

Supplementary Materials for “Estimation of Direct and Indirect Polygenic Effects and Gene-Environment Interactions using Polygenic Scores in Case-Parent Trio Studies”

Contents

1	Statistical Derivations	2
1.1	Derivation under the Assumption of No Indirect Effects of Parental PGS	2
1.1.1	Likelihood Derivation	3
1.1.2	Parameter Estimation	3
1.1.3	Asymptotic Variance Estimation when Only Considering Direct PGS Effect	4
1.1.4	Asymptotic Variance Estimation in the General Form	5
1.2	Derivation Incorporating Indirect Effects of Parental PGS	6
1.2.1	Derivation of Estimates of Parental Indirect Genetic Effects	7
1.2.2	Approximate Estimator of the Required Scale Factor	7
1.2.3	Asymptotic Variance Estimation	8
2	Details of Simulation Studies	8
2.1	Simulate PGS Values Based on the Assumed Model	8
2.2	Simulation using the UK Biobank Data	9
3	Details of the Data Applications	10
3.1	Data Analyses of Autism Spectrum Disorder (ASD) in the SPARK Study	10
3.1.1	Genotype Data Preprocessing and PGS Construction	10
3.2	Data Analyses of Non-Syndromic Orofacial Clefts (OFCs) in the GENEVA Study	10
3.2.1	Genotype Data Preprocessing and PGS Construction	10
3.3	Association Studies of Genetically Predicted Multi-Omics Data on ASD and OFCs	11

1 Statistical Derivations

1.1 Derivation under the Assumption of No Indirect Effects of Parental PGS

Let PGS_{iC} denote the PGS value for the child/offspring, PGS_{iM} the PGS for mother, and PGS_{iF} the PGS for father in family i , $i = 1, \dots, N$. Let D_{iC} denote the disease status for the offspring. For each family i , the prospective risk model for D_{iC} follows a log-linear model in the form of $\text{pr}(D_{iC}|PGS_{iC}, E_{iC}) = \exp(\alpha_i + \beta_G PGS_{iC} + \beta_E^T E_{iC} + \beta_{GE}^T E_{iC} PGS_{iC})$. Under rare disease assumptions, the parameters of the log-linear model can be interpreted as odds ratios, but more generally they correspond to relative risks. Assume that the probability of $(PGS_{iC}, PGS_{iM}, PGS_{iF})^T$ follows

$$\begin{pmatrix} PGS_{iC} \\ PGS_{iM} \\ PGS_{iF} \end{pmatrix} \sim \mathcal{N} \left(\mu_i \mathbf{1}_3, \sigma_i^2 \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix} \right), \quad (1)$$

where μ_i and σ_i are the mean and standard deviation for each family/strata i .

We will show that

$$PGS_{iM/iF}|D_{iC} = 1, \mathbf{E}_{iC} \sim \mathcal{N} \left((\mu_i + \frac{\sigma_i^2}{2}(\beta_G + \beta_{GE}^T \mathbf{E}_{iC})), \sigma_i^2 \right). \quad (2)$$

To derive (2), we consider a risk model of the form

$$\text{pr}(D = 1|U, \mathbf{E}) = \exp(\alpha + \beta_G U + \beta_{GE}^T \mathbf{E} U + \beta_E^T \mathbf{E}),$$

where U is the PGS value for offspring and we assume $U \sim \mathcal{N}(\mu, \sigma^2)$. Let R_i and R_j index pair of relatives and d_{ij} denote degree of relatedness. For parent-offspring, $d_{ij} = 1$. We have $\text{cov}(U_{R_i}, U_{R_j}) = \rho\sigma^2 = 0.5^{d_{ij}}\sigma^2$. We want to show that

$$U_{R_i}|D_{R_j} = 1, \mathbf{E}_{R_j} \sim \mathcal{N}(\mu + 0.5^{d_{ij}}\sigma^2(\beta_G + \beta_{GE}^T \mathbf{E}_{R_j}), \sigma^2). \quad (3)$$

We can write the probability as

$$\begin{aligned} \text{pr}(U_{R_i}|D_{R_j} = 1, \mathbf{E}_{R_j}) &= \int_{\mathbb{R}} \text{pr}(U_{R_i}, U_{R_j}|D_{R_j} = 1, \mathbf{E}_{R_j}) dU_{R_j} \\ &= \int_{\mathbb{R}} \frac{\text{pr}(D_{R_j} = 1|U_{R_i}, U_{R_j}, \mathbf{E}_{R_j}) \text{pr}(U_{R_i}, U_{R_j}|\mathbf{E}_{R_j})}{\text{pr}(D_{R_j} = 1|\mathbf{E}_{R_j})} dU_{R_j} \\ &= \int_{\mathbb{R}} \frac{\text{pr}(D_{R_j} = 1|U_{R_j}, \mathbf{E}_{R_j}) \text{pr}(U_{R_i}, U_{R_j})}{\int_{\mathbb{R}} \text{pr}(D_{R_j} = 1, U_{R_j}|\mathbf{E}_{R_j}) dU_{R_j}} dU_{R_j} \\ &= \int_{\mathbb{R}} \frac{\text{pr}(D_{R_j} = 1|U_{R_j}, \mathbf{E}_{R_j}) \text{pr}(U_{R_i}, U_{R_j})}{\int_{\mathbb{R}} \text{pr}(D_{R_j} = 1|U_{R_j}, \mathbf{E}_{R_j}) f(U_{R_j}) dU_{R_j}} dU_{R_j} \\ &= \int_{\mathbb{R}} \frac{\exp(\alpha + \beta_G U_{R_j} + \beta_{GE}^T \mathbf{E}_{R_j} U_{R_j} + \beta_E^T \mathbf{E}_{R_j}) \text{pr}(U_{R_i}, U_{R_j})}{\int_{\mathbb{R}} \exp(\alpha + \beta_G U_{R_j} + \beta_{GE}^T \mathbf{E}_{R_j} U_{R_j} + \beta_E^T \mathbf{E}_{R_j}) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(U_{R_j} - \mu)^2}{2\sigma^2}) dU_{R_j}} dU_{R_j} \\ &= \int_{\mathbb{R}} \frac{\exp(\beta_G U_{R_j} + \beta_{GE}^T \mathbf{E}_{R_j} U_{R_j}) \text{pr}(U_{R_i}, U_{R_j})}{\int_{\mathbb{R}} \exp(\frac{1}{2}\sigma^2(\beta_G + \beta_{GE}^T \mathbf{E}_{R_j})^2 + \mu(\beta_G + \beta_{GE}^T \mathbf{E}_{R_j})) dU_{R_j}} dU_{R_j}, \end{aligned} \quad (4)$$

we know that $\text{pr}(U_{R_i}, U_{R_j})$ follows a bivariate normal distribution with $\text{cov}(U_{R_i}, U_{R_j}) = \rho\sigma^2$:

$$\text{pr}(U_{R_i}, U_{R_j}) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(U_{R_i} - \mu)^2}{\sigma^2} - 2\rho \frac{(U_{R_i} - \mu)(U_{R_j} - \mu)}{\sigma^2} + \frac{(U_{R_j} - \mu)^2}{\sigma^2} \right] \right\}.$$

