Title: Searching across-cohort relatives via encrypted genotype regression

Authors: Qixin Zhang[1], Tianzi Liu[2], Xinxin Guo[3], Jianxin Zhen[3,4], Kan Bao[5], Meng-yuan Yang[6], Saber Khederzadeh[6], Fang Zhou[7], Xiaotong Han[8], Qiwen Zheng[2], Peilin Jia[2], Xiaohu Ding[8], Mingguang He[8,9,10], Xin Zou[11], Hongxin Zhang[11], Ji He[12], Xiaofeng Zhu[13], Yangyun Zou[5], Sijia Lu[5], Daru Lu[14,15], Hongyan Chen[7], Changqing Zeng[2,16], Fan Liu[2,17], Hou-Feng Zheng[6], Siyang Liu[3], Hai-Ming Xu[1] and Guo-Bo Chen[18,19,*]

Affiliations:

1. Institute of Bioinformatics, Zhejiang University, Hangzhou, Zhejiang 310058, China
2. CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
3. School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, Guangdong 510006, China
4. Central Laboratory, Shenzhen Baoan Women's and Children's Hospital, Shenzhen, Guangdong 518102, China
5. Department of Clinical Research, Yikon Genomics Company, Ltd., Suzhou, Jiangsu 215000, China
6. Diseases & Population (DaP) Geninfo Lab, School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China
7. State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China
8. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, Guangdong 510060, China
9. Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, Melbourne, Victoria, Australia
10. Ophthalmology, Department of Surgery, University of Melbourne, Melbourne, Victoria, Australia
11. Stake Key Laboratory of CAD & GC, Zhejiang University, Hangzhou, Zhejiang 310058, China
12. Department of Neurology, Peking University Third Hospital, Beijing 100191, China
13. Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA
14. State Key Laboratory of Genetic Engineering and MOE Engineering Research Center of Gene Technology, School of Life Sciences and Zhongshan Hospital, Fudan University, Shanghai 200438, China

15. NHC Key Laboratory of Birth Defects and Reproductive Health, Chongqing Population and Family Planning Science and Technology Research Institute, Chongqing 401120, China
16. College of Life Science, University of Chinese Academy of Sciences, Beijing 100049, China
17. Department of Forensic Sciences, College of Criminal Justice, Naif Arab University of Security Sciences, Riyadh, 11452, Kingdom of Saudi Arabia
18. Clinical Research Institute, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang 310014, China
19. Key Laboratory of Endocrine Gland Diseases of Zhejiang Province, Hangzhou, Zhejiang 310014, China

*Correspondence:**

Guo-Bo Chen: chenguobo@gmail.com

ABSTRACT

Identifying relatives across cohorts makes one of the basic routines for genomic data. As conventional such practice often requires explicit genomic data sharing, it is easily hampered by privacy or ethical constraints. In this study, using our proposed scheme for genomic encryption we developed ***encG-reg***, a regression approach that is able to detect relatives of various degrees based on encrypted genomic data. The encryption properties of ***encG-reg*** is built on random matrix theory, which masks the original genotypic matrix but still provides controllable precision to that of direct individual-level genotype data. After having found tractable eighth-order moments for encrypted genotype, we established connection between the dimension of a random matrix and the required precision of a study. ***encG-reg*** consequently led to balanced i) false positive and false negative rates and ii) the computational cost and the degree of relatives to be searched. We validated ***encG-reg*** in 485,158 UKBiobank multi-ethnical samples, and the resolution of ***encG-reg*** was comparable with the conventional method such as KING. In a more complex application, we launched a fine-devised multi-center collaboration across 6 research institutes in China, covering 11 cohorts of 64,091 GWAS samples. In both examples, ***encG-reg*** robustly identified and validated relatives existing across the cohorts even under various ethnical background and different genotypic qualities.

Introduction

Genomic datasets have been reaching millions of individuals and often encapsulated in well protected cohorts, in which relatives more than often, given increasing genotyped individuals, spread across cohorts and can be identified once the genomic data are compared[1]. Finding relatives often has clear scientific reasons, such as controlling false positive rates in genome-wide association study (GWAS) or reducing overfitting in polygenic risk score prediction[2–4]. Social benefits are recently promoted for available individual genomic data in such as relativeness testing and forensic genetic genealogy[5]. However, direct-to-consumer (DTC) genetic testing activities along with third-party services pose new privacy and ethnic concerns[6]; law enforcement authorities have exploited some of consumer genomic databases to identify suspects by finding their distant genetic relatives, which has brought privacy concerns to the attention of the general public[7,8]. For regulating forensic genetic genealogy, laws, policies and privacy-protection techniques such as homomorphic encryption are in parallel development[9–11].

The above progress, nevertheless, often requires individual-level data to be shared which may often be beyond the permitted range of data sharing because of privacy concerns. Directly processing raw genotype in genetic tests may be vulnerable to attacks[12]. We developed a novel mitigation strategy called “encrypted genotype regression”, hereby encG-reg, which does not require direct genotype data but is able to identify relatedness with highly controllable precision of balanced Type I and Type II error rates. As only encrypted genotype data is exchanged in performing encG-reg, a pair or a group of collaborated cohorts are able to minimize their concerns of privacy breach. In this study we explore the properties of encG-reg in theory, simulations, and 485,158 UK Biobank (UKB) samples of various ethnicity. In a collaboration that includes 6 genomic centers from north to south China (Beijing, Suzhou, Shanghai, Hangzhou, Guangzhou, and Shenzhen) totaling 64,091 genetically diverse samples genotyped based on different platforms, intriguing relatedness were identified between cohorts by encG-reg. Often, the logistic complex of a human genetic study is exacerbated by the number of cohorts involved[13]; the presented study, however, establishes an expert-driven constitution and technical innovations for driving multi-cohort collaborations that is now in demand for genomic studies.

Materials and Methods

Overview of GRM

A pair of collaborators, who concern the privacy of their genomic data, seek identification of relatedness between their cohorts using GWAS data. Using whole genome-wide markers, inter-cohort relatedness for pairs of individuals can be inferred from genetic relationship matrix (GRM), which requires matrix multiplication between two genotype matrices, say $\mathbf{X_1}$ and $\mathbf{X_2}$; where $\mathbf{X_1}$ is a matrix of n_1 individuals (rows) and m markers (columns), so is $\mathbf{X_2}$. We define $\mathbf{G}_{12} = \frac{1}{m}\mathbf{X}_1\mathbf{X}_2^{\mathrm{T}} = \{g_{ij}\}_{n_1 \times n_2}$ as the real inter-cohort GRM. Here, the genotype matrices are standardized by SNP allelic frequencies to have zero mean and a unit variance. Under the assumption of multivariate normal distribution, the expectation and variance of g_{ij}, using Isserlis's theorem are[14]

$$E(g_{ij}) = \theta_r \text{ and } var(g_{ij}) = \frac{1+\theta_r^2}{m} \qquad \textbf{(Eq 1)}$$

respectively, where r is the degree of relatives and θ_r is relatedness score, which has $E(\theta_r) = \left(\frac{1}{2}\right)^r$, say $E(\theta_r) =$ 0.5, 0.25, and 0.125 for the first, second, and third degree of relatives, respectively.

Encrypted genotype (encG) and encG regression (encG-reg)

To extend GRM into its encrypted form, one insight from approximate matrix decomposition is that we can find a $\mathbf{Q}_{m\times m}$ matrix, which satisfies $\mathbf{X}_1\mathbf{Q}\mathbf{X}_2^{\mathrm{T}} \approx \mathbf{X}_1\mathbf{X}_2^{\mathrm{T}}$[15]. $\mathbf{Q}$ matrix can be decomposed as $\mathbf{Q} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$, where $\mathbf{S}$ is an $m \times k$ matrix and its elements are dependently sampled from a normal distribution, $N(0, \sigma^2)$. We show that $E(\mathbf{S}\mathbf{S}^{\mathrm{T}}) = \mathbf{I}$ and $E\left(\mathbf{X}_1\mathbf{S}\mathbf{S}^{\mathrm{T}}\mathbf{X}_2^{\mathrm{T}}\right) = \mathbf{X}_1\mathbf{X}_2^{T}$, with the choice of $\sigma^2 = \frac{1}{k}$. When two collaborators provide $\widehat{\mathbf{X}}_1 = \mathbf{X}_1\mathbf{S}$ and $\widehat{\mathbf{X}}_2 = \mathbf{X}_2\mathbf{S}$, it leads to $E\left(\widehat{\mathbf{X}}_1\widehat{\mathbf{X}}_2^{\mathrm{T}}\right) = \mathbf{X}_1\mathbf{X}_2^{\mathrm{T}}$ the approximated precision of which relies on the sampling variance. In this study, we attack the question that if relatives are involved between $\mathbf{X}_1$ and $\mathbf{X}_2$, how precisely k should be to control sampling variance that is able to identify relatives of certain degree. The products of matrix multiplication present an ideal one-way encryption technique in private genetic data sharing, and this is what we call $\widehat{\mathbf{X}}_1$ "encrypted genotype", hereby encG. As discussed, it is computationally impossible to recover $\mathbf{X}$ from $\widehat{\mathbf{X}}$ without the knowledge of $\mathbf{S}$[16].

