

Article

# The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population

Anurag Verma <sup>1,2,\*</sup>, Scott M. Damrauer <sup>2,3,4</sup>, Nawar Naseer <sup>2</sup>, JoEllen Weaver <sup>2</sup>, Colleen M Kripke <sup>1,2</sup>, Lindsay Guare <sup>5</sup>, Giorgio Sirugo <sup>1,2</sup>, Rachel L. Kember <sup>6</sup>, Theodore G Drivas <sup>1</sup>, Scott M. Dudek <sup>3</sup>, Yuki Bradford <sup>3</sup>, Anastasia Lucas <sup>3</sup>, Renae Judy <sup>4</sup>, Shefali S. Verma <sup>5</sup>, Emma Meagher <sup>1,2</sup>, Katherine L. Nathanson <sup>1,7</sup>, Michael Feldman <sup>5</sup>, Marylyn D. Ritchie <sup>3</sup>, Daniel J. Rader <sup>1,2,3,\*</sup> and The Penn Medicine BioBank <sup>†</sup>

<sup>1</sup> Department of Medicine, Division of Translational Medicine and Human Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>2</sup> Institute for Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>3</sup> Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>4</sup> Department of Surgery, Division of Vascular Surgery and Endovascular Therapy, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>5</sup> Department of Pathology, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>6</sup> Department of Psychiatry, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>7</sup> Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

\* Correspondence: anurag.verma@pennmedicine.upenn.edu (A.V.); rader@pennmedicine.upenn.edu (D.J.R.)

† A full list of contributions from Penn Medicine BioBank Team is provided in the supplement.

**Citation:** Verma, A.; Damrauer, S.M.; Naseer, N.; Weaver, J.; Kripke, C.M.; Guare, L.; Sirugo, G.; Kember, R.L.; Drivas, T.G.; Dudek, S.M.; et al. The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **2022**, *12*, 1974. <https://doi.org/10.3390/jpm12121974>

Academic Editor: Michael F. Murray

Received: 8 October 2022

Accepted: 19 November 2022

Published: 29 November 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** The Penn Medicine BioBank (PMBB) is an electronic health record (EHR)-linked biobank at the University of Pennsylvania (Penn Medicine). A large variety of health-related information, ranging from diagnosis codes to laboratory measurements, imaging data and lifestyle information, is integrated with genomic and biomarker data in the PMBB to facilitate discoveries and translational science. To date, 174,712 participants have been enrolled into the PMBB, including approximately 30% of participants of non-European ancestry, making it one of the most diverse medical biobanks. There is a median of seven years of longitudinal data in the EHR available on participants, who also consent to permission to recontact. Herein, we describe the operations and infrastructure of the PMBB, summarize the phenotypic architecture of the enrolled participants, and use body mass index (BMI) as a proof-of-concept quantitative phenotype for PheWAS, LabWAS, and GWAS. The major representation of African-American participants in the PMBB addresses the essential need to expand the diversity in genetic and translational research. There is a critical need for a “medical biobank consortium” to facilitate replication, increase power for rare phenotypes and variants, and promote harmonized collaboration to optimize the potential for biological discovery and precision medicine.

**Keywords:** genomics; electronic health records; biobank; PMBB; precision medicine; learning health system

## 1. Introduction

Precision medicine incorporates clinical, environmental, lifestyle, family, and genomic data to tailor disease management and optimize disease prevention and health maintenance. Since the completion of the human genome project, large-scale genomic information linked to individual-level phenotype data has fueled biomedical discovery and has become an integral component of precision medicine. Large amounts of clinical data

are generated daily in clinical care and stored in the electronic health record (EHR). Linking the phenomic data encompassed in the EHR with biospecimens and genomic data from appropriately consented individuals at scale represents a tremendous opportunity for biomedical discovery and precision medicine. Penn Medicine, part of the University of Pennsylvania, is a large integrated academic health system with six hospitals and ten multispecialty centers that serve south-central Pennsylvania, south-central New Jersey, and northern Delaware. The Penn Medicine BioBank (PMBB) was launched with the intent of harnessing clinical data for discovery, creating a genomics-enabled learning healthcare system [1], and facilitating precision medicine for disease prevention and personalized therapy. Patients are enrolled under a single IRB-approved protocol that enables the acquisition of biospecimens, generation of genomic and multi-omic data, linkage to clinical information included in the EHR, and permission to recontact participants for future studies and/or the return of clinically relevant results. The scientific goal of the PMBB is to promote the integration of clinical and genomic data to power biomedical discovery and precision medicine. This report describes the operations and architecture of the PMBB and summarizes the information on the first ~170,000 participants recruited.

## 2. Materials and Methods

### 2.1. Planning and Development of PMBB

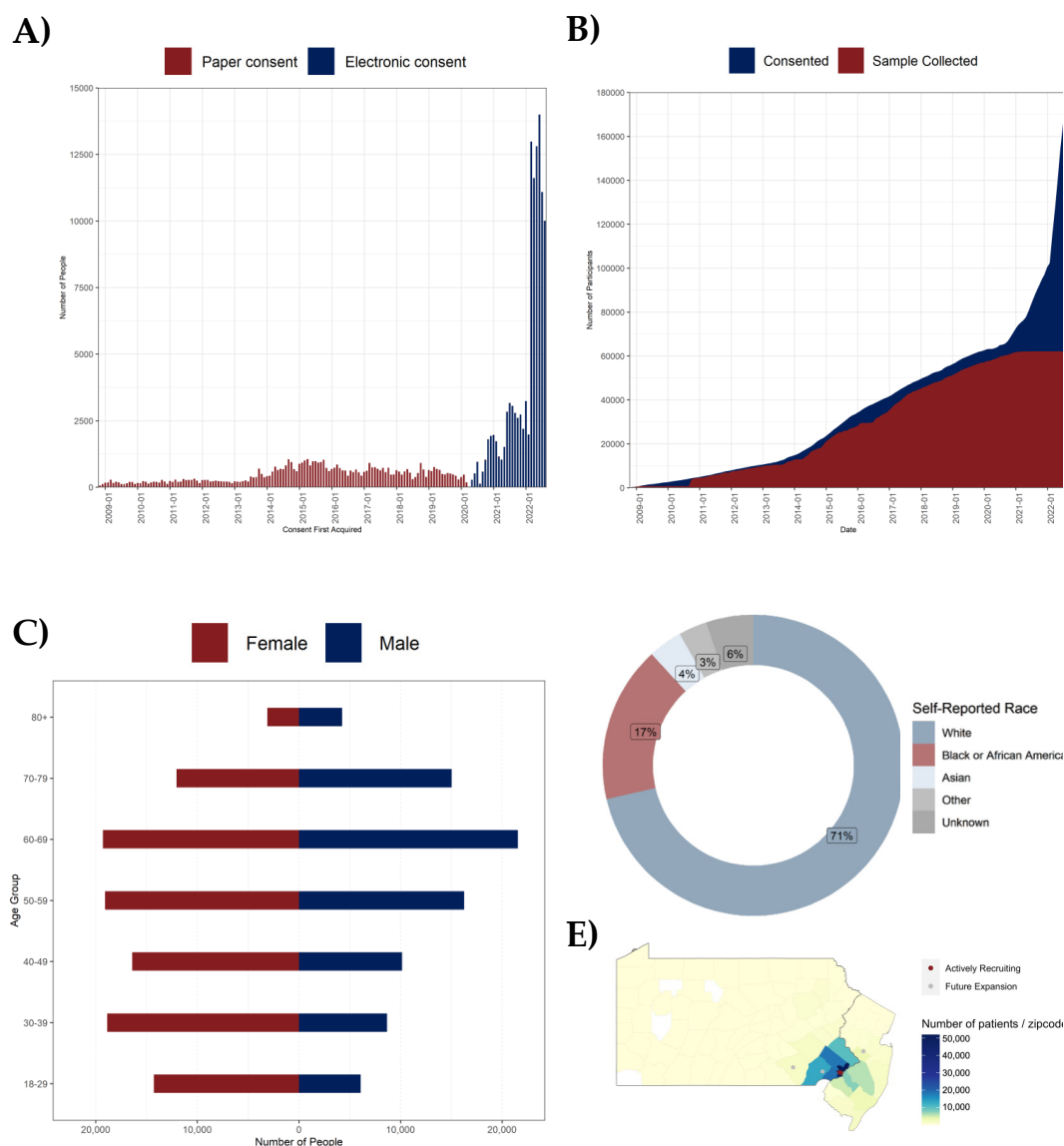
Recognizing the need for access to large numbers of appropriately consented and well-characterized human biospecimens to conduct translational research, in 2008, Penn Medicine established an IRB protocol and process for consenting patients and obtaining blood for genomic and biomarker research. In 2013, after a strategic planning process identified a pressing institutional mandate for an expanded biobank resource, the Penn Medicine BioBank (PMBB) was formally constituted, funded, and launched under the Institute of Translational Medicine and Therapeutics (ITMAT) in order to ensure it was integrated with critical infrastructure relevant to clinical and translational research and precision medicine.

