Accurate cross-platform GWAS analysis via two-stage imputation

Anya Greenberg[1,2], Kaylia Reynolds[1,2], Michelle T. McNulty[1,2], Matthew G. Sampson[1,2,3,4], Hyun Min Kang[5], Dongwon Lee[1,2,3,6,#]

Affiliations

1. Division of Pediatric Nephrology, Boston Children's Hospital, Boston, MA, USA
2. Kidney Disease Initiative and Medical Population Genetics Group, Broad Institute, Cambridge, MA, USA
3. Department of Pediatrics, Harvard Medical School, Boston, MA, USA
4. Division of Renal Medicine, Brigham and Women's Hospital, Boston, MA, USA
5. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
6. Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, USA

[#] *Send all correspondence to:*

Dongwon Lee, Ph.D.

Division of Nephrology

The Manton Center for Orphan Disease Research

Boston Children's Hospital & Harvard Medical School

300 Longwood Ave, Enders 505

Boston, MA 02115, USA

(T) 617-919-7691

(E) dongwon.lee@childrens.harvard.edu

Abstract

In genome-wide association studies (GWAS), combining independent case-control cohorts has been successful in increasing power for meta and joint analyses. This success sparked interest in extending this strategy to GWAS of rare and common diseases using existing cases and external controls. However, heterogeneous genotyping data can cause spurious results. To harmonize data, we propose a new method, two-stage imputation (TSIM), where cohorts are imputed separately, merged on intersecting high-quality variants, and imputed again. We show that TSIM minimizes cohort-specific bias while controlling imputation-derived errors. Merging arthritis cases and UK Biobank controls using TSIM, we replicated known associations without introducing false positives. Furthermore, GWAS using TSIM performed comparably to the meta-analysis of nephrotic syndrome cohorts genotyped on five different platforms, demonstrating TSIM's ability to harmonize heterogeneous genotyping data. With the plethora of publicly available genotypes, TSIM provides a GWAS framework that harmonizes heterogeneous data, enabling analysis of small and case-only cohorts.

Introduction

Genome-wide association studies (GWAS) are common for studying the genetic determinants of traits and diseases, especially when large sample sizes are available. Smaller studies or case-only cohorts are often combined with external controls to enable GWAS and increase statistical power[1–3]. When combining cohorts, batch effects can arise due to differences in genotyping platform. To reduce these effects, the current practice for combining cohorts prior to imputation is to use genotyped single nucleotide polymorphisms (SNPs) that are shared between cohorts[1,4–7]. This approach not only reduces the number of SNPs available for analysis but may also adversely affect genotype imputation by decreasing accuracy.

Genotype imputation approximates whole genome sequencing (WGS) by leveraging linkage disequilibrium (LD)[8,9]. It is an essential step in GWAS to increase the genome-wide coverage of the analysis and enable post-GWAS analyses such as fine mapping and meta-analysis[1,9]. By restricting the SNPs used for imputation to the intersection, haplotype-informative SNPs may be lost, thus reducing the accuracy of imputation and downstream analyses. Given this, it is currently recommended to only combine cohorts genotyped on the same platform prior to imputation[4,5]. Finding external cohorts that meet these criteria is often difficult, especially for underrepresented populations. While it is becoming easier to find large external control cohorts with the advent of diverse biobanks, they are often genotyped on custom platforms designed for their population of interest[10,11] making it difficult to combine them with other cohorts.

Previously, several approaches have been proposed to overcome the challenges in combining heterogeneous genotyping data. Some studies argued that cohorts can be imputed separately, and

then combined by restricting analysis to the shared SNPs with high imputation quality[2,12]. This approach reduced type I error in GWAS at the expense of reduced power and fewer SNPs for testing. GAWMerge presented another solution to this problem: merge the array-based genotypes of a study cohort with the WGS of publicly available controls[13]. However, this method requires the external cohort to have WGS available which may not be feasible for many underrepresented populations, particularly in regions with limited funding and resources. Another study sought to harmonize multiple control cohorts by removing low-quality genotyped SNPs, then merging genotypes after a single imputation[7]. These SNPs were identified through an iterative process of GWAS using genotyping platform as the outcome, to identify spurious genome-wide significant loci, and re-imputation on array genotypes, to identify low-quality SNPs. This approach showed promise in reducing batch effects when merging multiple cohorts. However, it requires internal controls genotyped on the same platform as cases. It may also result in substantial loss of imputed SNPs after filtering out low-quality variants from re-imputation.

Here, we introduce two-stage imputation (TSIM), a method to address these issues, which includes a second imputation on high-quality SNPs after merging separately imputed cohorts. In brief, cohorts are imputed separately (Stage 1) and then merged on high-quality genotyped and imputed SNPs (hq-SNPs) present in both imputations. Finally, a second imputation (Stage 2) is performed on the merged data. Essentially, TSIM utilizes a first stage of high-quality imputation to improve a second stage of imputation which covers a larger proportion of the genome. Similar strategies have been used to improve genotype imputation of ancient DNA[14] and rare variants[15], but no studies have applied this concept to GWAS of common variants. We illustrate the validity and utility of TSIM using data from the 1000 Genomes Project (1KGP)[16], UK Biobank (UKB)[10],

a psoriatic arthritis case-control cohort[17], and two cohorts from a published pediatric steroid sensitive nephrotic syndrome (pSSNS) meta-analysis[18]. Along with this research, we provide a command line tool, tsim, for others to easily implement this method in their analyses.

Results

Overview of two-stage imputation method (TSIM)

TSIM implements two primary steps bookended by two rounds of imputation (**Fig 1**). In the first round of imputation (Stage 1), we impute cohorts separately using the same reference panel, following current GWAS practices for quality control (QC) of a single cohort. This Stage 1 imputation serves to increase the number of intersecting SNPs for merging and reduce cohort-specific bias by identifying high-quality genotyped and imputed SNPs (hq-SNPs) present in all imputed cohorts. We define hq-SNPs as common SNPs (minor allele frequency (MAF) ≥ 0.01) with high imputation quality ($R^2 \geq 0.99$; $ER^2 \geq 0.9$ if genotyped). We discuss the determination of these thresholds in detail below (also see ***Methods***). This step can be considered as *"in silico* genotyping" simulating the "same platform" for all cohorts to be merged. After Stage 1, cohorts are merged on the intersection of hq-SNPs. Then, we perform the second round of imputation (Stage 2) on the merged cohort to achieve greater genome-wide coverage for GWAS and downstream analyses.

Determination of high-quality SNPs

We evaluated the quality of Stage 1 imputation results to determine what criteria were needed to define our hq-SNPs. We used unrelated individuals from 1KGP across four ancestry groups (European (EUR), African (AFR), East Asian (EAS), and South Asian (SAS)) that were

genotyped on two different array platforms, Affymetrix SNP 6.0 (AFFY) and HumanOmni2.5 (OMNI), and had WGS. We imputed AFFY and OMNI genotypes using the TOPMed Imputation Server and its associated reference panel (TOPMed)[19–21]. Within each ancestry group and for each platform, we identified imputation-derived errors based on the absolute difference in allele frequency (AF_{diff}) between Stage 1 imputed dosages and the corresponding WGS genotypes, which we considered the ground truth. We also calculated AF_{diff} between WGS and "good-quality" genotyped SNPs ($ER^2 \geq 0.9$) for comparison. We found that SNPs with $R^2 \geq 0.99$ generally had the lowest AF_{diff} from WGS and were better correlated with the unimputed vs. WGS line across all ancestry groups for both platforms. The other R^2 bins followed in descending order indicating that SNPs with higher R^2 had lower discordance with WGS (**Fig 2A,B**; **Supp Table 1**).

