

RESEARCH ARTICLE

Direct-to-consumer genetic testing in China and its role in GWAS discovery and replication

Kang Kang^{1,2}, Xue Sun², Lizhong Wang², Xiaotian Yao², Senwei Tang², Junjie Deng², Xiaoli Wu², WeGene Research Team^{1,2}, Can Yang³, Gang Chen^{1,2,4,*}

¹ Shenzhen WeGene Clinical Laboratory, Shenzhen 518118, China

² WeGene, Shenzhen Zaozhidao Technology Co., Ltd., Shenzhen 518042, China

³ Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

⁴ Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China

* Correspondence: cg@wegene.com

Received November 11, 2019; Revised March 27, 2020; Accepted April 30, 2020

Background: The direct-to-consumer genetic testing (DTC-GT) industry has exploded in recent years, initiated by market pioneers from the United States and quickly followed by companies from Europe and Asia. In addition to their primary objective of providing ancestry and health information to customers, DTC-GT services have emerged as a valuable data resource for large-scale population and genetics studies.

Methods: We assessed DTC-GT market leaders in the U.S. and China, user participation in research, and academic reports based on this information. We also investigated DTC-GT end-user value by tracing key updates of companies provided via health risk reports and evaluating their predictive power. We then assessed the replicability of several genome-wide association studies (GWAS) based on a Chinese DTC-GT biobank.

Results: As recent entrants to the market, Chinese DTC-GT service providers have published less academic research than their Western counterparts; however, a larger proportion of Chinese users consent to participate in research projects. Dramatic increases in user volume and resultant report updates led to reclassification of some users' polygenic risk levels, but within a reasonable scale and with increased predictive power. Replicability among GWAS using the Chinese DTC-GT biobank varied by studied trait, population background, and sample size.

Conclusions: We speculate that the rapid growth in DTC-GT services, particularly in non-Caucasian populations, will yield an important and much-needed resource for biobanking, large-scale genetic studies, clinical trials, and post-clinical applications.

Keywords: DTC-GT; biobank; Chinese population; polygenic risk; GWAS replication

Author summary: Direct-to-consumer genetic testing in China has exploded over the past five years. Chinese DTC-GC users are overwhelmingly willing to participate in research initiated by service providers. As most of these users are non-Caucasian, we evaluated the reliability of GWAS-derived polygenic disease reports using populations of predominantly European ancestry and found that prediction power increased alongside new GWAS loci integration. In assessing the outcomes of different GWAS, replicability varied among studies with different ethnic backgrounds and sample sizes. We speculate that Chinese DTC-GT databases represent valuable biobanks for genetic studies and clinical applications.

INTRODUCTION

In the framework of basic medical services, genetic tests are rarely offered by medical professionals other than genetic counselors and justice services; from a medical

perspective, such tests are focused on the prevention of severe Mendelian disorders or birth defects, or tumor genotyping or paternity tests. In recent decades, however, technological innovation has enabled large-scale genetic screening, including population genomics and genome-

wide association studies (GWAS), significantly expanding our knowledge as to the genetic underpinnings of common diseases and traits [1,2]. In light of identified associations between genomic variants and polygenic traits and the decreased cost of high-throughput genotyping, several direct-to-consumer genetic testing (DTC-GT) companies now offer genetic reports without requiring a medical professional intermediary.

In 2007, 23andMe, a company based in the United States, became the first DTC-GT service to provide personal genomics services, sending saliva sampling kits to users who returned them for analysis against a discrete subset of several thousand genetic markers. According to an estimation from MIT Technology Review [3], by January 2019, 26 million U.S. consumers, or roughly 8% of the country's population, had tested their DNA via one or more such DTC-GT services. DTC-GT user willingness to participate in research has substantially increased the biobank-scale genotype and phenotype database; starting in 2012, these companies became active participants in population genomics and health studies. From 2013 to 2018, Chinese DTC-GT services delivered around one million DNA tests to Chinese users [4]. Considering the population volume in East Asia alongside its non-European background, the Chinese DTC-GT market represents a potentially valuable contribution to the academic community with room for development.

DTC-GT services have also aroused concerns and fomented debates regarding risk assessment reliability, clinical utility, consumer perceptions, and ethical issues [5–8]. In some investigations, the reproducibility of DTC-GT assessments was evaluated among different DTC-GT companies, reporting high concordance rates for SNP data but inconsistent disease risk predictions [9–12]. The predictive power of some reports has also been evaluated by case studies, and Graves' disease, type 2 diabetes, lupus, Alzheimer's disease, restless leg syndrome, Crohn disease, age-related macular degeneration, and celiac disease were found to have the highest prediction power [7,13]. To the best of our knowledge, large-scale systematic evaluations of DTC-GT's genetic assessments, in particular from the service providers themselves, have yet to be published. Currently, updates to health reports intended to integrate new GWAS outcomes may result in reclassification of one's predicted risk levels long after DNA results were delivered, leading to user confusion and complicating clinical practice. Investigations therein suggest a reclassification rate ranged from 16.3% to 24.4% [14]. A reassessment of their predictive power, followed by risk level adjustment, has yet to be undertaken and thereby undercuts the reliability of different DTC-GT services.

Compounding this issue, the applicability of GWAS results and corresponding polygenic risk assessment to

non-Caucasian populations range from uncertain to downright misleading [15–17] since nearly all GWAS data comes from studies performed on predominantly European-ancestry (white) populations [18]. This situation presents risks and opportunities for DTC-GT services of non-European populations, such as the Chinese. While current knowledge may not be in alignment with local users, large cohort studies with Chinese populations are on the rise [19,20]. This may enable future GWAS outcomes from local research with relatively small sample sizes to be validated by biobank-scale datasets. Increasing the size and diversity of the global DTC-GT database could essentially improve GWAS replication efficacy and new association discovery.

This investigation analyzes user growth, rates of user participation in research, and academic outcomes of investigations conducted by major DTC-GT service providers in the U.S. and China. GWAS-derived DTC-GT reports were systematically evaluated for reliability through analysis of the distribution of polygenic risk scores (PRS) for multiple WeGene polygenic disease reports and tracing of risk level reclassifications over time. Results were directed towards assessing purported increases in predictive power from these companies. We also evaluated the reproducibility of several GWAS outcomes from the WeGene Biobank, including trans-ethnic or cross-ethnic studies, and of investigations with relatively small sample sizes.

RESULTS

Rapid growth of Chinese DTC-GT market

Founded in 2014 and 2015, respectively, WeGene and 23Mofang are leading the genetic testing upsurge in China, providing microarray-based, high-throughput genotyping products to customers similar to analyses provided by 23andMe and AncestryDNA in Western nations. Five-year user growth patterns for Chinese DTC-GT providers (Fig. 1B) are comparable to those for U.S. pioneers (Fig. 1A). The first wave of rapid growth in DTC-GT in China emerged in 2017 and 2018 and benefited from cost reductions in genotyping and a dynamic capital market in medical and health fields. In 2017, WeGene was the first to offer a whole-genome sequencing (WGS) service directly to consumers, and 23Mofang followed with a whole-Y chromosome sequencing and analysis service. In 2019, Hong-Kong S.A.R.-based CircleDNA entered the Chinese mainland market, offering a whole-exome sequencing (WES) service. Recent research estimates that around one million Chinese had availed themselves of a DTC-GT service as of 2018 [4].

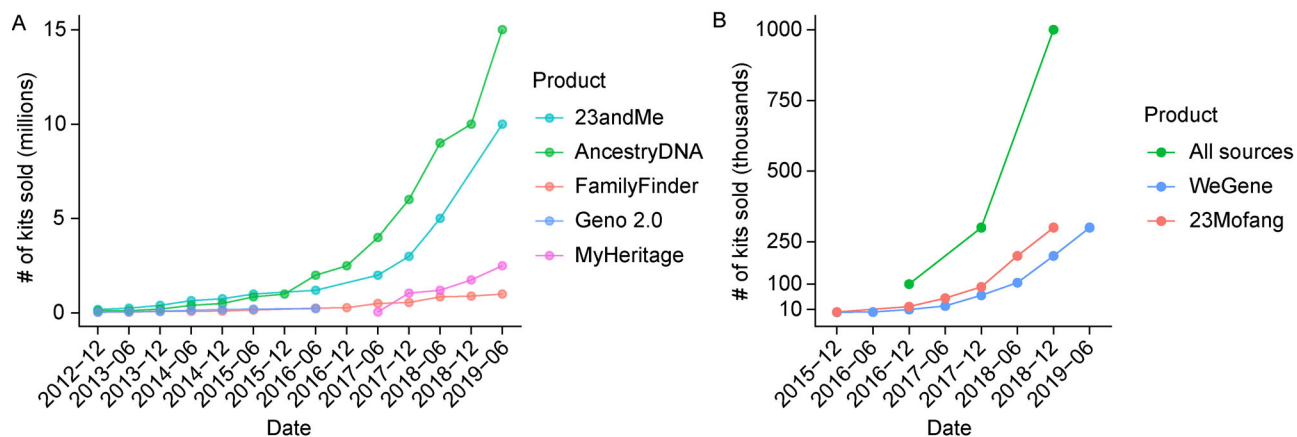


Figure 1. Customer growth for major DTC-GT service providers in the U.S. (A) and China (B). User information for the U.S. was acquired from the International Society of Genetic Genealogy Wiki [21]. Chinese data were acquired from a market research report [4] and news [22].

