Supplemental Material

Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations.

Supplemental Tables 2
Supplemental Table S1: Study information and sample sizes used in Figure 1. 2
Supplemental Table S2: Quality control considerations based on ancestry. 3
Supplemental Table S3: New and ongoing initiatives to expand available imputation reference panels. 4
Supplemental Table S4: Listing of select methods and software implementations cited. 5

Supplemental Figures 7
Supplemental Figure S1: Detailed flow chart for QC, imputation, and association analysis in diverse population samples. 7

Supplemental Methods 8
I. Quality Control 8
   I.1 Considerations for quality control in diverse cohorts 8
   I.2 Example workflow for stratified meta-analysis approach 9
   I.3 Example workflow for joint mixed model approach 10
II. Inferring Population Structure 11
III. Imputation and Population Reference Panels 12
   III.1 Genotype imputation methods 12
   III.2 Factors affecting imputation accuracy 13
   III.3 Preparing imputed data for GWAS 13
IV. Genome-wide Association 14
   IV.1 Mixed model implementations 14
   IV.2 Evaluating control of population stratification 15
   IV.3 Principal components in admixed samples 16
V. Meta-analysis of GWAS Summary Statistics 16
VI. Heritability and Genetic Correlation 17
   VI.1 Estimating SNP heritability from individual-level genetic data 18
   VI.2 Estimating SNP heritability from GWAS summary statistics 19
   VI.3 Estimating genetic correlation 19
VII. Phenotypic Measurement 20

References 21
Supplemental Tables

**Supplemental Table S1:** Study information and sample sizes used in Figure 1.

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Case N</th>
<th>Control N</th>
<th>Ancestry</th>
<th>First Author</th>
<th>PMID or Biorxiv*</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDD</td>
<td>135458</td>
<td>344901</td>
<td>European</td>
<td>Wray</td>
<td>29700475</td>
</tr>
<tr>
<td>MDD</td>
<td>5303</td>
<td>5337</td>
<td>East Asian</td>
<td>CONVERGE</td>
<td>26176920</td>
</tr>
<tr>
<td>SCZ</td>
<td>36265</td>
<td>111256</td>
<td>European</td>
<td>Ripke</td>
<td>25056061</td>
</tr>
<tr>
<td>SCZ</td>
<td>22778</td>
<td>35362</td>
<td>East Asian</td>
<td>Lam</td>
<td>2018/10/18/445874</td>
</tr>
<tr>
<td>PTSD</td>
<td>21032</td>
<td>144432</td>
<td>European</td>
<td>Nievergelt</td>
<td>2018/11/01/458562</td>
</tr>
<tr>
<td>PTSD</td>
<td>4510</td>
<td>12607</td>
<td>African American</td>
<td>Nievergelt</td>
<td>2018/11/01/458562</td>
</tr>
<tr>
<td>PTSD</td>
<td>2186</td>
<td>4165</td>
<td>Latino</td>
<td>Nievergelt</td>
<td>2018/11/01/458562</td>
</tr>
<tr>
<td>BIP</td>
<td>20352</td>
<td>31358</td>
<td>European</td>
<td>Stahl</td>
<td>31043756</td>
</tr>
<tr>
<td>ADHD</td>
<td>19099</td>
<td>34194</td>
<td>European</td>
<td>Demontis</td>
<td>30478444</td>
</tr>
<tr>
<td>ADHD</td>
<td>1012</td>
<td>925</td>
<td>East Asian</td>
<td>Demontis</td>
<td>30478444</td>
</tr>
<tr>
<td>AUT</td>
<td>18381</td>
<td>27969</td>
<td>European</td>
<td>Grove</td>
<td>30804558</td>
</tr>
<tr>
<td>AD</td>
<td>11569</td>
<td>34999</td>
<td>European</td>
<td>Walters</td>
<td>30482948</td>
</tr>
<tr>
<td>AD</td>
<td>3335</td>
<td>2945</td>
<td>African American</td>
<td>Walters</td>
<td>30482948</td>
</tr>
<tr>
<td>AN</td>
<td>3495</td>
<td>10982</td>
<td>European</td>
<td>Duncan</td>
<td>28494655</td>
</tr>
</tbody>
</table>

MD=major depression, SCZ=schizophrenia, PTSD=post-traumatic stress disorder, BIP=bipolar disorder, ADHD=attention deficit hyperactivity disorder, AUT=autism, AD=alcohol dependence, AN=anorexia.

*Link includes the following prefix: https://www.biorxiv.org/content/early/
Supplemental Table S2: Quality control considerations based on ancestry.

UPDATED:
https://docs.google.com/document/d/1Q0leWGl77oXGHysl-Lsk8omlncqC-Ow_7n4fju91EY/edit#
### Supplemental Table S3: New and ongoing initiatives to expand available imputation reference panels.

