Generating synthetic data

The synthetic datasets include the paired synthetic SNPs data $\mathcal{T}$ and synthetic QTs data $\mathcal{H}$ for subjects. The details of generating synthetic data are described as follows.

1) Generating the synthetic SNPs data $\mathcal{T}$

The synthetic SNPs data is defined as $\mathcal{T}=[\mathcal{T}_{\text{normal}},\mathcal{T}_{\text{disease}}]$, where $\mathcal{T}_{\text{normal}}$ is the synthetic SNPs data from the normal control (NC) subjects and $\mathcal{T}_{\text{disease}}$ is the synthetic SNPs data from the Alzheimer's disease (AD) subjects.

For the NC subjects, we first extract the n_{ts} risk SNPs from the ADNI dataset, then calculate the mutation rate $\varphi_n(g_n)$ of each $(g_n)^{\text{th}}$ subject by adopting the following formula, $\varphi_n(g_n)=\{\frac{1}{n_{ts}}\sum_{ts=1}^{n_{ts}}\sum_{i=1}^{2}t_{g_n}^{(ts)_i}\,|\,t_{g_n}^{(ts)_1}=1,\ t_{g_n}^{(ts)_2}=2,\ g_n=1,2,...,G_n\}$, where $t_{g_n}^{ts}$ is the value of $(ts)^{\text{th}}$ SNP in $(g_n)^{\text{th}}$ subject, G_n is the number of NC subjects. We assume the mutation rate φ_n following a normal distribution, *i.e.* $\varphi_n \sim \mathcal{N}(\mu_n,\sigma_n^2)$ (p-value = 2.2E-16) as shown in Fig. S1 (a), where $\mu_n=\frac{1}{G_n}\sum_{g_n=1}^{G_n}\varphi_n(g_n)$ and $\sigma_n^2=\frac{1}{G_n}[\varphi_n(g_n)-\mu_n]^2$. Using φ_n, we synthetic SNPs data $t_{m_1,v}$ of m_1 SNP in subject v following a Bernoulli distribution, *i.e.* $t_{m_1,v} \sim Bernoulli(\varphi_n)$. In this work, we generate 500 NC subjects, each with 1000 SNPs.

For the AD subjects, we first extract n_{ts} risk SNPs from ADNI dataset, and then calculate the mutation rate $\varphi_d(g_d)$ of each $(g_d)^{\text{th}}$ subject by adopting the following formula, $\varphi_d(g_d)=\{\frac{1}{n_{ts}}\sum_{ts=1}^{n_{ts}}\sum_{i=1}^{2}t_{g_d}^{(ts)_i}\,|\,t_{g_d}^{(ts)_1}=1,\ t_{g_d}^{(ts)_2}=2,\ g_d=1,2,...,G_d\}$, where $t_{g_d}^{ts}$ is the value of $(ts)^{\text{th}}$ SNP in $(g_d)^{\text{th}}$ subject, G_d is the number of AD subjects. We assume the mutation rate φ_n following a normal distribution, *i.e.* $\varphi_d \sim \mathcal{N}(\mu_d,\sigma_d^2)$ (p-value = 2.7E-18) as shown in Fig. S1 (b). Using φ_d, we synthetic SNPs data $t_{m_2,d}$ of m_2 SNP in subject d following a Bernoulli distribution, *i.e.* $t_{m_2,d} \sim Bernoulli(\varphi_d)$. In this work, we generate 500 AD subjects, each with 1000 SNPs.

In this work, $n_{ts} = 2000$, $G_n = 223$, $\mu_n = 0.146$, $\sigma_n^2 = 0.0292$, $G_d = 277$, $\mu_d = 0.481$, $\sigma_d^2 = 0.0444$.

2) Calculating the related mean μ_1 and nonrelated mean μ_2

First, based on the empirical distribution as shown in Fig. S1 (c), we obtain the related SNPs-QTs distribution $\beta_1 \sim \mathcal{N}(-0.3, 0.1)$ from AD subjects, the unrelated SNPs-QTs distribution $\beta_2 \sim \mathcal{N}(0, 0.02)$ from NC subjects, respectively.

Second, we assume the random effect variable $b_{qt,v}$ of $(qt)^{\text{th}}$ QT in subject v following a Gaussian distribution, *i.e.* $b_{qt,v} \sim \mathcal{N}(\mu_{b'}, \sigma_{b'}^2)$, and then use the generalized linear mixed model (glmmTMB) to estimate the parameters $\mu_{b'}$ and $\sigma_{b'}^2$ for each subject based on the matrix composed of QT n_r $(r = 1, 2, ..., 119)$ and the top selected n_{ts} risk SNPs (Fig. S1 d-e). In this work, $\mu_{b'} = 3.07$ and $\sigma_{b'}^2 = 1.48$.

Third, after generating the synthetic NC SNPs data $t_{m_1,v}$, the synthetic AD SNPs data $t_{m_2,d}$, and obtaining the related regression coefficients β_1, the unrelated regression coefficients β_2, and the random effect variable $b_{qt,v}$, we calculate the normalized expected related mean $\mu_{1,d} = e^{\beta_1 t_{m_2,d} + b_{qt,v}}$, and the unrelated mean $\mu_{2,v} = e^{\beta_2 t_{m_1,v} + b_{qt,v}}$, $\mu_{2,d} = e^{\beta_2 t_{m_2,d} + b_{qt,v}}$.

