

1 **The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in**
2 **Michigan Medicine patients**

3 Matthew Zawistowski¹, Lars G. Fritsche¹, Anita Pandit¹, Brett Vanderwerff¹, Snehal Patil¹, Ellen M.
4 Schmidt¹, Peter VandeHaar¹, Chad M. Brummett², Sachin Keterpal², Xiang Zhou¹, Michael Boehnke¹,
5 Gonçalo R. Abecasis^{1,3}, Sebastian Zöllner^{1,4}

6 ¹ Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor
7 Michigan

8 ² Department of Anesthesiology, University of Michigan, Ann Arbor Michigan

9 ³ Regeneron Genetics Center, 777 Old Saw Mill River Road, Tarrytown, NY 10591, USA

10 ⁴ Department of Psychiatry, University of Michigan, Ann Arbor Michigan

11 **Abstract** The recent wave of biobank repositories linking individual-level genetic data with dense clinical
12 health history has introduced a dramatic paradigm shift in phenotyping for human genetic studies. The
13 mechanism by which biobanks recruit participants can vary dramatically according to factors such as
14 geographic catchment and sampling strategy. These enrollment differences leave an imprint on the
15 cohort, defining the demographics and the utility of the biobank for research purposes. Here we
16 introduce the Michigan Genomics Initiative (MGI), a rolling enrollment, single health system biobank
17 currently consisting of >85,000 participants recruited primarily through surgical encounters at Michigan
18 Medicine. A strong ascertainment effect is introduced by focusing recruitment on individuals in
19 Southeast Michigan undergoing surgery. MGI participants are, on average, less healthy than the general
20 population, which produces a biobank enriched for case counts of many disease outcomes, making it
21 well suited for a disease genetics cohort. A comparison to the much larger UK Biobank, which uses
22 population representative sampling, reveals that MGI has higher prevalence for nearly all diagnosis-
23 code-based phenotypes, and larger absolute numbers of cases for many phenotypes. GWAS of these
24 phenotypes replicate many known findings, validating the genetic and clinical data and their proper
25 linkage. Our results illustrate that single health-system biobanks that recruit participants through
26 opportunistic sampling, such as surgical encounters, produce distinct patient profiles that provide an
27 ideal resource for exploring the genetics of complex diseases.

28

29

30 **Introduction**

31 Genome-Wide Association Studies (GWAS) have identified thousands of genetic variants associated with
32 a wide range of human phenotypes (Buniello et al., 2019) . Traditionally, GWAS have been designed with
33 one or a few related traits in mind, where participants are specifically recruited on the basis of those
34 traits. This design strategy optimizes power for those particular traits but has limited reuse potential for
35 studying additional outcomes.

36 The recent wave of biobank repositories linking individual-level genetic data with dense clinical health
37 history has introduced a dramatic paradigm shift in phenotyping for genetic studies (Beesley et al.,
38 2020). Such biobanks allow broad phenotyping based on patient Electronic Health Records (EHRs) across
39 a common set of genotyped samples, allowing investigation of a wide range of clinically important traits
40 within the same cohort. Rather than being optimized for a single trait, the EHR-linked biobank design
41 creates a resource for repeated use across diverse traits and study questions. The rich clinical data
42 provide the ability to fine-tune inclusion criteria and phenotype definitions on a per-study basis using
43 combinations of diagnoses, clinical lab results, medication usage, imaging results, and more. Thus, the
44 same biobank cohort can yield GWAS for thousands of traits, with each GWAS being cost and time
45 effective since participant recruitment, consent, and genotyping are completed in advance and
46 phenotyping is performed on existing clinical data. In addition, biobanks have spawned novel analytic
47 methods that leverage the unique feature of having the entire phenome measured on the same set of
48 samples. For example, the Phenome-Wide Association Study, or PheWAS, tests individual genetic
49 variants for associations across the phenome allowing investigation of comorbid outcomes and
50 pleiotropic genetic effects, again without the need for additional participant recruitment or data
51 collection (Denny et al., 2010).

52 Although biobanks share a common theme of linked clinical and biological data, they are otherwise
53 remarkably heterogeneous across health systems. Differences in population demographics, recruitment
54 strategy and criteria, consent procedures, and data sharing introduce distinct benefits and drawbacks.
55 Large nationwide biobanks such as UK Biobank (UKB) (Bycroft et al., 2017), BioBank Japan (Nagai et al.,
56 2017), and All of Us (Denny et al., 2019) aim to capture a diverse set of individuals across their
57 respective nations using broad geographical recruitment strategies. This population-based recruitment
58 is effective at generating very large sample sizes, with UKB notably containing >500K participants and All
59 of Us aiming for >1 million participants. To achieve these massive sizes, participants are recruited from

60 across multiple sites and/or health systems and can require substantial effort to merge and harmonize
61 the heterogeneous sources of clinical data.

62 An alternative biobank design is localized recruitment within a single site or healthcare system. In this
63 paper we describe the Michigan Genomics Initiative (MGI), a single-healthcare system biobank recruited
64 from patients receiving care at Michigan Medicine, the University of Michigan health system. MGI
65 recruitment began in 2012 with the driving scientific goal of creating a resource to accelerate biomedical
66 and precision health research at the University of Michigan. Recruitment has primarily occurred through
67 the Department of Anesthesiology during inpatient surgical procedures at Michigan Medicine. The
68 preoperative encounter provides a convenient opportunity to obtain patient consent, complete
69 questionnaires, and collect a blood sample. MGI participants consent to linkage of their blood sample,
70 which is subsequently stored in the University of Michigan Central Biorepository, to their existing and
71 future clinical data, including their Michigan Medicine EHR. The consent form, which covers broad
72 research purposes and re-contact potential, is intentionally brief and accompanied by an easy-to-read
73 pamphlet describing the risks and benefits in terms and pictorial descriptions accessible to a public
74 audience to maximize participant understanding of the project (Supplementary Material). Participants
75 complete a baseline questionnaire capturing socio-demographic, pain, and lifestyle information often
76 not captured in traditional EHR data.

77 To date, >85K Michigan Medicine patients have enrolled in MGI. Recruitment is ongoing recruitment
78 and has expanded to include additional studies that complement preoperative enrollment and target
79 other patient populations, thereby broadening the demographic and clinical profile of the cohort.
80 Already, MGI has yielded numerous research contributions including the discovery of novel variants for
81 clinical laboratory traits (Goldstein et al., 2020), PheWAS-based identification of polygenic risk score-
82 trait associations (Fritsche et al., 2018), pharmacogenetic analysis of chemotherapeutic toxicity (Shakeel
83 et al., 2021), and the integration of MGI participants as “external” controls within GWAS (Y. Li & Lee,
84 2021). The large cohort size in conjunction with the collection of rich clinical phenotypes have also
85 allowed for non-genetic studies, such as evaluating the phenotypic characteristics among participants in
86 relationship to preoperative opioid use (Hilliard et al., 2018).

87 As a single-health system biobank, MGI is smaller than most national biobanks and reflects the
88 demographics of a tertiary health system in Ann Arbor, Michigan rather than the demography of the
89 broader US. Moreover, the opt-in recruitment through preoperative encounters produces a non-
90 random sampling of the overall Michigan Medicine health system population (Spector-Bagdady et al.,

91 2021). Although these ascertainment effects distort population measures such as disease prevalence, it
92 introduces distinct advantages as a genetic research resource. Specifically, we show that MGI is enriched
93 for nearly all disease outcomes, even containing larger case counts than UKB for some diseases. This
94 case enrichment mirrors non-random sampling techniques routinely used in GWAS, for example case-
95 control and extreme phenotype designs, that are specifically designed to increase statistical power.
96 Thus, MGI compares favorably as a genetic analysis resource to much larger biobanks, despite being of
97 substantially smaller overall sample size.

98 We provide a description of the MGI cohort, detail our rigorous quality control procedures, and describe
99 GWAS results for 1,547 phenotypes based on diagnosis codes (Figure 1). Our GWAS analysis yielded
100 1,901 genome-wide significant associations across a wide range of traits. Our strongest associations
101 replicate known genotype-phenotype associations, validating genetic and clinical data quality. Our
102 results highlight the important role that single-health system biobanks provide to genetic research, at
103 both the local institution and by broader collaborative efforts such as the Global Biobank Meta-Analysis
104 Consortium and a wide range of specific-trait-focused GWAS meta-analysis consortia.

105 **Methods**

106 *MGI Recruitment and Consent*

107 The Michigan Genomics Initiative (MGI) participants consent to research use of their biospecimens and
108 EHR data, callback for future studies, and linking of their EHR data to national data sources such as
109 medical and pharmaceutical claims data. As of October 15, 2021, 87,623 participants have enrolled in
110 the study. Participants are primarily recruited through the MGI - Anesthesiology Collection Effort (n =
111 71,168) while awaiting a diagnostic or interventional procedure either at a preoperative appointment or
112 on the day of their operative procedure at Michigan Medicine. Additional participants are recruited
113 through the Michigan Predictive Activity and Clinical Trajectories (MIPACT) Study (n = 7,616), the
114 Michigan Genomics Initiative-Metabolism, Endocrinology, and Diabetes (MGI-MEND) Study (n = 4,108),
115 the Mental Health BioBank (MHB2; n = 2,360), the Biobank to Illuminate the Genomic Basis of Pediatric
116 Disease (BIGBiRD; n = 226). The primary Anesthesiology Collection Effort collects blood samples, but the
117 secondary studies collect either blood or saliva.

118 We collect various self-reported demographic data provided by participants as part of routine
119 appointment questionnaires for the health system. Participant age is computed based on self-reported
120 date of birth and defined as age as of April 2020 or age at death if the participant is deceased. Self-

121 reported race is based on a multiple-choice question with options: Caucasian, African American, Asian,
122 American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and Other/Unknown. Self-
123 reported ethnicity is based on a multiple-choice question with options: Hispanic or Latino, Not Hispanic
124 or Latino, and Unknown. Data are collected according to the Declaration of Helsinki principles (World
125 Medical Association, 2013). MGI study participants' consent forms and protocols were reviewed and
126 approved by the University of Michigan Medical School Institutional Review Board (IRB IDs
127 HUM00071298, HUM00148297, HUM00099197, HUM00097962, and HUM00106315). Opt-in written
128 informed consent for broad research purposes and re-contact potential was obtained. The consent form
129 is intentionally brief and accompanied by an easy-to-read pamphlet describing the risks and benefits in
130 terms and pictorial descriptions catered to a general public audience in order to maximize participant
131 understanding of the project (Supplementary Material). Additional details about MGI can be found at
132 (<https://precisionhealth.umich.edu/our-research/michigangenomics/>).

133 *Genetic Data*

134 DNA samples were genotyped by the University of Michigan Advanced Genomics Core on one of two
135 customized versions of the Illumina Infinium CoreExome-24 bead array platform. These array versions
136 have nearly identical 570K marker backbones synthesized in two batches. The array design contains
137 customized probes incorporated to detect candidate variants from GWAS for multiple diseases and
138 traits (~2,700), nonsense and missense variants (~49,000), ancestry informative markers (~3,300), and
139 Neanderthal variants (~5,300) (Surakka et al., 2020).

140

141 We perform sample-level quality control (QC) on a rolling basis, typically in batches of ~576 samples
142 corresponding to six 96 well plates. We estimate pairwise relatedness using KING (v2.1.3) (Manichaikul
143 et al., 2010), and cross-sample contamination using VICES (Zajac et al., 2019). We use PLINK (v1.9) to
144 determine sample level call-rates (Purcell et al., 2007). We exclude individual samples for any of the
145 following : (1) the participant withdraws from the study, (2) genotype-inferred sex does not match the
146 self-reported gender or self-reported gender was missing, (3) sample has an atypical sex chromosomal
147 aberration, (4) kinship coefficient > 0.45 with another participant with a different study ID, (5) sample-
148 level call-rate <99%, (6) sample is a technical duplicate or twin of another sample with a higher call-rate
149 either within the same array or across arrays, (7) estimated contamination level exceeds 2.5%, (8)
150 missingness on any chromosome exceeds 5%, or (9) sample is processed in a DNA extraction batch that
151 is flagged for severe technical problems.

152

153 We merge samples across genotyping batches and apply SNP-level QC procedures. We exclude SNPs
154 with poor intensity separation based on metrics from the GenomeStudio Genotyping Module (GenTrain
155 score < 0.15 or Cluster Separation score < 0.3). We further drop SNPs with overall call-rate < 99% or
156 Hardy Weinberg $p < 10^{-4}$ within each array. To identify potential batch effects between arrays, we test
157 for differences in allele frequency between array versions among unrelated participants of PCA-inferred
158 European ancestry (see below) using the Fisher's Exact Test and exclude variants with p -value < 10^{-3} ,
159 then merging genotype data from the two arrays.

160

161 We estimate the genetic ancestry of participants passing QC using principal component analysis (PCA)
162 and admixture analysis using SNP data for 938 unrelated individuals of known worldwide ancestry from
163 the Human Genome Diversity Panel (HGDP) as ancestry reference samples (J. Z. Li et al., 2008; Wang et
164 al., 2014). We define continental labels for the individual populations based on mappings available from
165 the Center for the Study of Human Polymorphism's website (https://cephb.fr/en/hgdp_panel.php). We
166 first calculate a reference space of worldwide principal components (PCs) for the HGDP samples using
167 PLINK. We then project MGI samples into this space and broadly infer the genetic ancestry of MGI
168 samples based on their proximity to the known HGDP continental labels. We define MGI participants to
169 be of European ancestry if their first two PCs are contained within a circle defined by a radius $1/8$ the
170 distance between the centroid formed by European HGDP samples and the centroid formed between
171 European, East Asian, and African HGDP samples in the PC1 vs. PC2 space (Fritsche et al., 2018). We also
172 estimate the fraction of each MGI participant's genome that originates from European, African, East
173 Asian, Central/South Asian, West Asian, Native American, or Oceanian ancestral HGDP continental
174 populations using ADMIXTURE (v1.3.0) (Alexander et al., 2009). We merge genotypes of MGI
175 participants with the HGDP reference individuals to run ADMIXTURE in supervised mode using the total
176 number of HGDP continental population labels ($K=7$) as a template. We define the ADMIXTURE-based
177 majority global ancestry for each MGI participant as the largest Q value (ancestry fraction) reported by
178 ADMIXTURE (Supplementary Figure 2).