<table>
<thead>
<tr>
<th>Reference Panels</th>
<th>N</th>
<th>Ancestries</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Icelanders</td>
<td>15,220</td>
<td>Icelandic population</td>
<td><a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5607473/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5607473/</a></td>
</tr>
<tr>
<td>African Genome Variation Project (AGVP)</td>
<td>100</td>
<td>African</td>
<td><a href="https://www.sanger.ac.uk/science/collaboration/african-genome-variation-project">https://www.sanger.ac.uk/science/collaboration/african-genome-variation-project</a></td>
</tr>
<tr>
<td>Human Genome Diversity Project (HGDP)</td>
<td>1,043</td>
<td>Africa, Europe, Middle East, South &amp; Central Asia, East Asia, Oceania and the Americas</td>
<td><a href="http://www.hagsc.org/hgdp/">http://www.hagsc.org/hgdp/</a></td>
</tr>
<tr>
<td>23andMe</td>
<td>200,000</td>
<td>African American</td>
<td><a href="http://grantome.com/grant/NIH/R44-HG009460-01">http://grantome.com/grant/NIH/R44-HG009460-01</a></td>
</tr>
<tr>
<td>Western Indian Project</td>
<td>407</td>
<td>Western Indian</td>
<td><a href="https://www.nature.com/articles/s41598-017-06905-6">https://www.nature.com/articles/s41598-017-06905-6</a></td>
</tr>
<tr>
<td>Simons Genomic Diversity Project</td>
<td>300</td>
<td>Africa, America, Central Asia/Siberia, East Asia, Oceania, South Asia, West Eurasia</td>
<td>PMID:27654912; most data public through EBI, some requires a release; <a href="http://reichdata.hms.harvard.edu/pub/datasets/sgdp/">http://reichdata.hms.harvard.edu/pub/datasets/sgdp/</a></td>
</tr>
<tr>
<td>GenomeDenmark</td>
<td>150</td>
<td>Danish population</td>
<td>PMID:28746312</td>
</tr>
<tr>
<td>Korean National Standard Reference Variome (KoVariome)</td>
<td>80</td>
<td>Korean population</td>
<td>PMID:29618732; ftp://biodisk.org/Release/VariomeData/</td>
</tr>
<tr>
<td>Estonian Biocentre Human Genome Diversity Panel (EGDP)</td>
<td>402</td>
<td>Africa, West Asia, Caucasus, Europe, Central Asia, South Asia, East Asia, Siberia, Island Southeast Asia, Sahul, Americas</td>
<td>PMID: 27654910; <a href="http://evolbio.ut.ee/CGgenomes.html">http://evolbio.ut.ee/CGgenomes.html</a></td>
</tr>
<tr>
<td>Genome Aggregation Database (gnomAD)</td>
<td>15,708</td>
<td>African/African American, Latino, Ashkenski Jewish, East Asian, Finnish, Non-Finnish European, Other</td>
<td><a href="https://www.biorxiv.org/content/10.1101/531210v2">https://www.biorxiv.org/content/10.1101/531210v2</a>; <a href="https://gnomad.broadinstitute.org">https://gnomad.broadinstitute.org</a></td>
</tr>
</tbody>
</table>
### Supplemental Table S4: Listing of select methods and software implementations cited.

<table>
<thead>
<tr>
<th>Software</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR-MEGA</td>
<td>Mägi, R. et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Human Molecular</td>
</tr>
<tr>
<td>Method</td>
<td>Reference</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
Supplemental Figures

Supplemental Figure S1: Detailed flow chart for QC, imputation, and association analysis in diverse population samples.

This flowchart depicts the general analysis framework for genome-wide association studies of participants with diverse ancestral backgrounds. Note: boxes with red headers indicate analyses done in samples with diverse ancestral backgrounds and blue denotes analysis done within samples in major population groups. The left panel shows a strategy for the stratified meta-analysis approach and the right panel shows steps for the joint mixed model approach (see Supplemental Table 2 for more detailed QC considerations).
Supplemental Methods

I. Quality Control

I.1 Considerations for quality control in diverse cohorts

Genome-wide association studies (GWAS) rely on rigorous quality control (QC) of genotype data prior to analysis to help ensure that technical artifacts do not lead to false positive results in the analysis. Several guides to quality control of genotype data exist in the literature (Anderson et al., 2010; de Bakker et al., 2008; Laurie et al., 2010; Turner et al., 2011). Large modern GWAS also routinely report their QC criteria (Bycroft et al., 2018; Walters et al., 2018; Wojcik et al., 2019). Briefly, QC criteria generally focus on measures of missingness and heterozygosity that may reflect technical confounds, as well as identifying cryptically related individuals or low allele frequency variants that may lead to statistical artifacts in association testing. Studies with multiple genotyping batches or using multiple genotyping arrays in the same sample may also check for batch effects. We focus here on how standard QC criteria, such as those implemented within the Psychiatric Genomic Consortium using Ricopili (https://sites.google.com/a/broadinstitute.org/ricopili/; (Lam et al., 2019)), are influenced by the presence of diverse population structure within the cohort (Supplementary Table S2). We then suggest example workflows for either the stratified meta-analysis or joint mixed model approach (Figure 2).

The primary concern for QC in diverse cohorts is that some conventional QC metrics expect the allele frequency of a variant to be consistent within a sample. However, many QC criteria, do not depend on allele frequency, including per variant or per person missingness rates, Mendelian error rates in trio data, and comparison of case and control missingness rates for binary traits. These criteria can therefore be applied without modification or special concern in QC of diverse cohorts.

For the remaining QC criteria that are sensitive to allele frequency differences between ancestries, including individual heterozygosity rates, inference of sex from chromosome X heterozygosity, deviations from Hardy-Weinberg Equilibrium, and minor allele frequency, their application in the diverse cohort depends on the choice between two typical strategies for analyzing samples from multiple major populations and/or admixed populations: (1) Empirically define and assign samples to major/admixed populations using genome-wide data, analyze each population separately, and conduct cross-ancestry meta-analysis (stratified meta-analysis approach), or (2) analyze samples from multiple populations together with a mixed model (joint mixed model approach). The genotype QC procedures should fit the association analysis strategy (Figure 2, Supplemental Figure S1).

