Michigan Genomics Initiative: Freeze 6 PheWeb

Emily Bertucci-Richter[1*], Matthew Zawistowski[1], Lars G. Fritsche[1], Brett Vanderwerff[1], Snehal Patil[1], Michael Boehnke[1], Xiang Zhou[1], and Sebastian Zöllner[1,3]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. [3]Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed: embricht@umich.edu

1. Overview

The Michigan Genomics Initiative (MGI)[1] has a wealth of genotype and electronic health record data available for research. To aid in the exploration of these high dimensional data, we provide the MGI PheWeb[2] which visualizes pre-computed multi-ancestry genome wide association studies (GWAS) for 1,728 Phecode[3] phenotypes using the most recent MGI Genetic Data Freeze 6[4]. Previous releases of the MGI PheWeb were based solely on genetically inferred European participants[1]. The results of multi-ancestry analyses have the potential to lend insight into relationships between clinical phenotypes and genetic variants which are shared across populations[5]. The PheWeb[2] is an online interface where GWAS and phenome wide association study (PheWAS) results can be explored through interactive Manhattan, regional association plots (Locuszoom), and Q-Q plots. Links to the GWAS Catalog, dbSNP, and the UCSC Genome Browser are provided for additional information on individual genetic variants. Results can be searched by phecode, genetic variant, or gene name to explore the associations between 52 million imputed genetic variants and 1,728 phecode phenotypes. Users can request access to summary statistics by contacting phdatahelp@umich.edu.

2. Study Population

We included participants from the MGI Data Freeze 6 for which we had International classification of diseases ICD9-CM or ICD10-CM diagnosis codes available (n = 80,381). The MGI cohort is primarily recruited during inpatient surgical procedures at Michigan Medicine (**Figure 1**)[1,4]. MGI Freeze 6 consists of primarily of individuals of majority European descent (EUR; 69,505; 86.5%) with the remaining participants being majority African (AFR; 4,980; 6.2%), West Asian (WAS; 2,221; 2.8%), East Asian (EAS; 1,782; 2.2%), Central/Southern Asian (CSA; 1,175; 1.5%), and Native American (AMR; 718; 0.9%)[4]. Genetically inferred females and males make up 53.8% and 46.2% of the study cohort, respectively. The mean participant age is 57.4 years (SD = 16.8). Additional cohort demographics are described in detail elsewhere[4].

