype error rate for SNP arrays. We restricted our analyses to biallelic variants. The positions and alleles of 1000G and TOPMed variants were matched using the liftOver software tool (KUHN *et al.* 2013).

We evaluated all tests as described above. For meta-analysis with Stouffer's method, we divided the samples into 5 strata, using the five 1000G super population code labels – African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). To obtain PC coordinates for 1000G samples, we estimated 4 PCs from the aligned sequence reads (BAM) with verifyBamID2 (ZHANG *et al.* 2020), using PCs from 936 samples from the Human Genome Diversity Project (HGDP) panel as reference coordinates. The RUTH score test and LRT used these PCs as inputs, along with genotypes in raw GT, LD-aware GT, and GL formats. For PCAngsd, we used GLs from all variants tested as the input. We limited the analysis to a single chromosome due to the heavy computational requirements of PCAngsd.

Application to TOPMed Data

We analyzed variants from 53,831 individuals from the TOPMed sequencing study (TALIUN *et al.* 2019). These samples came from multiple studies from a diverse spectrum of ancestries, leading to substantial population structure. Using the same criteria as our 1000G analysis, we identified 17,524 high-quality variants and 329,699 low-quality variants across chromosome 20. Since TOPMed genomes were deeply sequenced at $37.2 \times (\pm 4.5 \times)$, LD-aware genotype

refinement was not necessary to obtain accurate genotypes. Therefore, we used two genotype representations – raw GT and GL – in our evaluations. This genotype data contained no missingness.

Similar to 1000G, for best-guess genotypes (raw GT), we used PLINK for the unadjusted test. For meta-analysis, we assigned each sample to one of the five 1000G super populations as follows. First, we summarized the genetic ancestries of aligned sequenced genomes with verifyBamID2 by estimating 4 PCs using HGDP as reference. Second, we used Procrustes analysis (DRYDEN AND MARDIA 1998; WANG *et al.* 2010) to align the PC coordinates of HGDP panels (to account for different genome builds) so that the PC coordinates were compatible between TOPMed and 1000G samples. Third, for each TOPMed sample, we identified the 10 closest corresponding individuals from 1000G using the first 4 PC coordinates with a weighted voting system (assigning the closest individual a score of 10, next closest a score of 9, and so on until the 10th closest individual is assigned a score of 1, then adding up the scores for each super population) to determine the super population code that had the highest sum of scores, and therefore best described that sample. In this way, we classified 15,580 samples as AFR, 4,836 as AMR, 29,943 as EUR, 2,960 as EAS, and 716 as SAS. Among these samples, 94.5% had the same super population code for all 10 nearest 1000G neighbors. To evaluate the RUTH score test and LRT for both raw GT and GL, we used 4 PCs estimated by verifyBamID2 (ZHANG *et al.* 2020), consistent with the method applied for the 1000G data.

Impact of Ancestry Estimates on Adjusted HWE Tests

We examined the effect of changing the number of PCs used as input for RUTH tests by using 2 PCs as opposed to 4 PCs. We also evaluated the impact of using different approaches to classify

ancestry when adjusting for population structure with meta-analysis. By default, our analysis classified the 1000 Genomes subjects into 5 continental super populations based on published information (THE 1000 GENOMES PROJECT CONSORTIUM *et al.* 2015). For TOPMed, the best-matching 1000 Genomes continental ancestry was carefully determined using the PCA-based matching strategy described above. However, in practice, ancestry classification may be performed with a coarser resolution (JIN *et al.* 2019). To mimic plausible scenarios in which sample ancestries are not carefully determined, we used k-means clustering on the first 2 PCs of our samples to divide individuals into 3 distinct groups, roughly corresponding to East Asian, European, and African populations, and performed meta-analyses based on this coarse classification for both 1000G and TOPMed data.

RESULTS

Simulation: Effect of Genotype Uncertainty

To evaluate the impact of genotype uncertainty, we first compared tests in the absence of population structure (i.e. single ancestry). For the unadjusted test, we used only best-guess genotypes (GTs). For PCAngsd, we used only genotype likelihoods (GLs). For RUTH score and likelihood ratio tests, we used both.

Using GLs over GTs substantially reduced Type I errors in HWE tests, especially in low-coverage data (Figure 1A-C). For example, the standard HWE test based on GTs resulted in a 229-fold inflation (22.9%) at $p < .001$ (Figure 1B, Table S1), a threshold which allows the evaluation of Type I error with reasonable precision with 50,000 variants (50 expected false positives under the null). GT-based RUTH-Score and RUTH-LRT tests showed similar inflation.

When GLs were used instead of best-guess genotypes, RUTH-Score and RUTH-LRT had Type I errors close to the null expectation (.001 for RUTH-Score and .0012 for RUTH-LRT). PCAngsd, which also accounts for genotype uncertainty (MEISNER AND ALBRECHTSEN 2019), had similar performance. The severely inflated Type I errors with best-guess genotypes can largely be attributed to high uncertainty and bias towards homozygote reference genotypes in single site calls from low-coverage sequence data, resulting in apparent deviations from HWE. For high-coverage sequence data, inflation of Type I error with GTs was substantially attenuated; inflation nearly disappeared when using GLs (.004 for RUTH-Score and .002 for RUTH-LRT; Figure 1D-F).

Next, we evaluated the power to identify variants truly deviating from HWE at various levels of inbreeding (θ). For low-coverage sequence data, we skip interpretation of power of GT-based tests owing to their extremely inflated false positive rates. All GL-based tests behaved similarly, achieving ~19-21% power at $p < .001$ with moderate excess heterozygosity ($\theta = -0.05$) (Figure 2B, Table S1). For high-coverage sequence data, the power of GL-based tests at the same p-value threshold increased to ~56-60%, comparable to corresponding GT-based tests. Interestingly, the unadjusted GT-based test showed much lower power than RUTH and PCAngsd tests under excess heterozygosity ($\theta < 0$) while demonstrating much higher power with excess homozygosity ($\theta > 0$). Upon further investigation, we observed that the tests have lower power than the exact test specifically for rare variants with excess homozygosity due to the mismatch between the empirical and asymptotic null distributions (see Discussion for details).

We also generated precision-recall curves (PRC) and receiver-operator characteristic (ROC) curves to better understand the tradeoff between the Type I errors and power under

moderate excess heterozygosity (θ = -.05) (Figure S3C-D). Again, accounting for genotype uncertainty resulted in better empirical power and Type I error, especially for low-coverage data: at an empirical false positive rate of 1%, GL-based tests had 41-45% power, as opposed to 4-10% for GT-based tests. For high-coverage data, GL-based tests had 1-2% greater power than GT-based tests at the same false positive rate. These results suggest that ignoring genotype uncertainty in HWE tests is reasonable for high-coverage sequence data.

Simulation: Impact of Population Structure on HWE Test Statistics

As expected, the unadjusted HWE test had substantially inflated Type I errors under population structure based on the Balding-Nichols (1995) model (Figure 1, Table S1). Even for an intra-continental level of population differentiation (F_{ST} = .01), the Type I errors at $p < .001$ were inflated 13.5-fold even for high-coverage data. With an inter-continental level of differentiation (F_{ST} = .1), we observed orders of magnitude more Type I errors across different simulation conditions. This inflation is expected to increase with larger sample sizes, suggesting that adjustment for population structure is important even if a study focuses on a single continental population.