179

180 We phase the full set of merged genotype samples using EAGLE (v2.4.1) (Loh et al., 2016) without the
181 use of a reference panel ("within-cohort" phasing). We then impute samples with both the Haplotype
182 Reference Consortium (HRC) reference panel (64,940 predominantly European haplotypes containing
183 40,457,219 genetic variants) (McCarthy et al., 2016) and the Trans-Omics for Precision Medicine
184 (TOPMed) reference panel (194,512 ancestrally diverse haplotypes containing 308,107,085 genetic

185 variants) (Taliun et al., 2021). We measure imputation quality using the estimate of imputation accuracy
186 (Rsq) and the squared correlation between imputed and true genotypes (EmpRsq) metrics produced by
187 the imputation software Minimac4 (v1.0.0) (Howie et al., 2012).

188

189 *Clinical Phenotype Data*

190 We extract all available ICD 9 and 10 diagnosis codes for MGI participants from the Michigan Medicine
191 EHR. These codes are mapped to binary phecode phenotypes based on ICD inclusion and exclusion
192 criteria using the PheWAS R package v0.99.5.-5 (Carroll et al., 2014). We use the default PheWAS
193 package requirements for case and control definitions: cases require two instances of an inclusion ICD
194 code and controls have neither inclusion nor exclusion ICD codes. We also account for sex-specific
195 phenotypes using the restrictPhecodesByGender() function and the genotype-inferred sex.

196 *Genetic Analysis*

197 We performed GWAS in MGI samples of genetically inferred European ancestry on 1,712 phecode traits
198 with case count ≥ 20 . The GWAS cohort contains 51,583 MGI participants, including 49,689 with inferred
199 European ancestry by the HGDP projection PCA and an additional 1,894 participants with inferred
200 majority European ancestry by ADMIXTURE, but not identified as East Asian or African by the projection
201 PCA. GWAS were run on the TOPMed-imputed genetic dataset using a mixed model implemented in
202 SAIGE v0.43.3 to account for relatedness and case-control imbalance (Zhou et al., 2020). For each
203 phecode trait, we analyze variants with minor allele frequency (MAF) $>0.01\%$ and adjusted for age,
204 inferred sex, genotyping array, and the first ten genetic PCs. We compute the genomic control inflation
205 factor for the GWAS of each phecode trait to assess stratification and test inflation (Devlin & Roeder,
206 1999). To identify near-independent genome-wide significant loci within each GWAS, we extract all SNPs
207 with p-value $< 5e-8$ and create 1Mb intervals centered around each resulting SNP. Overlapping intervals
208 are combined and we report the SNP with the lowest p-value from each of the resulting intervals as the
209 genome-wide significant peak SNP.

210 We compared the 30 associations with smallest p-value for variants with MAF $>1\%$ with associations
211 reported in the GWAS Catalog (flat file downloaded August 16, 2021) (Buniello et al., 2019). We
212 considered only associations in the GWAS Catalog that had a minimum reported p-value $< 5e-10$ to limit
213 potential false positives within the Catalog. We defined an exact regional match as Catalog associations
214 reported at the same chromosomal location as the peak SNP. If an exact positional match was found, we
215 manually scanned the list of Catalog associations for the same or a clinically similar phenotype to the

216 corresponding phecode trait that produced the genome-wide significant association in MGI. If multiple
217 related traits were reported in the Catalog for that SNP, we reported the trait with lowest p-value *except*
218 in one case where the top association appeared to be a sub-analysis that was more specific than our
219 definition (e.g. for rs4148325 associated with "Disorders of bilirubin excretion," we reported "Bilirubin
220 levels" as the GWAS Catalog match which had $p=5e-62$ in the Catalog, even though the Catalog also
221 listed this SNP for "Bilirubin levels in extreme obesity" at $p=5e-93$). If an exact positional match was not
222 found, we expanded our search to a 50kb window surrounding the peak SNP and followed the same
223 protocol. In only one case was an association not found within a 50kb window and we expanded to a
224 1MB region for this association.

225 Genetic effect sizes computed using SAIGE are biased for low frequency variants. We therefore
226 estimated effect size for these 30 top associations using the exact firth logistic regression implemented
227 by REGENIE v2.2.4 (Mbatchou et al., 2021). We ran REGENIE with settings for $-b$ size 100 in step 1 and $-$
228 b size 200, $--p$ Thresh 0.99, and $-f$ irth in step 2. For both steps, we provided a covariate file with age,
229 inferred sex, genotyping array, and the first ten genetic PCs.

230 *Phecodes in UK Biobank*

231 We computed phecodes for 408,595 individuals of White British ancestry with high-quality genetic data
232 in the UK Biobank (UKB). We used ICD codes and genotyped derived data from open-access UK Biobank
233 data. UK Biobank received ethical approval from the NHS National Research Ethics Service North West
234 (11/NW/0382). We conducted these analyses under UK Biobank data application number 24460.

235 We excluded samples which were flagged by the UK Biobank quality control documentation (Resource
236 531) as (1) "het.missing.outliers", (2) "putative.sex.chromosome.aneuploidy", (3) "excess.relatives", (4)
237 "excluded.from.kinship.inference", (5) the reported gender ("Submitted.Gender") did not match the
238 inferred sex ("Inferred.Gender"), (6) withdrew from the UKB study and (7) were not included in the phased
239 and imputed genotype data of chromosomes 1-22, and X ("in.Phasing.Input.chr1_22 and
240 in.Phasing.Input.chrX"). Furthermore, we reduced the data to samples of White British ancestry (see UK
241 Biobank Resource 531, "in.white.British.ancestry.subset"). We used the PheWAS R package to aggregate
242 the ICD9 and ICD10 codes into phecode traits, requiring one inclusion code for case definitions.

243

244

245 **Results**

246 As of October 15, 2021, 87,623 patients receiving care at the Michigan Medicine health system have
247 consented to participate in the Michigan Genomics Initiative. Participants are recruited on a rolling basis
248 and genotyped in batches at the University of Michigan's Advanced Genomics Core. Enrollment has
249 steadily increased since project initiation, beginning at approximately 730 newly enrolled participants
250 per month in 2013 to just over 1000 per month in 2019, prior to suspension of enrollment in 2020 due
251 to the pandemic (Figure 2A). Notably, enrollment of individuals who self-report their race as something
252 other than Caucasian has likewise increased, from 71 participants per month in 2013 to 292 per month
253 in 2019. In this paper, we describe the genetic and clinical data for MGI freeze 3 comprised of 57,055
254 participants and present results from GWAS for 1,547 traits in a set of 51,583 inferred European
255 samples.

256 **Demographic and Clinical Description of the Cohort**

257 MGI participants range in age from 18 to over 90 years (Table 1). There are slightly more female
258 participants (53%), and male participants are slightly older (58.4 vs 54.7 years, Figure 2B). Most
259 participants self-report race as Caucasian (N=49,605, 87%), with African American (N=3,223, 5.6%) and
260 Asian (N=1,324, 2.3%) next most common; 805 (1.4%) individuals report Hispanic or Latino ethnicity.

261 A wealth of clinical data recorded in the Michigan Medicine EHR are available to develop phenotypes for
262 MGI participants. In this paper we consider a broad set of traits defined using ICD 9 and 10 codes, but
263 clinical laboratory results, medication history and additional contents of the electronic medical files are
264 also available to approved researchers. The number of ICD codes differed across participants (median:
265 604; mean: 1494; 25th percentile: 229; 75th percentile: 1643), reflecting inter-individual differences in
266 overall health and utilization of the health system. We computed a follow-up time measurement,
267 defined as the difference in time between the oldest and most recent ICD diagnoses for an individual, to
268 measure the length of time each participant has interacted with the Michigan Medicine healthcare
269 system. The distribution of follow-up time is U-shaped, with the most frequent follow-up times being <1
270 year and ~19 years (Figure 2C). The upper bound of 20 years corresponds to the beginning of electronic
271 capture of diagnosis codes beginning at Michigan Medicine in 2000. Age distribution among individuals
272 is almost identical among all categories of follow-up time (Figure 2C), suggesting that follow-up time is
273 largely independent of participant age.

274

275 **Phecode Traits**

276 Due to the granularity and redundancy of ICD codes, we mapped individual ICD codes to broader binary
277 phecode traits using the PheWAS software (Carroll et al., 2014). Individual phecode traits can be
278 grouped into 17 general categories of clinically similar traits. For example, hypertension (phecode 401),
279 myocardial infarction (411.2) and myocarditis (420.1) are each mapped to the ‘Circulatory System’
280 phecode group. In total, we observed case samples for 1,817 phecode traits, with 1,712 traits having at
281 least 20 cases (Table 2, Supplementary Table 1). The most common traits are related to high prevalence
282 diseases (Figure 2D), including hypertension (phecode 401, 401.1), lipid disorders (272, 272.1), obesity
283 (278, 278.1), esophagus/GERD (530, 530.1, 530.11), and mental health disorders (mood disorders: 296;
284 anxiety: 300, 300.1; depression: 296.2). Several pain related traits (pain in joint: 745; abdominal pain:
285 785; pain: 338; back pain: 760) also appear among the most common phecodes, likely due in part to the
286 enrollment of surgical patients through anesthesiology. The number of phecodes per sample was
287 strongly right skewed (median: 31; mean: 44.2; maximum: 435) and positively correlated with both age
288 (Figure 2E) and follow-up time (Figure 2F).

289 As a single-health system biobank, MGI is smaller than some biobanks employing broader national
290 recruitment strategies. UKB, for example, boasts nearly half a million participants. When comparing
291 GWAS cohorts of European inferred ancestry in each biobank (see Methods), MGI has a higher
292 prevalence for nearly all phecode traits compared to UKB (Supplementary Figure 1). Of the 1,772
293 phecode traits for which either MGI or UKB had at least one case, UKB has no cases for 354 and MGI has
294 no cases for 22, many of which are common conditions. For example, there are no phecode-defined
295 cases in UKB for basal cell carcinoma (172.21), insulin pump user (250.3), and hypo- (275.51) and
296 hypercalcemia (275.6). The missing cases for these traits reflect different ICD code systems or
297 differential use of ICD codes between the two biobanks rather than an actual lack of these traits in the
298 cohorts.

299 As power of association studies depends strongly on the number of cases, it is more helpful to compare
300 the overall number of cases between MGI and UKB. MGI has a higher case count for 557 (41%) of the
301 1,358 phecodes for which both biobanks have cases (Figure 3). MGI has traits with greater case counts
302 across all phecode categories, particularly within endocrine/metabolic and neurological categories.
303 There are 48 phecode traits for which MGI has over 10-fold number of cases found in UKB
304 (Supplementary Table 2), including “Vitamin D deficiency” (phecode: 261.4), “pain” (phecode: 338),
305 “migraine with aura” (phecode: 340.1), “insomnia” (phecode: 327.4), and “varicella infection” (phecode:

306 079.1). Phecode traits for which MGI has more cases than UKB and a case count >10K, including
307 overweight/obesity (278, 278.1), mood disorders (296), depression (296.2), anxiety (300, 300.1), sleep
308 apnea (327.3), allergic rhinitis (476), other symptoms of respiratory system (512), pain (338), pain in
309 joint (745), and back pain (760).

310 **Genetic Data**

311 Overall, genetically inferred ancestry is consistent with self-reported race and ethnicity obtained from
312 appointment intake surveys (Figure 4A). The majority of participants that self-report as Caucasian
313 clustered with European HGDP populations at the top of the familiar continental PCA plot. Nearly all
314 self-reported African American participants in MGI cluster between the HGDP African and European
315 reference populations, consistent with admixture between those groups. Self-reported Asian
316 participants show two distinct clusters corresponding to East Asian and Central/Southern Asian HGDP
317 populations. As expected, participants that reported Hispanic/Latino ethnicity overwhelmingly appear
318 between European and Asian continental populations (Bryc et al., 2010).

319 Genotype imputation increases the number of variants in the dataset and is dependent upon the
320 haplotype reference panel. Our current data freeze uses the TOPMed reference panel which
321 substantially increases both the number and quality of imputed variants. Imputation using TOPMed
322 produces ~52 million variants post QC-filtering compared to ~32 million using the HRC reference panel,
323 with the largest gain in imputable variants at the lower end of the allele frequency spectrum (Figure 4B);
324 TOPMed imputation results in 45,399,294 variants with MAF between 0.01% and 5% and imputation R_{sq}
325 > 0.3, compared to 26,769,074 of such variants based on HRC. Moreover, TOPMed-imputed variants are
326 more accurate across the frequency spectrum, particularly for variants with MAF < 5% (Figure 4C).
327 Comparing the reference panels across samples from different ancestries reveals that the increased
328 diversity in TOPMed reference haplotypes leads to increased imputation accuracy in all non-European
329 samples (Figure 4D). MGI samples with majority African ancestry based on ADMIXTURE showed the
330 largest improvement in imputation accuracy, even for common variants, reflecting the large proportion
331 of African American individuals in TOPMed compared to HRC. We observe a more modest increase in
332 accuracy among majority Asian ancestry MGI samples, likely because TOPMed contains comparatively
333 fewer Asian haplotypes.

334

335

336 GWAS Results

337 We initially conducted GWAS for the 1,712 phecode traits with at least 20 cases in the set of 51,583 MGI
338 samples with genetically inferred European ancestry, across 51.8M SNPs with MAF > 0.01% and
339 imputation score $R_{sq} > 0.3$. We evaluated genomic control values and find that traits with less than 60
340 cases were highly susceptible to inflation (Supplementary Figure 3-4). Thus, we present results for the
341 1,547 traits with ≥ 60 cases (Table 2). We identified 1,901 distinct genome-wide significant loci across
342 977 phecode traits, including at least one genome-wide significant association within each of the
343 seventeen phecode categories. Many of the associations occur at low frequency SNPs which have higher
344 false positive rates at the standard $5e-8$ threshold for genome-wide significance (Annis et al., 2021).
345 Among SNPs with MAF > 1%, we observe 606 associations across the 340 traits. The complete set of
346 genetic analyses described in this paper are viewable through an interactive “PheWeb” tool that
347 includes GWAS summary statistics, regional association plots and PheWAS analyses that can be used for
348 replication and hypothesis-driven look-ups by the research community (See Resources).