For ease of notation, let $x = U_{R_j}$ and $y = U_{R_i}$. To calculate $\int_{\mathbb{R}} \exp(\beta_G x + \beta_{GE}^T E x) \text{pr}(x, y) dx$, we use conditional distribution of $\text{pr}(X|Y = y)$, i.e.,

$$\begin{aligned} \int_{\mathbb{R}} \exp(\beta_G x + \beta_{GE}^T E x) \text{pr}(x, y) dx &= \int_{\mathbb{R}} \exp(\beta_G x + \beta_{GE}^T E x) \text{pr}(x|Y = y) \text{pr}(y) dy \\ &= \mathbb{E}(e^{(\beta_G + \beta_{GE}^T E)x} | Y = y) \text{pr}(y) \end{aligned}$$

We can derive $X|Y = y \sim \mathcal{N}((1 - \rho)\mu + \rho y, (1 - \rho^2)\sigma^2)$, then we use the moment generating function for normal distribution and we have

$$\begin{aligned} \int_{\mathbb{R}} \exp(\beta_G x + \beta_{GE} E x) \text{pr}(x, y) dx &= \exp \left\{ \frac{1}{2} (1 - \rho^2) \sigma^2 (\beta_G + \beta_{GE} E)^2 + [(1 - \rho)\mu + \rho y] (\beta_G + \beta_{GE} E) \right\} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{\{y - [\mu + \rho\sigma^2(\beta_G + \beta_{GE} E)]\}^2}{2\sigma^2} \right\} \exp \left(\mu(\beta_G + \beta_{GE} E) + \frac{\sigma^2(\beta_G + \beta_{GE} E)^2}{2} \right). \end{aligned}$$

Plug into (4) and we have proved (3).

From (1), we can derive the conditional probability of PGS_{iC} as

$$PGS_{iC}|PGS_{iM}, PGS_{iF} \sim \mathcal{N} \left(\frac{1}{2}(PGS_{iM} + PGS_{iF}), \frac{\sigma_i^2}{2} \right), \quad (5)$$

denote $\mu_{iC} = \frac{1}{2}(PGS_{iM} + PGS_{iF})$ and $\sigma_{iC}^2 = \frac{\sigma_i^2}{2}$.

1.1.1 Likelihood Derivation

The conditional likelihood for each family i is:

$$\begin{aligned} L_i &= \text{pr}(PGS_{iC}, PGS_{iM}, PGS_{iF} | \mathbf{E}_{iC}, D_{iC} = 1) \\ &= L_{iC} \times L_{iP} \\ &= \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF}, \mathbf{E}_{iC}, D_{iC} = 1) \times \text{pr}(PGS_{iM}, PGS_{iF} | \mathbf{E}_{iC}, D_{iC} = 1) \end{aligned}$$

Here,

$$\begin{aligned} L_{iC} &= \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF}, \mathbf{E}_{iC}, D_{iC} = 1) \\ &= \frac{\text{pr}(D_{iC} = 1 | PGS_{iC}, \mathbf{E}_{iC}) \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF})}{\text{pr}(D_{iC} = 1 | PGS_{iM}, PGS_{iF}, \mathbf{E}_{iC})} \\ &= \frac{\exp(\alpha_i + \beta_G PGS_{iC} + \beta_{GE}^T \mathbf{E}_{iC} PGS_{iC} + \beta_E^T \mathbf{E}_{iC}) \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF})}{\int_{-\infty}^{+\infty} \exp(\alpha_i + \beta_G x + \beta_{GE}^T \mathbf{E}_{iC} x + \beta_E^T \mathbf{E}_{iC}) f_{PGS_C}(x | PGS_{iM}, PGS_{iF}) dx} \\ &= \frac{\exp(\beta_G PGS_{iC} + \beta_{GE}^T \mathbf{E}_{iC} PGS_{iC}) \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF})}{\exp \{ \mu_{iC}(\beta_G + \beta_{GE}^T \mathbf{E}_{iC}) + \frac{1}{2} \sigma_{iC}^2 (\beta_G + \beta_{GE}^T \mathbf{E}_{iC})^2 \}} \end{aligned}$$

From (2), we have

$$L_{iP} = \frac{1}{2\pi\sigma_i^2} \exp \left\{ -\frac{\left\{ PGS_{iM} - [\mu_i + \frac{\sigma_i^2}{2}(\beta_G + \beta_{GE}^T \mathbf{E}_{iC})] \right\}^2 + \left\{ PGS_{iF} - [\mu_i + \frac{\sigma_i^2}{2}(\beta_G + \beta_{GE}^T \mathbf{E}_{iC})] \right\}^2}{2\sigma_i^2} \right\}. \quad (6)$$

1.1.2 Parameter Estimation

We observe that for fixed values of β_G , β_{GE} and $\sigma_i^2, i = 1, \dots, N$, an unbiased estimator of μ_i is given by $\hat{\mu}_i = \frac{1}{2}(PGS_{iM} + PGS_{iF}) - \frac{1}{2}\sigma_i^2(\beta_G + \beta_{GE}^T \mathbf{E}_{iC}), i = 1, \dots, N$.

Now by plugging in $\hat{\mu}_i$ to L_{iP} , we obtain the profile-likelihood:

$$L_{iP}^* = \frac{1}{2\pi\sigma_i^2} \exp \left\{ -\frac{\frac{1}{2}(PGS_{iM} - PGS_{iF})^2}{2\sigma_i^2} \right\}.$$

Therefore the profile-likelihood of each family i is of the form below

$$\begin{aligned} L_i^* &= \text{pr}(PGS_{iC}, PGS_{iM}, PGS_{iF} | \mathbf{E}_{iC}, D_{iC} = 1) \\ &= L_{iC} \times L_{iP}^* \\ &= \frac{1}{\sqrt{2\pi\sigma_{iC}^2}} \exp \left\{ -\frac{[PGS_{iC} - \mu_{iC} - \sigma_{iC}^2(\beta_G + \beta_{GE}^T \mathbf{E}_{iC})]^2}{2\sigma_{iC}^2} \right\} \times \frac{1}{2\pi\sigma_i^2} \exp \left\{ -\frac{\frac{1}{2}(PGS_{iM} - PGS_{iF})^2}{2\sigma_i^2} \right\} \end{aligned}$$

The score functions for β_G and β_{GE} are given by

$$\begin{aligned}\frac{\partial \log L^*}{\partial \beta_G} &= \sum_{i=1}^N \left(PGS_{iC} - \mu_{iC} - \frac{1}{2} \sigma_i^2 (\beta_G + \beta_{GE}^T \mathbf{E}_{iC}) \right) \\ \frac{\partial \log L^*}{\partial \beta_{GE}} &= \sum_{i=1}^N \left(PGS_{iC} \mathbf{E}_{iC} - \mu_{iC} \mathbf{E}_{iC} - \frac{1}{2} \sigma_i^2 (\beta_G + \beta_{GE}^T \mathbf{E}_{iC}) \mathbf{E}_{iC} \right),\end{aligned}$$

with $\mu_{iC} = \frac{1}{2}(PGS_{iM} + PGS_{iF})$.