One simple approach to account for population structure is to stratify individuals into distinct subpopulations and apply HWE tests separately, as was done in UK Biobank (Bycroft *et al.* 2018), then meta-analyze the results (Figure 3B). Type I errors were appropriately controlled with this approach in high-coverage but not low-coverage data, likely due to unmodeled genotype uncertainty (Figure 1, Table S1). Instead of classifying individuals into distinct subpopulations, RUTH incorporates PCs to jointly perform HWE tests (Figure 3C). By estimating individual-specific allele frequencies, RUTH was able to adjust for the simulated population

structure. For both low- or high-coverage data, GL-based RUTH tests and PCAngsd showed well-controlled Type I errors, while GT-based tests showed slight (high-coverage) to severe (low-coverage) inflation.

Although meta-analysis resulted in well-controlled Type I errors for high-coverage data, it was considerably less powerful than RUTH. For example, with moderate excess heterozygosity (θ = -.05) across five ancestries (F_{ST} = .1), RUTH tests identified 20-27% more variants as significant at $p < .001$ (Figure 2, Table S1) compared to meta-analysis. PRCs also clearly showed better operating characteristics for RUTH and PCAngsd compared to meta-analysis (Figure S4). For example, at an empirical false positive rate of 1%, RUTH showed much greater power (66-68%) than meta-analysis (43%), even though the simulation scenario favors meta-analysis because samples were perfectly classified into distinct subpopulations. When stratified by allele frequency, RUTH showed better operating characteristics for common variants compared to rare variants due to a difference in power (Figure S5).

Application to 1000 Genomes WGS data

Next, we evaluated the performance of various HWE tests in low-coverage (~6x) sequence data from the 1000 Genomes Project. We evaluated three representations of genotypes - (1) raw GT, (2) LD-aware GT, and (3) GL, as described in Materials and Methods. Among chromosome 20 variants, we selected 17,740 high-quality (HQ) variants that are polymorphic in GWAS arrays, and 10,966 low-quality (LQ) variants enriched for genotype discordance in duplicates and trios. Unlike simulation studies, not all LQ variants are expected to violate HWE, so we consider the proportion of significant LQ variants as a lower bound for the sensitivity to identify significant

variants. Similarly, not all HQ variants are expected to follow HWE, so the proportion of significant HQ variants serves as an upper bound for the false positive rate.

Consistent with simulation results, all tests based on raw GTs generated from low-coverage sequence data had severe inflations of false positives (Figure 4A, Table 1). This was true even for HQ variants, presumably due to genotyping errors and bias in raw GTs. Standard HWE tests, which model neither genotype uncertainty nor population structure, showed the highest inflation of false positives at 44% for $p < 10^{-6}$, a threshold commonly used for HWE testing in large genetic studies (LOCKE *et al.* 2015; FRITSCHE *et al.* 2016). Modeling population structure substantially reduced inflation, with RUTH tests showing fewer false positives (0.7-1.0% at $p < 10^{-6}$) than meta-analysis (2.0% at $p < 10^{-6}$). False positives were inflated across all methods when using raw GTs.

Similarly, GL-based RUTH tests further reduced false positives (0.034% at $p < 10^{-6}$). In contrast to our simulations, however, PCAngsd demonstrated considerably higher false positives than RUTH (2.1% at $p < 10^{-6}$) because PCAngsd estimates PCs from the input data without the ability to use externally provided PCs (see Discussion). The sensitivity for detecting significant LQ variants was also consistent with our simulations (Figure 4B, Table 1). GL-based tests, which showed better control of false positives, identified 22-25% of LQ variants as significant at $p < 10^{-6}$.

Strikingly, while using LD-aware GTs reduced false positives with adjusted tests, it was at the expense of substantially reduced sensitivity to detect LQ variants. The false positive rates of any adjusted test with LD-aware GTs were uniformly lower than those of any GL- and raw GT-based tests across all p-value thresholds (Figure 4A). However, sensitivity was also substantially

reduced with LD-aware genotypes (Figure 4B). For example, at $p < 10^{-6}$, GL-based RUTH tests identified 22-23% of LQ variants as significant, while using LD-aware GTs halved the proportions. Meta-analysis with LD-aware GTs had even lower sensitivity, likely because the implicit HWE assumption in LD-aware genotype refinement altered the LD-aware genotypes to conform to HWE, further reducing both false positives and sensitivity.

We evaluated PRCs between HQ and LQ variants to further evaluate this tradeoff. The results clearly demonstrated that HWE tests using LD-aware GTs are substantially less robust than tests using other genotype representations (Table S2, Figure S6A). For example, for the RUTH score test, when LD-aware GTs identified 0.1% of HQ variants as significant, 17% of LQ variants were identified as significant. However, with raw GT and GL, 24~27% were identified as significant at the same threshold. Even fewer were significant in meta-analysis with LD-aware GTs (13%). Similar trends were observed across all thresholds, suggesting that using LD-aware GTs results in substantially poorer operating characteristics. As more accurate genotyping in LD-aware genotype refinement is expected to improve the performance of QC metrics compared to raw GTs, these results are quite striking, and highlight a potential oversight in using LD-aware genotypes in various QC metrics for sequence-based genotypes. It should also be noted the significance threshold we used can be subjective (see Discussion), but the relative trends between the methods largely remained similar (Table 1).

Application to TOPMed Deep WGS data

We evaluated the various HWE tests on a subset of the Freeze 5 variant calls from high-coverage (~37×) whole genome sequence (WGS) data in the TOPMed Project (Taliun *et al.* 2019). We identified 17,524 HQ variants and 329,699 LQ variants using the same criteria used

for 1000G variants and evaluated raw GTs and GLs. We did not evaluate PCAngsd due to excessive computational time (see "Computational cost" below).

We first evaluated the false positive rates of different HWE tests indirectly by using HQ variants. With a >20-fold larger sample size than 1000G, we identified more significant HQ variants, while the false positive rates were still reasonable with adjusted tests. At $p < 10^{-6}$, 74% of HQ variants were significant with unadjusted tests, while the adjusted GL-based tests identified ~0.3% at $p < 10^{-6}$ (Figure 4C-D, Table 2). Adjusted GT-based tests had only slightly higher levels of false positives at $p < 10^{-6}$. However, inflation was more noticeable at less stringent p-value thresholds, suggesting that GL-based tests may be needed for larger sample sizes.

Next, we evaluated the proportions of LQ variants found to be significant by different tests to indirectly evaluate their statistical power. GT- and GL-based RUTH tests showed similar power, while meta-analysis showed considerably lower power. For example, at $p < 10^{-6}$, meta-analysis identified 47% of LQ variants as significant, while RUTH tests identified 54-58%. This pattern was similar across different p-value thresholds (Figure 4C-D) or choices of LQ variants (Table S3, Figure S7). Our results suggest that GL-based RUTH tests are suitable for testing HWE for tens of thousands of deeply sequenced genomes with diverse ancestries, and that using raw GTs will also result in a comparable performance at typically used HWE p-value thresholds (e.g. $p < 10^{-6}$).

We used PRCs to evaluate the tradeoff between empirical false positive rates and power. Consistent with previous results, the GL-based RUTH test showed the best tradeoff between false positives and power, while the GT-based RUTH test and meta-analysis were

slightly less robust but largely comparable (Figure S6). Notably, when we evaluated the different methods at an empirical false positive rate of 0.1%, RUTH score tests had ~4% higher power than RUTH LRT for both raw GTs and GLs (Figure S8-9).