For the joint mixed model approach, QC should be done with respect to the entire sample, though this may include evaluation of QC criteria within major population subgroups as well (Figure 2, Supplemental Figure S1). Appropriate modifications to QC criteria that rely on allele frequencies are described here (Supplementary Table S2). Both heterozygosity measures (Bycroft et al., 2018) and associations with batch can be adjusted for ancestry differences using principal components (PC). HWE can be evaluated with relaxed thresholds for excluding variants or can be adjusted for predicted allele
frequencies as proposed (Hao and Storey, 2017). Filtering variants on MAF should also be considered in the context of population structure, since a single MAF threshold for the cohort may not fully account for variations in data quality in population subgroups where MAF is lower. When comparing MAF for a variant in a particular sample to an external reference panel as a QC check, the population background of the study and the reference samples should be matched (Bycroft et al., 2018). Relatedness checks should be performed using methods robust to population structure, such as PCRelate (Conomos et al., 2018), REAP (Thornton et al., 2012), and KING (Manichaikul et al., 2010). Of these, PCRelate and REAP can also model individuals with admixed ancestries. Note however that the mixed model approach does not require removal of related samples since genetic relatedness can be modelled appropriately (Conomos et al., 2018; Zhang and Pan, 2015).

For the stratified meta-analysis approach, the QC criteria affected by allele frequencies can be applied to each population separately (Supplemental Figure S1) (Duncan et al., 2018; Peterson et al., 2017; Walters et al., 2018). For admixed populations, the QC criteria that depend on allele frequency may still need to be adjusted for ancestry or admixture proportions as described above. To identify and remove related individuals that will confound conventional regression-based association tests (i.e., based on familial relatedness), the above relatedness estimation methods robust to population structure may also be recommended, although simpler relatedness estimation as implemented in PLINK (Purcell et al., 2007) may be sufficient for samples from a single major population.

**I.2 Example workflow for stratified meta-analysis approach**

We provide here one possible workflow for QC of a diverse cohort following the stratified meta-analysis approach (Figure 2, Supplementary Figure 1). *This example is not intended to be prescriptive*, since each dataset presents unique challenges to ensure robust control of technical artifacts and to avoid false positive associations. Instead, we aim to illustrate one possible application of the considerations above for QC of a diverse ancestry cohort for studies aiming to follow the stratified meta-analysis approach. Where applicable, we suggest thresholds that are in line with default filter values in Ricopili (Lam et al., 2019).

Beginning with the full diverse ancestry cohort, perform initial QC using criteria that are not influenced by population allele frequencies. Specifically, remove variants with >5% missingness, then remove samples with >2% missingness. For family-based samples, identify Mendelian errors and remove samples with excessive errors (e.g., > 10,000) and variants with multiple errors (e.g., > 4). Lastly, remove variants whose missingness is associated with the phenotype (e.g., if the difference in missingness proportion between cases and controls is > 0.02).

After this initial QC, merge the cohort with a population reference panel (e.g., 1000 Genomes Project; (Sudmant et al., 2015)), and verify that genome builds and alleles are properly matched. Then identify a set of variants for use in principal components analysis (PCA), prioritizing high quality common (allele frequency > 0.05) autosomal variants pruned for linkage disequilibrium (LD; \( r^2 > 0.2 \)) and excluding regions of long-range LD (Price et al., 2008). Filtering these variants for strong deviations from Hardy-Weinberg Equilibrium (HWE) may also be considered. Using these variants, use PC-AiR (Conomos et al., 2015) to estimate the ancestry structure for all samples controlling for relatedness.
From the PC-AiR results, stratify the cohort into ancestrally homogeneous subgroups. Depending on the structure of the sample, it may be possible to stratify the sample based on threshold values for the estimated principal components (PCs), using cluster analysis, or based on estimated ancestry proportions from ADMIXTURE (Alexander et al., 2009) or a random forest classifier trained using the reference panel samples. In all cases, visual inspection of the PCA results is recommended to evaluate the matching of the cohort samples to the reference panel.

Once the cohort has been stratified, perform additional QC within each assigned ancestry group. Filter samples for autosomal heterozygosity (e.g., $|F_{het}| > 0.2$) and for discordance between reported sex and genetically inferred sex based on chromosome X heterozygosity. Remove variants that deviate from HWE (e.g., $p < 1e-6$ in controls), that are invariant, or that are associated with study batch or genotyping platform. Optionally, consider repeating the filters on missingness and mendelian errors applied prior to PCA. Lastly, using a high quality variant set (described above), estimate genetic relatedness between the samples and remove cryptically related individuals (e.g., relatedness $\hat{\pi} > 0.2$) if present.

After the above process, it may be desirable to repeat the process of ancestry inference in order to verify that the QCed cohort matches well to the population reference panel intended for imputation. Once the researcher is content with the sample QC, the ancestry group can then be aligned to the reference panel for imputation, again being sure to verify that the variant names, alleles and genome build are well aligned. Comparisons of allele frequencies between the stratified sample and the matched ancestry group in the reference panel may aid this alignment. Imputation, including filtering based on imputation quality and allele frequency, and association analysis can then proceed as described below.

I.3 Example workflow for joint mixed model approach

As noted above, we provide an example workflow here to illustrate the joint mixed model approach but we do not intend this example to be prescriptive. We draw heavily from recently reported QC protocols applied in large, diverse cohorts (e.g., (Bycroft et al., 2018; Wojcik et al., 2019)). As noted elsewhere, evaluation of methods for joint analysis of diverse ancestry cohorts remains an active area of research, with a lack of consensus on what current options are best and a high likelihood that new methods will be proposed to replace current standards. Nevertheless, we hope this example is useful for illustrating one possible approach to the issues noted in this paper when aiming to analyze a diverse cohort jointly.

Begin with initial QC and ancestry inference as described above in the example workflow for the stratified meta-analysis approach. Instead of stratifying the cohort by ancestry, evaluate whether the PCA solution is consistent with the expected ancestry composition of the cohort and retain the PCs for use in adjusting remaining QC metrics for ancestry structure. Remove samples that are outliers for autosomal heterozygosity after regressing out PCs as described by (Bycroft et al., 2018). Similarly, remove samples whose sex inferred from visual inspection of heterozygosity on chromosome X after controlling for PCs is discordant with their reported sex. Next remove variants that show deviations from HWE based on the predicted allele frequencies (Hao and Storey, 2017), variants associated with batch or genotyping platform after controlling for ancestry PCs, and variants with low minor allele counts across ancestries. Relatedness should be evaluated using a method robust to ancestry structure (e.g. PCRelate; (Conomos et al., 2018)) in order to confirm expected relatedness in family-based cohorts or identify individuals with unusual relatedness patterns. It is not necessary to remove individuals with
cryptic relatedness from the mixed model analysis, though using a mixed model method that models both recent and ancestral relatedness, as well as effects of shared environment, may be advisable (e.g., (Conomos et al., 2018)).