We can write the solution in closed form by letting the score functions equal to 0:

$$\sum_{i=1}^N \sigma_i^2 \mathbf{E}_i \mathbf{E}_i^T \boldsymbol{\beta} = 2 \sum_{i=1}^N (PGS_{iC} - \mu_{iC}) \mathbf{E}_i,$$

where $\mathbf{E}_i = (1, \mathbf{E}_{iC}^T)^T$ and $\boldsymbol{\beta} = (\beta_G, \beta_{GE}^T)^T$. Since $\mathbf{E}_i \mathbf{E}_i^T$ is positive definite (assuming that there is no collinearity between the environmental variables), it is invertible. Therefore, the solution is

$$\hat{\boldsymbol{\beta}} = 2 \left(\sum_{i=1}^N \sigma_i^2 \mathbf{E}_i \mathbf{E}_i^T \right)^{-1} \sum_{i=1}^N (PGS_{iC} - \mu_{iC}) \mathbf{E}_i.$$

In matrix form, this is

$$\hat{\boldsymbol{\beta}} = (\mathbf{E}^T \mathbf{W} \mathbf{E})^{-1} \mathbf{E}^T \mathbf{Z},$$

where $\mathbf{E}_{N \times K} = (\mathbf{E}_1^T, \dots, \mathbf{E}_N^T)^T$, $\mathbf{W} = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$, and $\mathbf{Z} = (2(PGS_{1C} - \mu_{1C}), \dots, 2(PGS_{NC} - \mu_{NC}))^T$.

Finally, we note that in the absence of parental indirect genetic effect, from L_{iP} we can easily show that $E\{\frac{1}{2}(PGS_{iM} - PGS_{iF})^2\} = \sigma_i^2$ and thus throughout we plug in $\hat{\sigma}_i^2 = \frac{1}{2}(PGS_{iM} - PGS_{iF})^2$ for the final estimation.

1.1.3 Asymptotic Variance Estimation when Only Considering Direct PGS Effect

When we only consider the direct PGS effect β_G , the MLE for β_G can be easily derived in closed form by plugging in the method of moments estimator for σ_i^2 :

$$\hat{\beta}_G = \frac{2 \sum_{i=1}^N (PGS_{iC} - \frac{1}{2}(PGS_{iM} + PGS_{iF})) / N}{\sum_{i=1}^N \hat{\sigma}_i^2 / N}.$$

Since we assume that the within family variance σ_i^2 is finite for $i = 1, \dots, N$, then it satisfies that

$$\max_{i=1, \dots, N} \frac{\sigma_i^2}{\sum_{i=1}^N \sigma_i^2} \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Therefore by the Lindeberg-Feller central limit theorem, the numerator of the above expression converges in distribution to normal distribution, allowing for the fact that the families are independent but not identically distributed. This holds the same to the denominator since we assume σ_i^2 is finite. We denote $X := 2 \sum_{i=1}^N (PGS_{iC} - \frac{1}{2}(PGS_{iM} + PGS_{iF})) / N$ and $Y := \sum_{i=1}^N \frac{1}{2}(PGS_{iM} - PGS_{iF})^2 / N$. By definition, the variance of $f(X, Y) = \frac{X}{Y}$ is

$$\text{var}(f(X, Y)) = \mathbb{E}\{[f(X, Y) - \mathbb{E}(f(X, Y))]^2\}.$$

The first-order Taylor approximations for $f(X, Y) = \frac{X}{Y}$ around $\boldsymbol{\mu} = (\mu_x, \mu_y) = (\mathbb{E}(X), \mathbb{E}(Y))$ give

$$\mathbb{E}(f(X, Y)) \approx f(\boldsymbol{\mu}),$$

we then have

$$\begin{aligned}
\text{var}(f(X, Y)) &\approx \mathbb{E} \left\{ [f(X, Y) - f(\boldsymbol{\mu})]^2 \right\} \\
&\approx \mathbb{E} \left\{ \left[f(\boldsymbol{\mu}) + \frac{\partial f(\boldsymbol{\mu})}{\partial x} (X - \mu_x) + \frac{\partial f(\boldsymbol{\mu})}{\partial y} (Y - \mu_y) - f(\boldsymbol{\mu}) \right]^2 \right\} \\
&= \frac{1}{\mu_y^2} \text{var}(X) + \frac{\mu_x^2}{\mu_y^4} \text{var}(Y) - \frac{2\mu_x}{\mu_y^3} \text{cov}(X, Y)
\end{aligned}$$

Since we have $\mu_x = \beta_G \sum_{i=1}^N \sigma_i^2 / N$ and $\mu_y = \sum_{i=1}^N \sigma_i^2 / N$, $\text{cov}(X, Y) = 0$ since X is derived from children's probability conditional on parents, and Y is derived from parents' likelihood term. We can further derive $\text{var}(X) = 2 \sum_{i=1}^N \sigma_i^2 / N^2$ and $\text{var}(Y) = 2 \sum_{i=1}^N \sigma_i^4 / N^2$. Therefore, we have $\text{var}(\hat{\beta}_G) = \frac{2}{\sum_{i=1}^N \sigma_i^2} + \frac{2\beta_G^2 \sum_{i=1}^N \sigma_i^4}{(\sum_{i=1}^N \sigma_i^2)^2}$. For implementations, we plug-in the estimated $\hat{\sigma}_i^2$ and $\hat{\beta}_G$ and the unbiased estimator for $\hat{\sigma}_i^4 = \frac{1}{12} (PGS_{iM} - PGS_{iF})^4$ into $\text{var}(\hat{\beta}_G)$.