Using a similar framework, we also evaluated cohort-specific bias using samples genotyped on both AFFY and OMNI. Here, we define "cohort-specific bias" as differences resulting from mismatched genotyping platforms, but the term encompasses all technical differences between the cohorts being merged. Within each ancestry group, we calculated the AF_{diff} between the AFFY- and OMNI-imputed results. As expected, we observed a strong correlation between imputation quality and cohort-specific bias, with minimal bias achieved when $R^2 \geq 0.99$ (**Fig 2C**). Despite this high threshold, we were able to retain over 3,000,000 SNPs for both EUR and AFR, and around 1,000,000 for EAS and SAS (**Supp Table 1**). The fewer number of SNPs for EAS and SAS can be partially attributed to the limitations of the TOPMed reference panel, where only 8% of samples have Asian ancestry (both East and South)[19], and demonstrates the importance of a complementary reference for imputation. These results suggest that an $R^2 \geq 0.99$

cutoff is sufficient for AFR and EUR cohorts imputed on TOPMed, but that an alternate reference panel with more Asian representation may be beneficial for EAS and SAS cohorts.

TSIM is robust against imputation-derived errors

We assessed the effect of TSIM on imputation-derived error by extending the previous analysis to Stage 2 imputations. With the 1KGP samples used previously, we compared AFFY- and OMNI-based imputations to WGS genotypes using AF_{diff}. Here, we use "S1" and "S2" to represent imputation results following Stage 1 and Stage 2, respectively. S1 is more similar to current practices of imputing once whereas S2 implements TSIM and imputes data twice. Generally, for both platforms across ancestries, the AF_{diff} is similar in S1 and S2 for SNPs with higher imputation quality ($R^2 \geq 0.8$) and worse (i.e., higher) in S2 for SNPs with lower imputation quality ($R^2 < 0.6$). For R^2 bins in between, AF_{diff} progressively increases as imputation quality decreases (**Supp Fig 1**). For example, the 514 EUR_{AFFY} samples with WGS had 2.10% of SNPs (7,064,852) with $AF_{diff} \geq 1\%$ in S1, decreasing to 1.98% in S2. Looking at a more relaxed threshold of $AF_{diff} \geq 2\%$, S1 had 0.39% of SNPs meeting this criterion whereas S2 slightly increased to 0.56% (**Supp Table 2**). Similar marginal differences were observed in AFR_{AFFY}, AFR_{OMNI}, and EUR_{OMNI} samples (**Supp Table 2,3**). However, for EAS and SAS, S2 had higher AF_{diff} indicating S2 was more discordant from WGS than S1 (**Supp Fig 1**). Stratifying these calculations by imputation quality, the higher AF_{diff} in S2 was primarily driven by SNPs with low imputation quality (**Supp Fig 1; Supp Table 2,3**). These low-quality SNPs ($R^2 < 0.6$) account for a small percentage of total SNPs (EUR_{AFFY}=2%, EUR_{OMNI}=2%, AFR_{AFFY}=2%, AFR_{OMNI}=2%, EAS_{AFFY}=11%, EAS_{OMNI}=12%, SAS_{AFFY}=12%, SAS_{OMNI}=11%) and can be removed (**Supp Table 2,3**). The increase for EAS and SAS is likely due to the limitations of the

TOPMed reference panel mentioned before. Thus, despite using a second stage of imputation where new errors might be introduced, TSIM produces comparable imputation-derived error to using a single stage of imputation when poor imputation quality is accounted for and the reference panel is well-matched.

TSIM shows substantially reduced cohort-specific bias

We next evaluated the effect of TSIM on cohort-specific bias. In these analyses, we refer to a method of separately imputing cohorts once (Stage 1), followed by simple merging on intersecting SNPs, as "*separately-imputed*" and our new method implementing TSIM as "*two-stage*". Thus, we have four imputed dosage datasets for this analysis: Stage 1 and Stage 2 imputed dosages for each of AFFY and OMNI. We then calculated the AF_{diff} for each variant between the AFFY- and OMNI-imputed dosages. We further stratified these per variant calculations based on the average imputation quality for each stage across datasets (see ***Methods***). For EUR samples, 1.49% of SNPs (8,693,470) had $AF_{diff} \geq 1\%$ in *separately-imputed*. This bias was ameliorated in *two-stage* with only 0.03% of SNPs with $AF_{diff} \geq 1\%$ (**Fig 3; Supp Table 4**). The substantially smaller AF_{diff} between cohorts in *two-stage* illustrates that cohort-specific bias is greatly reduced after the second stage of imputation. This pattern is most pronounced for SNPs with lower imputation quality (**Supp Table 4**; **Supp Fig 2**). We saw similar results when analyzing other 1KGP ancestries (AFR: samples=328, SNPs=14,383,283, *separately-imputed*=1.46%, *two-stage*=0.07%; EAS: samples=464, SNPs=6,899,281, *separately-imputed*=5.85%, *two-stage*=0.63%; SAS: samples=100, SNPs=7,325,035, *separately-imputed*=7.13%, *two-stage*=3.02%) (**Fig 3**; **Supp Fig 2; Supp Table 4**). This demonstrates that TSIM successfully reduces cohort-specific biases between cohorts that were genotyped on

different genotyping platforms, an issue that has long been a challenge in robust GWAS analysis[3–6].

TSIM robustly controls for type I error

To evaluate type I error we performed 1KGP vs. UKB GWAS for each ancestry group (AFR, EAS, EUR, and SAS). For each ancestry, we randomly selected up to 5,000 ancestry-matched individuals from UKB, genotyped on UKB Axiom, to be merged with unrelated samples in 1KGP (**Supp Table 5**). Notably, only 90,530 of UKB Axiom SNPs (total= 784,256) are shared with AFFY (total= 905,788) and only 273,045 are shared with OMNI (total= 2,458,861) before QC. We performed GWAS using SAIGE[22] with two different methods of merging data: (1) a simple merging of separately imputed data on intersecting SNPs ("*separately-imputed*"), and (2) a full implementation of TSIM ("*TSIM*") (see ***Methods***). Since 1KGP samples were genotyped on two different platforms (AFFY and OMNI), we assessed both data separately. Principal component analysis (PCA) showed that the genetic backgrounds of the two different cohorts were well matched for all ancestry groups (**Supp Fig 3**). Because no specific phenotypes are enriched in either 1KGP or UKB (i.e., control cohorts), we would expect no genome-wide significant loci (GWAS loci) from this analysis. Thus, any GWAS loci we found would be false positives, indicating a cohort-specific bias in the analysis. In all GWAS, the *separately-imputed* method had many GWAS loci whereas *TSIM* had little to none (**Fig 4**; **Supp Fig 4**; **Supp Table 5**). Specifically, for the EUR GWAS using 1KGP-AFFY and the matched UKB subset, 170 independent GWAS loci in *separately-imputed* were reduced to none in *TSIM*; using OMNI, 99 in *separately-imputed* were reduced to one barely passing genome-wide significance in *TSIM* ($p=3.85\times10^{-8}$, $R^2=0.54$). Similarly, for AFR, over 500 GWAS loci in *separately-imputed* were

reduced to one for AFFY ($p=2.56\times10^{-8}$, $R^2=0.59$) and 403 became two for OMNI ($p=1.61\times10^{-8}$, 6.58×10^{-9}; $R^2=0.66, 0.33$). We saw similar patterns for EAS and SAS. Of note, the total number of SNPs in the *TSIM* GWAS was reduced for EAS and SAS ($EAS_{AFFY\text{-}UKB}$=3,033,659; $EAS_{OMNI\text{-}UKB}$ =5,646,177; $SAS_{AFFY\text{-}UKB}$=3,380,760; $SAS_{OMNI\text{-}UKB}$ =6,340,030). This is likely due to the lack of sufficient genome-wide coverage of hq-SNPs in Stage 1 imputation which negatively impacted Stage 2 imputation results.

We also evaluated the GWAS of UKB samples vs. the control group (n=1,406) from a GWAS of psoriatic arthritis ("ART"), genotyped on the HumanOmni1-Quad BeadChip[17]. We used a random subset of 5,000 UKB individuals ("UKB5K") of EUR descent and UKB individuals with psoriatic arthritis were removed (see ***Methods***). ART controls were merged with UKB5K after separate imputations and using TSIM. In the *TSIM* GWAS, no variants reached genome-wide significance (**Supp Fig 5**, top), while the *separately-imputed* GWAS yielded over 500 GWAS loci (**Supp Fig 5**, bottom). This drastic decrease in GWAS loci between the *separately-imputed* and *TSIM* GWAS, and the noticeable lack of associations in *TSIM* demonstrates that our method effectively controls for type I error and batch effects across different genotyping platforms, provided there is sufficient genome-wide coverage of hq-SNPs.