The PMBB protocol was established as an institutional umbrella protocol under which any registered patient of Penn Medicine aged 18 or older was eligible, with no exclusions except an inability to provide informed consent. The core features of the consent include: (1) provision of a blood sample for biobanking and broad use for data generation, including genomic data and permission to bank any other residual tissues obtained in the context of clinical care; (2) permission to access data from the EHR for the purpose of research; and (3) permission to recontact participants for potential future studies or to return results.

### 2.2. Patient-Participant Recruitment

Initially, PMBB enrollment was accomplished through face-to-face encounters with clinical research coordinators (CRCs) in outpatient clinical areas, prioritizing locations where procedures that involved access to blood samples (phlebotomy labs, imaging procedures involving IV placement, cardiac catheterization labs, etc.) were being performed. After the onset of the COVID-19 pandemic, in August of 2020, the PMBB transitioned to remote recruitment efforts to prioritize patient-participant and staff safety by initiating an electronic consent and enrollment process through REDCap, a secure web platform for building and managing online databases and surveys. Simultaneously, a process for consent utilizing the EHR (PennChart, Epic) was developed, which was initially done in person at the time of patient check-in, and then also expanded to include consent at the time of pre-check-in through their myPennMedicine online patient portal, available through web browsers and mobile devices. Eligible patients scheduled for an upcoming outpatient office visit at one of the UPHS clinic sites actively recruiting (Figure 1E) receive the PMBB consent form as part of their electronic pre-visit check-in process and have the option to complete the form online through this portal. The three-page consent form includes a link

to the PMBB website, the PMBB email address, and the PMBB enrollment telephone number, which is staffed by CRCs on weekdays from 7:30 a.m. to 5 p.m. local time to answer questions potential participants may have as they are completing the consent procedure. A small percentage (~5%) of patients who either are not eligible to receive the online pre-check-in (e.g., certain surgical patients), or patients who skip the consent form during their online pre-check-in, are consented in person by registration desk staff when the patient physically reports for their appointment. PMBB brochures, with basic information about the PMBB, the PMBB website link, and contact information, are also available for registration desk staff to distribute to patients during the consent process. The website also contains a short video for patients explaining how PMBB participants contribute to scientific research.



**Figure 1.** Recruitment and Demographics. (A) Distribution of enrollment through paper and electronic consent. (B) Cumulative numbers of participants consented and biospecimen sample collection. (C) Distribution of participants by age group and self-reported sex. (D) Distribution of participants by self-reported race. (E) Density of recruitment around the six clinical sites of UPHS in Pennsylvania and New Jersey: Hospital of the University of Pennsylvania, Penn Presbyterian Medical Center, Pennsylvania Hospital, Chester County Hospital, Lancaster General Health, Princeton Health.

### 2.3. Sample Collection

The PMBB patient participants consent to the collection of identifying information (e.g., name, date of birth, medical record number, and contact information), information from medical records (e.g., test results, medical procedures, medical diagnosis and procedure codes, lab values, images such as X-rays, and medicines), blood samples (up to four tablespoons), urine, saliva and/or respiratory specimens, and residual samples from clinical pathology. There is no limit on the length of time samples may be kept in the biobank. All blood samples are banked as whole blood, plasma, serum, buffy coat, and DNA for future studies following stringent standard operating procedures. Samples are collected in sterile vacutainer tubes barcoded with an identification number and scanned into a system for sample tracking. Using the institution's adopted laboratory information system, LabVantage (LabVantage Solutions, Inc., Somerset, NJ, USA), biospecimens are processed and tracked with time-date stamps to document processing and freezing times throughout the laboratory workflow. Sample inventory is robustly supported with real-time, adaptive sample pull lists and images of sample pull locations, as well as simultaneous creation of distribution boxes and decrement of sample aliquots.

Prior to the COVID-19 pandemic, blood samples were collected from patients at the point of enrollment by CRCs cross-trained as certified phlebotomists. With the transition to electronic consenting, the consenting and sample collection processes have been decoupled. For sample collection, an automated weekly report is generated containing a list of consented patients for whom a blood sample is absent who have an upcoming phlebotomy appointment the following week at select Penn Medicine sites. Every Friday, three PMBB physicians place the electronic lab order for PMBB blood draws in the EHR of these patients. When they report to their phlebotomy appointment, the phlebotomist adds on the PMBB blood order to the patient's existing orders and collects the blood sample. One 6 mL EDTA tube is collected per patient. All blood samples are transported to the centralized clinical laboratory and logged in prior to being transferred to the PMBB core laboratory, where the blood is processed and stored following standard operating procedures.

Additionally, the PMBB banks residual biospecimens and tissues (for example, blood, urine, cerebrospinal fluid, or tissue collected as part of clinical care) when available as fresh, frozen or fixed dependent upon the tissue histology, following standard procedures. Residual tissues are released by the Department of Pathology after examination if the specimen or tissue is determined to be in excess of that required for patient care, or for tissue or bodily fluid that would otherwise be waste material.

### 2.4. Genomic Data Generation

#### 2.4.1. Genotyping and Imputation

DNA samples on approximately 44,000 PMBB participants have been genotyped to date on an Illumina Global Screening Array v.2.0 (GSAv2) by the Regeneron Genomics Center (RGC). The genotyping array chip has a backbone of 654,027 genetic markers as well as additional ancestry informative markers. In addition, approximately 8595 of the 44,000 PMBB participants were also genotyped in the Center for Applied Genomics (CAG) at the Children's Hospital of Philadelphia on the GSAv1 and GSAv2 genotyping array. After performing sample-level quality control (QC), genotype imputation was performed using Eagle2 [2] and Minimac [3] software on the TOPMed Imputation Server [3]. Imputation was performed for all autosomes, with TOPMed version R2 on a GRCh38 reference panel. Cosmopolitan post-imputation QC included imputation score filtering ( $R^2 > 0.7$ ), removal of palindromic variants, biallelic variant check, sex check, genotype call rate ( $>99\%$ ) and sample call rate ( $>99\%$ ) filtering, minor allele frequency filtering ( $MAF > 1\%$ ), and a Hardy–Weinberg equilibrium test ( $p\text{-value} > 1 \times 10^{-08}$ ). We generated PCAs to adjust for population structure and to identify genetically informed ancestry (GIA) from EIGENSOFT [4].