Published studies from DTC-GT service providers

DTC-GT databases have also emerged as a valuable resource for population genomics and genotype-phenotype association studies. 23andMe has been an active participant in the academic community since 2010 [23] (Fig. 2A), with research focused on human health, traits, and behaviors. 23andMe is also heavily involved in ethical, legal and social implications (ELSI) topics surrounding genetic testing. Chinese service providers followed suit beginning in 2017. Like their U.S. counterparts, Chinese service providers have published studies on biogeographic ancestry and population genomics [24–29], and on method development [28,30] (Fig. 2B); these assessments, however, do not rely on the ethically problematic large-scale collection of phenotypic data. To the best of our knowledge, the only published GWAS study of a phenotypic trait in the Chinese population was on photic sneeze reflex [31], with no health-related studies having been published as of 2019.

User participation in genetic research

By October 2019, 98.0% of WeGene profiles (Note: a single user account may include multiple genetic profiles) were accompanied by consent to use the genetic and phenotypic data for research purposes. This is much higher than the 80% participation rate for 23andMe users [23]. Among WeGene users, 97.5% of the company's profiles allowed access for genealogy matching. Of these, 56.6% have participated in at least one of 662 third-party trait reports (PocketDNA), contributed by WeGene customers via the open API platform. User activation and retention are high, with 86.0% of customers having

visited at least one WeGene online platform (website, mobile app, WeChat platform) over the past six months, and 77.1% over the past three months. Averages for 23andMe are approximately 60% over three months [32].

Basic information sharing is also high, with 99.6% of consented profiles reporting biological sex and 94.9% reporting date of birth (Fig. 3A). Current residence, ancestral home (defined as the birthplace of the participant's father's father which is recorded in the citizen residence registry in China), ethnic group, and surname were provided by 46.6% to 47.0% of profiles (Fig. 3A). In phenotype collection, the most provided traits are height and weight, with over 41,000 reports submitted (Fig. 3B). Among the 33 research projects on the WeGene open platform, 15 were self-conducted investigations, and 18 were collaborative studies. Among consented profiles, 48.4% have participated in at least one research project. Over 10,000 responses were collected from questionnaires about stress and mental traits, eyelid type, sleep pattern, color recognition, and blood type (Fig. 3B).

Reclassification of risk level for polygenic health reports

Three polygenic diseases, Alzheimer's disease (AD) [33–66], type 2 diabetes (T2D) [67–92], and schizophrenia (SP) [20,93–106], were selected to assess user risk level reclassification, as were growth in user numbers and report updates following synthesis of added GWAS outcomes. Normalized PRS and risk levels were calculated for the first 100, 300, 1,000, 3,000, and 10,000 users using the first version of WeGene from 2015, and then for each key update through October 2019 (Supplementary

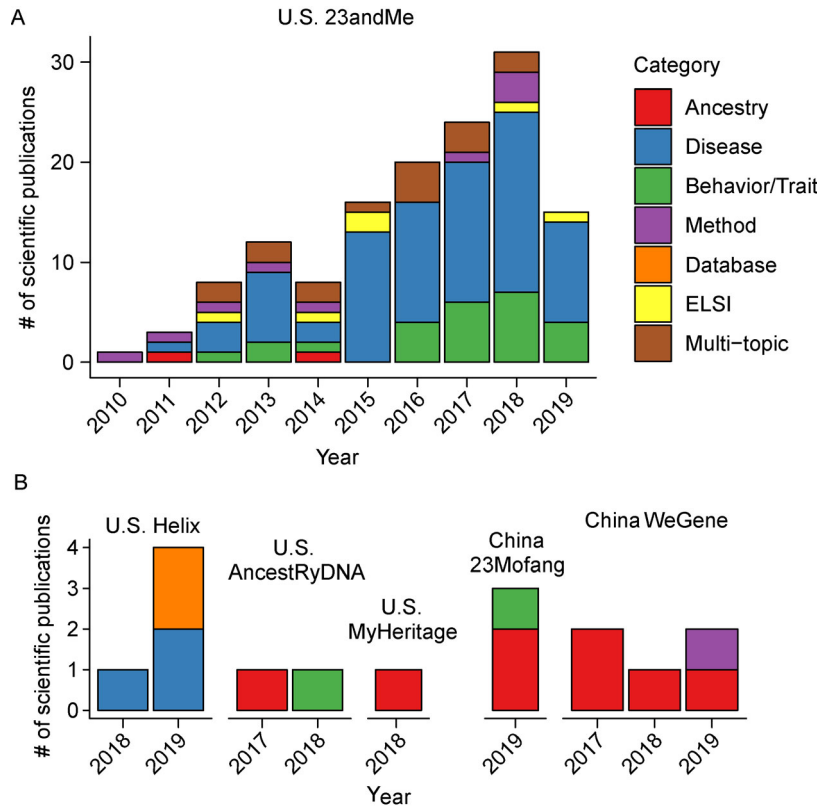


Figure 2. Scientific publications from DTC-GT service providers from the U.S. and China. (A) Publications by 23andMe. (B) Publications from other U.S. and Chinese companies.

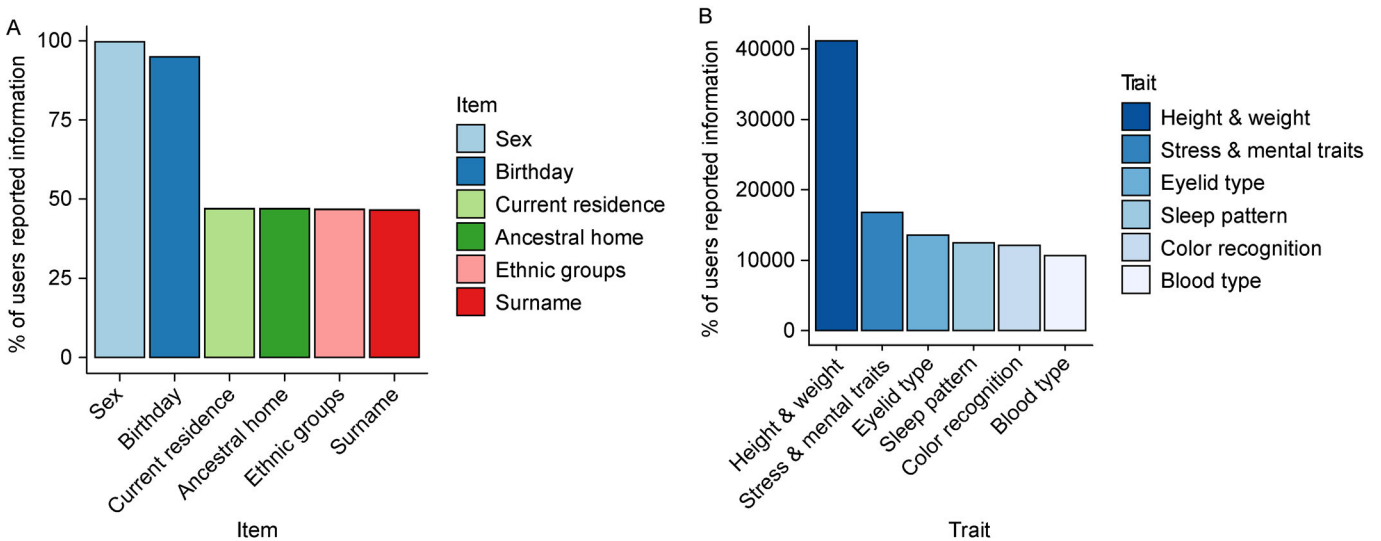


Figure 3. WeGene profiles sharing basic information and phenotypes. (A) Percentages of profiles with basic information, including date of birth, sex, current residence, ancestral home, ethnic group, and surname. (B) The most participated trait collections, led by height and weight, stress and mental traits, eyelid type, sleep pattern, color, and blood type.

Table S1, Data S1). Increased numbers of users did not change the overall PRS distribution per Pearson's correlation between user amount and PRS interquartile range (IQR) at $p = 0.66$ but did add smoothness. A two-tailed F -test for the variances of the PRS of two adjacent sampling points resulted in $p > 0.05$, with the exception of a single significant case wherein user numbers increased from 1,000 to 3,000 for the AD report, which came out to $p = 0.004$ (Fig. 4A). As the number of loci increased along with report updates, the PRS distribution broadened concomitantly (Pearson's coefficient: 0.74, $p = 0.0035$; two-tailed F -test: $p < 0.0001$).

Observed risk level reclassification after report updates occurred on a reasonable scale (Fig. 5). Between two adjacent versions of reports, $76.2\% \pm 12.0\%$ users' risk levels remain unchanged, $22.1\% \pm 10.7\%$ were reclassified to an adjacent level, and extreme alterations (high to low or medium-low, or low to high or medium-high) were only observed in $0.010\% \pm 0.019\%$ of cases.

As most GWAS data were derived from European populations, cross-ethnic replication of GWAS loci and applicability of the GWAS-based PRS models remain

unclear. We evaluated the predictive power of polygenic disease reports by analyzing user feedback on AD family history. In total, 10,435 individuals reported a family history of AD, and 1,763 individuals (16.9%) reported at least one positive AD case in a parent or grandparent. The predictive power of correlation testing between disease risk levels and family history is expected to improve following increasing numbers of users and integration of new GWAS outcomes, as shown by a trend towards increased odds ratios and narrower confidence intervals (CIs) (Fig. 6).