TSIM replicates psoriatic arthritis GWAS results and enables the use of case-only cohorts

Using the ART cohort, which contained 1,410 psoriatic arthritis cases in addition to the 1,406 controls, and UKB5K, we also evaluated TSIM's ability to accurately identify GWAS loci in a case-control study. We compared GWAS results from *TSIM* and *separately-imputed* for two different scenarios: (1) *both-controls*: ART cases and controls merged with UKB5K to represent

a common situation where external controls may be used to increase the power of GWAS and (2) *external-controls*: only cases from ART merged with UKB5K to represent a situation where only cases were available and merging with external cohorts would enable GWAS to be performed. We also ran GWAS using only the ART cases and controls, referred to as *internal-controls*, to generate reference results to better compare GWAS loci (see ***Methods***). Of note, one case and four controls were removed from the *internal-controls* GWAS due to lack of matching controls and cases, respectively, and were recovered when ART was merged with UKB5K using TSIM (**Supp Table 6**). PCA and QQ plots for these GWAS are available in **Supp Fig 6,7**. We found that for both *both-controls* and *external-controls* scenarios, the TSIM GWAS was better able to replicate the *internal-controls* GWAS than the *separately-imputed* GWAS (**Fig 5A**; **Supp Fig 8,9**). Specifically, in the *both-controls* GWAS, three out of four genome-wide significant loci from the *internal-controls* GWAS had similar λ_{GC}-adjusted p-values, odds ratios, and directions of effect (*HLA-B* – rs36058333, rs4418214; *TRAF3IP2* – rs33980500; *IL12B* – rs918520, rs1582515) with the fourth reaching suggestive significance (*TYK2* – rs11085727 ($p_{\lambda GC\text{-}adj.}=8.84\times10^{-8}$), rs35251378 ($p_{\lambda GC\text{-}adj.}=1.10\times10^{-7}$)) (**Supp Table 7**). The *both-controls* GWAS also replicated and improved results for some suggestive loci by bringing them over the genome-wide significance threshold (*HLA-G* – rs3115628 ($p_{\lambda GC\text{-}adj.}=5.80\times10^{-9}$); *TNIP1* – rs75851973 ($p_{\lambda GC\text{-}adj.}=4.66\times10^{-14}$), rs76956521 ($p_{\lambda GC\text{-}adj.}=4.62\times10^{-14}$), rs17728338 ($p_{\lambda GC\text{-}adj.}=1.08\times10^{-13}$)) (**Fig 5B**; **Supp Fig 9; Supp Table 7**). These suggestive loci were implicated in previously published psoriatic arthritis GWAS and many had follow-up analyses further connecting the loci to the disease risk[17,23–26]. Similarly, the *external-controls* GWAS replicated three out of four signals from the *internal-controls* GWAS (*HLA-B* – rs36058333, rs4418214; *TRAF3IP2* – rs33980500; *IL12B* – rs1582515) and the other reached suggestive significance (*IL12B* – rs918520 ($p_{\lambda GC\text{-}}$

$_{adj.}$=5.34×10^{-7}); *TYK2* – rs11085727 ($p_{\lambda GC\text{-}adj.}$=3.92×10^{-6}), rs35251378 ($p_{\lambda GC\text{-}adj.}$=4.84×10^{-6})). We also found improved results for one suggestive loci (*TNIP1* – rs75851973 ($p_{\lambda GC\text{-}adj.}$=4.66×10^{-14}), rs76956521 ($p_{\lambda GC\text{-}adj.}$=4.62×10^{-14}), rs17728338 ($p_{\lambda GC\text{-}adj.}$=1.08×10^{-13})) (**Fig 5B**; **Supp Fig 9; Supp Table 7**). On the other hand, *separately-imputed* resulted in many false positives, similar to the previous 1KGP vs. UKB control GWAS (**Fig 4**,**5A**; **Supp Fig 4,5**). These GWAS demonstrate that our TSIM method can accurately replicate and improve GWAS results by enabling the inclusion of external controls to increase the power of detection. It is especially effective for case-only cohorts for which GWAS analysis was previously impossible due to a lack of controls matching both genetic population structure and genotyping platform.

TSIM facilitates standard GWAS of multi-platform cohorts at the genotype level

Next, we evaluated whether a single joint GWAS of two European pSSNS cohorts using TSIM could accurately replicate results from a meta-analysis on the same data conducted by Barry et al.[18]. In the published analysis, these two cohorts (designated “EU” and “US” based on the geographical origin of each dataset) were meta-analyzed because they were genotyped on two distinct array platforms across five different sub-cohorts (**Supp Table 8**). When merged on hq-SNPs after the first stage of imputation, we found that the two cohorts clustered together in the PCA space (**Supp Fig 10**), indicating that the samples had similar genetic population structure and might benefit from being analyzed together in a single TSIM Joint GWAS (“*TSIM Joint*”). We also ran a TSIM meta-analysis (“*TSIM Meta*”), using TSIM to merge the four sub-studies comprising the EU cohort, to assess the impact of using TSIM to merge smaller cohorts. In total, there were 674 cases (EU=313, US=361) and 6,817 controls (EU=2,508, US=4,309). The TSIM GWAS included more SNPs (Joint=8,531,980, Meta=8,209,112) compared to the 8,014,298

included in the published GWAS and had similar results (**Fig 6A**; **Supp Fig 11**; **Supp Table 9,10**). In both *TSIM Joint* and *TSIM Meta*, we replicated three out of four GWAS loci from the published meta-analysis (*HLA-DQB1* – rs17211699; *NFKBIL1* – rs2857607; *CALHM6* – rs2637678) (**Supp Table 10**). Using TSIM, we were able to identify eight additional SNPs in high LD with rs2637678 in the *CALHM6* locus (**Fig 6B**) compared to 32 SNPs in the published analysis. The fourth published GWAS loci (*MORF4L1* – rs12911841) failed to achieve genome-wide significance after correcting for genomic control (**Supp Table 10**). This SNP had low imputation quality ($R^2 < 0.6$) and the effect allele (T) frequency is 0.007 in Europeans[27]. Considering that the SNP is rare and has low imputation quality in our study population, we cannot conclude if TSIM shows an improvement by removing a false positive or if we simply lack the power to detect this association with this dataset. Nevertheless, TSIM provides a robust analysis of this multi-cohort pSSNS study. Furthermore, the low inflation and high concordance of *TSIM Joint* and *TSIM Meta* results compared to the published GWAS demonstrate that TSIM is an effective strategy for combining multiple cohorts and may be used as an alternative to meta-analysis for smaller cohorts, provided cohorts have similar genetic population structure.

Discussion

In summary, we present TSIM, a method to extend the applicability of GWAS to previously understudied or underpowered cohorts. TSIM consists of two primary steps for efficiently harmonizing separately imputed cohorts bookended by two stages of imputation: (1) identify hq-SNPs and (2) merge cohorts based on hq-SNPs present in all cohorts. The identification of hq-SNPs corrects cohort-specific bias and the second stage of imputation increases the number of SNPs available for GWAS and downstream analyses. The development and evaluation of TSIM

was done with data from 1KGP, UKB, a psoriatic arthritis case-control GWAS, and a pSSNS case-control GWAS meta-analysis. We evaluated cohort-specific bias, imputation-derived error, and GWAS results when implementing TSIM for two cohorts in our psoriatic arthritis analysis (ART and UKB) and for multiple cohorts in our pSSNS analysis (four in TSIM EU, five in TSIM Joint). In our validation, we showed that TSIM is an effective method to reduce cohort-specific bias without increasing imputation-derived error. Thus, TSIM can effectively harmonize heterogeneous genotyping data. However, it cannot account for sample heterogeneity arising from differences in genetic population structure and other relevant phenotypic characteristics such as sex, age, and potential covariates for the phenotype of interest[1,3]. Similarly, in order to use TSIM for continuous phenotypes, robust harmonization of phenotype measurements must be performed separately.