Impact of ancestry estimation accuracy on HWE tests

So far, our evaluations relied on genetic ancestry estimates carefully determined with sophisticated methods (see Materials and Methods). However, using simpler approaches instead during the variant QC step may affect the performance of adjusted HWE tests. We evaluated whether the number of PC coordinates affected the performance of RUTH tests by comparing the use of 2 vs. 4 PCs (default). The results from both simulated and real datasets consistently demonstrated that using 4 PCs led to substantially reduced Type I errors compared to using 2 PCs at a similar level of power (Table S2, Table S4, Figure S10). PRCs also clearly showed that using 4 PCs was more robust against population structure across both simulated and real datasets (Figure S11).

We also evaluated whether the classification accuracy of subpopulations affected the performance of meta-analysis. Instead of assigning 1000 Genomes individuals into five continental populations, we used the k-means algorithm on those samples' top 2 PCs to classify them into 3 crude subpopulations (Figure S12). This led to a much higher false positive rate with virtually no increase in true positives (Figure S13, Table S2). We saw the same pattern in simulated data (Figure S11, Table S5).

Computational cost

We compared the computational costs of RUTH and PCAngsd for simulated and real data. RUTH has linear time complexity to sample size, while PCAngsd appears to have quadratic time complexity due to joint estimation of PCs (Tables 3, S6). RUTH also has low memory requirements compared to PCAngsd (for example, 14 MB vs 2 GB for 1000G data). Extrapolating our results to the whole genome scale, analyzing 1000G (i.e. 80 million variants) is expected to take 120 CPU-hours for RUTH and 3,200 CPU-hours for PCAngsd (with >1 TB memory consumption). Additionally, RUTH can be parallelized into smaller regions in a straightforward manner.

DISCUSSION

RUTH is a unified, flexible, and robust approach to incorporate genetic ancestry and genotype uncertainty for testing Hardy-Weinberg Equilibrium capable of handling large amounts of genotype data with structured populations. Sha and Zhang (2011) proposed HWES, an HWE test for structured populations, to address some of these challenges, but it has not been widely used due to the lack of an implementation that supports popular genotype data formats (e.g. PED, BED, VCF, or BCF) and inability to handle imputed or uncertain genotypes. Hao and colleagues (2016) proposed sHWE which can only handle best-guess (hard call) genotypes (i.e. 0, 1, or 2 for biallelic variants) and does not account for genotype uncertainty. MEISNER AND ALBRECHTSEN (2019) proposed PCAngsd to address some of these issues, but it does not support the standard VCF/BCF formats for sequence-based genotypes, and its current implementation scales poorly with genome-wide analyses of large samples.

Similar to previous studies (SHA AND ZHANG 2011; HAO *et al.* 2016), our proposed framework uses individual-specific allele frequencies rather than allele frequencies pooled

across all samples to systematically account for population structure in HWE tests. Unlike those previous studies, we model genotype uncertainty in sequence-based genotypes using a likelihood-based framework. We implemented two RUTH tests – a score test and a likelihood ratio test (LRT) – to test for HWE under population structure for genotypes with uncertainty. While RUTH LRT is similar to the independently developed PCAngsd, the software implementation of RUTH is more flexible, scales much better to large studies, and supports the standard VCF format.

We provide a comprehensive evaluation of various approaches for testing HWE using simulated and real data. Our results demonstrate that modeling population stratification is necessary for HWE tests on heterogeneous populations. We showed that accounting for genotype uncertainty via genotype likelihoods performs substantially better than using best-guess genotypes, especially for low-coverage sequenced genomes. Importantly, we included evaluations for an unpublished but commonly used approach – meta-analysis across stratified subpopulations, cohorts, or batches. Our results demonstrate that while meta-analysis may be effective in reducing false positives, it does so at the expense of substantially reduced power compared to RUTH.

We observed that the current implementation of PCAngsd does not scale well to large-scale sequencing data, though in principle it can be implemented more efficiently, because the underlying HWE test itself is similar to RUTH LRT. PCAngsd requires loading all genotypes into memory, which is often infeasible for large sequencing studies. For example, loading all of 1000 Genomes will require ~4.8 TB of memory. In our evaluation of 1000G chromosome 20 variants, the inability of PCAngsd to estimate PCs from the whole genome may have contributed to the observed difference in results from RUTH compared to our simulation studies. Moreover,

PCAngsd does not offer an option to externally provide PCs or exclude false positive variants when calculating PCs, so it performs poorly when false positive variants confound PC estimation as demonstrated in the 1000 Genomes examples.

Although our 1000G experiments demonstrated the unexpected result that using raw GTs had better sensitivity than using LD-aware GTs at the same empirical false positive rates for low-coverage data, we do not advocate using raw GTs for low-coverage sequence data. First, the results for raw GTs were still consistently less robust than GL-based RUTH tests. Moreover, it would be tricky to determine an appropriate p-value threshold when false positives are severely inflated. Therefore, we strongly advocate using GL-based RUTH tests for robust HWE tests with low-coverage sequence data. For the now more typical high-coverage sequence data, GL-based tests are still preferred, but GT-based RUTH tests should be acceptable for cases in which genotype likelihoods are unavailable.

Our experiment compared using 2 vs 4 PCs only because the *verifyBamID2* software tool estimated up to 4 PCs projected onto the HGDP panel by default (ZHANG *et al.* 2020). Because our method focuses on testing HWE during the QC steps in sequence-based variant calls, a curated version of PCs, estimated from the sequenced cohort themselves, may not be readily available. However, it is possible to use a larger number of PCs (e.g. >10 PCs) if available at the time of HWE test. We expect that a larger number of PCs will account for finer-grained population structure and may improve the performance of HWE tests, but additional experiments are needed to quantify the effect.

Our results demonstrate that RUTH score and LRT tests perform similarly in simulated and experimental datasets. Overall, the RUTH-LRT was slightly more powerful than the RUTH-score test at the expense of slightly greater false positive rates, although this tendency was not

consistent. We observed that the RUTH tests tended to be slightly more powerful in identifying deviation from HWE in the direction of excess heterozygosity than excess homozygosity when compared to adjusted meta-analysis. These results might be caused by the difference between our model-based asymptotic tests compared to the exact test used in meta-analysis.

We did not evaluate our methods on imputed genotypes in this manuscript. Because imputed genotypes implicitly assume HWE, we suspect that HWE tests based on imputed genotypes may have reduced power compared to directly genotyped variants. It is possible to use approximate genotype likelihoods instead of best-guess genotypes for imputed genotypes, but this requires genotype probabilities, not just genotype dosages. If genotype probabilities $\Pr(g_i = G|Data_i)$ are available, they can be converted to genotype likelihoods $L_i^{(G)} = \Pr(Data_i|g_i = G)$ using Bayes' rule by modeling $\Pr(g_i = G)$ as a binomial distribution based on allele frequencies (which implicitly assumes HWE). However, similar to LD-aware genotypes in low-coverage sequencing, the power of HWE tests with imputed genotypes may be poor. Further evaluation is needed to understand the effect of using imputed genotypes on the behavior of HWE tests.