Replicability of GWAS results

Seven association studies were selected for validation in WeGene Biobank, including single nucleotide polymorphism (SNPs) originally identified in European populations and studies with relatively small sample sizes.

Studies in Caucasian populations indicated that apolipoprotein E (*ApoE*) genotypes are associated with late-onset AD [107,108]. Among the 10,435 WeGene profiles

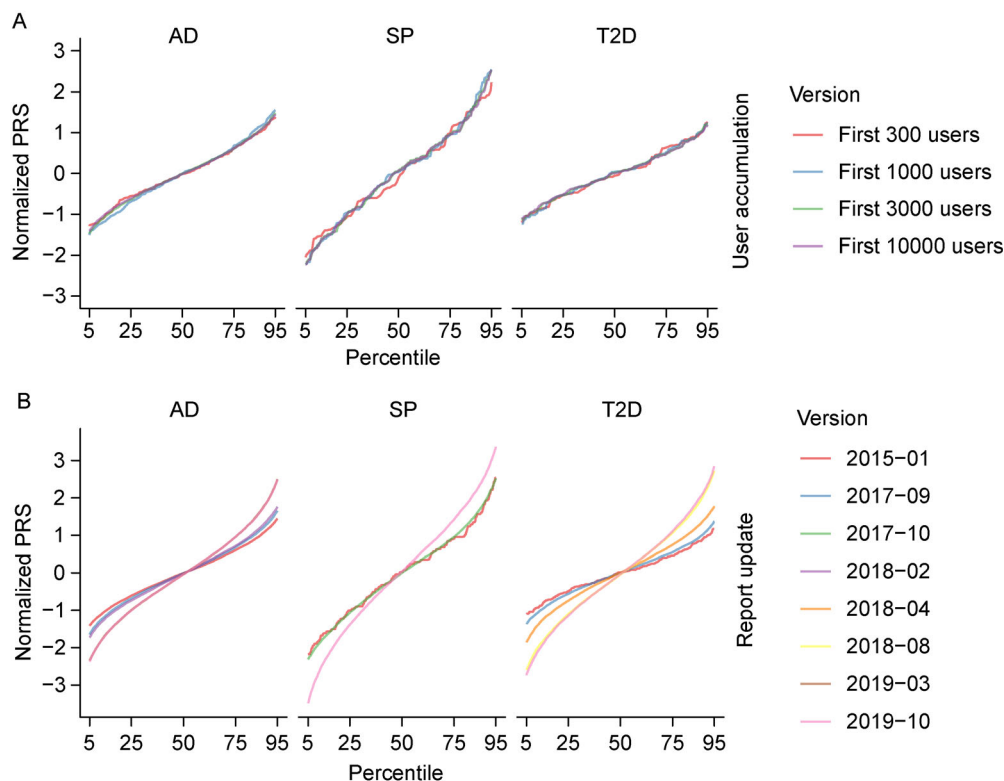


Figure 4. PRS distribution (5th to 95th percentile) changes along with increased numbers of users and report updates for three polygenic diseases (AD, SP, and T2D). (A) PRS distribution modified after significant increase in amount of user data from the first version of reports. (B) PRS distribution changes by report update. AD report included 53 loci in the first version and increased to 84 loci by Oct. 2019. The T2D report increased from 18 to 213 loci. The SP report increased from 18 to 300 loci.

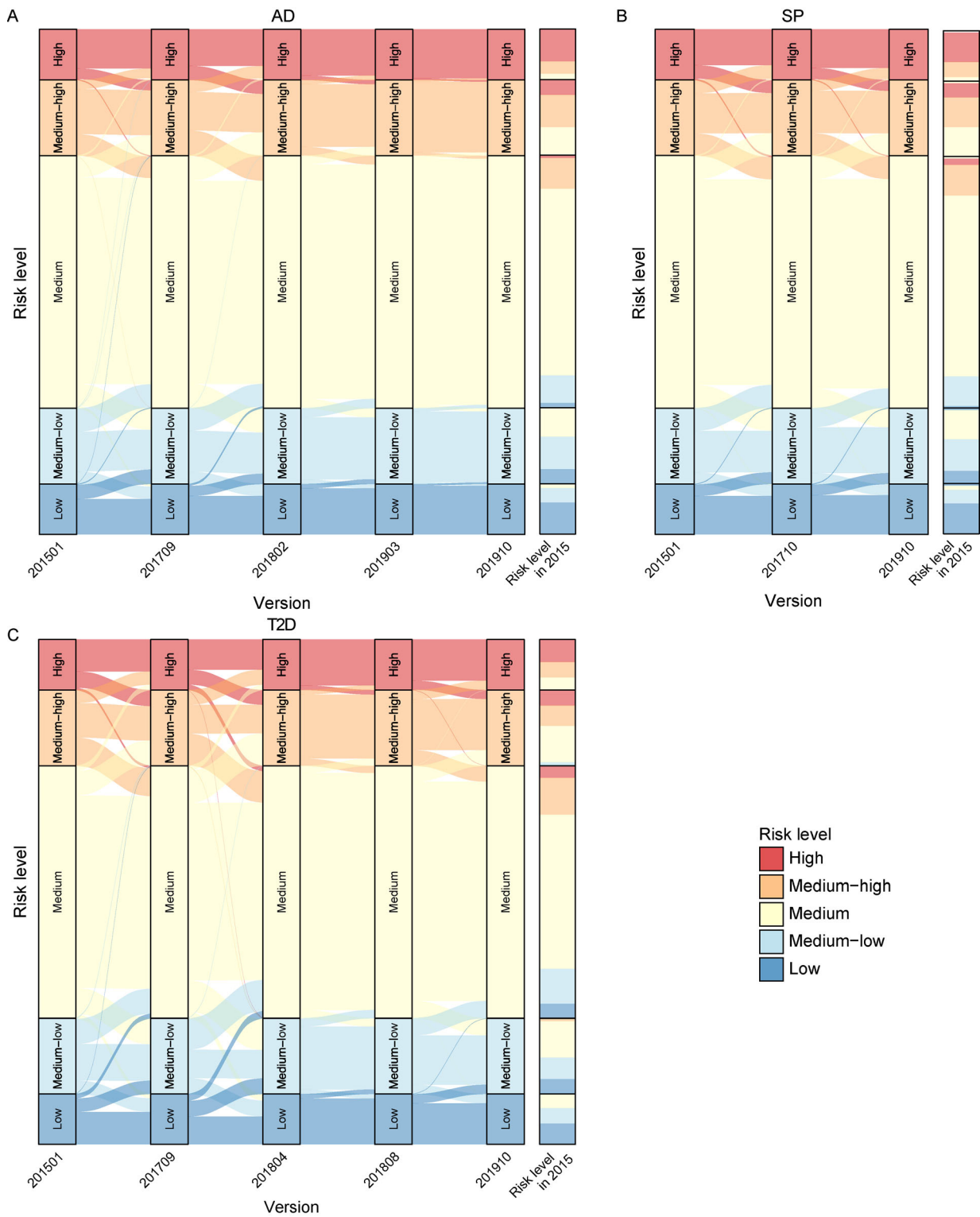


Figure 5. Users' risk level reclassification for three polygenic disease reports: AD (A), SP (B), and T2D (C). Comparisons conducted between temporally adjacent reports. Connections are colored by risk level in the report before update. Original risk levels in the report from the first version (2015-01) are shown in the right-hand color bar.

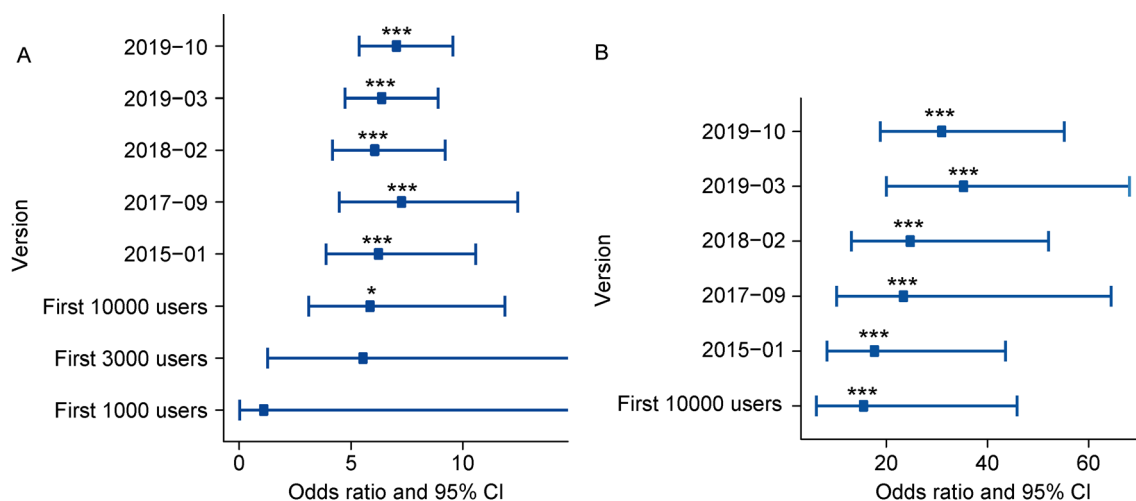


Figure 6. Odds ratio and 95% confidence interval (CI) testing of different AD risk levels with self-reported AD family history along with increased user numbers and report updates. (A) Odds ratios for 2×2 table of users with relatively high risk (High, Medium-high) and relatively low risk (Low, Medium-low). (B) Results for 2×2 table for High and Low risk users. Fisher test significance level: *** $p < 0.0001$; ** $p < 0.01$; * $p < 0.05$.

that participated in AD report feedback, both the risk allele type $\epsilon 4$ (OR: 1.43, 95% CI: 1.27 to 1.61, two-tailed Fisher's, $p = 2.6 \times 10^{-9}$) and protective allele type $\epsilon 2$ (OR: 0.76, 95% CI: 0.64 to 0.98, two-tailed Fisher's, $p = 7.9 \times 10^{-4}$) were significantly associated with AD family history. We also tried to replicate per-locus associations of the 84 loci used in the WeGene AD report (Supplementary Table S2) and found 35 loci with low minor allele frequency (MAF) (< 0.05) among WeGene users; among the remaining 49 loci, 12 could be replicated with normal significance (one-tailed Fisher's, $p < 0.05$), led by rs7412 ($p = 0.002$), an *ApoE*-determining SNP.