TSIM addresses many issues with current practices for harmonizing genotype data which are insufficient and impractical for many datasets. In current practices, many SNPs are often removed when merging on the intersection and many samples may be dropped due to lack of matched controls. By merging separately imputed cohorts, TSIM enables the aggregation of case-only cohorts with external controls. TSIM can even recover cases for inclusion in GWAS that may not have had well-matched controls in a case-control study cohort, as we saw when analyzing the ART cohort with UKB. TSIM also shows potential for combining datasets, at the genotype level, from different study centers to conduct a single joint GWAS analysis which performs just as well as a meta-analysis, as demonstrated with our pSSNS analysis. With the growing availability of public databases and biobanks, such as dbGaP, UKB, All of US, FinnGen, and others, TSIM offers the opportunity for researchers to apply GWAS to existing

case-only cohorts using external controls. With the growing access and coverage of large biobanks and publicly available cohorts, our method provides an avenue through which cohorts for previously understudied phenotypes may be investigated in the GWAS framework.

TSIM also has some limitations, primarily the computational costs of imputing large genetic datasets twice and all the known limitations of imputation itself[9]. All cohorts should be imputed using the same reference panel and imputation algorithm. This may require researchers to run imputation on a large amount of data to ensure that all cohorts are processed appropriately. However, some biobanks, such as UKB, are imputing their data using publicly available reference panels and imputation methods, such as TOPMed, which may alleviate this burden for researchers in the future[10]. Additionally, because TSIM is largely reliant on high-quality imputation, the accuracy of the first stage of imputation and effectiveness of merging are dependent on key factors impacting imputation results, such as demographics and size of the reference panel and how well ancestry is matched to the study cohort. We found that the TOPMed reference panel performed better for the African and European populations than East and South Asian. The size of the study sample may also impact imputation as smaller sample sizes are more likely to have less accurate imputation quality calculations resulting in inaccurate classification of hq-SNPs[9,28,29]. Many of these issues with imputation have been addressed in recent research. For instance, Sun et al. have developed MagicalRsq[28], a machine-learning-based genotype imputation quality calibration, which takes a sample size agnostic approach to calculating imputation quality. Additionally, meta-imputation, which combines imputation results from multiple imputations with different reference panels, has been shown to improve imputation accuracy and results, especially for admixed individuals[30]. Both these methods may

be implemented in TSIM after or during the first stage of imputation, respectively, and before identifying hq-SNPs.

There are two data processing steps which will have major impacts on results. First is the pre-imputation QC. With two stages of imputation, there is the potential for genotype errors made in the raw data to propagate if insufficient QC is done. We've found that running "Imputation preparation and checking" from the McCarthy Group Tools (https://www.chg.ox.ac.uk/~wrayner/tools/) as well as filtering out poorly genotyped SNPs is crucial to ensure homogeneity among all cohorts-to-be-merged. Second is to keep in mind any pre-processing that was done on the data, particularly when the cohorts are received in different formats. For example: PLINK[31], a tool often used for QC in GWAS, codes minor and major alleles based on allele frequency while VCF files code reference and alternative alleles based on a reference. As the reference allele isn't always the major allele, some SNPs could have flipped allele codes when attempting to merge cohorts. Fortunately, these biases may also be mitigated by McCarthy Group Tools. Additional research applying TSIM to other datasets involving different genotyping platforms, genetic population structures, and phenotypes will be beneficial to further understand TSIM's limitations and applications.

Methods

Two-stage imputation method (TSIM)

In TSIM, after cohorts are separately processed and imputed using the same reference panel and imputation algorithm, high-quality genotyped and imputed SNPs (hq-SNPs) are identified in each cohort. Then, the cohorts are merged based on SNPs present in both hq-SNPs sets. Lastly, this merged dataset undergoes a second stage of imputation, after which post-imputation analysis proceeds as normal. Hq-SNPs are defined as SNPs with imputation quality (R^2) ≥ 0.99, minor allele frequency (MAF) ≥ 0.01, Hardy-Weinberg equilibrium (HWE) p-value $\geq 1\times10^{-6}$ for controls, and, if genotyped, empirical imputation quality (ER^2) ≥ 0.9. Ideally, there should be at least 300,000 hq-SNPs for the second stage of imputation[6]. If there are fewer hq-SNPs, the second stage of imputation will result in lower imputation quality and potentially more false positives in downstream analyses. Our command line tool, tsim, implements these two steps (i.e., hq-SNPs identification and merging) and outputs per chromosome VCFs ready for imputation. Additional sample QC to remove outliers or unmatched cases or controls may be done on merged unimputed genotypes if there is sufficient SNP overlap or following cohort merging on hq-SNPs after the first stage of imputation. The TOPMed Imputation Server v1.6.6 (Minimac4 for imputation, Eagle v2.4 for phasing, r2 for reference panel) was used for all our analyses.

Quality control

Before imputation, all cohorts underwent similar quality control using PLINK 1.9[31]. SNPs with missingness ≥ 0.02, MAF < 0.01, and/or HWE p-value $\leq 1\times10^{-6}$ for controls were removed. Samples with heterozygosity greater than four standard deviations from mean, missingness $\geq$ 0.02, closer than second-degree relation to other samples, and/or outlying in the principal

component analysis (PCA) space (by visual inspection) were also removed. Following imputation, we filtered out low-quality and rare imputed variants from all datasets. This included variants with $R^2 < 0.3$, MAF < 0.01, HWE p-value $\leq 1\times10^{-6}$ for controls, and if genotyped, $ER^2 < 0.9$. Further quality control was done according to the analysis conducted (see following methods for specific details on each analysis).

Determination of high-quality SNPs

We evaluated the cohort-specific bias and imputation-derived error for SNPs with high imputation quality ($R^2 \geq 0.95$) stratified by R^2 bins. Only SNPs with $R^2 \geq 0.95$ and MAF ≥ 0.01 in both AFFY- and OMNI-imputed datasets were analyzed. R^2 bins were defined based on whether the SNP's imputation quality met thresholds in both AFFY-imputed and OMNI-imputed results. See **Supp Table 1** for the total number of SNPs evaluated in each ancestry group for each R^2 bin.

Evaluating imputation-derived error

Unrelated samples from 1KGP with both array-based genotyping on AFFY and WGS were used to evaluate imputation-derived error (see **Supp Table 2** for number of samples). Here, we used the WGS genotypes to represent "ground truth." Each of the four ancestry groups (defined by 1KGP's superpopulation labels) European (EUR), African (AFR), East Asian (EAS), and South Asian (SAS), were evaluated separately. For each ancestry group, we filtered WGS for good quality biallelic SNPs with MAF ≥ 0.01, PASS in the FILTER column of the VCF, sample missingness < 0.01, and alternative allele frequency differences (determined from INFO column) between Phase 3 and 30X WGS < 0.01. We use S1 to refer to results from the first imputation

and S2 to refer to results from the second imputation. For both S1 and S2, the imputed dosages for those genotyped on AFFY were compared to their WGS genotypes using alternative allele frequency difference (AF_{diff}), defined as follows:

$$AF_{diff} = \left| \frac{\sum_{j=1}^{m} a_j - w_j}{2m} \right|$$

where:

m is the number of samples

a_j is the dosage of sample j in the AFFY imputation

w_j is the genotype of sample j in the WGS

Additionally, we stratified per SNP calculations into R^2 bins based on the imputation quality. This analysis was restricted to variants that were present in either S1 or S2 imputations for AFFY with MAF ≥ 0.01 and in our "good quality" WGS SNP set. The same analysis was repeated for samples genotyped on the OMNI platform (see **Supp Table 3** for number of samples).

Evaluating cohort-specific bias

Unrelated samples from the 1000 Genomes Project (1KGP) with array-based genotyping conducted with both Affymetrix SNP 6.0 (AFFY) and HumanOmni2.5 (OMNI) were used to evaluate cohort-specific bias (see **Supp Table 4** for number of samples) with two different merging methods: (1) separately imputing cohorts once, followed by simple merging on intersecting SNPs ("*separately-imputed*") and (2) implementing TSIM ("*two-stage*"). For both *separately-imputed* and *two-stage*, the imputed dosages for those genotyped on AFFY were compared to the imputed dosages for those genotyped on OMNI using similar AF_{diff} calculations

described previously. Only variants present in both *separately-imputed* and *two-stage* for AFFY and OMNI with MAF $\geq$ 0.01 were analyzed. Because the same samples were genotyped on both platforms, we recalculated R^2 in *two-stage* to accurately measure imputation quality for each platform, separately. The average R^2 across each stage was used to classify variants into R^2 bins. Thus, the average R^2 of a SNP in *separately-imputed* may be slightly different from the one in *two-stage*.