As described in our results, we observed that the current implementations of RUTH (and PCAngsd) tests relying on asymptotic distributions do not work more robustly than the exact test when testing for excess homozygosity ($\theta > 0$). This is mainly because the empirical null distribution becomes increasingly asymmetric between the two directions of effects for rarer variants, but the asymptotic approximation assumes symmetry between them, causing loss of power for excess homozygosity. Using RUTH score test will further reduce power because score tests are known to have reduced power than LRT when θ strongly deviates from zero, which

happens in rare variants with excess homozygosity. Applying Saddlepoint approximation (Dey et al. 2017) or similar techniques may help address this issue.

In practice, when we examined low quality (LQ) variants, determined by high Mendelian errors, the vast majority (65% for 1000G, 82% for TOPMed) of them deviated from HWE towards excess heterozygosity ($\theta < 0$) as opposed to excess homozygosity ($\theta > 0$) when we examined the direction of deviation from HWE regardless of its significance. On the other hand, the majority of high quality (HQ) variants (77% for 1000G, 64% for TOPMed) mildly deviated from HWE towards excess homozygosity ($\theta > 0$), presumably due to residual population structure and cryptic relatedness. These observations suggest that detecting excess heterozygosity is practically more important for variant QC, on which RUTH tests are expect to perform well.

Our methods have room for further improvement. First, we used a truncated linear model for individual-specific allele frequencies for computational efficiency. Although such an approximation was demonstrated to be effective in practice (ZHANG *et al.* 2020), applying a logistic model or some other more sophisticated model may be more effective in improving the precision and recall of RUTH tests. Second, we did not attempt to model or evaluate the effect of admixture in our method. Because HWE is reached in two generations with random mating, accounting for admixed individuals may only have a marginal impact. On the other hand, admixture can lead to higher observed heterozygosity. It may be possible to improve RUTH by explicitly modeling and adjusting for the effect of admixture on individual-specific allele frequencies. Systematic evaluations focusing on admixed populations are needed to evaluate whether an admixture adjustment is necessary. Third, RUTH tests do not account for family structure or individual-level inbreeding. We suspect that the apparent inflation of Type I error

for the TOPMed data was partially due to sample relatedness. Accounting for family structure or individual-level inbreeding in other ways, for example using variance components models, will require much longer computational times and may not be feasible for large-scale datasets. Fourth, RUTH currently does not directly support imputed genotypes or genotype dosages. In principle, it is possible to convert posterior probabilities for imputed genotypes into genotype likelihoods to account for genotype uncertainty (by using individual-specific allele frequencies). However, because most genotype imputation methods implicitly assume HWE, we suspect that HWE tests on imputed genotypes will be underpowered, similar to our observations with LD-aware genotypes in the 1000 Genomes dataset, even though explicitly modeling posterior probabilities may slightly mitigate this reduction in power.

The choice of a p-value threshold to indicate deviation from HWE remains an open question. In previous studies, stringent p-value thresholds were used to prevent high-quality variants from being filtered due to population structure. Adjusting for population structure with RUTH helps mitigate this problem, allowing the use of less stringent thresholds to improve test performance, but the choice of p-value threshold remains subjective, based on the tradeoff between sensitivity and specificity. Future development of more robust methods to determine significance thresholds would help further improve the use of HWE tests for variant quality control.

In summary, we have developed and implemented robust and rapid methods and software tools to enable HWE tests that account for population structure and genotype uncertainty. We comprehensively evaluated both our methods and alternative approaches. Our tools can be used to evaluate variant quality in very large-scale genetic data sets, with the

ability to handle standard VCF formats for storing sequence-based genotypes. Our software tools are publicly available at http://github.com/statgen/ruth.

Software and data availability

RUTH is available at https://github.com/statgen/ruth. Genotype data from 1000G is available from the International Genome Sample Resource at https://www.internationalgenome.org. TOPMed data is available via a dbGaP application for controlled-access data (see https://www.nhlbiwgs.org for details). Supplementary materials have been uploaded to figshare: https://doi.org/10.25386/genetics.14068970.

Acknowledgements

TOPMed source studies and sample counts are described in Table S7. Acknowledgements for TOPMed omics support are detailed in Table S8. Full TOPMed study acknowledgements are listed in Supplementary File S1.

Funding

This work was supported by NIH grants HL137182 (from NHLBI), HG009976 (from NHGRI), HG007022 (from NHGRI), DA037904 (from NIDA), HL117626-05-S2 (from NHLBI), and MH105653 (from NIMH). Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general

program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Competing Interests

K.C.B. receives funds from the NIH and receives royalties from UpToDate. E.G.B. has received funds from the following: National Institute of Health, Lung, Blood Institute, National Institute of Health, General Medical Sciences, National Institute on Minority Health and Health Disparities, The Tobacco-Related Disease Research Program, Food and Drug Administration, and The Sandler Family Foundation. L.A.C. spends part of her time consulting for Dyslipidemia Foundation, a non-profit company, as a statistical consultant. P.T.E. is supported by a grant from Bayer to the Broad Institute focused on the genetics and therapeutics of cardiovascular diseases. P.T.E. has also served on advisory boards or consulted for Quest Diagnostics and Novartis. S.A.L. receives sponsored research support from Bristol Myers Squibb/Pfizer, Bayer, Boehringer Ingelheim and Fitbit, has consulted for Bristol Myers Squibb/Pfizer and Bayer, and participates in a research collaboration with IBM. M.E.M. is an inventor on a patent that was published by the United States Patent and Trademark Office on 6 December 2018 under Publication Number US 2018-0346888, and an international patent application that was published on 13 December 2018 under Publication Number WO-2018/226560 regarding B4GALT1 Variants and uses thereof. S.T.W. receives royalties from UpToDate. G.R.A. and H.M.K. are employees of Regeneron Pharmaceuticals, they own stock and stock options for Regeneron Pharmaceuticals.

REFERENCES

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol 63: 221-230.

Balding, D. J., and R. A. Nichols, 1995 A Method for Quantifying Differentiation between Populations at Multi-Allelic Loci and Its Implications for Investigating Identity and Paternity. Genetica 96: 3-12.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott *et al.*, 2018 The UK Biobank resource with deep phenotyping and genomic data. Nature 562: 203-209.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

Delaneau, O., J. F. Zagury and J. Marchini, 2013 Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10: 5-6.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39: 1-22.

Dryden, I. L., and K. V. Mardia, 1998 *Statistical shape analysis*. John Wiley & Sons, Chichester ; New York.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194.

Fritsche, L. G., W. Igl, J. N. Bailey, F. Grassmann, S. Sengupta *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.

Graffelman, J., and V. Moreno, 2013 The mid p-value in exact tests for Hardy-Weinberg equilibrium. Stat Appl Genet Mol Biol 12: 433-448.

Hao, W., M. Song and J. D. Storey, 2016 Probabilistic models of genetic variation in structured populations applied to global human studies. Bioinformatics 32: 713-721.

Hao, W., and J. D. Storey, 2019 Extending Tests of Hardy-Weinberg Equilibrium to Structured Populations. Genetics 213: 759-770.

Hardy, G. H., 1908 Mendelian Proportions in a Mixed Population. Science 28: 49-50.

Holsinger, K. E., 1999 Analysis of Genetic Diversity in Geographically Structured Populations: A Bayesian Perspective. Hereditas 130: 245-255.

Holsinger, K. E., P. O. Lewis and D. K. Dey, 2002 A Bayesian approach to inferring population structure from dominant markers. Mol Ecol 11: 1157-1164.

Jin, Y., A. A. Schaffer, M. Feolo, J. B. Holmes and B. L. Kattman, 2019 GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis. G3 (Bethesda) 9: 2447-2461.