Similarly, rs9939609, a SNP in the fat mass and obesity-associated protein (FTO) gene reported to be associated with body mass index (BMI) [109], was also replicated by a correlation test between body weight categories (overweight: BMI > 28 , lean: BMI < 18.5) and genotypes (Chi-square, $p = 0.011$), and the rs9939609-AA carriers presented with significantly higher BMIs than individuals with rs9939609-GG (23.3 vs. 22.4, one-tailed t -test, $p = 3.4 \times 10^{-11}$). In another case, among the top 20 loci (ranked by OR) associated with male pattern baldness (Supplementary Table S3), six out of the eight loci with MAF $\geq 5\%$ were significantly associated with self-reported hair loss levels (Chi-square for all genotypes, one-tailed Fisher's for high-risk genotypes, $p < 0.05$ for both) in WeGene Biobank (Supplementary Table S2).

Conversely, SNPs reportedly associated with cilantro dislike and soap taste in European populations [110], were not replicated in WeGene users: rs72921001 was

non-significant (Chi-square test, $p = 0.060$), and no rs78503206 polymorphism was found in the database. Similarly, none of the four SNPs associated with handedness in a recent study in UK Biobank participants [111] were found in a Chinese sample size of 7,644 (Chi-square, $p > 0.05$).

For certain small sample GWAS on East Asians, the WeGene Biobank could be valuable as a dataset for GWAS discovery and validation. In a study of 96 Han Chinese individuals [112], three loci were identified as significantly or inconsistently significantly associated with eyelid traits, although none of these presented among the 13,715 participants reporting eyelid type in the WeGene Biobank (Chi-square, $p > 0.05$). Conversely, a study on 2,980 Han Chinese did not turn up any significant markers for petaloid toenails [113] despite the fact that it is a signature trait among Han Chinese. Similar genotyping methods in over 8,000 individuals reporting fifth toenail types in the WeGene Biobank, however, uncovered 32 SNPs that met the genome-wide significance threshold ($p < 5 \times 10^{-8}$) (unpublished data).

DISCUSSION

The years 2017 and 2018 saw rapid growth in the DTC-GT market in the U.S. and China. Although less than 0.1% of the Chinese population has performed a self-assisted DNA test to date, as home to the world's largest population, China has the potential to be a powerful force in the emerging DTC-GT market.

User composition limits research diversity and value of current data

As is to be expected with novel technology, acceptance and popularity of DTC-GT in China was originally heavily skewed towards young people. WeGene users have an average age of 31, and approximately 50% of their clients are aged 26 to 38. Over 80% of these users live in first-tier metropolitans. Biased age and residence compositions have limited clinical research opportunities and applications of Chinese DTC-GT biobanks and cohort recruitment and GWAS for less-common diseases, with a dataset of diagnosed users insufficient to support these measures. We encountered data limitations in our evaluation of the PRS model and replication of GWAS loci for late-onset diseases, such as AD, and were unable to directly link positive AD cases to particular genetic profiles. We instead had to rely on user family history as an alternative, somewhat limiting the replication, extension, and reliability of our findings. As such, similar to early 23andMe research endeavors, publications from the Chinese DTC-GT service providers remain focused on population genomics and tools [24–30].

Heavy user activity and research initiatives promote future outcomes

The skewed age composition also confers a benefit in terms of the openness and willingness to engage in data sharing among current users, including feedback reporting, third-party report participation, and phenotype collection. DTC-GT companies use comprehensively connected web-based platforms that include an official website, mobile app, and social media profiles on WeChat API and other official media. This likely contributes to robust user activation and retention and thereby promotes phenotype collection for research purposes. These advantages suggest promising academic contributions from Chinese DTC-GT companies in the future. Chinese companies interested in replicating 23andMe's model could yield Chinese GWAS on human health and other traits within three to five years. As user numbers increase, phenotype collection shifts in user composition will render Chinese DTC-GT-derived biobanks more valuable, particularly for studies on disease.

Report reliability and optimization

The core mission of DTC-GT service providers is to provide increasingly accurate and understandable genetic reports to customers. Current DTC-GT reports for polygenic diseases and traits are predominantly generated by frontloading GWAS outcomes in the absence of systematical examination and validation [6–8]. Our

examination of the predictive power of multiple polygenic disease reports using normalized PRS distribution indicates a trend towards increasing prediction accuracy alongside concomitant user growth and the synthesis of new GWAS results. In the meantime, the scale of risk level reclassification was shown to be within normal parameters and is unlikely to cause distress in customers lacking professional knowledge of genetics.

It is important to note, however, that a large number of reported SNP associations from other GWAS could not be replicated in the WeGene user database, likely due to the application of European-based GWAS against a non-Caucasian population. We also found that the overall reproducibility of loci included in such reports remains unclear, confounding verifiable evaluations. We thereby propose significant improvements could be made to phenotype and disease risk predictive models, and necessary follow-up tasks should include an evaluation of the reproducibility of different GWAS, an implementation of new GWAS studies to identify new loci, new SNP ranking and weight, predictive model selection and adjustment, precise covariant selection (such as biological sex, age, and family history) in a complex predictive model, and ethnicity-specific modeling.

Opportunities for Chinese DTC-GT biobanks

Biobanks are an important data resource for human genetic research projects, particularly medical cohort studies and GWAS discovery and replication. The UK Biobank, with more than 500,000 genotyped participants, is the largest biobank that is publicly accessible [114], and has been mined for genetic research across the globe. Among UK Biobank samples, around 150,000 individuals were genotyped with microarrays from Affymetrix that search for 600,000 to 800,000 SNPs and indels. UK Biobank investigations have produced 944 scientific papers [114]. Apart from government-financed biobanks, commercial biobanks like 23andMe have also become valuable resources for large-scale studies. All 23andMe users were genotyped with high-throughput arrays from Illumina and Affymetrix, covering from 500,000 to 900,000 SNPs and indels across versions, similar to the arrays used by WeGene. 23andMe datasets have been mined for 130 scientific publications [23], and the biobanking of 23andMe also possess commercial value via data purchase and trading with the pharmaceutical industry [115]. Currently, whole-genome genotyping (WGG) is used by most biobanks to balance costs, sample size, scientific interest, and cross-biobank compatibility of biobanking.

Our trans-ethnic GWAS replication analyses recapitulated previous studies demonstrating that population background is a crucial factor influencing the reproducibility of

GWAS outcomes [15–17]. A biobank with WGG data from a majority Chinese population is in high demand for health-related studies and commercial purposes such as drug development. The most famous open-to-public human biobank in East Asia is BioBank Japan; no UK Biobank-like dataset for Chinese or even East Asian populations generally currently exists. In the absence of an official biobank, and light of rapid demand growth for commercial and research datasets alongside robust user study participation, Chinese DTC-GT-based biobank shows strong potential in both academic and industrial contexts.

MATERIALS AND METHODS

Research participants

Participants in the genome-wide association study (GWAS) validation and health risk level analyses were drawn from consenting WeGene customers from Shenzhen Zaozhidao Technology Co. Ltd., a direct-to-consumer genetic testing service provider. User statistics, genotypes, and phenotypes were collected in October 2019.

Ethical approval

Informed consent for online research was obtained from all individual participants included in the study. The study was approved by the Ethical Committee of Shenzhen WeGene Clinical Laboratory. The study was conducted in accordance with the human and ethical research principles of The Ministry of Science and Technology of the People's Republic of China (Regulation of the Administration of Human Genetic Resources, July 1, 2019).

DNA sampling and genotyping assay

Saliva samples for DNA extraction were collected and stored with an Oragene DNA Sample Collection Kit (OG-250 or OG-510, DNA Genotek, Canada). DNA isolation and purification were performed with the Magnetic Saliva Fast DNA kit DP703-73A (Tiangen, China). Samples were genotyped at WeGene Clinical Laboratory on one of two custom arrays: Affymetrix WeGene V1 Array (596,744 SNPs) by Affymetrix GeneTitan MC Instrument, and Illumina WeGene V2 Array (742,762 SNPs) by Illumina iScan System.

