Within-Family GWAS

ESSGN Workshop March 2026

Sjoerd van Alten

Vrije Universiteit Amsterdam

March 26, 2026

Introduction

- GWAS estimates the association of each SNP with a phenotype y, conditional on some controls
- $y = \beta_0 + \beta_1 g_i + \sum_{k=1}^{k=10} \gamma_{ik} PC_i + X_i' \delta + \varepsilon_i$
- Ideally, we would like to think of such associations as the effect of (hypothetically) altering some part of the genome on the outcome y
- However, in practice, associations can also be driven by population stratification, assortative mating, genetic effects from relatives (IGEs)

Introduction

- Within-family GWAS to estimate direct genetic effects (DGEs)
- Random segregation during meiosis results in a natural experiment, as offspring genotype g_i randomly varies around the expectation, determined by parental genotype $(g_{m,i}, g_{p,i})$
  - Genetic differences among biological siblings are as good as random, because they play the same genetic lottery at conception
  - Alternatively, after controlling for the genome of one's parents, the remaining genetic variation in the proband is as good as random
- But within-family GWAS comes with its own challenges:
  - Unbiased estimator, but loss of power (bias-variance tradeoff)
  - Summary statistics need to be handled differently in some post-GWAS analysis
  - Indirect genetic effects from siblings might still be an issue
  - Some biases remain

Outline

Literature

Within-family GWAS studies:

- Howe, L. J., Nivard, M. G., Morris, T. T., Hansen, A. F., Rasheed, H., Cho, Y., ... & Davies, N. M. (2022). Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. Nature genetics, 54(5), 581-592.
- Tan, T., Jayashankar, H., Guan, J., Nehzati, S. M., Mir, M., Bennett, M., ... & LifeLines Cohort Study. (2025). Family-GWAS reveals effects of environment and mating on genetic associations. MedRxiv, 2024-10.

Methods:

- Young, A. I., Nehzati, S. M., Benonisdottir, S., Okbay, A., Jayashankar, H., Lee, C., ... & Kong, A. (2022). Mendelian imputation of parental genotypes improves estimates of direct genetic effects. Nature genetics, 54(6), 897-905.
- Benjamin, D. J., Cesarini, D., Turley, P., & Young, A. S. (2024). Social-science genomics: Progress, challenges, and future directions.
- Veller, C., & Coop, G. M. (2024). Interpreting population-and family-based genome-wide association studies in the presence of confounding. PLoS biology, 22(4), e3002511.

Theory

Theory

We define causal effects according to Rubin's (1974) counterfactual framework

- Let i's genotype at genetic variant j be $x_{ij} \in \{0, 1, 2\}$
- Such that i's genome be characterized as $\mathbf{x}_i = (x_{i1}, x_{i2}, ...x_{iJ})$
- The average causal effect in the population of changing from genotype $\mathbf{x}$ to $\mathbf{x}'$ is (Benjamin et al., 2024):

$$\mathbb{E}[y_i(\mathbf{x}') - y_i(\mathbf{x})]$$

- We remain agnostic about mechanisms: active and evocative rGE might be part of the causal effect

Theory

- Consider the regression

$$Y_{ij} = \delta g_{ij} + \alpha_p g_{p(i)} + \alpha_m g_{m(i)} + \varepsilon_{ij} \tag{1}$$

- δ has a causal interpretation: the change in $E(Y)$ due to a hypothetical increase of g_{ij} *or a correlated locus* by 1 *at conception*: a direct genetic effect
- Estimates of α_p and α_m do not have a causal interpretation: they can reflect IGEs of the father and mother, but also bias due to assortative mating or population stratification (sometimes called dynastic effects).
- In pop. GWAS, we do not control for $g_{p(i)}, g_{m(i)}$, and $E(\hat{\beta}) = \delta + \frac{\alpha_p + \alpha_m}{2}$