### 2.4.2. Whole Exome Sequencing

Whole exome sequencing (WES) has been performed on approximately 44,000 participants to date by the RGC. DNA samples were processed with the custom IDT xGen v1 exome capture platform and sequenced on the Illumina NovaSeq 6000 system on S4 flow cells. Sequence alignment, variant identification, and genotype assignment were performed using a WeCall variant caller. Sample level QC steps were then applied and sample sex errors, high rates of heterozygosity/contamination ( $D\text{-stat} > 0.4$ ), low sequence coverage (less than 85% of targeted bases achieving 20X coverage), or genetically identified sample duplicates, were excluded. Additional filters were applied to pVCFs. Any SNV genotype with a read depth of less than seven reads ( $DP < 7$ ) was changed to a no-call. After the application of the DP genotype filter, only the SNV variant sites that met at least one of the following two criteria were retained: (1) at least one heterozygous variant genotype with an allele balance ratio greater than or equal to 15% ( $AB \geq 0.15$ ); (2) at least one homozygous variant genotype. The same filtering was applied to INDEL variants but with an INDEL depth filter of  $DP < 10$  and an INDEL allele balance cutoff of  $AB \geq 0.20$ . Multi-allelic variant sites in the pVCF file were normalized by left-alignment and represented as bi-allelic. The variant frequency data for exome sequences and imputed data are available here: <https://pmbb.med.upenn.edu/allele-frequency> (accessed on 20 November 2022). The PMBB has a return of actionable results program for genomic findings that have a potential impact on participant health; this program is beyond the scope of this manuscript and will be described in a separate manuscript.

### 2.5. Clinical Data and Clinical Informatics Core

Clinical data are obtained through multiple sources, including a questionnaire completed at the time of recruitment and the Penn Medicine Clinical Data Warehouse and PennG&P (Penn Genotype & Phenotype) platform. PennG&P contains over 5.6 million patient records and other discrete clinical information amalgamated from 12 different source systems throughout the enterprise. PennG&P is based on a standard research data model called the Observational Medical Outcomes Partnership (OMOP), Common Data Model (CDM) [5], which is used worldwide by the Observational Health Data Sciences and Informatics (OHDSI) research consortium. It uses standardized language from national coding systems, such as SNOMED, LOINC [6], and RxNorm [7], for consistent terms and the labeling of information. Additionally, the PMBB maps International Classification of Diseases (ICD)-9 and ICD-10 codes to 1866 discrete disease traits using Phecode groupings [8].

### 2.6. Access to Data and Biospecimens

In keeping with the expansive mission of the PMBB, data and biospecimens are available to investigators throughout the Perelman School of Medicine and Penn Medicine for a broad range of research. External collaborations, including those with other academic institutions as well as biopharma, are encouraged and proceed through scientific collaboration with identified local Penn investigators. All research studies must have study specific IRB approval because the umbrella PMBB IRB protocol covers only sample and data acquisition, processing, storage, and dissemination. Researchers request access to data and biospecimens using a simple REDCap project proposal intake form which is then reviewed by the PMBB Steering Committee. Proposals are evaluated for scientific merit and priority, as well as the efficient use of data and biospecimens, as well as the ability to impact the care provided by Penn Medicine clinicians. The PMBB Steering Committee provides feedback to the investigator with either approval to move forward or with concerns to be addressed. This REDCap project also includes a Data Access Agreement form that must be signed by investigators prior to gaining access to any PMBB data. Per the terms of this agreement, the sharing of PMBB data with additional collaborators, whether

they are internal or external to Penn, must be handled by the PMBB and not the investigators themselves, to maintain the confidentiality and integrity of any protected health information (PHI) included within PMBB datasets.

Standardized EHR clinical data are deidentified and provided to approved investigators in a secure computing environment. For assistance with the creation of more complicated phenotypes, researchers have access to the Clinical Informatics Core (CIC), a shared resource that is managed by the Institute for Biomedical Informatics in collaboration with the PMBB. The CIC is staffed by clinical data scientists with expertise in data extraction, natural language processing (NLP), and data analysis.

### 3. Results

#### 3.1. Enrollment

During the initial phase of recruitment from 2008 to 2013, 13,366 Penn Medicine patients were enrolled (Figure 1A,B). In 2013, recruitment was expanded, resulting in a steady increase in enrollment to ~71,000 participants by the end of 2019 (Figure 1A,B). In 2020, the transition to electronic consenting triggered a rapid expansion in PMBB enrollment, with the total number of participants more than doubling between 2020 and 2022. There were 174,712 total participants enrolled in the PMBB as of September 2022 (Table 1, Figure 1A,B). Presently, nearly 3500 new participants are being enrolled weekly across two Penn Medicine hospitals that are actively recruiting through all their ambulatory sites, and recruitment at the other four Penn Medicine hospitals is targeted to begin in 2023. The PMBB currently has obtained and processed blood biospecimens from approximately 50% of enrolled participants; the recent shift to an electronic consent process has resulted in enrollment outpacing sample collection (Figure 1B). Active processes are underway to obtain biospecimens on the remainder of enrolled individuals. The goal is to enroll >1 million Penn Medicine patient participants, with >90% providing a blood sample for DNA and biomarker studies.

The PMBB participant population currently represents approximately 2.5% of active Penn Medicine patients. Similar to the general Penn Medicine patient population, a slightly higher percentage of PMBB participants are female (55.9%) as compared to male (44.1%) (Table 1, Figure 1C). The age distribution of PMBB participants also tracks with that of Penn Medicine patients, with participants ranging in age from 18 years to >100 years (Table 1 and Figure 1C). The age distributions differ slightly between sex, with males trending towards an older age (Figure 1C). The PMBB cohort is diverse: 17% of enrolled PMBB participants (and 25% of genotyped/sequenced participants) are identified as African American, 4% as Asian, and 3.3% as Hispanic (Table 1; Figure 1D). With nearly 30,000 African-American patient participants currently enrolled, the PMBB has, to our knowledge, the largest number of African-American participants of any single-institutional medical biobank in the US. As shown in Figure 1E, most biobank participants reside within the Philadelphia metropolitan area including New Jersey and Delaware (53.6%); there is also a total of 3.3% of participants from other states across the US.

**Table 1.** Comparison of PMBB Participant Characteristics with UPHS Patient Characteristics.

	<b>PMBB Participants n (%)</b>	<b>Genotyped Participants n (%)</b>	<b>UPHS Patients n (%) <sup>†</sup></b>
<b>Total</b>	174,712	43,884	3,688,610
<b>Gender</b>			
Female	97,674 (55.9%)	21,965 (50.1%)	2,042,868 (55.4%)
Male	77,055 (44.1%)	21,918 (49.9%)	1,604,210 (43.5%)
Other	17 (<1%)	1 (<1%)	107 (<1%)
<b>Age Range (Years)</b>	18–103	18–103	0–121

<b>Age Groups</b>			
18–29	17,815 (10.2%)	4302 (9.8%)	392,532 (10.5%)
30–39	27,355 (15.7%)	5406 (12.3%)	518,679 (14.1%)
40–49	25,819 (14.8%)	5688 (13.0%)	430,489 (11.7%)
50–59	33,827 (19.4%)	9519 (21.7%)	458,952 (12.4%)
60–69	40,268 (23.0%)	10,839 (24.7%)	507,397 (13.8%)
70–79	28,582 (16.4%)	5941 (13.5%)	400,016 (10.8%)
80+	8811 (5.0%)	2189 (5.0%)	333,781 (9.0%)
<b>Self-reported Race</b>			
African American	29,372 (16.8%)	10,815 (24.6%)	672,461 (18.2%)
White	124,406 (71.2%)	29,329 (66.8%)	2,029,684 (55.0%)
Asian	7156 (4.1%)	979 (2.2%)	152,615 (4.1%)
Other	9386 (5.4%)	1372 (3.1%)	370,313 (10.0%)
Unknown	7499 (4.3%)	1761 (4.0%)	463,537 (12.6%)
<b>Self-reported Ethnicity</b>			
Hispanic	5715 (3.3%)	1112 (2.5%)	174,179 (4.7%)
Non-Hispanic	165,713 (94.8%)	42,425 (96.7%)	3,290,018 (89.2%)
Unknown	3284 (1.9%)	347 (0.8%)	183,723 (5.0%)
<b>Genetically-Inferred Ancestry</b>			
African	N/A	11,300 (25.7%)	N/A
European	N/A	30,360 (69.2%)	N/A
East Asian	N/A	680 (1.5%)	N/A
South Asian	N/A	573 (1.3%)	N/A
Admixed American	N/A	711 (1.6%)	N/A
Other	N/A	301 (0.7%)	N/A
<b>Median period of EHR follow-up since enrollment</b>			
	7 years	5.7 years	N/A

<sup>†</sup> Demographics based on EHR data on UPHS patients who have been seen at least once from 2008 to present. Data from following UPHS sites were included: Hospital of the University of Pennsylvania, Penn Presbyterian Medical Center, Pennsylvania Hospital, Chester County Hospital, Princeton Health.