Quality control of genotype data

Quality control (QC) was performed with PLINK V1.9 [116]. Individuals and SNPs with an overall genotype call

rate lower than 98.5% were excluded. In polygenic risk score (PRS) distribution and health risk level reclassification analyses, individuals with AD, T2D, or SP, and SNPs with a genotype call rate lower than 80.0% were excluded.

Phenotype and family disease history

Self-reported phenotypes and family histories were provided by participants via web-based questionnaires. Customers who did not fill out these questionnaires were eliminated from the dataset used for statistical analysis of the target disease or phenotype.

Body mass index (BMI) Individuals' BMIs were calculated from self-reported height and weight using the following formula:

$$BMI = weight(kg) / height(m)^2$$

Only participants aged from 18 to 65 and with BMI values from the 5th to 95th percentile were used in statistics.

Hair loss Respondents were asked if they were bald, and example images for different levels were given for selecting one of four responses: "no," "slight," "medium," or "severe"; these were used to classify the respondent's phenotype. They were then asked if their father and mother were bald, with the same four options plus a fifth for "not sure" for each. Respondents were then asked to provide dates of birth for themselves and their parents. "Slight" and "medium" were quantified as "hair loss." "Severe" was quantified as "bald" in GWAS replication analysis.

Family disease history Respondents were asked whether they have any family members diagnosed with a specific disease. The family members include the respondent, the respondent's father, the respondent's mother, the respondent's grandfathers, and grandmothers. A participant was marked as positive for disease family history in any of these family members was reported as a diagnosed case, otherwise, the respondent was marked as negative for the disease history. The disease family histories of T2D, AD, and SP were used in this study.

Cilantro preference Respondents were asked whether they did or did not like cilantro, and whether or not they thought it had a pleasant or soap-like taste or aroma; "not sure" was also provided as an option. These answers were used to classify phenotypes for GWAS replication.

Handedness Participants were asked to provide their biological sex and whether or not they were a twin with options for identical, fraternal same-sex, and fraternal opposite-sex. They were then asked about handedness with options for "right-handed," "left-handed," "ambidextrous," and "not sure"; these were used to classify GWAS phenotype. Subsequently, they were asked for

their preferred hand in multiple behaviors, including writing, drawing, throwing, using scissors, tooth-brushing, using a knife, using a spoon, using chopsticks, using a hand broom, and unscrewing caps with the following five options provided for each: “right hand only,” “right hand mostly,” “no preference,” “left hand mostly,” and “left hand only.” Finally, they were asked about each parent’s handedness with options for “right-handed,” “left-handed,” “ambidextrous,” and “not sure.”

Eyelids Participants were asked to classify single- or double-fold eyelids for each eye, with an additional option of “difficult to classify” for both; these responses were used for GWAS phenotype classification. The participants were also asked to classify the eyelid types for the right and left eye of each parent with the added option of “not sure.”

Petaloid toenail Participants were asked whether their fifth pedal digit (“little toe”) had a petaloid toenail for each foot. Phenotypes were classified as Petaloid_E (petaloid toenail on one foot) and Petaloid_D (petaloid toenail on both feet) in accordance with established standards [113]. GWAS for Petaloid_E and Petaloid_D were performed separately.

Odds ratio (OR) normalization, PRS and risk level

All participants were included in OR and PRS calculations before the participant volume reached 10,000. Participants for version 2015-01 were acquired from users up to the first key update (September or October 2017). The impact of report updates was evaluated by randomly selecting 10,000 more participants plus all those who provided a corresponding family disease history and received genetic testing before the first report update. Risk level reclassification was assessed using 10,000 randomly selected subjects genotyped with the WeGene V2 Array at all time points.

Allele ORs were converted to genotype ORs before PRS calculations. If a biallelic OR was not specified in the original literature, a single risk/protective allele OR was assigned to the heterozygous genotype, and both risk/protective alleles were assigned to a homozygous genotype with the squared allele’s OR. Each SNP’s OR distribution was \log_2 -transformed and adjusted to be zero-centered in the population using the following formula:

$$adjOR_{j,a} = \log_2 OR_{j,a} - \frac{\sum_{j,n} \log_2 OR_{j,n}}{n}$$

where $adjOR_{j,a}$ is the adjusted OR for genotype a of locus j ; $OR_{j,n}$ is the OR of locus j for individual n ; and $OR_{j,a}$ is the original OR of the genotype a of locus j .

For a single health risk report, the PRS incorporating all risk loci for individual n was:

$$PRS_n = \sum_j adjOR_{j,n}$$

Participant PRS values were classified into five risk level categories by percentile: Low = $PRS < 10^{th}$, Medium-low = $10^{th} \leq PRS < 25^{th}$; Medium = $25^{th} \leq PRS \leq 75^{th}$; Medium-high = $75^{th} < PRS \leq 90^{th}$; High = $PRS > 90^{th}$. Participant PRS-based health risk levels were subject to change according to increased numbers of users, OR adjustments, and health report update following incorporation of new GWAS-identified SNPs.

Genome-wide association study

Initial genome-wide association analyses on ordinal or binary phenotype were performed with PLINK 1.9 [116] using multiple linear regression models of additive allelic effects with sex and an appropriate number of genetic principal components (PCs) as covariates. Detailed methods will be released when the corresponding GWAS published.

Statistics and visualization

Statistics were conducted in *Python* and *R* with packages including *scipy* and *numpy*. Data visualization was performed with *R* and corresponding packages, including *ggplot2*, *RColorBrewer*, *ggalluvial*, and *qqman*. Fisher’s exact test (2×2 table) or Chi-square tests (3×2 genotype table) were performed to assess independence. A *t*-test was performed for mean value comparisons between parametric statistics. Pearson’s correlation was used to evaluate correlations between parametric data. *P*-value correction for multiple testing was performed with a Bonferroni adjustment. The significance threshold was set to $p < 0.05$ and false discovery rate (FDR) < 0.05 . During GWAS discovery, the genome-wide significance threshold was set to $p < 5 \times 10^{-8}$ for SNPs. In GWAS replication, $p < 0.05$ was used as the threshold for statistical significance.

Data availability

In light of our commitment to customer privacy and privacy regulations from the Administration of Human Genetic Resource of China, we will not be publishing user health reports or detailed genotype or phenotype distributions. For questions about the analyses in this research or academic collaboration opportunities with WeGene, please contact the WeGene Research Team by email (research@wegene.com).

ABBREVIATIONS

AD	Alzheimer's disease
API	Application programming interface
ApoE	Apolipoprotein E
BMI	Body mass index
CI	Confidence interval
DTC	Direct-to-consumer
ELSI	Ethical, Legal and Social Implications
EMR	Electronic medical record
FDR	False discovery rate
FTO	Fat mass and obesity-associated protein
GT	Genetic testing
GWAS	Genome-wide association study
IDR	Interdecile range
IQR	Interquartile range
ISOGG	International Society of Genetic Genealogy
MAF	Minor allele frequency
OR	Odds ratio
PC	Principal component
PRS	Polygenic risk score
QC	Quality control
T2D	Type II diabetes
SNP	Single nucleotide polymorphism
SP	Schizophrenia
WES	Whole-exome sequencing
WGG	Whole-genome genotyping
WGS	Whole-genome sequencing

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at <https://doi.org/10.1007/s40484-020-0209-2>.

ACKNOWLEDGEMENTS

We thank all WeGene users who consented to share their genotype and phenotype data for research purposes.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Kang Kang, Xue Sun, Lizhong Wang, Xiaotian Yao, Senwei Tang, Junjie Deng, Xiaoli Wu, and Gang Chen work for WeGene (Shenzhen Zaozhidao Technology Co. Ltd. or Shenzhen WeGene Clinical Laboratory).

All procedures performed in studies were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

REFERENCES

1. Gibson, G. (2010) Hints of hidden heritability in GWAS. *Nat.*

- Genet.*, 42, 558–560
2. Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. and Cox, N. J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, 6, e1000888
3. More than 26 million people have taken an at-home ancestry test. <https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/>. Accessed: 12 October, 2019
4. 2018 China dtc-gt market research report. <https://www.iyiou.com/intelligence/report605.html>. Accessed: 12 October, 2019
5. Covolo, L., Rubinelli, S., Ceretti, E. and Gelatti, U. (2015) Internet-based direct-to-consumer genetic testing: A systematic review. *J. Med. Internet Res.*, 17, e279
6. Frueh, F. W., Greeley, H. T., Green, R. C., Hogarth, S. and Siegel, S. (2011) The future of direct-to-consumer clinical genetic tests. *Nat. Rev. Genet.*, 12, 511–515
7. Kalf, R. R., Mihaescu, R., Kundu, S., de Knijff, P., Green, R. C. and Janssens, A. C. (2014) Variations in predicted risks in personal genome testing for common complex diseases. *Genet. Med.*, 16, 85–91
8. Kolor, K., Duquette, D., Zlot, A., Foland, J., Anderson, B., Giles, R., Wrathall, J. and Khoury, M. J. (2012) Public awareness and use of direct-to-consumer personal genomic tests from four state population-based surveys, and implications for clinical and public health practice. *Genet. Med.*, 14, 860–867
9. Adams, S. D., Evans, J. P. and Aylsworth, A. S. (2013) Direct-to-consumer genomic testing offers little clinical utility but appears to cause minimal harm. *N C Med. J.*, 74, 494–498
10. Buitendijk, G. H., Amin, N., Hofman, A., van Duijn, C. M., Vingerling, J. R. and Klaver, C. C. (2014) Direct-to-consumer personal genome testing for age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.*, 55, 6167–6174
11. Imai, K., Kricka, L. J. and Fortina, P. (2011) Concordance study of 3 direct-to-consumer genetic-testing services. *Clin. Chem.*, 57, 518–521
12. Kido, T., Kawashima, M., Nishino, S., Swan, M., Kamatani, N. and Butte, A. J. (2013) Systematic evaluation of personal genome services for Japanese individuals. *J. Hum. Genet.*, 58, 734–741
13. Bloss, C. S., Topol, E. J. and Schork, N. J. (2012) Association of direct-to-consumer genome-wide disease risk estimates and self-reported disease. *Genet. Epidemiol.*, 36, 66–70
14. Krier, J., Barfield, R., Green, R. C. and Kraft, P. (2016) Reclassification of genetic-based risk predictions as GWAS data accumulate. *Genome Med.*, 8, 20
15. Lewis, C. M. and Vassos, E. (2017) Prospects for using risk scores in polygenic medicine. *Genome Med.*, 9, 96
16. Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M. and Daly, M. J. (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 51, 584–591
17. Reisberg, S., Iljasenko, T., Läll, K., Fischer, K. and Vilo, J. (2017) Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations.