Jun, G., M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny *et al.*, 2012 Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91: 839-848.

Kuhn, R. M., D. Haussler and W. J. Kent, 2013 The UCSC genome browser and associated tools. Brief Bioinform 14**:** 144-161.

Laurie, C. C., K. F. Doheny, D. B. Mirel, E. W. Pugh, L. J. Bierut *et al.*, 2010 Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34**:** 591-602.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 319**:** 1100-1104.

Li, M., and C. Li, 2008 Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. Genet Epidemiol 32**:** 589-599.

Locke, A. E., B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers *et al.*, 2015 Genetic studies of body mass index yield new insights for obesity biology. Nature 518**:** 197-206.

McCarroll, S. A., T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody *et al.*, 2006 Common deletion polymorphisms in the human genome. Nat Genet 38**:** 86-92.

Meisner, J., and A. Albrechtsen, 2019 Testing for Hardy-Weinberg Equilibrium in Structured Populations using Genotype or Low-Depth NGS Data. Mol Ecol Resour.

Mosteller, F., and R. A. Fisher, 1948 Questions and Answers. The American Statistician 2**:** 30-31.

Nielsen, D. M., M. G. Ehm and B. S. Weir, 1998 Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 63**:** 1531-1540.

Nielsen, R., J. S. Paul, A. Albrechtsen and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12**:** 443-451.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38**:** 904-909.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81**:** 559-575.

Rohlfs, R. V., and B. S. Weir, 2008 Distributions of Hardy-Weinberg equilibrium test statistics. Genetics 180**:** 1609-1616.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science 298**:** 2381-2385.

Sha, Q., and S. Zhang, 2011 A test of Hardy-Weinberg equilibrium in structured populations. Genet Epidemiol 35**:** 671-678.

Stouffer, S. A., 1949 *The American soldier*. Princeton University Press, Princeton,.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star and R. M. Williams Jr, 1949 The American soldier: Adjustment during army life.(Studies in social psychology in World War II), Vol. 1.

Taliun, D., D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech *et al.*, 2019 Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv.

The 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526**:** 68-74.

The International HapMap Consortium, D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler *et al.*, 2010 Integrating common and rare genetic variation in diverse human populations. Nature 467**:** 52-58.

Van Oosterhout, C., W. F. Hutchinson, D. P. M. Wills and P. Shipley, 2004 MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4**:** 535-538.

Wang, C., Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton *et al.*, 2010 Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat Appl Genet Mol Biol 9**:** Article 13.

Waples, R. S., 2015 Testing for Hardy-Weinberg proportions: have we lost the plot? J Hered 106**:** 1-19.

Weinberg, W., 1908 Uber den nachweis der vererbung beim menschen. Jh. Ver. vaterl. Naturk. Wurttemb. 64**:** 369-382.

Wigginton, J. E., D. J. Cutler and G. R. Abecasis, 2005 A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76**:** 887-893.

Yang, W. Y., J. Novembre, E. Eskin and E. Halperin, 2012 A model-based approach for analysis of spatial structure in genetic data. Nat Genet 44**:** 725-731.

Zhang, F., M. Flickinger, S. A. G. Taliun, P. P. G. C. In, G. R. Abecasis *et al.*, 2020 Ancestry-agnostic estimation of DNA sample contamination from sequence reads. Genome Res 30**:** 185-194.

Figure 1 Evaluation of Type I Errors between various HWE tests on simulated genotypes.
Under each combination of simulation conditions (number of ancestries, sequencing coverage, and fixation index), we simulated 5,000 samples with 50,000 variants that follow HWE within each of the subpopulations and determined the Type I error performances of different HWE tests based on the proportion of variants labeled as having significant p-values. Five HWE tests – (1) Unadjusted HWE test (WIGGINTON *et al.* 2005) implemented in PLINK-1.9 (PURCELL *et al.* 2007) using hard genotypes, (2) meta-analysis using Stouffer's method across ancestries using hard genotypes (GT), (3) RUTH test using hard genotypes, (4) RUTH test using phred-scale likelihood (GL) computed from simulated sequence reads, and (5) PCAngsd (MEISNER AND ALBRECHTSEN 2019) – were tested under HWE with various parameter settings. Gray dotted lines indicate targeted Type I Error rates. Top panels (A-C) represent results from shallow sequencing (5x), and the bottom panels (D-F) represent results from deep sequencing (30x). Using GL-based genotypes resulted in Type I Error rates closer to the targeted rate than using GT-based genotypes across different numbers of ancestries (A, D), P-value thresholds (B, E), and fixation indices (C, F). The difference is especially large for low-coverage genotypes.

Figure 2 Evaluation of power between different HWE tests on simulated genotypes.
Under each combination of simulation conditions (number of ancestries, sequencing coverage, fixation index, and deviation from HWE), we simulated 50,000 variants for 5,000 samples and evaluated the ability of different HWE tests to find the variants significant. Unless otherwise specified, the default simulation parameters are 5 ancestries, with F_{ST}=.1, P-value threshold=.001, and Theta=-0.05. Tests that can find a larger proportion of significant variants are considered more powerful. Five HWE tests – (1) Unadjusted HWE test (WIGGINTON *et al.* 2005) implemented in PLINK-1.9 using hard genotypes (2) RUTH test using hard genotypes, (3) RUTH test using phred-scale likelihood (PL) computed from simulated sequence reads, (4) meta-analysis using Stouffer's method across ancestries using hard genotypes, and (5) PCAngsd (MEISNER AND ALBRECHTSEN 2019) – were tested for variants deviating from HWE with various parameter settings, for low coverage (A-D) and high coverage (E-H) data. (A, E) Theta controls the degree of deviation from HWE, with negative values indicating excess heterozygosity and positive values indicating heterozygote depletion. The high Type I Error rates in GT-based tests (Figure 2) lead to those methods appearing to have higher power in some scenarios. The unadjusted test suffers from this problem the most. GL-based methods have slightly lower powers than GT-based methods in exchange for a much better controlled Type I error rate. This pattern mostly holds across different numbers of ancestries (B, F), p-value thresholds (C, G), and fixation indices (D, H). Meta-analysis had the lowest power in the presence of excess heterozygosity.

Figure 3 Schematic diagrams of different methods to test HWE under population structure.
Three different methods to test HWE under population structure are described. (A) In the standard (unadjusted) HWE test, all samples are tested together using best-guess genotypes. This test does not adjust for sample ancestry. (B) In a meta-analysis of stratified HWE tests, the samples must first be categorized into discrete subpopulations, determined a priori based on their genotypes or self-reported ancestries. Next, standard HWE tests (based on best-guess genotypes) are performed on each of these subpopulations. Then, the resulting HWE statistics are converted into Z-scores and combined in a meta-analysis using Stouffer's method, with the sample sizes of the subpopulations as weights. (C) In our proposed method (RUTH), either best-guess genotypes or genotype likelihoods can be used as input for HWE test. We assume that the genetic ancestries of each sample are estimated a priori, typically as principal components (PCs). We combine the genotypes and PCs to perform either a score test or a likelihood ratio test to obtain a joint ancestry-adjusted HWE statistic for each variant across all samples.

