False discovery rates for genome-wide association tests in biobanks with thousands of phenotypes

Aubrey Annis (acannis@umich.edu)
University of Michigan Department of Biostatistics https://orcid.org/0000-0002-4095-9051

Anita Pandit
University of Michigan

Jonathon LeFaive
Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health https://orcid.org/0000-0003-3668-6086

Sarah Gagliano Taliun
University of Michigan Department of Biostatistics

Lars Fritsche
University of Michigan-Ann Arbor https://orcid.org/0000-0002-2110-1690

Peter VandeHaar
University of Michigan-Ann Arbor https://orcid.org/0000-0002-8072-9461

Michael Boehnke
University of Michigan https://orcid.org/0000-0002-6442-7754

Matthew Zawistowski
University of Michigan - Department of Biostatistics https://orcid.org/0000-0002-3005-083X

Gonçalo Abecasis
University of Michigan https://orcid.org/0000-0003-1509-1825

Sebastian Zöllner
University of Michigan

Article

Keywords: association testing, biobank data, single-iteration permutation method, false discovery rate

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-873449/v1

False discovery rates for genome-wide association tests in biobanks with thousands of phenotypes

Aubrey C Annis[a,b,*,†], Anita Pandit[a,b,*], Jonathon LeFaive[a,b], Sarah A Gagliano Taliun[a,b,c,d], Lars G Fritsche[a,b], Peter VandeHaar[a,b], Michael Boehnke[a,b], Matthew Zawistowski[a,b,**], Gonçalo R Abecasis[a,b,e,**], Sebastian Zöllner[a,b,**]

[a] Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA
[b] Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA
[c] Faculty of Medicine, Université de Montréal, Montréal, QC H3T 1J4, Canada
[d] Montréal Heart Institute, Montréal, QC H1T 1C8, Canada
[e] Regeneron Genetics Center, Regeneron Pharmaceuticals Inc., Tarrytown, NY 10591, USA

*co-first author

**co-last author

†corresponding author

ABSTRACT

Biobanks housing genetic and phenotypic data for thousands of individuals introduce new opportunities and challenges for genetic association studies. Association testing across many phenotypes increases the multiple-testing burden and correlation between phenotypes makes appropriate multiple-testing correction uncertain. Moreover, analysis including low-frequency variants results in inflated type I error due to the much larger number of tests and the elevated importance of each individual minor allele carrier in those tests. Here we demonstrate that standard Bonferroni and permutation-based methods for multiple testing correction are inadequate for a holistic analysis of biobank data because ideal significance thresholds vary across datasets and minor allele frequencies. We propose a single-iteration permutation method that is computationally feasible and provides false discovery rate (FDR) estimates tailored to individual datasets and variant frequencies. Each dataset's unique FDR estimates provide customized levels of confidence for association results and enable informed interpretation of genetic association studies across the phenome.

MAIN

Biobanks are rapidly growing in size and scope, with some now reaching 100,000s of individuals characterized across thousands of phenotypes.[1–6] The large number of phenotypes combined with extensive genetic data results in challenges for the proper interpretation of genetic association studies: (i) more association tests, which increases the multiple-testing burden; (ii) complex correlation structure among phenotypes, which varies between datasets; (iii) simultaneous common- and rare-variant testing, with rare variants vastly exceeding common variants in number;[7–9] and (iv) choices of analytical models and software.[10] Consequently, the analysis of genome- and

phenome-wide biobank data is a substantially larger and more complex endeavor than a classical genome-wide association study.

The established genome-wide significance threshold of $p \leq 5 \times 10^{-8}$ arose from Bonferroni correction accounting for the equivalent of ~1,000,000 independent tests across the genome,[11–15] and it is easy to imagine extending this approach to account for multiple testing across the phenome as well. However, given the strong correlations among phenotypes in biobank data, Bonferroni correction for phenotypes is likely unnecessarily conservative. Moreover, as common-variant testing is often more powerful than rare-variant testing, it may be unsuitable to apply the same significance threshold to both variant types.[16] Finally, differences among biobanks in sample size, density of genetic variants, phenotypes and phenotype correlation structure, and analytical choices suggest a one-size-fits-all significance threshold may be inadequate. Permutation-based methods are a common alternative to Bonferroni correction, but typically require thousands or even millions of replicates for each association test. A permutation analysis on this scale is computationally impractical even on a high performance cluster. Here we propose a computationally feasible single-iteration permutation analysis that works well despite potential variability among permutations and provides significance criteria tailored to individual datasets.

Our key insight rests on taking advantage of analyzing many phenotypes across the biobank simultaneously. When a large number of phenotypes are analyzed in parallel, a single permutation across all phenotypes followed by genetic association analyses of the permuted data enables us to understand false discovery rates (FDRs) across the phenome. Our FDR estimates in turn help us to interpret genetic associations in a biobank context.

The single-iteration permutation analysis is straightforward. We begin by separating each participant's data vector, denoted as primary data, into two subvectors: one containing phenotypes

and the other containing genotypes. We then break any true phenotype-genotype associations in the primary data by shuffling the phenotype subvectors among individuals. The shuffling process randomly merges phenotype and genotype subvectors into new data vectors, which constitute the permuted data. Any association identified in the permuted data is consequently due to chance and represents a false discovery. We include age in the phenotype subvector to avoid nonsensical data combinations in our permuted data (e.g. diagnosing a young person with Alzheimer's disease) and to ensure that we properly control for age-specific effects by incorporating age in our regression model. Similarly, we control for sex-specific effects by only permuting within sexes, thereby preventing nonsensical data combinations related to sex-specific phenotypes (e.g. diagnosing a male with ovarian cancer). When there is substantial relatedness in the data, we recommend grouping vectors into blocks (for example, nuclear families or pairs of individuals with similar kinship) and then randomly swapping vectors between blocks (Appendix A). After permuting the data, we perform genetic association studies in both the primary and permuted data and identify independent associations in each dataset by doing p-value clumping in a 1 MB region around each association signal. We estimate the FDR as the ratio of the number of independent associations across all phenotypes in the permuted-data genetic association studies (presumed false due to the phenotype-genotype dissociation) to the number of independent associations across all phenotypes in the primary-data genetic association studies (Figure 1a; Online Methods).

We applied our permutation method to individuals in two biobanks: ~408,000 "White British" participants from the UK Biobank[1] (UKB) and ~42,000 European-ancestry participants from the Michigan Genomics Initiative (MGI). The phenotypes in both datasets were derived from ICD diagnosis codes in patient electronic health records (EHRs).[17] We analyzed 1,418 binary EHR-derived UKB phenotypes (https://pheweb.org/UKB-TOPMed/)[18] and 1,659 binary EHR-derived MGI phenotypes (http://pheweb.org/MGI-freeze2/) with case counts >50; 1,365 of these phenotypes

were common to both datasets (Online Methods). To obtain FDR estimates, we performed association testing in both datasets on the primary and permuted data, in which all associations are generated by chance through shuffling the phenotype vectors. We used SAIGE[10] for the UKB analysis and both SAIGE and fastGWA[19] for the MGI analysis, which allowed for a comparison of different genetic association software (Appendix B). In addition, to assess the precision of a single-permutation FDR estimation, we repeated our SAIGE analysis across five independent permutations of the MGI data (Figure 1b). Although our analysis focuses on binary traits analyzed primarily with SAIGE software, our method is also compatible with quantitative data and with any association software that is well-calibrated for the data being analyzed (Appendix B).

To understand FDRs in biobank-scale association studies, we first evaluated the number of signals with $p \leq 5 \times 10^{-8}$ in the primary and permuted data of both biobanks. The association study of the primary UKB data yielded 5,279 independent associations with $p \leq 5 \times 10^{-8}$ while the permuted UKB data yielded 1,452, giving an overall FDR_{UKB} estimate of 28% (1,452/5,279) across the 1,418 UKB phenotypes (Figure 2a; Supplementary Table 1). The primary MGI data yielded 1,400 independent associations with $p \leq 5 \times 10^{-8}$ and the permuted data yielded 880-989 (average: 935) over the five permutations, yielding FDR_{MGI} estimates of 63-71% (average 67%) across the 1,659 MGI phenotypes (Figure 2b; Supplementary Table 1; Supplementary Figure 1). In both cases, the results immediately suggest that the standard genome-wide significance threshold of 5×10^{-8} would result in an unacceptably high false positive rate.

Next, we evaluated FDR estimates for variants in four MAF partitions (common variants: $MAF \geq 0.05$, mid-frequency variants: $0.01 \leq MAF < 0.05$, low-frequency variants: $0.001 \leq MAF < 0.01$, rare variants: $0.0001 \leq MAF < 0.001$). We found that FDRs for significant associations were substantially lower among common variants than among low-frequency and rare variants, with FDR_{UKB} ranging from 2% to 83% and FDR_{MGI} ranging from 21% to 100% (Supplementary Table 1).