Based on encG, it is now trustworthy to construct encrypted GRM (encGRM) inter-cohort. We define $\widehat{\mathbf{G}}_{12} = \frac{\mathbf{1}}{k}(\mathbf{X_1}\mathbf{S})\left(\mathbf{S}^{\mathrm{T}}\mathbf{X}_2^{\mathrm{T}}\right) = \{\hat{g}_{ij}\}_{n_1\times n_2}$, and elements of the random matrix $\mathbf{S}$ are sampled from a normal distribution $N\left(0, \frac{1}{m}\right)$ to provide a good transformation of expectation from $E\left(\frac{\mathbf{X}_1\mathbf{X}_2^{T}}{m}\right)$ to $E(\frac{(\mathbf{X_1}\mathbf{S})(\mathbf{S}^{\mathrm{T}}\mathbf{X}_2^{\mathrm{T}})}{k})$. In terms of the matrix element $\hat{g}_{ij}$ by eight-order moments approximation, its expectation and variance are $E\left(\hat{g}_{ij}\right) = \theta_r$ and $var\left(\hat{g}_{ij}\right) \simeq$

$\frac{1+\theta_r^2}{k}+\frac{1+\theta_r^2}{m}$, in which $\frac{1+\theta_r^2}{k}$ is crept in $var(\hat{g}_{ij})$ compared with that of $var(g_{ij})$. As SNPs are often in linkage disequilibrium (LD), we introduce the effective number of markers (m_e), which is a parameter engaged in various genetic analyses[17]. The variance of g_{ij} and $\hat{g}_{ij}$ turns to $\frac{1+\theta_r^2}{m_e}$ and $\frac{1+\theta_r^2}{k}+\frac{1+\theta_r^2}{m_e}$, respectively.

Another interpretation on encGRM is from the perspective of regression. The regression is also based on encG and we call it encG regression, which regresses one individual's encrypted genotype against another. For a pair of individuals, say individual i and individual j, the slope b_{ij} of a simple regression model $\hat{\mathbf{x}}_j = b_{ij}\hat{\mathbf{x}}_i + \mathbf{e}$, also known as regression coefficient, indicates the identical by descent (IBD) score between these two individuals. Here $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ are vectors of encrypted genotypes for two individuals. $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ are scaled to zero mean and unit variance. The expectation and the sampling variance of $\hat{b}_{ij} = \frac{cov(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j)}{var(\hat{\mathbf{x}}_i)}$ can be approximated as

$$E(\hat{b}_{ij}) = \theta_r \text{ and } var(\hat{b}_{ij}) \simeq \frac{1-\theta_r^2}{k}+\frac{1-\theta_r^2}{m_e} \qquad \textbf{(Eq 2)}$$

Compared to encGRM, encG-reg generates smaller sampling variance and thus conceals improved power in identifying relatives from unrelated pairs.

A minimal number of m_e and k

For a pair of individuals I) whose relatedness is estimated by GRM and follows the distribution of $N(\theta_r, \frac{1+\theta_r^2}{m_e})$, we ask how to identify them from unrelated pairs with a distribution of $N(0, \frac{1}{m_e})$; II) whose relatedness is estimated by encG-reg and follows the distribution of $N(\theta_r, \frac{1-\theta_r^2}{k}+\frac{1-\theta_r^2}{m_e})$, we ask how to differentiate them from unrelated pairs with a distribution of $N(0, \frac{1}{k}+\frac{1}{m_e})$. This question is analogous to the conventional pattern recognition, which can be solved under the power calculation in the statistical test framework for null verse alternative hypotheses. We consequently need to determine two key parameters. I) the effective number of markers, m_e, a population statistic that sets the resolution of GRM itself in detecting relatives. II) the column number of the random matrix, k, an iteration dimension that sets the precision of encG-reg. To determine m_e and k, upon Type I error rate (α, false positive rate as aforementioned) and Type II error rate (β, false negative rate), m_e should satisfy below

$$m_{e|\alpha,\beta,\theta_r} > \left(\frac{z_{1-\beta}\sqrt{1+\theta_r^2}+z_{1-\alpha}}{\theta_r}\right)^2 \qquad \textbf{(Eq 3)}$$

Similar to m_e, the minimal number of k is also responsible for a certain Type I and Type II error rates, and the degree of relatives to be detected, while corresponding to m_e as well,

$$k_{|\alpha,\beta,\theta_r,m_e} > \frac{1}{\left(\frac{\theta_r}{z_{1-\beta}\sqrt{1-\theta_r^2}+z_{1-\alpha}}\right)^2 - \frac{1}{m_e}} \qquad \textbf{(Eq 4)}$$

In particular, α should be under experiment-wise control, say after Bonferroni correction, and consequently upon the total comparisons $\mathcal{N} = \sum_{i<j}^{C} n_i n_j$, where there are C cohorts and n_i is the sample size of cohort i, or just pair-wise comparisons $\mathcal{N}_{ij} = n_i n_j$.

Validation for theoretical results

We validated the variance of GRM, encGRM and encG-reg in simulations. 1,000 pairs of relatives were separated in cohort 1 and cohort 2. m = 1,000, 1,250, 1,500, 1,750 and 2,000 independent markers were simulated, and their minor allele frequency (MAF) was sampled from a uniform distribution $U(0.05, 0.5)$. Genotype matrices from two cohorts were encrypted with the same $m \times k$ random matrix **S**, whose elements drew from a normal distribution $N(0, \frac{1}{m})$. We set k to be 1,000, 2,000, 3,000, 4,000 and 5,000, respectively. Both real and encrypted genotype matrices were standardized based on the description for the three methods. Observed and theoretical variances were examined among four different degrees of relatedness ($\theta_r = 0.5^r$, in which $r = 0, 1, 2, \text{and } 3$ for r^{th} degree of relatives). Besides, to testify how allele frequency can influence the variance of GRM – which should be modeled by conditional binomial distribution as discussed above, we simulated 1,000 pairs of relatives of certain degrees, and 2,000 markers with the same MAF from 0.05 to 0.45 per increase in 0.1. We compared the observed variance of relatedness with the theoretical relatedness in 10 repeats.

We also examined how m and k affect the identification of various relatedness in simulations. We simulated 200 individuals each for cohort 1 and cohort 2 ($n_1 = n_2 = 200$); between cohort 1 and cohort 2 we generated 10 pairs of identical samples for each relative, i.e., 1st-degree, 2nd-degree, and 3rd-degree relatives, respectively. We set the desired number of markers (m) two times of that given by **Eq 3** and the corresponding size of k as given by **Eq 4** at the experiment-wise Type I error rate of 0.05 and Type II error rate of 0.1 – statistical power of 0.9 accordingly. We simulated individual-level genotype matrices with the dimension of $n_1 \times m$ and $n_2 \times m$ and the encrypted genotype matrices with the dimension of $n_1 \times k$ and $n_2 \times k$. Relatedness scores for GRM, encGRM and encG-reg were calculated accordingly and theoretical distributions were derived under the assumption of multivariate distribution for each degree of relatedness. In this case, we ignored the difference between m and m_e, because SNPs were generated independently here.

More detailed theoretical work for **Eq** 1~2 of GRM (**SNote 1 and 2**, and **SNote 3** for conditional binomial distribution properties of GRM), encGRM (**SNote 4**), and encG-reg (**SNote 5**) is summarized in supplementary notes and **Table S1-2** which was validated in simulation (**Figure S1-3**). Details on statistical power calculation for **Eq** 3~4 please see **SNote 6**.

Protocol for encG-reg for biobank-scale application

Figure 1 presents the workflow of encG-reg algorithm and its detailed implementation from cohort assembly to final relatedness identification. After the assembly of cohorts, there are options in choosing SNPs upon the experimental design. An exhaustive design denotes the use of intersected SNPs between each pair of cohorts, thus a specific random matrix will be shared to each pair of cohorts. Given C cohorts, there are $C(C-1)/2$ **S** matrices generated and each cohort is likely to receive $C-1$ different **S** matrices. Adopting exhaustive design is possibly to maximize the statistical power with maximized number of SNPs, but the computational, as well as communicational, efforts may overwhelm the organization of a study. In contrast, a parsimony design denotes the use of intersected SNPs among all assembled cohorts, as long as the number of SNPs satisfies the resolution in **Eq 3** and **Eq 4**. Exhaustive design and parsimony design are both validated in the 19 UKB cohorts, which had sample size greater than 10,000 each, and parsimony design are further tested in the real-world for 11 Chinese cohorts in this study.