Family GWAS: Unbiased Estimation

Equation (1) can easily be estimated in data with offspring, father and mother genotyped (Trios)

- Such data is scarce. Examples are:
  - Norwegian, Mother, Father and Child Cohort (MoBa), $\sim$ 30k trios
  - ALSPAC, $\sim$ 3k trios
  - Millenium Cohort Study, $\sim$ 3k trios

Data on genotyped siblings more widely available

- Examples are:
  - Norwegian, Mother, Father and Child Cohort (MoBa), $\sim$ 80k siblings
  - UK Biobank $\sim$ 40k siblings
  - Lifelines $\sim$ 30k siblings

Within-Sibling Estimation

(1) Write model for both siblings i and i' in family j

$$Y_{ij} = \delta\, g_{ij} + \alpha_p\, g_{p(i)} + \alpha_m\, g_{m(i)} + \varepsilon_{ij}$$
$$Y_{i'j} = \delta\, g_{i'j} + \alpha_p\, g_{p(i')} + \alpha_m\, g_{m(i')} + \varepsilon_{i'j}$$

(2) siblings share both parents, so $g_{p(i)} = g_{p(i')}$ and $g_{m(i)} = g_{m(i')}$

$$Y_{ij} - Y_{i'j} = \delta(g_{ij} - g_{i'j}) + \alpha_p \underbrace{(g_{p(i)} - g_{p(i')})}_{=0} + \alpha_m \underbrace{(g_{m(i)} - g_{m(i')})}_{=0} + (\varepsilon_{ij} - \varepsilon_{i'j})$$

(3) Parental genotypes cancel ✓

$$Y_{ij} - Y_{i'j} = \delta(g_{ij} - g_{i'j}) + (\varepsilon_{ij} - \varepsilon_{i'j})$$

(4) OLS on differenced data ⇒ unbiased

$$E(\hat{\delta}) = \delta \qquad \text{since} \quad E\big[(g_{ij} - g_{i'j})(\varepsilon_{ij} - \varepsilon_{i'j})\big] = 0$$

Within-Family Estimators I: Family Fixed Effect

Model: Add a family-specific intercept μ_j to absorb all family-level confounders (parental genotypes, shared environment):

$$Y_{ij} = \delta\, g_{ij} + \mu_j + \varepsilon_{ij}$$

Define: $\tilde{g}_{ij} = g_{ij} - \bar{g}_j, \qquad \tilde{Y}_{ij} = Y_{ij} - \bar{Y}_j$ where $\bar{g}_j = \frac{1}{n_j}\sum_i g_{ij}$ and $\bar{Y}_j = \frac{1}{n_j}\sum_i Y_{ij}$ are family means.

Estimator: OLS on the full model with family dummies. By the Frisch-Waugh-Lovell theorem, this is equivalent to OLS of $\tilde{Y}_{ij}$ on $\tilde{g}_{ij}$, such that

$$\hat{\delta}_{\mathrm{FE}} = \frac{\sum_j \sum_i \tilde{g}_{ij}\, \tilde{Y}_{ij}}{\sum_j \sum_i \tilde{g}_{ij}^2}$$

and $E(\hat{\delta}_{FE}) = \delta$

works, but has strong computational burden

Within-Family Estimators II: Within-Group Demeaning

Procedure: Subtract the family mean from Y_{ij} and g_{ij}, then regress demeaned outcomes on demeaned genotypes:

$$\tilde{Y}_{ij} = \delta\, \tilde{g}_{ij} + \tilde{\varepsilon}_{ij}$$

Family-level terms drop out in the demeaning step. For any family-level confounder μ_j:

$$\widetilde{(\mu_j)} = \mu_j - \bar{\mu}_j = \mu_j - \mu_j = 0$$

Estimator:

$$\hat{\delta}_{\mathrm{DM}} = \frac{\sum_j \sum_i \tilde{g}_{ij}\, \tilde{Y}_{ij}}{\sum_j \sum_i \tilde{g}_{ij}^2} = \hat{\delta}_{FE}$$