349 To assess the quality of our genetic data and phecode traits, we compared our thirty most significant
350 associations among MAF>1% variants to previously identified associations reported in the GWAS Catalog
351 (Table 3). These thirty associations occur among 15 unique SNPs because seven SNPs were associated
352 with multiple related phecode traits, reflecting the hierarchical nature of ICD coding. For 10 of the SNPs,
353 we observed an association with a related trait in the GWAS Catalog at the exact chromosomal location.
354 Four SNPs had a relevant association in the GWAS Catalog within a 50kb window. The one association
355 for which we did not observe a close phenotypically relevant association within the GWAS Catalog was
356 for the indel rs113993960 (chr7:117559590:ATCT:A) and cystic fibrosis (phecode 499). The indel is a
357 known low frequency, pathogenic inframe shift within *CFTR* (NC_000007.14(CFTR_i001):p.(Phe538del),
358 (VCV000007105.43 - ClinVar - NCBI, 2021).

359 Our strongest association occurred between rs6025 (chr1:169549811;
360 NC_000001.11(F5_i001):p.(Arg397Gln)) and primary hypercoagulable state (phecode: 286.81). This
361 missense variant in the *F5* gene is among our top associations for multiple phecode traits related to
362 coagulation (286.8: hypercoagulable state, 286: coagulation defects, 286.7: other and unspecified
363 coagulation defects, 286.12: congenital deficiency of other clotting factors (including factor VII)).
364 Associations between rs6025 and venous thromboembolism (Klarin et al., 2019) and thrombosis have
365 previously been reported (Hinds et al., 2016). rs143260331 was associated with two nested atrial
366 fibrillation phecode traits (427.2 and 427.21) and was nearby previous associations for atrial fibrillation

367 and flutter. We also observed several strong associations between SNPs in the HLA locus and phecodes
368 related to type 1 diabetes. These associations have been reported for related traits in the GWAS
369 Catalog. For example, we observed an association between rs9273368 (chr6:32658525;
370 NC_000006.12(HLA-DQB1_v001):c.*1711A>C) with the phecode 250.1: type 1 diabetes (4.23e-106),
371 which has been previously reported for diabetes medication use (Wu et al., 2019). Broadly, our results
372 replicate known signals, indicating that phenotyping and genotyping in MGI enable well-calibrated
373 GWAS.

374 **Discussion**

375 Biobanks are an efficient strategy to generate large samples for modern genetics research. The biobank
376 approach leverages central genotyping and QC to provide a single resource that can be used for a wide
377 range of research questions. Not only does this result in cost-efficient research in general, but it also
378 particularly empowers researchers who lack resources to create large datasets of their own. To provide
379 such a broadly useful resource requires using state-of-the-art QC as each error may affect multiple
380 independent analyses. Consistent with previous findings (Taliun et al., 2021), we show that e.g. using
381 TOPMed reference panels rather than HRC provides a boost in both number and accuracy of imputed
382 markers, with particularly meaningful gains in non-European samples. Importantly, our results highlight
383 the value of diverse haplotype imputation in a real-world dataset of US samples recruited without
384 regard to ancestry.

385
386 MGI exists within a broad family of single-health system biobanks, e.g. at Vanderbilt University (Roden
387 et al., 2008), Geisinger (Carey et al., 2016) and UCLA (Johnson et al., 2021). Even within this group major
388 differences exist in recruitment strategy. For example, BioVU recruitment includes opportunistic
389 inclusion of patients with existing blood specimens collected during prior clinical testing at the
390 Vanderbilt University Medical Center (Roden et al., 2008). This strategy has implications for the size and
391 composition of the cohort that will differ from MGI enrollment that requires prospective collection of
392 blood samples. Like all single-health system biobanks, the cohort demographics will naturally reflect the
393 patient population served by the health system. In the case of MGI, the cohort largely comes from the
394 community and thus overrepresents individuals of European ancestry relative to both the population of
395 Michigan and the US. Moreover, the MGI cohort itself is less diverse in terms of age, sex, race, ethnicity,
396 and socioeconomic status than the overall clinical population at Michigan Medicine (Spector-Bagdady et
397 al., 2021). Underrepresentation of minority individuals in particular can lessen generalizability of results

398 and exacerbate existing health inequities (Landry et al., 2018; Sirugo et al., 2019). For these reasons, we
399 are currently seeking to enrich enrollment of underrepresented populations, notably the Middle Eastern
400 – North African and African American populations of southeast Michigan, by leveraging epidemiological
401 studies in minority populations and by using the Michigan Health Care patient portal for targeted
402 recruitment.

403 Our results show that recruitment within a tertiary care center, primarily among surgical patients,
404 results in substantial case enrichment compared to the general health system population as well as
405 larger population-sampled biobanks. This case enrichment mirrors non-random sampling techniques
406 routinely used in GWAS, for example case-control and extreme phenotype designs, that are specifically
407 designed to increase statistical power. The implication is that MGI provides powerful GWAS testing
408 despite not being among the largest biobanks (Zhou et al., 2021). Although some of the case count
409 differences identified between MGI and UKB are likely the result of differing diagnostic coding criteria,
410 they nevertheless still reflect the ability to identify cases within the data.

411 Our analysis identified unique features of the cohort that can in part be connected to the strong surgical
412 enrollment bias. The distribution of participant follow-up time suggests that MGI is a mixture of long-
413 time, regular users of Michigan Medicine with lengthy follow-up times, and new possibly one-time
414 patients with modest follow-up times. It is possible that participants with short follow-up times are
415 utilizing the health system for the first time during the surgical procedure in which they enrolled in MGI.
416 Patient age was relatively consistent across follow-up times, but patients with longer follow-up times
417 had higher numbers of phecode case assignments. It is possible that participants with longer follow-up
418 times, despite being of similar age, simply have more health problems. Alternatively, participants with
419 shorter follow-up times might have incomplete medical history within the Michigan Medicine EHR, a
420 plausible scenario for out-of-system enrollees receiving one-time specialized care at UM. For these
421 participants, we may be misclassifying them as controls for traits missing diagnoses in the Michigan
422 Medicine EHR.

423 Phenotype development from EHRs requires interpretation of dense administrative data. For first-pass
424 phenotyping, the PheWAS software provides a convenient approach to map the granular ICD codes to
425 broader phecode traits. The advantage of this technique is rapid and automated generation of the
426 phenome across all individuals in a biobank. Our GWAS results indicate that phecodes are an effective
427 tool for broad phenotyping at the phenome-scale. Importantly the PheWAS software provides a realistic
428 strategy for consistent large-scale phenotyping across biobanks. The ICD mappings however are often

429 not sufficiently precise to correctly identify cases or controls with perfect sensitivity. The phecode
430 system also neglects clinical data sources like laboratory results, physician notes and medication history
431 that can be informative for elucidating true disease status. Further, ICD usage differs among health
432 systems, which impacts the sensitivity of the phecode approach. To maximize power and obtain
433 unbiased effect size estimates for specific traits, it may be advantageous to carefully extract all relevant
434 information from the EHR data and apply validated electronic phenotype algorithms, for example, as
435 described by the Phenotype KnowledgeBase (<https://phekb.org>).

436

437 MGI represents the important class of single-health system biobank in the emerging field of EHR-based
438 genomics. We have shown that a biobank recruited from within a single-health system can strategically
439 recruit large sample sizes and provide an excellent multi-purpose resource for genetic research. With a
440 sample size expected to top 100,000 participants by 2022, we anticipate that MGI will play an important
441 role in future research both at the University of Michigan as well as the broader community. To date,
442 MGI data has been used in over 30 peer reviewed publications, which can be viewed at our website:
443 <https://precisionhealth.umich.edu/our-research/michigangenomics/publications/>

444

445

446 **Acknowledgements**

447 The authors acknowledge the Michigan Genomics Initiative participants, Precision Health at the
448 University of Michigan, the University of Michigan Medical School Central Biorepository, and the
449 University of Michigan Advanced Genomics Core for providing data and specimen storage, management,
450 processing, and distribution services, and the Center for Statistical Genetics in the Department of
451 Biostatistics at the School of Public Health for genotype data curation, imputation, and management in
452 support of the research reported in this publication. We thank Ruth Johnson for her careful reading of
453 the manuscript and thoughtful feedback.

454 **Resources**

455 The GWAS results presented in this paper are viewable in an interaction PheWeb browser tool:
456 <https://pheweb.org/MGI-freeze3/>. Email mgipheweb@umich.edu to receive access to the pheweb.
457 A curated list of research papers using the MGI resource is available at:
458 <https://precisionhealth.umich.edu/our-research/michigangenomics/publications/>.

459 **Author Contributions**

460 Conceptualization: CMB, SK, GRA; Software: PV, SP; Formal Analysis: MZ, LGF, AP, BV, EMS; Resources:
461 CMB, SK, GRA; Data Curation: LGF, AP, BV; Writing - Original Draft: MZ, LGF, AP, BV, XZ MB, SZ;
462 Visualization: MZ, LGF, AP, BV, EMS; Supervision: MZ, LGF, XZ, MB, SZ; Funding Acquisition: CMB, GRA

463

464 References

- 465 1. Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in
466 unrelated individuals. *Genome Research*, *19*(9), 1655–1664.
467 <https://doi.org/10.1101/gr.094052.109>
- 468 2. Annis, A., Pandit, A., LeFaive, J., Taliun, S. G., Fritsche, L., VandeHaar, P., Boehnke, M.,
469 Zawistowski, M., Abecasis, G., & Zöllner, S. (2021). *False discovery rates for genome-wide*
470 *association tests in biobanks with thousands of phenotypes*. [https://doi.org/10.21203/rs.3.rs-](https://doi.org/10.21203/rs.3.rs-873449/v1)
471 [873449/v1](https://doi.org/10.21203/rs.3.rs-873449/v1)
- 472 3. Beesley, L. J., Salvatore, M., Fritsche, L. G., Pandit, A., Rao, A., Brummett, C., Willer, C. J.,
473 Lisabeth, L. D., & Mukherjee, B. (2020). The emerging landscape of health research based on
474 biobanks linked to electronic health records: Existing resources, statistical challenges, and
475 potential opportunities. *Statistics in Medicine*, *39*(6), 773–800.
476 <https://doi.org/10.1002/sim.8445>
- 477 4. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M.,
478 Bustamante, C. D., & Ostrer, H. (2010). Colloquium paper: Genome-wide patterns of population
479 structure and admixture among Hispanic/Latino populations. *Proceedings of the National*
480 *Academy of Sciences of the United States of America*, *107* Suppl 2, 8954–8961.
481 <https://doi.org/10.1073/pnas.0914618107>
- 482 5. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon,
483 A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousitou, O., Whetzel, P. L., Amode, R.,
484 Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., ... Parkinson, H. (2019). The NHGRI-
485 EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary
486 statistics 2019. *Nucleic Acids Research*, *47*(Database issue), D1005–D1012.
487 <https://doi.org/10.1093/nar/gky1120>
- 488 6. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D.,
489 Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., McVean, G., Leslie, S., Donnelly, P., & Marchini,
490 J. (2017). *Genome-wide genetic data on ~500,000 UK Biobank participants* (p. 166298).
491 <https://doi.org/10.1101/166298>
- 492 7. Carey, D. J., Fetterolf, S. N., Davis, F. D., Faucett, W. A., Kirchner, H. L., Mirshahi, U., Murray, M.
493 F., Smelser, D. T., Gerhard, G. S., & Ledbetter, D. H. (2016). The Geisinger MyCode Community
494 Health Initiative: An electronic health record-linked biobank for Precision Medicine research.
495 *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *18*(9), 906–
496 913. <https://doi.org/10.1038/gim.2015.187>
- 497 8. Carroll, R. J., Bastarache, L., & Denny, J. C. (2014). R PheWAS: Data analysis and plotting tools for
498 phenome-wide association studies in the R environment. *Bioinformatics*, *30*(16), 2375–2376.
499 <https://doi.org/10.1093/bioinformatics/btu197>
- 500 9. Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang,
501 D., Masys, D. R., Roden, D. M., & Crawford, D. C. (2010). PheWAS: Demonstrating the feasibility

