Supplemental Material

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

Supplemental Tables

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

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

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.

QC step	Motivation	Ancestry-related considerations
Genotype missing percentage per variant	Identify variants with poor genotyping quality	Typically applied without alterations.
Genotype missing percentage per sample	Identify samples with poor data quality	Typically applied without alterations.
Difference in variant missingness between groups (e.g., cases vs. controls, batches, arrays)	Prevent confounding of association results by technical artifacts related to data quality or batch effects	Consider covarying missingness with PCs if there are concerns about confounding with ancestry.
Mendelian error rate per variant (family data only)	Identify variants with poor genotyping quality (assuming rate of true *de novo* mutations is low)	None.
Mendelian error rate per sample (family data only)	Identify samples with poor data quality; may also indicate reported pedigree relationship is incorrect	None.
Relatedness between pairs of samples	Confirm expected relationships (in family-based cohorts) and identify cryptic relationships that may confound tests of association (conventional regression analyses only)	Within-ancestry pairs will show greater relatedness than between ancestry pairs, beyond pedigree-based expectations for IBD sharing. For admixed samples, an admixture-aware tool is needed: PCRelate or REAP. KING is suitable for mixed homogeneous populations, not admixed populations.
Heterozygosity rate	Identify potential sample contamination or other unusual artifacts	Influenced by allele frequency differences across ancestry. Consider covarying heterozygosity rate estimates with PCs, or stratifying into homogeneous populations (if no admixture). Higher rates of consanguinity in certain populations may also affect estimation of heterozygosity percentage.
Predicted sex from chromosome X heterozygosity	Identify potential sample swaps based on discordance from reported sex	Influenced by allele frequency differences across ancestry. Need to control for differences in allele frequency, or visually inspect distribution.
Variant association with batch	Prevent confounding of association results by technical artifacts related to data quality or batch effects	Consider covarying with PCs if there are concerns about confounding with ancestry.
Hardy-Weinberg equilibrium (HWE)	Identify variants with poor genotyping quality (assumes most HWE deviations are from technical issues rather than population structure)	Deviations from HWE are expected in admixed and diverse ancestry samples. Relax significance thresholds or use modified HWE test. Consider testing HWE in more homogeneous subgroups and applying SNP exclusions to the full sample.
Minor allele frequency threshold	Identify variants likely to have poor quality on genotyping arrays, and likely to perform poorly in association analysis due to not reaching asymptotic behavior of association test	Assess on a case-by-case basis. Single thresholds will have differential effects on subsamples that differ in ancestry and frequency within a sample. (Note: invariant SNPs may not be called properly by many genotype calling algorithms.)

Supplemental Table S3: New and ongoing initiatives to expand available imputation reference panels.

Reference Panels	N	Ancestries	Note
SG10K	4,810	Chinese, Malays, Indians	https://www.biorxiv.org/content/biorxiv/early/2018/08/11/390070.full.pdf
Icelanders	15,220	Icelandic population	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5607473/
African Genome Variation Project (AGVP)	100	African	https://www.sanger.ac.uk/science/collaboration/african-genome-variation-project
Human Genome Diversity Project (HGDP)	1,043	Africa, Europe, Middle East, South & Central Asia, East Asia, Oceania and the Americas	http://www.hagsc.org/hgdp/
23andMe	200,000	African American	http://grantome.com/grant/NIH/R44-HG009460-01
Western Indian Project	407	Western Indian	https://www.nature.com/articles/s41598-017-06905-6
Latin-American Genomic Initiative	6,487	Latin Americans	http://www.ashg.org/2014meeting/abstracts/fulltext/f140122398.htm
Simons Genomic Diversity Project	300	Africa, America, Central Asia/Siberia, East Asia, Oceania, South Asia, West Eurasia	PMID:27654912; most data public through EBI, some requires a release; http://reichdata.hms.harvard.edu/pub/datasets/sgdp/
GenomeDenmark	150	Danish population	PMID:28746312
Korean National Standard Reference Variome (KoVariome)	80	Korean population	PMID:29618732; ftp://biodisk.org/Release/VariomeData/
Estonian Biocentre Human Genome Diversity Panel (EGDP)	402	Africa, West Asia, Caucasus, Europe, Central Asia, South Asia, East Asia, Siberia, Island Southeast Asia, Sahul, Americas	PMID: 27654910; http://evolbio.ut.ee/CGgenomes.html
Genome Aggregation Database (gnomAD)	15,708	African/African American, Latino, Ashkensi Jewish, East Asian, Finnish, Non-Finnish European, Other	https://www.biorxiv.org/content/10.1101/531210v2; https://gnomad.broadinstitute.org

Supplemental Table S4: Listing of select methods and software implementations cited.