Standard error must use $N - J - 1$ degrees of freedom, with J the number of families

Within-Family Estimators III: First Differences

$$\tilde{g}_{ij} = g_{ij} - \bar{g}_j, \qquad \tilde{Y}_{ij} = Y_{ij} - \bar{Y}_j$$

Procedure: For each sibling pair (i, i') in family j, subtract one sibling from the other:

$$\Delta Y_j = Y_{ij} - Y_{i'j}, \qquad \Delta g_j = g_{ij} - g_{i'j}$$

Family-level terms cancel: $\mu_j - \mu_j = 0$. Regress ΔY_j on Δg_j (no intercept):

$$\Delta Y_j = \delta \, \Delta g_j + \Delta \varepsilon_j$$

Estimator:

$$\hat{\delta}_{\mathrm{FD}} = \frac{\sum_j \Delta g_j \, \Delta Y_j}{\sum_j \Delta g_j^2}$$

If we have exactly two siblings in each family, then $\hat{\delta}_{\mathrm{FD}} = \hat{\delta}_{\mathrm{FE}}$

Within-Family Estimators IV: Mundlak

Decomposition of genotype:

$$g_{ij} = (g_{ij} - \bar{g}_j) + \bar{g}_j \qquad \Rightarrow \qquad g_{ij} = \tilde{g}_{ij} + \bar{g}_j$$

Mundlak specification:

$$Y_{ij} = \delta\, \tilde{g}_{ij} + \eta\, \bar{g}_j + \varepsilon_{ij}$$

- $\tilde{g}_{ij} = g_{ij} - \bar{g}_j$: within-family (Mendelian) variation
- $\bar{g}_j$: between-family variation (captures parental genotypes, stratification, etc.)

Then:

$$\hat{\delta}_{\text{Mundlak}} = \hat{\delta}_{FE}$$

- No need to include family dummies or demean Y_{ij}

Interpretation

δ is identified purely from within-family genetic differences. γ captures between-family associations (subject to confounding).

Bias Under Sibling Indirect Genetic Effects (IGE)

$$Y_{ij} = \delta g_{ij} + \alpha_p g_{p(i)} + \alpha_m g_{m(i)} + \gamma g_{i'j} + \varepsilon_{ij} \qquad \text{(and symmetrically for } Y_{i'j})$$

(1) First difference $Y_{ij} - Y_{i'j}$

$$Y_{ij} - Y_{i'j} = \delta(g_{ij} - g_{i'j}) + \gamma(g_{i'j} - g_{ij}) + (\varepsilon_{ij} - \varepsilon_{i'j}) \qquad [g_p g_m \text{ cancel as before}]$$

(2) Re-write: sibling IGE term has opposite sign to Δg_j

$$\Delta Y_j = \delta\ \Delta g_j - \gamma\ \Delta g_j + \Delta\varepsilon_j = (\delta - \gamma)\ \Delta g_j + \Delta\varepsilon_j$$

(3) OLS estimand on differenced data $\rightarrow$ biased

$$E(\hat{\delta}) = \delta - \gamma\ \neq\ \delta \qquad (\text{unless } \gamma = 0)$$

Note that estimating the following regression results in unbiased δ (but biased α_m, α_p)

$$Y_{ij} = \delta g_{ij} + \alpha_p g_{p(i)} + \alpha_m g_{m(i)} + \varepsilon_{ij}$$

Within-family estimation: Notes and caveats

Siblings share identical parental genotypes $g_{p(\cdot)}$ and $g_{m(\cdot)}$ — these cancel in the within-family comparison, any biases due to population stratification bias, indirect genetic effects, and assortative mating are controlled for

Caveats:

