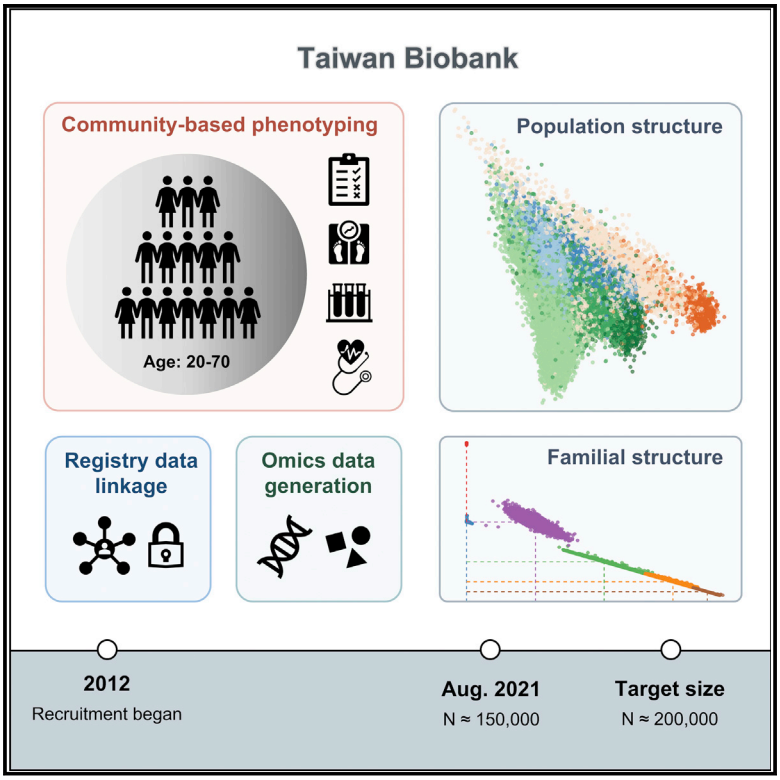
Cell Genomics

Resource

Taiwan Biobank: A rich biomedical research database of the Taiwanese population

Graphical abstract



Highlights

- Taiwan Biobank (TWB) is a prospective population study with >150,000 individuals
- Rich phenotypic and genetic data were collected in the Han-Chinese-based cohort
- Future linkage to national health registries will empower longitudinal genetic studies
- TWB studies contribute to population diversity in global genetics research

Authors

Yen-Chen Anne Feng, Chia-Yen Chen, Tzu-Ting Chen, ..., Tian Ge, Hailiang Huang, Yen-Feng Lin

Correspondence

ajfeng@ntu.edu.tw (Y.-C.A.F.), yflin@nhri.edu.tw (Y.-F.L.)

In brief

Feng. et al. describe an impressive biomedical resource in Taiwan that measures a wide range of characteristics in >150,000 individuals along with their genetic makeup. This resource enables large-scale epidemiological and genetic investigations and enriches population diversity in human genomic research.

Feng et al., 2022, Cell Genomics 2, 100197
November 9, 2022
https://doi.org/10.1016/j.xgen.2022.100197

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Resource

Taiwan Biobank: A rich biomedical research database of the Taiwanese population

Yen-Chen Anne Feng,[1,22,*] Chia-Yen Chen,[2] Tzu-Ting Chen,[3] Po-Hsiu Kuo,[1,4] Yi-Hsiang Hsu,[5,6,7,8] Hwai-I Yang,[9,10,11,12] Wei J. Chen,[1,3,4] Ming-Wei Su,[13] Hou-Wei Chu,[13] Chen-Yang Shen,[14,15] Tian Ge,[16,17] Hailiang Huang,[16,18,19] and Yen-Feng Lin[3,20,21,*]

[1]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan
[2]Biogen, Cambridge, MA, USA
[3]Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan
[4]Department of Psychiatry, College of Medicine and National Taiwan University Hospital, Taipei, Taiwan
[5]Marcus Institute for Aging Research and Harvard Medical School, Boston, MA, USA
[6]Beth Israel Deaconess Medical Center, Boston, MA, USA
[7]Harvard School of Public Health, Boston, MA, USA
[8]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[9]Genomics Research Center, Academia Sinica, Taipei, Taiwan
[10]Institute of Clinical Medicine, National Yang-Ming University, Taipei, Taiwan
[11]Graduate Institute of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
[12]Biomedical Translation Research Center, Academia Sinica, Taipei, Taiwan
[13]Taiwan Biobank, Academia Sinica, Taipei, Taiwan
[14]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
[15]Graduate Institute of Environmental Science, China Medical University, Taichung, Taiwan
[16]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[17]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[18]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[19]Department of Medicine, Harvard Medical School, Boston, MA, USA
[20]Department of Public Health & Medical Humanities, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[21]Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[22]Lead contact
*Correspondence: ajfeng@ntu.edu.tw (Y.-C.A.F.), yflin@nhri.edu.tw (Y.-F.L.)
https://doi.org/10.1016/j.xgen.2022.100197

SUMMARY

The Taiwan Biobank (TWB) is an ongoing prospective study of >150,000 individuals aged 20–70 in Taiwan. A comprehensive list of phenotypes was collected for each consented participant at recruitment and follow-up visits through structured interviews and physical measurements. Biomarkers and genetic data were generated from blood and urine samples. We present here an overview of TWB's genetic data quality, population structure, and familial relationship, which consists of predominantly Han Chinese ancestry, and highlight its important attributes and genetic findings thus far. A linkage to Taiwan's National Health Insurance database of >25 years and other registries is underway to enrich the phenotypic spectrum and enable deep and longitudinal genetic investigations. TWB provides one of the largest biobank resources for biomedical and public health research in East Asia that will contribute to our understanding of the genetic basis of health and disease in global populations through collaborative studies with other biobanks.

INTRODUCTION

Situated in East Asia, Taiwan is an island of 36,000 km^2 comprising a population of 23 million people with a well-established public health infrastructure. To advance epidemiological and biomedical research, Taiwan launched its own biobank, the Taiwan Biobank (TWB), in 2012, which prospectively collected a wide variety of lifestyle behaviors, environmental risk factors, and family history of common, complex diseases in the general Taiwanese population and measured genetic information in the biobank participants. One powerful aspect of TWB is the ability to link biobank subjects to Taiwan's own health insurance database and other longitudinal registries, which, similar to the Nordic countries, provide population-wide coverage of lifelong health information and events of nearly all of its citizens.