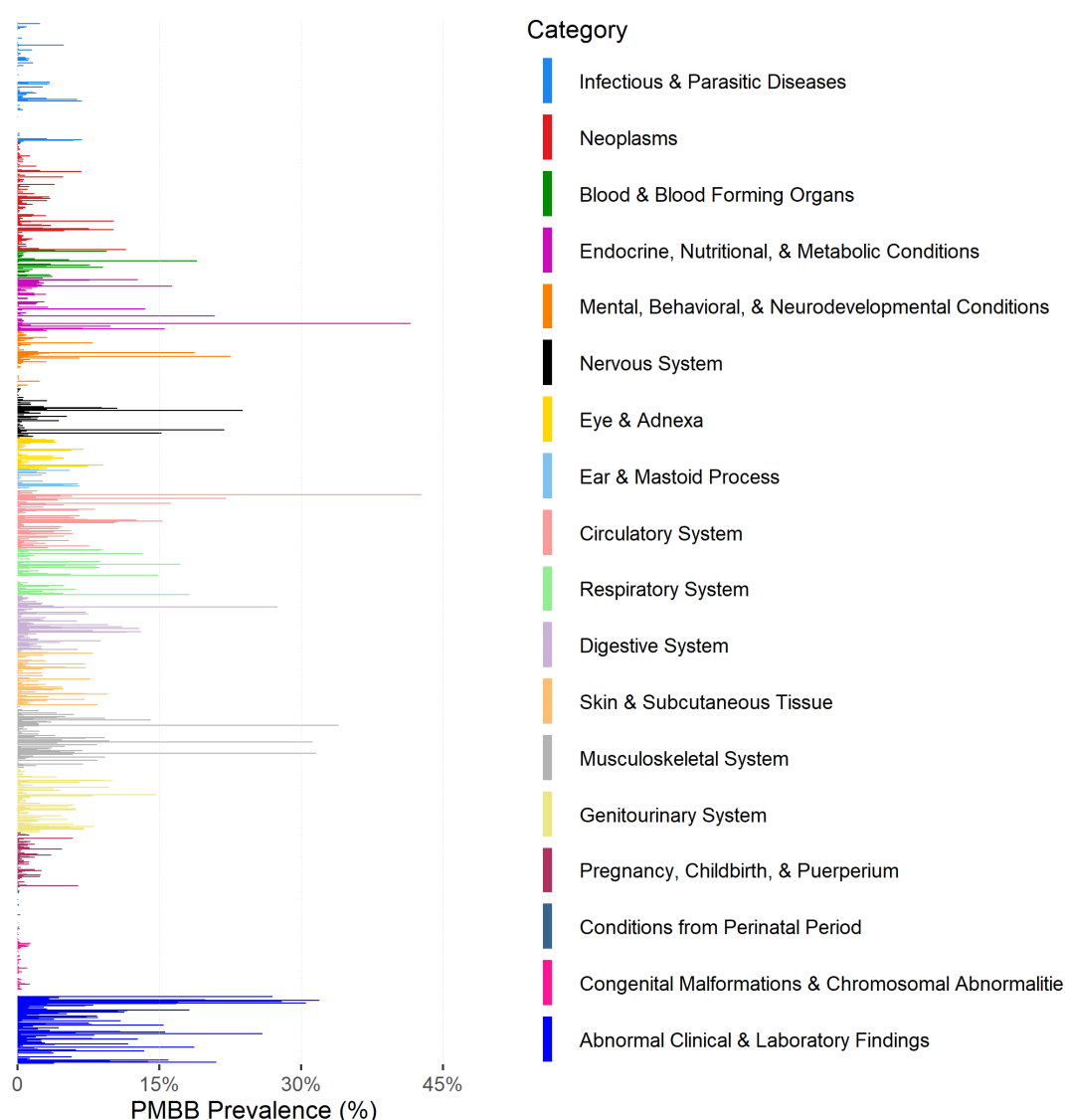
### 3.2. Clinical Data Availability

There are over 66.7 million data points collected in the form of encounters, diagnosis codes, procedure codes, and medication orders (Table 2). Across the PMBB cohort, over 46.7 million encounters, 10 million diagnosis codes, 3.6 million procedure codes, and 6.3 million medication orders have been recorded, averaging 268 encounters, 57 diagnosis codes, 21 procedure codes, and 36 medication orders per individual (Table 2). The most common diagnoses codes include hypertension, hypercholesterolemia, obesity, and insomnia (Figure 2).

**Table 2.** Clinical Data Availability for PMBB Participants.

Event Type	Total Number of Events	Mean Number of Events (SD) *
Encounter	46,738,773	268 (321)
Diagnosis Code (number of condition-related visits)	10,023,922	57.4 (58.7)
Procedure Code	3,621,056	20.7 (26.7)
Medication Order	27,914,486	159.8 (271.8)

\* Average number of events each patient has for each event type.

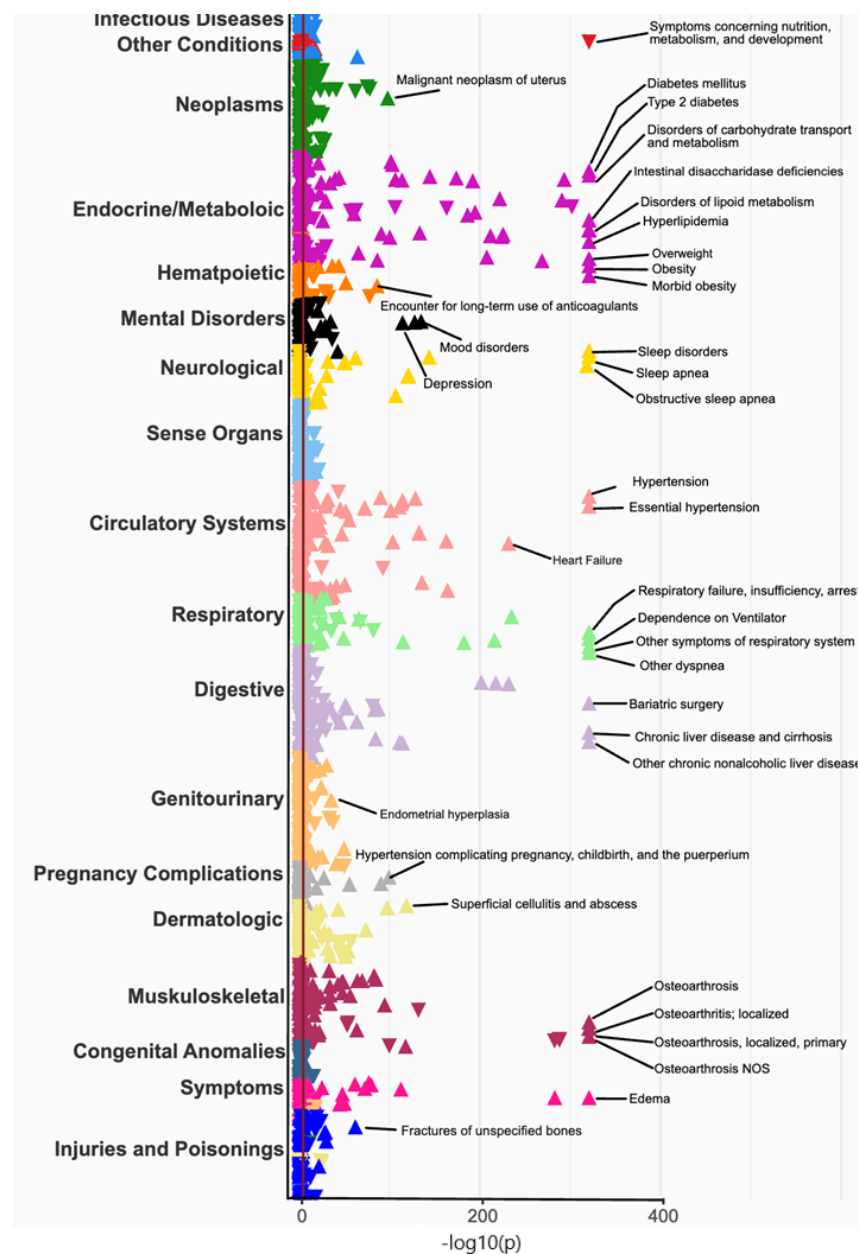
**Figure 2.** Prevalence of diagnoses code among PMBB participants grouped by broader disease domain.

