Supplementary Material of "A data harmonization pipeline to leverage external controls and boost power in GWAS"

Danfeng Chen[1], Katherine Tashman[2,3], Duncan S. Palmer[2,3], Benjamin Neale[2,3,4], Kathryn Roeder[5], Alex Bloemendal[2,3], Claire Churchhouse[2,3,4,*] and Zheng Tracy Ke[6,*]

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, 08544, New Jersey, United States

[2] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, 02114, Massachusetts, United States

[3] Stanley Center for Psychiatric Research, Broad Institute of of MIT and Harvard, Cambridge, 02142, Massachusetts, United States

[4] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, 02142, Massachusetts, United States

[5] Department of Statistics, Carnegie Mellon University, Pittsburgh, 15213, Pennsylvania, United States

[6] Department of Statistics, Harvard University, Cambridge, 02138, Massachusetts, United States

Appendix A: The 16 genotype data collections in the harmonized control dataset

In the Results section, we created a resource of external controls using genotype data taken from 27,517 individuals of self-reported European-descent in 16 studies in the dbGaP repository. Information regarding these cohorts can be found in Table S1.

Table S1. The 16 studies used for creating a resource of external controls, taken from the dpGAP repository.

Study	dbGaP Accession Number	Ancestry	#Samples	Genotyping Chip
Late Onset Alzheimer's Disease Family Study: Genome-Wide Association Study for Susceptibility Loci	phs000168.v2.p2	European	1292	HumanHap 610
GENEVA Genes and Environment Initiatives in Type 2 Diabetes	phs000091.v2.p1	European	3148	Affymetrix 6.0
National Eye Institute Glaucoma Human Genetics Collaboration (NEIGHBOR) Consortium Glaucoma Genome-Wide Association Study	phs000238.v1.p1	European	1240	HumanHap 610
A Study of the Genetic Causes of Complex Pediatric Disorders	phs000490.v1.p1	European	2631	HumanHap 610/ HumanHap 550
Genome-Wide Association Study of Schizophrenia	phs000021.v3.p2	African/ European	2332	Affymetrix 6.0
Neurodevelopmental Genomics: Trajectories of Complex Phenotypes	phs000607.v2.p2	European	5715	HumanHap 610/ HumanHap 550
Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing	phs000473.v2.p2	European	773	HumanHap550
National Human Genome Research Institute (NHGRI) GENEVA Genome-Wide Association Study of Venous Thrombosis	phs000289.v2.p1	European	1310	HumanHap 610
Whole Genome Scan for Pancreatic Cancer Risk in the Pancreatic Cancer CohortConsortium and Pancreatic Cancer Case-Control Consortium (PanScan)	phs000206.v5.p3	European	4133	HumanHap 610/ HumanHap 550
Research Program on Genes, Environment and Health (RPGEH)	phs000788.v1.p2	European	10000	Axiom KP
NIH Exome Sequencing of Familial Amyotrophic Lateral Sclerosis Project	phs000101.v4.p1	European	773	HumanHap 550

Appendix B: The complete QC Pipeline for data harmonization

We give a step-by-step description of the QC pipeline. All the details are summarized in Figure S1.

First, we conducted pre-processing:

(1) All cohorts were lifted over from hg18 to hg19 using the liftOver tool. Marker names across all cohorts were made consistent with the HRC reference panel based on chromosome, position, and alleles. Indels and missing alleles, duplicated markers, and non-matching SNPs were excluded.

Next, we conducted cohort-level pre-imputation QC:

(2) *Sample QC*: Samples with missing rate greater than 2% or with mismatching sex information between genotypes and provided phenotype information were excluded. Samples with abnormal inbreeding coefficients (3 standard deviations away from the mean) were removed. Related samples (pi-hat > 0.0625) were identified and the sample in each related pair with higher missingness was removed from the dataset.

(3) *Variant QC*: Variant-level missingness was calculated and variants with missing rate > 1% were excluded. Additionally, all A/T and C/G SNPs were excluded at this point to avoid downstream issues related to strandness.