- 502 of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9), 1205–
503 1210. <https://doi.org/10.1093/bioinformatics/btq126>
- 504 10. Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., & Dishman,
505 E. (2019). The “All of Us” Research Program. *The New England Journal of Medicine*, 381(7), 668–
506 676. <https://doi.org/10.1056/NEJMs1809937>
- 507 11. Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4), 997–
508 1004. <https://doi.org/10.1111/j.0006-341x.1999.00997.x>
- 509 12. Fritsche, L. G., Gruber, S. B., Wu, Z., Schmidt, E. M., Zawistowski, M., Moser, S. E., Blanc, V. M.,
510 Brummett, C. M., Kheterpal, S., Abecasis, G. R., & Mukherjee, B. (2018). Association of Polygenic
511 Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics
512 Initiative. *American Journal of Human Genetics*, 102(6), 1048–1061.
513 <https://doi.org/10.1016/j.ajhg.2018.04.001>
- 514 13. Goldstein, J. A., Weinstock, J. S., Bastarache, L. A., Larach, D. B., Fritsche, L. G., Schmidt, E. M.,
515 Brummett, C. M., Kheterpal, S., Abecasis, G. R., Denny, J. C., & Zawistowski, M. (2020). LabWAS:
516 Novel findings and study design recommendations from a meta-analysis of clinical labs in two
517 independent biobanks. *PLOS Genetics*, 16(11), e1009077.
518 <https://doi.org/10.1371/journal.pgen.1009077>
- 519 14. Hilliard, P. E., Waljee, J., Moser, S., Metz, L., Mathis, M., Goesling, J., Cron, D., Clauw, D. J.,
520 Englesbe, M., Abecasis, G., & Brummett, C. M. (2018). Prevalence of Preoperative Opioid Use
521 and Characteristics Associated With Opioid Use Among Patients Presenting for Surgery. *JAMA*
522 *Surgery*, 153(10), 929–937. <https://doi.org/10.1001/jamasurg.2018.2102>
- 523 15. Hinds, D. A., Buil, A., Ziemek, D., Martinez-Perez, A., Malik, R., Folkersen, L., Germain, M.,
524 Mälarstig, A., Brown, A., Soria, J. M., Dichgans, M., Bing, N., Franco-Cereceda, A., Souto, J. C.,
525 Dermitzakis, E. T., Hamsten, A., Worrall, B. B., Tung, J. Y., & Sabater-Lleal, M. (2016). Genome-
526 wide association analysis of self-reported events in 6135 individuals and 252 827 controls
527 identifies 8 loci associated with thrombosis. *Human Molecular Genetics*, 25(9), 1867–1874.
528 <https://doi.org/10.1093/hmg/ddw037>
- 529 16. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and
530 accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature*
531 *Genetics*, 44(8), 955–959. <https://doi.org/10.1038/ng.2354>
- 532 17. Johnson, R., Ding, Y., Venkateswaran, V., Bhattacharya, A., Chiu, A., Schwarz, T., Freund, M.,
533 Zhan, L., Burch, K. S., Caggiano, C., Hill, B., Rakocz, N., Balliu, B., Sul, J. H., Zaitlen, N., Arboleda,
534 V. A., Halperin, E., Sankararaman, S., Butte, M. J., ... Pasaniuc, B. (2021). *Leveraging genomic*
535 *diversity for discovery in an EHR-linked biobank: The UCLA ATLAS Community Health Initiative* (p.
536 2021.09.22.21263987). <https://doi.org/10.1101/2021.09.22.21263987>
- 537 18. Klarin, D., Busenkell, E., Judy, R., Lynch, J., Levin, M., Haessler, J., Aragam, K., Chaffin, M., Haas,
538 M., Lindström, S., Assimes, T. L., Huang, J., Lee, K. M., Shao, Q., Huffman, J. E., Kabrhel, C.,
539 Huang, Y., Sun, Y. V., Vujkovic, M., ... Natarajan, P. (2019). Genome-wide association analysis of

- 540 venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular
541 disease. *Nature Genetics*, 51(11), 1574–1579. <https://doi.org/10.1038/s41588-019-0519-3>
- 542 19. Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack Of Diversity In
543 Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health*
544 *Affairs (Project Hope)*, 37(5), 780–785. <https://doi.org/10.1377/hlthaff.2017.1595>
- 545 20. Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M.,
546 Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide Human
547 Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 319(5866), 1100–
548 1104. <https://doi.org/10.1126/science.1153717>
- 549 21. Li, Y., & Lee, S. (2021). Novel score test to increase power in association test by integrating
550 external controls. *Genetic Epidemiology*, 45(3), 293–304. <https://doi.org/10.1002/gepi.22370>
- 551 22. Loh, P.-R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK
552 Biobank cohort. *Nature Genetics*, 48(7), 811–816. <https://doi.org/10.1038/ng.3571>
- 553 23. Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust
554 relationship inference in genome-wide association studies. *Bioinformatics*, 26(22), 2867–2873.
555 <https://doi.org/10.1093/bioinformatics/btq559>
- 556 24. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C.,
557 O’Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis,
558 G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for
559 quantitative and binary traits. *Nature Genetics*, 53(7), 1097–1103.
560 <https://doi.org/10.1038/s41588-021-00870-7>
- 561 25. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M.,
562 Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S.,
563 Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., ... the Haplotype Reference Consortium. (2016). A
564 reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10), 1279–
565 1283. <https://doi.org/10.1038/ng.3643>
- 566 26. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi,
567 A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T.,
568 Ohnishi, Y., Nakamura, Y., & Kubo, M. (2017). Overview of the BioBank Japan Project: Study
569 design and profile. *Journal of Epidemiology*, 27(3 Suppl), S2–S8.
570 <https://doi.org/10.1016/j.je.2016.12.005>
- 571 27. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar,
572 P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome
573 association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3),
574 559–575. <https://doi.org/10.1086/519795>
- 575 28. Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., & Masys,
576 D. R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized
577 medicine. *Clinical Pharmacology and Therapeutics*, 84(3), 362–369.
578 <https://doi.org/10.1038/clpt.2008.89>

- 579 29. Shakeel, F., Fang, F., Kwon, J. W., Koo, K., Pasternak, A. L., Henry, N. L., Sahai, V., Kidwell, K. M.,
580 & Hertz, D. L. (2021). Patients carrying DPYD variant alleles have increased risk of severe toxicity
581 and related treatment modifications during fluoropyrimidine chemotherapy.
582 *Pharmacogenomics*, 22(3), 145–155. <https://doi.org/10.2217/pgs-2020-0154>
- 583 30. Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic
584 Studies. *Cell*, 177(1), 26–31. <https://doi.org/10.1016/j.cell.2019.02.048>
- 585 31. Spector-Bagdady, K., Tang, S., Jabbour, S., Price, W. N., Bracic, A., Creary, M. S., Kheterpal, S.,
586 Brummett, C. M., & Wiens, J. (2021). Respecting Autonomy And Enabling Diversity: The Effect Of
587 Eligibility And Enrollment On Research Data Demographics. *Health Affairs (Project Hope)*, 40(12),
588 1892–1899. <https://doi.org/10.1377/hlthaff.2021.01197>
- 589 32. Surakka, I., Fritsche, L. G., Zhou, W., Backman, J., Kosmicki, J. A., Lu, H., Brumpton, B., Nielsen, J.
590 B., Gabrielsen, M. E., Skogholt, A. H., Wolford, B., Graham, S. E., Chen, Y. E., Lee, S., Kang, H. M.,
591 Langhammer, A., Forsmo, S., Åsvold, B. O., Styrkarsdottir, U., ... Willer, C. J. (2020). MEPE loss-of-
592 function variant associates with decreased bone mineral density and increased fracture risk.
593 *Nature Communications*, 11(1), 4093. <https://doi.org/10.1038/s41467-020-17315-0>
- 594 33. Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G.,
595 Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning,
596 B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., ... Abecasis, G. R. (2021). Sequencing of
597 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845), 290–299.
598 <https://doi.org/10.1038/s41586-021-03205-y>
- 599 34. *VCV000007105.43—ClinVar—NCBI*. (2021).
600 <https://www.ncbi.nlm.nih.gov/clinvar/variation/7105/>
- 601 35. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., Branham, K. E.,
602 Heckenlively, J., FUSION Study, Fulton, R., Wilson, R. K., Mardis, E. R., Lin, X., Swaroop, A.,
603 Zöllner, S., & Abecasis, G. R. (2014). Ancestry estimation and control of population stratification
604 for sequence-based association studies. *Nature Genetics*, 46(4), 409–415.
605 <https://doi.org/10.1038/ng.2924>
- 606 36. World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical
607 Principles for Medical Research Involving Human Subjects. *JAMA*, 310(20), 2191–2194.
608 <https://doi.org/10.1001/jama.2013.281053>
- 609 37. Wu, Y., Byrne, E. M., Zheng, Z., Kemper, K. E., Yengo, L., Mallett, A. J., Yang, J., Visscher, P. M., &
610 Wray, N. R. (2019). Genome-wide association study of medication-use and associated disease in
611 the UK Biobank. *Nature Communications*, 10(1), 1891. <https://doi.org/10.1038/s41467-019-09572-5>
- 613 38. Zajac, G. J. M., Fritsche, L. G., Weinstock, J. S., Dagenais, S. L., Lyons, R. H., Brummett, C. M., &
614 Abecasis, G. R. (2019). Estimation of DNA contamination and its sources in genotyped samples.
615 *Genetic Epidemiology*, 43(8), 980–995. <https://doi.org/10.1002/gepi.22257>
- 616 39. Zhou, W., Kanai, M., Wu, K.-H. H., Humaira, R., Tsuo, K., Hirbo, J. B., Wang, Y., Bhattacharya, A.,
617 Zhao, H., Namba, S., Surakka, I., Wolford, B. N., Faro, V. L., Lopera-Maya, E. A., Läll, K., Favé, M.-

618 J., Chapman, S. B., Karjalainen, J., Kurki, M., ... Neale, B. M. (2021). *Global Biobank Meta-analysis*
619 *Initiative: Powering genetic discovery across human diseases* (p. 2021.11.19.21266436).
620 <https://doi.org/10.1101/2021.11.19.21266436>

621 40. Zhou, W., Zhao, Z., Nielsen, J. B., Fritsche, L. G., LeFaive, J., Gagliano Taliun, S. A., Bi, W.,
622 Gabrielsen, M. E., Daly, M. J., Neale, B. M., Hveem, K., Abecasis, G. R., Willer, C. J., & Lee, S.
623 (2020). Scalable generalized linear mixed model for region-based association tests in large
624 biobanks and cohorts. *Nature Genetics*, 52(6), 634–639. [https://doi.org/10.1038/s41588-020-](https://doi.org/10.1038/s41588-020-0621-6)
625 0621-6

626

627

Figure 1: Overview of the Michigan Genomics Initiative (MGI) cohort. MGI currently consists of >85K participants recruited while seeking care at the Michigan Medicine Health Center. Recruitment is predominantly through the Department of Anesthesiology during surgical encounters. Participants consent to link a blood sample with their electronic health records for broad research purposes. Genotypes for ~570K genetic variants are obtained from DNA extracted from the blood sample using a customized Illumina Infinium CoreExome-24 array. In this paper, we describe the Freeze 3 MGI cohort consisting of ~57K samples having passed sample-level quality control filtering and imputed for >50 million variants using the TOPMed reference panel. We extracted all available International Classification of Disease (ICD) diagnosis codes from patient electronic health records and mapped to broader dichotomous phecode traits using the *PheWAS* software. We performed GWAS within a subset of ~50K European-inferred samples from the Freeze 3 cohort using a linear mixed effect regression model implemented in the *SAIGE* software. We report results and share GWAS summary statistics for 1,547 traits with ≥ 60 cases.

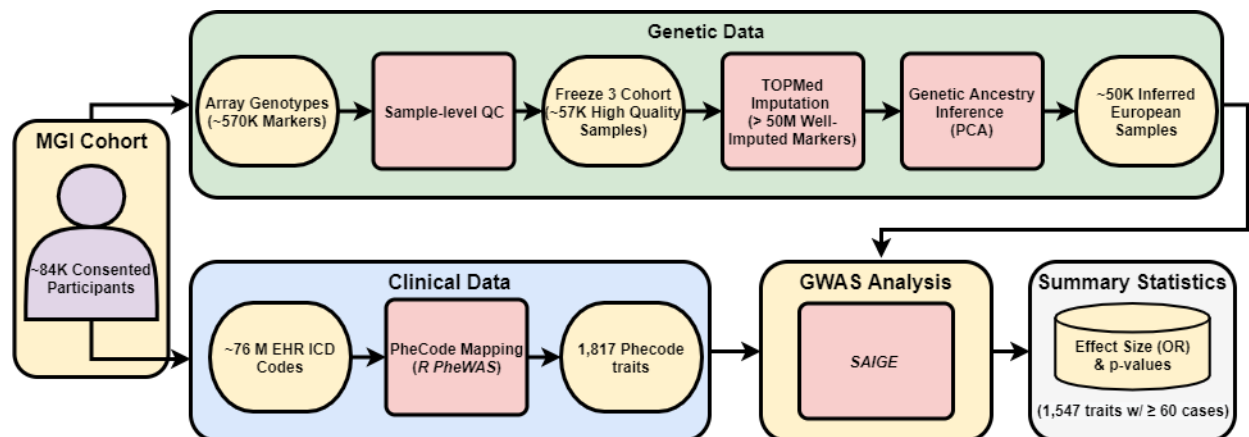


Figure 2: MGI Demographics and Clinical Data.

A) MGI recruitment over time. The solid line gives overall participant recruitment, and the dashed line is participants with self-reported race other than Caucasian.

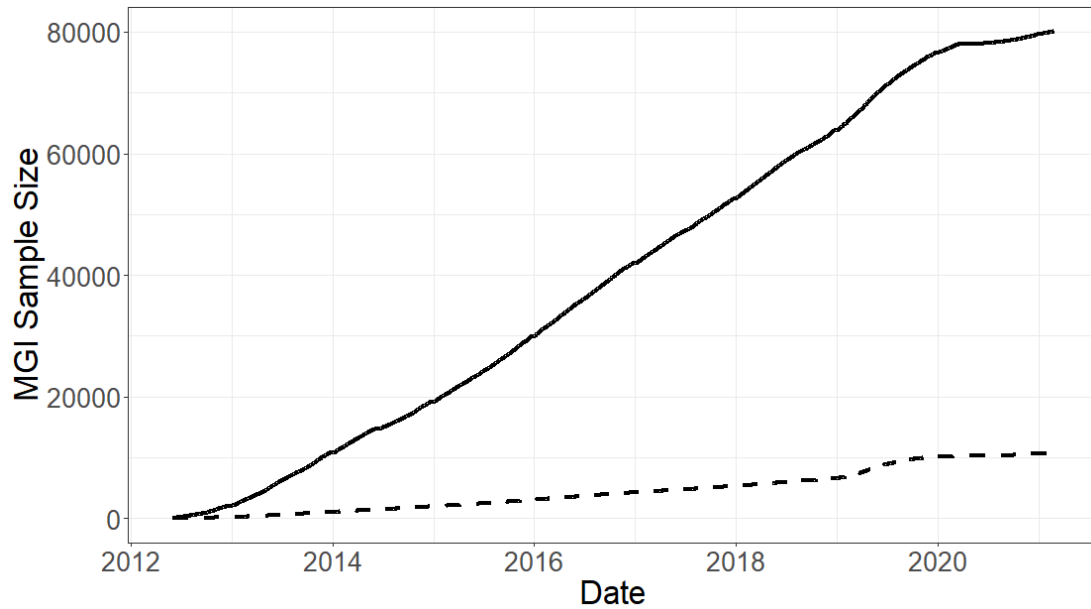


Figure 2

B) Age and sex distribution of MGI Participants.