1.1.4 Asymptotic Variance Estimation in the General Form

The variance $\text{var}(\hat{\boldsymbol{\beta}})$ is obtained using Taylor approximations. Let $\mathbf{X} := 2 \sum_{i=1}^N (PGS_{iC} - \mu_{iC}) \mathbf{E}_i / N$ and $\mathbf{Y} := \sum_{i=1}^N \hat{\sigma}_i^2 \mathbf{E}_i \mathbf{E}_i^T / N$. Then we have $f(\mathbf{X}, \mathbf{Y}) = \mathbf{Y}^{-1} \mathbf{X}$. Similarly to section 1.1.3, we have

$$\mathbb{E}(f(\mathbf{X}, \mathbf{Y})) \approx f(\boldsymbol{\mu}),$$

with $\boldsymbol{\mu} = (\boldsymbol{\mu}_x, \mu_{y_{11}}, \dots, \mu_{y_{KK}}) = (\mathbb{E}(\mathbf{X}), \mathbb{E}(\frac{1}{N} \sum_{i=1}^N \hat{\sigma}_i^2 e_{i1}^2), \mathbb{E}(\frac{1}{N} \sum_{i=1}^N \hat{\sigma}_i^2 e_{i1} e_{i2}), \dots, \mathbb{E}(\frac{1}{N} \sum_{i=1}^N \hat{\sigma}_i^2 e_{iK}^2))$, with e_{ij} being the j -th ($j = 1, \dots, K$) environmental variable for children in family i ($i = 1, \dots, N$). Here, each $y_{hl}, h \leq l = 1, \dots, K$ is an element in the symmetric matrix \mathbf{Y} , so there are finite number of parameters in \mathbf{Y} , i.e., $K(1+K)/2, K < \infty$.

By definition,

$$\begin{aligned}
\text{cov}(f(\mathbf{X}, \mathbf{Y})) &= \mathbb{E} \left\{ [f(\mathbf{X}, \mathbf{Y}) - \mathbb{E}(f(\mathbf{X}, \mathbf{Y}))][f(\mathbf{X}, \mathbf{Y}) - \mathbb{E}(f(\mathbf{X}, \mathbf{Y}))]^T \right\} \\
&\approx \mathbb{E} \left\{ [f(\mathbf{X}, \mathbf{Y}) - f(\boldsymbol{\mu})][f(\mathbf{X}, \mathbf{Y}) - f(\boldsymbol{\mu})]^T \right\} \\
&\approx \mathbb{E} \left\{ \left[\frac{\partial f(\boldsymbol{\mu})}{\partial \mathbf{X}} (\mathbf{X} - \boldsymbol{\mu}_x) + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{11}} (y_{11} - \mu_{y_{11}}) + \dots + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{KK}} (y_{KK} - \mu_{y_{KK}}) \right] \cdot \right. \\
&\quad \left. \left[\frac{\partial f(\boldsymbol{\mu})}{\partial \mathbf{X}} (\mathbf{X} - \boldsymbol{\mu}_x) + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{11}} (y_{11} - \mu_{y_{11}}) + \dots + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{KK}} (y_{KK} - \mu_{y_{KK}}) \right]^T \right\} \\
&= \frac{\partial f(\boldsymbol{\mu})}{\partial \mathbf{X}} \text{cov}(\mathbf{X}) \left(\frac{\partial f(\boldsymbol{\mu})}{\partial \mathbf{X}} \right)^T + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{11}} \text{var}(y_{11}) \left(\frac{\partial f(\boldsymbol{\mu})}{\partial y_{11}} \right)^T + \dots + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{KK}} \text{var}(y_{KK}) \left(\frac{\partial f(\boldsymbol{\mu})}{\partial y_{KK}} \right)^T \\
&\quad + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{11}} \text{cov}(y_{11}, y_{12}) \left(\frac{\partial f(\boldsymbol{\mu})}{\partial y_{12}} \right)^T + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{12}} \text{cov}(y_{11}, y_{12}) \left(\frac{\partial f(\boldsymbol{\mu})}{\partial y_{11}} \right)^T + \dots \\
&\quad + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{K-1, K}} \text{cov}(y_{K-1, K}, y_{KK}) \left(\frac{\partial f(\boldsymbol{\mu})}{\partial y_{KK}} \right)^T + \frac{\partial f(\boldsymbol{\mu})}{\partial y_{KK}} \text{cov}(y_{K-1, K}, y_{KK}) \left(\frac{\partial f(\boldsymbol{\mu})}{\partial y_{K-1, K}} \right)^T \\
&= \boldsymbol{\mu}_y^{-1} \text{cov}(\mathbf{X}) (\boldsymbol{\mu}_y^{-1})^T + \sum_{i=1}^N \boldsymbol{\mu}_y^{-1} \mathbf{E}_i \mathbf{E}_i^T \boldsymbol{\mu}_y^{-1} \boldsymbol{\mu}_x \text{var}(\hat{\sigma}_i^2) (\boldsymbol{\mu}_y^{-1} \mathbf{E}_i \mathbf{E}_i^T \boldsymbol{\mu}_y^{-1} \boldsymbol{\mu}_x)^T
\end{aligned}$$

Since \mathbf{X} and \mathbf{Y} are independent, $cov(\mathbf{X}, y_{hl}) = 0$. We also have

$$\begin{aligned}\boldsymbol{\mu}_x &= \sum_{i=1}^N \sigma_i^2 (\boldsymbol{\beta}^T \mathbf{E}_i) \mathbf{E}_i / N = \sum_{i=1}^N \sigma_i^2 \mathbf{E}_i \mathbf{E}_i^T \boldsymbol{\beta} / N \\ \boldsymbol{\mu}_y &= \sum_{i=1}^N \sigma_i^2 \mathbf{E}_i \mathbf{E}_i^T / N \\ cov(\mathbf{X}) &= 4 \sum_{i=1}^N \mathbf{E}_i var(PGS_{iC} - \mu_{iC}) \mathbf{E}_i^T / N^2 = 2 \sum_{i=1}^N \mathbf{E}_i \mathbf{E}_i^T \sigma_i^2 / N^2 \\ var(\hat{\sigma}_i^2) &= 2\sigma_i^4 \\ \hat{\sigma}_i^4 &= \frac{1}{3}(\hat{\sigma}_i^2)^2 = \frac{1}{12}(PGS_{iM} - PGS_{iF})^4\end{aligned}$$

1.2 Derivation Incorporating Indirect Effects of Parental PGS

Suppose that the disease outcome of the offspring is also affected by indirect parental PGS effects (IDE), the disease risk model takes the form

$$\begin{aligned}\text{pr}(D_{iC} = 1 | PGS_{iC}, PGS_{iM}, PGS_{iF}, \mathbf{E}_{iC}) \\ = \exp(\alpha_i + \beta_G PGS_{iC} + \beta_M PGS_{iM} + \beta_F PGS_{iF} + \boldsymbol{\beta}_E^T \mathbf{E}_{iC} + \boldsymbol{\beta}_{GE}^T \mathbf{E}_{iC} PGS_{iC}).\end{aligned}$$

Now the likelihood for each family i becomes

$$\begin{aligned}L_i &= \text{pr}(PGS_{iC}, PGS_{iM}, PGS_{iF} | \mathbf{E}_{iC}, D_{iC} = 1) \\ &= L_{iC} \times L_{iP} \\ &= \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF}, \mathbf{E}_{iC}, D_{iC} = 1) \times \text{pr}(PGS_{iM}, PGS_{iF} | \mathbf{E}_{iC}, D_{iC} = 1) \\ &= \frac{\text{pr}(D_{iC} = 1 | PGS_{iC}, PGS_{iM}, PGS_{iF}, \mathbf{E}_{iC}) \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF})}{\text{pr}(D_{iC} = 1 | PGS_{iM}, PGS_{iF}, \mathbf{E}_{iC})} \times L_{iP} \\ &= \frac{\exp(\beta_G PGS_{iC} + \boldsymbol{\beta}_{GE}^T \mathbf{E}_{iC} PGS_{iC}) \text{pr}(PGS_{iC} | PGS_{iM}, PGS_{iF})}{\exp\{\mu_{iC}(\beta_G + \boldsymbol{\beta}_{GE}^T \mathbf{E}_{iC}) + \frac{1}{2}\sigma_{iC}^2(\beta_G + \boldsymbol{\beta}_{GE}^T \mathbf{E}_{iC})^2\}} \times L_{iP},\end{aligned}$$

with L_{iC} unchanged given that the parental effects are canceled out.