- Estimates of δ are still biased in the presence of sibling indirect genetic effects ($\gamma \neq 0$)
- Reduced power: In general, $Var(G_{ij}|g_{p(i)}, g_{m(i)}) = \frac{1}{2}Var(G_{ij})$
- All our proofs rely on the *linearity of the data-generating process.*
  - Fixed effects estimation is in general biased for non-linear estimators.
  - For example, when y is binary and you are using logistic regression, conditioning on family dummies does **not** guarantee unbiasedness. Use linear probability models instead

FGWAS I - Howe et al., (2022)

Howe et al., (2022)

- First large-scale GWAS on within-family variation
- 178,086 siblings, 77,832 sibships, 19 cohorts (UK Biobank, HUNT, QIMR, ...), 25 phenotypes
- compares within-family and population GWAS
- for individual i in family j:
  - **Population GWAS:** $y_{ij} \sim g_{ij} + Sex_{ij} + Age_{ij} + \sum_{k=1}^{20} PC_{ij}^{k}$
  - **Within-sib GWAS:** $y_{ij} \sim \tilde{g}_{ij} + \bar{g}_{j} + Sex_{ij} + Age_{ij} + \sum_{k=1}^{20} PC_{ij}^{k}$
  - Standard errors clustered by sibship
- European Ancestry, except for China Kadoorie Biobank
- Meta-analysis in METAL, like any other GWAS

Shrinkage of Within-sib effect vs. pop. GWAS effects

- Create a *Summary-based score* of most significant SNPs for both pop. and WF GWAS:

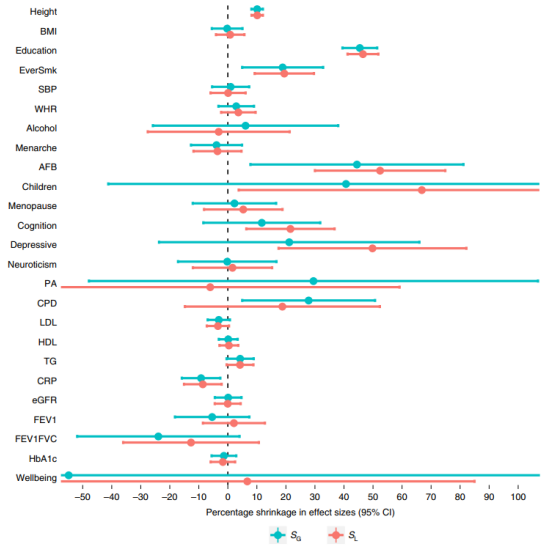
$$S = \frac{\sum_k^M (w_k \beta_k / \sigma_k^2)}{(w_k^2 / \sigma_k^2)}$$

- Using two thresholds for SNP inclusion (UKB discovery analysis): $P < 5 \times 10^{-8}$ or $P < 1 \times 10^{-5}$
- Note that we are weighting the analysis towards SNPs with a larger population effect
- Note that if we have

$$\mathbb{E}(\beta^{pop} k) = \mathbb{E}(\beta_k^{WF})$$

then: $\mathbb{E}(S_{pop}) = \mathbb{E}(S_{WF})$, even if $\sigma^2_{WF,k} > \sigma^2_{pop,k}$

Shrinkage of Within-sib effect vs. pop. GWAS effects



- 8 out of 25 phenotypes show substantial shrinkage:
  - Number of children
  - Age at first birth
  - Depressive symptoms
  - Educational attainment
  - Cognitive ability
  - Ever smoking
  - Height
- Little evidence of heterogeneity in shrinkage across SNPs
- Consistent with bias arising from assortative mating or indirect genetic effects (not population stratification)

Estimating SNP-based heritability within-family

- In population GWAS, we can estimate SNP-based heritability (h^2) using LD-score regression

$$\mathbb{E}[\chi_j^2] = 1 + N\alpha + Nh^2 l_j / M$$

  - χ_j^2, the GWAS test statistic for SNP j
  - l_j the LD score
  - M number of SNPs
- The slope is informative of heritability, but the proof of LD-score regression relies on OLS being BLUE
- How can it be applied in inefficient estimators such as within-family?