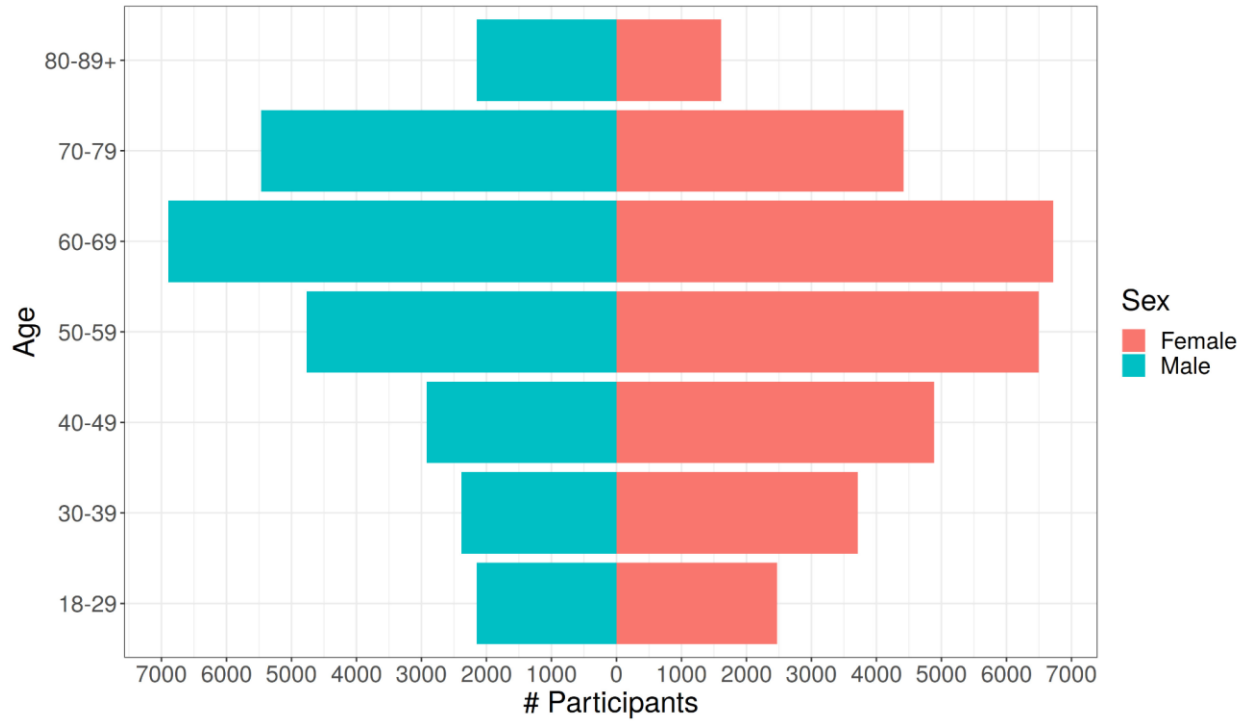
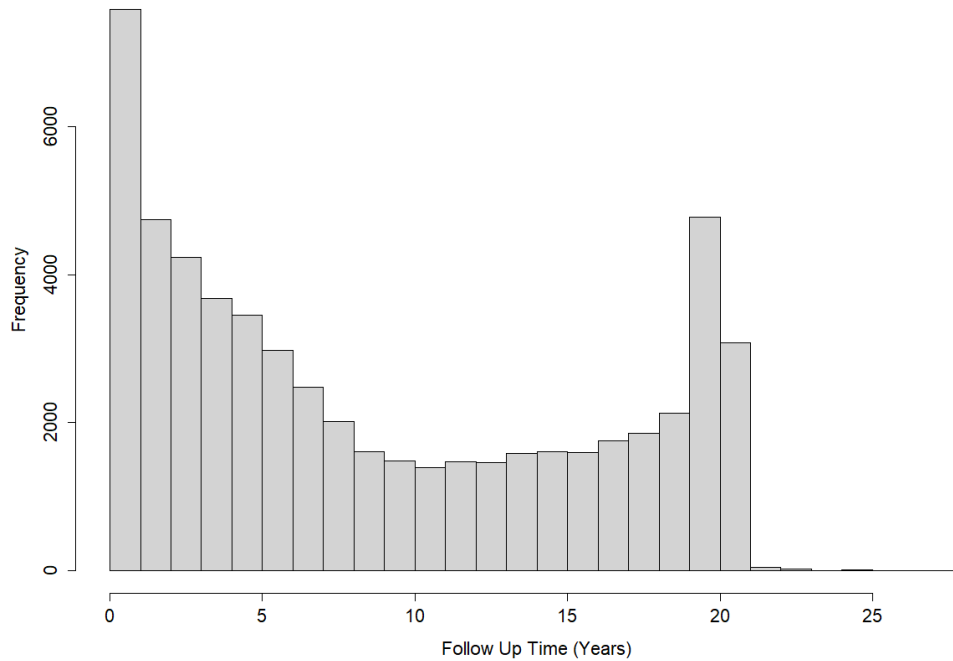


Figure 2

C) Clinical follow-up time for MGI participants. Follow-up is the amount of time between a participant's first and most recent diagnosis codes in the Michigan Medicine EHR. Insert: Distribution of ages for MGI participants is nearly identical across follow-up times.



Insert:

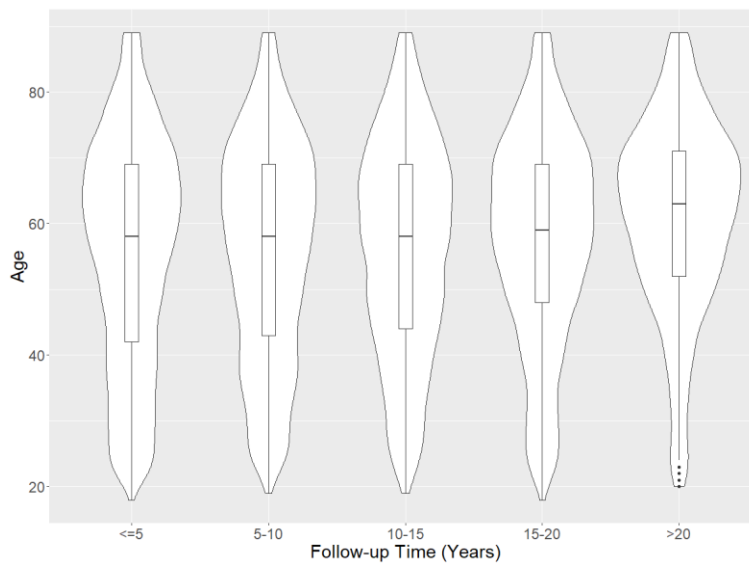


Figure 2

D) Most common phecodes traits among MGI participants.

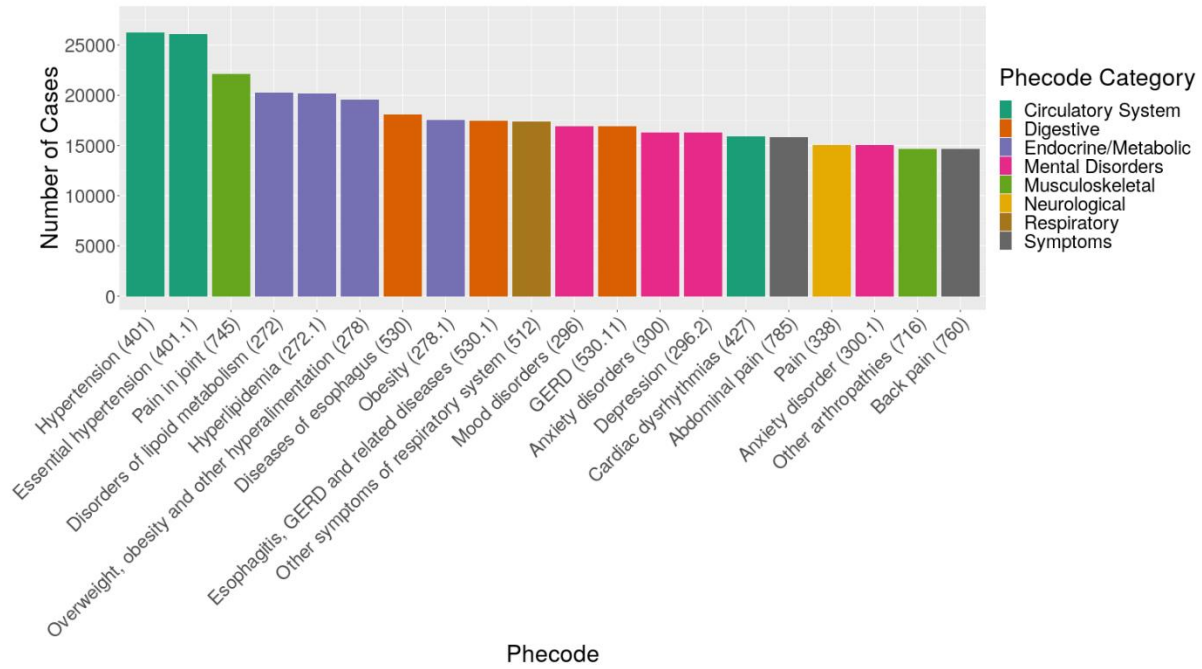


Figure 2

E) Number of phecode case assignments per sample increases with participant age (boxplot outliers excluded for readability).

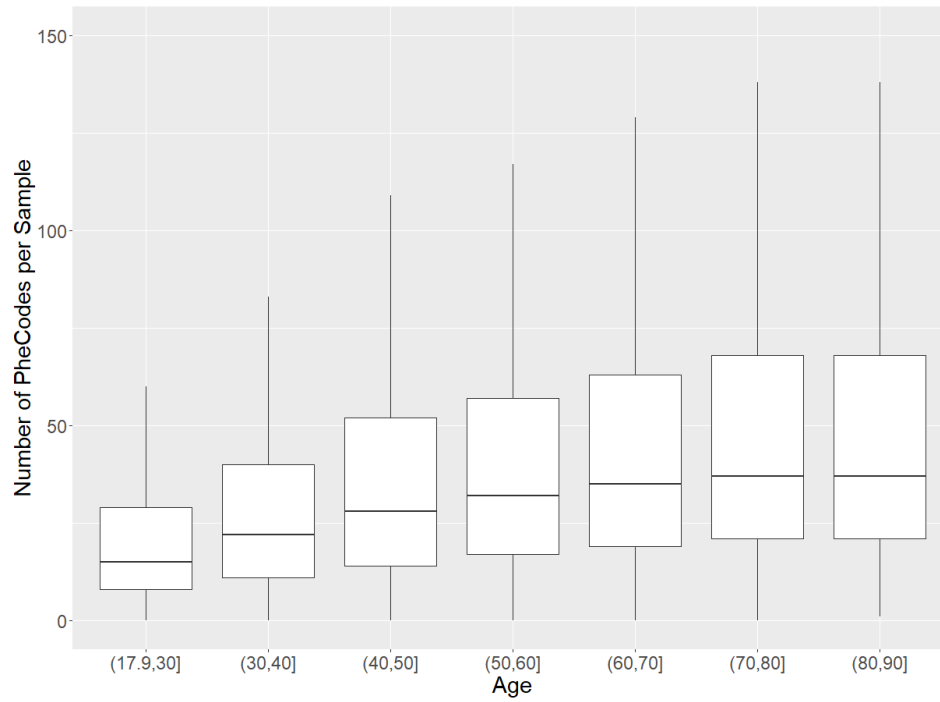


Figure 2

F) Number of phecode case assignments per sample increases with participant follow-up time (boxplot outliers excluded for readability).

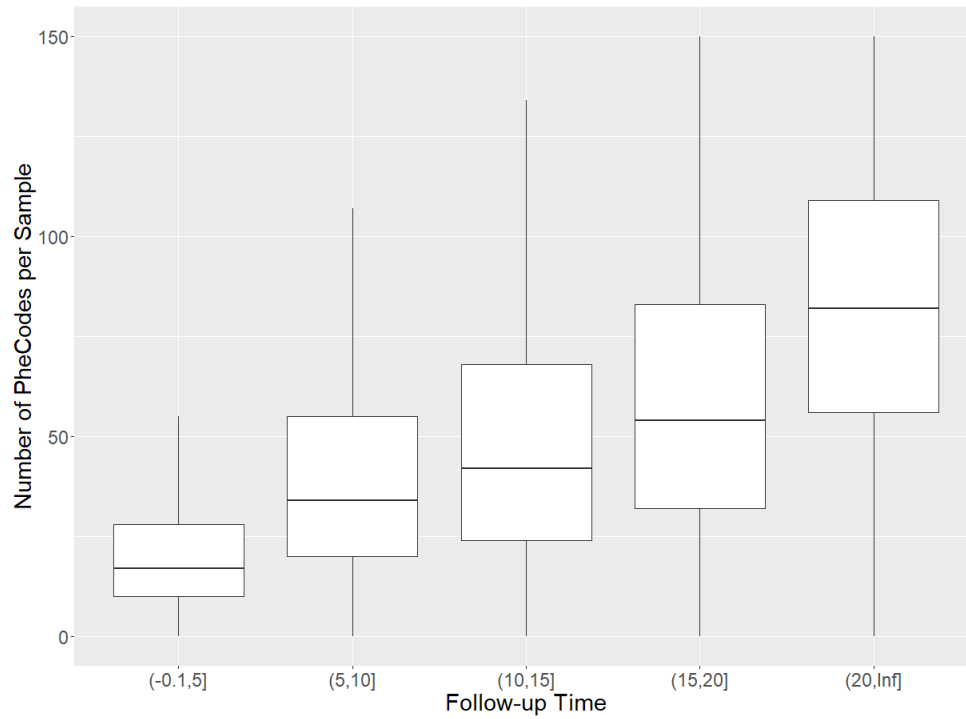


Figure 3: Comparison of case counts for phecode traits between MGI and UKB European GWAS cohorts by phecode categories.