Once this QC is complete, the cohort can be prepared for imputation and association testing as described below. Alignment of the cohort to the reference panel for imputation should follow the same precautions described in the stratified meta-analysis approach, although comparisons to expected allele frequencies may need to be evaluated in subsets of the cohort.

II. Inferring Population Structure

An important factor for understanding the effects of ancestry on GWAS is preventing spurious findings due to population stratification. Population stratification occurs when both disease prevalence and allelic frequency differences exist in the subpopulations sampled, which may lead to false positive associations of genetic signals (Marchini et al., 2004). A common approach to examining population structure is the use of principal component analysis (PCA), which has been applied to adjust for global and local population structures. The principal components (PCs) can be used to empirically assign samples to more homogenous ancestry groups based on reference populations, to exclude outliers, and can be used as covariates in association analyses to reduce the effects of population stratification. Note that PCA should be done with genome-wide independent single-nucleotide polymorphisms (SNPs) with SNPs in long range LD regions excluded (Price et al., 2008).

For cohorts with diverse ancestral backgrounds, we can empirically assign samples into groups with more homogeneous genetic backgrounds. For example, using PCA programs such as Eigenstrat (Price et al., 2006), ancestry-informative PCs can be estimated from an external reference (e.g., The 1000 Genomes Project (1KGP))(Sudmant et al., 2015) and the GWAS sample can be projected onto the PC axes defined by the external reference. GWAS samples are then assigned to the closest reference population match based on PCs (Peterson et al., 2017). Common GWAS strategies remove those subjects that are missing self-reported race, those endorsing more than one census category, or those ancestral groups with small sample sizes. By applying reference population matching, more genetically homogenous groups can be identified for downstream genetic analyses while retaining those groups that are commonly removed from analyses. However, this approach is limited by the populations represented on the reference panel and can be problematic if samples are not well represented on the panel. Indeed, as used in this example, the 1000 Genomes reference panel is missing populations, including those from Oceania and the Middle East. As noted in the GWAS section, once samples are assigned to more homogeneous groups, a second round of PCA is performed within each group and the resulting PCs are used as covariates in association analyses to control for residual population stratification or technical artifacts.

Other population assignment approaches include clustering of PCA (Purcell et al., 2007), ADMIXTURE proportions (Alexander et al., 2009), snpweights (Chen et al., 2013; Duncan et al., 2018), and machine learning methods (e.g., random forest) (see review (Hellwege et al., 2017)). These assignment methods will not provide - and are not intended to provide - accurate, detailed ancestral population background information for each sample; but rather to provide a working solution to reduce
population stratification. We stress that sample group assignment and identifying appropriate reference population panels can be difficult, particularly for admixed ancestry, thus requiring careful inspection of data and methods (Medina-Gomez et al., 2015).

III. Imputation and Population Reference Panels

III.1 Genotype imputation methods
Genotype imputation is a cost-effective computational approach for inferring genotypes or genotype probabilities at variants that have not been directly genotyped on GWAS arrays, based on comparisons to genetic data from external reference samples. The most widely used approach for genotype imputation is based on haplotype estimation methods using hidden Markov models (HMM). Commonly used imputation tools are IMPUTE2 (Howie et al., 2009), Minimac ((Howie et al., 2012); default for Michigan Imputation Server, https://imputationserver.sph.umich.edu), Beagle (Browning and Browning, 2009), PBWT ((Durbin, 2014); default for Sanger Imputation Server, https://imputation.sanger.ac.uk/) and MACH-Admix (Liu et al., 2013a). Unlike statistical imputation techniques (i.e., linear regression and regression trees) these modern frameworks model important characteristics of genetic data (i.e., linkage patterns, recombination hotspots, and genotyping errors) and are ideally suited for haplotype estimation and inferring genotypes using multi-ancestry reference panels. To enhance the speed of imputation, GWAS data are typically pre-phased, estimate haplotypes from genotyped markers in the cohort before beginning imputation, using programs like EAGLE (Loh et al., 2016) or SHAPEIT (Delaneau et al., 2011). Although splitting the process into pre-phasing and imputation can slightly reduce the imputation accuracy in admixed populations (Howie et al., 2012; Roshyara et al., 2016), ensuring the computational feasibility of using larger multi-ancestry reference panels with pre-phasing can lead to better imputation accuracy than imputing without pre-phasing with a smaller reference panel.

Imputation servers, such as the Michigan (https://imputationserver.sph.umich.edu) and Sanger (https://imputation.sanger.ac.uk) servers mentioned above, provide a standardized and easy to use platform for performing imputation of genotyped cohorts, facilitating collaborative efforts. These servers have the benefit of allowing secure use of large multi-ancestry reference panels whose data sharing restrictions would otherwise prevent access to the panel’s individual genetic data used for imputation. On the other hand, use of these servers does require that the data use restrictions of the genotyped cohort permit submission of the data to the secure server for imputation.