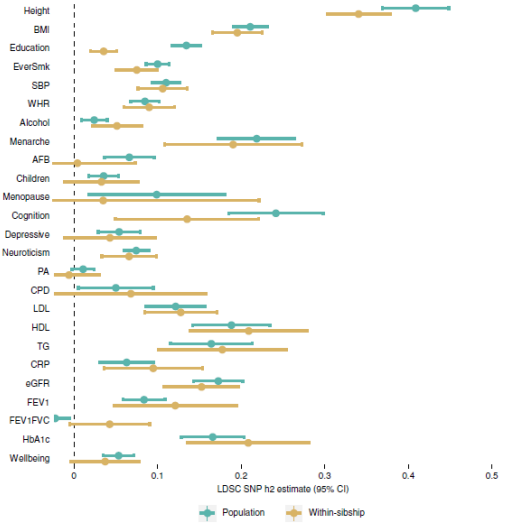
Estimating SNP-based heritability within-family

- Howe et al., go around the problem by estimating *effective N*

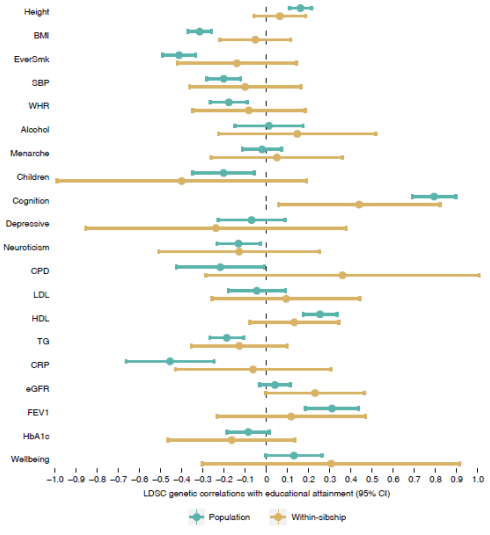
$$\text{Effective N} = \frac{1}{s.e.^2} \frac{s.d.\_\text{resid}^2}{2 \times MAF(1 - MAF)}$$

- Effective N is the N in population-GWAS that would give equally precise estimates as the obtained within-fam GWAS estimates
- "We used simulations to investigate the applicability of LDSC when using within-sibship GWAS data [...] if effective sample sizes are used to account for differences in power between the models"
- But, assortative mating might bias heritability based on within-sib GWAS downwards