We show the below formula

$$PGS_{iM} | \mathbf{E}_{iC}, D_{iC} = 1 \sim N\left(\mu_i + \sigma_i^2\left(\beta_M + \frac{1}{2}(\beta_G + \boldsymbol{\beta}_{GE}^T \mathbf{E}_{iC})\right), \sigma_i^2\right), \quad (7)$$

$$PGS_{iF} | \mathbf{E}_{iC}, D_{iC} = 1 \sim N\left(\mu_i + \sigma_i^2\left(\beta_F + \frac{1}{2}(\beta_G + \boldsymbol{\beta}_{GE}^T \mathbf{E}_{iC})\right), \sigma_i^2\right). \quad (8)$$

We derive formula (7) here: for ease of notation, let's denote $x = PGS_{iC}$, $y = PGS_{iM}$, $z = PGS_{iF}$, $E = \mathbf{E}_{iC}$, $D = D_{iC}$. Note that y is independent of z and $f(y, z) = f(y)f(z)$.

$$\begin{aligned}\text{pr}(y, z | E, D = 1) &= \int_{\mathbb{R}} \text{pr}(x', y, z | E, D = 1) dx' \\ &= \int_{\mathbb{R}} \frac{\text{pr}(D = 1 | x', y, z, E) \text{pr}(x', y, z | E)}{\text{pr}(D = 1 | E)} dx' \\ &= \frac{\int_{\mathbb{R}} \text{pr}(D = 1 | x', y, z, E) \text{pr}(x', y, z | E) dx'}{\int \int \int_{\mathbb{R}} \text{pr}(D = 1 | x'', y', z', E) f(x'', y', z') dx'' dy' dz'} \\ &= \frac{\int_{\mathbb{R}} \exp(\beta_G x' + \boldsymbol{\beta}_{GE}^T E x' + \beta_M y + \beta_F z) f(x', y, z) dx'}{\int \int \int_{\mathbb{R}} \exp(\beta_G x'' + \boldsymbol{\beta}_{GE}^T E x'' + \beta_M y' + \beta_F z') f(x'', y', z') dx'' dy' dz'}\end{aligned}$$

Here,

$$\begin{aligned} \text{numerator} &= \int \exp(\beta_G x' + \beta_{GE} E x' + \beta_M y + \beta_F z) f(x'|y, z) dx' f(y) f(z) \\ &= \exp\left\{\frac{1}{4}\sigma_i^2(\beta_G + \beta_{GE}E)^2 + \frac{1}{2}(y+z)(\beta_G + \beta_{GE}E)\right\} \exp(\beta_M y + \beta_F z) f(y) f(z) \end{aligned}$$

and

$$\begin{aligned} \text{denominator} &= \exp\left\{\frac{1}{2}\sigma_i^2\left(\beta_M + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)^2 + \mu_i\left(\beta_M + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)\right\} \\ &\quad \times \exp\left\{\frac{1}{2}\sigma_i^2\left(\beta_F + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)^2 + \mu_i\left(\beta_F + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)\right\} \\ &\quad \times \exp\left\{\frac{1}{4}\sigma_i^2(\beta_G + \beta_{GE}E)^2\right\}, \end{aligned}$$

Then we have

$$\begin{aligned} \text{pr}(y, z|E, D=1) &= \frac{\exp\left\{\left[\frac{1}{2}(\beta_G + \beta_{GE}E) + \beta_M\right]y\right\} f(y)}{\exp\left\{\frac{1}{2}\sigma_i^2\left(\beta_M + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)^2 + \mu_i\left(\beta_M + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)\right\}} \\ &\quad \times \frac{\exp\left\{\left[\frac{1}{2}(\beta_G + \beta_{GE}E) + \beta_F\right]z\right\} f(z)}{\exp\left\{\frac{1}{2}\sigma_i^2\left(\beta_F + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)^2 + \mu_i\left(\beta_F + \frac{1}{2}(\beta_G + \beta_{GE}E)\right)\right\}}, \end{aligned}$$

by plugging in $f(y)$ and $f(z)$ ($y, z \sim N(\mu_i, \sigma_i^2)$), we can easily show formula (7).

Therefore we have

$$L_{iP} = \frac{1}{2\pi\sigma_i^2} \exp\left\{-\frac{\{PGS_{iM} - [\mu_i + \sigma_i^2(\beta_M + \frac{1}{2}(\beta_G + \beta_{GE}^T \mathbf{E}_{iC})]\})^2 + \{PGS_{iF} - [\mu_i + \sigma_i^2(\beta_F + \frac{1}{2}(\beta_G + \beta_{GE}^T \mathbf{E}_{iC})]\})^2}{2\sigma_i^2}\right\}.$$

1.2.1 Derivation of Estimates of Parental Indirect Genetic Effects

To estimate β_M and β_F , we try to estimate them using the information in L_{iP} alone. We construct a new random variable

$$X_i = PGS_{iM} - PGS_{iF}$$

Denote $\delta_{MF} = \beta_M - \beta_F$, we have $X_i \sim N(\delta_{MF}\sigma_i^2, 2\sigma_i^2)$ and $\sum_{i=1}^N X_i \sim N(\delta_{MF}\sum_{i=1}^N \sigma_i^2, 2\sum_{i=1}^N \sigma_i^2)$, and thus we have

$$\mathbb{E}(\bar{X}) = \frac{\delta_{MF}\sum_{i=1}^N \sigma_i^2}{N}.$$

Therefore we have the estimator

$$\hat{\delta}_{MF} = \frac{\sum_{i=1}^N X_i/N}{\sum_{i=1}^N \sigma_i^2/N} = \frac{\bar{X}}{\sum_{i=1}^N \sigma_i^2/N}.$$