### 3.3. Phenome-Wide Association Study (PheWAS) of BMI

To evaluate the validity of the PMBB clinical data, a phenome-wide association study (PheWAS) was conducted using mean body mass index (BMI) as the exposure and 1856 disease traits derived from grouping encounter diagnoses using phecodes as the outcome. All the models were adjusted with age, sex, and self-reported race in the EHR. A total of 662 associations of BMI with at least one disease trait across the 18 disease categories



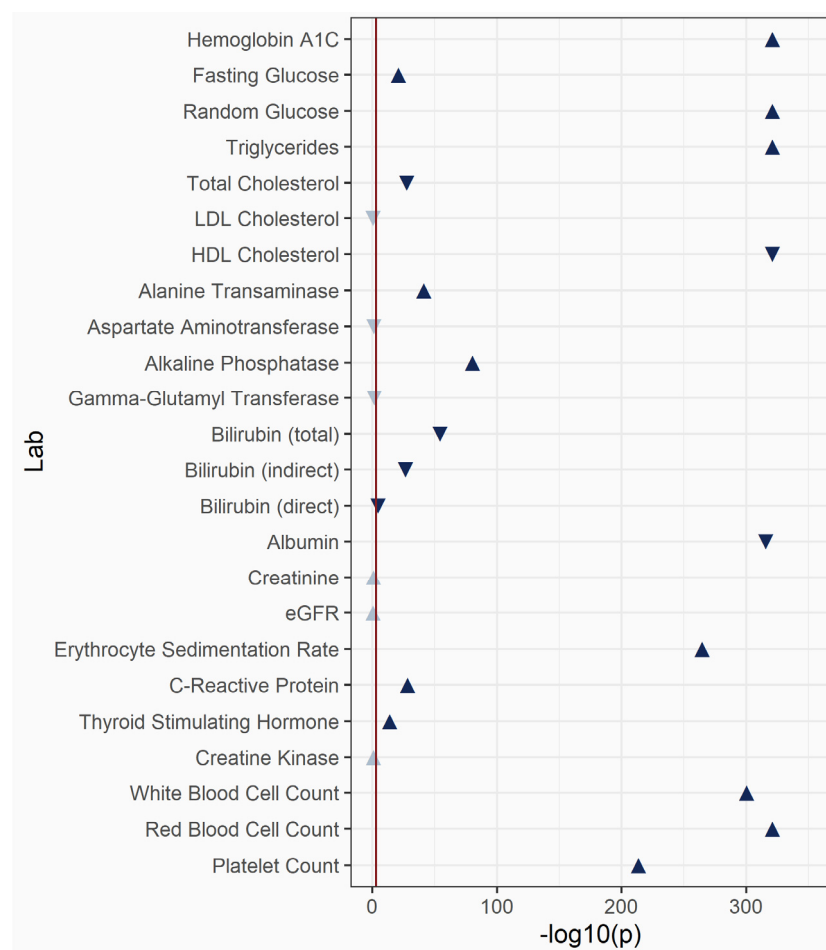
passed Bonferroni correction for multiple hypothesis testing ( $p < 2.6 \times 10^{-5}$ ). The strongest associations with BMI were with type 2 diabetes, hyperlipidemia, overweight, obesity, sleep apnea, and osteoarthritis (all  $p < 1 \times 10^{-331}$ , Figure 3, Supplementary Table S1). The associations with mean BMI show evidence of its effect on all organ systems, covering associations with the spectrum of disease categories (Figure 3). Additional associations include hypertension, heart failure, endometrial hyperplasia, bone fracture, depression, pregnancy complications, and respiratory failure (Figure 3).



**Figure 3.** A phenome-wide association between mean body mass index and 1856 EHR-derived phenotypes.

### 3.4. Laboratory-Wide Association Study (LabWAS) of BMI

We extracted 24 clinical laboratory measurements from the EHR for all the PMBB participants. These lab measurements were selected based on a common lab test in the health system and contained at least 1000 individuals within each lab which was measured. We computed median values for each individual within each lab and, as a proof-of-concept, evaluated their association with BMI. Linear regression was performed to test for association and all the models were adjusted with age, sex and self-reported race. We replicated many known associations between BMI and lab values (Figure 4, Table S2). For example, blood glucose measures were significantly associated with increased BMI. Triglyceride levels were significantly positively associated and high-density lipoprotein cholesterol (HDL-C) levels were significantly inversely associated with BMI, as expected. Markers of inflammation were also significantly associated with BMI. In this proof-of-concept analysis of the lab measurements, the associations with BMI support the association with the disease outcomes reported in the PheWAS.

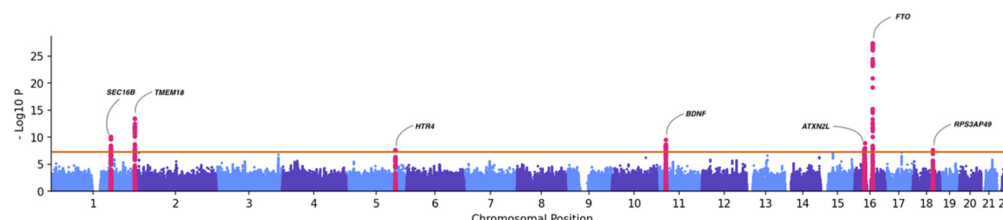


**Figure 4.** A laboratory-wide association between mean body mass index and 24 laboratory measurements derived from the electronic health records.

### 3.5. Genome-Wide Association (GWAS) with BMI

As a proof-of-concept, we conducted a GWAS for median BMI within five genetically inferred ancestry groups. These included 30,360 EUR, 11,300 AFR, 711 AMR, 680 EAS, and 573 SAS individuals in the PMBB (Table 1). The analysis tested the association of ~7.6 million SNPs with MAF > 1%, imputation R<sup>2</sup> > 0.3, using a linear mixed model implemented in REGNIE. All the models were adjusted with age, sex, and the first six ancestry specific principal components to account for population stratification. We then conducted cross-ancestry meta-analysis by integrating GWAS summary statistics from each ancestry

group using PLINK. Our meta-analysis identified 201 genome-wide significant SNP associations with BMI ( $p < 5 \times 10^{-8}$ , Figure 5), replicating several previously reported associations in published GWAS of BMI. The strongest association in our PMBB analysis was with *FTO* variant rs55872725 ( $p = 4.7 \times 10^{-28}$ ,  $\beta = 0.271$ ), which has been previously reported. Other known associations included rs7559547 ( $p = 3.92 \times 10^{-14}$ ,  $\beta = 0.41$ , *TMEM18*) and rs539515 ( $p = 9.02 \times 10^{-11}$ ,  $\beta = 0.36$ , *SEC16B*), among others. Summary statistics of results with  $p < 1 \times 10^{-4}$  are available in Supplementary Table S3.