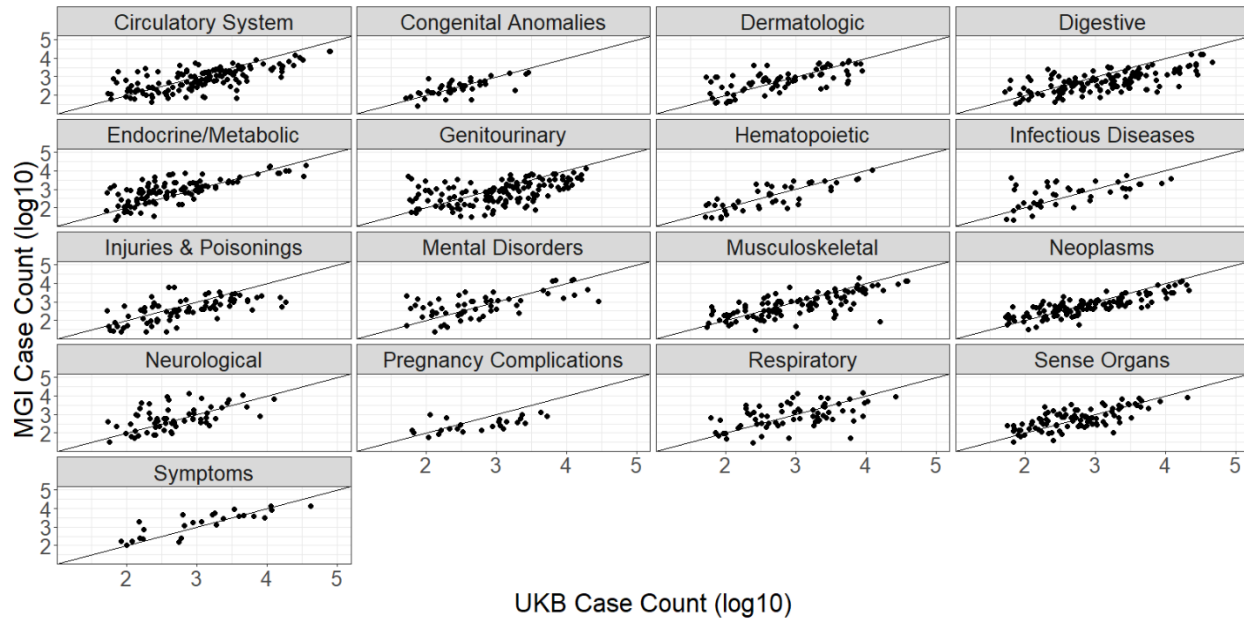
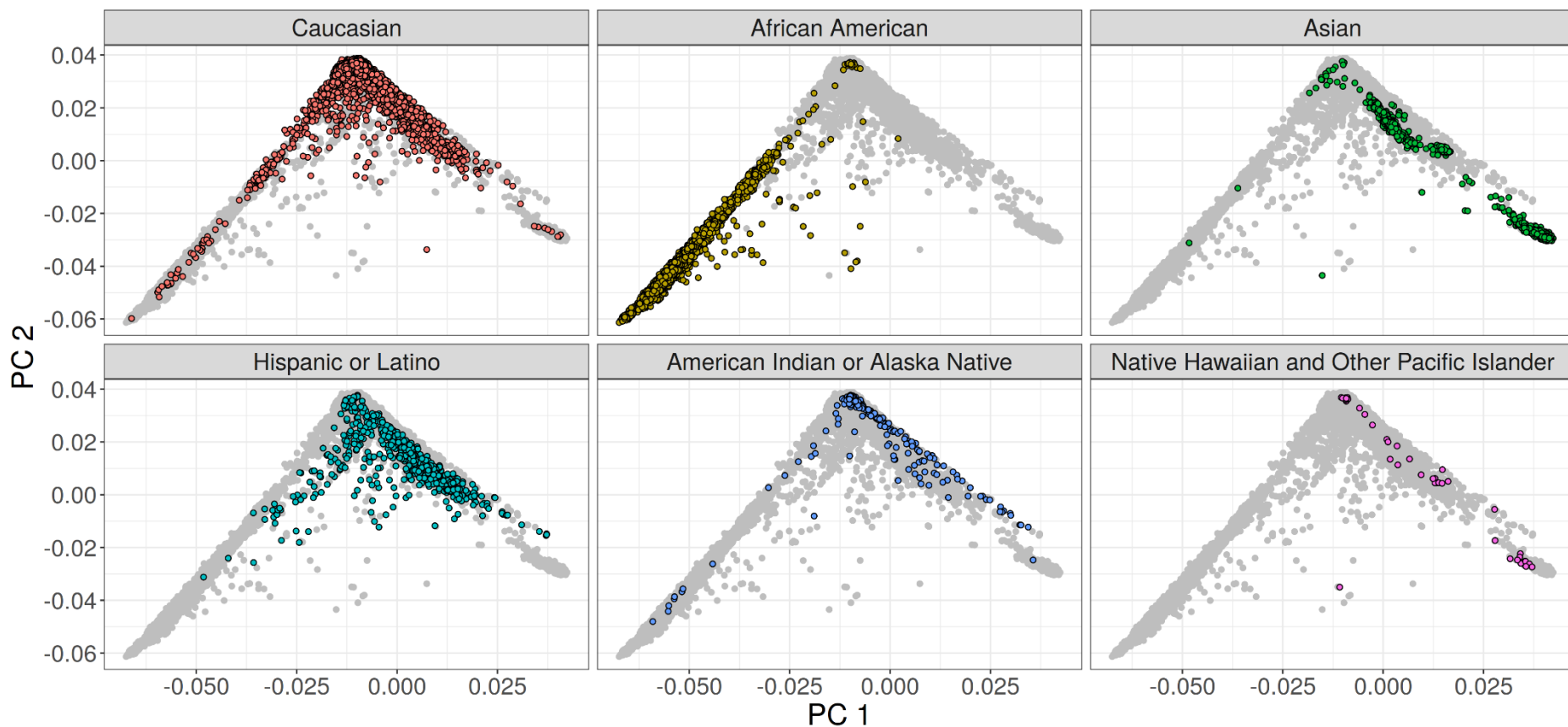
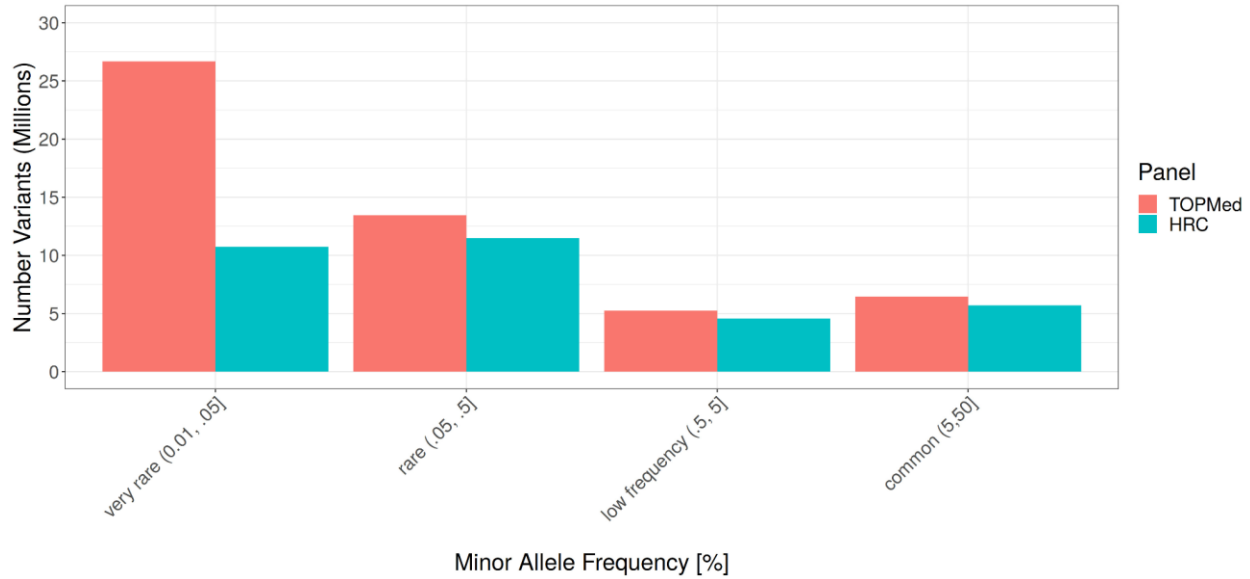


Figure 4: Summary of Genetic Data

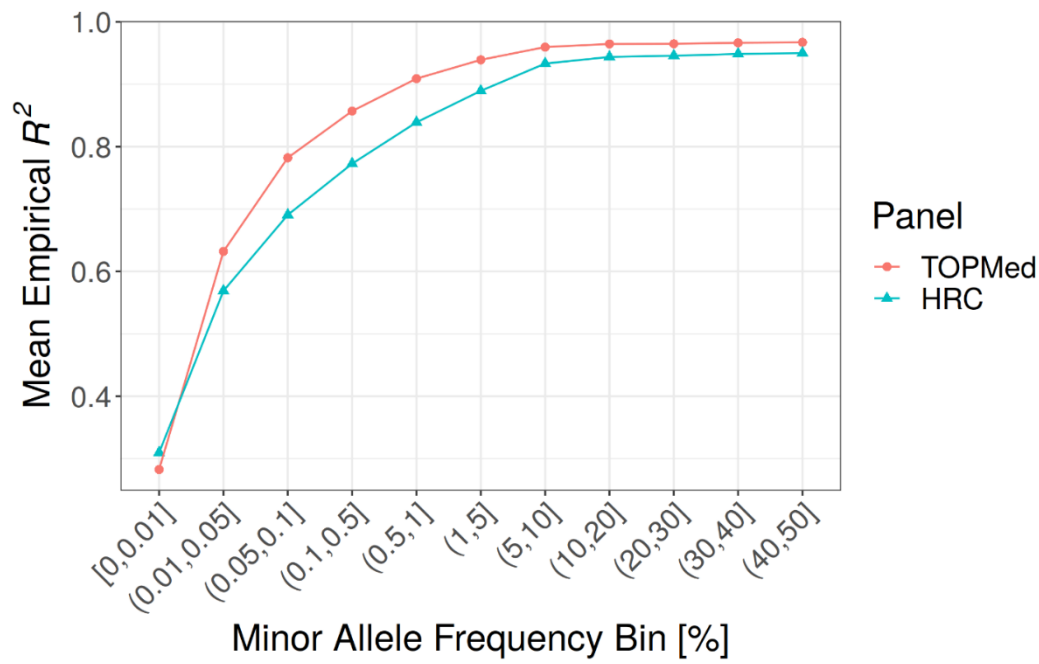
A) Comparison of self-reported race/ethnicity and genetically inferred ancestry. MGI samples are projected in the Principal Component (PC) reference space created by worldwide samples from the Human Genome Diversity Project (HGDP). Each panel shows all MGI participants (gray dots), with colored dots denoting participants of the indicated self-reported race or ethnicity (Hispanic or Latino).



B) Comparison of frequency spectrum for TopMed and HRC imputation. The largest gain in coverage for TopMed is at the lower end of the frequency spectrum.



C) Comparison of TopMed and HRC imputation accuracy by MAF. TopMed provides increased accuracy for all MAF>0.01 bins, with greatest improvement for SNPs < 5%.



D) Comparison of TopMed and HRC imputation accuracy by ADMIXTURE inferred ancestry groups. TopMed provides more accurate imputation in all populations with notable gains among MGI participants whose majority ancestry is non-European.

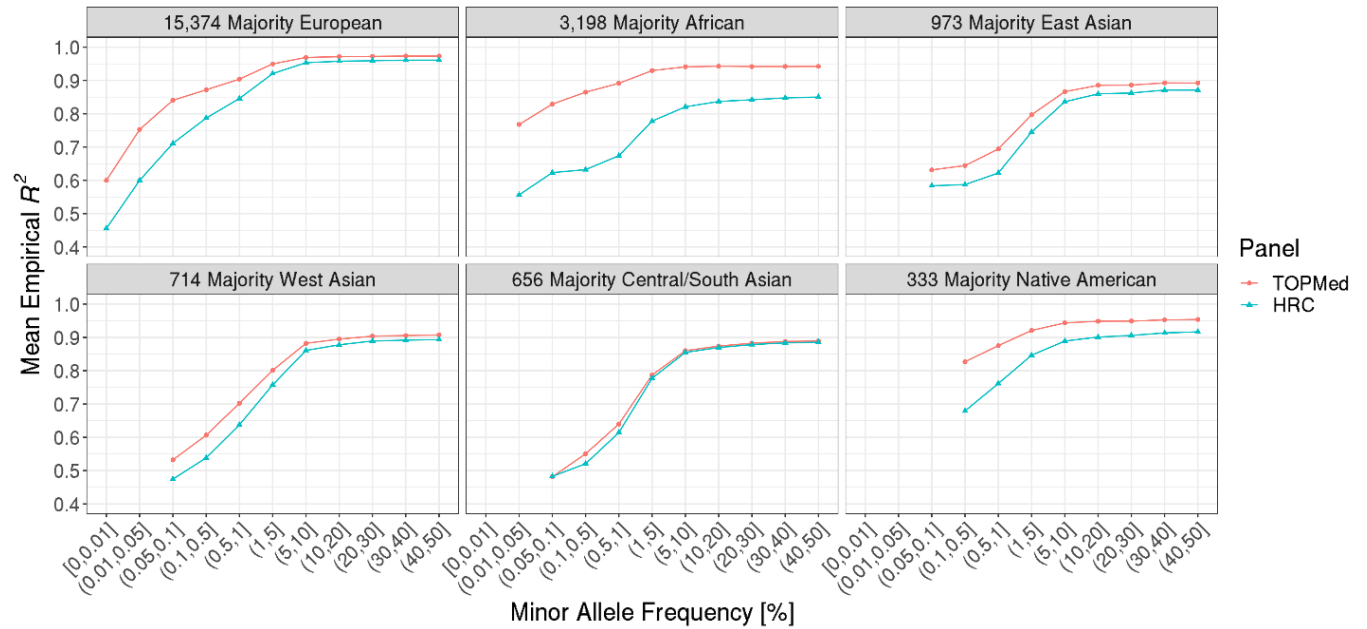


Table 1: Demographics of MGI Participants, Freeze 3.