Concerningly, large FDRs for low-frequency and rare variants persisted even among associations with p-values we would typically consider conclusive (e.g. ~70% FDR at $p \leq 5 \times 10^{-9}$ and $0.0001 \leq MAF < 0.001$) (Figure 3a, 3b; Supplementary Figure 2). Overall FDR estimates and the FDR estimates in each MAF partition also differed meaningfully between UKB and MGI (e.g. at $p \leq 5 \times 10^{-9}$ FDR_{UKB}=5% and average FDR_{MGI}=24%), demonstrating the variability that can exist among datasets due to their specific genotype and phenotype compositions and sample sizes (Supplementary Table 2). We believe the majority of FDR variation observed between UKB and MGI is due to greater power in UKB because of its larger sample size; increased power is expected to increase the number of true signals at any significance threshold even while the number of false signals remains constant, thus decreasing FDR. The large FDR among rare variants likely reflects the combination of decreased power among these variants and increased multiple testing burden (since the number of independent rare variants in these imputed datasets greatly exceeds the number of common variants after accounting for linkage disequilibrium). The variability among FDR estimates by dataset emphasizes the value of developing significance criteria tailored to the individual dataset.

Each FDR estimate provides an individualized level of confidence for a result by giving an approximate probability of the association being false; consequently we expect a negative correlation between FDRs and replication rates, though naturally this will depend on having sufficient power for replication as well. To assess the correlation between FDRs and replication rates phenome-wide, we performed reciprocal replication analyses of significant independent associations in UKB and MGI. In total, we evaluated 3,285 UKB and 1,010 MGI associations for replication in the other biobank across the 1,365 phenotypes common to both studies (Online Methods). As expected, we observed a gradual decrease in true replication ($p \leq 0.05$ and same direction of effect) for signals with higher FDRs (Figure 4a). In both replication analyses, most associations (~80%) with FDRs 0-50% replicated in direction of effect regardless of p-value (Supplementary Figure 3). Interestingly, in low-

FDR regions UKB replicated MGI at a much higher rate than MGI replicated UKB, most likely due to a power differential between the datasets that enabled UKB to replicate marginal MGI associations, but not vice-versa (Figure 4a; Supplementary Figure 3; Appendix C). Replication rates for signals with higher FDRs (51-100%) were much lower for both datasets (Supplementary Figure 3; Appendix C).

Bringing together the concepts of rare variants, FDRs, and replication rates, we looked at the interaction of these three elements in UKB and MGI globally and through selected examples. Across both datasets, we saw a general decrease in MAF and replication rates with increasing FDRs (Figure 4b). Median MAFs ranged from ~0.2 in low-FDR regions (<1% FDR) to ~0.0005 in high-FDR regions (81-100% FDR). Correspondingly, replication rates ranged from ~60% in low-FDR regions (<1% FDR) to ~1% in high-FDR regions (81-100%).

In a more detailed examination, we chose five representative associations in each dataset with generally comparable discovery p-values and reported their MAFs, case/control counts, estimated FDRs, and replication status in the other biobank (Table 1a, 1b). Both datasets revealed a correspondence between low FDRs, high MAFs, and replication as well as the converse correlation between high FDRs, low MAFs, and lack of replication (e.g. low FDRs: $UKB_{rs7328654\text{-Cancer of Larynx, Pharynx, Nasal Cavities}}$ — FDR<1%, MAF=0.48, replicated; $MGI_{rs7681423\text{-Pulmonary Heart Disease}}$ — FDR=3%, MAF=0.24, replicated; high FDRs: $UKB_{rs764706784\text{-Convulsions}}$ — FDR=83%, MAF=0.0005, not replicated; $MGI_{rs575967928\text{-Vitamin B-complex Deficiencies}}$ — FDR=86%, MAF=0.0008, not replicated).

These examples illustrate the importance of tailoring significance criteria to different MAFs. For instance, comparing two MGI associations with nearly identical p-values (rs3928325-Posttraumatic Stress Disorder: $cases_{MGI}$=536, $controls_{MGI}$=23,601 MAF_{MGI}=0.12, OR_{MGI}=1.74, $p_{MGI}=4.6\times10^{-8}$; rs1016111760-Osteoarthrosis: $cases_{MGI}$=9,522, $controls_{MGI}$=32,589, MAF_{MGI}=0.0006,

OR_{MGI}=11.01, $p_{MGI}=4.6\times10^{-8}$), we found that the former association involved a common variant and replicated in UKB ($cases_{UKB}$=113, $controls_{UKB}$=363,984, p_{UKB}=0.005) while the latter association involved a low-frequency variant and did not replicate in UKB ($cases_{UKB}$=28,225, $controls_{UKB}$=378,889, p_{UKB}=0.83). The FDRs for these associations corresponded with their replication status, with the former association possessing FDR_{MGI}=23% and the latter having FDR_{MGI}=92%. Similar findings held true across other examples.

Figure 5 illustrates how FDRs can help evaluate associations not only in an isolated, tabular context, but also when viewing Manhattan plots in which association signals look equally valid. The solitary signals for the MGI phenotypes "corneal opacity and other disorders of cornea" and "acquired hemolytic anemias" look almost identical, with well-formed peaks clearly exceeding the 5×10^{-8} threshold. At first glance the associations seem to have comparable chances of denoting a true signal; however, after considering the FDR estimate for the top hit in each peak (corneal opacity and other disorders of cornea: FDR=4%; acquired hemolytic anemias: FDR=72%), we concluded that while we have high confidence in the corneal phenotype association with rs11659764 (*TCF4*) on chromosome 18 ($p_{MGI}=1.9\times10^{-9}$), our confidence in the hemolytic anemias association with rs760692431 (*HS3ST4*) on chromosome 16 is attenuated despite having a similar p-value ($p_{MGI}=4.3\times10^{-9}$) that would often be considered sufficient evidence for association. A replication analysis of these two signals in the UKB confirmed the conclusions suggested by the FDRs: the association with "corneal opacity and other disorders of cornea" replicated in UKB ($p_{UKB}=2.4\times10^{-30}$) while the association with "acquired hemolytic anemias" did not replicate (p_{UKB}=0.71). These results are also in keeping with the definition of the two phenotypes: while disorders of the cornea are well-recognized as having a genetic component, acquired hemolytic anemias are less heritable. When comparing the two signals, a notable difference between them was the MAFs of the top variant in

each peak indicating either a common-variant ($MAF_{rs11659764}$=0.05 for corneal opacity) or rare-variant ($MAF_{rs760692431}$=0.0002 for acquired hemolytic anemias) association.

To address the accuracy of a single-iteration permutation, we performed five permutations of our MGI data and compared the number of independent associations yielded in each permutation. The results indicated that the number of independent associations was similar across all permutations, with the most variability occurring among low-frequency and rare variants (Figure 6; Supplementary Table 1). For associations with $p \leq 5 \times 10^{-8}$, the FDR estimates varied by only 8% across the genome (Supplementary Table 1), and variation rapidly decreased with more stringent p-value thresholds (Supplementary Table 2). Moreover, at $p \leq 5 \times 10^{-8}$ all five permutations suggested that association signals with common variants (MAF≥0.05) would have a modest FDR (21-24%), those with mid-frequency variants (0.01≤MAF<0.05) would have FDR≈50%, and those with lower MAF (MAF<0.01) would have a relatively high FDR (≥80% in all permutations). When we consider the interpretation of each FDR category (21-24%: likely to be a true association; ~50%: association could be true or false; ≥80%: likely to be a false association), we can easily see that this amount of variability in the FDR estimation achieves our goal of detecting likely reliable vs. unreliable associations and that a single permutation is adequate for estimating FDRs for associations with $p \leq 5 \times 10^{-8}$.

We also assessed the total computation time and cost for a single-iteration permutation analysis of the UKB and MGI data. Computation time for the permuted genetic association studies of 1,418 UKB and 1,659 MGI phenotypes using SAIGE were 1,752,160 and 48,221 CPU hours, respectively. Estimated computation costs for the UKB analysis ranged from ~$35,000 on Google Cloud Platform n1-standard machines to ~$47,000 for in-house computing; costs for the MGI analysis were ~$1,000 for both computing options (Table 2). It should be noted that our analysis included only European-ancestry individuals and that more computationally intensive analyses

188 employed to incorporate multi-ancestry data would increase computation time overall. A large
189 number of permutations would be prohibitively expensive and inefficient for analyzing single- or
190 multi-ancestry data in any large biobank, but a single permutation analysis has the same computation
191 time and cost as the primary data analysis and should therefore be feasible. Consequently, we suggest
192 that a single-iteration permutation analysis be performed alongside genetic association studies in a
193 biobank context and that the resulting FDR estimates will be a valuable resource for the proper
194 interpretation of association results.

195 Finally, extensions of the single-iteration permutation method can be used to evaluate other
196 analysis results. These include alternative stratification of association signals that considers features
197 like case count, minor allele count (MAC), or functional annotation of variants in addition to p-value
198 and minor allele frequency[16]. Partitioning by case count in UKB and MGI decreased variability
199 between the datasets, but it had comparatively little effect on the FDRs apart from that already
200 captured by the p-value partitions (Supplementary Figure 4). Partitioning by MAC proved more
201 useful in providing both increased consistency across the datasets and demonstrating a correlation
202 between the frequency of the variant and the FDR, yielding FDR_{UKB} ranging from 4% to 82% and an
203 average FDR_{MGI} ranging from 19% to 85% (Supplementary Table 3). Despite increased consistency
204 across MAC categories, we still observed noticeable variation between datasets (e.g. at $p \leq 5 \times 10^{-8}$ and
205 $1{,}000 \leq MAC < 5{,}000$ $FDR_{UKB}=79\%$ and average $FDR_{MGI}=57\%$), once again emphasizing the necessity
206 for calculating FDRs for the specific dataset under investigation (Supplementary Figure 5).