Figure 4 Evaluation of different HWE tests on 1000 Genomes and TOPMed variants.
In 1000 Genomes data (A, B), we identified 17,740 "high quality" (HQ) variants and 10,966 "low quality" (LQ) variants in chromosome 20. In TOPMed data (C, D), we identified 17,524 HQ variants and 329,699 LQ variants in chromosome 20. A well-behaved HWE test should maximize the proportion of significant LQ variants while controlling the false positive rate for HQ variants. Dotted gray lines represent targeted Type I error levels if we assume all HQ variants follow HWE. (A) Both the unadjusted test and PCAngsd found substantially more significant variants than expected in the 1000G HQ variant set, while both RUTH and meta-analysis were more conservative. Methods that used raw GTs showed substantial false positive rates, while methods that used GLs and LD-aware GTs had much better control of false positives. (B) In 1000G LQ variants, meta-analysis lagged behind RUTH and the unadjusted test in discovering significant deviation from HWE. RUTH behaved well for HQ variants while having

more power to find low-quality variants significantly deviating from HWE. (C) In TOPMed data, the unadjusted test resulted in an excess of false positives. Tests using GL-based genotypes outperformed tests using GT-based genotypes. (D) Methods using GL-based genotypes were able to discover more LQ variants than methods using GT-based genotypes, demonstrating the advantage of accounting for genotype uncertainty in HWE tests.

Table 1

Performance of the unadjusted test, meta-analysis, RUTH, and PCAngsd on 1000 Genomes chromosome 20 variants.

Variant Category	Genotype Format	HWE Test	Proportion of Significant Variants					Total Variant Count
			$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	
LQ Variants	raw GT	Unadjusted	0.487	0.432	0.394	0.366	0.339	10,966
		Meta-analysis	0.392	0.343	0.307	0.283	0.262	10,966
		RUTH-Score	0.418	0.367	0.333	0.305	0.284	10,966
		RUTH-LRT	0.431	0.373	0.335	0.305	0.280	10,966
	LD-aware GT	Unadjusted	0.479	0.395	0.336	0.292	0.259	10,966
		Meta-analysis	0.184	0.149	0.127	0.111	0.098	10,966
		RUTH-Score	0.211	0.172	0.147	0.130	0.112	10,966
		RUTH-LRT	0.215	0.177	0.151	0.131	0.115	10,966
	GL	RUTH-Score	0.336	0.295	0.264	0.242	0.223	10,966
		RUTH-LRT	0.358	0.306	0.270	0.243	0.225	10,966
		PCAngsd	0.380	0.331	0.300	0.275	0.255	10,920
HQ Variants	raw GT	Unadjusted	0.755	0.657	0.573	0.501	0.443	17,740
		Meta-analysis	0.298	0.161	0.084	0.042	0.020	17,740
		RUTH-Score	0.183	0.083	0.036	0.015	7.4×10^{-3}	17,740
		RUTH-LRT	0.200	0.095	0.044	0.021	0.010	17,740
	LD-aware GT	Unadjusted	0.623	0.507	0.422	0.361	0.311	17,740
		Meta-analysis	0.019	3.1×10^{-3}	5.6×10^{-4}	1.7×10^{-4}	1.1×10^{-4}	17,740
		RUTH-Score	0.011	1.9×10^{-3}	1.1×10^{-4}	0	0	17,740
		RUTH-LRT	0.011	1.1×10^{-3}	2.3×10^{-4}	5.6×10^{-5}	0	17,740
	GL	RUTH-Score	0.026	3.3×10^{-3}	7.9×10^{-4}	4.5×10^{-4}	3.4×10^{-4}	17,740
		RUTH-LRT	0.036	6.4×10^{-3}	1.3×10^{-3}	5.1×10^{-4}	3.4×10^{-4}	17,740
		PCAngsd	0.059	0.032	0.026	0.022	0.021	17,740

The numbers within cells represent the proportions of significant variants under the corresponding testing conditions at the given P-value threshold. We expect our LQ variants to violate HWE at a higher rate than our HQ variants. A well-behaved test is expected to find a high proportion of LQ variants to be significant while maintaining the targeted Type I Error rate in HQ variants. The unadjusted test consistently shows the highest false positive rate among all the tests. HWE tests that rely on raw GTs also show much higher false positive rates than tests that use other genotype representations. RUTH tests were the best at controlling false positives while still maintaining comparable power to the other methods. PCAngsd had a much higher false positive rate than RUTH-based methods, especially at more stringent p-value thresholds.

Table 2

Performance of the unadjusted test, meta-analysis, and RUTH on TOPMed freeze 5 chromosome 20 variants.

Variant set	Genotype Format	HWE Test	Proportion of Significant Variants					Total Variant Count
			$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	
LQ Variants	raw GT	Unadjusted	0.592	0.561	0.539	0.521	0.506	329,699
	raw GT	Meta-analysis	0.554	0.524	0.502	0.485	0.471	329,699
	raw GT	RUTH-Score	0.608	0.587	0.572	0.559	0.549	329,699
	GL	RUTH-Score	0.635	0.608	0.590	0.575	0.563	329,699
	raw GT	RUTH-LRT	0.610	0.580	0.556	0.538	0.522	329,699
	GL	RUTH-LRT	0.653	0.615	0.588	0.567	0.550	329,699
HQ Variants	raw GT	Unadjusted	0.890	0.842	0.800	0.766	0.736	17,524
	raw GT	Meta-analysis	0.065	0.022	9.0×10^{-3}	4.8×10^{-3}	3.3×10^{-3}	17,524
	raw GT	RUTH-Score	0.145	0.047	0.172	7.1×10^{-3}	3.5×10^{-3}	17,524
	GL	RUTH-Score	0.034	0.011	4.9×10^{-3}	3.1×10^{-3}	2.5×10^{-3}	17,524
	raw GT	RUTH-LRT	0.125	0.036	0.012	5.0×10^{-3}	2.7×10^{-3}	17,524
	GL	RUTH-LRT	0.041	0.018	8.5×10^{-3}	4.3×10^{-3}	3.1×10^{-3}	17,524