In companion with the Global Biobank Meta-analysis Initiative (GBMI) flagship paper,[1] we provide here a detailed description of the TWB resource for East Asian (EAS) genetics research and

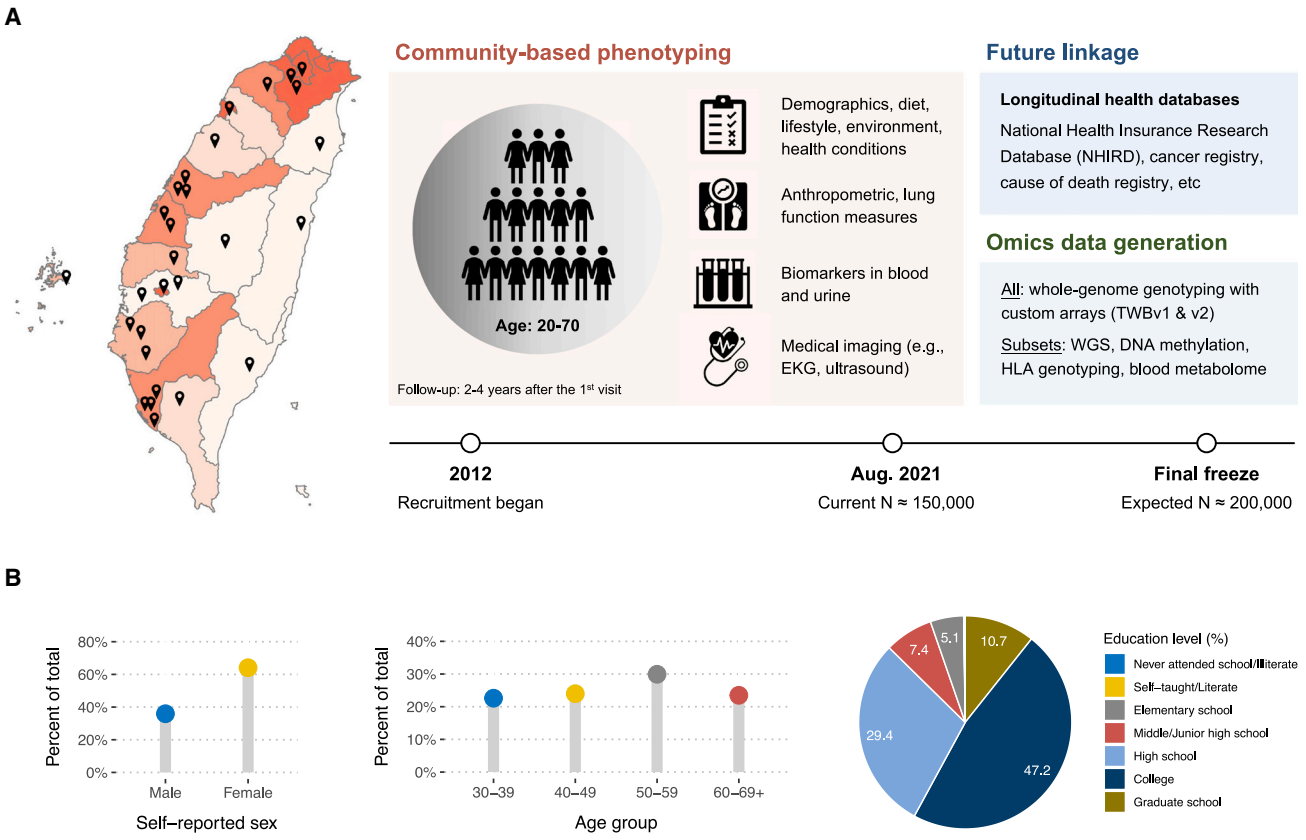


Figure 1. Design and demographics of the TWB

(A) TWB enrolled men and women aged 20–70 across recruitment centers in Taiwan that account for population density, with every county/city having at least one recruitment site (indicated by the location marks; a darker color represents a higher population density. The current release comprises individuals aged 30–70, but future releases will include participants down to age 20). Phenotypes were collected at baseline through a structured interview, physical examination, and blood/urine tests for each participant, with repeated measurements taken 2–4 years after the first visit. As of August 2021, 37,508 individuals have finished the first round of follow-up among all 150,000 participants. Multi-omics data were generated for all or subsets of the participants, including array genotyping, WGS, HLA typing, DNA methylation, and blood metabolome (see also Tables S1–S3). A linkage to the NHIRD and other health-related registries is underway to provide phenotypic information in addition to self-report for the TWB. Enrollment is expected to continue until reaching the target of 200,000 volunteers.

(B) The distribution of age, sex, and education level of study participants at baseline. More female than male participants were enrolled in TWB. Age of the participants was evenly distributed across every 10-year age bracket. People with a college degree accounted for the largest proportion in the current cohort.

beyond. Different from previous publications on TWB that focused on specific aspects or analyses of the TWB data, we aim at giving an overview of TWB—from cohort design, phenotype availability, genomic data generation, sample characteristics, genetic discoveries to date, and finally to its data access and sharing policy.

RESULTS

Overview of the TWB

The Taiwan Biobank (TWB) is a government-supported, prospective cohort study with a wide range of phenotypic measurements and genomic data collected on the Taiwanese population (https://www.twbiobank.org.tw).

Commenced in 2012 with a target sample size of 200,000 individuals, TWB employed a community-based approach and has been enrolling men and women aged 20–70 with no prior diagnosis of cancer from more than 30 recruitment sites across Taiwan, distributed based on population density of different counties and cities. At recruitment, participants provided a written informed consent and had their baseline data collected through questionnaires, physical examination, and blood and urine tests (Figure 1A; Table S1). Starting in 2016, repeated measurements of these phenotypes have been taken at prospectively planned visits, on average within 2–4 years since the initial visit. During the follow-up visits, participants would additionally undergo medical imaging examinations such as ultrasound and electrocardiogram scans. As of August 2021, a total of 151,406 volunteers have joined the biobank, among which 37,508 have completed the first round of follow-up. Whole-genome genotypes were measured for all participants, with a few other data types also available for selected subsets of the cohort, including whole-genome sequencing (WGS), DNA methylation, human leukocyte antigen (HLA) typing, and blood metabolome (Table S2). TWB has been releasing its data to the scientific community on a regular basis in a de-identified format.