207 Our analysis demonstrates that the current significance threshold ($p \leq 5 \times 10^{-8}$) results in an
208 unacceptable number of false positives when testing biobanks with thousands of phenotypes. A better
209 calibrated significance criterion is needed to account for increased testing, genetic and phenotypic
210 variation among datasets, and differing variant frequencies. Our analysis showed that FDRs for low-
211 frequency and rare variants were very high in both UKB and MGI at a p-value threshold of 5×10^{-8},

whereas at lower p-value thresholds (5×10^{-10} or 5×10^{-11}) the FDRs decreased substantially (Figure 3a, 3b; Supplementary Figure 2). These results suggest that for these two datasets a more appropriate cutoff for statistical significance among low-frequency and rare variants would be around 5×10^{-10} or 5×10^{-11} rather than 5×10^{-8}, which is generally used as the significance threshold for common variants. As shown in the differences between the UKB and MGI FDR estimates, FDRs will likely vary across datasets depending on the variant frequencies, sample size, and correlation structure of each dataset. Instead of recommending a universal significance threshold for biobank studies that does not take into account differences in biobank features, we suggest using FDRs to provide a customized level of confidence for each association given its specific discovery dataset, MAF, and p-value. Since only one permutation is required to achieve a stable FDR estimate, our permutation analysis can be run alongside a primary genetic association study with manageable additional computation time and cost. Moreover, our method is applicable to both binary and quantitative traits, any association software properly calibrated for the data being analyzed, datasets with related individuals, and multi-ancestry datasets, making it useful on a broad spectrum. We believe that publications of genetic association study findings should include the estimated probability of success suggested by FDR estimates along with the primary association study results whenever possible. This process will contextualize genetic association study results for any dataset regardless of its multiple testing context, correlation structures, or proportion of rare variants.

ONLINE METHODS

UKB

Our analysis included data from 407,753 "White British" participants drawn from the full UKB cohort released in July 2017.[1] Participants were genotyped on an Affymetrix Axiom array with 820,967 variants. Non-genotyped variants were imputed using the TOPMed imputation reference

panel and filtered to remove variants with $R^2 \leq 0.3$ and/or MAF≤0.01% for a total of ~37,000,000 variants analyzed across each phenotype.[20–22] We specified individuals of "White British" ancestry using the original definitions provided by UKB.[1] We drew all other phenotype and covariate data from participant electronic health records (EHRs), including diagnoses coded with the Ninth and Tenth Revision of the International Statistical Classification of Diseases with clinical modifications (ICD9-CM and ICD10-CM), sex, genotyping batch, and age. We aggregated the ICD9-CM and ICD10-CM codes to form up to 1,857 PheWAS phenotypes using the PheWAS R package,[17] which groups related ICD codes into hierarchical phenotypes. We used 1,418 of the resulting phenotypes having case count >50 in the analysis. 1,365 of these 1,418 phenotypes were also analyzed in MGI.

UKB received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382). The present analyses were conducted under UKB data application number 24460.

MGI

Our analysis included data from 42,167 European-ancestry participants in the second freeze of MGI (March 2019). Participants were genotyped on an Illumina HumanCoreExome array with 542,585 variants. Non-genotyped variants were imputed to the Haplotype Reference Consortium (HRC) panel using the Michigan Imputation Server and filtered to remove variants with $R^2 \leq 0.3$ and/or MAF≤0.01% for a total of ~32,000,000 variants analyzed across each phenotype.[20,23] We inferred recent ancestry by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project reference panel using PLINK1.9 (938 unrelated individuals) and defined individuals with European ancestry similarly to Fritsche et al. 2018.[24–27] We drew all other phenotype and covariate data from participant EHRs, including diagnoses coded with the Ninth and Tenth Revision of the International Statistical Classification of Diseases with clinical modifications (ICD9-CM and ICD10-CM), sex, genotyping batch, and age. We also aggregated the

ICD9-CM and ICD10-CM codes to form up to 1,857 PheWAS phenotypes using the PheWAS R package[17] and used 1,659 phenotypes having case count >50 in the analysis. 1,365 of these 1,659 phenotypes were also analyzed in UKB.

MGI data were collected according to Declaration of Helsinki principles. Study participants provided written informed consent, and protocols were reviewed and approved by local ethics committees (IRB ID HUM00099605).

Permutation and Association Analyses

Overview

Both the UKB and MGI analyses utilized genotype and EHR-derived phenotype data for n participants (n_{UKB}=407,753, n_{MGI}=42,167) and p phenotypes having case count >50 (p_{UKB}=1,418, p_{MGI}=1,659). Because the allele frequency filters applied in the association analyses depend on individuals labeled as cases and controls for each phenotype, every phenotype was analyzed with a slightly different set of variants (~37,000,000 for UKB phenotypes and ~32,000,000 for MGI phenotypes). Our permutation method stratifies by inferred genetic sex and then shuffles the phenotype data, along with any phenotypic covariates, to break the association with the genotype vectors and any genotypic covariates. In our analyses, we included only age as a phenotypic covariate and sex, PCs, and chip version as genotypic covariates, but it is possible that specific phenotypes could have additional phenotypic or genotypic covariates (e.g. specific clinical risk factor, batch effects). Our notation allows for refinement of the model to accommodate this scenario.

Notation

Let n be the number of participants in the dataset, m be the number of genotyped and imputed variants, and p be the number of phenotypes. Let Y_{ij} be the observation for the j^{th} phenotype in

individual i where $\mathbf{Y}$ is an $n \times p$ matrix. Participant outcome data for the j^{th} phenotype is stored in an n-element phenotype vector $\mathbf{Y_{*j}}$, and phenotype data for the i^{th} individual is stored in a p-element individual vector $\mathbf{Y_{i*}}$. Let participant genotype data be stored in $\mathbf{G}$, an $n \times m$ genotype matrix. Finally, let covariate data for the j^{th} phenotype be contained in matrices $\mathbf{Q_j}$ and $\mathbf{W_j}$, where $\mathbf{Q_j}$ is an $n \times r_j$ matrix with r_j genotypic covariates (e.g. sex, PCs, genotyping batch) and $\mathbf{W_j}$ is an $n \times l_j$ matrix with l_j phenotypic covariates (e.g. age, phenotyping batch).

A complete phenotype vector defining each participant's case-control status can be constructed by concatenating the participant's phenotypes and phenotypic covariates. Thus if a participant has a p-length phenotype vector $\mathbf{Y_{i*}}$, which includes all phenotypes for the i^{th} participant, and an l-length phenotypic covariate vector $\mathbf{W_{i*}}$, which includes all phenotypic covariates for the i^{th} participant, the i^{th} participant's complete phenotype vector can be defined as a $(p + l)$-length vector $(\mathbf{Y_{i*}}, \mathbf{W_{i*}})$. The $n \times (p + l)$ matrix $(\mathbf{Y}, \mathbf{W})$ denotes the collection of complete phenotype vectors for all participants.

To obtain the permuted data, we shuffle the participants' complete phenotype vectors, thereby permuting case-control status while preserving the correlation structure among phenotypes. Our first step in this process is to separate $(\mathbf{Y}, \mathbf{W})$ by sex into $(\mathbf{Y}, \mathbf{W})^M$ and $(\mathbf{Y}, \mathbf{W})^F$ to ensure permutation only among individuals of the same sex, which accomodates sex-specific phenotypes such as prostate or ovarian cancer. We then randomly permute the complete phenotype vectors by shuffling rows among participants in each group to obtain permuted complete phenotype matrices for males and females. We recombine the permuted data to make $(\mathbf{Y}, \mathbf{W})^P$, a permuted complete phenotype matrix containing both males and females that comprises permuted phenotype matrix $\mathbf{Y}^P$ and permuted phenotypic covariate matrix $\mathbf{W}^P$.

Using appropriate association software (SAIGE, fastGWA, etc.), we test for association between genetic markers and case-control status for each phenotype in both the primary and permuted data. When using SAIGE, we employed a logistic mixed model; when using fastGWA, we employed a linear mixed model:

SAIGE

$$\text{Primary:} \quad \boldsymbol{\mu} := \text{logit}(E[\mathbf{Y}_{*\mathbf{j}} \mid \mathbf{W_j}, \mathbf{G}, \mathbf{Q_j}]) = \mathbf{W_j\alpha} + \mathbf{G\beta} + \mathbf{Q_j\gamma} + \mathbf{v_j} \tag{1}$$

$$\text{Permuted:} \quad \boldsymbol{\mu}^P := \text{logit}(E[\mathbf{Y}_{*\mathbf{j}}{}^{\mathbf{P}} \mid \mathbf{W_j}^{\mathrm{P}}, \mathbf{G}, \mathbf{Q_j}]) = \mathbf{W_j}^{\mathbf{P}}\boldsymbol{\alpha} + \mathbf{G\beta} + \mathbf{Q_j\gamma} + \mathbf{v_j} \tag{2}$$

fastGWA

$$\text{Primary:} \quad \boldsymbol{\mu} := E[\mathbf{Y}_{*\mathbf{j}} \mid \mathbf{W_j}, \mathbf{G}, \mathbf{Q_j}] = \mathbf{W_j\alpha} + \mathbf{G\beta} + \mathbf{Q_j\gamma} + \mathbf{v_j} \tag{3}$$

$$\text{Permuted:} \quad \boldsymbol{\mu}^P := E[\mathbf{Y}_{*\mathbf{j}}{}^{\mathbf{P}} \mid \mathbf{W_j}^{\mathrm{P}}, \mathbf{G}, \mathbf{Q_j}] = \mathbf{W_j}^{\mathbf{P}}\boldsymbol{\alpha} + \mathbf{G\beta} + \mathbf{Q_j\gamma} + \mathbf{v_j} \tag{4}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are l_j-length, m-length, and r_j-length vectors of the unknown effects of the phenotypic covariates, genotypes, and genotypic covariates respectively, and $\mathbf{v_j}$ is the n-length random effects vector for the j^{th} phenotype.