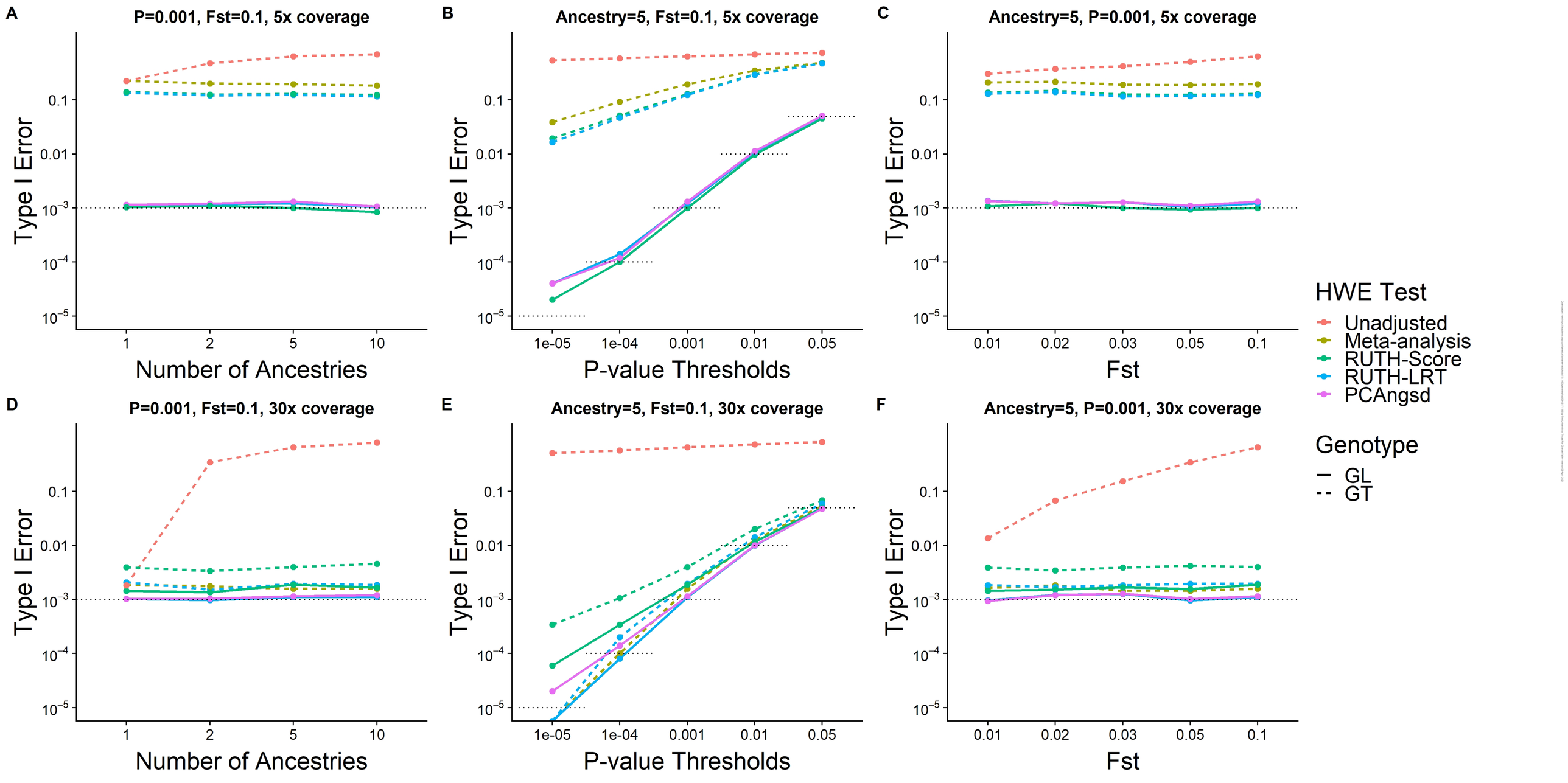
The numbers within cells represent the proportions of significant variants under the corresponding testing conditions at the given P-value threshold. These results are based on tests that used likelihood-based genotype representations as input. A well-behaved test should reduce the number of significant high-quality (HQ) variants while increasing the number of significant low-quality (LQ) variants. The unadjusted test had a greatly inflated false positive rate for HQ variants while showing a lower true positive rate for LQ variants. While meta-analysis performed better for HQ variants, it had reduced power to find LQ variants to be significant. RUTH performed the best, with fewer false positives (significant HQ variants) compared to both the unadjusted test and meta-analysis, while at the same time finding more true positives (significant LQ variants).