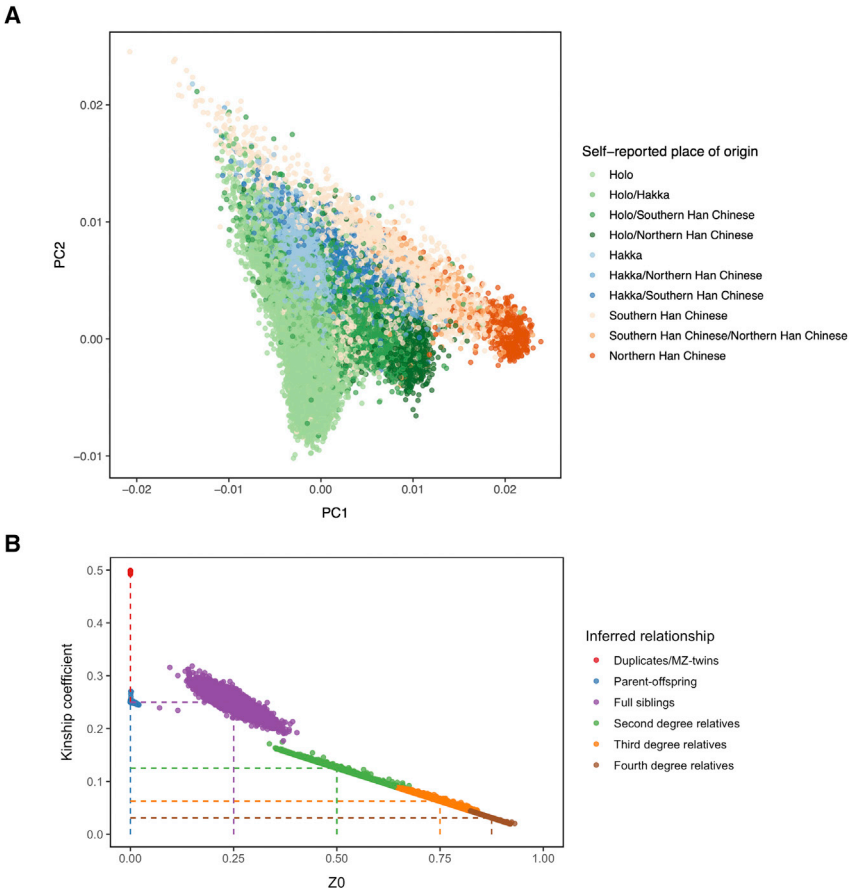


Figure 2. Population structure and familial relationship within TWB

(A) PCA on the QC'ed genotype data revealed a homogeneous population structure of primarily Han Chinese descent among the TWB participants, which can be further divided into three distinct subgroups representing different geographic and ancestral origins (Holo, Hakka, and Mainlanders). Mainlanders were roughly separated into Southern and Northern Chinese for visualization (details in Table S4, based on the GWAS sample in Chen et al.[3]). Participants with the same paternal and maternal place of ancestral origin were assigned into one single subgroup; those with mixed origins were assigned with A/B labels. Projection of TWB onto 1KG data showed tight clustering with the EAS superpopulation as well as the two Han Chinese populations (Figure S1). Shown here are results from batch 2 of 66,000 individuals. PC plots for batch 1, while not shown, were nearly identical to the batch 2 results.

(B) Kinship estimation of TWB participants showed a non-trivial number of relatedness, including over 25,000 pairs who are third-degree relatives or closer across batch 1 and batch 2 (Table S5, based on the GWAS sample in Chen et al.[3]). Plotted here are pairs of individuals in batch 2 with a kinship coefficient >0.02 (y axis; within fourth-degree relatedness), estimated using KING,[4] against the proportion of loci with 0 allele shared by descent (Z0; x axis). Dashed lines indicate the midpoints of cutoffs commonly used for determining the pairwise familial relationship based on kinship coefficients (duplicates/MZ twins: kinship > 0.353; parent-offspring: 0.177 < kinship < 0.353 and Z0 < 0.05; full siblings: 0.177 < kinship < 0.353 and Z0 > 0.05; second-degree relatives: 0.088 < kinship < 0.177; third-degree relatives: 0.044 < kinship < 0.088; fourth-degree relatives: kinship < 0.044). A substantial number of TWB participants were found to be fourth-degree relatives, among whom a random subset of 10,000 pairs were included for demonstration in the figure.

The current cohort comprises more females than males (64% versus 36%) and a similar number of participants across each 10-year age strata. 75% of the participants had a high school degree or higher at baseline (Figure 1B). Nearly all individuals are of Han Chinese descent (Figure 2A), reflective of the ancestry composition in Taiwan.[2] We demonstrate the biobank characteristics using a TWB release containing 108,955 individuals with genome-wide genotype data available at the time of our analysis.

Study population and participation

Designed as a prospective, population-based cohort for studying complex traits, TWB has been recruiting adult individuals (aged 20–70) with full capacity to consent who had not been previously diagnosed with cancer—a major non-communicable disease with an incidence expected to rise markedly after mid-life and a leading cause of death in Taiwan—at the time of enrollment.

Taiwan's population comprises an overwhelming majority of Han Chinese and a minority of indigenous tribes and new immigrants.[2] To ensure different ancestral origins are properly represented, TWB did not impose exclusion criteria on race/ethnicity groups. However, data collected from indigenous people are of highly restricted access, bound by strict ethical and legal guidelines in Taiwan. The public TWB release (including data described herein) does not contain information from indigenous peoples.

Since the launch of TWB, information sessions have been held regularly at enrollment sites across Taiwan to communicate its objectives, process, informed consent, and data security to the public. While TWB data and findings are primarily used for research activities, participants could apply for a report of health-related examinations and are free to request at any time not to be re-contacted, not to have their data linked to other databases, or to have a complete withdrawal.