Population and Within-sibship heritabilities



Genetic Correlations with EA

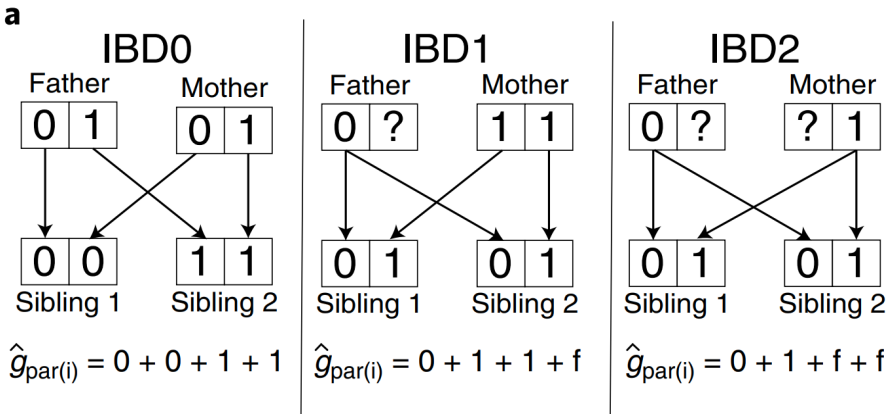


Mendelian Imputation

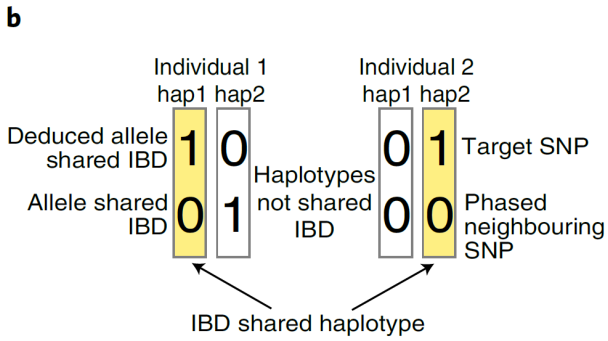
Mendelian Imputation (Young et al., 2022)

- What if we want to control for parental genotype, but parental genotypes are missing?
- Young et al., show that we can control for the *imputed* genotype
- Genotype can be imputed if:
  - at least one sibling genotyped
  - at least one parent genotyped
- Let $\widehat{g_i^{par}}$ be the imputed sum of SNPs of the parents of i
- $y_i = \beta_0 + \delta g_{i,j} + \alpha_l \widehat{g_i^{par}} + \gamma g_{i'j} + \varepsilon$ is unbiased, even in the presence of siblings
  - with $\alpha_l = (\alpha_p + \alpha_m)/2$
- Increase in power if we are willing to assume $\gamma = 0$, because
  - Fitting one coefficient for $\widehat{g}_i^{par}$ is more efficient than fitting J family fixed effects
  - Combining sibling data with trio data and parent-offspring data maximizes sample size

Mendelian Imputation in Practice – Siblings genotyped, parents missing



Mendelian Imputation in Practice – IBD1 ambiguity



- **Phasing:** Genotyping arrays measure *unordered* allele pairs at each SNP (e.g. $\{A, T\}$); **phasing** assigns each allele to one of the two chromosomal copies (haplotypes), recovering the ordered sequences $\langle \text{hap}_1, \text{hap}_2 \rangle$. Pipelines for this exist (e.g. `SHAPEIT`, `Eagle`) that exploit long-range LD patterns across many individuals to infer which alleles co-occur on the same chromosome.

Mendelian imputation of parental genotypes (phased)

Setup: siblings i_1, i_2 are genotyped; parent $\texttt{par}(i)$ is missing. IBD state $\in \{0, 1, 2\}$ is inferred from the siblings' genotypes via a hidden Markov model over the genome.

Imputed parental genotype:

$$\hat{g}_{\text{par}(i)} = \begin{cases} g_{i_1} + g_{i_2} & \text{if IBD} = 0 \\ g_{i_1} + g_{i_2}^{k} + f & \text{if IBD} = 1 \\ g_{i_1} + 2f & \text{if IBD} = 2 \end{cases}$$

where f is the allele frequency and $g_{i_2}^{k}$ is the allele of sibling i_2 that is *not* shared IBD with sibling i_1.

Mendelian imputation of parental genotypes (unphased)

Imputed parental genotype:

$$\hat{g}_{\text{par}(i)} = \begin{cases} g_{i_1} + g_{i_2} & \text{if IBD} = 0 \\ g_{i_1} + g_{i_2}^{k} + f & \text{if IBD} = 1 \text{ and } \neg H_i \\ 1 + 2f & \text{if IBD} = 1 \text{ and } H_i \\ g_{i_1} + 2f & \text{if IBD} = 2 \end{cases}$$

where H_i is the event that the siblings are heterozygous

Cost of not phasing

At IBD $= 1$ loci where both siblings are heterozygous, $g_{i_2}^{k}$ is unknown without phase information, so the imputation falls back on f — reducing accuracy.