Clumping and FDRs

After completing the association analyses for all phenotypes in both the primary and permuted data, we perform clumping of the summary statistics (using PLINK1.9[27]) with 500 kb flanks around the most significant signal in that region (--clump-kb 500). This clumping yields independent associations in 1 MB windows for both primary and permuted data.[28,29] We use these results to calculate the FDR for the phenome at a specified significance level L (e.g. $p \leq 5 \times 10^{-8}$) with the equation,

$$FDR_L = \frac{N_{L,permuted}}{N_{L,primary}}$$

where $N_{L,permuted}$ is the number of independent associations in the permuted phenome with p-value ≤ L and $N_{L,primary}$ is the number of independent associations in the primary phenome with p-value ≤ L.

Replications

Our replication analysis employed UKB and MGI as reciprocal discovery and replication datasets. After identifying all $p \leq 5 \times 10^{-8}$ independent associations in each dataset, we looked for nominal replication ($p \leq 0.05$) of each phenotype-variant association in the summary statistics of the other dataset.

Acknowledgments

We thank the following grants for supporting this project: HG000040 (AA), HG009976 (MB), and R01 HG005855 (SZ). The authors acknowledge the Michigan Genomics Initiative participants, Precision Health at the University of Michigan, the University of Michigan Medical School Central Biorepository, and the University of Michigan Advanced Genomics Core for providing data and specimen storage, management, processing, and distribution services, and the Center for Statistical Genetics in the Department of Biostatistics at the School of Public Health for genotype data curation, imputation, and management in support of the research reported in this publication.

Author Contributions

GRA, SZ, and MZ conceived and supervised the study. ACA, AP, GRA, MB, SZ, and MZ were responsible for statistical analysis and interpretation of results. LGF, SAGT, and AP developed pipelines for association analyses. AP, LGF, and ACA carried out the MGI analysis. JL, SAGT, and ACA carried out the UKB analysis. PV developed and implemented PheWeb software for displaying MGI and UKB analysis results. ACA prepared the manuscript. All authors critically reviewed the manuscript. ACA and AP are developing software for easy implementation of the method described in the paper.

Competing Interests

GRA is an employee of Regeneron Pharmaceuticals and owns stock options in the company. The authors declare no other competing financial interests.

Appendix A: Permutation with Related Individuals

Overview

Mbatchou and McPeek[30] have shown that permutations of datasets with large numbers of related individuals have inflated type 1 error rates, which would in turn lead to inflated FDRs. Although our single iteration permutation method assumes independence among individuals, it can be extended easily to accommodate groups of related individuals by first grouping the most highly-related participants and then proceeding to permute by groups (Supplementary Figure 6). This modification should yield similar results to the analysis presented above with unrelated individuals, and we recommend it for biobanks with substantial amounts of relatedness.

Notation

Let n be the number of participants in the dataset, m be the number of genotyped and imputed variants, and p be the number of phenotypes. Let Y_{ij} be the observation for the j^{th} phenotype in individual i where $\mathbf{Y}$ is an $n \times p$ matrix. Participant outcome data for the j^{th} phenotype is stored in an n-element phenotype vector $\mathbf{Y_{*j}}$, and phenotype data for the i^{th} individual is stored in a p-element individual vector $\mathbf{Y_{i*}}$. Let participant genotype data be stored in $\mathbf{G}$, an $n \times m$ genotype matrix. Finally, let covariate data for the j^{th} phenotype be contained in matrices $\mathbf{Q_j}$ and $\mathbf{W_j}$, where $\mathbf{Q_j}$ is an $n \times r_j$ matrix with r_j genotypic covariates (e.g. sex, PCs, genotyping batch) and $\mathbf{W_j}$ is an $n \times l_j$ matrix with l_j phenotypic covariates (e.g. age, phenotyping batch).

A complete phenotype vector defining each participant's case-control status can be constructed by concatenating the participant's phenotypes and phenotypic covariates. Thus if a participant has a p-length phenotype vector $\mathbf{Y_{i*}}$, which includes all phenotypes for the i^{th} participant, and an l-length phenotypic covariate vector $\mathbf{W_{i*}}$, which includes all phenotypic covariates for the i^{th} participant, the i^{th} participant's complete phenotype vector can be defined as a $(p + l)$-length vector

371 ($\mathbf{Y}_{i*}$, $\mathbf{W}_{i*}$). The $n \times (p + l)$ matrix ($\mathbf{Y}$, $\mathbf{W}$) denotes the collection of complete phenotype vectors for
372 all participants.

373 To obtain the permuted data, we find genetically related groups of individuals of the same
374 sex within our dataset and then shuffle the participants' complete grouped phenotype vectors, thereby
375 permuting case-control status among groups while preserving the correlation structure among
376 phenotypes. Our first step in this process is to separate ($\mathbf{Y}$, $\mathbf{W}$) by sex into $(\mathbf{Y}, \mathbf{W})^{M}$ and $(\mathbf{Y}, \mathbf{W})^{F}$ to
377 ensure permutation only among individuals of the same sex, which accomodates sex-specific
378 phenotypes such as prostate or ovarian cancer. We then use a genetic-relatedness software, such as
379 PLINK[27] or KING[31], to group participants within each sex with their closest relatives; this process
380 will produce g complete phenotype grouped matrices denoted $(\mathbf{Y}, \mathbf{W})_{k}$ each with dimension $g_k \times (p +$
381 $l)$, where g is the total number of groups, g_k is the number of participants in the k^{th} group, p is the
382 length of the phenotype vector, and l is the length of the phenotypic covariate vector. We then
383 randomly permute the phenotype data by shuffling the complete phenotype grouped matrices within
384 sex to obtain permuted complete phenotype matrices for males and females (N.B. there must be a
385 sufficient number of groups within each sex containing the same number of individuals to
386 accomplish random shuffling between groups). We recombine the permuted data to make $(\mathbf{Y}, \mathbf{W})^{P}$, a
387 permuted complete phenotype matrix containing both males and females that comprises permuted
388 phenotype matrix $\mathbf{Y}^{P}$ and permuted phenotypic covariate matrix $\mathbf{W}^{P}$.

389 The association analysis then proceeds in the same manner as the analysis for unrelated
390 individuals (Online Methods).

391 **Appendix B: fastGWA Analysis**

392 Researchers may wish to use faster and less computationally intensive association analysis
393 software, such as fastGWA[19], to aid in lessening the computational burden of analyzing two datasets

(primary and permuted) phenome-wide; however, they must use care to employ software that is appropriately calibrated to the data being analyzed because improper software choices may yield inaccurate FDR estimates. To illustrate the importance of utilizing software suited to the data being analyzed when calculating FDRs, we repeated our MGI analysis using fastGWA and compared it to our SAIGE results. SAIGE is calibrated to account for binary data and case-control imbalances while fastGWA performs best in datasets with quantitative data or binary data with balanced numbers of cases and controls. Many MGI phenotypes have large case-control imbalances (case-control ratio: mean=0.048, median=0.019), which led to the number of independent associations found using fastGWA for the primary and permuted genetic association studies to be highly inflated (at $p \leq 5 \times 10^{-8}$: $primary_{SAIGE}$=1,400, $permuted_{SAIGE}$=880; $primary_{fastGWA}$=4,597,051, $permuted_{fastGWA}$=4,528,660) (Supplementary Table 1; Supplementary Table 4; Supplementary Figure 7). The massive inflation in both primary and permuted data made FDR estimates for associations with all but relatively common variants unacceptably high (Supplementary Table 4; Supplementary Figure 7). When we restricted our analysis to variants with MAF≥0.05, we obtained sensible FDRs for each p-value category that corresponded well with analogous FDRs in the SAIGE analysis ($p \leq 5 \times 10^{-8}$: FDR_{SAIGE}=23%, $FDR_{fastGWA}$=19%; $p \leq 5 \times 10^{-9}$: FDR_{SAIGE}=4%, $FDR_{fastGWA}$=4%; $p \leq 5 \times 10^{-10}$: FDR_{SAIGE}<1%, $FDR_{fastGWA}$<1%) (Supplementary Table 4; Supplementary Figure 7; Figure 3b). Thus, to get accurate FDR estimates it is important to pair our permutation method with software that works well for the data being analyzed, and careful consideration should be given to data with binary outcomes, case control imbalances, and a large proportion of rare variants.

Appendix C: Lack of Replication

In our reciprocal replication analyses of all significant independent associations in UKB and MGI, we evaluated 3,285 UKB and 1,010 MGI associations for replication in the other biobank (Online Methods). In both replication analyses most associations (~80%) with FDRs 0-50%

replicated in direction of effect, regardless of p-value; a large proportion of these associations (UKB=44%, MGI=71%) also replicated at nominal significance ($p \leq 0.05$) (Supplementary Figure 3).