Investigating impact of TSIM on type I Error

We used 1KGP, subsets of UK Biobank (UKB), and controls from a psoriatic arthritis study cohort (see below) to conduct control vs. control GWAS. See **Supp Table 5** for the number of samples in each GWAS. ART and UKB samples were projected to the 1KGP PCA space using KING[32] to harmonize ancestry group annotations. All datasets underwent the standard quality control process as described above. SAIGE[22] v0.29.5 was used to conduct GWAS analysis on dosages and to account for case-control imbalances. For SAIGE Step 1, --LOCO=FALSE was used. See **Supp Table 5** for the number of principal components included. For Step 2, the following parameters were used: --minMAF=0.01, --minMAC=1, --LOCO=FALSE. LocusZoom[33] was used to identify independent genome-wide significant loci.

Investigating impact of TSIM on published psoriatic arthritis GWAS results

Our psoriatic arthritis study cohort (ART), containing both cases and controls, came from a published GWAS meta-analysis on psoriatic arthritis genotyped on the HumanOmni1-Quad BeadChip[17]. All study cohort samples were annotated as Caucasian. We randomly selected 5,000 individuals from 427,234 Europeans in UKB. Individuals with psoriatic arthritis or failing QC

were removed prior to random selection. This subset of UKB individuals comprised our external control cohort genotyped on UKB Axiom (UKB5K). There were 1,410 cases and 6,406 controls (1,406 from the study cohort; 5,000 from UKB5K) passing QC.

ART data was used in two separate analyses, the 1KGP vs. UKB control GWAS (see above) and in evaluating TSIM. We evaluated GWAS results from two different scenarios using different combinations of our cohorts:

1) *both-controls*: ART cases and controls merged with UKB5K
2) *external-controls*: ART cases merged with UKB5K

For *separately-imputed* GWAS, cohorts were merged after single, separate imputations (Stage 1). For *TSIM* GWAS, cohorts were merged using TSIM. We also conducted a GWAS of only ART cases and controls to generate reference results for comparison (*internal-controls*). All datasets underwent the standard quality control process as described above. SAIGE[22] v0.29.5 was used to conduct GWAS analysis on dosages and account for case-control imbalances, using the parameters described above. All reported p-values were adjusted for genomic control to accurately compare results. LocusZoom[33] was used to create locus plots in **Supp Fig 9**.

Of note, it was not feasible to perform a GWAS using a common practice, which combines separate cohorts before imputation, since ART and UKB were genotyped separately on two different genotyping platforms intersecting by only 100,000 SNPs after QC. This is insufficient coverage for accurate genotype imputation[6].

Replicating results from pSSNS GWAS meta-analysis

We applied TSIM and replicated results from a pediatric steroid-sensitive nephrotic syndrome (pSSNS) GWAS meta-analysis our lab previously published[18]. Specific details on the published pSSNS GWAS methods and cohort composition can be found in Barry et al.[18]. Briefly, our pSSNS GWAS consisted of a meta-analysis of two European cohorts consisting of five separate substudies. Samples were projected to 1KGP PCA space to infer ancestry using PEDDY[34]. The European Union (EU) cohort consists of four sub-cohorts: case and control data from Sorbonne Université in Paris and the NEPHROVIR study[35], as well as healthy controls from the Three Cités Study[36] and 1KGP. The United States (US) cohort contained cases and controls obtained from Columbia University in New York[18]. For EU, all individual studies were imputed separately in the first stage and then combined before the second stage of imputation with the US cohort. See **Supp Table 8** for more details on each study.

We evaluated TSIM in two different analyses: (1) GWAS was run separately in EU (using TSIM to merge studies) and US cohorts, then combined in a meta-analysis using the STDERR scheme in METAL[37], and (2) all studies in both EU and US cohorts were merged using TSIM and analyzed in one joint GWAS. For both analyses, PLINK 1.9[31] was used to conduct GWAS on genotypes in order to better compare to the published results which also used PLINK. All reported p-values were adjusted for genomic control to better compare results. LocusZoom[33] was used to create locus plots in **Supp Fig 11**.

Data Availability

Data used included genotype data from the 1000 Genomes Project, UK Biobank, dbGaP study phs000982.v1.p1, and data used in our previous pSSNS GWAS.

Code availability

The tsim command line tool is available at https://github.com/dongwonlee-lab/tsim.

Acknowledgements

This research has been conducted using UK Biobank Resource under Application Number 64945. The data used for the analyses described in this paper were obtained from the database of Genotypes and Phenotypes (dbGaP), at http://www.ncbi.nlm.nih.gov/gap. Genotype and phenotype data for the Genetic Analysis of Psoriasis and Psoriatic Arthritis study were provided by Dr. James T. Elder, University of Michigan, with collaborators Dr. Dafna Gladman, University of Toronto and Dr. Proton Rahman, Memorial University of Newfoundland, providing samples. This study was supported by grants from the National Institutes of Health, the Canadian Institute for Health Research, and the Krembil Foundation. Additional support was provided by the Babcock Memorial Trust and by the Barbara and Neal Henschel Charitable Foundation. JTE is supported by the Ann Arbor Veterans Affairs Hospital. We would like to acknowledge the Boston Children's Hospital High-Performance Computing Resources BCH HPC Clusters, Enkefalos 2 (E2), made available for conducting the research reported in this publication. Software used in the project was installed and configured by BioGrids[38]. We would like to also thank Pierre Ronco and Simone Sanna-Cherchi for allowing us to use the pSSNS

dataset in our analysis. This study was funded by NIH grants R01HG012871, R01DK119380, and RC2DK122397, as well as the Manton Center Endowed Scholar Award.

Figure Legends

Figure 1: An overview of the two-stage imputation method and its applications

(A) A detailed workflow for the two-stage imputation method applied to X number of cohorts. Dark shaded regions represent SNPs included in input to imputation. Light shaded regions represent imputed SNPs. Non-overlapping SNPs are colored in grey. (B) Illustrations of potential applications of TSIM.

Figure 2: Determination of high-quality SNPs

Line plots show AF difference (AF_{diff}) between (A) AFFY and WGS, (B) OMNI and WGS, and (C) AFFY and OMNI after the first round of imputation for Europeans (EUR), Africans (AFR), East Asians (EAS), and South Asians (SAS). SNPs with imputation quality (R^2) ≥ 0.95 were included, which were then stratified into R^2 bins. Dosage information from imputation were used for this calculation. Unimputed (with empirical $R^2 \geq 0.9$) vs. WGS (dotted line) shows AF_{diff} between SNPs genotyped on (A) AFFY or (B) OMNI compared to WGS for comparison. The total number of samples is shown in parentheses for each ancestry group. The y-axis shows the proportion of SNPs with AF_{diff} > threshold (shown in the x-axis). The y-axis is in the $\log_{10}$ scale. See **Supp Table 1** for the number of SNPs in each R^2 bin.

Figure 3: TSIM shows substantially reduced cohort-specific bias

Line plots show AF difference (AF_{diff}) between platforms in *separately-imputed* (red) and *two-stage* datasets (blue) for Europeans (EUR), Africans (AFR), East Asians (EAS), and South Asians (SAS). Dosage information from imputation were used for this calculation. AF_{diff}

threshold of zero represents no cohort-specific bias. The total number of samples (in parentheses) and SNPs (*m*) are shown for each ancestry group. The y-axis is in the $\log_{10}$ scale.

Figure 4: Manhattan plots for 1KGP vs. UKB GWAS using cohort as outcome

Manhattan plots show GWAS results of (A) 1KGP-AFFY vs. UKB samples and (B) 1KGP-OMNI vs. UKB samples. A simple merging method after separate imputations (*separately-imputed*, left) is compared to the TSIM method (*TSIM*) for each ancestry group. Red lines indicate the threshold for genome-wide significance (5×10-8).