Phenotyping procedures

Upon consenting to join the biobank, participants would undergo a physical examination and a structured interview with a well-trained researcher to report demographics, lifestyle behaviors, environmental exposures, dietary habits, family history, and health-related information in a questionnaire, lasting, on average, 1 to 1.5 h. TWB questionnaires were designed and vetted by expert epidemiologists; workgroups were formed to assess the clarity and logical flows of the questions as well as to conduct pretesting and a pilot study to validate

the questionnaire. Reliability of the questionnaire was also measured at pretesting and by comparing records at baseline and follow-up visits to ensure stability and consistency of the response over time. At baseline, participants provided blood and urine specimens for laboratory assays of biomarkers and omics data generation. Biospecimens were processed for experiments and underwent quality control (QC) procedures for release; the generated sensitive data were encrypted and stored at the Institute of Biomedical Sciences at Academia Sinica in Taiwan. The remaining DNA, plasma, and urine samples are actively monitored for their conditions and can be obtained at additional cost for research use.

At follow-up visits, a repeated assessment was carried out for phenotypes measured at baseline. Additionally, with feedback gathered from biobank users, efforts were expanded to include selected medical tests at follow-up to generate medical images such as abdominal ultrasound, bone mass density, and electrocardiograms for assessing relevant health conditions (Figure 1A) by medical professionals from collaborating hospitals. While the follow-up procedure is slowly expanding, the retention rate is currently high (70%–80%) among the re-contacted participants.

In sum, more than 1,000 phenotypes have been measured in TWB (Tables 1 and S1), including 21 blood and urine biomarkers indicative of a range of hematological, metabolic, kidney, and liver functions (Table S3).

Phenotype linkage and harmonization with national registries

With participants' full consent on linking their data to external databases—which can be opted out of at any time with no need of justification—individual disease and health information not recorded in the questionnaire can also be obtained through linkage to the National Health Insurance Research Database (NHIRD) and >70 additional databases that cover specific health-related outcomes in the Taiwanese population, such as the Taiwan Cancer Registry and cause of death registry.[5]

Taiwan's National Health Insurance (NHI) is a single-payer compulsory insurance program instituted in 1995 that provides accessible and affordable health care to all citizens in Taiwan (coverage > 99.99%).[6] The NHIRD is constructed for research purposes from registration and claims data in the NHI, actively maintained since 2002, and contains an extensive list of sorted data files, such as registry for beneficiaries, drug prescription registry, inpatient and outpatient claims, registry for medical facilities, and three embedded longitudinal health insurance databases.[5] Through linking to the unique personal identification numbers assigned to each individual in the NHIRD, data available only in other registries and regional hospitals may also be acquired. For example, founded in 1979, the Taiwan Cancer Registry includes a population-based collection of detailed cancer staging, treatment, and recurrence information for both in- and outpatients diagnosed with malignant neoplasms.[7] Cross-linkage between TWB and these health care databases currently only permits "on-site" analysis with a separate, study-specific institutional review board (IRB) approval. Efforts are underway to integrate health-related records from across these databases with the TWB phenotype collection.

Genotyping, QC, and imputation

TWB developed two custom single-nucleotide polymorphism (SNP) arrays for genotyping. Details of the design were described earlier.[8–10] Briefly, the TWBv1 array was designed in 2011 based on the Thermo Fisher Axiom Genome-Wide CHB Array with customized content that contains ~650,000 markers on the GRCh37 coordinates, providing a comprehensive coverage of common genetic variation for genome-wide association studies (GWASs). With a goal to capture not only GWAS markers but also functional variants, the TWBv2 array was later designed by Thermo Fisher Scientific in 2017 to deliberately enrich the content of rare coding risk alleles (e.g., protein-altering variants) based on WGS data from 946 TWB samples, including ~690,000 markers aligned to the GRCh38 reference build. ~28,000 and ~81,000 participants were genotyped on the TWBv1 array and TWBv2 array, respectively, with 1,462 duplicate individuals typed on both arrays to assess the concordance of SNP calls. The two arrays share ~100,000 markers.

We provide below an overview on data quality and structure of the TWB genetic data based on a subset of TWB participants, including all individuals genotyped on TWBv1 (batch 1) and ~85% of those genotyped on TWBv2 (batch 2; n ~ 69,000), as described in a related work[3] (see details in the STAR Methods; Table 1). Other studies utilizing the TWB data may follow a different QC pipeline with a different sample size, but the general rule should apply. Briefly, our QC was done separately for each batch, including pre-imputation QC that removed variants and samples with a high missing rate; principal-component analysis (PCA) to infer genetic ancestry and identify population outliers with 1000 Genomes (1KG; RRID:SCR_006828) samples[11] as the reference panel; within-East Asian (EAS) QC that discarded variants with low call rate or failing the Hardy-Weinberg equilibrium (HWE) test; phasing and imputation of TWB genotypes using the 1KG-EAS panel; and finally, post-imputation QC that removed additional variants of poor imputation quality (INFO score) or low minor allele frequency (MAF).

At INFO > 0.6, MAF distribution in the imputed data showed that each batch contained ~10% very rare variants (MAF < 0.001), ~15% rare variants (0.001 < MAF < 0.01), ~15% low-frequency variants (0.01 < MAF < 0.05), and ~60% common variants (0.05 < MAF < 0.5) (Figure S1). Retaining variants with an imputation quality INFO score > 0.6 and MAF > 0.5% after imputation, the final analytic sample consisted of ~8 million variants in ~93,000 individuals (n ~ 27,000 in batch 1 and n ~ 66,000 in batch 2;[3] Table 1). Previous studies have shown that greater genetic similarity between the reference panel and the target sample can improve genotype imputation accuracy, particularly for low-frequency variants.[10,12] With the 2,000-person WGS dataset in TWB (Table S2), internal efforts are in progress to develop a population-specific reference panel and server for genotype imputation of Han Chinese individuals.