Interestingly, in both analyses the associations with FDRs 51-100% replicated in direction of effect (regardless of p-value) less than 50% of the time, the proportion we would expect purely by chance (Supplementary Figure 3). This 51-100% FDR category primarily contains rare variants ($0.0001 \leq MAF < 0.001$: UKB=65%, MGI=63%). Since most traits have a much lower number of cases than controls (case-control ratio: $mean_{UKB}=0.007$, $mean_{MGI}=0.048$; $median_{UKB}=0.002$, $median_{MGI}=0.019$), any rare alleles that have no effect on the disease are expected to occur primarily in controls. Under this Null Hypothesis, the only way rare variants can show a highly significant association is by being over-abundant in cases. Thus, when we condition on rare variants having a significant association (i.e. when we attempt to replicate rare variants that have significant p-values in the discovery dataset), we implicitly condition on an increased frequency among cases, which yields a surplus of positive effect sizes (proportion of positive effect sizes among associations with $0.0001 \leq MAF < 0.001$: UKB=100%, MGI=100%) (Supplementary Figure 8). In contrast, the replication datasets have a smaller proportion of positive effect sizes for the minor allele among rare-variant associations (average proportion of positive effect sizes among associations with $0.0001 \leq MAF < 0.001$: UKB=28%, MGI=39%) (Supplementary Figure 9), which is unsurprising since we do not condition on a highly significant p-value when observing these associations in the replication datasets. Thus in the replication datasets, variants that do not affect the trait of interest will follow the Null Hypothesis and the rare allele will occur primarily in controls, resulting in a negative effect size. This combination of an excess of rare-variant tests in noncausal regions and a case-control imbalance in most phenotypes leads to many significant associations having positive effect sizes and many nonsignificant associations having negative effect sizes, resulting in fewer than

441 50% of the replication-dataset associations having the same direction of effect as the discovery-
442 dataset associations.

443 We also observed a lower replication ($p \leq 0.05$) rate than expected among UKB associations
444 with $FDR_{UKB}=0$, which we would expect to replicate nearly 100% of the time (Figure 4a). Only 65%
445 (389/597) of associations with $FDR_{UKB}=0$ replicated in MGI, contrasting with 99% (75/76) of MGI
446 associations with $FDR_{MGI}=0$ replicating in UKB. We believe this lack of replication of UKB
447 associations in MGI is due to a variety of factors, including different uses of ICD codes in the UK
448 and Michigan, dissimilarities in the aggressiveness of preventative treatments in these locations, and
449 a lack of power due to smaller sample sizes in MGI.

450 A clear difference in phecode definitions, most likely originating from different uses of ICD
451 codes, can be seen if we compare the UKB and MGI results for the related phenotypes "disorders of
452 iron metabolism" (phecode 275.1; $cases_{UKB}=666$, $controls_{UKB}=405{,}081$; $cases_{MGI}=149$,
453 $controls_{MGI}=39{,}037$) and "disorders of mineral metabolism" (phecode 275; $cases_{UKB}=2{,}118$,
454 $controls_{UKB}=405{,}081$; $cases_{MGI}=3{,}074$, $controls_{MGI}=39{,}037$). Both UKB and MGI demonstrated a
455 strong association with "disorders of iron metabolism" near *HFE* on chromosome 6, with MGI
456 replicating ($p \leq 0.05$) 71% (12/17) of UKB associations with $FDR_{UKB}=0$ near this signal
457 (Supplementary Figure 10); UKB also showed a similar signal near *HFE* for "disorders of mineral
458 metabolism," but MGI had no significant associations for "disorders of mineral metabolism" and
459 only replicated ($p \leq 0.05$) 8% (1/13) of UKB associations with $FDR_{UKB}=0$ near this signal
460 (Supplementary Figure 11). The phenotypes appear to be closely related in UKB, with associations
461 occurring in the same region of chromosome 6. This signal is also picked up in the association
462 analysis of MGI's "disorders of iron metabolism," but not in the analysis of MGI's "disorders of
463 mineral metabolism." The lack of signal in MGI's "disorders of mineral metabolism" suggests that
464 this association study is unexpectedly capturing a group of participants with an underlying phenotype

that is dissimilar to the other three association studies despite the similarity of their phecodes. This discrepancy among phecodes would make replication using MGI's "disorders of mineral metabolism" phenotype virtually impossible.

Another lack of replication occurred for the phenotype "benign neoplasm of colon" (phecode 208; $cases_{UKB}$=20,121, $controls_{UKB}$=384,292; $cases_{MGI}$=8,083, $controls_{MGI}$=33,652), which may be due in part to different treatment techniques in the UK and Michigan. The United States has traditionally taken a more aggressive stance towards colon screening than the UK, with the focus in the former being on cancer prevention and in the latter on cancer detection.[32] The preventative colon cancer treatments common in the US would result in not only including people who manifest concerning symptoms in the screening for and removal of benign neoplasms, but also including individuals who, though having no identifying symptoms or genetic predisposition for colon cancer, meet the US criteria for preventative care. These asymptomatic people who had surgery to remove a benign neoplasm of the colon would be included as cases in the MGI analysis. In contrast, the UK's focus on cancer detection would result in largely symptomatic people being included in the UKB analysis. The overall quality of the UKB data, therefore, would be superior to the MGI data for detecting a genetic predisposition towards neoplasms of the colon since the MGI sample is diluted by people who are undergoing routine preventative care. In light of the potentially different populations composing the studies — along with a lack of power arising from unequal sample sizes — it makes sense that only 22% (2/9) of UKB associations with FDR_{UKB}=0 replicated in MGI (Supplementary Figure 12).

Finally, many failures to replicate UKB signals are most likely due to a lack of power, which is unsurprising when the replication dataset is ~1/10 the size of the discovery dataset. Using the phenotype prevalences manifested in UKB and the corresponding allele frequencies, case and control numbers, and effect sizes in MGI, we calculated the power of replicating the UKB associations in

MGI.[33,34] Supplementary Figure 13 shows that as the replication power increased, so too did the proportion of associations replicated in MGI. Moreover, Supplementary Figure 13 reveals a negative association between the AF of the phenotype effect allele (i.e. the allele that contributes to phenotypic manifestation) and the power to replicate; no other changes in variables (case count, disease prevalence, or relative risk) showed a strong correlation with power. These results suggest that most of the differences in power between the MGI results are due to differences in AFs, where variants having lower phenotype effect AFs tend to have greater power for replication. However, it must be noted that this power gradation among phenotype effect AFs only exists among common variants (median AFs for all power categories range from 0.13-0.74) (Supplementary Figure 13). Among truly rare variants ($0.0001 \leq MAF < 0.001$) we observed low power and low replication (mean power=0.19, median power=0.07; 7% replicated), which most likely is due to the replication dataset's smaller sample size and the consequent lower incidence rate of the rare allele as compared with the discovery dataset. The rare-variant associations that did have higher power to replicate were generally accompanied by large relative risks, indicating that these variants have a particularly large effect on the phenotype. As rare variants gained power with increasing relative risks, we saw a corresponding increase in replication rates in MGI (Supplementary Figure 13).

REFERENCES

1. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
2. Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *J Pers Med* **6**, (2016).
3. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
4. FinnGen. FinnGen. *In: FinnGen Documentation of R3 release [Internet]. [cited 26 Jan 2021]* https://finngen.gitbook.io/documentation/.
5. Krokstad, S. *et al.* Cohort Profile: the HUNT Study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
6. Abul-Husn, N. S. & Kenny, E. E. Personalized Medicine and the Power of Electronic Health Records. *Cell* **177**, 58–69 (2019).
7. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24**, 1202–1205 (2016).
8. Sveinbjornsson, G. *et al.* Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).
9. Tabangin, M. E., Woo, J. G. & Martin, L. J. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc.* **3 Suppl 7**, S41 (2009).
10. 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).
11. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144 (1999).
12. Abecasis, G. R., Cardon, L. R., Cookson, W. O., Sham, P. C. & Cherny, S. S. Association analysis in a variance components framework. *Genet. Epidemiol.* **21 Suppl 1**, S341–6 (2001).

13. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

14. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).

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

16. Asif, H. *et al.* GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Mol. Psychiatry* (2020) doi:10.1038/s41380-020-0670-3.

17. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).

18. Gagliano Taliun, S. A. *et al.* Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).

19. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).

20. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

21. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. 563866 (2019) doi:10.1101/563866.

22. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).

23. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

24. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based

association studies. *Nat. Genet.* **46**, 409–415 (2014).

25. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
26. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* **102**, 1048–1061 (2018).
27. 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).
28. Choi, S. H. *et al.* Six Novel Loci Associated with Circulating VEGF Levels Identified by a Meta-analysis of Genome-Wide Association Studies. *PLoS Genet.* **12**, e1005874 (2016).
29. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
30. Mbatchou, J., Abney, M. & McPeek, M. S. Permutation methods for assessing significance in binary trait association mapping with structured samples. *bioRxiv* 451377 (2019) doi:10.1101/451377.
31. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
32. Meza, R., Jeon, J., Renehan, A. G. & Luebeck, E. G. Colorectal cancer incidence trends in the United States and United kingdom: evidence of right- to left-sided biological gradients with implications for screening. *Cancer Res.* **70**, 5419–5429 (2010).
33. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
34. Johnson, J. L. & Abecasis, G. R. GAS Power Calculator: web-based power calculator for genetic association studies. *bioRxiv* 164343 (2017) doi:10.1101/164343.