Figure 5: TSIM replicates psoriatic arthritis GWAS results and enables the use of case-only cohorts

(A) Manhattan plots show GWAS results of the arthritis cohort following common practices (cases vs. internal controls, top), after merging cases from arthritis cohort with UKB5K using TSIM (*TSIM* cases vs. external controls, middle), and after single, separate imputations (*separately-imputed* cases vs. external controls, bottom). Total number of SNPs are shown in upper right. Numbers of cases and controls are in the upper left of each plot. Red lines indicate the genome-wide significance threshold (5×10^{-8}). Genome-wide significant loci are labelled by nearest gene. See **Supp Fig 8** for Manhattan plots after merging cases and controls from arthritis cohort with UKB5K (cases vs. both controls). (B) Locus plots for *TNIP1* showing TSIM GWAS of cases vs. external controls (middle) and cases vs. both controls (bottom). GWAS results of cases vs. internal controls (top) is also shown.

Figure 6: TSIM facilitates standard GWAS of multi-platform cohorts at the genotype level

(A) Manhattan plots show results of pSSNS European GWAS after merging data from five genotyping platforms using TSIM (*TSIM Joint*, bottom), meta-analysis of EU and US cohorts using TSIM to merge EU studies (*TSIM Meta*, middle), and the published meta-analysis (*Published*, top). Total number of SNPs are shown in upper right. 684 cases and 6,817 controls were used for all GWAS. Black dashed lines indicate the genome-wide significance threshold (5×10^{-8}). Genome-wide significant loci are labelled by nearest gene. (B) Locus plot for *CALHM6* showing results of TSIM Joint, TSIM Meta, and Published GWAS.

References

1. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 1–21 (2021).
2. Sinnott, J. A. & Kraft, P. Artifact due to differential error when cases and controls are imputed from different platforms. *Hum Genet* **131**, 111–119 (2012).
3. Wojcik, G. L. *et al.* Opportunities and challenges for the use of common controls in sequencing studies. *Nat Rev Genet* **23**, 665–679 (2022).
4. Uh, H.-W. *et al.* How to deal with the early GWAS data when imputing and combining different arrays is necessary. *Eur J Hum Genet* **20**, 572–576 (2012).
5. Johnson, E. O. *et al.* Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum Genet* **132**, 509–522 (2013).
6. Appadurai, V. *et al.* Accuracy of haplotype estimation and whole genome imputation affects complex trait analyses in complex biobanks. *Commun Biol* **6**, 101 (2023).
7. Chen, D. *et al.* A data harmonization pipeline to leverage external controls and boost power in GWAS. *Human Molecular Genetics* **31**, 481–489 (2022).
8. Li, N. & Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**, 2213–2233 (2003).
9. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics* **19**, 73–96 (2018).
10. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
11. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214–223 (2016).

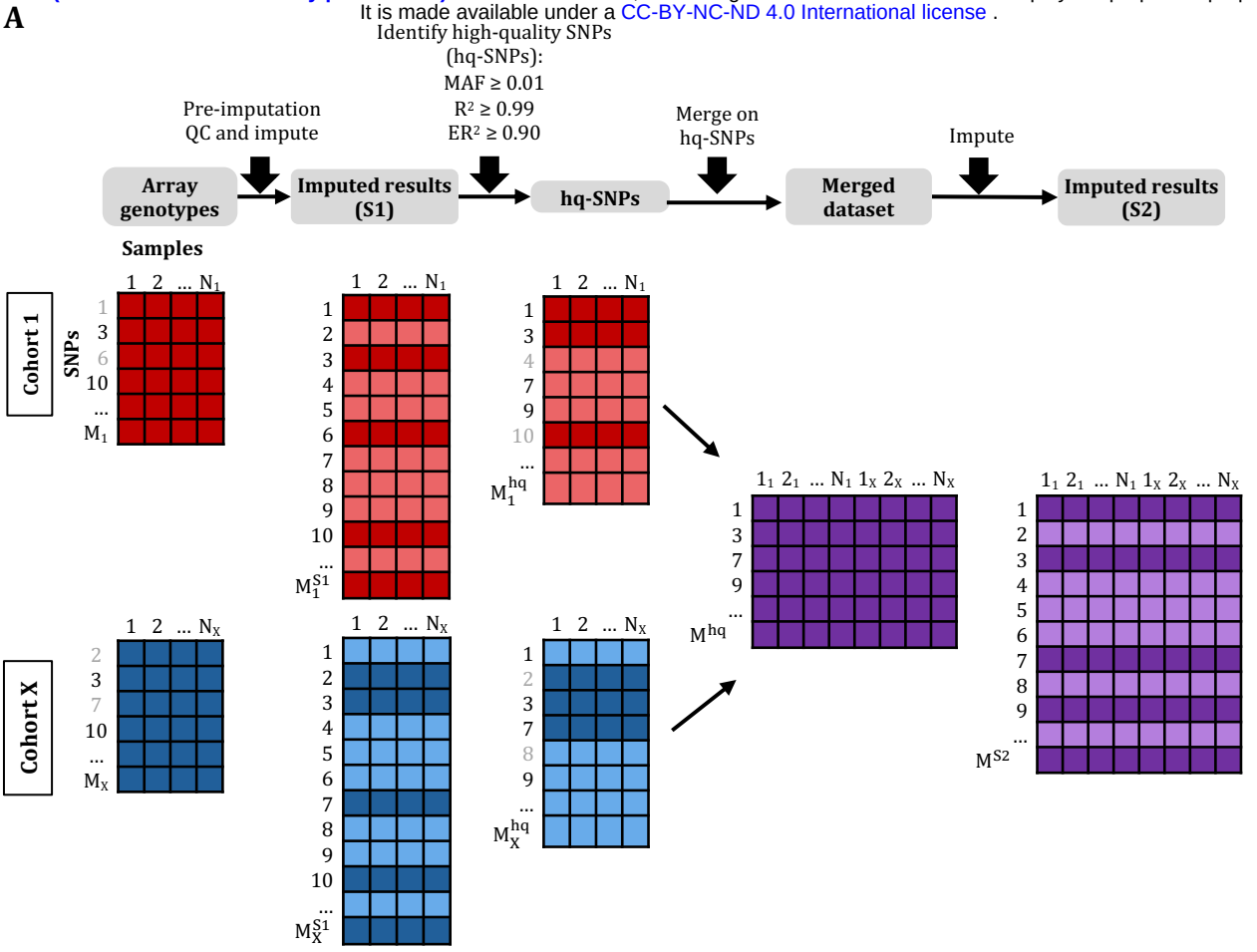
12. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics* **5**, (2014).
13. Mathur, R. *et al.* GAWMerge expands GWAS sample size and diversity by combining array-based genotyping and whole-genome sequencing. *Commun Biol* **5**, 806 (2022).
14. Hui, R., D'Atanasio, E., Cassidy, L. M., Scheib, C. L. & Kivisild, T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci Rep* **10**, 18542 (2020).
15. Kreiner-Møller, E., Medina-Gomez, C., Uitterlinden, A. G., Rivadeneira, F. & Estrada, K. Improving accuracy of rare variant imputation with a two-step imputation approach. *Eur J Hum Genet* **23**, 395–400 (2015).
16. Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research* **48**, D941–D947 (2020).
17. Stuart, P. E. *et al.* Genome-wide Association Analysis of Psoriatic Arthritis and Cutaneous Psoriasis Reveals Differences in Their Genetic Architecture. *Am J Hum Genet* **97**, 816–836 (2015).
18. Barry, A. *et al.* Multi-population genome-wide association study implicates immune and non-immune factors in pediatric steroid-sensitive nephrotic syndrome. *Nat Commun* **14**, 2481 (2023).
19. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
20. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).

21. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
22. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).
23. Aterido, A. *et al.* Genetic variation at the glycosaminoglycan metabolism pathway contributes to the risk of psoriatic arthritis but not psoriasis. *Ann Rheum Dis* **78**, e214158 (2019).
24. Baurecht, H. *et al.* Genome-wide Comparative Analysis of Atopic Dermatitis and Psoriasis Gives Insight into Opposing Genetic Mechanisms. *Am J Hum Genet* **96**, 104–120 (2015).
25. O'Rielly, D. D., Jani, M., Rahman, P. & Elder, J. T. The Genetics of Psoriasis and Psoriatic Arthritis. *The Journal of Rheumatology Supplement* **95**, 46–50 (2019).
26. Yin, X. *et al.* Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility. *Nat Commun* **6**, 6916 (2015).
27. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311 (2001).
28. Sun, Q. *et al.* MagicalRsq: Machine-learning-based genotype imputation quality calibration. *The American Journal of Human Genetics* **109**, 1986–1997 (2022).
29. Deng, T. *et al.* Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. *Front Genet* **12**, 704118 (2021).
30. Yu, K. *et al.* Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *The American Journal of Human Genetics* **0**, (2022).

31. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).
32. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
33. Boughton, A. P. *et al.* LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).
34. Pedersen, B. S. & Quinlan, A. R. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *The American Journal of Human Genetics* **100**, 406–413 (2017).
35. Debiec, H. *et al.* Transethnic, Genome-Wide Analysis Reveals Immune-Related Risk Alleles and Phenotypic Correlates in Pediatric Steroid-Sensitive Nephrotic Syndrome. *JASN* **29**, 2000–2013 (2018).
36. Vascular Factors and Risk of Dementia: Design of the Three-City Study and Baseline Characteristics of the Study Population. *Neuroepidemiology* **22**, 316–325 (2003).
37. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
38. Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).