Population structure and subgroups

Ancestral background is an important feature to study and account for in population-based epidemiological and genetic association analyses. Using a combination of self-report and a series of PCA, we modeled genetic diversity and inferred genetic ancestry of TWB participants. The results revealed that the

Table 1. An overview of sample characteristics of TWB participants at baseline

Characteristic	Samples with available genotype data[a]			The QC'ed analytical sample in Chen et al.[3,b]	
	All	Males	Females	Batch 1 (TWBv1)	Batch 2 (TWBv2)
Number of individuals	108,955	39,410	69,545	27,033	65,582
Age (mean ± SD)	49.9 ± 10.9	49.9 ± 11.3	49.9 ± 10.7	48.8 ± 11.1	50.5 ± 10.7
Sex (% women)	63.8%	–	–	50.0%	68.8%
BMI (kg/m^2; mean ± SD)	24.2 ± 3.8	25.4 ± 3.5	23.6 ± 3.7	24.3 ± 3.7	24.2 ± 3.8
Smoking experience (n; %)					
Never or rarely	87,742 (81%)	21,934 (56%)	65,808 (95%)	20,555 (76%)	55,190 (84%)
Ex-smoker	10,993 (10%)	9,282 (24%)	1,711 (2.5%)	3,272 (12%)	6,191 (9.4%)
Current smoker	10,218 (9.4%)	8,192 (21%)	2,026 (2.9%)	3,194 (12%)	5,623 (8.6%)
Alcohol consumption (n; %)					
Never or rarely	99,711 (92%)	32,071 (81%)	67,640 (97%)	24,298 (90%)	61,870 (94%)
Ex-drinker	2,790 (2.6%)	2,208 (5.6%)	582 (0.8%)	788 (2.9%)	1,578 (2.4%)
Current drinker	6,395 (5.9%)	5,106 (13.0%)	1,289 (1.9%)	1,913 (7.1%)	3,556 (5.4%)
Quantitative measurements (mean ± SD)					
Systolic blood pressure (mmHg)	119.3 ± 17.9	125.0 ± 16.5	116.1 ± 17.8	118.2 ± 17.4	119.7 ± 18.1
LDL cholesterol (mg/dL)	120.8 ± 31.7	121.7 ± 31.4	120.4 ± 31.8	120.9 ± 31.7	120.8 ± 31.6
FEV1.0/FVC (%)	79.8 ± 17.5	79.7 ± 17.4	79.9 ± 17.5	73.1 ± 18.8	81.1 ± 17.1
Fasting glucose (mg/dL)	95.9 ± 20.6	99.4 ± 23.4	93.9 ± 18.5	96.3 ± 21.0	95.7 ± 20.5
Creatinine (mg/dL)	0.72 ± 0.31	0.91 ± 0.37	0.62 ± 0.21	0.76 ± 0.36	0.71 ± 0.30
Albumin (g/dL)	4.52 ± 0.23	4.59 ± 0.24	4.48 ± 0.22	4.56 ± 0.24	4.50 ± 0.23
Most common self-reported diseases in TWB (n; %)					
Peptic ulcer	15,926 (15%)	6,548 (17%)	9,378 (13%)	3,923 (15%)	9,897 (15%)
Gastroesophageal reflux disease (GERD)	14,682 (13%)	5,054 (13%)	9,628 (14%)	3,306 (12%)	9,145 (14%)
Hypertension	13,432 (12%)	6,667 (17%)	6,765 (9.7%)	3,297 (12%)	8,445 (13%)
Drug medication	10,243 (9.4%)	2,934 (7.4%)	7,309 (11%)	2,403 (8.9%)	6,485 (9.9%)
Hyperlipidemia	8,019 (7.4%)	3,615 (9.2%)	4,404 (6.3%)	1,884 (7.0%)	5,020 (7.7%)
Kidney stone	7,003 (6.4%)	4,225 (11%)	2,778 (4.0%)	1,973 (7.3%)	4,116 (6.3%)

See also Table S1.

[a]TWB samples with available genome-wide genotyping data at baseline at the time of our analysis.

[b]TWB participants were genotyped using two custom genotyping arrays, TWBv1 (batch 1, n ~ 28,000) and TWBv2 (batch 2, n ~ 81,000). TWBv1 individuals and ~85% of TWBv2 individuals were included for GWASs.[3] Here shows the final analytical sample after batch-specific genotype QC.

population structure of TWB is highly homogeneous, comprising Taiwanese individuals of predominantly Han Chinese descent (>99%; Figure 2A, S2A, and S2B), largely consistent with the racial/ethnic demographic of Taiwan from the national census.[2] Based on the QC'ed genotype dataset of TWB, the first two principal components (PCs) distinguished individuals from various sub-continental geographic origins, including a majority (~78%) who self-identified as Holo (or Hoklo) or Hakka—the two largest Han Chinese groups in Taiwan who arrived from Southeast China during the last few centuries—as well as those who migrated to Taiwan after 1949 ("Mainlanders") and their descendants with a paternal/maternal ancestral origin that can be traced back to different provinces across Mainland China (~6%–7%; Figure 2A; Table S4). Interestingly, while Holo and Hakka people both originated from the Southeast coast of China in adjacent regions—mainly the Fujian and Guangdong provinces—PCA suggested a differentiation in genetic variation between the two ethnic groups, likely attributable to physical boundaries separating the two regions. Among the Mainlanders, as well as those who have mixed ancestral origins between Holo, Hakka, and the Mainlanders, a North-South gradient of genetic variation was also observed, which has been previously shown to correspond with the migration history of the Han Chinese population.[13] Further PCs formed a single ball-shaped cluster when plotted against each other, showing no additional signs of substructure (Figure S2C).

Familial structure and genetic analysis with related samples

Familial relationship among study participants provides critical analytical considerations for epidemiological and genetic research. Using genome-wide genotype data, we estimated pairwise kinship coefficients of TWB participants to identify related individuals not reported in the questionnaire. Among ~93,000 participants across the two batches (Table 1), 26,693 (28.8%) were found to have at least one related individual within third degree or closer that formed more than 25,000 related pairs, including 9 pairs of duplicates or monozygotic (MZ) twins, 4,154 parent-offspring pairs, 5,079 sibling pairs, 2,756 pairs of second-degree relatives, and >10,000 pairs of third-degree

relatives (Table S5; batch 2 results in Figure 2B). The non-trivial number of pairwise relatedness could be due to recruitment that took place in local communities, whereby family members might live in proximity and were more likely to be recruited together or invite one another to participate when informed of the TWB activity.