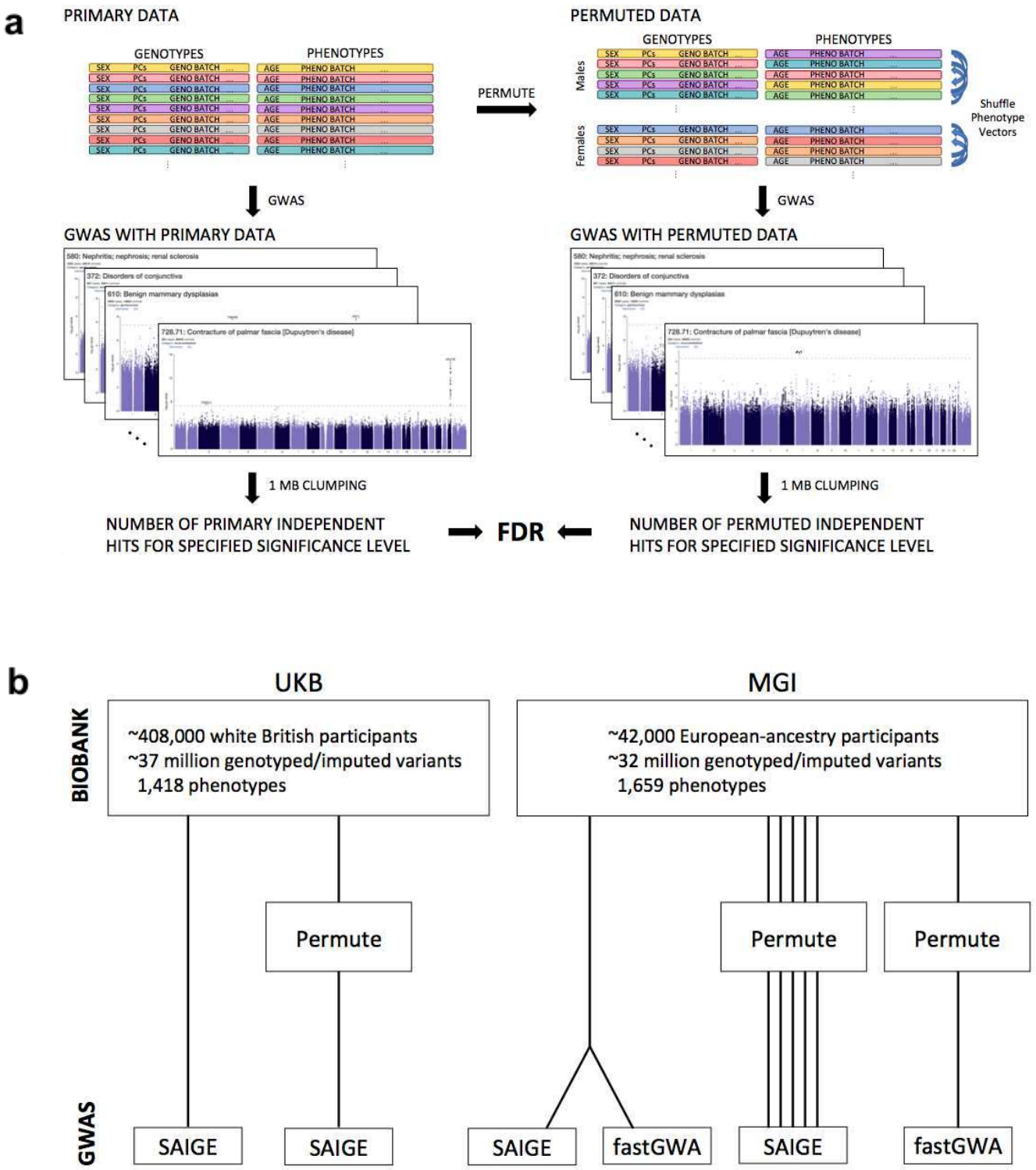

Figure 1. Single-iteration permutation method and study design. a) Single-iteration permutation method including the permutation process, association studies of primary and permuted data, 1 MB positional clumping, and calculation of the FDR. b) Study design for the single-iteration permutation method including analysis of primary and permuted UKB and MGI data and utilization of the two association software packages SAIGE and fastGWA.

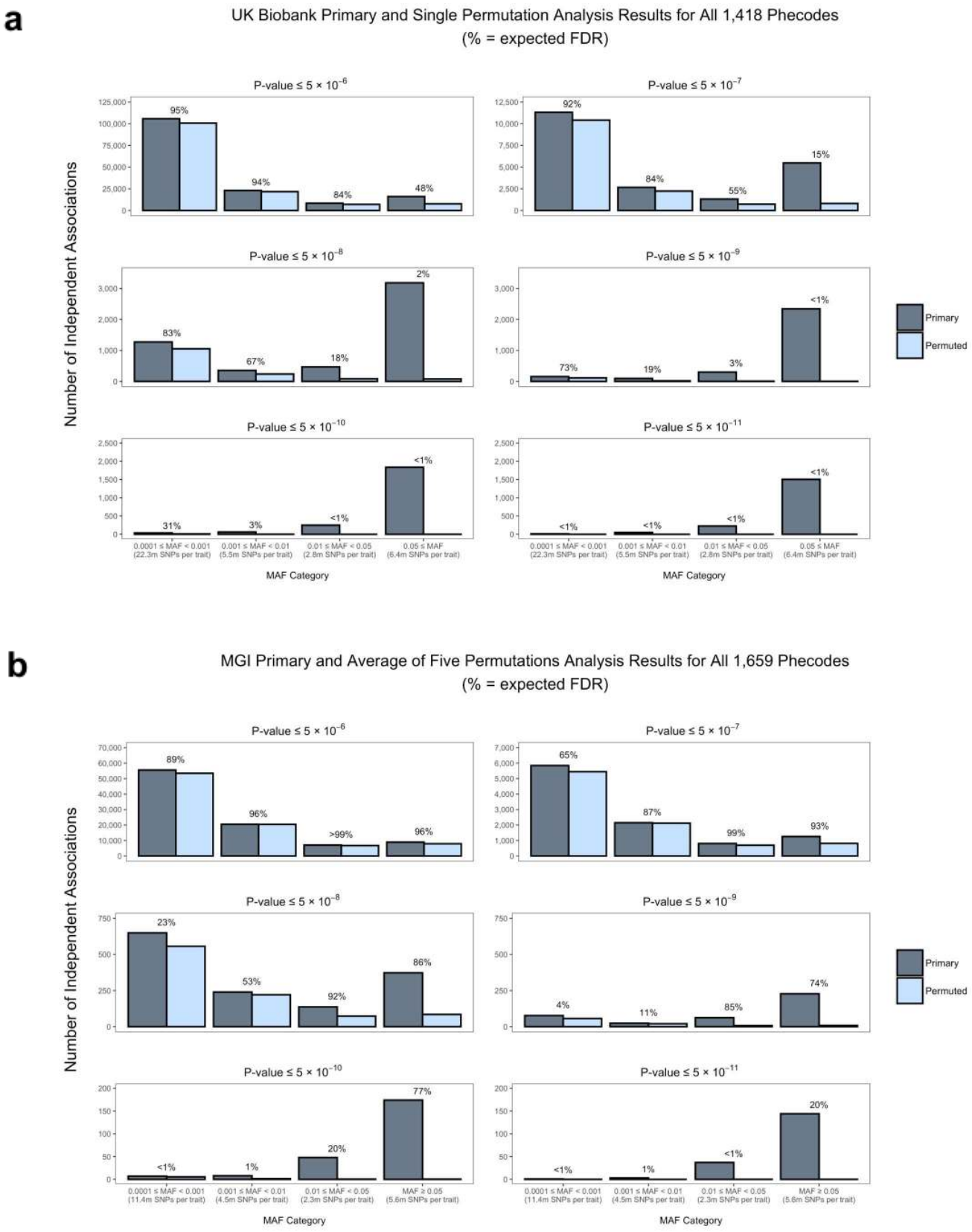

Figure 2. SAIGE independent association results (p≤5×10-6) for primary and permuted data. a) 153,360 primary and 137,224 permuted independent associations for a single permutation of 1,418 UKB phenotypes. b) 92,109 primary and an average of 88,585 permuted independent associations for five permutations of 1,659 MGI phenotypes.