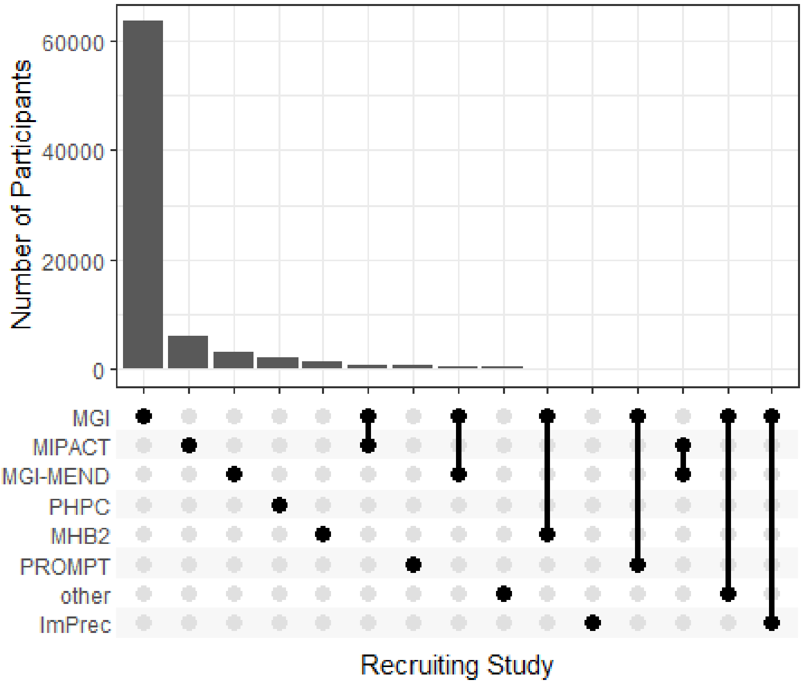

Figure 1. Study enrollment in MGI Freeze 6. Upset plot showing the contribution of recruitment studies for MGI Freeze 6. Only the largest 15 sets are plotted. Studies include the Michigan Genomics Initiative Anesthesiology Collection Effort (MGI), Michigan Predictive Activity and Clinical Trajectories (MIPACT), Metabolism Endocrinology & Diabetes (MGI-MEND), Michigan and You – Partnering to Advance Research Together (PHPC also known as MYPART), Mental Health BioBank (MHB2), PROviding Mental health Precision Treatment (PROMPT), and Immune Precision in Solid Organ Transplantation (ImPrec). Recruiting studies with less than 200 participants were combined into the “other” category for visualization and include the Biobank to Illuminate the Genomic Basis of Pediatric Disease (BIGBiRD), Michigan Neurological Disorders Precision Health Objective (MIND-PRO), Michigan eArly disease Progression cohort in COPD (MGI-MAP-COPD), Integration of Immune Phenotypes in Autoimmune Skin Disease (PerMIPA), Inflammatory Bowel Disease Databank (IBD-Biobank), and MGI-Dysplasia-Associated Arterial Disease Precision Health Network (MGI-DAAD).

3. Genetic Data

We used Trans-Omics for Precision Medicine (TOPMed) imputed genotypes for 52 million well imputed (Rsq ≥ 0.3) genetic variants. Analyses were restricted to variants with minor allele frequency (MAF) ≥ 0.001 and minor allele count (MAC) ≥ 20[4]. Genotyping, quality control, and genotype imputation methods are described in detail elsewhere[4].

4. Phenotyping

ICD9-CM and ICD10-CM billing codes codes were extracted from the De-Identified Research Data Warehouse (DeID RDW) on 09/18/2023. We mapped ICD codes to phecode phenotypes using the R PheWAS package (v0.99.6-1)[3,6,7]. We required a minimum ICD code count of 1 to determine phecode cases and exclusions. All default exclusions were applied including those for sex specific traits. Due to high p-value inflation observed previously in phenotypes with low case counts[1], we excluded any phecode phenotype with < 60 cases from further analyses. The number of phecodes per individual varied with age, with older individuals having a greater number of phecodes on average (**Figure 2A**). The number of phecodes per individual also varied by genetic ancestry with AFR and WAS having the greatest number of phecodes per sample (mean phecodes per sample: AFR = 87.5, AMR = 64.9, CSA = 58.2, EAS = 57.7, EUR = 74.4, WAS = 84.9; **Figure 2B**). Overall, we conducted GWAS for 1,728 traits across 17 phecode categories[3] (**Table 1**).