1.2.2 Approximate Estimator of the Required Scale Factor

We can further derive the expectation of sample variance of $\tilde{X} = (X_1, X_2, \dots, X_N)$ using information in L_{iP} alone:

$$\mathbb{E}\left(\frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X})^2\right) = \frac{1}{N-1}\mathbb{E}\left(\sum_{i=1}^N X_i^2 - \frac{(\sum_{i=1}^N X_i)^2}{N}\right) \quad (9)$$

$$= \frac{1}{N}\sum_{i=1}^N 2\sigma_i^2 + \frac{\delta_{MF}^2}{N-1}\left(\sum_{i=1}^N \sigma_i^4 - \frac{1}{N}\left(\sum_{i=1}^N \sigma_i^2\right)^2\right), \quad (10)$$

we observe the second term in the above formula will be close to zero if either δ_{MF} is small or the variability of σ_i^2 across families is small, or both. Therefore, we can use $\sum_{i=1}^N (X_i - \bar{X})^2 / [2(N-1)]$ as an approximately unbiased estimator of $\sigma_{sum}^2 = \sum_{i=1}^N \sigma_i^2 / N$, and thus we estimate δ_{MF} as

$$\hat{\delta}_{MF} = \frac{\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2 / [2(N-1)]}.$$

We observe that in the presence of indirect effects, the maximum likelihood estimate of $\beta = (\beta_G, \beta_{GE}^T)^T$ obtained from L_{iC} remain unchanged and takes the form $\hat{\beta} = (\mathbf{E}^T \mathbf{W} \mathbf{E})^{-1} \mathbf{E}^T \mathbf{Z}$, where $\mathbf{W} = \mathbf{diag}(\sigma_1^2, \dots, \sigma_N^2)$. As \mathbf{W} is unknown, we need to consider estimation of the elements of $\mathbf{E}^T \mathbf{W} \mathbf{E}$, which are of the form $\sum_{i=1}^N \sigma_i^2 e_{ik} e_{ik'}$, for $k = 1, \dots, K; k' = 1, \dots, K$. We observe that for any set of $v_i, i = 1, \dots, N$, where $v_i = e_{ik} e_{ik'}$ for some k and k' , the expectation of the weighted sample variance is

$$\begin{aligned} \mathbb{E} \left(\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 v_i \right) &= \frac{1}{N-1} \sum_{i=1}^N \mathbb{E} \left(X_i^2 v_i + \bar{X}^2 v_i - 2X_i \bar{X} v_i \right) \\ &= \frac{1}{N-1} \left(\sum_{i=1}^N 2\sigma_i^2 v_i - \frac{4}{N} \sum_{i=1}^N \sigma_i^2 v_i + \frac{2}{N^2} \sum_{i=1}^N \sigma_i^2 \sum_{i=1}^N v_i \right) + \frac{\delta_{MF}^2}{N-1} \sum_{i=1}^N (\sigma_i^2 - \frac{1}{N} \sum_{i=1}^N \sigma_i^2)^2 v_i \end{aligned}$$

Note that when $v_i = 1$, the above equation is the same as (9). Under the condition $\sum_{i=1}^N v_i = O(N)$,

$$\mathbb{E} \left(\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 v_i \right) \approx \frac{1}{N} \sum_{i=1}^N 2\sigma_i^2 v_i + \frac{\delta_{MF}^2}{N-1} \sum_{i=1}^N (\sigma_i^2 - \frac{1}{N} \sum_{i=1}^N \sigma_i^2)^2 v_i.$$

Above, we again note that if δ_{MF} is small or σ_i^2 across families are relatively constant, or both, the last term is expected to be negligible. Thus, we propose using $\sum_{i=1}^N (X_i - \bar{X})^2 v_i / [2(N-1)]$ as an approximate unbiased estimator for $\sum_{i=1}^N \sigma_i^2 v_i / N$.

1.2.3 Asymptotic Variance Estimation

Let $A = \bar{X} = \sum_{i=1}^N (PGS_{iM} - PGS_{iF}) / N$ and $B = \sum_{i=1}^N (X_i - \bar{X})^2 / [2(N-1)]$. The variance of $\hat{\delta}_{MF}$ can be approximated using first-order Taylor expansion for $\delta_{MF} = f(A, B) = A/B$ around $(\delta_{MF} \sum_{i=1}^N \sigma_i^2 / N, \sum_{i=1}^N \sigma_i^2 / N)$:

$$var(\hat{\delta}_{MF}) = \frac{2}{\sigma_{sum}^2} + \frac{2\delta_{MF}^2 \sum_{i=1}^N \sigma_i^4}{(\sum_{i=1}^N \sigma_i^2)^2}.$$

Note that the numerator A and the denominator B in $\hat{\delta}_{MF}$ are independent to each other (mean and sample variance of normal distribution). For small δ_{MF} , we can further approximate the variance formula as

$$var(\hat{\delta}_{MF}) = \frac{2}{\sigma_{sum}^2}.$$

Based on the above formula, we can obtain variance estimators by plugging in values for $\hat{\sigma}_{sum}^2, \hat{\delta}_{MF}$, and $\sum_{i=1}^N \hat{\sigma}_i^4 = \frac{N^2}{12(N-1)^2} \sum_{i=1}^N (X_i - \bar{X})^4$.

2 Details of Simulation Studies

2.1 Simulate PGS Values Based on the Assumed Model

We directly simulated PGS values for 1,300,000 parent-child trios in the population using a multivariate normal distribution as shown in (1) and prospectively simulated disease status in the children based on a logistic risk model with disease prevalence of 1%. We allow each family i to have specific mean μ_i and

variance σ_i^2 . We let σ_i^2 follow a mixture of 3-component gamma distributions, reflecting fluctuated variances of PGS values in different families, with the parameters

$$\sigma_i^2 \sim 0.6\Gamma(6, 15) + 0.3\Gamma(15, 30) + 0.1\Gamma(60, 150),$$

so that the mean of σ_i^2 is approximately 0.4 (this value is close to the population variance of 313-SNP PGS related to breast cancer in UK biobank). Furthermore, we simulated family-specific disease-risk parameters by using a model of the form $\alpha_i \sim N(\alpha + \rho_G \mu_i, 1)$. We varied $\rho_G = \text{cor}(\alpha_i, \mu_i)$ to create different scenarios of population-stratification bias, with a value of 0 indicating no relationship between variation in disease risk and PGS distribution across underlying substructure, a scenario where one would not expect any effect of underlying population substructure in creating spurious associations between disease risk and PGS at the population level. We selected different numbers of case-parent trios ($N=200, 500, 1000, 2000$) from a random sample of families by restricting to those families where the children were cases ($D_{iC} = 1$) in the population. The number of trios in the simulation study corresponded to the various sample sizes of different populations in the GENEVA study and the SPARK consortium. We compared PGS-TRI with the pTDT test for the performance of the PGS main effect. We also compared the family-based methods with the performance of population-based case-control studies by randomly sampling unrelated disease-free children from the same simulated family-based population. For the parental indirect effect difference, we included different magnitudes of maternal effects and no paternal effect in the underlying disease risk model and evaluated our model’s performance.