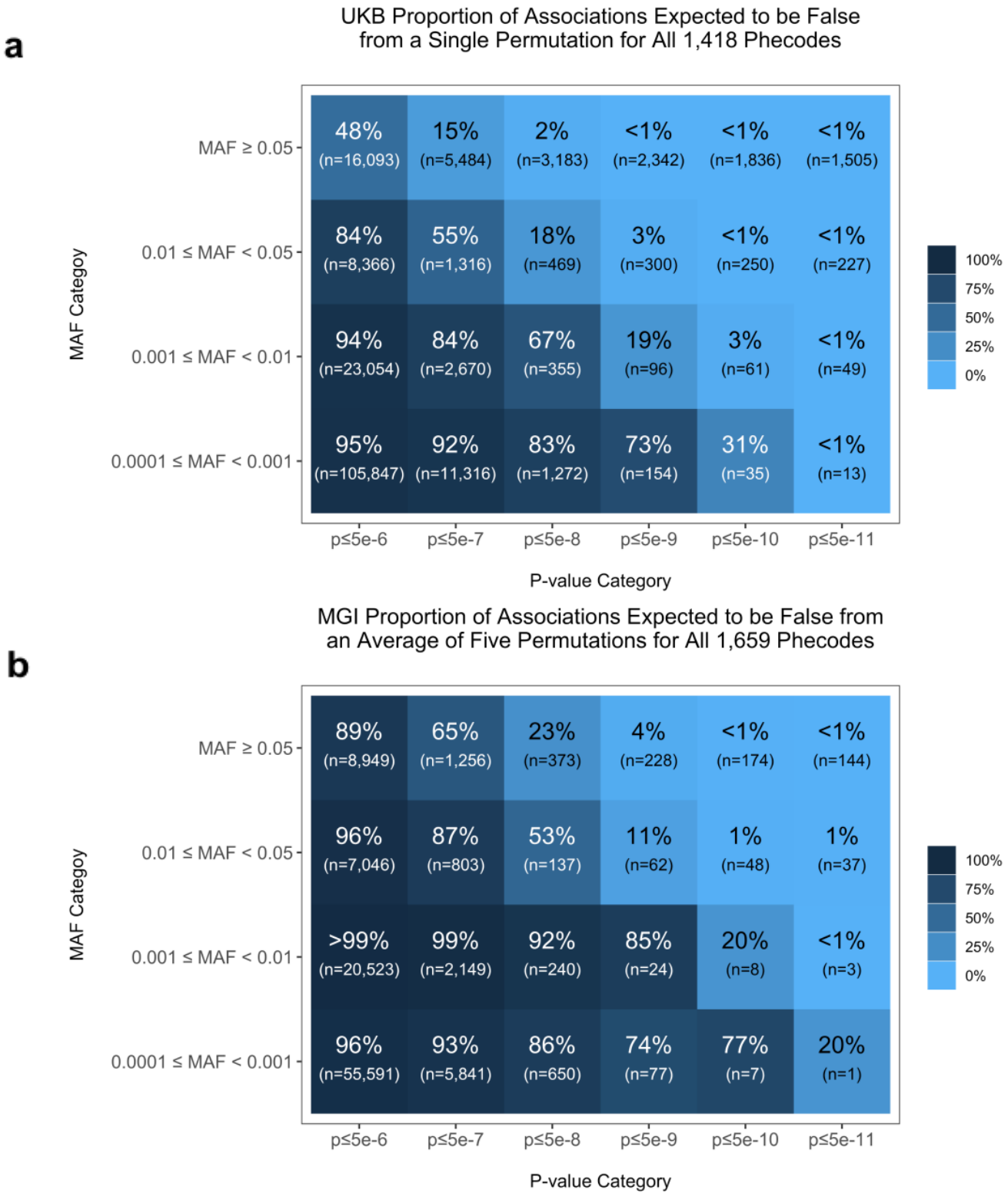

Figure 3. Heatmap of the proportion of independent associations ($p \leq 5 \times 10^{-6}$) expected to be false in the SAIGE analyses by minor allele frequency and p-value category. *n* is the number of primary independent associations in each category. a) Results for a single permutation of 1,418 UKB phenotypes. b) Results for an average of five permutations of 1,659 MGI phenotypes.

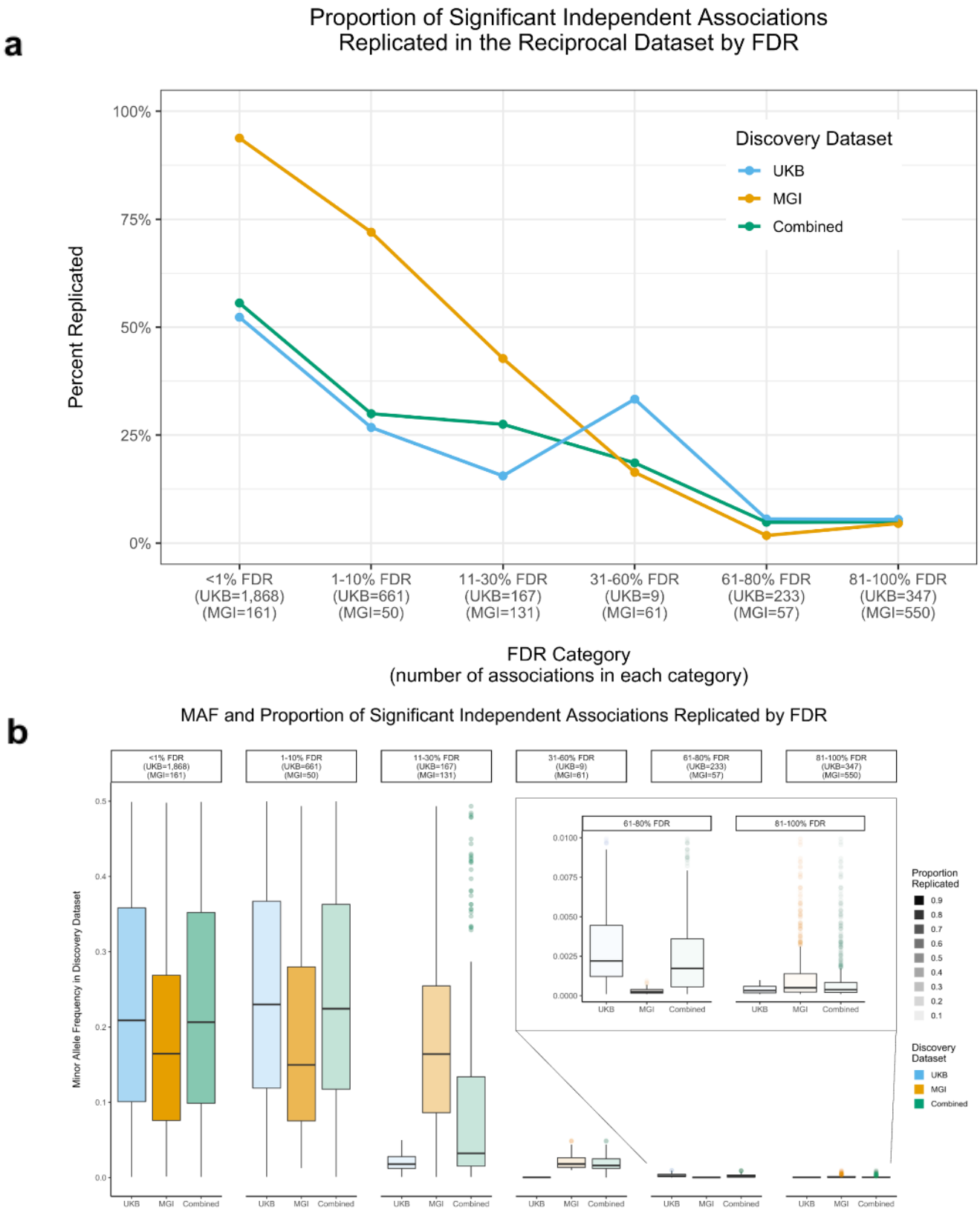

Figure 4. Reciprocal replication rates for significant (p≤5×10-8) independent associations in UKB and MGI. Replication analyses performed for significant independent associations among 1,365 phenotypes common to both datasets. a) Proportion replicated in each dataset by estimated FDR category. b) Proportion replicated in each dataset by minor allele frequency and estimated FDR category. The inset plot is a magnification of the 61-80% and 81-100% FDR group results.