We sketch encG-reg into a detailed technical protocol. This protocol can be automated, such as by a web server that coordinates the study. Once the cohorts are assembled, there are four steps in total, where steps 1 and 3 are performed by each collaborator and steps 2 and 4 are performed by a central analyst. We provide commands and simulated data in https://github.com/qixininin/encG-reg.

Step 1 Cohort assembly and intra-cohort quality controls Basic intra-cohort QCs should be conducted. Summary information such as SNP ID, reference allele, and its frequency are then requested by the central analyst.

Step 2 Inter-cohort quality controls and parameter set up Using "geo-geno" relationship, we suggested two inter-cohort QCs. One is called frequency-principle component analysis (fPCA) which illustrate the origins of cohorts, and another is called fStructure which explores genetic composition of each cohort in comparing with reference populations. The technical details of the employed methods can be found in our previous study[16]. Finally, the feasibilities of exhaustive and parsimony designs will be evaluated depending on the number of intersected SNPs and possible costs in communication. Central analyst determines m and k by **Eq 3** and **Eq 4** based on survived SNPs and passes parameter information to each collaborator along with an SNP list. The corresponding m_e will be estimated from, here, 1KG-EUR and 1KG-CHN as the reference populations for validation in the UKB cohorts and the Chinese cohorts, respectively.

Step 3 Encrypt genotype matrix The m-by-k random matrix, or matrices when an exhaustive design is chosen, is generated and sent to each cohort. As a positive control, reference samples will be merged to each cohort. Genotype encryption is realized by the matrix multiplication between the standardized genotype matrix and $\mathbf{S}$.

Step 4 Perform encG-reg Inter-cohort computing for relatedness will be conducted by the central analyst. A successful implementation will lead to at least positive controls consistently identified as inter-cohort "overlap" and if possible, various sporadic relatedness.

Validation I: UK Biobank in exhaustive and parsimony design

Both exhaustive and parsimony design were conducted for the validation of encG-reg on 485,158 UKB multi-ethnical samples from 19 assessment centers, which had sample size greater than 10,000 (**Table S3**). Identical/twins, 1st-degree and 2nd-degree relatedness were aimed to be detected by KING ("the rule of thumb") using the real genotypes and encG-reg using the encrypted genotypes, respectively. We conducted QC on the 784,256 chip SNPs within the 19 cohorts, and the inclusion criteria for autosome SNPs were: (1) MAF > 0.01; (2) Hardy-Weinberg equilibrium (HWE) test p-value > 1e-7; and (3) locus-level missingness < 0.05. In addition, taking account of cross-ethnicity nature in those UKB samples, only SNPs of ethnicity-insensitive frequency, which had indifferent allele frequencies statistically, were included.

For an exhaustive design, intersected SNPs were selected between each two cohorts, leading to generate 171 pairs of cohort combination for detecting relatedness. For a parsimony design, a total number of 12,858 intersected SNPs among all 19 cohorts were selected. The number of k for encG-reg were estimated by **Eq 4** at Type I error rate of 0.05 and Type II error rate of 0.1. To note that, experiment-wise Bonferroni correction is based on the number of paired samples between each two cohorts ($\mathcal{N}_{ij} = n_i n_j$) for exhaustive design and based on total number of paired samples among all cohorts ($\mathcal{N} = \sum_{i<j}^{\mathcal{C}} n_i n_j$) for parsimony design. The number of intersected SNPs were all given in **Table S4**.

To zoom in the performance of encG-reg, we took a close scrutiny at two assessment centers in Manchester (11,502 individuals) and Oxford (12,260 individuals) from UKB white British. We used KING to estimate relationship of any pair of individuals between two cohorts with the recommended thresholds of (0.354, 0.500), (0.177, 0.354), and (0.088, 0.177) in determining identical, 1st-degree, and 2nd-degree relatives[1]. 17 pairs of 1st-degree relatedness and 2 pairs of 2nd-degree relatedness detected (no identical samples detected) by KING were taken for a close scrutiny of encG-reg. As we have already known, in the discussion on **Eq** 1, that a relatively high MAF has smaller sampling variance and contributes more statistical power (**Figure S3**), we randomly sampled SNPs with different ranges of MAF (0.01 to 0.05, 0.05 to 0.15, 0.15 to 0.25, 0.25 to 0.35, 0.35 to 0.5, and 0.05 to 0.5) so as to compare the performance of encG-reg and KING. According to the minimal number of m_e and k at the experiment-wise Type I error rate of 0.05 and Type II error rate of 0.1 (**Table S5**), we selected 566 ($m_e = 566$) and 2,209 ($m_e = 2{,}023$) markers for detecting 1st-degree and 2nd-degree relatedness. m_e could be empirically estimated as $\frac{1}{var(\mathbf{G}_{off})}$, where $\mathbf{G}_{off}$ denotes the off-diagonal elements of GRM. Since m_e is asymptotically distributed as $N(m_e, \frac{4m_e^2}{n^2})$ according to our estimation, the sampling variance of m_e is negligible as long as the studying populations are of the similar ancestry, such as the case for Manchester and Oxford cohorts in UKB and the Chinese datasets employed in this study (**Table S6**). Against possible noise that may rust statistical power, we also increased k to $1.2k$ and denoted as encG-reg+. Average relatedness score, standard deviation and statistical power were calculated for each detected relative-pairs after resampling SNPs for 100 times.

Validation II: 10 multi-center Chinese datasets in parsimony design

We launched a national-scale test for encG-reg in 10 Chinese datasets under the parsimony design to avoid possible computational and communicational costs. 4 out of 10 datasets were publicly available, while the remaining datasets were recruited from 6 research centers, located in from north to south China, Beijing, Suzhou, Shanghai, Hangzhou, Guangzhou, and Shenzhen. As a proof of principle and brief validation of encG-reg in as civil as complex environment, these datasets agreed to detect identical samples or 1st-degree relatedness but without other exchange for medical information.

1KG-CHN (public): We considered two Chinese subpopulations in 1000 Genome Project (1KG)[18], CHB (Han Chinese in Beijing, 103 individuals) and CHS (Southern Han Chinese, 105 individuals) as reference population and positive control in the cross-cohort test in Chinese datasets. Individuals in the project were genotyped by whole-genome sequencing or whole-exon sequencing.

UKB-CHN (accessible after application): The UK Biobank (UKB) includes 1,653 individuals of self-reported Chinese[19]. After genomic assessment, 1,435 were considered from Chinese origin. Individuals in the project were genotyped using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix, followed by genotype imputation.

CONVERGE (public): The CONVERGE consortium aimed to investigate major depressive disorder (MDD)[20]. It included 5,303 Chinese women with recurrent MDD and 5,337 controls, all of whom were genotyped with low-coverage whole-genome sequencing and followed by imputation.

MESA (accessible after application): The Multi-Ethnic Study of Atherosclerosis (MESA) was to investigate subclinical cardiovascular disease[21]. 653 Chinese samples were included. Individuals were genotyped using Affymetrix Genome-Wide Human Single Nucleotide Polymorphism array 6.0, followed by genotype imputation.

SBWCH Biobank: The Shenzhen Baoan Women's and Children's Hospital (Baoan district, Shenzhen, Guangdong province) Biobank aims to investigate traits and diseases during pregnancy and at birth. 30,074 women were included in this study. Maternal genotypes were inferred from the non-invasive prenatal testing (NIPT) low depth whole genome sequencing data using STITCH[22] following the methodological pipeline that we previously published[23]. The average genotype imputation accuracy reaches 0.89 after filtration of INFO score 0.4.

CAS and ZOC: The Chinese Academy of Sciences (CAS) cohort is a prospective cohort study aiming to identify risk factors influencing physical and mental health of Chinese mental workers via a multi-omics approach. Since 2015, the study has recruited 4,109 CAS employees (48.2% male) located in Beijing, China. All participants belong to the research/education sector, and are characterized by a primary of Chinese Han origin (94.1%). DNA was extracted from peripheral blood samples and genotyped on the Infinium Asian Screening Array + MultiDisease-24 (ASA+MD) BeadChip, a specially designed genotyping array for clinical research of East Asian population with 743,722 variants. CAS study was approved by the Institutional Review Board of Beijing Institute of Genomics Chinese Academy of Sciences and Zhongguancun hospital. For validation purpose, samples were randomly split into CAS1 and CAS2. According to their records, ZOC was consisted of 19 homozygotic and heterozygotic siblings, who were evenly split into CAS1 and CAS2 as internal validation of encG-reg. ZOC is part of The Guangzhou Twin Eye Study (GTES), a prospective cohort study that included monozygotic and dizygotic twins born between 1987 and 2000 as well as their biological parents in Guangzhou, China. Baseline examinations were conducted in 2006, and all participants were invited to attend annual follow-up examinations. Non-fasting peripheral venous blood was collected by a trained nurse at baseline for DNA extraction, and genotyping was performed using the Affymetrix axiom arrays (Affymetrix) at the State Key Laboratory of Ophthalmology at Zhongshan Ophthalmic Center (ZOC)[24]. This study was approved by the ethics committee of Zhongshan Ophthalmic Center and was conducted in accordance with the tenets of the Declaration of Helsinki. Written informed consent was obtained for all participants from parents or their legal guardians. CAS and ZOC cohorts were deeply collaborated for certain studies, and consequently merged to fit this study.