- PLoS One, 12, e0179238
18. Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) The missing diversity in human genetic studies. *Cell*, 177, 26–31
 19. Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., Fang, L., Li, Z., Lin, L., Liu, R., *et al.* (2018) Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell*, 175, 347–359
 20. Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., *et al.* (2017) Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.*, 49, 1576–1583
 21. Autosomal DNA testing comparison chart. https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart. Accessed: 12 October, 2019
 22. Dtc-gt business sense: Data “fosse” behind free service. <https://new.qq.com/omn/20190830/20190830A00TIP00.html>. Accessed: 12 October, 2019
 23. 23andme’s new research overview page. <https://blog.23andme.com/23andme-research/23andmes-new-research-overview-page/>. Accessed: 12 October, 2019
 24. Chen, P., Wu, J., Luo, L., Gao, H., Wang, M., Zou, X., Li, Y., Chen, G., Luo, H., Yu, L., *et al.* (2019) Population genetic analysis of modern and ancient DNA variations yields new insights into the formation, genetic structure, and phylogenetic relationship of northern han chinese. *Front. Genet.*, 10, 1045
 25. Li, Y. C., Ye, W. J., Jiang, C. G., Zeng, Z., Tian, J. Y., Yang, L. Q., Liu, K. J. and Kong, Q. P. (2019) River valleys shaped the maternal genetic landscape of han chinese. *Mol. Biol. Evol.*, 36, 1643–1652
 26. Huang, X., Zhou, Q., Bin, X., Lai, S., Lin, C., Hu, R., Xiao, J., Luo, D., Li, Y., Wei, L. H., *et al.* (2018) The genetic assimilation in language borrowing inferred from Jing People. *Am. J. Phys. Anthropol.*, 166, 638–648
 27. Yao, H. B., Tang, S., Yao, X., Yeh, H. Y., Zhang, W., Xie, Z., Du, Q., Ma, L., Wei, S., Gong, X., *et al.* (2017) The genetic admixture in Tibetan-Yi Corridor. *Am. J. Phys. Anthropol.*, 164, 522–532
 28. Yao, X., Tang, S., Bian, B., Wu, X., Chen, G. and Wang, C. C. (2017) Improved phylogenetic resolution for Y-chromosome Haplogroup O2a1c-002611. *Sci. Rep.*, 7, 1146
 29. Zeng, Z., Tian, J., Jiang, C., Ye, W., Liu, K. and Li, Y. (2019) Inferring the history of surname Ye based on Y chromosome high-resolution genotyping and sequencing data. *J. Hum. Genet.*, 64, 703–709
 30. Zhao, J., Ming, J., Hu, X., Chen, G., Liu, J. and Yang, C. (2020) Bayesian weighted Mendelian randomization for causal inference based on summary statistics. *Bioinformatics*, btz749
 31. Wang, M., Sun, X., Shi, Y., Song, X. and Mi, H. (2019) A genome-wide association study on photic sneeze reflex in the Chinese population. *Sci. Rep.*, 9, 4993
 32. 23andme wants to solve the patient recruitment problem. <https://www.clinicalleader.com/doc/andme-wants-to-solve-the-patient-recruitment-problem-0001>. Accessed: 12 October, 2019
 33. Clark, R. F., Hutton, M., Fuldner, M., Froelich, S., Karran, E., Talbot, C., Crook, R., Lendon, C., Prihar, G., He, C., *et al.* (1995) The structure of the presenilin 1 (S182) gene and identification of six novel mutations in early onset AD families. *Nat. Genet.*, 11, 219–222
 34. Rogaev, E. I., Sherrington, R., Rogaeva, E. A., Levesque, G., Ikeda, M., Liang, Y., Chi, H., Lin, C., Holman, K., Tsuda, T., *et al.* (1995) Familial Alzheimer’s disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer’s disease type 3 gene. *Nature*, 376, 775–778
 35. Sherrington, R., Rogaev, E. I., Liang, Y., Rogaeva, E. A., Levesque, G., Ikeda, M., Chi, H., Lin, C., Li, G., Holman, K., *et al.* (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer’s disease. *Nature*, 375, 754–760
 36. Kwok, J. B., Taddei, K., Hallupp, M., Fisher, C., Brooks, W. S., Broe, G. A., Hardy, J., Fulham, M. J., Nicholson, G. A., Stell, R., *et al.* (1997) Two novel (M233T and R278T) presenilin-1 mutations in early-onset Alzheimer’s disease pedigrees and preliminary evidence for association of presenilin-1 mutations with a novel phenotype. *Neuroreport*, 8, 1537–1542
 37. Bruni, A. C. (1998) Cloning of a gene bearing missense mutations in early onset familial Alzheimer’s disease: a Calabrian study. *Funct. Neurol.*, 13, 257–261
 38. Harvey, R. J., Ellison, D., Hardy, J., Hutton, M., Roques, P. K., Collinge, J., Fox, N. C. and Rossor, M. N. (1998) Chromosome 14 familial Alzheimer’s disease: the clinical and neuropathological characteristics of a family with a leucineright-arrowserine (L250S) substitution at codon 250 of the presenilin 1 gene. *J. Neurol. Neurosurg. Psychiatry*, 64, 44–49
 39. Poorkaj, P., Sharma, V., Anderson, L., Nemens, E., Alonso, M. E., Orr, H., White, J., Heston, L., Bird, T. D. and Schellenberg, G. D. (1998) Missense mutations in the chromosome 14 familial Alzheimer’s disease presenilin 1 gene. *Hum. Mutat.*, 11, 216–221
 40. Lewis, P. A., Perez-Tur, J., Golde, T. E. and Hardy, J. (2000) The presenilin 1 C92S mutation increases abeta 42 production. *Biochem. Biophys. Res. Commun.*, 277, 261–263
 41. Tedde, A., Forleo, P., Nacmias, B., Piccini, C., Bracco, L., Piacentini, S. and Sorbi, S. (2000) A presenilin-1 mutation (Leu392Pro) in a familial AD kindred with psychiatric symptoms at onset. *Neurology*, 55, 1590–1591
 42. Avella, A. B., Teruel, B. M., Rodriguez, J. L., Viera, N. G., Martinez, I. B., e, S., M, J., Duijn, C., Baute, L. H. and P, H. (2002) A novel presenilin 1 mutation (L174 M) in a large Cuban family with early onset Alzheimer disease. *Neurogenetics*, 4, 97–104
 43. Queralt, R., Ezquerra, M., Lleó, A., Castellví, M., Gelpí, J., Ferrer, I., Acarín, N., Pasarín, L., Blesa, R. and Oliva, R. (2002) A novel mutation (V89L) in the presenilin 1 gene in a family with early onset Alzheimer’s disease and marked behavioural disturbances. *J. Neurol. Neurosurg. Psychiatry*, 72, 266–269
 44. Miklossy, J., Taddei, K., Suva, D., Verdile, G., Fonte, J., Fisher, C., Gnjec, A., Ghika, J., Suard, F., Mehta, P. D., *et al.* (2003) Two novel presenilin-1 mutations (Y256S and Q222H) are associated with early-onset Alzheimer’s disease. *Neurobiol. Aging*, 24, 655–662