(4) *Ancestry Inference*: We built a random forest classifier (see below for details) with the first six PCs from 1000 genomes to determine broad ancestry grouping. We then further separate population isolates from major European within the European ancestry cluster by building a second random forest classifier using the first four PCs calculated from a merged set of the 1000 genomes Europeans and controls from an Ashkenazi Jewish cohort. We then projected each sample onto the 1000 genomes PC space, and assigned broad ancestry-group labels using the first random forest classifier. Each sample was assigned to one continental ancestry group. A similar procedure was applied to samples assigned a European ancestry label to determine its finer

ancestral origin. Within each ancestral group, variants with minor allele frequencies lower than 0.01 or with p-values for Hardy-Weinberg Equilibrium test less than 1e-4 were excluded.

Random forest classifier details: We used the following Random Forest algorithm to perform ancestry assignment. We first assigned individuals into one of five super-populations: East Asian, South Asian, African, American, and European. To do so, we first calculated the PCs of the 1000 Genomes dataset. We then apply RandomForestClassifier from scipy package and train a random forest classifier with 100 bootstrap draws using the first six PCs as input. All parameters of the random forest training function were set at their defaults. We then used the output classifier to predict the population label of each individual. For individuals that are assigned as European, we further assigned them into major European Ancestry, Finnish Ancestry and Ashkenazi Jewish Ancestry. To do this, we merged 1000 Genomes European genotype data with genotype data from an Ashkenazi Jewish cohort as there are no Ashkenazi Jewish individuals among the 1000 genomes samples. We then evaluate PCs in this merged dataset. Non-Finnish Europeans are assigned the *Major European* population group label. The first four PCs are used to train another the random forest model. Again, all parameters of the random forest training function in RandomForestClassifier are set at their defaults. The resultant classifier was used to assign finer structure in the European subset.

Next, we conducted array-level pre-imputation QC:

(5) *Matching*: Cohorts from the same genotyping array and ancestry-group were merged together for downstream analysis.

(6) *Sample QC*: Samples with missing rate greater than 2% and the member of related pairs (pi-hat > 0.0625) of samples with higher missingness was removed.

(7) *Variant QC*: Within each array group, variants with minor allele frequencies lower than 0.01 or with p-values for a Hardy-Weinberg Equilibrium test less than 1e-4 were excluded. To further

identify variants susceptible to batch effects, a pseudo-GWAS comparison labelling array samples as cases and 1000 Genomes samples as controls was performed. Variants with p-value less than 1e-4 were excluded. Further pseudo-GWASes defeined by labelling samples from one cohort as cases and samples from all other cohorts as controls were carried out, and variants with p-values less than 1e-4 were also excluded. For pseudo-GWAS comparison, the first 20 PCs were included as covariates to control for population stratification.

The next step is imputation:

(8) We imputed each stratum separately using 1000 genomes as a reference panel via the Michigan Imputation server.

Then, we have a few steps of post-imputation QC: multiple sources, and enables external controls to be reliably integrated in GWAS.

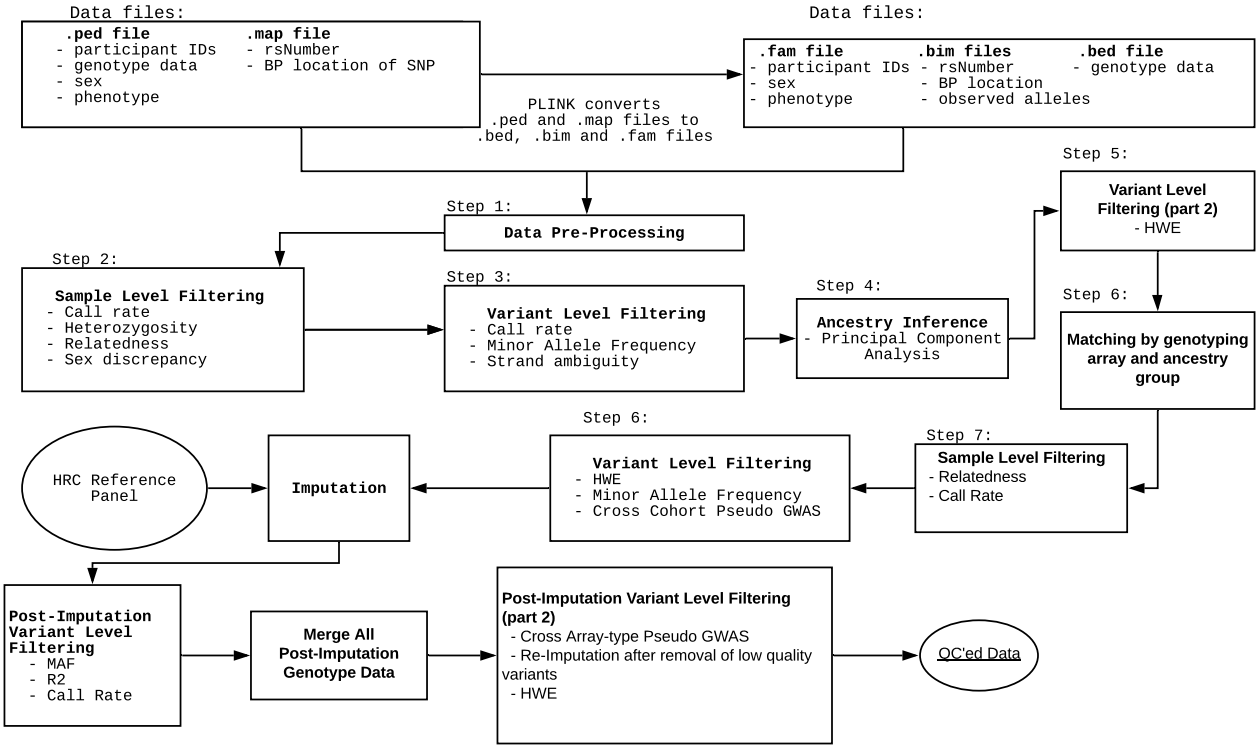
(9) *QC each array stratum*: After imputation, 46.8M variants had been imputed in each array stratum. Variants with minor allele frequencies < 0.01, or Hardy-Weinberg Equilibrium p-value < 1e-4, or an imputation info score < 0.8 were removed from the data.