These imputation servers also provide instructions and example code for aligning the genotyped cohort to the reference panel prior to starting imputation. This alignment process assures that the variant name, alleles, genomic location, human genome build, and strand (e.g., for strand ambiguous variants) are all harmonized between the reference panel and the genotyped cohort in order to ensure the proper data is compared by the imputation algorithm. Comparison of sample allele frequencies to appropriate ancestry samples in the reference panel can further aid this process. For imputation outside of these servers, scripts implementing similar checks are included as part of the imputation module in Ricopili (Lam et al., 2019).
III.2 Factors affecting imputation accuracy

Many factors can affect imputation accuracy. Increasing the sample size of the reference panel can increase the imputation accuracy when the genetic similarity between study samples and the reference panel is maintained. This is especially true for variants with low minor allele frequency (i.e., rare variants) where large sample sizes improve the representation of haplotypes containing the rare variant for matching to the genotyped cohort (McCarthy et al., 2016). Importantly, imputation frameworks do not require that the reference panel be restricted to only haplotypes that match the ancestry of the target cohort. Using the largest available multi-ancestry reference panel is expected to perform similar to subsetting that reference panel to match the target population (Howie et al., 2012). This property potentially facilitates the use of the same imputation reference panel for all cohorts in a study regardless of the ancestry composition of each cohort. Improving the genome sequencing depth and coverage of the reference panel can additionally improve imputation accuracy since the coverage is directly correlated to genotype accuracy, which in turn is correlated with the accuracy of inferred haplotypes (Das et al., 2018). These factors highlight the value of large, high-quality, ancestrally-diverse imputation reference panels for use in GWAS of diverse ancestry cohorts.

The density and distribution of markers in genotyping arrays can impact imputation accuracy by affecting the chances of finding the shared haplotype segments. Using multi-ancestry arrays helps to address this problem by improving the amount of genetic variation tagged by LD with the genotyped variants within different ancestry groups. This effect is most pronounced for variant with low allele frequencies, since these variants have lower LD to common variants that can be used for efficient tagging on genotyping arrays and have a higher chance of being population-specific (Wojcik et al., 2018). Genotyping with multi-ancestry arrays is thus potentially beneficial for multi-ancestry studies as long as the included populations are covered by the genotyping chip’s design.

Genomic locations with multi-allelic variance can be a challenge for most imputation algorithms. Representing such sites as multiple di-allelic variants can be problematic when imputation requires that the posterior allele probabilities of each variant must sum to 1.0. This concern is especially noteworthy for diverse cohorts and large reference panels, since the number of multi-allelic variants will increase with improved ancestry representation on larger reference panels. Currently Beagle is the only tool that supports imputing with multi-allelic marker representation (Browning et al., 2018).

Despite these issues, current reference panels yield high quality imputation results for most common variants in most major ancestry groups (Taliun et al., 2019). Imputation accuracy, especially for lower frequency variants, should continue to improve as multi-ancestry genotyping arrays, large diverse reference panels, and imputation methods continue to develop. Post-imputation QC remains important, however, for removing poorly imputed variants from consideration in GWAS.

III.3 Preparing imputed data for GWAS

After imputation is complete, post-imputation QC should be performed to identify variants suitable for association testing. This QC should include filters for imputation quality (e.g., INFO scores reported by the imputation software) and minor allele frequency or counts (e.g., at the level necessary to ensure the chosen association model will have valid p-values based on assumptions about the asymptotic
distribution of the test statistic), along with consideration of batch effects and other variant QC criteria described previously (Supplementary Methods I) as necessary.

IV. Genome-wide Association

After imputation, the data are ready to perform genome-wide association analysis, testing the association between each variant and a target phenotype. Ideally this relationship is tested either using the observed data likelihood based on the posterior probabilities of each genotype call estimated by imputation, or using imputed dosages (the expected number of alternative alleles) in linear or logistic regression. Both of these methods account for uncertainty in the imputation and yield similar power as long as effect sizes are small (Zheng et al., 2011), though analysis of the dosages is more common and computationally convenient. Using best-guess genotypes (the genotype call with the highest posterior probability for each individual) in regression can also give similarly acceptable performance when restricting to variants with high imputation quality (e.g., INFO scores above 0.8 or 0.9; (Zheng et al., 2011) as long as sample sizes are large and effect sizes are small. In our experience (Walters et al., 2018), when using best-guess genotypes it may also be desirable to set low certainty calls (e.g., posterior probabilities below 0.8) as missing and filter variants for call rate at this threshold, but the comparative performance of that approach at varying INFO and posterior probability thresholds has been less studied in the literature. In all cases, it’s recommended that variants be filtered for imputation quality and allele frequency or count prior to GWAS (Lam et al., 2019; Winkler et al., 2014).

IV.1 Mixed model implementations

Mixed models are ideally suited for joint analysis of diverse cohorts. The literature on the use of mixed models with genome-wide data is rapidly growing, but a useful introduction to the motivation and pitfalls of mixed models can be found elsewhere (Yang et al., 2014). Briefly, mixed models control for population structure and relatedness by modelling one or more random effects variance components based on the genetic similarity of individuals in the GWAS. Modelling the genetic similarity between individuals can reduce false positives and improve power, with the increase in power proportional to the amount of variance explained by the modelled random effects terms (Yang et al., 2014). On the other hand, because basic mixed models for genome-wide data only model effects proportional to genetic similarity, stratification that does not fully fit this simple model may not be fully controlled. For example, variants that are more or less differentiated between populations than the genome-wide average will not have correct control of false positive rates (Conomos et al., 2018). The performance of mixed models will similarly degrade if environmental stratification does not perfectly align with genetic similarity (Heckerman et al., 2016). Some of these concerns may potentially be addressed by using extended versions of the mixed models to improve control of population structure by modelling environmental stratification directly (Heckerman et al., 2016) or separately modelling recent and distant relatedness (Conomos et al., 2018; Zhang and Pan, 2015). Still, more methodological development is needed before mixed models or other strategies for joint GWAS of a diverse cohort can be confidently recommended.
Mixed models can be computationally intensive, thus multiple implementations exist to facilitate genome-wide analyses with both useful modelling features and computational efficiency (Supplementary Table S4). Common choices for linear mixed model with GWAS data for continuous phenotypes include BOLT-LMM (Loh et al., 2015), GCTA (Yang et al., 2011a), and GEMMA (Zhou and Stephens, 2012). Among these options, GEMMA provides the option of a Bayesian sparse mixed model and BOLT-LMM offers a Bayesian mixture-of-normals prior for the random effects component as an alternative to the infinitesimal prior implied by the conventional mixed model. Both GCTA and GEMMA include the option to model multiple random effects components. This more flexible modelling of genetic architecture may further improve power for GWAS, but it is not evident that the extra flexibility impacts control of stratification and the computationally efficient approximations used by BOLT-LMM may only be appropriate for sample sizes > 5,000 (Loh et al., 2015).