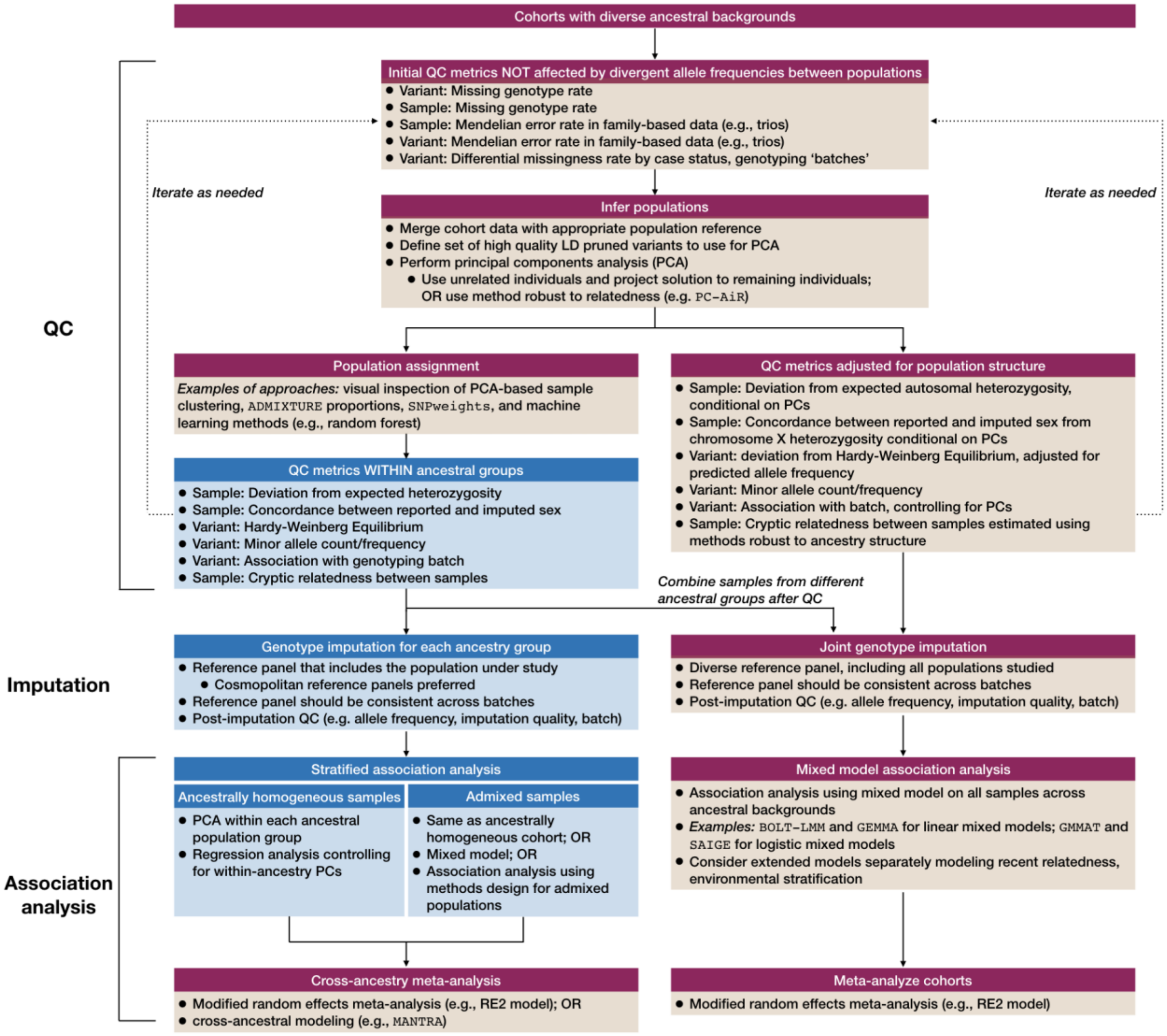
Software	Reference
ADMIXTURE	D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19, 1655–1664 (2009).
BEAGLE	Browning, B.L. & Browning, S.R., A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics, 84(2), 210–223, (2009).
BOLT-LMM	Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics, 47, 284–290 (2015).
EAGLE	Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. Nature Genetics, 48, 811–816 (2016).
Eigenstrat	Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38, 904–909 (2006).
GCTA	Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics. 88, 76–82 (2011).
GEMMA	Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 44, 821–824 (2012).
GMMAT	Chen, H. et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. American Journal of Human Genetics. 98, 653–666 (2016).
IMPUTE2	Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics. 5, e1000529 (2009).
KING	Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873 (2010).
LDAK	Speed D, et al. Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics, 91(6), 1011-21 (2012).
LD Score regression	Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. 47, 291–295 (2015).
MaCH-Admix	Liu, E.Y. et al. MaCH-admix: genotype imputation for admixed populations. Genetic Epidemiology, 37(1), 25–37 (2013).
MAGMA	de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Computational Biology 11, e1004219 (2015).
MANTRA	Morris, A.P. Transethnic meta-analysis of genomewide association studies. Genetic epidemiology, 35(8), 809–822 (2011)
Minimac	Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 44, 955–959 (2012).
MR-MEGA	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 Genetics 26, 3639–3650 (2017).
PAINTOR	Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping

	Studies. American Journal of Human Genetics, 97, 260–271 (2015).
PBWT	Richard Durbin, Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT), Bioinformatics, Volume 30, Issue 9, 1266–1272 (2014).
PC-AiR	Conomos, M. P., Miller, M., & Thornton, T. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genetic Epidemiology 39, 276–293 (2015).
PCRelate	Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. American Journal of Human Genetics, 98, 127–148 (2016).
Popcorn	Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. American Journal of Human Genetics, 99, 76–88 (2016).
RE2	Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. American Journal of Human Genetics, 88, 586-598 (2011).
RE2C	Lee, C. H., Eskin, E. & Han, B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. Bioinformatics 33, i379–i388 (2017).
REAP	Thornton, T. et al. Estimating kinship in admixed populations. American Journal of Human Genetics, 91, 122–138 (2012).
RICOPILI	Lam, M. et al. RICOPILI: Rapid Imputation for COnsortias PIpeLIne. bioRxiv. doi: 10.1101/587196
SAIGE	Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics. 50, 1335–1341 (2018).
SHAPEIT	Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. Nature Methods 9, 179–181 (2011).
S-PrediXcan	Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nature Communications. 9, 1825 (2018).

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).


Cohorts with diverse ancestral backgrounds
Initial QC metrics NOT affected by divergent allele frequencies between populations
• Variant: Missing genotype rate
• Sample: Missing genotype rate
• Sample: Mendelian error rate in family-based data (e.g., trios)
• Variant: Mendelian error rate in family-based data (e.g., trios)
• Variant: Differential missingness rate by case status, genotyping 'batches'
Iterate as needed
Iterate as needed
Infer populations
• Merge cohort data with appropriate population reference
• Define set of high quality LD pruned variants to use for PCA
• Perform principal components analysis (PCA)
• Use unrelated individuals and project solution to remaining individuals; OR use method robust to relatedness (e.g. PC-AiR)
QC
Population assignment
Examples of approaches: visual inspection of PCA-based sample clustering, ADMIXTURE proportions, SNPweights, and machine learning methods (e.g., random forest)
QC metrics adjusted for population structure
• Sample: Deviation from expected autosomal heterozygosity, conditional on PCs
• Sample: Concordance between reported and imputed sex from chromosome X heterozygosity conditional on PCs
• Variant: deviation from Hardy-Weinberg Equilibrium, adjusted for predicted allele frequency
• Variant: Minor allele count/frequency
• Variant: Association with batch, controlling for PCs
• Sample: Cryptic relatedness between samples estimated using methods robust to ancestry structure
QC metrics WITHIN ancestral groups
• Sample: Deviation from expected heterozygosity
• Sample: Concordance between reported and imputed sex
• Variant: Hardy-Weinberg Equilibrium
• Variant: Minor allele count/frequency
• Variant: Association with genotyping batch
• Sample: Cryptic relatedness between samples
Combine samples from different ancestral groups after QC
Imputation
Genotype imputation for each ancestry group
• Reference panel that includes the population under study
• Cosmopolitan reference panels preferred
• Reference panel should be consistent across batches
• Post-imputation QC (e.g. allele frequency, imputation quality, batch)
Joint genotype imputation
• Diverse reference panel, including all populations studied
• Reference panel should be consistent across batches
• Post-imputation QC (e.g. allele frequency, imputation quality, batch)
Association analysis
Stratified association analysis
Ancestrally homogeneous samples
• PCA within each ancestral population group
• Regression analysis controlling for within-ancestry PCs
Admixed samples
• Same as ancestrally homogeneous cohort; OR
• Mixed model; OR
• Association analysis using methods design for admixed populations
Mixed model association analysis
• Association analysis using mixed model on all samples across ancestral backgrounds
• Examples: BOLT-LMM and GEMMA for linear mixed models; GMMAT and SAIGE for logistic mixed models
• Consider extended models separately modeling recent relatedness, environmental stratification
Cross-ancestry meta-analysis
• Modified random effects meta-analysis (e.g., RE2 model); OR
• cross-ancestral modeling (e.g., MANTRA)
Meta-analyze cohorts
• Modified random effects meta-analysis (e.g., RE2 model)