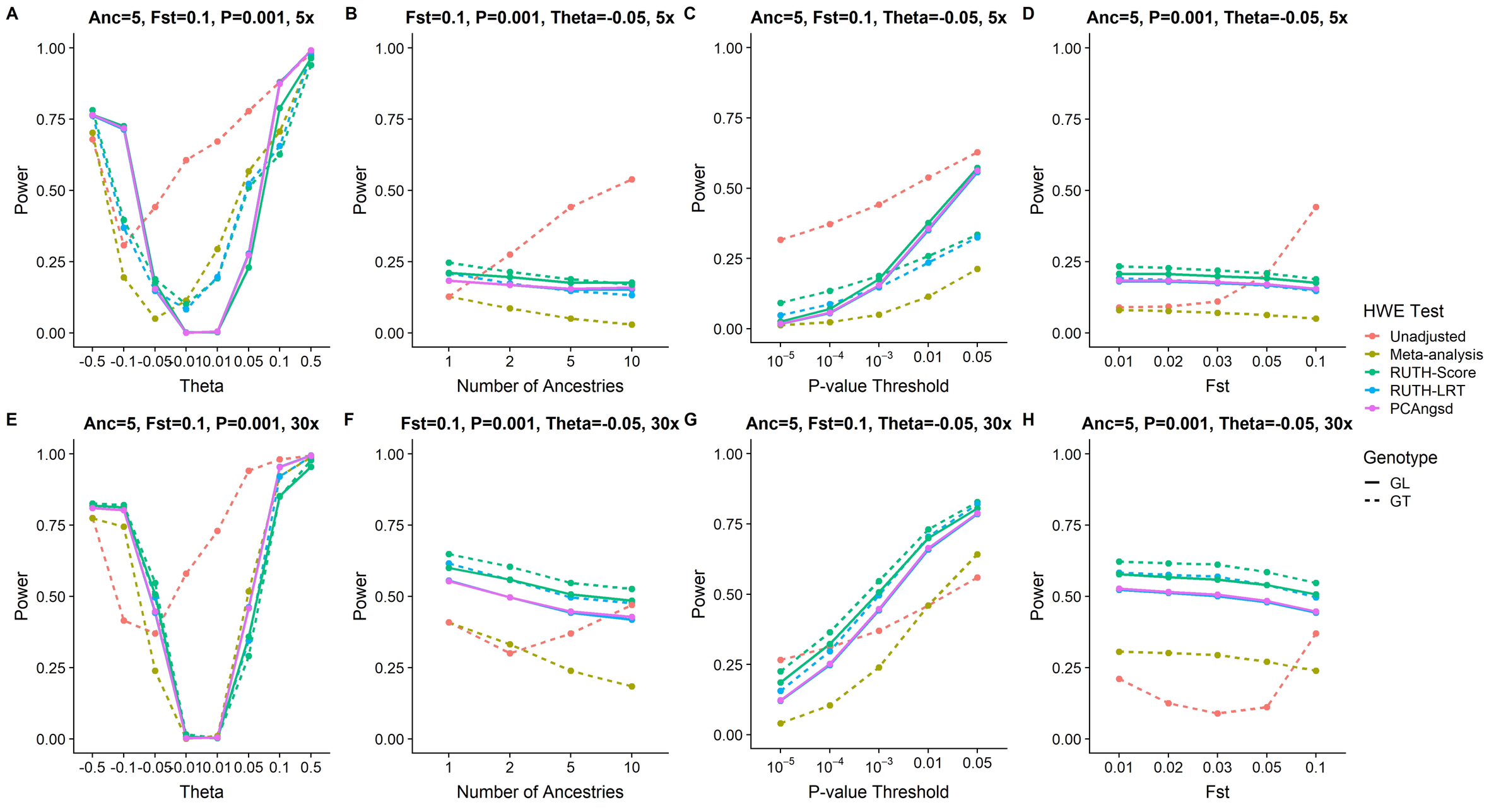
Table 3

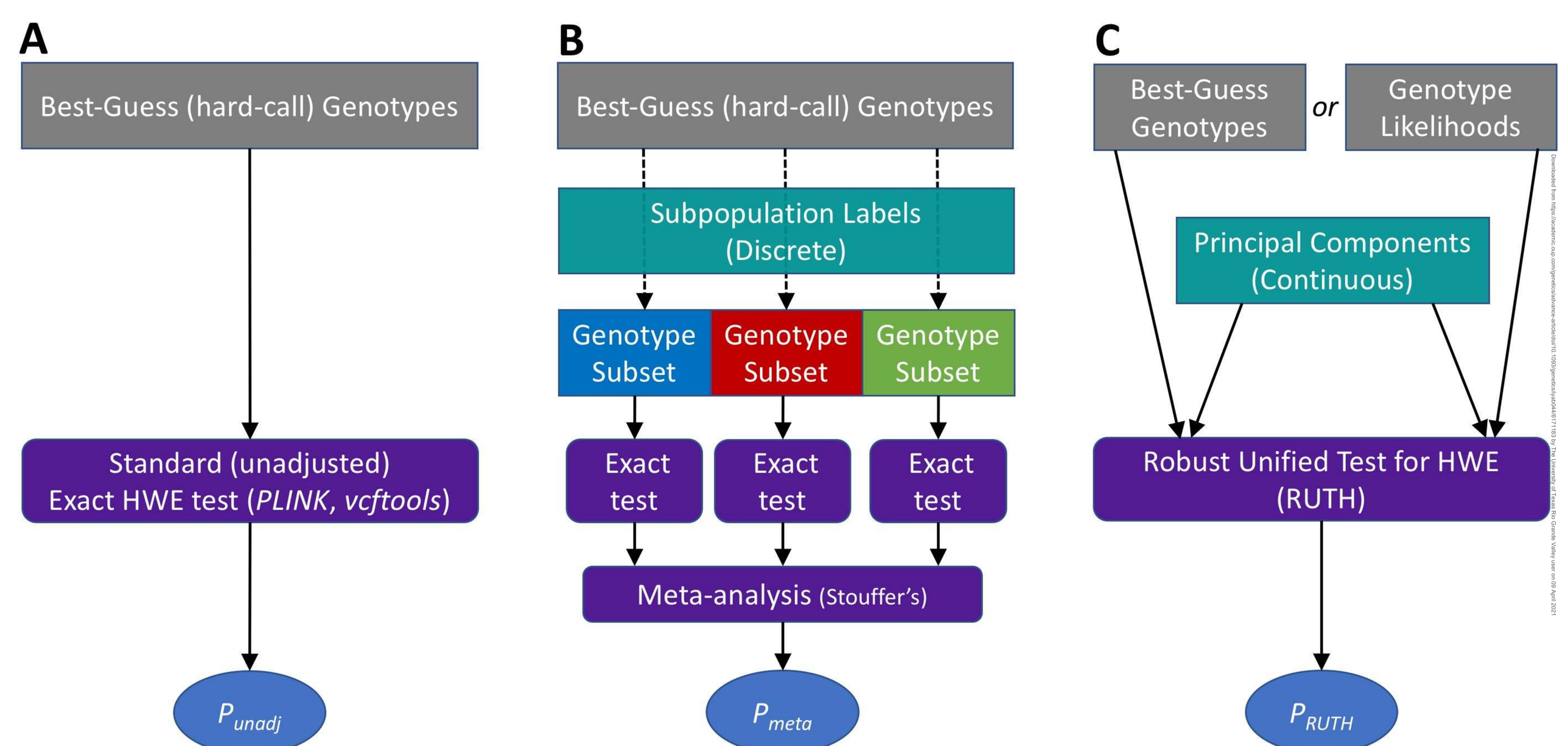
Runtimes for RUTH and PCAngsd on simulated data.

Sample Size	Wall Time (s)			User Time (s)		
	RUTH-LRT	RUTH-Score	PCAngsd	RUTH-LRT	RUTH-Score	PCAngsd
1,000	16.21	27.24	173.11	16.16	27.09	172.37
2,000	32.19	54.63	347.10	31.94	54.51	345.58
5,000	82.80	136.44	1,124.83	81.81	136.20	1,102.85
10,000	165.48	273.67	7,396.00	163.88	273.27	7,235.91
20,000	336.75	553.92	38,807.67	332.06	553.05	37,338.69
50,000	902.81	1,438.32	461,971.33	886.67	1,435.87	403,296.5

We simulated 10,000 genotype likelihood-based variants for varying numbers of samples. Wall time indicates total runtime, while user time is the amount of time the CPUs spent running each program. All programs were run in single-threaded mode. System processes make up the difference between the two values, with a majority consisting of file I/O. We used VCF files with GL fields in RUTH and converted them to Beagle3 format for PCAngsd. The RUTH likelihood ratio test (LRT) was the fastest method, with the score test about 60% slower. PCAngsd was about 10 times slower than RUTH-LRT with the smallest sample sizes and over 400 times slower with our largest tested size of 50,000 samples.


A
P=0.001, Fst=0.1, 5x coverage
B
Ancestry=5, Fst=0.1, 5x coverage
C
Ancestry=5, P=0.001, 5x coverage
D
P=0.001, Fst=0.1, 30x coverage
E
Ancestry=5, Fst=0.1, 30x coverage
F
Ancestry=5, P=0.001, 30x coverage
Type I Error
0.1
0.01
10^{-3}
10^{-4}
10^{-5}
1
2
5
10
Number of Ancestries
1e-05
1e-04
0.001
0.01
0.05
P-value Thresholds
0.01
0.02
0.03
0.05
0.1
Fst
HWE Test
Unadjusted
Meta-analysis
RUTH-Score
RUTH-LRT
PCAngsd
Genotype
GL
GT


A
Anc=5, Fst=0.1, P=0.001, 5x
B
Fst=0.1, P=0.001, Theta=-0.05, 5x
C
Anc=5, Fst=0.1, Theta=-0.05, 5x
D
Anc=5, P=0.001, Theta=-0.05, 5x
E
Anc=5, Fst=0.1, P=0.001, 30x
F
Fst=0.1, P=0.001, Theta=-0.05, 30x
G
Anc=5, Fst=0.1, Theta=-0.05, 30x
H
Anc=5, P=0.001, Theta=-0.05, 30x
Power
1.00
0.75
0.50
0.25
0.00
-0.5 -0.1 -0.05 -0.01 0.01 0.05 0.1 0.5
Theta
1 2 5 10
Number of Ancestries
10^{-5} 10^{-4} 10^{-3} 0.01 0.05
P-value Threshold
0.01 0.02 0.03 0.05 0.1
Fst
HWE Test
Unadjusted
Meta-analysis
RUTH-Score
RUTH-LRT
PCAngsd
Genotype
GL
GT



A

1000G Chr20, HQ Variants

Proportion of Significant Variants

0.05
0.01
10^{-3}
10^{-4}
10^{-5}

10^{-5} 10^{-4} 10^{-3} 0.01 0.05

P-value Threshold

B

1000G Chr20, LQ Variants

Proportion of Significant Variants

0.5
0.4
0.3
0.2
0.1

10^{-5} 10^{-4} 10^{-3} 0.01 0.05

P-value Threshold

HWE Test
Unadjusted
Meta-analysis
RUTH-Score
RUTH-LRT
PCAngsd

C

TOPMed Chr20, HQ Variants

Proportion of Significant Variants

0.05
0.01
10^{-3}
10^{-4}
10^{-5}

10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 0.01 0.05

P-value Threshold

D

TOPMed Chr20, LQ Variants

Proportion of Significant Variants

0.70
0.65
0.60
0.55
0.50
0.45

10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 0.01 0.05

P-value Threshold

Genotype
GL
raw GT
LD-aware GT