Available implementations for logistic mixed model for binary phenotypes, which are even more computationally intensive, include GMMAT (Chen et al., 2016) and SAIGE (Zhou et al., 2018). Alternatively, GEMMA offers a probit version of the Bayesian sparse mixed model. Note applying a linear mixed model to binary phenotypes is likely to result in inflated type I error rates (Chen et al., 2016; Wu et al., 2011). SAIGE utilizes computationally efficient approximations not present in GMMAT, as well as better approximations to calibrate the test for samples with small case counts. Both methods rely primarily on a score test for GWAS, but GMMAT does include the option of effect size estimation for a Wald test (at a much greater computational cost) while SAIGE approximates effect size estimation under a null model. This computational burden remains a challenge for applying mixed models in GWAS of binary phenotypes, especially at large sample sizes, and thus is an important consideration in constructing an analysis plan for diverse and admixed ancestry cohorts.

### IV.2 Evaluating control of population stratification

Multiple methods have been proposed for evaluating the degree of uncontrolled population stratification remaining in GWAS results. We briefly review the most common options here and note how they might be expected to perform in diverse ancestry studies.

One of the simplest metrics is $\lambda_{GC}$ (Devlin and Roeder, 1999), the ratio between the median observed chi square statistic in the GWAS results with the expected median value of the chi square statistic under the null hypothesis. Values greater than 1 indicate GWAS test statistics are inflated from the expected null distribution, though they do not distinguish whether the inflation is from confounding or polygenic signal. Because $\lambda_{GC}$ will scale with sample size when there is any inflation of the chi square statistics, it is sometimes reported after rescaling $\lambda_{GC}$ to correspond to a study with 1000 cases and 1000 controls, known as $\lambda_{1000}$, to facilitate interpretation on a more standardized scale (de Bakker et al., 2008; Freedman et al., 2004). Since $\lambda_{GC}$ and $\lambda_{1000}$ reflect inflation from true polygenic effects as well as population stratification it can be difficult to conclusively evaluate the presence of population stratification based on these measures. On the other hand, the simplicity of these measures using just the GWAS test statistics makes them generally applicable for use in multi-ancestry studies.

To distinguish inflation of GWAS results from population stratification as opposed to polygenic signal, LD Score regression (LDSC) models an expected relationship between polygenic signal and LD (Bulik-Sullivan et al., 2015). This regression yields an estimated intercept which has an expected value...
of 1 when no stratification or other confounding is present in the GWAS, and will be greater than 1 when there is inflation that is not explained by the modelled relationship of GWAS results with LD. This inflation in the LDSC intercept may reflect population stratification or other misspecification of the LDSC model (Bulik-Sullivan et al., 2015; Lee et al., 2018). The ratio between the inflation in this value and the inflation in the mean chi square statistic from the GWAS can be used as a measure of how much of the total inflation (as measured by $\lambda_{GC}$) is due to stratification or other confounders as opposed to polygenic signal. For multi-ancestry studies, it’s important to note that LDSC relies on having available LD scores for the studied population, usually from a sequenced reference panel. This may make it challenging to assess stratification in multi-ancestry studies, whether analyzed jointly or meta-analyzed across major population groups (though LDSC may be easier to apply within each ancestry group separately), especially in admixed samples requiring additional modelling of in-sample LD (Luo et al., 2019).

Finally, population stratification can be assessed by comparing GWAS results to ancestry structure in external samples. For example, correlations between the GWAS betas and SNPs’ loadings on PCs can provide evidence of confounding (Sohail et al., 2019). Comparison to within-family GWAS of siblings, which are more robust to population stratification, may also be informative (Berg et al., 2019). Although these approaches are more labor intensive and don’t yield a single numerical summary of the degree of stratification in the GWAS, they can provide stronger evidence for the existence of residual stratification in GWAS results, and how that stratification aligns with structure in known reference populations.

IV.3 Principal components in admixed samples

Investigators working with cohorts of admixed individuals may need to carefully consider the interpretation of PC covariates, since there is evidence that some PCs in those samples may reflect stochastic patterns in local ancestry tracts (i.e., which haplotypes were inherited from each ancestral population involved in the admixture; (Walters et al., 2018)). These local ancestry tracts can themselves be tested for association with the phenotype in admixture mapping analyses (Seldin et al., 2011). Directly controlling for local ancestry tracts in variant-level association analyses may further improve power and reduce certain type I errors in admixed samples (Li and Keating, 2014). Improving tools for inclusion of local ancestry information in GWAS is an area of active development (Atkinson et al., 2018).

V. Meta-analysis of GWAS Summary Statistics

Heterogeneity between cohorts is a separate concern from controlling for population structure within a cohort. Within-cohort differences related to ancestry or other factors will not be controlled by meta-analysis. Heterogeneity between cohorts, such as the differences in marginal effect sizes expected due to LD differences when cohorts are stratified by ancestry, violates the conventional fixed effects meta-analysis model. LD differences will also yield cross-cohort heterogeneity when analyzed with the joint mixed model GWAS approach, since the ancestry composition of each cohort is likely to differ.
Heterogeneity may similarly exist among cohorts from admixed populations due to differences in admixture proportions or different genome-wide distributions of ancestry tracts (Liu et al., 2013b).