Fudan: A multistage GWAS of glioma were performed in the Han Chinese population, with a total of 3,097 glioma cases and 4,362 controls. All Chinese Han samples used in this study were obtained through collaboration with multiple hospitals (Southern population from Huashan Hospital, Nanjing 1st Hospital, Northern population from Tiantan Hospital and Tangdu Hospital). DNA samples were extracted from blood samples and were genotyped using Illumina Human OmniExpress v1 BeadChips[25]. 2,008 samples were included for this study.

YiKon: YiKon cohort is striving for the research of reproductive medicine. 9,999 Chinese samples many with known pedigrees were included in this study. Individuals were genotyped using Illumina Infinium Asian Screening Array. For the validation of encG-reg, familial members were randomly split into YiKon1 (5,000 samples) and YiKon2 (4,999 samples).

WBBC: The Westlake BioBank for Chinese (WBBC) cohort is a population-based prospective study with its major purpose to better understand the effect of genetic and environmental factors on growth and development from youngster to elderly[26]. The mean age of the study samples were 18.6 years for males and 18.5 years for females, respectively. The Westlake BioBank WBBC pilot project have finished whole-genome sequencing (WGS) in 4,535 individuals and high-density genotyping in 5,841 individuals[27,28].

In total, based on 10 datasets, we reorganized, mostly retained, 11 Chinese cohorts (1KG-CHN, UKB-CHN, CONVERGE, META, SBWCH, CAS1, CAS2, Fudan, YiKon1, YiKon2 and WBBC) to be involved in the real-world test of encG-reg. Within CAS1 and CAS2 and within YiKon1 and YiKon2, relatedness if would be reported by encG-reg was verified by CAS and YiKon, respectively. Between other pairs of cohorts, sporadic relatedness might occur, as would have been found.

Results

Simulations

We performed a series of simulations to evaluate the robustness of encG-reg, accompanied by GRM and encGRM. The estimated sampling variance of GRM, encGRM and encG-reg matched with the theoretical variance at each level of relatedness (**Figure S2**). It was noticeable that larger MAFs could lead to a smaller variance of GRM score (**Figure S3)**, that further resulted in a smaller variance and a higher power of detecting relatives for encGRM and encG-reg. We also sketched up how m and k determined the resolution of encGRM and encG-reg (**Figure S4**). The results showed that for encG-reg, in each scenario, sufficient k was able to detect a certain degree of relatedness if m could support. As we evaluated in simulation, encG-reg stood out against encGRM with a smaller variance and a higher resolution as a good attempt in detecting relatives with encrypted genotypes.

Validation I: UKBiobank exercise for multi-ethnical samples

We verified the exhaustive design of encG-reg in 19 UKB cohorts by comparing with KING (**Figure 2A**). The average number of intersected SNPs between each two pairs of cohorts was 13,157. Relatedness was estimated and inferred up to the second degree, where KING used real genotypes and encG-reg used encrypted genotypes only. The same 38 pairs of identical samples (monozygotic twins in this case) were detected by KING and encG-reg, 7,965, and 6,632 pairs of 1st-degree and 2nd-degree relatedness were inferred by KING, the number of which went to 7,913 and 7,022 for encG-reg, respectively. It could be seen that encG-reg was quite comparable to KING in practice. Based on individual ID and their recorded ethnicity, consistent relatedness scores were estimated by KING and encG-reg

(**Figure 2B-D**). Combining geographic distance between 19 cohorts, we discovered that more relatives were detected between adjacent assessment centers, like Manchester and Bury, Newcastle and Middlesborough, and Leeds and Sheffield. Besides, consistent numbers of relatedness were inferred by the parsimony design of encG-reg (**Table S7**). The decrease in the number of detected 2nd-degree relatedness for parsimony design was possibly due to a smaller experiment-wise Type I error rate and thus a more stringent cutting threshold.

We took a closer look at two representative assessment centers in Manchester and Oxford. **Figure 2E** listed that of the $11{,}502 \times 12{,}260 = 141{,}014{,}520$ pairs of inter-cohort individuals, 17 pairs of so-called 1st-degree and 2 pairs of 2nd-degree relatives were found using overall QCed SNPs by KING. The bar plots compared relatedness scores of the known 1st-degree ($m_e = 566,\ k = 494$) and 2nd-degree ($m_e = 2023,\ k = 2{,}342$) relatives, estimated by KING, GRM, encG-reg, and encG-reg+ (using $1.2k$). In general, encG-reg and encG-reg+, still showed very similar estimations of relatedness score comparing with KING, even only encrypted genotypes were provided. When SNPs were sampled with MAFs between 0.05 and 0.5, the average statistical power reached 0.9 and 0.95 for detecting 1st-degree relatedness by encG-reg and encG-reg+. The overall statistical power increased as MAF increased; otherwise the MAF of the sampled SNPs was less than 0.05, the statistical power of encG-reg was practically as sufficient as devised (**Figure S5**).

Validation II: national-scale test in China

As summarized in **Figure 1**, the Chinese cohort study was swiftly organized and completed within about 7 weeks, demonstrating that encG-reg was easy to carry out. Following intra-cohort QCs and upon received summary information, we examined sample sizes and SNPs in each cohort (**Table 1**). In total, it included 64,091 samples and generated $\mathcal{N}$ =1,496,000,912 pairs of tests. When allele frequencies were compared with that in CONVERGE, the majority of SNPs had consistent allele frequencies across cohorts (**Table S8** and **Figure S6**). The missing rates and the intersected SNPs were also examined across cohorts (**Figure S7-8, and Table S9**), after which a total of 1,650 SNPs were in common among 11 cohorts for parsimony design of encG-reg (**Figure 3A**). The results of fPCA and fStucture matched with their expected "geo-geno" mirror in Chinese samples[23]. The first eigenvector of fPCA distinguished southern and northern Chinese samples in this study, the SBWCH Biobank (dominantly sampled from Shenzhen, the southmost metropolitan city in mainland China) and CAS cohort (dominantly sampled from Beijing)

(**Figure 3B and 3C**). Using a slightly different illustration strategy, the fStructure results, a counterpart to the well-known Structure plot in population genetics, were also consistent with the reported Chinese background of the 11 cohorts (**Figure 3C and 3D**). As the Chinese datasets showed little population structure, the choice of SNPs ignored the technical consideration for multi-ethnicity as in UKBiobank exercise.

We offered a list of 500 shared SNPs, whose m_e was 477 (evaluated in 1KG-CHN) and the corresponding minimal number of k was 757 given the experiment-wise Type I error rate of 0.05 and statistical power of 0.9. Each collaborator then encrypted their genotype matrix by the random matrix $\mathbf{S}$. As foolproof controls, 1KG-CHN samples were consistently identified as "identical" inter-cohort.

Anticipated relatives were identified between YiKon1 and YiKon2, and between CAS1 and CAS2 (**Figure 4A** and **4B**), and further validated by intra-cohort IBD calculation, respectively. Between YiKon1 and YiKon2, we reported 194 identical samples and 2,194 1st-degree relatedness, respectively. The pair-wise encG-reg distributions between cohorts were consistent to our theoretical expectation (**Figure 4C** and **Figure S9)**. Detected relatedness were confirmed by medical records (101 pairs were unknown among 2,388 identified pairs) in YiKon. However, for 20 inferred but unrecorded relatedness pairs, YiKon further verified them using real genotype data (**Figure 4D**). KING-inferred relatedness matched with encG-reg in 14 pairs. Of the rest six pairs that all identified as 1st-degree by encG-reg, three were inferred as 2nd-degree and one as unrelated by KING. In addition, due to possible adopted thresholds, KING reported two 1st-degree pairs as identical (their kinship scores were 0.390 and the suggested threshold for separating 1st-degree and identical pairs was 0.354), while encG-reg clearly separated identical pairs from 1st-degree (**Figure 4C)**.

Specifically, as each of 19 Guangzhou twins was split into CAS1 and CAS2, 18 pairs were identified as monozygotic (MZ) or dizygotic (DZ) by encG-reg and verified by intra-cohort IBD calculation in CAS Beijing team (**Figure 4E)**. Remarkably, one pair of so-called twins that was left out by encG-reg was verified as unrelated by IBD calculation, and ZOC team took further investigation on possible logistic errors. These results demonstrated that encG-reg was reliable with well controlled Type I and Type II error rates.