**Figure 5.** Manhattan plot showing association between common genetic variants (MAF > 1%) and BMI.

#### 4. Discussion

The Penn Medicine BioBank was created to harness clinical data and biospecimens for biomedical research within Penn Medicine, a large academic healthcare system. Within a decade, it has emerged as a critical resource for translational medicine that has fueled discovery science and facilitated precision medicine, empowering a genomics-enabled learning healthcare system. As of September 2022, the PMBB had enrolled over 174,000 participants, obtained biospecimens on >70,000 participants, and generated genomic data on ~44,000 of its participants. The PMBB intends to enroll >1 million participants, obtain biospecimens on >90% of participants, and generate genomic data on all participants for whom biospecimens have been obtained.

The 2015 Institute of Medicine (now National Academy of Medicine) report on Translating Genomic-Based Research for Health [1] advocated for the development of ‘genomics-enabled learning healthcare systems’ based on the systematic summarized collection and use of genomic data, integrated with phenotypic data, to make discoveries and enhance healthcare in large healthcare systems. More recently, in its strategic vision for genomics research and application of genomics to clinical care, the National Human Genome Research Institute (NHGRI) emphasized the design and implementation of genomics-enabled learning healthcare systems to include infrastructure, resources, and technology development for genomics; the inclusion of underrepresented and minority groups to make genomic research more equitable; the development of multi-omics studies to get a comprehensive view of disease biology and the progression of diseases; and building tools to implement the knowledge back into the EHR to improve healthcare [9]. Large disease-agnostic and diverse medical biobanks at academic medical centers, such as the PMBB, are a critical component of fulfilling this vision.

Despite recapitulating health and disease traits from structured diagnosis codes, and the successful integration of this with genomic data [10–16], it must be acknowledged that diagnosis codes are crude approximations of underlying biological traits. As such, the future of EHR-empowered genomics research lies in ‘advanced phenotyping’ beyond diagnosis codes. These approaches include laboratory data, medication data, and other forms of structured data, all of which are relatively straightforward to access and use for research. Laboratory data, procedure codes, family history, and billing codes are all being mapped to concepts from various vocabularies (MONDO [17], SNOMED) to develop phenotypic algorithms that characterize the outcome of interest. Even more exciting and informative is the extraction of meaningful quantitative information from unstructured data (e.g., clinical notes, procedure reports, imaging reports, and pathology reports) using nat-

ural language processing (NLP) methods. To this end, the PMBB has deployed NLP software (Linguamatics, Cambridge, UK) to extract phenotypes from clinical notes and other unstructured data in PMBB participants. Furthermore, there is an immense amount of clinical imaging performed in medical centers and these images are another potential source of phenotypic data, sometimes referred to as ‘imaging-derived phenotypes’ (IDPs). In several ongoing PMBB efforts, deep learning and machine learning techniques are being used to translate imaging data, such as CT, MRI, and fMRI, into quantitative IDPs to fuel new discovery.

Another approach to collecting additional phenotype and exposure data that are absent in the EHR is using participant questionnaires. During the COVID-19 pandemic, an initial COVID-19 survey [18] was deployed to PMBB participants to collect information on symptoms, co-morbidities, and outcomes related to COVID-19. As the pandemic has progressed, we now administer an active longitudinal survey to follow participants for up to 18 months from their first COVID-19 diagnosis, yielding insights into long COVID. Combining survey results with biospecimen and EHR-derived phenotypes can shed light on factors that predict the onset of disease, refine preventative care, and optimize the clinical trial design. Current efforts are focused on extending active data acquisition through integrating mobile devices for both real-time data collection from survey questions and biometric activity data.

A major focus of biobank research is the use of genomics to understand the genetic architecture of health and disease and its implications for clinical care by linking phenomic efforts with genomic data obtained from biospecimens. Leveraging these approaches, the PMBB has developed a robust and diverse genomics research enterprise. Studies using PMBB data have highlighted the utility of the ‘genome-first’ approach’s utility in studying rare variants at scale and identifying new associations between genes and disease [10,11], as well as refined the range of the phenotypic presentation of individuals carrying rare impactful variants in known disease genes [12,19].

To support equitable genomic research, a commitment to participant diversity has been a hallmark feature of the PMBB since its inception. Seventeen percent of PMBB participants (and 24% of those for whom biospecimens are currently available) are African Americans or immigrants of African ancestry. This diversity has led to novel genomic and biological insights that directly impact the health of underrepresented groups. For example, recent work in the PMBB highlighted that hereditary amyloid transthyretin cardiomyopathy was a common yet markedly underdiagnosed cause of heart failure among African-American individuals [20], with many cases of the disease remaining undiagnosed even at a tertiary medical center such as Penn Medicine. Given the availability of targeted therapy, this finding advocates for the aggressive utilization of genomic and precision medicine to diagnose transthyretin cardiomyopathy in this population. This ‘genome-first’ approach is revealing an under-diagnosis of other genetic conditions. A ‘return of actionable results’ program is underway, and the implications for the greater utilization of genetic testing in clinical practice are clear.

The integration of genomic data into clinical practice is essential for the next generation of healthcare. Penn Medicine is at the forefront of developing techniques to provide high-quality patient care based on real-world evidence and genomic discoveries [21,22]. An analysis of pharmacogenetic (PGx) variants in the PMBB concluded that anticipatory genotyping can efficiently lead to the effective communication of PGx results to patients [23]. Polygenic risk scores (PRS) have been posited as a novel approach to leverage common-variant genetics for clinical care to predict the genetic risk for complex diseases, although the clinical utility of this approach has yet to be fully determined [24,25]. Researchers utilizing PMBB data reported that PRS for psychiatric disorders [26] and substance use disorders [27] has shown cross-trait associations beyond traditional diagnostic boundaries, suggesting broad effects of genetic liability for these disorders. Furthermore, PRS in

PMBB participants significantly increased the ability to identify the cancer status of European individuals but not African Americans, underscoring the need for large-scale genomic studies on non-white populations [13].

A critical feature of the PMBB is the availability of stored plasma and serum for biomarker analyses and integration with clinical and genomic data. Although the effort and expense of generating and storing plasma/serum aliquots are substantial, the benefit of this approach is rapidly becoming apparent. Multiple investigators have utilized the genomic data to identify PMBB participants with genotypes of interest to then use stored samples in cases and controls to measure biomarkers of interest. During the COVID-19 pandemic, access to stored samples from PMBB participants enrolled pre-pandemic who subsequently developed COVID-19 permitted investigators to address a number of important research questions [28–31]. Now, with over seven years of median follow-up data since the recruitment of PMBB patient participants, the stored samples are increasingly precious for their use in assaying biomarkers that may be predictive of incident disease. As methods for large-scale proteomics and metabolomics improve and the costs come down, the opportunity to generate plasma-based large-scale omics and integrate with genomics and clinical data is increasingly feasible and promises to further enhance biological discovery and precision medicine.

All PMBB participants consent to be recontacted, a critical feature of the protocol that is useful for several purposes. Patient-reported surveys represent an important addition to the EHR data for certain phenotypes as well as lifestyle and exposure data. Permission to recontact facilitates ‘recall-by-genotype’ deep phenotyping studies, which represents a tremendous opportunity for ‘genome-first’ discovery. Several investigators are actively performing studies in which the genomic data are used to identify individuals that carry rare variants in genes of interest or have a high polygenic risk for certain conditions, and participants are contacted to consider participation in hypothesis-driven deep phenotyping studies. Deep phenotyping can include targeted imaging, immunological profiling, provocative testing (e.g., oral glucose or fat tolerance test), creation of induced pluripotent stem cells (iPSCs), or any number of other clinical phenotyping approaches driven by the specific scientific question. Finally, the era of precision medicine will certainly include many clinical trials that are targeted to individuals of a specific inherited genotype; large medical biobanks with pre-existing genomic data, such as the PMBB, offer a fertile opportunity for the recruitment of individuals for such genotype-directed clinical trials.