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 inspecition 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 λ_{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 λ_{GC} will scale with sample size when there is any inflation of the chi square statistics, it is sometimes reported after rescaling λ_{GC} to correspond to a study with 1000 cases and 1000 controls, known as λ_{1000}, to facilitate interpretation on a more standardized scale (de Bakker et al., 2008; Freedman et al., 2004). Since λ_{GC} and λ_{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. Ths 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 λ_{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 6×10^{-7} to 6×10^{-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_{Gg}) 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

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. *5*, 1564–1573.

Atkinson, E.G., Maihofer, A.X., Martin, A.R., Koenen, K.C., Neale, B.M., Nievergelt, C.M., and Daly, M.J. (2018). Association studies for all: A novel framework to allow for the well-calibrated genomic analysis of underrepresented admixed individuals. In American Society of Human Genetics,.

de Bakker, P.I.W., Ferreira, M.A.R., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. *17*, R122–R128.

Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., et al. (2019). Reduced signal for polygenic adaptation of height in UK Biobank.

Bromet, E., Andrade, L.H., Hwang, I., Sampson, N.A., Alonso, J., de Girolamo, G., de Graaf, R., Demyttenaere, K., Hu, C., Iwata, N., et al. (2011). Cross-national epidemiology of DSM-IV major depressive episode. BMC Med. *9*, 90.

Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes, Consortium, Ye, C.J., Price, A.L., and Zaitlen, N. (2016). Transethnic Genetic-Correlation Estimates from Summary Statistics. Am. J. Hum. Genet. *99*, 76–88.

Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.

Browning, S.R., and Browning, B.L. (2011). Population structure can inflate SNP-based heritability estimates. Am. J. Hum. Genet. *89*, 191–193; author reply 193–195.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. *103*, 338–348.

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

de Candia, T.R., Lee, S.H., Yang, J., Browning, B.L., Gejman, P.V., Levinson, D.F., Mowry, B.J., Hewitt, J.K., Goddard, M.E., O’Donovan, M.C., et al. (2013). Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. Am. J. Hum. Genet. *93*, 463–470.

Chen, C.-Y., Pollack, S., Hunter, D.J., Hirschhorn, J.N., Kraft, P., and Price, A.L. (2013). Improved ancestry inference using weights from external reference panels. Bioinformatics *29*, 1399–1406.

Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., et al. (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am. J. Hum. Genet. *98*, 653–666.

Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.

Conomos, M.P., Reiner, A.P., McPeek, M.S., and Thornton, T.A. (2018). Genome-Wide Control of Population Structure and Relatedness in Genetic Association Studies via Linear Mixed Models with Orthogonally Partitioned Structure (bioRxiv).

Das, S., Abecasis, G.R., and Browning, B.L. (2018). Genotype Imputation from Large Reference Panels. Annu. Rev. Genomics Hum. Genet. *19*, 73–96.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.

Duncan, L.E., Ratanatharathorn, A., Aiello, A.E., Almli, L.M., Amstadter, A.B., Ashley-Koch, A.E., Baker, D.G., Beckham, J.C., Bierut, L.J., Bisson, J., et al. (2018). Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Mol. Psychiatry *23*, 666–673.