(10) *Inner-join*: A master imputed genotype dataset was generated through an inner join (i.e., taking the intersection of SNPs) across all stratum.

(11) *Cross-array comparison*: For each genotyping array, samples that were genotyped on that array were coded as cases and samples that were genotyped on any other array were coded as controls to make a pseudo-case control comparison. To run association testing, we added first 20 PCs as covariates and drop variants found to be associated with any given array with p-value < 1e-4.

(12) *Re-imputation*: We removed variants with empRsq < 0.6 from the typed data, and re-imputed them back so as to increase the number of SNPs as well as to improve the quality. The re-imputation procedure repeats steps (8)-(11).

Figure S1. The complete QC pipeline.



Appendix C: The QC pipeline on Cohron's disease data

We conducted GWAS on five Crohn's disease genetics data sets from the IBD Genetics Consortium, to evaluate the quality of the collection of harmonized controls. We applied a QC pipeline to these five data sets. This QC pipeline is a subset of the data harmonization pipeline in Appendix A, where we skip the steps for cross-array comparison and re-imputation.

We now describe the QC pipeline on Crohn's disease data.%It is also the QC pipeline we provide to users in the UNICORN framework:

- We first conducted data pre-processing to align cases to hg19 Genome Build. This is the same as step (1) of the data harmonization pipeline in Appendix A.

- Next, we conducted cohort-level pre-imputation QC:
    - *Sample QC*: We filtered out samples with high missing rate, mismatching sex, abnormal inbreeding coefficients and related individuals. This is the same as step (2) of the data harmonization pipeline.
    - *Variant QC*: We filtered out high missing rate and strand ambiguous SNPs. This is the same as step (3) of the data harmonization pipeline.
    - *Ancestry Inference*: In step (4) of the data harmonization pipeline, we built a Random Forest Classifier for classifying any sample into one of the five super populations: East Asian, South Asian, African, American, and European. We applied this classifier to assign each sample into a super population. In step (4) of the data harmonization pipeline, we also built a Random Forest Classifier to classify any European sample into one of the three groups, European Mainland, Finnish, and Ashkenazi Jewish. We applied this classifier to further divide those samples assigned to the European super population.
    - We removed the ‘blacklist’ of SNPs obtained in applying the harmonization pipeline on the

public controls.

- We then imputed case cohort against 1000 Genomes panel. This is the same as step (8) of the data harmonization pipeline.
- Lastly, we performed post-imputation QC: We remove variants out of Hardy-Weinberg Equilibrium with low minor alleles frequency, or with low imputation info score. This is the same as step (9) of the data harmonization pipeline.

After the above QC pipeline is completed, if the GWAS results still have many suspicious signals, the user can add a further re-imputation step, the same as step (12) in the data harmonization pipeline. In the case of the Crohn's disease data, we do not require re-imputation.