45. Snider, B. J., Norton, J., Coats, M. A., Chakraverty, S., Hou, C. E., Jervis, R., Lendon, C. L., Goate, A. M., McKeel, Jr., D. W. and Morris, J. C. (2005) Novel presenilin 1 mutation (S170F) causing Alzheimer disease with Lewy bodies in the third decade of life. *Arch. Neurol.*, 62, 1821–1830
46. Lerner, A. J. and Doran, M. (2006) Clinical phenotypic heterogeneity of Alzheimer's disease associated with mutations of the presenilin-1 gene. *J. Neurol.*, 253, 139–158
47. Kauwe, J. S., Jacquart, S., Chakraverty, S., Wang, J., Mayo, K., Fagan, A. M., Holtzman, D. M., Morris, J. C. and Goate, A. M. (2007) Extreme cerebrospinal fluid amyloid beta levels identify family with late-onset Alzheimer's disease presenilin 1 mutation. *Ann. Neurol.*, 61, 446–453
48. Meng, Y., Lee, J. H., Cheng, R., St George-Hyslop, P., Mayeux, R. and Farrer, L. A. (2007) Association between SORL1 and Alzheimer's disease in a genome-wide study. *Neuroreport*, 18, 1761–1764
49. Aidaralieva, N. J., Kamino, K., Kimura, R., Yamamoto, M., Morihara, T., Kazui, H., Hashimoto, R., Tanaka, T., Kudo, T., Kida, T., *et al.* (2008) Dynamins 2 gene is a novel susceptibility gene for late-onset Alzheimer disease in non-APOE-epsilon4 carriers. *J. Hum. Genet.*, 53, 296–302
50. Piscopo, P., Marcon, G., Piras, M. R., Crestini, A., Campeggi, L. M., Deiana, E., Cherchi, R., Tanda, F., Deplano, A., Vanacore, N., *et al.* (2008) A novel PSEN2 mutation associated with a peculiar phenotype. *Neurology*, 70, 1549–1554
51. Rademakers, R., Eriksen, J. L., Baker, M., Robinson, T., Ahmed, Z., Lincoln, S. J., Finch, N., Rutherford, N. J., Crook, R. J., Josephs, K. A., *et al.* (2008) Common variation in the miR-659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia. *Hum. Mol. Genet.*, 17, 3631–3642
52. Schjeide, B. M., Hooli, B., Parkinson, M., Hogan, M. F., DiVito, J., Mullin, K., Blacker, D., Tanzi, R. E. and Bertram, L. (2009) GAB2 as an Alzheimer disease susceptibility gene: follow-up of genomewide association results. *Arch. Neurol.*, 66, 250–254
53. Bennet, A. M., Reynolds, C. A., Gatz, M., Blennow, K., Pedersen, N. L. and Prince, J. A. (2010) Pleiotropy in the presence of allelic heterogeneity: alternative genetic models for the influence of APOE on serum LDL, CSF amyloid- β 42, and dementia. *J. Alzheimers Dis.*, 22, 129–134
54. Sanders, A. E., Wang, C., Katz, M., Derby, C. A., Barzilai, N., Ozelius, L. and Lipton, R. B. (2010) Association of a functional polymorphism in the cholesteryl ester transfer protein (CETP) gene with memory decline and incidence of dementia. *JAMA*, 303, 150–158
55. Xu, X., Wang, Y., Wang, L., Liao, Q., Chang, L., Xu, L., Huang, Y., Ye, H., Xu, L., Chen, C., *et al.* (2013) Meta-analyses of 8 polymorphisms associated with the risk of the Alzheimer's disease. *PLoS One*, 8, e73129
56. Cruchaga, C., Karch, C. M., Jin, S. C., Benitez, B. A., Cai, Y., Guerreiro, R., Harari, O., Norton, J., Budde, J., Bertelsen, S., *et al.* (2014) Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*, 505, 550–554
57. Floudas, C. S., Um, N., Kamboh, M. I., Barmada, M. M. and Visweswaran, S. (2014) Identifying genetic interactions associated with late-onset Alzheimer's disease. *BioData Min.*, 7, 35
58. Medway, C. and Morgan, K. (2014) Review: The genetics of Alzheimer's disease; putting flesh on the bones. *Neuropathol. Appl. Neurobiol.*, 40, 97–105
59. Gao, Y., Tan, M. S., Wang, H. F., Zhang, W., Wang, Z. X., Jiang, T., Yu, J. T. and Tan, L. (2016) ZCWPW1 is associated with late-onset Alzheimer's disease in Han Chinese: a replication study and meta-analyses. *Oncotarget*, 7, 20305–20311
60. Wang, H. Z., Bi, R., Hu, Q. X., Xiang, Q., Zhang, C., Zhang, D. F., Zhang, W., Ma, X., Guo, W., Deng, W., *et al.* (2016) Validating GWAS-identified risk loci for Alzheimer's disease in Han Chinese populations. *Mol. Neurobiol.*, 53, 379–390
61. Sims, R., van der Lee, S. J., Naj, A. C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle, B. W., Boland, A., Raybould, R., Bis, J. C., *et al.* (2017) Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet.*, 49, 1373–1384
62. Zhang, X. Y., Wang, H. F., Tan, M. S., Wan, Y., Kong, L. L., Zheng, Z. J., Tan, C. C., Zhang, W., Wang, Z. X., Tan, L., *et al.* (2017) Association of disc1 polymorphisms with late-onset Alzheimer's disease in northern Han Chinese. *Mol. Neurobiol.*, 54, 2922–2927
63. Zhu, X. C., Cao, L., Tan, M. S., Jiang, T., Wang, H. F., Lu, H., Tan, C. C., Zhang, W., Tan, L. and Yu, J. T. (2017) Association of parkinson's disease GWAS-linked loci with Alzheimer's disease in Han Chinese. *Mol. Neurobiol.*, 54, 308–318
64. Zhou, X., Chen, Y., Mok, K. Y., Zhao, Q., Chen, K., Chen, Y., Hardy, J., Li, Y., Fu, A. K. Y., Guo, Q., *et al.* (2018) Identification of genetic risk factors in the Chinese population implicates a role of immune system in Alzheimer's disease pathogenesis. *Proc. Natl. Acad. Sci. USA*, 115, 1697–1706
65. Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., *et al.* (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.*, 51, 404–413
66. Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., van der Lee, S. J., Amlie-Wolf, A., *et al.* (2019) Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.*, 51, 414–430
67. Rubin, D., Helwig, U., Pfeuffer, M., Schreiber, S., Boeing, H., Fisher, E., Pfeiffer, A., Freitag-Wolf, S., Foelsch, U. R., Doering, F., *et al.* (2006) A common functional exon polymorphism in the microsomal triglyceride transfer protein gene is associated with type 2 diabetes, impaired glucose metabolism and insulin levels. *J. Hum. Genet.*, 51, 567–574
68. Takeuchi, F., Serizawa, M., Yamamoto, K., Fujisawa, T., Nakashima, E., Ohnaka, K., Ikegami, H., Sugiyama, T., Katsuya, T., Miyagishi, M., *et al.* (2009) Confirmation of multiple risk loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. *Diabetes*, 58, 1690–1699