3) Calculating random variance σ^2

After getting the QT matrix $\mathrm{M_R} \in \mathbb{R}^{G_n \times 119}$ for NC subjects, where G_n is the number of NC subjects, we generate the scatter plot of the mean μ and the variance σ^2 based on the $\mathrm{M_R}$ as shown in Fig. S1 (f). From Fig. S1 (f), we can see that the variance σ^2 follows a binomial distribution, *i.e.* $\sigma^2 \sim \textit{binomial distribution}(\mu, 2),\ i = 1, 2$. Therefore, the random variance corresponding to μ_1 is $\sigma_1^2 \sim \textit{binomial distribution}(\mu_1, 2)$ and the random variance corresponding to μ_2 is $\sigma_2^2 \sim \textit{binomial distribution}(\mu_2, 2)$.

4) Generating the synthetic QTs data

The synthetic QTs data is defined as $\mathcal{H}=[\mathcal{H}_{\text{normal}},\mathcal{H}_{\text{disease}}]$. $\mathcal{H}_{\text{normal}}$ is generated by $h_{normal}\sim\mathcal{N}(\mu_1,\sigma_1^2)$ and $\mathcal{H}_{\text{disease}}$ is generated by $h_{disease}\sim\mathcal{N}(\mu_2,\sigma_2^2)$.

For the AD subjects, let two sparse vectors: $\boldsymbol{Q}\in\mathbb{R}^{s_a\times 1}$ and $\boldsymbol{T}\in\mathbb{R}^{s_b\times 1}$ represent AD-related QTs and SNPs, respectively, s_a denotes the dimension number for QTs randomly selected from 30% of the 119 QTs, and s_b denotes the number of SNPs related to QTs of AD subjects. In this work, we set $s_a=36$, and s_b to 10, 20, 25. For example, we set $s_a=36$, $s_b=10$, and $QT_{18},QT_{19},...,QT_{53}$ are the QTs related with AD, i.e., $\underbrace{QT_1,QT_2,...,QT_{17}}_{\text{unrelated QTs}},\underbrace{QT_{18},QT_{19},...,QT_{53}}_{\text{related QTs}},\underbrace{QT_{54},QT_{55},...,QT_{119}}_{\text{unrelated QTs}}$

where, $QT_{18}:\underbrace{\beta_1^{18},\beta_2^{18},...,\beta_{10}^{18}}_{\text{10 related SNPs}},\underbrace{...,\beta_{1000}^{18}}_{\text{unrelated SNPs}}$,

$QT_{19}:\underbrace{\beta_1^{19},\beta_2^{19},...,\beta_{10}^{19}}_{\text{unrelated SNPs}},\underbrace{\beta_{11}^{19},\beta_{12}^{19},...,\beta_{20}^{19}}_{\text{10 related SNPs}},\underbrace{...,\beta_{1000}^{19}}_{\text{unrelated SNPs}}$, and so on.

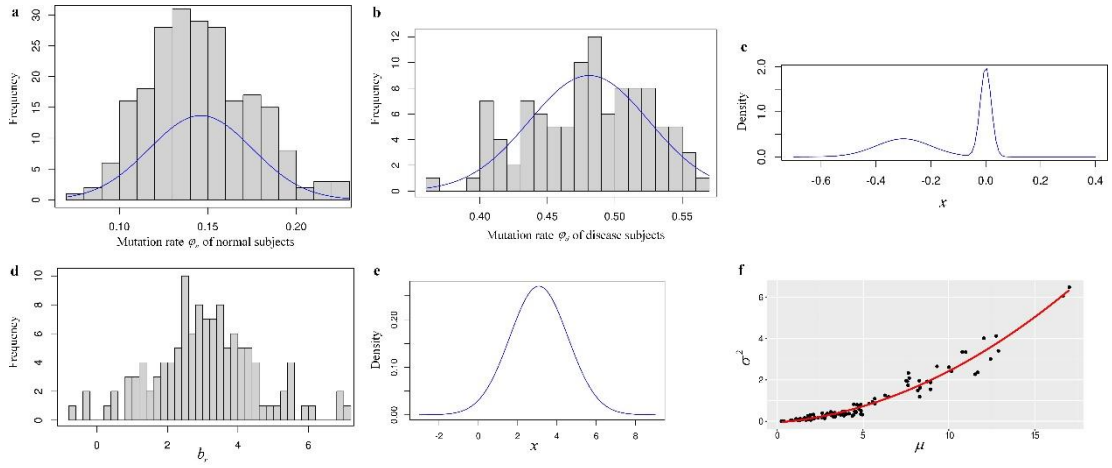

Fig. S1. a. The mutation rate φ_n of NC subjects from ADNI dataset and its gaussian distribution map (blue line). b. The mutation rate φ_d of AD subjects from ADNI dataset and its gaussian distribution map (blue line). c. The gaussian distribution map of SNPs related to QT ($\beta_1\sim\mathcal{N}(-0.3,0.1)$) and SNP unrelated to QT ($\beta_2\sim\mathcal{N}(0,0.02)$). d. The histogram of the random effect variable b_r. e. The gaussian distribution map based on b_r. f. The binomial distribution map of random variance σ^2 based on the μ.