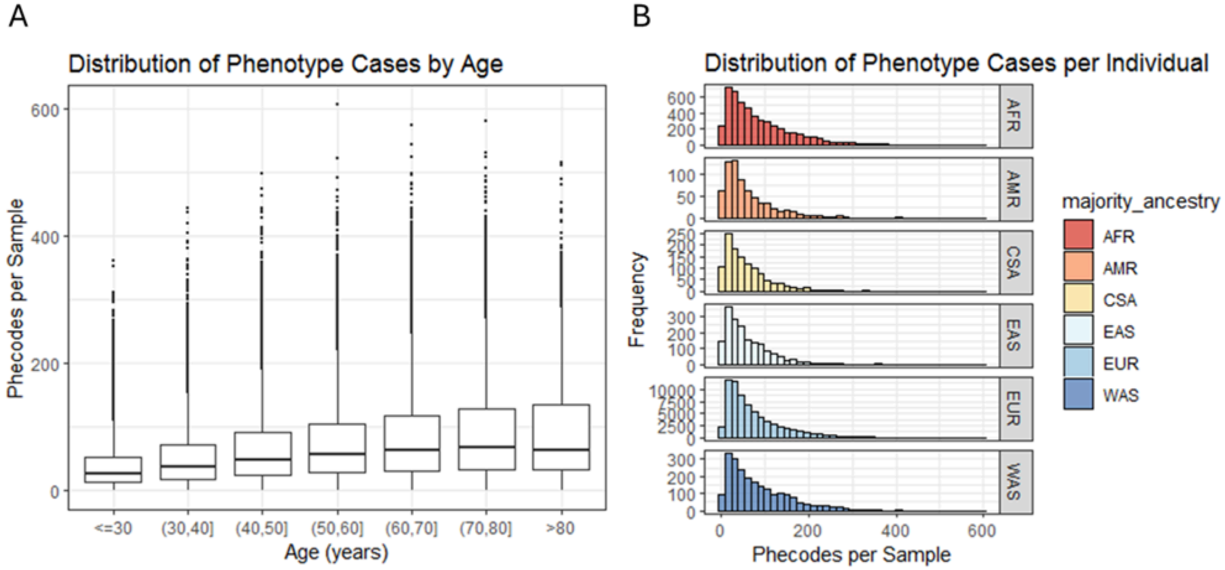

Figure 2. Phenotype distribution in MGI Freeze 6. The distribution of phecode cases across (A) participant age and (B) majority genetic ancestry. Colors represent majority genetic ancestry: red - African (AFR), orange - Native American (AMR), yellow - Central/Southern Asian (CSA), light blue - East Asian (EAS), blue - European (EUR), and dark blue - West Asian (WAS).

Phecode Category	Number of Traits
genitourinary	168
circulatory system	164
digestive	158
endocrine/metabolic	149
neoplasms	138
musculoskeletal	126
sense organs	119

injuries & poisonings	113
dermatologic	93
respiratory	85
neurological	84
mental disorders	73
infectious diseases	64
hematopoietic	58
congenital anomalies	54
symptoms	45
pregnancy complications	37

Table 1. Summary of phecode traits available in MGI Freeze 6. The number of phecode GWASs per phecode category.

5. Genome Wide Association Studies

We used SAIGE v1.3.0[8] to run a logistic mixed model with saddle point approximation and included age (as of January 1, 2023 for living participants and at death for deceased participants), genotype-inferred sex, genotyping array (CoreExome v1.0, CoreExome v1.1, CoreExomev1.3, or GSA v1.3), and the first 20 global principal components of ancestry[4] as covariates. To control for sample relatedness when fitting the null model in SAIGE step 1, we used a sparse genetic relatedness matrix (GRM) with a relatedness cutoff of 0.05. The sparse GRM was calculated using directly assayed autosomal genotypes. Prior to calculating the sparse GRM, we used PLINK for LD pruning by setting a squared correlation > 0.5, a walking window of 500 variants, and a step length of 5 variants. For association tests in SAIGE step 2, we used TOPMed imputed genotypes and excluded variants with low imputation quality (Rsq < 0.3) or very rare minor alleles (MAF < 0.001 or MAC < 20). Firth's test was applied to refine p-values < 0.01. We calculated the median genomic control values for variants with MAF > 1% for all phecode GWASs. If the genomic control value was greater than or equal to 1.05, we re-ran the GWAS using a full GRM generated on the fly in SAIGE step 1. The full GRM was also used for phenotypes with sample sizes ≥ 80,220 due to computational errors in SAIGE when performing factorization on the sparse GRM. When using the full GRM, the leave-one-chromosome out (LOCO) strategy was applied for autosomal variants, whereby the association test is conditional on the null model predictions made without using the chromosome where the variant is located. This was done to avoid proximal contamination[8].

After association analysis, we created genomic regions by including all variants 500 kilobases upstream and downstream of variants with $p<5x10^{-8}$. We then combined overlapping regions, identified the most significant variant within each region to be the top hit in the region, and refer to these top hits as independent associations here. This approach has been used to identify quasi-independent associations in previous phenome-level analyses of MGI[1].