In particular, we illustrated how sporadically related pairs were captured by encG-reg. We detected 6 pairs of inter-cohort relatedness, including 2 pairs of identical samples and 4 pairs of 1st-degree relatives (**Table 2**). For these sporadic related inter-cohort samples, encG-reg exhibited their relatedness in forms of regression plots and estimated regression coefficients (**Figure 4F**). Obviously, compared with the regression plot for 2 pairs of identical samples, the higher missing rate of SBWCH then introduced more noise but was still captured by encG-reg. Nevertheless, its largest sample size provided SBWCH more linked with other cohorts. To avoid possible breaching of privacy we did not explore their relationship further here.

DISCUSSION

Individual genome sequencing is likely to be the trend and deserves well preserved privacy. The purpose of genomic data sharing often leads to cross-cohort tasks, such as finding relatives as occurred but of various purposes. Privacy-protection issues are raised during these tasks. One attempt on detecting cross-cohort relatives, limited to only overlapping individuals, employed one-way cryptographic hashes, which offered qualitative but not quantitative conclusions on false positive and false negative rates[29]. To settle the question of exact encryption precision, we focused more on the intrinsic consequence after genotype encryption with random matrix. Given our current knowledge in random matrix theory, we described its properties in how k and m_e influence the encryption precision for encrypted genotypes. This property is well testified in GRM which can be considered as a basis for a multiparty, or say cross-population genotype sharing. To note that the random matrix encryption, also called "random orthogonal keys", has been applied in performing GWAS[30,31]. They claimed that random orthogonal keys provide an encryption scheme where it is very difficult to recover individual genetic or phenotypic data. However, our investigation led to controllable encryption precision even under varying genotype platforms and data quality.

As demonstrated in UKB multi-ethnical samples, encG-reg could be applied for biobank-scale datasets with very high precision compared with conventional individual-level benchmark methods such as KING and GRM. Our real-world test in Chinese cohorts present an unprecedent attempt on developing safe method that can be applied in large-scale searching relatives with encrypted genomic data. In a real-world setup, for the sake of convenience and manageability, we only considered parsimony design of using shared SNPs across the 11 Chinese cohorts. Switching to exhaustive design will be a better choice if each pair of cohorts conducts encG-reg for their customized degree of relatives. Compared with UKB, which has relatives more frequently found in nearby assessment centers, the assembled Chinses cohorts are unanticipatedly fused a "functional cascade". The cohorts SBWCH, YiKon, and CONVERGE could be engaged in a much bigger network on human production medicine. Consequently, close relatives were detected between them. Likely was a person to join one or another genomic service under the influence of relatives who has already been included in a such service.

For either exhaustive design or parsimony design of encG-reg, the core algorithm is algebraic and asks little human information in its implementation, so developing an automatic central analysis facility that can significantly host and synchronize more cohorts will be in the near future. An exhaustive implementation of encG-reg will search even deeper relatedness across cohorts in a highly mobilizing nation like China, in which relatives were used to live nearby but now are more distantly due to industrialization[32]. A much deeper implementation of encG-reg will bring out unique resource for conducting biomedical research at large scale as including familial information as demonstrated[33]. Last but not least, encG-reg is developed a tool that, under much better protected genomic privacy, can facilitate necessary relative searching when it is needed but not for the purpose of penetrate membership or other unethical activities.

Data availability statement

Public datasets used in this study can be freely downloaded from the following URLs. Access to certain public databases may require researchers to submit their access requests.

1000 Genome Project: https://www.internationalgenome.org/home.

UK Biobank: https://www.ukbiobank.ac.uk/.

CONVERGE: http://dx.doi.org/10.5524/100155.

MESA: https://www.mesa-nhlbi.org/.

All codes for simulation study and practical protocol are available in https://github.com/qixininin/encG-reg.

Acknowledgements

We thank the participants of the included cohorts and of UK Biobank for making this work possible (UKB application 41376). This work is supported by National Natural Science Foundation of China (31771392 to GBC, 31900487 to

SL, 31871707 to HMX, 32061143019 to HFZ, and 81974197 to JH), Chinese Academy of Sciences (KFJ-STS-ZDTP-079, XDB38010400, KJZD-EW-L14), China Postdoc Council (E1QTJP0201), and QZ was partially supported by a student research fellowship of Zhejiang provincial people's hospital; the funders played no role in designing, preparation, and submission of the paper. Thank Qiu Feng, Mu Wentao, and Mei Lixiao for various assistance in making this work possible.

Author contributions

GBC conceived and initiated the study. GBC, SL (SWBCH), FL (CAS), YY (YiKon), HFZ (WBBC), MH (ZOC), DL (Fudan), and HMX designed the part of study for 11 Chinese datasets; each cohort team conducted intra-cohort analyses. GBC and QZ derived the analytical results. QZ conducted simulation, analyzed UKBiobank samples, and QZ developed the toolkit for encG-reg. GBC and QZ wrote the first draft of the paper, ZX, HZ, JH, XZ, and HM contributed to the writing and discussion that improved earlier versions of the paper. All authors contributed to the writing, discussion of the paper, and validation of the results.

SWBCH team: XG, JZ, and SL;

CAS team: LT, QZ, PJ, CZ and FL;

ZOC team: XH, XD, and MH;

WBBC team: MY, SK, and HFZ;

YiKon Genomics: KB, YY, and SLu;

Fudan team: FZ, HC, and DL.

Declare of Interests

None.

References

1. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–73 (2010).
2. Thomson, R. & McWhirter, R. Adjusting for Familial Relatedness in the Analysis of GWAS Data. *Methods Mol. Biol.* **1526**, 175–190 (2017).
3. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
4. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–15 (2013).
5. Guerrini, C. J. *et al.* Family secrets: Experiences and outcomes of participating in direct-to-consumer genetic relative-finder services. *Am. J. Hum. Genet.* **109**, 486–497 (2022).
6. Nelson, S. C., Bowen, D. J. & Fullerton, S. M. Third-Party Genetic Interpretation Tools: A Mixed-Methods Study of Consumer Motivation and Behavior. *Am. J. Hum. Genet.* **105**, 122–131 (2019).
7. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. Identity inference of genomic data using long-range familial searches. *Science* **362**, 690–694 (2018).
8. Ram, N., Guerrini, C. J. & McGuire, A. L. Genealogy databases and the future of criminal investigation. *Science* **360**, 1078–1079 (2018).
9. Ram, B. N., Murphy, E. E. & Suter, S. M. Regulating forensic genetic genealogy. *Science* **373**, 1444–1446 (2021).
10. Bonomi, L., Huang, Y. & Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* **52**, 646–654 (2020).
11. Wan, Z. *et al.* Sociotechnical safeguards for genomic data privacy. *Nat. Rev. Genet.* **23**, 429–445 (2022).
12. Ney, P., Ceze, L., Kohno, T. & Allen, P. G. Genotype Extraction and False Relative Attacks: Security Risks to Third-Party Genetic Genealogy Services Beyond Identity Inference. *Annu. Netw. Distrib. Syst. Secur. Symp.* (2020). doi:10.14722/ndss.2020.23049
13. Yu, H. & Xue, L. Shaping the evolution of regime complex: The case of multiactor punctuated equilibrium in governing human genetic data. *Glob. Gov.* **25**, 645–669 (2019).
14. Isserlis, L. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika* **12**, 134–139 (1918).
15. Halko, N., Martinsson, P. G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for

constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–288 (2011).

16. Chen, G. B. *et al.* Across-cohort QC analyses of GWAS summary statistics from complex traits. *Eur. J. Hum. Genet.* **25**, 137–146 (2016).

17. Chen, G.-B. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front. Genet.* **5**, 107 (2014).

18. Altshuler, D. L. *et al.* A map of human genome variation from population scale sequencing. *Nature* **467**, 1061–1073 (2010).

19. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

20. Cai, N. *et al.* Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).

21. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *Am. J. Epidemiol.* **156**, 871–881 (2002).

22. Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969 (2016).

23. Liu, S. *et al.* Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* **175**, 347–359 (2018).

24. Zheng, Y., Ding, X., Chen, Y. & He, M. The Guangzhou twin project: An update. *Twin Res. Hum. Genet.* **16**, 73–78 (2013).

25. Chen, H. *et al.* Two novel genetic variants in the STK38L and RAB27A genes are associated with glioma susceptibility. *Int. J. Cancer* **145**, 2372–2382 (2019).

26. Zhu, X. W. *et al.* Cohort profile: the Westlake BioBank for Chinese (WBBC) pilot project. *BMJ Open* **11**, e045564 (2021).

27. Cong, P. K. *et al.* Identification of clinically actionable secondary genetic variants from whole-genome sequencing in a large-scale Chinese population. *Clin. Transl. Med.* **12**, e866 (2022).

28. Cong, P. K. *et al.* Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**, 2939 (2022).

29. Turchin, M. C. & Hirschhorn, J. N. Gencrypt: one-way cryptographic hashes to detect overlapping individuals across samples. *Bioinformatics* **28**, 886–8 (2012).