For the investigation of the performance of the proposed method for the estimation of gene-environment interaction parameters, for each family, we simulated a binary variable E_1 and a continuous variable E_2 independent of the underlying PGS values for all three family members. We assume a latent continuous variable S_1 for binary E_1 , and allow S_1 and E_2 to have family-specific mean values $\gamma_{i1}, \gamma_{i2} \sim N(0, 1)$. The mean distribution of family-specific random effect term $\alpha_i \sim N(\alpha + \rho_G \mu_i + \rho_{GE} \mu_i \gamma_{i1} + \rho_{GE} \mu_i \gamma_{i2}, 1)$ indicates the effect of E and PGS differ by population substructures, with an underlying disease risk model incorporating PGS-environment interaction terms. We compared PGS-TRI with the population-based case-only method to assess the performance of the PGS-E interaction terms. We further let γ_{i1} and γ_{i2} to co-vary systematically with μ_i following $\text{cor}(\mu_i, \gamma_{ij}) \sim \text{uniform}(0, 0.5), j = 1, 2$ to allow for potential population-level correlations between PGS and E due to the effect of population stratification and assortative mating.

2.2 Simulation using the UK Biobank Data

To create realistic population substructures, we simulated offspring genotypes conditional on pairs of independent individuals’ genotypes of British white ancestry using the UK Biobank (UKB) genotype data. Each pair was matched within the same assessment centre (UKB Field ID: 54), based on the individual’s propensity score generated from the place of birth north and east co-ordinates (UKB Field ID: 129 and 130). Specifically, we used the nearest available Mahalanobis metric matching within 0.1 calipers defined by the propensity score. To further assess the model performances under assortative mating of a single trait, we performed a separate set of matching based on educational attainment (EA) (UKB Field ID: 6138) in addition to the geographical regions. It has been previously reported that EA is a common trait in assortative mating and the EA-PGS are heavily confounded by geographical regions. We built EA-PGS using independent SNPs ($R^2 < 0.01$ within 1000kb) and weights reported in previous work (PGS Catalog ID: PGS002012). We prospectively simulated disease status in the 150 253 unrelated children based on a logistic risk model with a disease prevalence of 2%. We let the intercept term $\alpha_i \sim N(\alpha + \rho_G \text{BMI}, 1)$, where BMI values are the baseline values of the mothers in each independent simulated family in UKB (Field ID: 21001). We then compared PGS-TRI with the pTDT test, logistic regression of unrelated individuals, logistic regression adjusting for top 10 genetic principal components (PCs), and additionally adjusting for birth locations and assessment centres for the performance of the PGS main effect by selecting different numbers of case-parent trios ($N = 1000, 2000$) from a random sample of families and the same number of unrelated random controls for the comparisons with logistic regression.

We further grouped the parents into 100 clusters based on their east and north co-ordinates of birthplaces using the K-means clustering. We observed significant correlations ($\text{cor} = -0.48$) between BMI and EA-PGS between clusters but not within clusters ($\text{cor} = -0.018$). This demonstrated population structure and BMI as the hidden confounding variable which affected the random intercept term in disease risk. We

reached the same observations as our first simulation results for PGS main effects. We found that the adjustments of PCs and geographical regions in unrelated individuals showed an improvement compared with logistic regression alone. However, there are still residual biases in unrelated logistic regressions after population substructure adjustments, due to assortative mating, non-linear effects from population structures and geographical regions. Further, PGS-TRI had similar efficiency as logistic regression model adjusted for multiple covariates. We demonstrated that PGS-TRI remained the most unbiased method and produced correct type I error rates.

3 Details of the Data Applications

3.1 Data Analyses of Autism Spectrum Disorder (ASD) in the SPARK Study

3.1.1 Genotype Data Preprocessing and PGS Construction

We analyzed case-parent trio data from the Simons Foundation Powering Autism Research for Knowledge (SPARK) study¹. The genotype phasing and imputation followed previous research². Specifically, the imputation was performed on the Michigan imputation server using the Trans-Omics for Precision Medicine (TOPMed) Freeze 5b reference panel, which consisted of 125,568 haplotypes from multi-ancestry population. The SNPs with imputation quality $R^2 < 0.8$, missing call rates $> 1\%$, minor allele frequencies (MAF) $< 1\%$ were excluded.

The study population comprises 5 genetic ancestral groups determined using the HapMap3 reference panel: African (AFR), Americas (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). After eliminating trios with missing parent or shared parent(s), a total of 1,250 EUR, 63 AFR, 153 AMR, 23 SAS, and 28 EAS independent case-parent trios with available imputed genotype data were used for subsequent analysis. The PGS scores were constructed using 28 017 SNPs, with ambiguous SNPs removed, and their associated weights provided by recent external GWAS, as reported in the PGS catalog. To enable fair comparisons of estimated effect sizes across diverse ancestral populations, we further standardized the PGS values within each ancestry group using 1000 Genomes (1000G) Phase 3 healthy individuals by constructing ASD PGS scores following the same procedures. In particular, these included 503 EAS, 498 EUR, 487 SAS, 659 AFR, and 347 AMR independent individuals from the 1000 Genomes (1000G) Phase 3 Project³.

Variables we considered for the PGS-E interaction effects included maternal variables before pregnancy, including asthma, depression and other severe mental illness (defined as requiring medication or hospitalization), vitamin intake 3 months before pregnancy; variables during pregnancy including whether the mother experienced fever, eclampsia and preeclampsia, gestational diabetes, hyperemesis, pre-term or early labor; variables both before and during pregnancy including alcohol consumption and frequency, and smoking status; mother’s age at birth, mother’s educational attainment levels, and whether the child had low birth weight (defined as $< 2.5\text{kg}$). Mother’s pregnancy period was defined as from 3 months before pregnancy to the end of breastfeeding.

We further used our model and the individual-level genotype data of ASD case-parent trios to estimate the ASD risk associated with several common polygenic predictors of cognitive-related traits and diseases, including education attainment⁴, schizophrenia⁵, strictly defined lifetime major depressive disorder⁶, bipolar disorder⁷, neuroticism⁴, sleeplessness/insomnia⁴, and attention-deficit/hyperactivity disorder (ADHD)⁸. As a negative control, we used body mass index (BMI)⁴. We constructed PGS using reported GWAS summary statistics of these traits, following the same procedures as described above. In particular, for ADHD, we used PRScs⁹ and GWAS summary statistics following the steps reported previously⁸ to construct PGS in our study. We normalized the PGS scores within each ancestral group for fair comparisons.