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics *30*, 1266–1272.

Evans, L.M., Tahmasbi, R., Vrieze, S.I., Abecasis, G.R., Das, S., Gazal, S., Bjelland, D.W., de Candia, T.R., Haplotype Reference Consortium, Goddard, M.E., et al. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nat. Genet. *50*, 737–745.

Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., et al. (2004). Assessing the impact of population stratification on genetic association studies. Nat. Genet. *36*, 388–393.

Galinsky, K.J., Reshef, Y.A., Finucane, H.K., Loh, P.-R., Zaitlen, N., Patterson, N.J., Brown, B.C., and Price, A.L. (2019). Estimating cross-population genetic correlations of causal effect sizes. Genet. Epidemiol. *43*, 180–188.

Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am. J. Hum. Genet. *88*, 586–598.

Hao, W., and Storey, J.D. (2017). Extending Tests of Hardy-Weinberg Equilibrium to Structured Populations.

Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., Ekoru, K., Nsubuga, R.N., Ssenyomo, G., Kamali, A., et al. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc. Natl. Acad. Sci. U. S. A. *113*, 7377–7382.

Hellwege, J.N., Keaton, J.M., Giri, A., Gao, X., Velez Edwards, D.R., and Edwards, T.L. (2017). Population Stratification in Genetic Association Studies. Curr. Protoc. Hum. Genet. *95*, 1.22.1–1.22.23.

Henrich, J., Heine, S.J., and Norenzayan, A. (2010). The weirdest people in the world? Behav. Brain Sci. *33*, 61–83; discussion 83–135.

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

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Kanai, M., Tanaka, T., and Okada, Y. (2016). Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. J. Hum. Genet. *61*, 861–866.

Kendler, K.S., Aggen, S.H., Li, Y., Lewis, C.M., Breen, G., Boomsma, D.I., Bot, M., Penninx, B.W.J.H., and Flint, J. (2015). The similarity of the structure of DSM-IV criteria for major depression in depressed women from China, the United States and Europe. Psychol. Med. *45*, 1945–1954.

Lam, M., Awasthi, S., Watson, H.J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C.-C., De Witte, W., et al. (2019). RICOPILI: Rapid Imputation for COnsortias PIpeLIne.

Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. *34*, 591–602.

Lee, C.H., Eskin, E., and Han, B. (2017). Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. Bioinformatics *33*, i379–i388.

Lee, J.J., McGue, M., Iacono, W.G., and Chow, C.C. (2018). The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. Genet. Epidemiol. *42*, 783–795.

Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.

Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. *6*, 91.

Li, M.-X., Yeung, J.M.Y., Cherny, S.S., and Sham, P.C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum. Genet. *131*, 747–756.

Liu, E.Y., Li, M., Wang, W., and Li, Y. (2013a). MaCH-admix: genotype imputation for admixed populations. Genet. Epidemiol. *37*, 25–37.

Liu, J., Lewinger, J.P., Gilliland, F.D., Gauderman, W.J., and Conti, D.V. (2013b). Confounding and heterogeneity in genetic association studies with admixed populations. Am. J. Epidemiol. *177*, 351–360.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.

Loh, P.-R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.

Luo, Y., Li, X., Wang, X., Gazal, S., Mercader, J.M., 23andMe Research Team, SIGMA Type 2 Diabetes Consortium, Neale, B.M., Florez, J.C., Auton, A., et al. (2019). Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations.

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M.I., COGENT-Kidney Consortium, T2D-GENES Consortium, and Morris, A.P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum. Mol. Genet. *26*, 3639–3650.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

Marchini, J., Cardon, L.R., Phillips, M.S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. Nat. Genet. *36*, 512–517.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.

Medina-Gomez, C., Felix, J.F., Estrada, K., Peters, M.J., Herrera, L., Kruithof, C.J., Duijts, L., Hofman, A., van Duijn, C.M., Uitterlinden, A.G., et al. (2015). Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. Eur. J. Epidemiol. *30*, 317–330.

Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.

Ni, G., Moser, G., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Wray, N.R., and Lee, S.H. (2018). Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. Am. J. Hum. Genet. *102*, 1185–1194.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. *32*, 381–385.