30. Mott, R., Fischer, C., Prins, P. & Davies, R. W. Private Genomes and Public SNPs : Homomorphic Encryption of Genotypes and Phenotypes for Shared Quantitative Genetics. *Genetics* **215**, 359–372 (2020).

31. Yang, M. *et al.* TrustGWAS : A full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorandom number perturbation Methods TrustGWAS : A full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorand. *Cell Syst.* 1–16 (2022). doi:10.1016/j.cels.2022.08.001

32. Chen, G. B. Where is the friend's home. *Front. Genet.* **5**, 400 (2014).

33. Kaplanis, J. *et al.* Quantitative analysis of population-scale family trees with millions of relatives. *Science* **360**, 171–175 (2018).

Table 1 Summary information for the cohorts participated in this study

Cohort ID		Genotyping platform	Sample size	SNPs (after QC)	Description
1KG-CHN[18]		NGS	208	5,578,934	Chinese in 1000 Genome Project
UKB-CHN[19]		Affymetrix Chip + imputation	1,435	5,033,920	Chinese in UK Biobank
CONVERGE[20]		Low-coverage WGS + imputation	10,640	5,215,820	Chinese women in study of major depression
MESA[21]		Affymetrix Chip + imputation	653	4,950,239	Chinese samples in the multi-ethnic study of atherosclerosis
SBWCH[22,23]		Noninvasive prenatal testing (low-coverage WGS + imputation)	30,074	1,237,941	Chinese pregnancies recruited from the Shenzhen Baoan Women and Children's Hospital
CAS & ZOC[24]	CAS1	Illumina Chip; Affymetrix Chip	1,497	288,684	Unpublished Chinese samples mainly collected in Beijing, with which 19 pairs of twins (ZOC) were mixed in separately
	CAS2		1,497	288,539	
Fudan[25]		Illumina Chip	2,008	311,384	Chinese samples in the study of glioma
YiKon	YiKon1	Illumina Chip + single cell WGA	5,000	89,084	Chinese samples in the study of reproductive medicine
	YiKon2	Illumina Chip + single cell WGA	4,999	89,084	
WBBC[26–28]		Illumina Chip	6,080	319,930	The Westlake BioBank for Chinese pilot project
			64,091 (all)	1,650 (intersection)	

Table 2 Supporting evidence for the related pairs

Pair	Cohort 1	ID 1	Cohort 2	ID 2	Score (SD[a])	Score[b] (SD)	Inferred relatedness
1	SBWCH	SBWCH_21253	YiKon2	YKB1693	0.890 (0.017)	0.993 (0.019)	Identical
2	CAS1	2009111148	YiKon2	YKB570	0.985 (0.002)	0.999 (0.002)	Identical
3	SBWCH	SBWCH_2988	YiKon1	YKA1770	0.397 (0.033)	0.434 (0.036)	1st-degree
4	SBWCH	SBWCH_28165	YiKon1	YKA3820	0.406 (0.033)	0.479 (0.039)	1st-degree
5	SBWCH	SBWCH_200	WBBC	WBBC3849	0.427 (0.033)	0.533 (0.041)	1st-degree
6	YiKon2	YKB1046	CONVERGE	MD_CHW_AAD_11728	0.511 (0.031)	0.512 (0.031)	1st-degree

Notes: IDs were de-identified by each cohort.

[a] Standard deviation (SD) is calculated from $SD_{b_{ij}} = \sqrt{\frac{cov(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j)}{var(\hat{\mathbf{x}}_i)}}$, where $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ are the vectors of the encrypted genotypes for two individuals.

[b] Due to missing data, the corrected score, is adjusted for the genotype missing rate between the pair of individuals.

Figure 1 Workflow of encG-reg and its practical timeline as exercised in Chinese cohorts

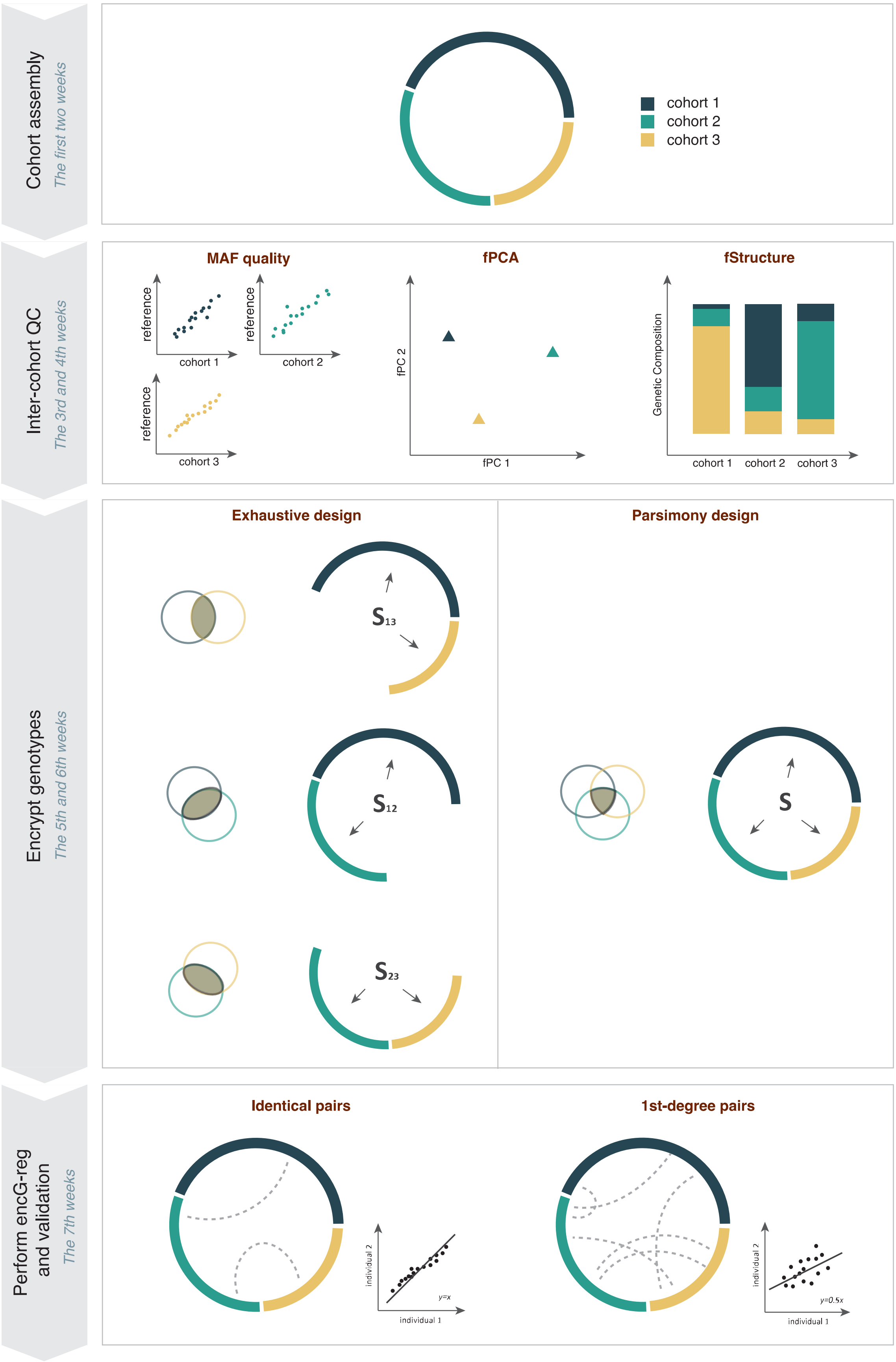

Figure notes: The mathematical details of encG-reg is simply algebraic, but its inter-cohort implementation involves coordination. We illustrate its key steps, the time cost of which was adapted from the present exercise for 10 Chinese datasets (here simplified as three cohorts). **Cohort assembly**: It took us about a week to call and got positive responses from our collaborators (See **Table 1**), who agreed with our research plan. **Inter-cohort QC**: we received allele frequencies reports from each cohort and started to implement inter-cohort QC according to "geo-geno" analysis (see **Figure 2**). This step took about two weeks. **Encrypt genotypes**: upon the choice of the exercise, it could be exhaustive design (see UKB example), which may maximize the statistical power but with increased logistics such as generating pairwise $\mathbf{S}_{ij}$; in the Chinese cohorts study we used parsimony design, and generated a unique $\mathbf{S}$ given 500 SNPs that were chosen from the 1,650 common SNPs. It took about a week to determine the number of SNPs and the dimension of k according to **Eq 3** and **4**, and to evaluate the effective number of markers. **Perform encG-reg and validation**: we conducted inter-cohort encG-reg and validated the results (see **Figure 3** and **Table 2**). It took one week.