6. Results

The majority of traits were run using the sparse GRM (n = 1,624) with a small number requiring the use of the full GRM (n = 104). The mean genomic control lambda[9] across all traits was 1.011 (± 0.013; **Figure 3**).

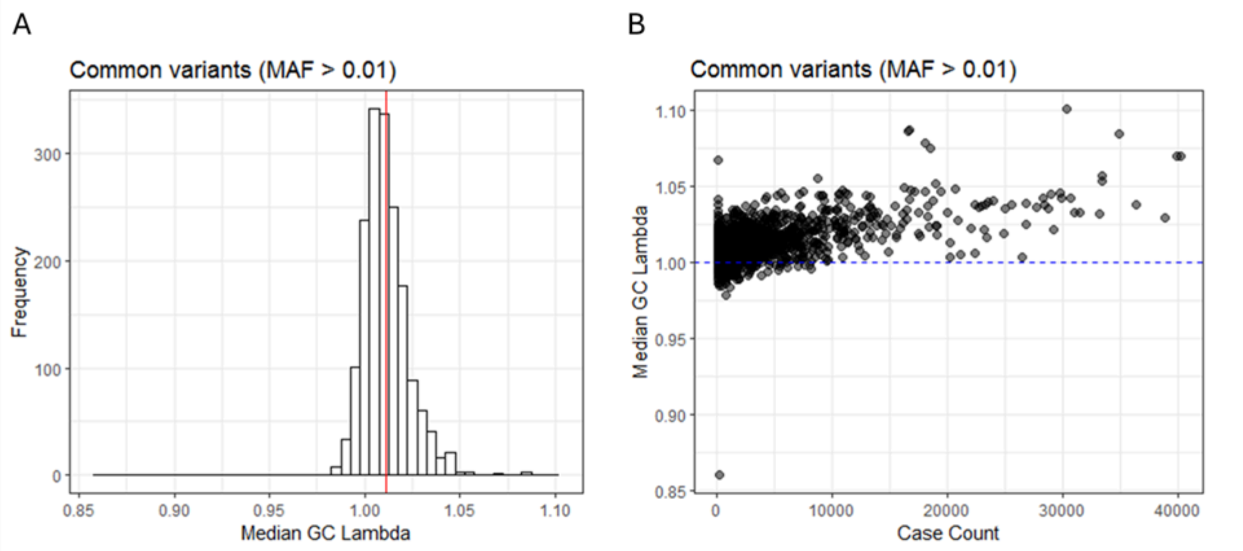

Figure 3. Distribution of Genomic Control Values in MGI Freeze 6 GWASs. (A) Frequency of the median genomic control lambda values across Freeze 6 GWASs. Red line in (A) represents a mean GC lambda of 1.011 across all traits. (B) Distribution of median GC lambda values across phecode case counts. The blue dashed line represents a GC lambda of 1.

We identified 1,516 independent top hits across 766 phecode phenotypes. Of these, 988 independent top hits across 472 phecode phenotypes have a MAF > 0.01. Caution should be taken when assessing associations where the MAF < 0.01 as they are more likely to be false positives[10]. An example of this can be seen in **Table 2**, where 4 likely spurious signals are identified for cancer of lip (X145.1) primarily due to the low MAF (≤ 0.001).