## 5. Conclusions

The PMBB is a disease-agnostic institutional biobank under a single umbrella protocol based at a large academic health system with the purpose of promoting a genomics-enabled learning healthcare system to fuel scientific discovery, translational science, and precision medicine. A comprehensive biobank of DNA, plasma, and serum on all participants with selected other specimens and tissues on a subset of participants is linked to rich EHR clinical data, imaging, and survey data. The clinical database is standardized to OMOP and contains demographic, diagnoses (e.g., ICD-9/ICD-10 codes), procedures (e.g., Current Procedure Terminology—CPT codes), laboratory data, medication data, encounter types, socioeconomic factors, and survey data. The initiation of e-consenting has led to a substantial increase in the rate of enrollment. As of September 2022, genome-wide genomic data have been generated on ~44,000 participants and plasma multi-omics data on several thousand participants. The substantial representation of African-American patient participants in the PMBB addresses the urgent need to increase diversity in human genetic studies. Researchers with approved IRB protocols can request access to biobank samples and data through a data access portal. Publications supported by PMBB data and specimens can be found here: <https://pmbb.med.upenn.edu/pmbb/publications.html> (accessed on 20 November 2022). The PMBB is one of several large medical biobanks at academic medical centers in the US and is strongly supportive of the creation of a ‘medical biobank consortium’ to facilitate replication, increase power for rare phenotypes and variants, and

promote harmonized collaboration around genotype-directed deep phenotyping and recruitment into clinical trials.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jpm12121974/s1>, Table S1: Phenome-wide association study between body mass index (mean values) and EHR-derived phecodes. Table S2: Laboratory-wide association study between body mass index (mean values) and clinical labs from the EHR. Table S3: Summary statistics for genome-wide association study for BMI. Supplemental document: A full list of PMBB author contributions.

**Author Contributions:** Conceptualization, D.J.R., M.D.R., A.V., N.N., S.M.D. (Scott M. Damrauer), M.F. and K.L.N.; methodology, A.V., N.N., D.J.R., J.W., M.F., K.L.N. and M.D.R.; validation: A.V., C.M.K. and L.G.; software A.V. and S.M.D. (Scott M. Dudek); formal analysis, A.V., N.N., C.M.K. and L.G.; data curation, N.N., C.M.K., L.G. and A.L.; writing—original draft preparation, J.W., A.V., N.N., D.J.R. and M.D.R.; writing—review and editing, S.M.D. (Scott M. Damrauer), S.S.V., G.S., R.L.K., T.G.D., S.M.D. (Scott M. Dudek), A.L., Y.B., R.J., E.M., K.L.N. and M.F.; visualization, N.N., C.M.K. and L.G.; supervision, D.J.R. and M.D.R.; project administration, J.W.; funding acquisition, D.J.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** The PMBB is supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. KLN is supported by the Basser Center for BRCA.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of University of Pennsylvania (protocol codes: 808346 approved 07/01/2008, 813913 approved 4/3/2013, and 817977 approved 6/6/2013" for studies involving humans.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** All the data used to generate the figures were made available in supplementary.

**Acknowledgments:** We thank the patient-participants of Penn Medicine who consented to participate in this research program. We acknowledge the efforts of the PMBB staff (a full list of contributors is provided in the supplement). We thank the outstanding Penn Medicine Corporate IS team (Jessica Chen, Christine Vanzandbergen, Jeffrey Landgraf, Colin Wollack, Ned Haubein) for its major efforts to implement e-consenting in the EHR as well as biospecimen acquisition and tracking. We thank the Regeneron Genetics Center for partnership in generating genetic variant data and for scientific interactions. We thank the Smilow family for their generous gift that made the launch of the PMBB possible. Finally, we would like to thank Kevin Mahoney, Jon Epstein, Larry Jameson (Penn Medicine leadership), Michael Parmacek (Department of Medicine), Garret FitzGerald (Institute for Translational Medicine and Therapeutics), and David Roth (Penn Center for Precision Medicine) for their vision and support.

**Conflicts of Interest:** S.M.D. receives research funding from RenalytixAI, in-kind research support from Novo Nordisk, and personal consulting fees from Calico Labs. D.J.R. serves on scientific advisory boards for Alnylam, Novartis, Pfizer, and Verve. Regeneron has generated genomic data in PMBB participants. These entities had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of this manuscript; or in the decision to publish these results.

## References

1. *Medicine I. of Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research: Workshop Summary*; National Academies Press (US): Washington, DC, USA, 2015; ISBN 978-0-309-37112-4.
2. Loh, P.-R.; Danecek, P.; Palamara, P.F.; Fuchsberger, C.; AReshef, Y.; Finucane, H.K.; Schoenherr, S.; Forer, L.; McCarthy, S.; Abecasis, G.R.; et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **2016**, *48*, 1443–1448. <https://doi.org/10.1038/ng.3679>
3. Das, S.; Forer, L.; Schönerr, S.; Sidore, C.; Locke, A.E.; Kwong, A.; Vrieze, S.I.; Chew, E.Y.; Levy, S.; McGue, M.; et al. Next-generation genotype imputation service and methods. *Nat Genet.* **2016**, *48*, 1284–1287. <https://doi.org/10.1038/ng.3656>
4. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **2006**, *38*, 904–909. <https://doi.org/10.1038/ng1847>