Figure 2 Resolution for detecting relatives in UKB cohorts by KING and encG-reg at exhaustive design

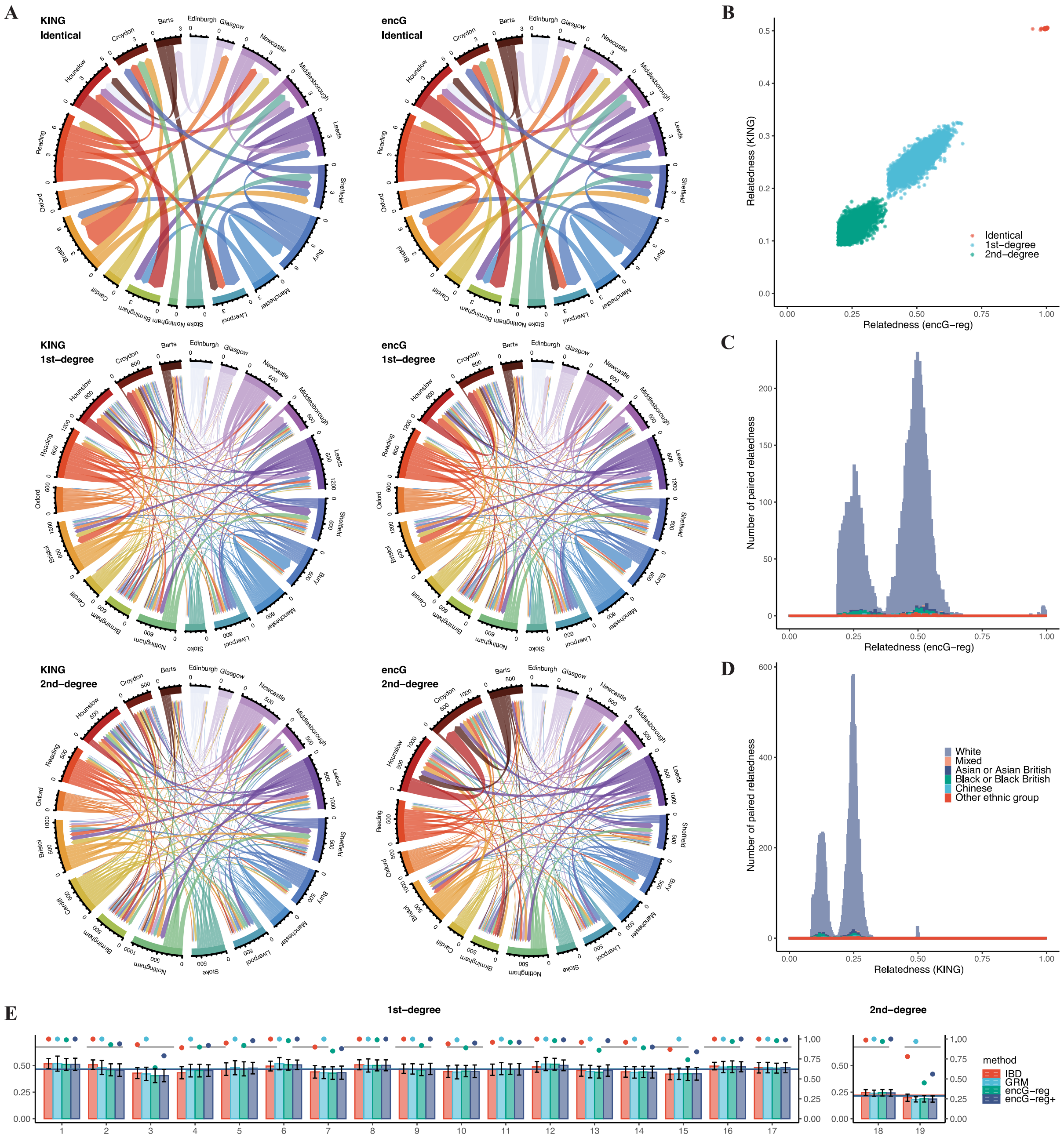

Figure notes: (**A**) Chord diagrams shows the number of inter-cohort identical/twins, 1st-degree and 2nd-degree relatedness for 19 UKB assessments which had more than 10,000 samples. Relatedness were detected and compared between KING and encG-reg under an exhaustive design, totaling 171 inter-cohort analyses. In each chord plot, the length of its side edge was proportional to the count of detected relatives between this cohort with other cohorts. (**B**) Scatter plot showed estimated relatedness score by KING and encG-reg. The inter-cohort links for the three relative clusters were as shown in A. (**C**) and (**D**) are the respective relatedness score distributions. (**E**) The bar plot compared relatedness scores of the known 1st-degree and 2nd-degree relatives estimated by KING, GRM, encG-reg and encG-reg+ across two representative assessment centers (Manchester and Oxford). 566 and 2,209 SNPs were randomly selected with MAF between 0.05 and 0.5. Here, encG-reg+ denotes the use of 1.2-fold of the minimal number of k and IBD denotes twice of the relatedness score estimated by KING. Average GRM score, standard deviation and statistical power were calculated for each detected relative-pair after resampling SNPs for 100 times. The grey dash line indicates the expected statistical power of 0.9. Colored solid lines indicate the average relatedness scores of certain degrees by the four methods. 17 pairs of so-called 1st-degree and 2 pairs of 2nd-degree relatives were approved using overall SNPs by KING.

Figure 3 Cohort-level genetic background analyses for Chinese cohorts under parsimony encG-reg analysis.