Phecode	Trait	CHR	Position	REF/ ALT	rsID	BETA (SE)	p.value	Case/ Control	MAF
X286.81	Primary hypercoagulable state	1	169549811	C/T	rs6025	2.330 (0.100)	4.5E-239	1449/ 64649	0.026
X286.8	Hypercoagulable state	1	169549811	C/T	rs6025	2.255 (0.099)	7.2E-231	1544/ 64649	0.026
X277.4	Disorders of bilirubin excretion	2	233759924	C/T	rs887829	1.193 (0.04)	1.17E-215	1650/ 69875	0.334

X654.2	Rhesus isoimmunization in pregnancy	1	25235176	G/A	rs55794721	2.372 (0.093)	4.35E-142	452/ 40843	0.360
X573.5	Jaundice (not of newborn)	2	233757337	A/G	rs1976391	0.814 (0.034)	1.93E-126	1969/ 59449	0.331
X286	Coagulation defects	1	169549811	C/T	rs6025	0.973 (0.047)	3.87E-94	7079/ 64649	0.026
X145.1	Cancer of lip	14	72285952	A/G	rs150669715	7.061 (0.344)	7.91E-94	118/ 75275	0.001
X286.12	Congenital deficiency of other clotting factors (including factor VII)	1	169549811	C/T	rs6025	3.004 (0.147)	2.73E-93	225/ 64649	0.023
X250.1	Type 1 diabetes	6	32658698	G/A	rs9273368	0.634 (0.032)	2.34E-88	4171/ 53267	0.272
X250.13	Type 1 diabetes with ophthalmic manifestations	6	32658698	G/A	rs9273368	1.046 (0.055)	7.27E-80	1330/ 53267	0.267
X145.1	Cancer of lip	X	87230155	T/C	rs144796369	3.674 (0.201)	6.65E-75	118/ 75275	0.001
X270.34	Alpha-1-antitrypsin deficiency	14	94378610	C/T	rs28929474	3.896 (0.217)	5.61E-72	143/ 74794	0.016
X172.2	Other non-epithelial cancer of skin	6	396321	C/T	rs12203592	0.370 (0.021)	4.65E-71	11082/ 65663	0.147
X250.12	Type 1 diabetes with renal manifestations	6	32658698	G/A	rs9273368	1.153 (0.065)	5.16E-70	956/ 53267	0.266
X286.7	Other and unspecified coagulation defects	1	169549811	C/T	rs6025	0.984 (0.056)	1.59E-69	4805/ 64649	0.025
X731.1	Osteitis deformans [Paget's disease of bone]	2	142181421	A/G	rs118023866	6.721 (0.385)	3.13E-68	61/ 58556	0.001
X172.21	Basal cell carcinoma	6	396321	C/T	rs12203592	0.462 (0.0267)	5.56E-67	5817/ 65663	0.144
X286.1	Congenital coagulation defects	1	169549811	C/T	rs6025	2.137 (0.124)	4.14E-66	464/ 64649	0.023

X275.1	Disorders of iron metabolism	6	26092913	G/A	rs1800562	1.706 (0.101)	2.30E-64	468/ 70712	0.053
X499	Cystic fibrosis	7	117559590	ATCT /A	rs113993960	2.813 (0.166)	4.23E-64	263/ 80118	0.013
X172	Skin cancer	6	396321	C/T	rs12203592	0.325 (0.019)	2.64E-63	13234/ 65663	0.148
X571.5	Other chronic nonalcoholic liver disease	22	43928850	C/T	rs738408	0.320 (0.019)	3.45E-61	8395/ 59449	0.232
X278.11	Morbid obesity	16	53767042	T/C	rs1421085	0.226 (0.014)	8.94E-61	16694/ 45492	0.388
X250.14	Type 1 diabetes with neurological manifestations	6	32706117	C/T	rs1794269	0.896 (0.055)	8.79E-60	1080/ 53267	0.372
X571	Chronic liver disease and cirrhosis	22	43928850	C/T	rs738408	0.308 (0.019)	1.94E-58	8735/ 59449	0.231
X145.1	Cancer of lip	10	104925395	A/C	rs117685299	6.959 (0.433)	4.72E-58	118/ 75275	0.001
X715.2	Ankylosing spondylitis	6	31368220	C/T	rs146683910	1.770 (0.112)	7.44E-56	414/ 54877	0.039
X272.1	Hyperlipidemia	19	44908822	C/T	rs7412	-0.330 (0.021)	1.56E-55	33356/ 46948	0.080
X272	Disorders of lipoid metabolism	19	44908822	C/T	rs7412	-0.328 (0.021)	6.40E-55	33433/ 46948	0.080
X145.1	Cancer of lip	10	2525677	G/A	rs117915764	6.354 (0.413)	2.16E-53	118/ 75275	0.001

Table 2. Top thirty strongest associations in MGI Freeze 6 GWASs. The top thirty most significant independent associations. Note that the same association is sometimes detected for multiple related subphenotypes (e.g. rs6025 is associated with 6 coagulation phenotypes).