5. Klann, J.G.; Joss, M.A.H.; Embree, K.; Murphy, S.N. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. *PLoS ONE* **2019**, *14*, e0212463. <https://doi.org/10.1371/journal.pone.0212463>.
6. McDonald, C.J.; Huff, S.M.; Suico, J.G.; Hill, G.; Leavelle, D.; Aller, R.; Forrey, A.; Mercer, K.; DeMoor, G.; Hook, J.; et al. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clin. Chem.* **2003**, *49*, 624–633. <https://doi.org/10.1373/49.4.624>.
7. Nelson, S.J.; Zeng, K.; Kilbourne, J.; Powell, T.; Moore, R. Normalized names for clinical drugs: RxNorm at 6 years. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 441–448. <https://doi.org/10.1136/amiajnl-2011-000116>.
8. Wu, P.; Gifford, A.; Meng, X.; Li, X.; Campbell, H.; Varley, T.; Zhao, J.; Carroll, R.; Bastarache, L.; Denny, J.C.; et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **2019**, *7*, e14325. <https://doi.org/10.2196/14325>.
9. Green, E.D.; Gunter, C.; Biesecker, L.G.; Di Francesco, V.; Easter, C.L.; Feingold, E.A.; Felsenfeld, A.L.; Kaufman, D.J.; Ostrander, E.A.; Pavan, W.J.; et al. Strategic vision for improving human health at The Forefront of Genomics. *Nature* **2020**, *586*, 683–692. <https://doi.org/10.1038/s41586-020-2817-4>.
10. Park, J.; Levin, M.G.; Haggerty, C.M.; Hartzel, D.N.; Judy, R.; Kember, R.L.; Reza, N.; Ritchie, M.D.; Owens, A.T.; Damrauer, S.M.; et al. A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes. *Genet. Med.* **2020**, *22*, 102–111. <https://doi.org/10.1038/s41436-019-0625-8>.
11. Park, J.; Packard, E.A.; Levin, M.G.; Judy, R.L.; Regeneron Genetics Center; Damrauer, S.M.; Day, S.M.; Ritchie, M.D.; Rader, D.J. A genome-first approach to rare variants in hypertrophic cardiomyopathy genes MYBPC3 and MYH7 in a medical biobank. *Hum. Mol. Genet.* **2022**, *31*, 827–837. <https://doi.org/10.1093/hmg/ddab249>.
12. Damrauer, S.M.; Hardie, K.; Kember, R.L.; Judy, R.; Birtwell, D.; Williams, H.; Rader, D.J.; Pyeritz, R.E. FBN1 Coding Variants and Nonsyndromic Aortic Disease. *Circ. Genom. Precis. Med.* **2019**, *12*, e002454. <https://doi.org/10.1161/CIRCGEN.119.002454>.
13. Wang, L.; Desai, H.; Verma, S.S.; Le, A.; Hausler, R.; Verma, A.; Judy, R.; Doucette, A.; Gabriel, P.E.; Nathanson, K.L.; et al. Performance of polygenic risk scores for cancer prediction in a racially diverse academic biobank. *Genet. Med.* **2022**, *24*, 601–609. <https://doi.org/10.1016/j.gim.2021.10.015>.
14. Kember, R.L.; Levin, M.G.; Cousminer, D.L.; Tsao, N.; Judy, R.; Schur, G.M.; Lubitz, S.A.; Ellinor, P.T.; McCormack, S.E.; Grant, S.F.A.; et al. Genetically Determined Birthweight Associates With Atrial Fibrillation: A Mendelian Randomization Study. *Circ. Genom. Precis. Med.* **2020**, *13*, e002553. <https://doi.org/10.1161/CIRCGEN.119.002553>.
15. Zhang, C.; Verma, A.; Feng, Y.; Melo, M.C.R.; McQuillan, M.; Hansen, M.; Lucas, A.; Park, J.; Ranciaro, A.; Thompson, S.; et al. Impact of natural selection on global patterns of genetic variation and association with clinical phenotypes at genes involved in SARS-CoV-2 infection. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2123000119. <https://doi.org/10.1073/pnas.2123000119>.
16. Bajaj, A.; Iheword, A.; Qiu, C.; Small, A.M.; Wei, W.-Q.; Bastarache, L.; Feng, Q.; Kember, R.L.; Risan, M.; Bloom, R.D.; et al. Phenome-wide association analysis suggests the APOL1 linked disease spectrum primarily drives kidney-specific pathways. *Kidney Int.* **2020**, *97*, 1032–1041. <https://doi.org/10.1016/j.kint.2020.01.027>.
17. Shefchek, K.A.; Harris, N.L.; Gargano, M.; Matentzoglou, N.; Unni, D.; Brush, M.; Keith, D.; Conlin, T.; Vasilevsky, N.; Zhang, X.A.; et al. The Monarch Initiative in 2019: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **2020**, *48*, D704–D715. <https://doi.org/10.1093/nar/gkz997>.
18. Verma, S.S.; Chung, W.K.; Dudek, S.; Williamson, J.L.; Verma, A.; Robinson, S.; Rader, D.J.; Reilly, M.P.; Sengupta, S.; FitzGerald, G.A.; et al. Research on COVID-19 through patient-reported data: A survey for observational studies in the COVID-19 pandemic. *J. Clin. Transl. Sci.* **2021**, *5*, e17. <https://doi.org/10.1017/cts.2020.509>.
19. Drivas, T.G.; Lucas, A.; Zhang, X.; Ritchie, M.D. Mendelian pathway analysis of laboratory traits reveals distinct roles for ciliary subcompartments in common disease pathogenesis. *Am. J. Hum. Genet.* **2021**, *108*, 482–501. <https://doi.org/10.1016/j.ajhg.2021.02.008>.
20. Damrauer, S.M.; Chaudhary, K.; Cho, J.H.; Liang, L.W.; Argulian, E.; Chan, L.; Dobbyn, A.; Guerraty, M.A.; Judy, R.; Kay, J.; et al. Association of the V122I Hereditary Transthyretin Amyloidosis Genetic Variant With Heart Failure Among Individuals of African or Hispanic/Latino Ancestry. *JAMA* **2019**, *322*, 2191. <https://doi.org/10.1001/jama.2019.17935>.
21. Lau-Min, K.S.; Asher, S.B.; Chen, J.; Domchek, S.M.; Feldman, M.; Joffe, S.; Landgraf, J.; Speare, V.; Varughese, L.A.; Tuteja, S.; et al. Real-world integration of genomic data into the electronic health record: The PennChart Genomics Initiative. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **2021**, *23*, 603–605. <https://doi.org/10.1038/s41436-020-01056-y>.
22. Lau-Min, K.S.; McKenna, D.; Asher, S.B.; Bardakjian, T.; Wollack, C.; Bleznuck, J.; Biros, D.; Anantharajah, A.; Clark, D.F.; Condit, C.; et al. Impact of integrating genomic data into the electronic health record on genetics care delivery. *Genet. Med.* **2022**, *24*, 2338–2350. <https://doi.org/10.1016/j.gim.2022.08.009>.
23. Verma, S.S.; Keat, K.; Li, B.; Hoffecker, G.; Risan, M.; Regeneron Genetics Center; Sangkuhl, K.; Whirl-Carrillo, M.; Dudek, S.; Verma, A.; et al. Evaluating the frequency and the impact of pharmacogenetic alleles in an ancestrally diverse Biobank population. *medRxiv* **2022**.08.26.22279261; doi: <https://doi.org/10.1101/2022.08.26.22279261>
24. Martin, A.R.; Kanai, M.; Kamatani, Y.; Okada, Y.; Neale, B.M.; Daly, M.J. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **2019**, *51*, 584–591. <https://doi.org/10.1038/s41588-019-0379-x>.
25. Sirugo, G.; Williams, S.M.; Tishkoff, S.A. The Missing Diversity in Human Genetic Studies. *Cell* **2019**, *177*, 26–31. <https://doi.org/10.1016/j.cell.2019.02.048>.

26. Kember, R.L.; Merikangas, A.K.; Verma, S.S.; Verma, A.; Judy, R.; Regeneron Genetics Center; Damrauer, S.M.; Ritchie, M.D.; Rader, D.J.; Bućan, M. Polygenic Risk of Psychiatric Disorders Exhibits Cross-trait Associations in Electronic Health Record Data From European Ancestry Individuals. *Biol. Psychiatry* **2021**, *89*, 236–245. <https://doi.org/10.1016/j.biopsych.2020.06.026>.
27. Hartwell, E.E.; Merikangas, A.K.; Verma, S.S.; Ritchie, M.D.; Regeneron Genetics Center; Kranzler, H.R.; Kember, R.L. Genetic liability for substance use associated with medical comorbidities in electronic health records of African- and European-ancestry individuals. *Addict. Biol.* **2022**, *27*, e13099. <https://doi.org/10.1111/adb.13099>.
28. Alanio, C.; Verma, A.; Mathew, D.; Gouma, S.; Liang, G.; Dunn, T.; Oldridge, D.A.; Weaver, J.; Kuri-Cervantes, L.; Pampena, M.B.; et al. Cytomegalovirus Latent Infection is Associated with an Increased Risk of COVID-19-Related Hospitalization. *J. Infect. Dis.* **2022**, *226*, 463–473. <https://doi.org/10.1093/infdis/jiac020>.
29. Banday, A.R.; Stanifer, M.L.; Florez-Vargas, O.; Onabajo, O.O.; Papenberg, B.W.; Zahoor, M.A.; Mirabello, L.; Ring, T.J.; Lee, C.-H.; Albert, P.S.; et al. Genetic regulation of OAS1 nonsense-mediated decay underlies association with COVID-19 hospitalization in patients of European and African ancestries. *Nat. Genet.* **2022**, *54*, 1103–1116. <https://doi.org/10.1038/s41588-022-01113-z>.
30. Anderson, E.M.; Goodwin, E.C.; Verma, A.; Arevalo, C.P.; Bolton, M.J.; Weirick, M.E.; Gouma, S.; McAllister, C.M.; Christensen, S.R.; Weaver, J.; et al. Seasonal human coronavirus antibodies are boosted upon SARS-CoV-2 infection but not associated with protection. *Cell* **2021**, *184*, 1858–1864.e10. <https://doi.org/10.1016/j.cell.2021.02.010>.
31. Flannery, D.D.; Gouma, S.; Dhudasia, M.B.; Mukhopadhyay, S.; Pfeifer, M.R.; Woodford, E.C.; Gerber, J.S.; Arevalo, C.P.; Bolton, M.J.; Weirick, M.E.; et al. SARS-CoV-2 seroprevalence among parturient women in Philadelphia. *Sci. Immunol.* **2020**, *5*, eabd5709. <https://doi.org/10.1126/sciimmunol.abd5709>.