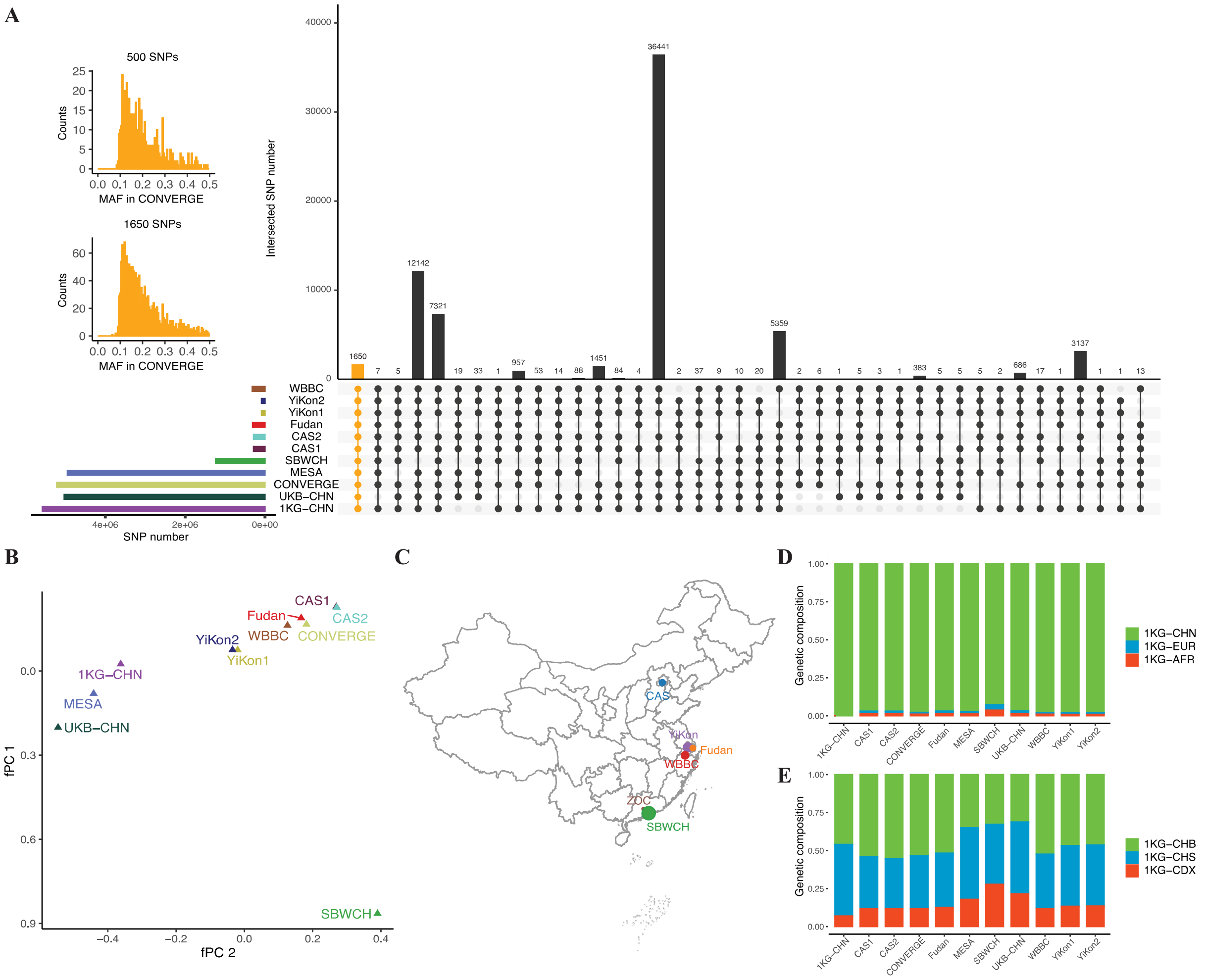

Figure notes: (**A**) Overview of the intersected SNPs across cohorts, a black dot indicated its corresponding cohort was included. Each row represented one cohort while each column represented one combination of cohorts. Dots linked by lines suggested cohorts in this combination. The height of bars represented the cohort's SNP numbers (rows) or SNP intersection numbers (columns). Inset histogram plot showed the distribution of the 1,650 intersected SNPs and the 500 SNPs chosen from the 1,650 SNPs for encG-reg analysis. (**B**) 1,650 SNPs were used to estimate fPC from the intersection of SNPs for the 11 cohorts. Each triangle represented one Chinese cohort and was placed according to their first two principle component score (fPC1 and fPC2) derived from the received allele frequencies. (**C**) A Chinese map had 6 private datasets pinned on it, according to the location of data owners. The size of point indicated the sample size of each dataset. (**D**) Global fStructure plot indicated global-level F_{st}-derived genetic composite projected onto the three external reference populations: 1KG-CHN (CHB and CHS), 1KG-EUR (CEU and TSI), and 1KG-AFR (YRI), respectively; 1,041 of the 1,650 SNPs intersected with the three reference populations were used. (**E**) Within Chinese fStructure plot indicated within-China genetic composite. The three external references were 1KG-CHB (North Chinese), 1KG-CHS (South Chinese), and 1KG-CDX (Southwest minority Chinese Dai), respectively; 1,164 of the 1,650 SNPs intersected with these three reference populations were used. Along x axis were 11 Chinese cohorts and the height of each bar represented its proportional genetic composition of the three reference populations. Cohort codes: YRI, Yoruba in Ibadan representing African samples; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; CHN, CHB and CHS together; CEU, Utah Residents with Northern and Western European Ancestry; TSI, Tuscani in Italy; CDX, Chinese Dai in Xishuangbanna.

Figure 4 Detected identical pairs and 1st-degree pairs between Chinese cohorts

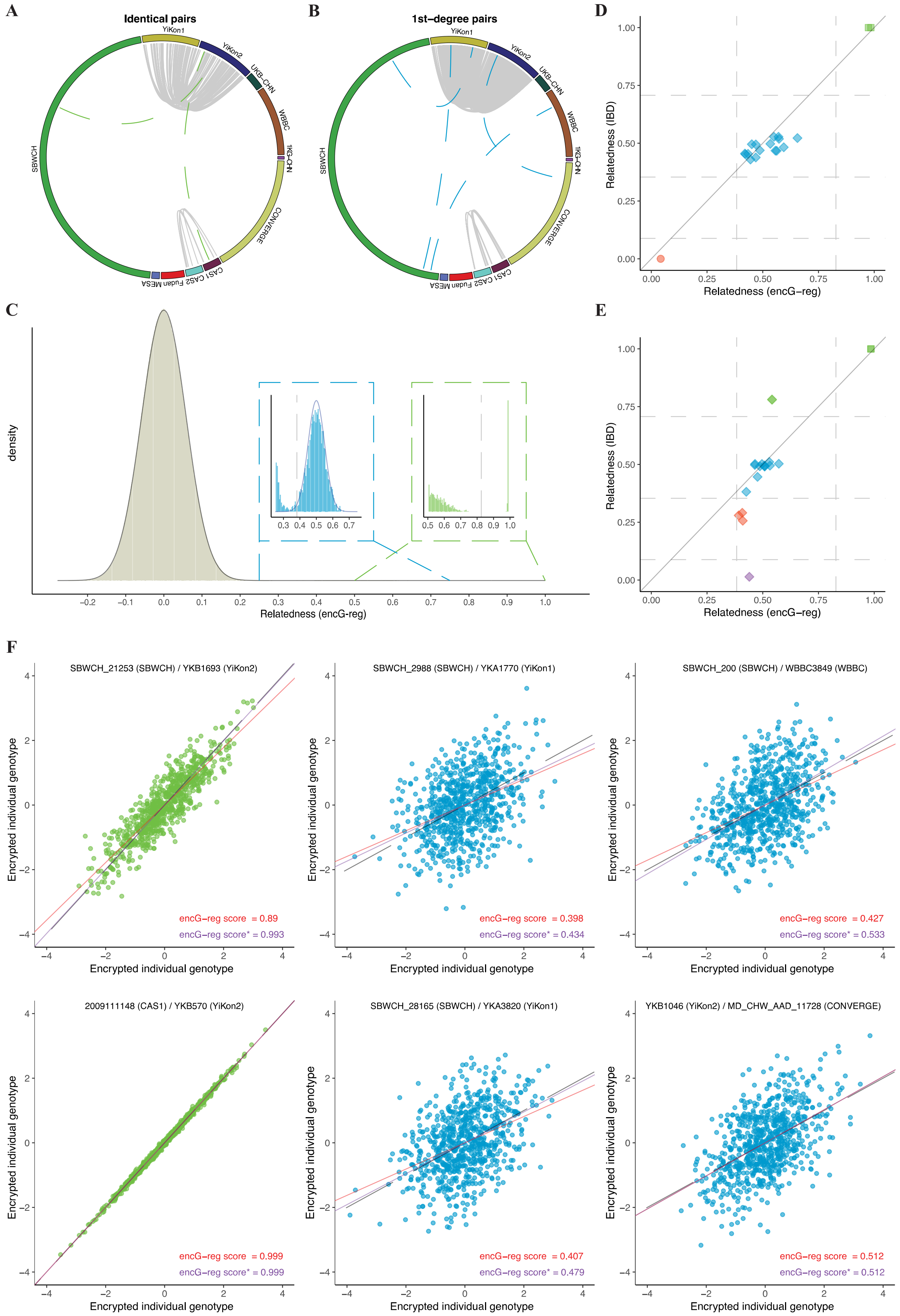

Figure notes: **(A)** The circle plot illustrated identical pairs and **(B)** 1st-degree pairs across 11 Chinses cohorts. The solid links indicated anticipated relatedness between the CAS cohorts and between the YiKon cohorts. The dashed links were sporadic relatedness found between the cohorts. The length of each cohort bar was proportional to their respective sample sizes. **(C)** The histogram showed all estimated relatedness using encG-reg, most of which were unrelated pairs and the theoretical probability density function was given as the normal distribution $N\left(0, \frac{1}{m_e} + \frac{1}{k_1}\right)$ (grey solid curve). The inset histogram on the left showed the estimated relatedness around 0.5 and the theoretical probability density function was given as the normal distribution $N\left(\theta_r, \frac{1-{\theta_r}^2}{m_e} + \frac{1-{\theta_r}^2}{k_1}\right)$ (blue solid curve). The threshold (grey dot line) for rejecting H_0 was calculated by $z_{1-\alpha/\mathcal{N}}\sqrt{\frac{1}{m_e} + \frac{1}{k_1}}$. The inset histogram on the right showed estimated relatedness around 1. The threshold (grey dot line) for rejecting H_0 was calculated by $z_{1-\alpha/\mathcal{N}}\sqrt{\frac{1}{m_e} + \frac{1}{k_0}}$. Here we included 208 controls merged from 1KG-CHN. $m_e = 477$, $k_0 = 72$, $k_1 = 757$, $\mathcal{N} = 1{,}496{,}000{,}912$. **(D)** Relationship verification for 20 YiKon pairs that had mismatched medical records with encG-reg inference. Relatedness score (y axis) was estimated in KING by YiKon. Dashed lines indicated inference criteria for detecting a range of relatedness. Solid line of $y = x$ indicated the agreement between encG-reg and IBD. Points were colored with KING-inferred relatedness (identical in green, 1st-degree in blue, 2nd-degree in red and unrelated in purple) and shaped with encG-reg-inferred relatedness (identical in square and 1st-degree in diamond). **(E)** Relationship verification for 19 Guangdong twins split in CAS cohorts. Dashed lines indicated inference criteria for detecting relatedness of different degrees. Solid line of $y = x$ indicated the agreement between encG-reg and IBD. Points were colored with IBD-inferred, in KING, relatedness (identical in green, 1st-degree in blue and unrelated in red) and was shaped according to encG-reg-inferred relatedness (identical in square, 1st-degree in diamond and unrelated in circle). **(F)** Illustration for encG-reg estimation for sporadic related inter-cohort samples. In each plot the grey line was the criterion for identical pairs (slope of 1) or 1st-degree pairs (slope of 0.5). The solid lines coloured in red were without adjustment for missing values (engG-reg score), and in the bottom (coloured in purple) were adjusted for missing values (encG-reg score*). The first two pairs (coloured in green) were inferred as identical samples, whose encG-reg scores were close to 1, and the rest four pairs (coloured in blue) were 1st-degree pairs, whose encG-reg scores were close to 0.5.