Given the degree of genetic relatedness in TWB, we considered genetic association tests that directly model relatedness in all samples to maximize power for GWASs. An overview of relatedness-aware analytical approaches in genetic studies is nicely summarized in Brumpton et al.,[14] including methods developed to efficiently analyze biobank-scale datasets (e.g., the mixed-model-based SAIGE [Scalable and Accurate Implementation of GEneralized mixed model] and SAIGE-GENE). To this end, we used a similar method, REGENIE, a two-step whole-genome regression that accounts for sample relatedness and population structure.[15] Assessment of the GWAS results showed adequate control of type I error and improved power from analysis restricting to unrelated individuals, highlighting the value of applying scalable mixed-model approaches to analyze large genomics datasets with a certain level of participant relatedness.[1,3]

Unique attributes of the biobank

Aside from phenotypes commonly shared across different biobanks, TWB measured lifestyle traits and environmental exposures that are more specific to the Taiwanese population as part of East and Southeast Asia. For example, included in the questionnaires were questions regarding exposure to different types of incense burning, such as mosquito repellent coils and joss sticks used in religious rituals, and betel nut chewing, a cultural tradition that has been associated with adverse health risk.[16] In addition, TWB employed a 44-item Body Constitution Questionnaire[17] that, according to traditional Chinese medicine theories, classifies individuals into one or more body constitution types, based upon which one's susceptibility to specific symptoms or diseases can be inferred to make personalized recommendations for promoting health-conscious lifestyle.

Genetic discoveries based on TWB

Several works utilizing TWB data have demonstrated its value in human genetic studies and expanded population diversity in genomic research in East Asia. Based on the initial release of 10,000 individuals, an earlier investigation provided a deeper dive into the population structure of TWB, showing that, in addition to Northern and Southern Han Chinese, there was a third cluster genetically similar to the Southern Han Chinese but with an extended haplotype in the major histocompatibility complex region, potentially reflecting a more recent evolutionary event.[8] Leveraging different kinds of physical activity measures in TWB, researchers have found that several exercises, such as regular jogging and mountain climbing, can attenuate the genetic influences on obesity and body mass index (BMI), suggesting converging evidence with previous findings from European (EUR)-centered genetic studies.[18] Measured in a subset of TWB participants, the WGS data of at least 30× coverage among 1,445 individuals from TWB have been demonstrated to outperform the 1KG EAS samples as an imputation panel for Han-Chinese-ancestry individuals, providing a moderate yet consistent improvement.[10] Exploring this Han-Chinese-specific reference panel also led to the identification of a great number of population-specific rare variants, many of which have not been seen in existing large allele frequency databases (e.g., gnomAD[19]), which has the potential to facilitate discovery of rare disease-causing variants, and, combined with the available information on common genetic variation from TWB array data, to improve genomic prediction for disease (e.g., hypertension).[9] More recently, we examined the genetic architecture of 36 quantitative traits, including anthropometric traits and biomarkers, across TWB, Biobank Japan, and UK Biobank, in which we identified TWB-specific novel genetic loci, fine-mapped those to putative causal variants, showed that genetic architecture of these traits was largely consistent within EAS as well as between EAS and EUR populations, and suggested the utility of biomarker GWASs in polygenic risk prediction for complex diseases in cross-population, multi-trait settings.[3]

DISCUSSION

The TWB is transforming biomedical research into the genetic basis of health and disease in the Taiwanese population. Here, we present an overview of the TWB resource (n ∼ 150,000 in 2021), illustrate the population structure and familial relationship within TWB, and highlight unique aspects and current genetic findings from TWB. In the next few years, TWB will continue to grow in sample size, accumulate longitudinal and repeated measures, and generate molecular phenotypes on the enrolled subjects. With the endeavor to link subjects to the NHIRD, which has covered almost all citizens of Taiwan since 1995, and other registries, TWB will develop into an invaluable resource integrating rich phenotypes and multi-omics data across a long time frame, delivering its mission of improving public health in Taiwan. Being one of the largest biobanks in East Asia, TWB will also contribute to the insights of human complex disorders and traits in world populations through integrative and comparative study with other global biobanks, as demonstrated by a series of studies in the GBMI special issue and beyond.[1,3]

Limitations of the study

One noticeable issue of TWB is that data might not be population-representative from community-based sampling, with visibly more women than men and no children in the cohort. Based on self-report alone, several common diseases also showed an in-sample prevalence lower than that in the general adult population (e.g., hypertension, 12.5% in TWB versus 26% reported from the Ministry of Health and Welfare in Taiwan for adults aged 20 and above; Table S6), which may reflect biases from self-report or similar to what was observed in the UK Biobank of a healthy volunteer bias.[20] On the other hand, with future linkage to the NHIRD and other registries, TWB will provide a useful resource for studying genetic and environmental liability to a broad spectrum of diseases and longitudinal outcomes beyond self-report in the Han Chinese population, including those traditionally found to be more prevalent in

Taiwan and its neighboring areas, such as nasopharyngeal carcinoma and liver diseases.[21,22]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Genotype data quality control
  - Estimation of familial relatedness
  - Relatedness-aware GWAS analysis in TWB

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100197.

ACKNOWLEDGMENTS

This research has been conducted using the Taiwan Biobank resource. We thank all the participants and investigators of the Taiwan Biobank. We thank the National Center for Genome Medicine of Taiwan for the technical support in genotyping. We thank the National Core Facility for Biopharmaceuticals (NCFB, MOST 106-2319-B-492-002) and National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) of Taiwan for providing computational and storage resources. Y.F.L. is supported by the National Health Research Institutes (NP-109-PP-09; NP-110-PP-09) and the Ministry of Science and Technology (109-2314-B-400-017; 110-2314-B-400-028-MY3) of Taiwan. T.G. is supported by NIA K99/R00AG054573. H.H. acknowledges support from NIDDK K01DK114379, NIMH U01MH109539, Brain & Behavior Research Foundation Young Investigator Grant, the Zhengxu and Ying He Foundation, and the Stanley Center for Psychiatric Research. Y.A.F. and P.H.K. are supported by the "National Taiwan University Higher Education Sprout Project" (NTU-110L8810) within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. Y.A.F. acknowledges support from the Yushan Young Fellow Program by the MOE at the National Taiwan University (NTU-111V1020-1). H.Y.H. acknowledges support from NIAMS R01AR072199.

The access to and use of the Taiwan Biobank data in the present work was approved by the Ethics and Governance Council (EGC) of Taiwan Biobank (approval number: TWBR10907-05) and the IRB of National Health Research Institutes, Taiwan (approval number: EC1090402-E). The data collection of Taiwan Biobank was approved by the EGC of Taiwan Biobank and the Department of Health and Welfare, Taiwan (Wei-Shu-I-Tzu No. 1010267471). Taiwan Biobank obtained informed consent from all participants for research use of the collected data and samples.