69. Bouatia-Naji, N., Bonnefond, A., Cavalcanti-Proença, C., Sparsø, T., Holmkvist, J., Marchand, M., Delplanque, J., Lobbens, S., Rocheleau, G., Durand, E., *et al.* (2009) A variant near *MTNR1B* is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nat. Genet.*, 41, 89–94
70. The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, 661–678
71. Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., *et al.* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, 40, 638–645
72. Horikawa, Y., Miyake, K., Yasuda, K., Enya, M., Hirota, Y., Yamagata, K., Hinokio, Y., Oka, Y., Iwasaki, N., Iwamoto, Y., *et al.* (2008) Replication of genome-wide association studies of type 2 diabetes susceptibility in Japan. *J. Clin. Endocrinol. Metab.*, 93, 3136–3141
73. Prudente, S., Morini, E., Marselli, L., Baratta, R., Copetti, M., Mendonca, C., Andreozzi, F., Chandalia, M., Pellegrini, F., Bailetti, D., *et al.* (2013) Joint effect of insulin signaling genes on insulin secretion and glucose homeostasis. *J. Clin. Endocrinol. Metab.*, 98, E1143–E1147
74. Tabara, Y., Osawa, H., Kawamoto, R., Onuma, H., Shimizu, I., Miki, T., Kohara, K. and Makino, H. (2009) Replication study of candidate genes associated with type 2 diabetes based on genome-wide screening. *Diabetes*, 58, 493–498
75. Prudente, S., Scarpelli, D., Chandalia, M., Zhang, Y. Y., Morini, E., Del Guerra, S., Peticone, F., Li, R., Powers, C., Andreozzi, F., *et al.* (2009) The *TRIB3* Q84R polymorphism and risk of early-onset type 2 diabetes. *J. Clin. Endocrinol. Metab.*, 94, 190–196
76. Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., Hirota, Y., Mori, H., Jonsson, A., Sato, Y., *et al.* (2008) Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. *Nat. Genet.*, 40, 1092–1097
77. Friedrich, B., Weyrich, P., Stancáková, A., Wang, J., Kuusisto, J., Laakso, M., Sesti, G., Succurro, E., Smith, U., Hansen, T., *et al.* (2008) Variance of the *SGK1* gene is associated with insulin secretion in different European populations: results from the TUEF, EUGENE2, and METSIM studies. *PLoS One*, 3, e3506
78. Keramati, A. R., Fathzadeh, M., Go, G. W., Singh, R., Choi, M., Faramarzi, S., Mane, S., Kasaei, M., Sarajzadeh-Fard, K., Hwa, J., *et al.* (2014) A form of the metabolic syndrome associated with mutations in *DYRK1B*. *N. Engl. J. Med.*, 370, 1909–1919
79. Tsai, F. J., Yang, C. F., Chen, C. C., Chuang, L. M., Lu, C. H., Chang, C. T., Wang, T. Y., Chen, R. H., Shiu, C. F., Liu, Y. M., *et al.* (2010) A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet.*, 6, e1000847
80. Murad, A. S., Smith, G. D., Lewis, S. J., Cox, A., Donovan, J. L., Neal, D. E., Hamdy, F. C. and Martin, R. M. (2010) A polymorphism in the glucokinase gene that raises plasma fasting glucose, rs1799884, is associated with diabetes mellitus and prostate cancer: findings from a population-based, case-control study (the ProtecT study). *Int J Mol Epidemiol Genet*, 1, 175–183
81. Zheng, J. S., Arnett, D. K., Parnell, L. D., Smith, C. E., Li, D., Borecki, I. B., Tucker, K. L., Ordovás, J. M. and Lai, C. Q. (2013) Modulation by dietary fat and carbohydrate of *IRS1* association with type 2 diabetes traits in two populations of different ancestries. *Diabetes Care*, 36, 2621–2627
82. Uma Jyothi, K., Jayaraj, M., Subburaj, K. S., Prasad, K. J., Kumuda, I., Lakshmi, V. and Reddy, B. M. (2013) Association of *TCF7L2* gene polymorphisms with T2DM in the population of Hyderabad, India. *PLoS One*, 8, e60212
83. Pei, Q., Huang, Q., Yang, G. P., Zhao, Y. C., Yin, J. Y., Song, M., Zheng, Y., Mo, Z. H., Zhou, H. H. and Liu, Z. Q. (2013) *PPAR-γ2* and *PTPRD* gene polymorphisms influence type 2 diabetes patients' response to pioglitazone in China. *Acta Pharmacol. Sin.*, 34, 255–261
84. Cho, Y. S., Chen, C.-H., Hu, C., Long, J., Hee Ong, R. T., Sim, X., Takeuchi, F., Wu, Y., Go, M. J., Yamauchi, T., *et al.* (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in East Asians. *Nat. Genet.*, 44, 67–72
85. Sokolova, E. A., Bondar, I. A., Shabelnikova, O. Y., Pyankova, O. V. and Filipenko, M. L. (2015) Replication of *kcj11* (p.E23k) and *abcc8* (p.S1369a) association in Russian diabetes mellitus 2 type cohort and meta-analysis. *PLoS One*, 10, e0124662
86. Williams, A. L., Jacobs, S. B., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., García-Ortiz, H., Gómez-Vázquez, M. J., Burt, N. P., Aguilar-Salinas, C. A., *et al.* (2014) Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature*, 506, 97–101
87. Sim, X., Ong, R. T., Suo, C., Tay, W. T., Liu, J., Ng, D. P., Boehnke, M., Chia, K. S., Wong, T. Y., Seielstad, M., *et al.* (2011) Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genet.*, 7, e1001363
88. Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., Yengo, L., Lloyd-Jones, L. R., Sidorenko, J., Wu, Y., *et al.* (2018) Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.*, 9, 2941
89. Morris, A. P. (2018) Progress in defining the genetic contribution to type 2 diabetes susceptibility. *Curr. Opin. Genet. Dev.*, 50, 41–51
90. Imamura, M., Takahashi, A., Yamauchi, T., Hara, K., Yasuda, K., Grarup, N., Zhao, W., Wang, X., Huerta-Chagoya, A., Hu, C., *et al.* (2016) Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nat. Commun.*, 7, 10531
91. Suzuki, K., Akiyama, M., Ishigaki, K., Kanai, M., Hosoe, J., Shojima, N., Hozawa, A., Kadota, A., Kuriki, K., Naito, M., *et al.* (2019) Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat. Genet.*, 51, 379–386
92. Cheng, L., Zhang, D., Zhou, L., Zhao, J. and Chen, B. (2015) Association between *slc30a8* rs13266634 polymorphism and type

- 2 diabetes risk: A meta-analysis. *Med. Sci. Monit.*, 21, 2178–2189
93. Chen, X., Wang, X., Chen, Q., Williamson, V., van den Oord, E., Maher, B. S., O'Neill, F. A., Walsh, D. and Kendler, K. S. (2008) MEGF10 association with schizophrenia. *Biol. Psychiatry*, 63, 441–448
94. Lohoff, F. W., Weller, A. E., Bloch, P. J., Buono, R. J., Doyle, G. A., Ferraro, T. N. and Berrettini, W. H. (2008) Association between polymorphisms in the vesicular monoamine transporter 1 gene (*VMAT1/SLC18A1*) on chromosome 8p and schizophrenia. *Neuropsychobiology*, 57, 55–60
95. Monakhov, M., Golimbet, V., Abramova, L., Kaleda, V. and Karpov, V. (2008) Association study of three polymorphisms in the dopamine D2 receptor gene and schizophrenia in the Russian population. *Schizophr. Res.*, 100, 302–307
96. Pal, P., Mihanović, M., Molnar, S., Xi, H., Sun, G., Guha, S., Jeran, N., Tomljenović, A., Malnar, A., Missoni, S., *et al.* (2009) Association of tagging single nucleotide polymorphisms on 8 candidate genes in dopaminergic pathway with schizophrenia in Croatian population. *Croat. Med. J.*, 50, 361–369
97. Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O. P., Mors, O., Mortensen, P. B., *et al.* (2009) Common variants conferring risk of schizophrenia. *Nature*, 460, 744–747
98. Cichon, S., Mühleisen, T. W., Degenhardt, F. A., Mattheisen, M., Miró, X., Strohmaier, J., Steffens, M., Meesters, C., Herms, S., Weingarten, M. (2011) Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *Am. J. Hum. Genet.*, 88, 372–381
99. Tsutsumi, A., Glatt, S. J., Kanazawa, T., Kawashige, S., Uenishi, H., Hokyo, A., Kaneko, T., Moritani, M., Kikuyama, H., Koh, J., *et al.* (2011) The genetic validation of heterogeneity in schizophrenia. *Behav. Brain Funct.*, 7, 43
100. Xiao, B., Li, W., Zhang, H., Lv, L., Song, X., Yang, Y., Li, W., Yang, G., Jiang, C., Zhao, J., *et al.* (2011) To the editor: Association of znf804a polymorphisms with schizophrenia and antipsychotic drug efficacy in a Chinese Han population. *Psychiatry Res.* 190, 379–381
101. Yue, W., Yang, Y., Zhang, Y., Lu, T., Hu, X., Wang, L., Ruan, Y., Lv, L. and Zhang, D. (2011) A case-control association study of NRXN1 polymorphisms with schizophrenia in Chinese Han population. *Behav. Brain Funct.*, 7, 7
102. Fineberg, A. M. and Ellman, L. M. (2013) Inflammatory cytokines and neurological and neurocognitive alterations in the course of schizophrenia. *Biol. Psychiatry*, 73, 951–966
103. Yan, P., Qiao, X., Wu, H., Yin, F., Zhang, J., Ji, Y., Wei, S. and Lai, J. (2016) An association study between genetic polymorphisms in functional regions of five genes and the risk of schizophrenia. *J. Mol. Neurosci.*, 59, 366–375
104. Yu, H., Yan, H., Li, J., Li, Z., Zhang, X., Ma, Y., Mei, L., Liu, C., Cai, L., Wang, Q., *et al.* (2017) Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in Han Chinese population. *Mol. Psychiatry*, 22, 954–960
105. Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., *et al.* (2018) Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.*, 50, 381–389
106. Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511, 421–427
107. Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, Jr, P. C., Rimmler, J. B., Locke, P. A., Conneally, P. M., Schmechel, K. E., *et al.* (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat. Genet.*, 7, 180–184
108. Saunders, A. M., Strittmatter, W. J., Schmechel, D., George-Hyslop, P. H., Pericak-Vance, M. A., Joo, S. H., Rosi, B. L., Gusella, J. F., Crapper-MacLachlan, D. R., Alberts, M. J., *et al.* (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, 43, 1467–1472
109. Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., *et al.* (2007) A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316, 889–894
110. Eriksson, N. W., Wu, S., Do, C. B., Kiefer, A. K., Tung, J. Y., Mountain, J. L., Hinds, D. A. and Francke, U. (2012) A genetic variant near olfactory receptor genes influences cilantro preference. *Flavour (Lond.)*, 1, 22
111. Wiberg, A., Ng, M., Al Omran, Y., Alfaro-Almagro, F., McCarthy, P., Marchini, J., Bennett, D. L., Smith, S., Douaud, G. and Furniss, D. (2019) Handedness, language areas and neuropsychiatric diseases: insights from brain imaging and genetics. *Brain*, 142, 2938–2947
112. Jin, B., Zhu, J., Wang, H., Chen, D., Su, Q., Wang, L., Liang, W. B. and Zhang, L. (2015) A primary investigation on SNPs associated with eyelid traits of Chinese Han adults. *Forensic Sci. International. Genet. Suppl. Ser.*, 5, e669–e670
113. Zhang, M., Wu, S., Zhang, J., Yang, Y., Tan, J., Guan, H., Liu, Y., Tang, K., Krutmann, J., Xu, S., *et al.* (2016) Large-scale genome-wide scans do not support petaloid toenail as a Mendelian trait. *J. Genet. Genomics*, 43, 702–704
114. About UK biobank. <https://www.ukbiobank.ac.uk/about-biobank-uk/>. Accessed: 12 October, 2019
115. 23andme sells access to biobank to more than 13 drug companies. <https://privacyinternational.org/examples-abuse/1854/23andme-sells-access-biobank-more-13-drug-companies>. Accessed: 12 October, 2019
116. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81, 559–575