	All	Male	Female
Number of Individuals	57,055	26,732 (47%)	30,323 (53%)
Age, yr (range 18-90+; mean \pm SD)	56.44 \pm 17	58.38 \pm 17	54.74 \pm 16
BMI, kg/m²	29.87 \pm 7.0	29.69 \pm 6.1	30.03 \pm 7.7
Self-Reported Race, N			
African American	3,223	1,264	1,959
American Indian or Alaska Native	237	112	125
Asian	1,324	601	723
Caucasian	49,605	23,534	26,071
Native Hawaiian or Pacific Islander	43	14	29
Other	1,023	463	560
Patient Refused	200	102	98
Unknown/Missing	1,400	643	757
Self-Reported Ethnicity, N			
Hispanic or Latino	805	337	468
Non-Hispanic or Latino	34,982	16,809	18,173
Patient Refused	155	70	85
Unknown/Missing	21,113	9,517	11,596

Table 2. Summary of GWAS results by phecode categories in European MGI participants. The table contains counts of associations across phecode traits with at least sixty cases and for all markers tested as well as minor allele frequency >1%. (GWS=genome-wide significant, see Methods for definition)

Phecode Category	Total Phecode Traits	Analyzed Traits (≥ 60 cases)	Traits with ≥1 GWS loci (MAF>1%)	Number of GWS Loci (MAF>1%)	Strongest Association (MAF >1%)
<i>circulatory system</i>	171	160	108 (43)	200 (72)	Atrial fibrillation (427.21), p=1.2e-37, chr4:110762205
<i>congenital anomalies</i>	56	44	18 (3)	36 (3)	Genitourinary congenital anomalies (751), p= 4.0e-09, chr2:161318326
<i>dermatologic</i>	95	77	53 (17)	93 (22)	Psoriasis vulgaris (696.41), p=4.7e-28, chr6:31274954
<i>digestive</i>	162	149	95 (39)	198 (59)	Other chronic nonalcoholic liver disease (571.5), p=3.0e-54, chr22:43928975
<i>endocrine/metabolic</i>	169	129	92 (65)	277 (180)	Type 1 diabetes (250.1), p=4.2e-106, chr6:32658525
<i>genitourinary</i>	173	157	101 (25)	191 (39)	Nephritis and nephropathy in diseases classified elsewhere (580.31), p=1.4e-19, chr6:32706117
<i>hematopoietic</i>	62	45	32 (16)	65 (26)	Primary hypercoagulable state (286.81), p=2.8e-157, chr1:169549811
<i>infectious diseases</i>	69	54	28 (8)	37 (8)	Aspergillosis (117.4), p=4.3e-17, chr7:117559590
<i>injuries & poisonings</i>	122	93	49 (5)	79 (6)	Salicylates causing adverse effects in therapeutic use (965.3), p=2.4e-10, chr6:33091097
<i>mental disorders</i>	76	63	39 (11)	64 (12)	Dementias (290.1), p=2.1e-18, chr19:44908684
<i>musculoskeletal</i>	132	114	71 (19)	121 (20)	Ankylosing spondylitis (715.2), p=2.9e-35, 6:31357491
<i>neoplasms</i>	141	129	76 (29)	194 (85)	Other non-epithelial cancer of skin (172.2), p=1.8e-38, chr6:396321
<i>neurological</i>	85	74	50 (11)	79 (14)	Restless legs syndrome (327.71), p=6.8e-29, chr2:66523432
<i>pregnancy complications</i>	46	28	18 (7)	23 (7)	Rhesus isoimmunization in pregnancy (654.2),

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .

					p=1.4e-54, chr1:25257119
<i>respiratory</i>	85	78	57 (22)	96 (26)	Cystic fibrosis (499), p=9.8e-49, chr7:117559590
<i>sense organs</i>	127	112	65 (18)	105 (25)	Fuchs' dystrophy (364.51), p=2.0e-31, chr18:55543071
<i>symptoms</i>	46	41	25 (2)	43 (2)	Fever of unknown origin (783), p= 2.9e-08, chr7:37808912
<i>Total</i>	1817	1547	977 (340)	1901 (606)	

Table 3: Top 30 GWAS associations among the 1,547 phecode traits with at least 60 cases in MGI, Freeze 3.

rs id (Position)	Alleles	Allele2 Frequency	Log(OR)	p-value	Trait Description (Phecode)	Cases	Controls	SNP Regional Match	GWAS Catalog (Pubmed ID)
rs6025 (chr1:169549811)	C/T	0.0282	2.44	2.81E-157	Primary hypercoagulable state (286.81)	727	43826	Exact	Venous thromboembolism (31676865)
			2.40	1.19E-153	Hypercoagulable state (286.8)	755	43826		
			1.33	1.80E-83	Coagulation defects (286)	2693	43826		
			1.17	6.73E-50	Other and unspecified coagulation defects (286.7)	1942	43826		
			2.67	5.24E-39	Congenital deficiency of other clotting factors (including factor VII) (286.12)	94	43826		
			0.77	1.82E-36	Other venous embolism and thrombosis (452)	4201	36930		
			0.86	3.01E-34	Deep vein thrombosis (452.2)	3162	36930		Thrombosis (26908601)
rs72660908 (chr1:25257119)	C/G	0.3856	2.48	1.40E-54	Rhesus isoimmunization in pregnancy (654.2)	145	26348	Exact	Blood protein levels (29875488)
rs4148325 (chr2:233764663)	C/T	0.3272	1.64	6.00E-82	Disorders of bilirubin excretion (277.4)	321	48830	Exact	Bilirubin levels (21646302)
rs1800562 (chr6:26092913)	G/A	0.0602	1.73	1.07E-51	Disorders of iron metabolism (275.1)	201	47321	Exact	Hemoglobin (32888494)
rs185937162 (chr6:31357491)	T/G	0.0428	1.79	2.92E-35	Ankylosing spondylitis (715.2)	190	35793	50kb window	Ankylosing spondylitis (20062062)
rs2040410 (chr6:32634921)	C/T	0.1260	1.04	5.91E-39	Celiac disease (557.1)	407	37236	50kb window	Celiac disease (20190752)
rs9273364 (chr6:32658525)	T/G	0.2769	0.87	4.23E-106	Type 1 diabetes (250.1)	2266	36631	Exact	Medication use - drugs used in diabetes (31015401)
			0.55	1.32E-34	Type 2 diabetes with ophthalmic manifestations (250.23)	1522	36631		

Table 3 (Con't): Top 30 GWAS associations among the 1,547 phecode traits with at least 60 cases in MGI, Freeze 3.

rs id (Position)	Alleles	Allele2 Frequency	Log(OR)	p-value	Trait Description (Phecode)	Cases	Controls	SNP Regional Match	GWAS Catalog (Pubmed ID)
rs9273368 (chr6:32658698)	G/A	0.2713	1.21	2.91E-101	Type 1 diabetes with ophthalmic manifestations (250.13)	760	36631	Exact	Latent autoimmune diabetes vs. type 1 diabetes (30254083)
			1.32	4.02E-80	Type 1 diabetes with renal manifestations (250.12)	509	36631		
			1.22	6.99E-76	Type 1 diabetes with neurological manifestations (250.14)	559	36631		
			1.42	1.23E-40	Type 1 diabetes with ketoacidosis (250.11)	205	36631		
rs1794269 (chr6:32706117)	C/T	0.3760	0.64	4.53E-52	Diabetic retinopathy (250.7)	1544	43849	50kb window	Type 2 Diabetes (32541925)
			0.41	1.04E-39	Insulin pump user (250.3)	3155	36631		
rs12203592 (chr6:396321)	C/T	0.1616	0.35	1.83E-38	Other non-epithelial cancer of skin (172.2)	6627	41896	Exact	Basal cell carcinoma (31174203)
			0.32	1.65E-36	Skin cancer (172)	8228	41896		
			0.44	2.36E-36	Basal cell carcinoma (172.21)	3509	41896		
rs113993960 (chr7:117559590)	ATCT/A	0.0146	2.58	9.80E-49	Cystic fibrosis (499)	97	51358	1MB window	Lung function - FEV1/FVC (30595370)
rs28929474 (chr14:94378610)	C/T	0.0179	3.36	1.71E-52	Alpha-1-antitrypsin deficiency (270.34)	60	48887	Exact	Serum albumin level (33462484)
rs1421085 (chr16:53767042)	T/C	0.4156	0.24	1.65E-36	Morbid obesity (278.11)	7255	32074	Exact	Body mass index (30595370)
rs3747207 (chr22:43928975)	G/A	0.2296	0.48	2.95E-54	Other chronic nonalcoholic liver disease (571.5)	2973	41006	Exact	Alanine transaminase levels in high alcohol intake (32561361)
			0.45	7.98E-53	Chronic liver disease and cirrhosis (571)	3150	41006		

Supplementary Table 1. Expanded Table 1 including descriptive statistics for quantitative laboratory measures and select phecode traits.

	All	Male	Female
Number of Individuals	57,055	26,732 (47%)	30,323 (53%)
Age, yr (range 18-90+; mean \pm SD)	56.44 \pm 17	58.38 \pm 17	54.74 \pm 16
BMI, kg/m²	29.87 \pm 7.0	29.69 \pm 6.1	30.03 \pm 7.7
Self-Reported Race, N			
African American	3,223	1,264	1,959
American Indian or Alaska Native	237	112	125
Asian	1,324	601	723
Caucasian	49,605	23,534	26,071
Native Hawaiian or Pacific Islander	43	14	29
Other	1,023	463	560
Patient Refused	200	102	98
Unknown/Missing	1,400	643	757
Self-Reported Ethnicity, N			
Hispanic or Latino	805	337	468
Non-Hispanic or Latino	34,982	16,809	18,173
Patient Refused	155	70	85
Unknown/Missing	21,113	9,517	11,596
Quantitative Measurements (mean \pm SD)			
SBP, mm Hg (N=56,293)	70.52 \pm 7.4	72.45 \pm 7.4	68.83 \pm 7.0
LDL-C, mg/dL (N=24,688)	100.38 \pm 35.7	95.92 \pm 35.2	104.65 \pm 35.7
Albumin (N=43,558)	3.96 \pm 0.6	3.93 \pm 0.6	3.99 \pm 0.6
Creatinine (N=50,255)	1.21 \pm 1.2	1.37 \pm 1.3	1.04 \pm 1.0
Glucose (N=50,327)	123.37 \pm 56.2	127.06 \pm 56.6	119.40 \pm 55.4
Mean corpuscular hemoglobin (N=50,246)	29.8 \pm 2.6	29.98 \pm 2.6	29.60 \pm 2.6
Thyroid stimulating hormone (N=27,753)	3.19 \pm 10.6	3.21 \pm 9.1	3.17 \pm 11.4
Cardiometabolic Phenotypes, Phecode (N cases, % of sample)			
Hypertension (401)	26,223 (46%)	13,858 (52%)	12,365 (41%)
Obesity (278.1)	17,547 (31%)	7,455 (28%)	10,092 (33%)
Cardiac Dysrhythmias (427)	15,896 (28%)	7,853 (29%)	8,043 (27%)
Type 2 Diabetes (250.2)	11,377 (20%)	6,036 (23%)	5,341 (18%)
Neurodegenerative and Neurological Phenotypes (N cases, % of sample)			
Sleep Apnea (327.3)	12,418 (22%)	6,884 (26%)	5,534 (18%)
Epilepsy (345.1)	606 (1%)	275 (1%)	331 (1%)
Parkinson's Disease (332)	399 (0.7%)	272 (1%)	127 (0.4%)
Multiple Sclerosis (335)	415 (0.7%)	109 (0.4%)	306 (1%)
Respiratory and Immunological Phenotypes (N cases, % of sample)			
Asthma (495)	10,135 (18%)	3,446 (13%)	6,689 (22%)
Pneumonia (480)	5,327 (9%)	2,467 (9%)	2,860 (9%)
Rheumatoid Arthritis (714.1)	1,750 (3%)	478 (2%)	1,272 (4%)