When cross-cohort heterogeneity is present, using random-effects meta-analysis allows for the true marginal effect size of each variant to vary across studies, testing the null hypothesis that the mean true effect size across cohorts is zero (Han and Eskin, 2011). This approach is not necessarily more conservative than a fixed-effects meta-analysis, but in practice it does tend to give less significant p-values than fixed-effect models in the presence of heterogeneity (Han and Eskin, 2011). To avoid a conservative null hypothesis, it may be preferable to test whether there is a non-zero effect of the variant in at least one cohort, as implemented by the RE2 and RE2C model, yielding higher statistical power than fixed effects model in the presence of heterogeneity (Han and Eskin, 2011; Lee et al., 2017; Periyasamy et al., 2019).

Alternatively, a pair of methods has been proposed to directly model cross-population differences in meta-analysis. MANTRA uses a Bayesian approach that assumes variant effect sizes may differ between sets of GWAS in the meta-analysis, with higher heterogeneity between more genetically diverged populations (Morris, 2011). MR-MEGA similarly uses meta-regression in a frequentist framework to model a variant’s effect size in each study as a function of a cohort’s position in ancestry space (Mägi et al., 2017). Like the modified random effects models, both of these methods focus on the alternative hypothesis that the variant has a non-zero effect in at least one cohort. Notably, both methods rely on having single-population GWAS as input to the meta-analysis. Simulations suggest both MANTRA and MR-MEGA similarly increased power compared to RE2 (Mägi et al., 2017; Wang et al., 2013). MR-MEGA also provides convenient effect size estimation and naturally extends to fine-mapping analyses. Still, the models are distinct enough that the optimal choice between these methods is likely to ultimately depend on the researcher’s hypothesis of interest, the statistical assumptions they’re willing to make about the distribution of effect sizes across populations, and their preference for Bayesian or frequentist inference.

Furthermore, GWAS meta-analyses should be evaluated using a genome-wide significance threshold that is appropriate for the genetic diversity of the studied populations. The current consensus significance threshold of \( p < 5 \times 10^{-8} \) was developed with a focus on analysis of common variants in European ancestry samples, though in practice it is often applied in other ancestries. However the multiple testing burden for GWAS in African populations, for example, may be more than twice the testing burden in European populations due to a higher number of effectively independent genetic variants (Pe’er et al., 2008). Currently recommended significance thresholds range from \( 6 \times 10^{-7} \) to \( 6 \times 10^{-9} \) depending on allele frequency threshold, imputation panel, and ancestry of the GWAS, as well as the method used to estimate the appropriate threshold (Kanai et al., 2016; Li et al., 2012; Pe’er et al., 2008).

VI. Heritability and Genetic Correlation

Heritability is the proportion of a phenotype’s variance explained by inherited genetic factors. Individual-level genotype data or GWAS summary statistics can be used to estimate the SNP heritability \( h^2_{SNP} \), the proportion of a phenotype’s variance that can be explained by the additive effects of common
SNPs present in a GWAS. This quantity is useful as a measure of the total polygenic signal existing in GWAS variants for a trait, as an estimated upper bound on the expected predictive power of polygenic risk scores (PRS) based on those variants with increasing GWAS sample size, and as a lower bound estimate of the heritability from additive genetic effects for comparison to twin- and family-based estimates. For binary (disease) outcomes, $h^2_{SNP}$ is commonly transformed to the liability scale, i.e. to account for the dichotomization of an underlying continuous liability distribution and over-sampling of cases relative to the population prevalence (Lee et al., 2011). Genetic correlation ($r_g$) describes the extent to which genetic architecture is shared between traits, involving both the degree of colocalization of causal variants and the similarity of the direction and magnitude of the causal effect sizes. Unlike heritability, $r_g$ is scale invariant and does not require scale transformation for binary phenotypes (van Rheenen et al., 2019). Estimates of $r_g$ are generally robust to uncertainty regarding the true underlying heritability model (Lee et al., 2018; Ni et al., 2018). See recent reviews of methods and applications of $h^2_{SNP}$ (Yang et al., 2017) and $r_g$ (van Rheenen et al., 2019).

For diverse cohorts, it is important to consider both the estimation and interpretation of $h^2_{SNP}$. Estimation is discussed at length below. Interpretation, especially in comparing estimates from different populations, requires careful evaluation of phenotyping (Supplementary Methods VII), study design, and environmental factors that will all have a role in shaping $h^2_{SNP}$ as a fraction of the total variability of the phenotype.

**VI.1 Estimating SNP heritability from individual-level genetic data**

Genomic relatedness matrix (GRM) restricted maximum likelihood (GREML) is a statistical method for variance component estimation that estimates $h^2_{SNP}$ from a GRM of pairwise inter-individual genetic similarity values. As popularized by the GCTA software tool (Yang et al., 2011a) and other implementations of related models (e.g., LDAK-MS (Speed et al., 2017); BOLT (Loh et al., 2015)), this approach can be straightforwardly applied to ancestrally homogeneous study cohorts. For application to diverse cohorts, we note two primary concerns: First, GREML estimates of $h^2_{SNP}$ can be biased in the presence of population stratification (Browning and Browning, 2011), and population substructure has the potential to manifest as genetic similarity, thereby confounding $h^2_{SNP}$ with phenotypic similarity that is attributable to ancestry correlations. This bias can be reduced by controlling for fixed effects of ancestry structure (Conomos et al., 2018) or by using a more robust, partitioned GREML approach such as GCTA-LDMS-I (Evans et al., 2018; Yang et al., 2015). However, further modelling may be required when applying GREML to populations with significant recent admixture (Zaitlen et al., 2014). A second concern is that modelling confounding arising from population stratification may not fully control for the effects of environmental stratification; approaches that directly model environmental structure (Heckerman et al., 2016) or shared environmental effects among close family members (Zaitlen et al., 2013) can reduce bias due to environmental stratification. Comparison of GREML $h^2_{SNP}$ estimates to the sum of estimates of $h^2_{SNP}$ from each chromosome may aid in identifying instances where the $h^2_{SNP}$ estimate has been biased by some form of confounding (Yang et al., 2011b).
VI.2 Estimating SNP heritability from GWAS summary statistics