Mendelian imputation, parent-offspring

Suppose we have one parent genotyped and one offspring genotyped, then then

		$g_{\mathrm{m(i)}}$		
		0	1	2
g_i	0	f	f	-
	1	$1+f$	$2f$	f
	2	-	$1+f$	$1+f$

Supplementary Note Table 7: $\mathbb{E}[g_{\mathrm{p(i)}}|g_i, g_{\mathrm{m(i)}}]$

In general, parent-offspring imputation adds less information as compared to imputation based on siblings

Mendelian imputation, more than two siblings

If more than 2 siblings, we either have that two siblings are IBD $= 0$, or all siblings are IBD $= 2$ and

Imputed parental genotype:

$$\hat{g}_{\mathsf{par}(i)} = \begin{cases} g_{i_1} + g_{i_2} & \text{if IBD} = 0 \\ g_{i_1} + 2f & \text{if IBD} = 2 \end{cases}$$

Imputation becomes even more accurate when we also observe 1 parent

Mendelian Imputation – unbiasedness

After imputing, simply estimate

$$y_i = \beta_0 + \delta g_i + \alpha_l \widehat{g}_i^{par} + \gamma g_{i'j} + \varepsilon$$

- Normally, imputed control variables might result in bias. However, here, the unobserved covariate (g_i^{par}) is replaced with its *expectation given observed covariates* (g_{ij}). As a result, estimates remain unbiased and consistent, and the empirical sampling variance-covariance matrix of the estimates is an unbiased estimate of the true sampling variance-covariance matrix
- The proof relies on the linearity of OLS and the law of iterated expectations
  - So it only works for OLS: not for non-linear specifications like e.g., logistic regression

Mendelian Imputation – unbiasedness

After imputing, simply estimate

$$y_i = \beta_0 + \delta g_{i,j} + \alpha_l \widehat{g}_i^{par} + \gamma g_{i'j} + \varepsilon$$

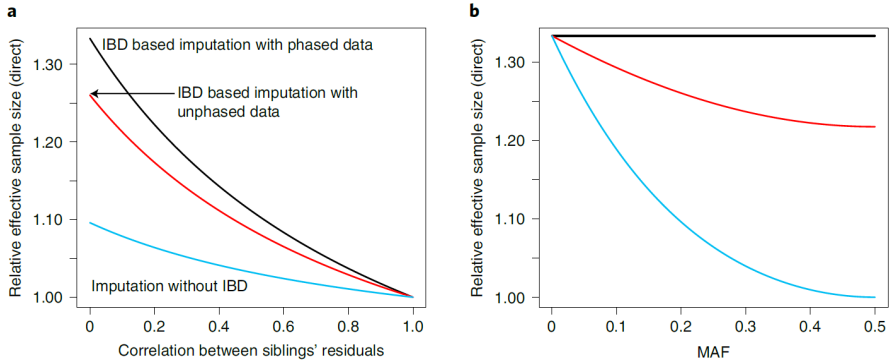
$cor(g_{i,j}, \widehat{g}_i^{par}) \approx 0.75$, $cor(g_{i'j}, \widehat{g}_i^{par}) \approx 0.75$, $cor(g_i, \widehat{g}_{i'j}) \approx 0.5$

- Because of the strong correlations in g_i, $\hat{g}_i^{par}$ and $g_{i'j}$, DGEs will not be very precise
- If we assume $\gamma = 0$, we can estimate

$$y_i = \beta_0 + \delta g_i + \alpha_l \widehat{g}_i^{par} + \gamma g_{i'j} + \varepsilon$$

- IF $\gamma \neq 0$, the bias in δ is $-\frac{(1+2r)}{(2+r)}\gamma$, with r the correlation in sibling residuals
- Still bias due to sibling effects, but less bias then within-family estimation (unless $r = 1$).