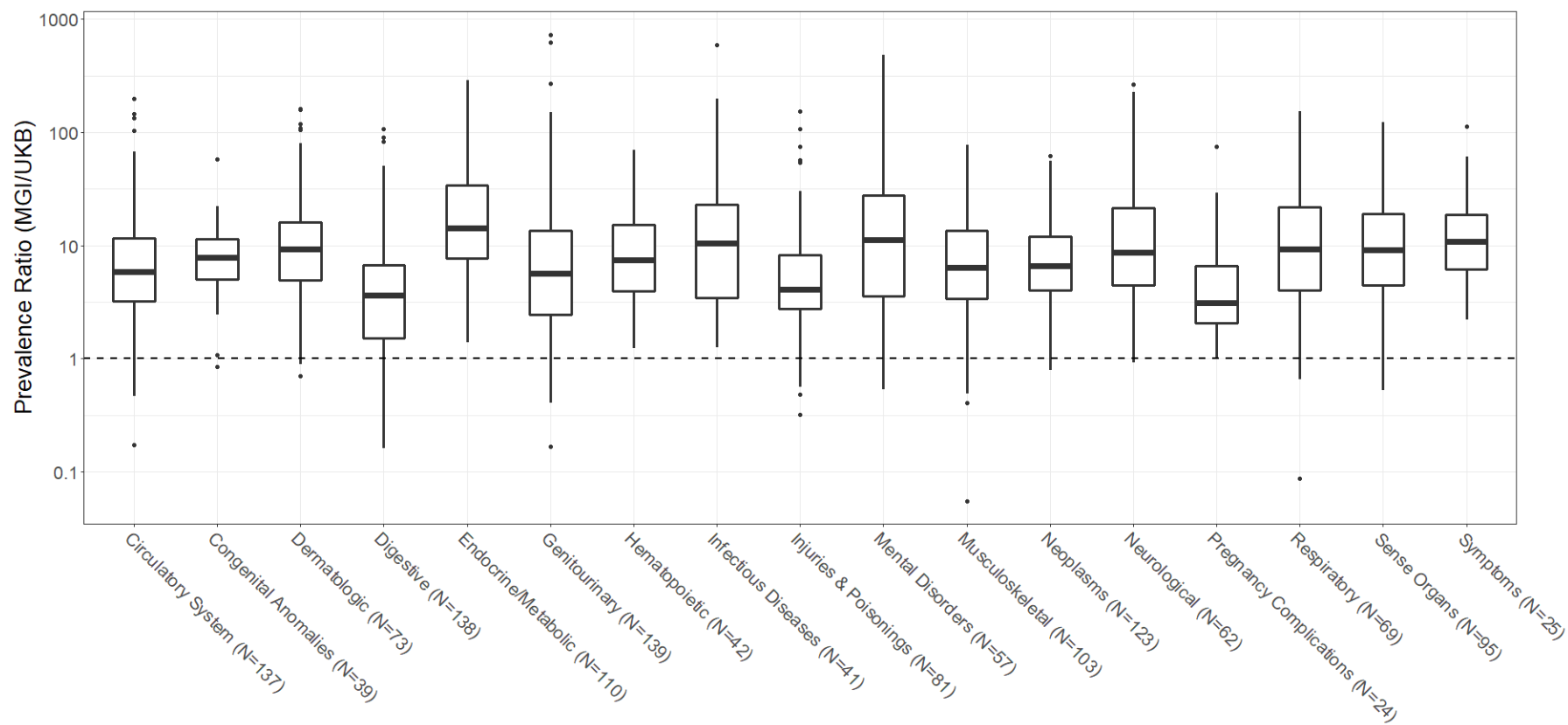
Ulcerative Colitis (555.2)	925 (2%)	428 (2%)	497 (2%)
Oncology Phenotypes (N cases, % of sample)			
Skin Cancer (172)	8,307 (15%)	4,580 (17%)	3,727 (12%)
Melanoma (172.11)	3,081 (5%)	1,772 (7%)	1,309 (4%)
Basal Cell Carcinoma (172.21)	3,522 (6%)	1,921 (7%)	1,601 (5%)
Breast Cancer (174)	3,418 (6%)	46 (0.2%)	3,372 (11%)
Prostate Cancer (185)	3,002 (5%)	3,002 (11%)	0 (0%)
Bladder Cancer (189.2)	1,746 (3%)	1,343 (5%)	403 (1%)
Mental Health Phenotypes (N cases, % of sample)			
Depression (296.2)	14,889 (26%)	4,992 (19%)	9,897 (33%)
Anxiety disorders (300)	14,922 (26%)	4,810 (18%)	10,112 (33%)
Bipolar (296.1)	1,134 (2%)	357 (1%)	777 (3%)
Schizophrenia (295.1)	115 (0.2%)	50 (0.1%)	65 (0.2%)

Supplementary Table 2. Phecode Traits with >10-fold enrichment of case samples in MGI compared to UKB.

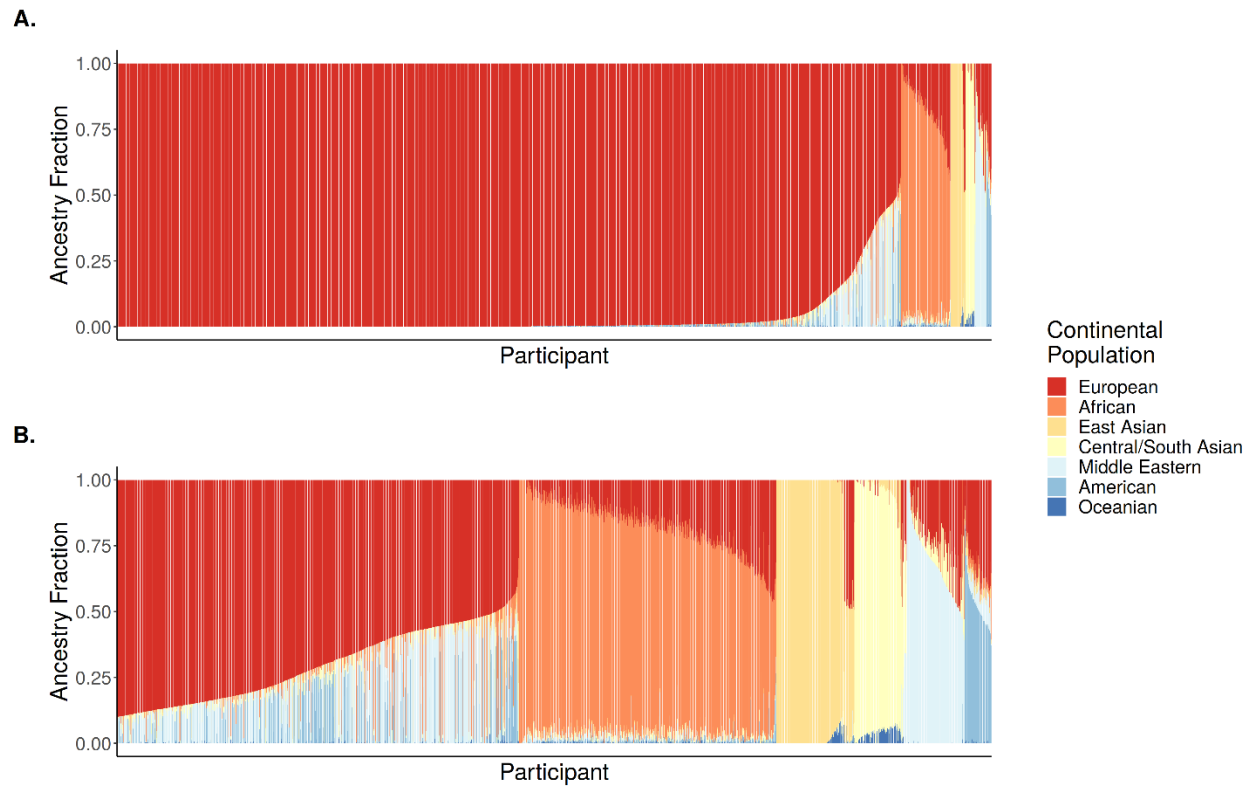
Phecode	Description	Category	MGI Cases	MGI Controls	UKB Cases	UKB Controls	MGI:UKB Case Ratio
611.1	Abnormal mammogram	genitourinary	4,851	42,782	56	400,113	86.6
079.1	Varicella infection	infectious diseases	4,143	39,472	64	401,601	64.7
599.8	Other symptoms involving urinary system	genitourinary	3,888	35,037	62	383,297	62.7
300.4	Dysthymic disorder	mental disorders	2,050	27,399	53	363,984	38.7
792	Abnormal Papanicolaou smear of cervix and cervical HPV	genitourinary	2,788	17,968	97	193,707	28.7
327.4	Insomnia	neurological	4,252	35,287	163	401,998	26.1
278.4	Abnormal weight gain	endocrine/metabolic	1,625	32,094	66	396,288	24.6
415.11	Pulmonary embolism and infarction, acute	circulatory system	1,825	47,402	75	400,268	24.3
338.2	Chronic pain	neurological	9,127	34,356	379	406,436	24.1
110.11	Dermatophytosis of nail	infectious diseases	1,664	44,101	74	404,551	22.5
272.12	Hyperglyceridemia	endocrine/metabolic	1,210	31,466	58	371,432	20.9
272.13	Mixed hyperlipidemia	endocrine/metabolic	3,221	31,466	164	371,432	19.6
313	Pervasive developmental disorders	mental disorders	1,449	49,104	76	406,624	19.1
279.7	Other immunological findings	endocrine/metabolic	4,513	44,339	248	406,707	18.2
613.5	Mastodynia	genitourinary	1,396	47,335	77	405,229	18.1
338	Pain	neurological	13,542	34,356	766	406,436	17.7
690.1	Seborrheic dermatitis	dermatologic	924	43,414	53	400,950	17.4
340.1	Migraine with aura	neurological	3,640	43,699	211	397,071	17.3
464	Acute sinusitis	respiratory	3,174	39,262	196	404,740	16.2
949	Allergies, other	injuries & poisonings	6,371	35,760	396	403,085	16.1
938.2	Chronic dermatitis due to solar radiation	dermatologic	3,992	35,760	251	403,085	15.9
338.1	Acute pain	neurological	6,081	34,356	393	406,436	15.5
246	Other disorders of thyroid	endocrine/metabolic	6,055	39,246	397	389,743	15.3
250.42	Other abnormal glucose	endocrine/metabolic	6,284	36,652	418	387,082	15.0
261.4	Vitamin D deficiency	endocrine/metabolic	6,612	40,977	446	404,730	14.8
456	Chronic venous insufficiency [CVI]	circulatory system	2,633	36,953	183	367,908	14.4

356	Hereditary and idiopathic peripheral neuropathy	neurological	2,114	47,411	152	405,102	13.9
367.2	Astigmatism	sense organs	2,622	43,024	189	404,787	13.9
627.2	Symptomatic menopause	genitourinary	2,534	17,353	183	189,130	13.8
090	Sexually transmitted infections (not HIV or hepatitis)	infectious diseases	953	50,467	71	407,129	13.4
690	Erythematous squamous dermatosis	dermatologic	929	43,414	71	400,950	13.1
790	Nonspecific findings on examination of blood	symptoms	1,930	44,032	149	400,617	13.0
401.3	Other hypertensive complications	circulatory system	1,827	26,620	144	328,788	12.7
840	Sprains and strains	injuries & poisonings	5,944	42,380	469	406,668	12.7
605	Erectile dysfunction [ED]	genitourinary	3,436	16,004	276	167,468	12.4
704.8	Other specified diseases of hair and hair follicles	dermatologic	943	46,497	76	400,627	12.4
476	Allergic rhinitis	respiratory	12,834	31,134	1,052	388,553	12.2
136	Other infectious and parasitic diseases	infectious diseases	2,459	47,630	205	406,672	12.0
277.51	Lipoprotein disorders	endocrine/metabolic	620	48,876	52	405,784	11.9
429.1	Heart transplant/surgery	circulatory system	762	45,087	64	400,869	11.9
573.9	Abnormal serum enzyme levels	digestive	2,452	41,044	209	398,276	11.7
279	Disorders involving the immune mechanism	endocrine/metabolic	2,845	44,339	249	406,707	11.4
483	Acute bronchitis and bronchiolitis	respiratory	2,122	42,165	193	396,438	11.0
505	Other pulmonary inflammation or edema	respiratory	659	43,653	61	395,582	10.8
938	Dermatitis due to solar radiation	dermatologic	4,174	35,760	387	403,085	10.8
527.7	Disturbance of salivary secretion	digestive	667	47,262	62	401,593	10.8
260	Protein-calorie malnutrition	endocrine/metabolic	2,103	40,977	202	404,730	10.4
250.4	Abnormal glucose	endocrine/metabolic	7,061	36,652	681	387,082	10.4

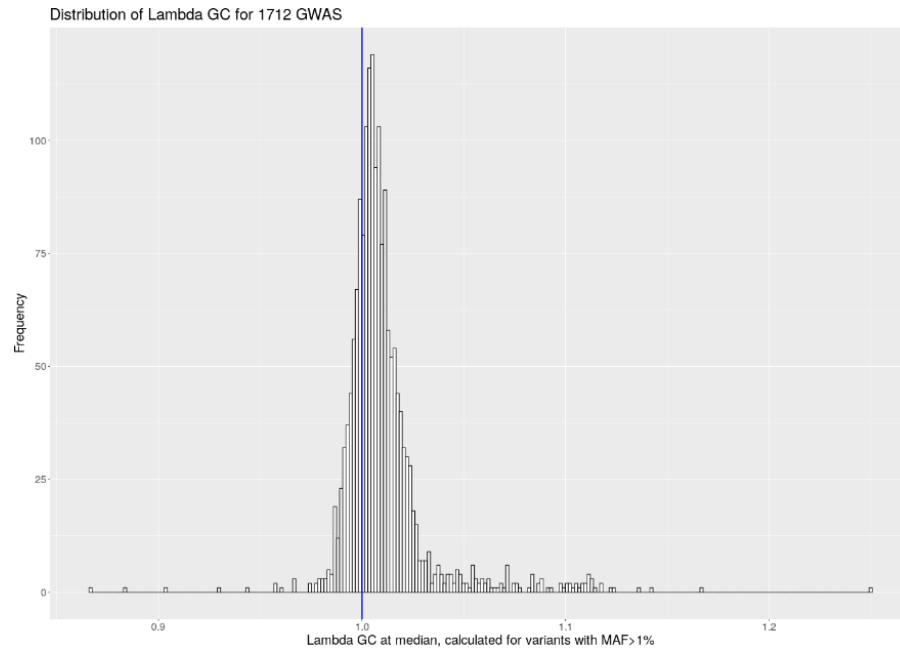
Supplementary Figure 1: Ratio of disease prevalence in phecode traits in MGI and UKB European GWAS cohorts by phecode category.



Supplementary Figure 2: Ancestry proportions based on ADMIXTURE for (A) all MGI participants and (B) only participants with less than 90% inferred EUR ancestry.



Supplemental Figure 3: Distribution of genomic control values for 1712 phecode traits with at least 20 cases.



Supplementary Figure 4: Genomic control and case count (log10 scale) for 1712 phecode traits with at least 20 cases. Traits with fewer than N=60 cases (vertical line) showed evidence of severe inflation. We report GWAS results for phecode traits with at least 60 cases.