3.2 Data Analyses of Non-Syndromic Orofacial Clefts (OFCs) in the GENEVA Study

3.2.1 Genotype Data Preprocessing and PGS Construction

We investigated OFCs using case-parent trio data from the Gene Environment Association Studies initiative (GENEVA)¹⁰. GENEVA is a multi-ethnic study with data collected from Europe (Norway), the United States, and Asia (China, South Korea, Singapore, and the Philippines). Detailed genotype data imputation

and quality control steps are described in previous work¹¹. We additionally excluded SNPs with MAF < 1% and missing call rates > 1%. After eliminating ambiguous SNPs, PGS values associated with cleft lip with or without cleft palate (CL/P) were computed using 24 SNPs and their respective weights sourced from the PGS catalog. These weights were based on summary statistics derived from multiple preceding GWAS studies. Note that due to the restricted data resources available in previous GWAS efforts, certain SNPs uncovered in specific studies either conducted meta-analyses or used data that partially intersected with samples in the GENEVA study. Nonetheless, these SNPs underwent subsequent validation using separate and independent data sources. We standardized the PGS values within each specific ancestry group using 503 EAS and 498 EUR independent individuals from the 1000G project.

For PGSxE interaction analysis, the maternal environmental exposures were collected through maternal interviews focused on the period from 3 months before pregnancy through the first trimester, which includes the first 8-9 weeks of gestation when palatal development is completed. The difference between maternal and paternal indirect PGS effects was also analyzed using our model. In the end, our analysis incorporated independent and complete 575 self-reported EUR and 891 Asian ancestry CL/P case-parent trios, and 203 EUR and 235 Asian CP case-parent trios.

3.3 Association Studies of Genetically Predicted Multi-Omics Data on ASD and OFCs

We first built genetic scores for 12 539 whole blood gene expression levels and 140 serum metabolomic traits using summary statistics reported in OMICSPRED. We then excluded biomolecular traits with variance explained $R^2 < 0.1$ by the genetic score in the internal validation or traits that contain fewer than 5 SNPs in the genetic score reported by OMICSPRED. So in the end we analyzed 27 metabolomics traits and 4 991 genes using PGS-TRI. All omics summary statistics were trained based on the INTERVAL EUR healthy blood cohort, and validated using multiple independent studies consisting of multi-ancestry populations. Specifically, the RNAseq summary statistics were trained based on the Illumina RNAseq platform using 4 136 individuals, and metabolomics summary statistics were based on the Nightingale platform trained using 37 359 individuals. Subsequently, we used PGS-TRI to conduct transcriptome-wide association studies and metabolome-wide association studies respectively to understand the potential molecular causal effects on ASD and OFCs risks using data from case-parent trios in SPARK and GENEVA studies. For each omics data type in each disease association study, we used the Benjamini-Hochberg false discovery rate (FDR) of 5% for multiple hypotheses testing adjustments.

References

- [1] Pamela Feliciano, Amy M Daniels, LeeAnne Green Snyder, Amy Beaumont, Alexies Camba, Amy Esler, Amanda G Gulsrud, Andrew Mason, Anibal Gutierrez, Amy Nicholson, et al. Spark: A us cohort of 50,000 families to accelerate autism research. *Neuron*, 97(3):488–493, 2018.
- [2] Nana Matoba, Dan Liang, Huaigu Sun, Nil Aygün, Jessica C McAfee, Jessica E Davis, Laura M Raffield, Huijun Qian, Joseph Piven, Yun Li, et al. Common genetic risk variants identified in the spark cohort support *ddhd2* as a candidate risk gene for autism. *Translational psychiatry*, 10(1):265, 2020.
- [3] 1000 Genomes Project Consortium Corresponding authors Auton Adam adam. auton@ gmail. com 1 b Abecasis Gonçalo R. goncalo@ umich. edu 2 c, Production group Baylor College of Medicine Gibbs Richard A.(Principal Investigator) 14 Boerwinkle Eric 14 Doddapaneni Harsha 14 Han Yi 14 Korchina Viktoriya 14 Kovar Christie 14 Lee Sandra 14 Muzny Donna 14 Reid Jeffrey G. 14 Zhu Yiming 14, Broad Institute of MIT, Harvard Lander Eric S.(Principal Investigator) 13 Altshuler David M. 3 Gabriel Stacey B.(Co-Chair) 13 Gupta Namrata 13, Coriell Institute for Medical Research Gharani Neda 31 Toji Lorraine H. 31 Gerry Norman P. 31 Resch Alissa M. 31, Illumina Bentley David R.(Principal Investigator) 5 Grocock Russell 5 Humphray Sean 5 James Terena 5 Kingsbury Zoya 5, McDonnell Genome Institute at Washington University Mardis Elaine R.(Co-Principal Investigator)(Co-Chair) 22 Wilson Richard K.(Co-Principal Investigator) 22 Fulton Lucinda 22 Fulton Robert 22, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.
- [4] Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F O’Reilly, and Bjarni J Vilhjálmsson. Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109(1):12–23, 2022.
- [5] Amanda B Zheutlin, Jessica Dennis, Richard Karlsson Linnér, Arden Moscati, Nicole Restrepo, Peter Straub, Douglas Ruderfer, Victor M Castro, Chia-Yen Chen, Tian Ge, et al. Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four health care systems. *American Journal of Psychiatry*, 176(10):846–855, 2019.
- [6] Na Cai, Joana A Revez, Mark J Adams, Till FM Andlauer, Gerome Breen, Enda M Byrne, Toni-Kim Clarke, Andreas J Forstner, Hans J Grabe, Steven P Hamilton, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nature genetics*, 52(4):437–447, 2020.
- [7] Yuanyuan Gui, Xiaocheng Zhou, Zixin Wang, Yiliang Zhang, Zhaobin Wang, Geyu Zhou, Yize Zhao, Manhua Liu, Hui Lu, and Hongyu Zhao. Sex-specific genetic association between psychiatric disorders and cognition, behavior and brain imaging in children and adults. *Translational Psychiatry*, 12(1):347, 2022.
- [8] Ditte Demontis, G Bragi Walters, Georgios Athanasiadis, Raymond Walters, Karen Therrien, Trine Tollerup Nielsen, Leila Farajzadeh, Georgios Voloudakis, Jaroslav Bendl, Biau Zeng, et al. Genome-wide analyses of *adhd* identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nature genetics*, 55(2):198–208, 2023.
- [9] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications*, 10(1):1776, 2019.
- [10] Terri H Beaty, Jeffrey C Murray, Mary L Marazita, Ronald G Munger, Ingo Ruczinski, Jacqueline B Hetmanski, Kung Yee Liang, Tao Wu, Tanda Murray, M Daniele Fallin, et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near *mafb* and *abca4*. *Nature genetics*, 42(6):525–529, 2010.
- [11] Wanying Zhang, Sowmya Venkataraghavan, Jacqueline B Hetmanski, Elizabeth J Leslie, Mary L Marazita, Eleanor Feingold, Seth M Weinberg, Ingo Ruczinski, Margaret A Taub, Alan F Scott,

et al. Detecting gene-environment interaction for maternal exposures using case-parent trios ascertained through a case with non-syndromic orofacial cleft. *Frontiers in cell and developmental biology*, 9:621018, 2021.