Figure 1

A



B

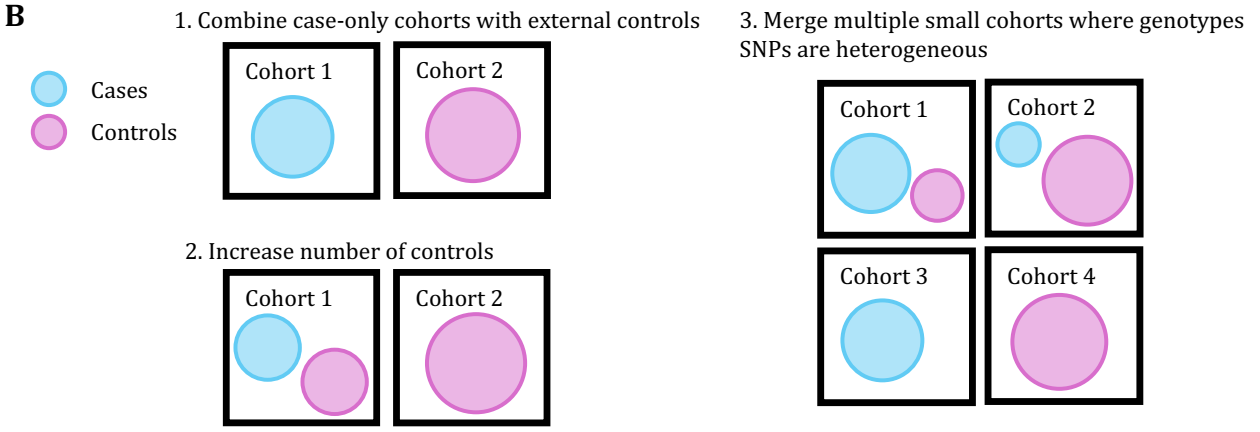


Figure 2

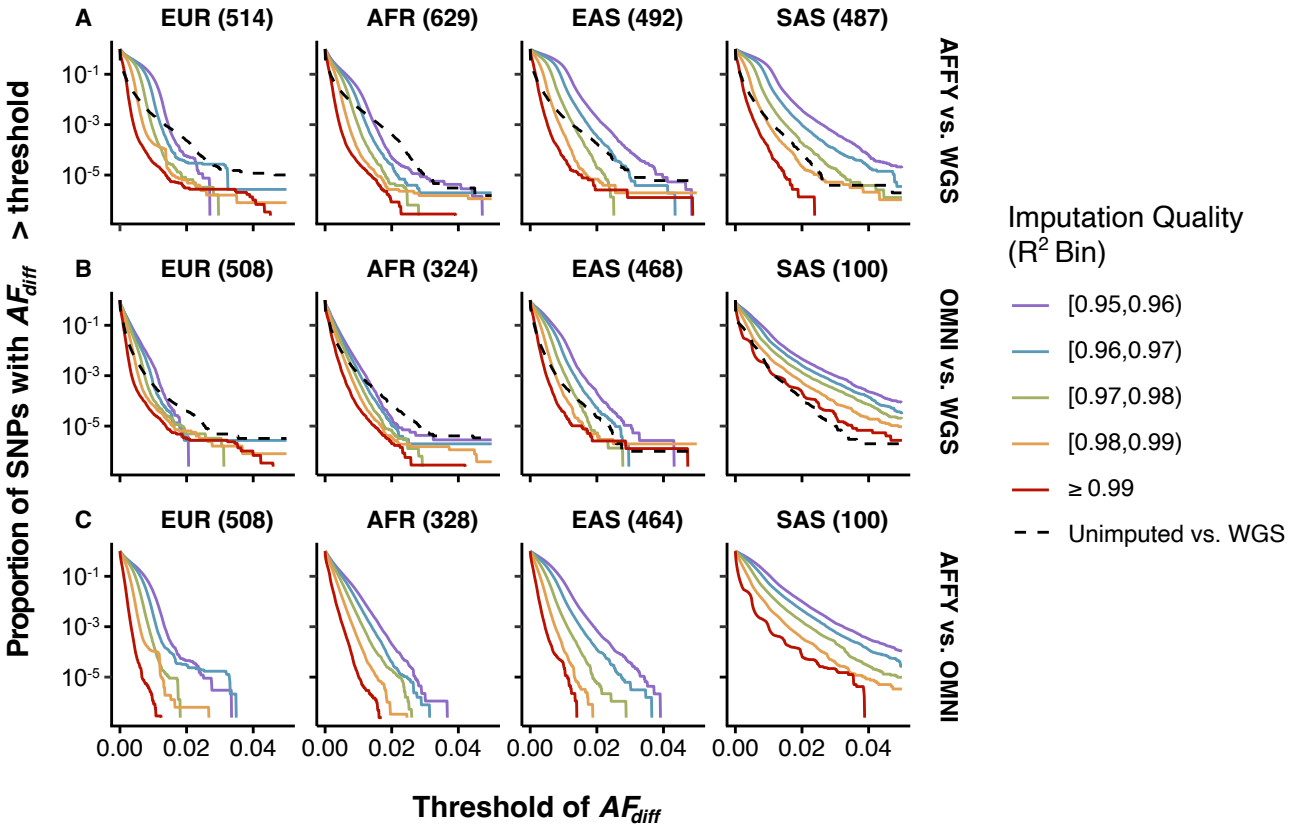


Figure 3

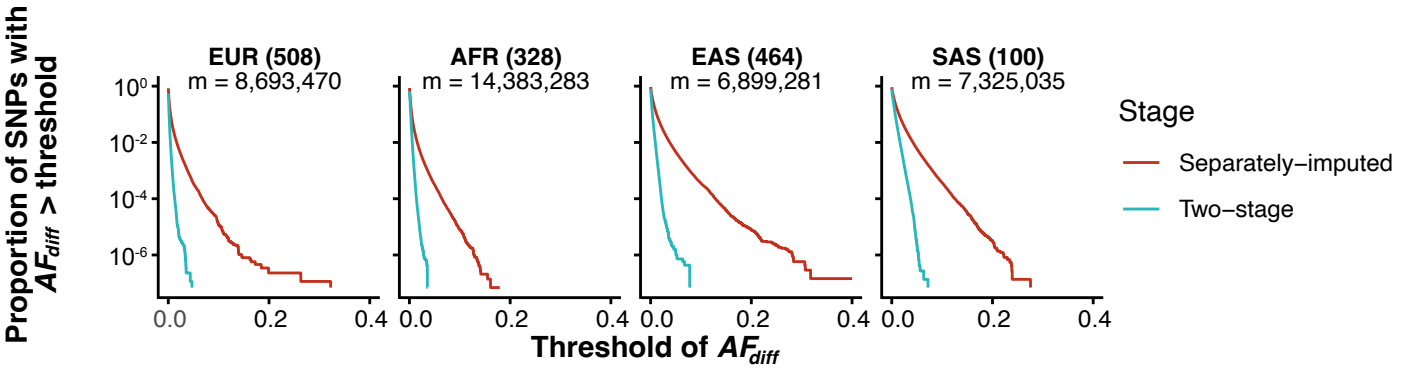


Figure 4

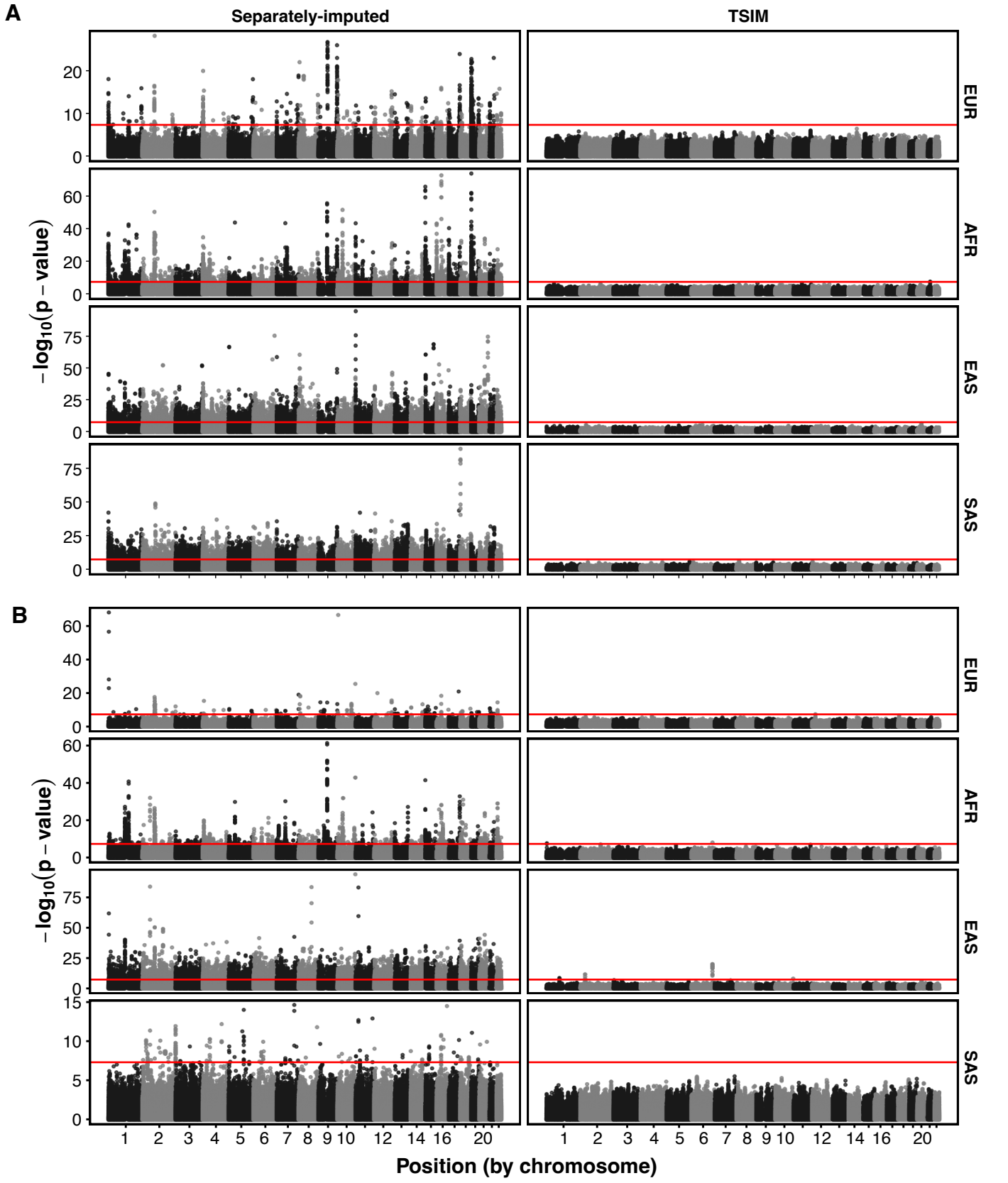


Figure 5

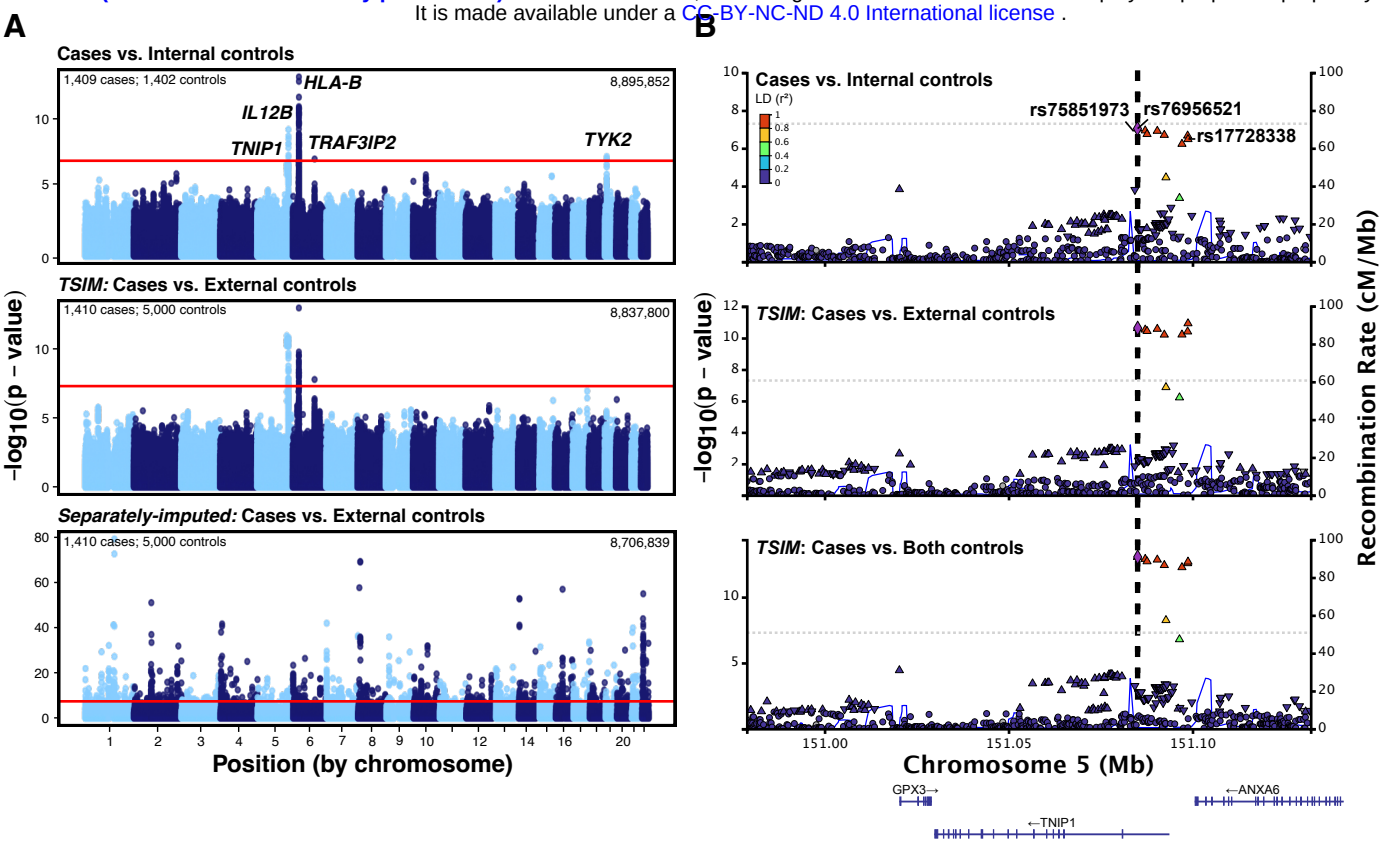


Figure 6