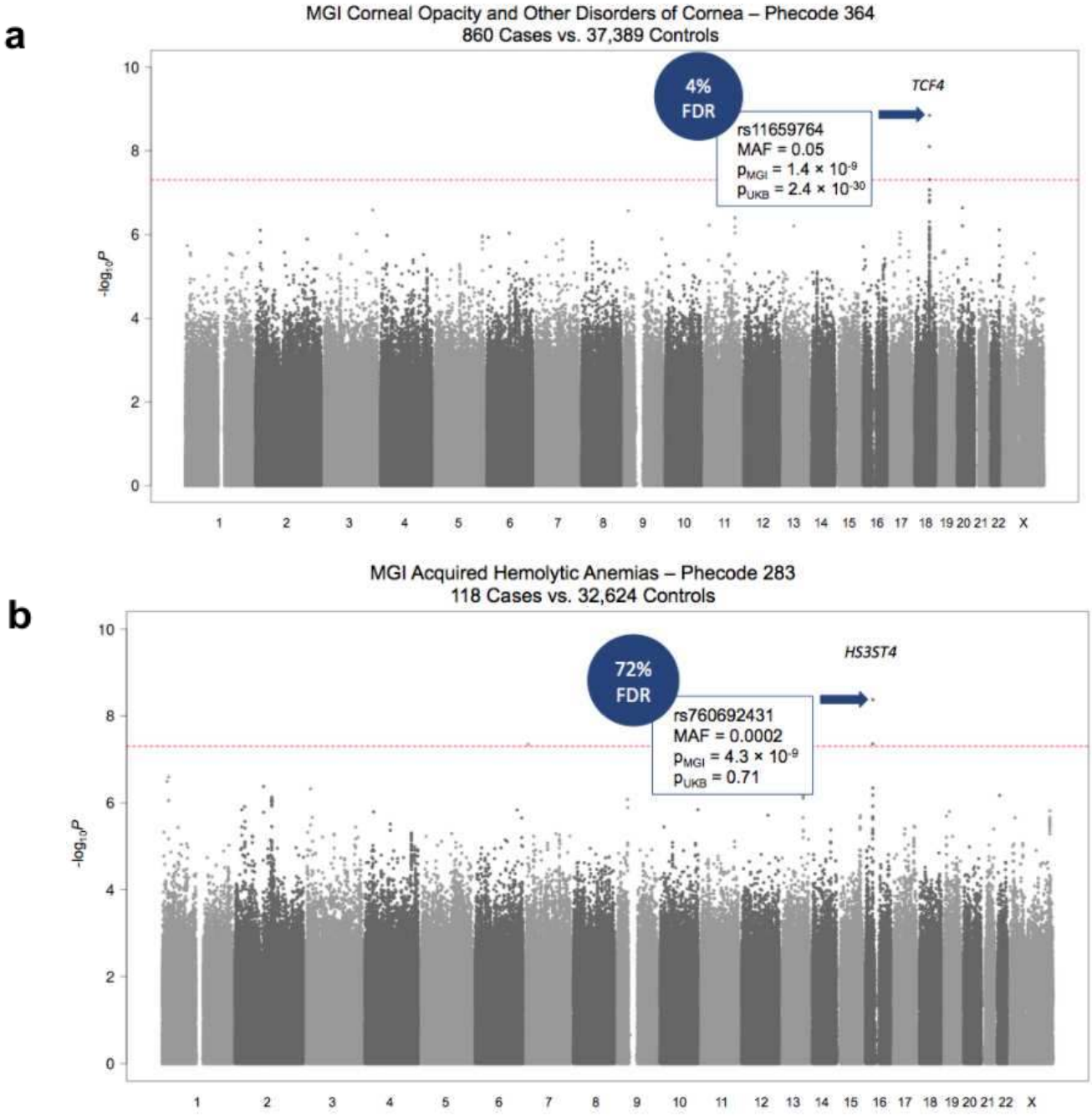

Figure 5. Manhattan plots and estimated FDRs for two MGI phenotypes. rsIDs, MAFs, p-values, and estimated FDRs given for the most significant association in each peak. a) Results for corneal opacity and other disorders of cornea. b) Results for acquired hemolytic anemias.

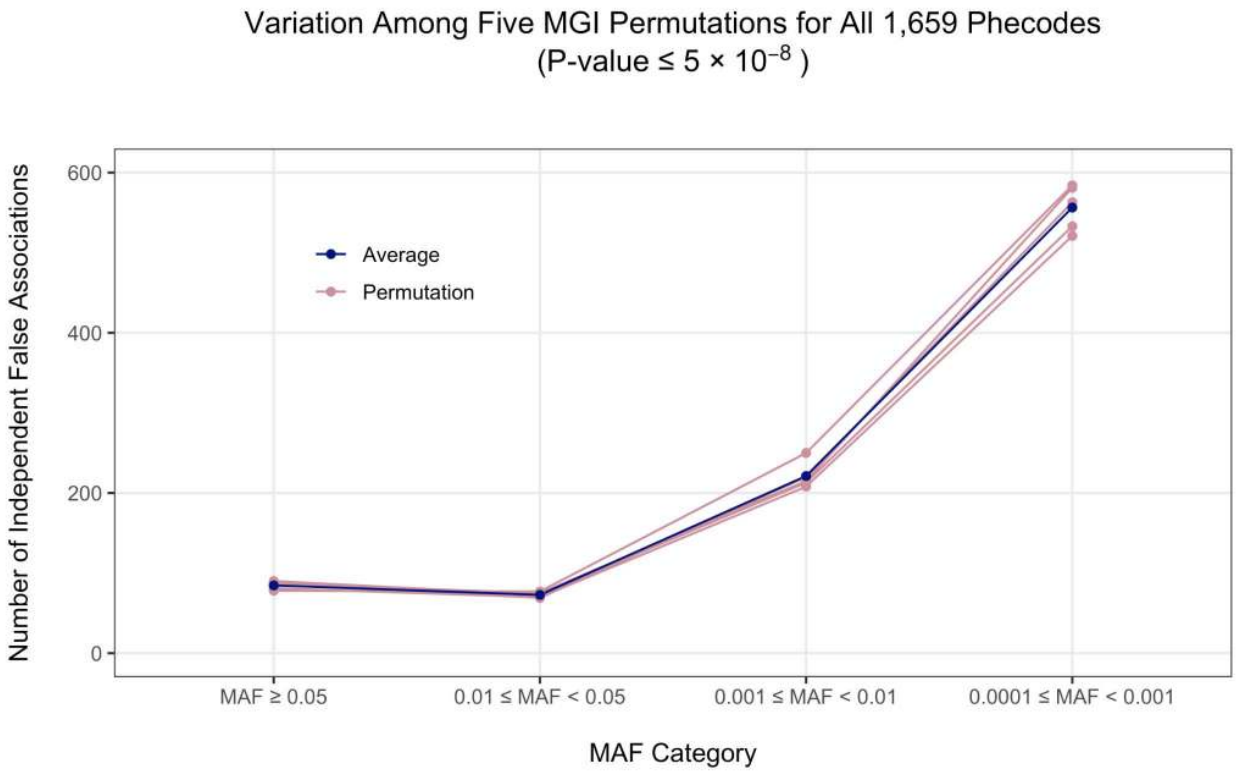

Figure 6. Comparison of significant ($p \leq 5 \times 10^{-8}$) independent false association counts in permuted MGI data. Plot demonstrates the extent of variability among permutations. Pink lines represent false association counts for each of the five permutations by MAF category. Blue line represents average false association counts across the five permutations by MAF category.

a

Replication of Selected UKB Associations in MGI									
rsID	Phenotype	Nearest Gene	UKB Number of Cases (Controls)	MGI Number of Cases (Controls)	UKB MAF	UKB P-value	Estimated FDR	Replicated at $p \leq 0.05$ in MGI	MGI P-value
rs7328654	Cancer of Larynx, Pharynx, Nasal Cavities	*BRCA2*	635 (405,057)	1,047 (39,344)	0.48	3.2×10^{-9}	<1%	✓	0.009
rs2790049	Primary Open Angle Glaucoma	*TMCO1*	1,028 (396,051)	307 (37,389)	0.13	5.4×10^{-9}	2%	✓	0.033
rs121918166	Skin Cancer	*OCA2*	13,603 (393,366)	6,437 (34,899)	0.009	1.7×10^{-9}	19%	✓	0.0003
rs532780200	Hyperosmolality / Hypernatremia	*PALMD*	238 (399,783)	412 (34,452)	0.001	8.0×10^{-9}	67%	✗	0.234
rs764706784	Convulsions	*ACO1*	2,238 (393,432)	1,548 (34,578)	0.0005	6.0×10^{-9}	83%	✗	0.473

b

Replication of Selected MGI Associations in UKB									
rsID	Phenotype	Nearest Gene	MGI Number of Cases (Controls)	UKB Number of Cases (Controls)	MGI MAF	MGI P-value	Estimated FDR	Replicated at $p \leq 0.05$ in UKB	UKB P-value
rs7681423	Pulmonary Heart Disease	*LRAT*	2,619 (38,906)	4,475 (400,268)	0.24	2.3×10^{-9}	3%	✓	2.5×10^{-29}
rs3928325	Posttraumatic Stress Disorder	*H3F3C*	536 (23,601)	113 (363,984)	0.12	4.6×10^{-8}	23%	✓	0.005
rs16891982	Actinic Keratosis	*SLC45A2*	3,403 (36,361)	2,582 (401,728)	0.05	3.8×10^{-8}	51%	✓	0.012
rs575967928	Vitamin B-complex Deficiencies	*ZNF800*	592 (33,939)	753 (404,730)	0.0008	2.6×10^{-8}	86%	✗	0.36
rs1016111760	Osteoarthrosis	*DOCK9*	9,522 (32,589)	28,225 (378,889)	0.0006	4.5×10^{-8}	86%	✗	0.83

Table 1. Replication of selected significant ($p \leq 5 \times 10^{-8}$) independent associations in the UKB and MGI association analyses. Each dataset alternatively acts as a discovery and replication dataset. The rsIDs, case and control counts, MAFs, FDR, and p-values are given for the most significant association within a 1 MB window in the discovery dataset. a) Replication of selected UKB associations in MGI. b) Replication of selected MGI associations in UKB.

Estimated Computation Time and Cost Using SAIGE Software				
	Average CPU Hours per Trait	Total CPU Hours	Total In-House Computation Cost Estimate	Total Web-Based Service Computation Cost Estimate
UKB (1,418 phenotypes)	1,236	1,752,160	$47,307	$35,042
MGI (1,659 phenotypes)	29	48,221	$1,302	$964

Table 2. Estimated CPU hours and cost for permutation analyses of 1,418 UKB and 1,659 MGI phenotypes. In-house computing cluster located at the University of Michigan. Web-based computing cluster is the Google Cloud Platform. Estimates for both clusters given for n1-standard machines.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pdf
- SupplementaryTables.pdf