LD Score regression (LDSC) (Bulik-Sullivan et al., 2015) allows estimation of \( h^2_{SNP} \) from genome-wide summary statistics by modelling an expected relationship between the strength of signal observed for a SNP in a GWAS of a polygenic trait and the amount of genetic variation that SNP tags in the population (i.e., its LD score). Estimating \( h^2_{SNP} \) using LDSC in diverse cohorts currently poses two major issues. The first is the reliance on LD scores calculated for ancestry-matched reference data (Bulik-Sullivan et al., 2015), which may not be available when analyzing diverse or admixed cohorts. A recently proposed extension of LDSC uses in-sample estimates of LD structure to more accurately estimate \( h^2_{SNP} \) from GWAS summary statistics for admixed populations (Luo et al., 2019). A second concern is that the LDSC model may not fully differentiate stratification from polygenic effects. Model misspecification may deflate \( h^2_{SNP} \) estimates and inflate the intercept, while stratification effects correlated with LD (e.g., due to background selection) may inflate \( h^2_{SNP} \) estimates (Evans et al., 2018; Holmes et al., 2019; Lee et al., 2018). Some caution is warranted when interpreting results from LDSC and related methods in stratified samples.

VI.3 Estimating genetic correlation

Genetic correlation (\( r_g \)) describes the extent to which genetic effects are correlated between traits, involving both the degree of colocalization of causal variants and the similarity of the direction and magnitude of the causal effect sizes. Unlike heritability, \( r_g \) is scale invariant and does not require scale transformation for binary phenotypes (van Rheenen et al., 2019). Estimates of \( r_g \) are generally robust to uncertainty regarding the true underlying heritability model (Lee et al., 2018; Ni et al., 2018).

Defining and estimating \( r_g \) between divergent populations is complicated by differences in allele frequencies and LD structure. As noted by Brown et al., allele frequency differences create the need to distinguish between the correlation of per-allele causal effect sizes (genetic-effect correlation) and the correlation of effect sizes in terms of the variance explained per SNP after accounting for MAF differences between the two populations (genetic-impact correlation (Brown et al., 2016). Galinsky et al. similarly distinguish between the correlation of causal effect sizes and the correlation of observed associations as a function of population differences in patterns of LD, and provide a method for estimation of the difference between these values (Galinsky et al., 2019).

To estimate \( r_g \) between populations, the bivariate extension of GREML — as implemented in GCTA for estimation of \( r_g \) between traits observed in disjoint sets of individuals (Visscher et al., 2014) — models the relationship between the genetic similarity of a cross-population pair of individuals and their phenotypic similarity. This approach has been applied successfully in two studies, with evidence that strict controls for the influence of ancestry on genetic similarity across populations are required (de Candia et al., 2013). Modeling of environmental differences across ancestries using the GxE function within GCTA may also be valuable in this context. Further methods evaluation of GREML-based models for estimating \( r_g \), including their behavior in the context of LD and MAF differences and the same considerations for \( h^2_{SNP} \) estimation, would be valuable. Meanwhile, for estimation of \( r_g \) from summary statistics Popcorn extends the LDSC model estimates cross-population genetic correlations based on modified LD scores that reflect the similarity of LD between the two populations (Brown et al., 2016). Similar to the considerations for LDSC estimation of \( h^2_{SNP} \), the Popcorn method requires that each
set of GWAS summary statistics come from a homogeneous population and that LD can be computed in a reference panel containing a set of individuals from each of the respective populations. Given the interest in this area, it is likely that future methods development will continue to add flexibility or identify alternatives for estimating $r_g$ in diverse cohorts.

VII. Phenotypic Measurement

Most psychiatric classification systems and diagnostic measures have been developed and validated in individuals from industrialized, Western societies (Henrich et al., 2010). This is a challenge for global and cross-cultural collaborations, where measurement invariance, or equivalence in the underlying construct an assessment measures across groups, is not assured; bias in this context occurs when differences in scores across groups do not correspond to differences in the underlying construct or trait. Potential sources of bias in cross-cultural research are extensive and can be classified into three major types: 1) construct bias occurs when a construct is not identical across cultures; 2) method bias occurs when features such as sampling, instrument characteristics, response styles, and administration weaken the validity of responses; 3) item bias, also called differential item functioning, occurs when instrument items have different meanings across groups (Vijver et al., 2004). For instance, investigations into cross-cultural differences in the prevalence and expression of major depression, have suggested that although the underlying disorder construct may appear to be equivalent across groups (Kendler et al., 2015; Simon et al., 2002), individuals may differ culturally in terms of the level of symptomatology reached prior to seeking help (Bromet et al., 2011; Simon et al., 2002). Additionally, certain items within commonly-used assessment tools can be shown to have varying utility in different cultural contexts (Uebelacker et al., 2009). The resulting cross-cultural phenotypic differences could affect both gene discovery and the transferability of genetic findings between populations. More importantly, without the inclusion and consideration of diverse populations in the development, validation, and deployment of diagnostic measures used in genetic studies, an imperfect, culturally constrained picture of disease etiology will emerge.
 references


Holmes, J.B., Speed, D., and Balding, D.J. (2019). Summary statistic analyses can mistake confounding bias for...
heritability. bioRxiv.


tissue-specific gene sets in admixed populations.