Periyasamy, S., John, S., Padmavati, R., Rajendren, P., Thirunavukkarasu, P., Gratten, J., Vinkhuyzen, A., McRae, A., Holliday, E.G., Nyholt, D.R., et al. (2019). Association of Schizophrenia Risk With Disordered Niacin Metabolism in an Indian Genome-wide Association Study. JAMA Psychiatry.

Peterson, R.E., Edwards, A.C., Bacanu, S.-A., Dick, D.M., Kendler, K.S., and Webb, B.T. (2017). The utility of empirically assigning ancestry groups in cross-population genetic studies of addiction. Am. J. Addict. *26*, 494–501.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet. *83*, 132–135; author reply 135–139.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

van Rheenen, W., Peyrot, W.J., Schork, A.J., Lee, S.H., and Wray, N.R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. Nat. Rev. Genet.

Roshyara, N.R., Horn, K., Kirsten, H., Ahnert, P., and Scholz, M. (2016). Comparing performance of modern genotype imputation methods in different ethnicities. Sci. Rep. *6*, 34386.

Seldin, M.F., Pasaniuc, B., and Price, A.L. (2011). New approaches to disease mapping in admixed populations. Nat. Rev. Genet. *12*, 523–528.

Simon, G.E., Goldberg, D.P., Von Korff, M., and Ustün, T.B. (2002). Understanding cross-national differences in depression prevalence. Psychol. Med. *32*, 585–594.

Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife *8*.

Speed, D., Cai, N., UCLEB Consortium, Johnson, M.R., Nejentsev, S., and Balding, D.J. (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. *49*, 986–992.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program (bioRxiv).

Thornton, T., Tang, H., Hoffmann, T.J., Ochs-Balcom, H.M., Caan, B.J., and Risch, N. (2012). Estimating kinship in admixed populations. Am. J. Hum. Genet. *91*, 122–138.

Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G., et al. (2011). Quality control procedures for genome-wide association studies. Curr. Protoc. Hum. Genet. *Chapter 1*, Unit1.19.

Uebelacker, L.A., Strong, D., Weinstock, L.M., and Miller, I.W. (2009). Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. Psychol. Med. *39*, 591–601.

Vijver, F. van de, van de Vijver, F., and Tanzer, N.K. (2004). Bias and equivalence in cross-cultural assessment: an overview. Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology *54*, 119–135.

Visscher, P.M., Hemani, G., Vinkhuyzen, A.A.E., Chen, G.-B., Lee, S.H., Wray, N.R., Goddard, M.E., and Yang, J. (2014). Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. PLoS Genet. *10*, e1004269.

Walters, R.K., Polimanti, R., Johnson, E.C., McClintick, J.N., Adams, M.J., Adkins, A.E., Aliev, F., Bacanu, S.-A., Batzler, A., Bertelsen, S., et al. (2018). Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat. Neurosci. *21*, 1656–1669.

Wang, X., Chua, H.-X., Chen, P., Ong, R.T.-H., Sim, X., Zhang, W., Takeuchi, F., Liu, X., Khor, C.-C., Tay, W.-T., et al. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. Hum. Mol. Genet. *22*, 2303–2311.

Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E., Mägi, R., Ferreira, T., Fall, T., Graff, M., Justice, A.E., et al. (2014). Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. *9*, 1192–1212.

Wojcik, G.L., Fuchsberger, C., Taliun, D., Welch, R., Martin, A.R., Shringarpure, S., Carlson, C.S., Abecasis, G., Kang, H.M., Boehnke, M., et al. (2018). Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies. G3 *8*, 3255–3267.

Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.

Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A comparison of association methods correcting for population stratification in case-control studies. Ann. Hum. Genet. *75*, 418–427.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011a). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.

Yang, J., Zeng, J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2017). Concepts, estimation and interpretation of SNP-based heritability. Nat. Genet. *49*, 1304–1310.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. *9*, e1003520.

Zaitlen, N., Pasaniuc, B., Sankararaman, S., Bhatia, G., Zhang, J., Gusev, A., Young, T., Tandon, A., Pollack, S., Vilhjálmsson, B.J., et al. (2014). Leveraging population admixture to characterize the heritability of complex traits. Nat. Genet. *46*, 1356–1362.

Zhang, Y., and Pan, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? Genet. Epidemiol. *39*, 149–155.

Zheng, J., Li, Y., Abecasis, G.R., and Scheet, P. (2011). A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet. Epidemiol. *35*, 102–110.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics *44*, 821–824.

Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.