7. Phenotype Genotype Reference Map

To assess how well our GWAS analyses were in replicating known associations, we used the Phenotype Genotype Reference Map[11] using both the European only and multi-ancestry GWAS catalog maps. We were well-powered (power > 80%) to detect 1,509 and 2,157 previously reported associations from the European and multi-ancestry GWAS Catalogs, respectively. Of these, we successfully replicated 69.9% of the European associations and

62.3% of the multi-ancestry associations (**Table 3**). The actual:expected ratio (AER), calculated as the number of replicated associations divided by the sum of the power, was 0.7 for European associations and 0.65 for multi-ancestry associations (**Table 3**). These values are somewhat lower than were reported previously in Freeze 3 MGI GWASs[1,11] (well powered replication rate = 76.1% and AER = 0.79), potentially due to the increase in heterogeneity when using a multi-ancestry study cohort.

PGRM population	Replication Rate well powered	Replication Rate all	Actual:Expected Ratio (AER)
EUR	69.6% (1,107 of 1509)	42.9% (1,881 of 4,388)	0.7
ALL	62.3% (1,343 of 2157)	41.1% (2,224 of 5,416)	0.65

Table 3. Phenotype Genotype Reference Map. Results from the PGRM for European only and multi-ancestry GWAS catalogs. The well powered replication rate refers to an estimated power of >80%. Numbers in parentheses report the number of variants replicated of the total number of PGRM associations.

9. Limitations of these Data

While multi-ancestry analyses are important for understanding relationships between genetics and disease risk in diverse populations, uncontrolled population stratification can result in both false positive and false negative GWAS signals[5]. Here we use a mixed model approach with a GRM[8] to control for sample relatedness and 20 PCs to correct for population stratification. While population stratification is not always adequately controlled when using a mixed model approach[5], it is less computationally expensive and potentially more powerful than conducting ancestry stratified GWAS and subsequent meta-analysis for the 1,728 traits in our PheWeb. We have carefully examined the results for systematic test statistic inflation and observed a mean genomic control of 1.01, in line with typical GWAS results from single population studies (**Figure 3**). We verified that well established genotype-phenotype associations from the PGRM[11] were replicated (**Table 3**). However, given that less than 14% of the MGI population is of non-European genetic ancestry[1,4], some care should be taken in interpreting the multi-ancestry results.

10. Accessing the PheWeb

Results can be viewed online on our PheWeb and summary statistics can be requested by contacting phdatahelp@umich.edu.

References

1. Zawistowski, M. *et al.* The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genomics* **3**, (2023).

2. Gagliano Taliun, S. A. *et al.* Exploring and visualizing large-scale genetic associations by using PheWeb. *Nature Genetics* vol. 52 550–552 (2020).

3. 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).

4. Vanderwerff, B. *et al. Michigan Genomics Initiative Freeze 6 Genome-Wide Genotypes*. (2023).

5. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* vol. 179 589–603 (2019).

6. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* **4**, 1–19 (2021).

7. Wei, W. Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* **12**, (2017).

8. 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).

9. Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997–1004 (1999).

10. Annis, A. *et al.* False discovery rates for genome-wide association tests in biobanks with thousands of phenotypes. (2021) doi:10.21203/rs.3.rs-873449/v1.

11. Bastarache, L. *et al.* The phenotype-genotype reference map: Improving biobank data science through replication. *Am J Hum Genet* **110**, 1522–1533 (2023).