AUTHOR CONTRIBUTIONS

Y.A.F. led the manuscript with key input from Y.-F.L., C.-Y.C., H.H., and T.G. C.-Y.S. and H.-I.Y. played a critical role in conceiving and designing the Taiwan biobank; M.-W.S. and H.-W.C. led the implementation and maintenance effort of the biobank. T.-T.C., P.-H.K., Y.-H.H., and W.J.C. provided analysis or institutional support necessary for the work. Y.A.F. wrote the manuscript, which was critically reviewed, revised, and approved by all the authors for submission.

DECLARATION OF INTERESTS

C.-Y.C. is an employee of Biogen.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: December 21, 2021
Revised: August 26, 2022
Accepted: September 24, 2022
Published: October 12, 2022

REFERENCES

1. Zhou, W., Kanai, M., Wu, K.-H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2022). Global Biobank Meta-analysis Initiative: Powering genetic discovery across human diseases. Cell Genomics *2*, 100192. https://doi.org/10.1016/j.xgen.2022.100192.
2. Executive Yuan and Republic of China (Taiwan) (2016). The Republic of China Yearbook.
3. Chen, C.-Y., Chen, T.-T., Feng, Y.-C.A., Longchamps, R.J., Lin, S.-C., Wang, S.-H., Hsu, Y.-H., Yang, H.-I., Kuo, P.-H., Daly, M.J., et al. (2021). Analysis across Taiwan Biobank, Biobank Japan and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. Preprint at medRxiv. https://doi.org/10.1101/2021.04.12.21255236.
4. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). Minimac2: Faster genotype imputation. Bioinformatics *31*, 782–784. https://doi.org/10.1093/bioinformatics/btu704.
5. Hsieh, C.-Y., Su, C.-C., Shao, S.-C., Sung, S.-F., Lin, S.-J., Yang Kao, Y.-H., and Lai, E.C.-C. (2019). Taiwan's National Health Insurance Research Database: Past and future. Clin. Epidemiol. *11*, 349–358. https://doi.org/10.2147/clep.s196293.
6. National Health Insurance Administration, Ministry of Health and Welfare, Taiwan, R.O.C. (2014). Annual Report 2014-2015.
7. Chiang, C.-J., Wang, Y.-W., and Lee, W.-C. (2019). Taiwan's nationwide cancer registry system of 40 years: Past, present, and future. J. Formos. Med. Assoc. *118*, 856–858. https://doi.org/10.1016/j.jfma.2019.01.012.
8. Chen, C.-H., Yang, J.-H., Chiang, C.W.K., Hsiung, C.-N., Wu, P.-E., Chang, L.-C., Chu, H.W., Chang, J., Song, I.W., Yang, S.L., et al. (2016). Population structure of Han Chinese in the modern Taiwanese population based on 10, 000 participants in the Taiwan Biobank project. Hum. Mol. Genet. *25*, 5321–5331. https://doi.org/10.1093/hmg/ddw346.
9. Juang, J.-M.J., Lu, T.-P., Su, M.-W., Lin, C.-W., Yang, J.-H., Chu, H.-W., Chen, C.-H., Hsiao, Y.-W., Lee, C.-Y., Chiang, L.-M., et al. (2021). Rare variants discovery by extensive whole-genome sequencing of the Han Chinese population in Taiwan: Applications to cardiovascular medicine. J. Advert. Res. *30*, 147–158. https://doi.org/10.1016/j.jare.2020.12.003.
10. Wei, C.-Y., Yang, J.-H., Yeh, E.-C., Tsai, M.-F., Kao, H.-J., Lo, C.-Z., Chang, L.-P., Lin, W.-J., Hsieh, F.-J., Belsare, S., et al. (2021). Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. NPJ Genom. Med. *6*, 10. https://doi.org/10.1038/s41525-021-00178-9.
11. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
12. Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C.C., Soon, W.W.J., Bertin, N., Lin, C.B., Khor, C.C., et al. (2019). Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. Cell *179*, 736–749.e15. https://doi.org/10.1016/j.cell.2019.09.019.

13. Chen, J., Zheng, H., Bei, J.-X., Sun, L., Jia, W.-H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.-X., Zhang, X., and Liu, J. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am. J. Hum. Genet. *85*, 775–785. https://doi.org/10.1016/j.ajhg.2009.10.016.

14. Brumpton, B.M., Graham, S., Surakka, I., Skogholt, A.H., Løset, M., Fritsche, L.G., Wolford, B., Zhou, W., Nielsen, J.B., Holmen, O.L., et al. (2022). The HUNT Study: A population-based cohort for genetic research. Cell Genomics *2*, 100193. https://doi.org/10.1016/j.xgen.2022.100193.

15. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. *53*, 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

16. Gupta, P.C., and Warnakulasuriya, S. (2002). Global epidemiology of areca nut usage. Addict. Biol. *7*, 77–83. https://doi.org/10.1080/13556210020091437.

17. Lin, J.-D., Lin, J.-S., Chen, L.-L., Chang, C.-H., Huang, Y.-C., and Su, Y.-C. (2012). BCQs: A body constitution questionnaire to assess stasis in traditional Chinese medicine. Eur. J. Integr. Med. *4*, e379–e391. https://doi.org/10.1016/j.eujim.2012.05.001.

18. Lin, W.-Y., Chan, C.-C., Liu, Y.-L., Yang, A.C., Tsai, S.-J., and Kuo, P.-H. (2019). Performing different kinds of physical exercise differentially attenuates the genetic effects on obesity measures: evidence from 18, 424 Taiwan Biobank participants. PLoS Genet. *15*, e1008277. https://doi.org/10.1371/journal.pgen.1008277.

19. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

20. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am. J. Epidemiol. *186*, 1026–1034. https://doi.org/10.1093/aje/kwx246.

21. Chang, E.T., and Adami, H.-O. (2006). The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol. Biomarkers Prev. *15*, 1765–1777. https://doi.org/10.1158/1055-9965.epi-06-0353.

22. Sarin, S.K., Kumar, M., Eslam, M., George, J., Al Mahtab, M., Akbar, S.M.F., Jia, J., Tian, Q., Aggarwal, R., Muljono, D.H., et al. (2020). Liver diseases in the Asia-Pacific region: A Lancet Gastroenterology & Hepatology Commission. Lancet Gastroenterol. Hepatol. *5*, 167–228. https://doi.org/10.1016/s2468-1253(19)30342-5.

23. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

24. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

25. Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.-Y., Popejoy, A.B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R.J., Brick, L., et al. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. Cell *179*, 589–603. https://doi.org/10.1016/j.cell.2019.08.051.

26. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448. https://doi.org/10.1038/ng.3679.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
1000 Genomes Phase 3 data	1000 Genomes Project Consortium et al.[11]	RRID:SCR_006828
Taiwan Biobank	This paper	https://taiwanview.twbiobank.org.tw/data_appl; https://www.biobank.org.tw/
Software and algorithms		
Regenie	Mbatchou et al.[15]	https://github.com/rgcgithub/regenie
PLINK2	Chang et al.[23]	https://www.cog-genomics.org/plink/2.0
KING	Manichaikul et al.[24]	https://www.kingrelatedness.com/
Analysis codes generated for this manuscript	This paper	https://doi.org/10.5281/zenodo.7091819

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yen-Chen Anne Feng (ajfeng@ntu.edu.tw).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data including individual genotypes and phenotypes of TWB participants is available upon application for research purposes. A detailed description of TWB data availability and the application process can be found at https://taiwanview.twbiobank.org.tw/data_appl and https://www.biobank.org.tw/. In brief, investigators who are interested in obtaining the TWB data would need to submit an application that includes a detailed research proposal and an IRB approval from the applicant's home institute to the TWB Data Release Group (contact e-mail: biobank@gate.sinica.edu.tw). The application will go through scientific and ethical reviews by external experts in the relevant scientific fields Ethics and Governance Council and the Ethics and Governance Committee of TWB. Once approved, researchers will be able to obtain the data for the approved research projects during the approved time period. For researchers who are interested in applying but reside outside of Taiwan, or any cross-country collaborations, an additional international data transfer agreement needs to be filed to the Ministry of Health and Welfare of Taiwan to enable sharing of the TWB individual-level data or any of its derivatives.
- All original code used in this manuscript has been deposited on Zenodo and is publicly available at https://doi.org/10.5281/zenodo.7091819 (also listed in the key resources table), including scripts for genotype data QC, figure generation, and genome-wide association analysis.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Genotype data quality control

To provide an overview on data quality and structure of the TWB genetic data, we describe below the QC and analysis procedures performed in a related work.[3] Specifically, as the analysis was initially designed to investigate the polygenic architecture of complex traits in TWB, we retained all individuals genotyped on TWBv1 (batch 1) and only ~85% of those genotyped on TWBv2 (batch 2; n~69,000) as the discovery sample for GWAS, while using the remaining samples for polygenic prediction and validation.[3] The following QC was described for the discovery sample subset of TWB (Table 1).

Considering differences in marker content between the two genotyping arrays, our quality QC procedures were performed for each batch separately using custom scripts adapted from https://github.com/Annefeng/PBK-QC-pipeline using PLINK2,[23] which largely followed the recommendations laid out in Peterson et al.[25] for samples with varying ancestral backgrounds to minimize the impact of population stratification.

For pre-imputation QC, we first removed variants with call rate <0.98, samples with call rate <0.98, as well as variants that were monomorphic, duplicated, or not confidently mapped to a genomic position. Next, we inferred genetic ancestry of TWB participants using 1000 Genomes (1KG) phase 3 samples[11] as the population reference panel. Specifically, we merged the TWB data with the 1KG data and performed PCA based on high-quality, common SNPs (biallelic; non-strand ambiguous; call rate >0.98; MAF >5%; outside the long-range LD regions [chr6:25-35Mb; chr8:7-13Mb]) in approximate linkage equilibrium. Using the 1KG data as the training set, we then fit a Random Forest classifier with the first 6 PCs to assign ancestry of TWB participants at a prediction probability >0.8 into each of the five 1KG super-populations: European (EUR), African (AFR), Admixed/Latino American (AMR), East Asian (EAS), and South Asian (SAS).

Focusing on the predicted EAS ancestral group that contained the majority of the TWB participants, we further removed duplicate samples between the two batches, samples whose self-reported and genetic sex did not match and those with outlying values of autosomal heterozygosity rate. To identify any residual population outliers, we performed three rounds of within-EAS PCA, which sequentially removed samples with any of the top 10 PCs deviating from the mean by more than 6 standard deviations. Lastly, we filtered out variants with call rate <0.98 and HWE test p value < $1x10^{-10}$. Rather than directly removing related individuals in QC, we estimated a kinship (relatedness) matrix among TWB participants that can be filtered for custom GWAS analysis strategies (e.g., using mixed effects models to account for population structure and cryptic relatedness). Based on this homogeneous EAS sample, we computed PCA among approximately unrelated individuals across the two batches (with one from each pair of second-degree or closer relatives removed) and projected the rest of the individuals onto this PC space to obtain the final in-sample PCs for subsequent analyses. Based on self-reported ancestral origins, we classified TWB participants into three broad clusters (''Holo'', ''Hakka'', and ''Mainlanders'') and those with mixed origins for visualizing PCA (Table S4; Figure 2A). PCA without restricting to the unrelated individuals revealed a consistent structure.

TWB genotypes were then pre-phased with Eagle v2.4[26] and imputed using Minimac4[4] with 1KG phase3 EAS dataset as the reference panel. After imputation, we retained variants with an imputation quality INFO score >0.6 and MAF >0.5%.

Estimation of familial relatedness

We estimated kinship coefficients for all pairs of participants using KING[24] to measure the degree of familial relatedness in TWB. Different levels of relatedness were inferred using commonly suggested cutoffs: Duplicates/MZ twins: kinship >0.353; Parent–offspring: 0.177 < kinship <0.353 & Z0 < 0.05; Full siblings: 0.177 < kinship <0.353 & Z0 > 0.05; Second-degree relatives: 0.088 < kinship <0.177; Third-degree relatives: 0.044 < kinship <0.088; Fourth-degree relatives: kinship <0.044.

Relatedness-aware GWAS analysis in TWB

We used REGENIE, a two-step whole-genome regression that accounts for sample relatedness and population structure with type I error properly controlled,[15] to maximize power for GWAS in TWB. Specifically, at step 1, we computed a leave-one-chromosome out (LOCO) polygenic score combining estimates across blocks of SNPs using Ridge regression. At step 2, we then performed single-variant association tests adjusting for covariates and the LOCO predictors as an offset. Batch-specific SNP association tests were conducted (logistic regression model score test for binary traits and linear regression for quantitative) and meta-analyzed via the inverse-variance-weighted fixed-effect model based on overlapping SNPs.