Mendelian Imputation increases power



- Here, we assume that the data only consists out of paired genotyped siblings, and $\gamma = 0$
- In practice, power gains can be larger because you can now combine sibling data, trio data, and duo data, such that you also increase the sample size

How to do it: snipar (Single Nucleotide Imputation of PARents)

- Imputation of missing parental genotypes
- Can be used to infer IBD states of siblings
- Family-PGI analyses, assortative mating correction for parental PGI
- Genetic correlations between direct and indirect genetic effects

FGWAS II - Tan et al., (2025)

Tan et al. (2025): *Family-GWAS reveals effects of environment and mating on genetic associations*

Advance over Howe et al. (2022):

- Howe et al. used *sibling differences* — identifies DGEs but discards parental genotype information
- Tan et al. use Mendelian Imputation, combined with observed trios
- Better-powered
- More strict QC
- Investigates difference in DGEs and population effects genome-wide, rather than top hits
- Allows for decomposition of confounding into stratification vs. assortative mating components
- Estimation less biased by possible sibling effects

Quality control in family-based GWAS

Imputation error reduces sibling genotype correlation

- Under Mendelian inheritance:

$$\text{corr}(g_{sib1}, g_{sib2}) = 0.5$$

- With imputed genotypes (low INFO scores):
  - INFO $= 0.30$ (Howe et al., threshold):

$$\text{corr} \approx 0.437 \quad \text{(dosages)}$$

$$\text{corr} \approx 0.376 \quad \text{(hard calls)}$$

- Only when INFO $\to 1$:

$$\text{corr} \to 0.5$$

- Low-quality imputation violates Mendelian assumptions
- Leads to bias in family-based GWAS

Solution: Strict quality control

- INFO score threshold: **INFO** > 0.99
- MAF $> 1\%$

The FGWAS regression

FGWAS estimates:

$$Y_i \;=\; \delta\, g_i \;+\; \alpha_p\, g_p \;+\; \alpha_m\, g_m \;+\; \mathbf{X}_i'\boldsymbol{\beta} \;+\; \varepsilon_i$$

when trio data is available

or:

$$Y_i \;=\; \delta\, g_i \;+\; \frac{(\alpha_p + \alpha_m)}{2}\hat{g}_i^{par} \;+\; \mathbf{X}_i'\boldsymbol{\beta} \;+\; \varepsilon_i$$

when sibling/parent-offspring data is available (Mendelian Imputation),

FGWAS: multivariate meta-analysis

We estimate the common parameter vector:

$$\theta_l = \begin{bmatrix} \delta_l \\ \alpha_{pl} \\ \alpha_{ml} \end{bmatrix}$$

For each cohort we have now estimated (at SNP l)

$$z_{jl} = [\hat{\delta}_l, \hat{\alpha}_{pl}, \hat{\alpha}_{ml}]^T$$

or

$$z_{jl} = [\hat{\delta}_l, \hat{\alpha}_l]^T$$

with $(\alpha_l = \frac{\alpha_{pl}+\alpha_{ml}}{2})$
under random mating: $\beta_l = \delta_l + \alpha_l$, where $\alpha_l = (\alpha_{pl} + \alpha_{ml})/2$

FGWAS: multivariate meta-analysis

For each cohort j, we observe:

$$z_{jl} = A_j \theta_l + \varepsilon_{jl}$$

- A_j: maps θ_l to what cohort j identifies
- Σ_{jl}: covariance matrix of z_{jl}

Meta-analysis estimator:

$$\hat{\theta}_l = \left(\sum_j A_j^\top \Sigma_{jl}^{-1} A_j \right)^{-1} \left(\sum_j A_j^\top \Sigma_{jl}^{-1} z_{jl} \right)$$

FGWAS: multivariate meta-analysis

We can even obtain pop-GWAS parameters via:

$$\hat{\theta}_l^* = B\hat{\theta}_l = \begin{bmatrix} \hat{\delta}_l \\ \hat{\alpha}_{pl} \\ \hat{\alpha}_{ml} \\ \hat{\alpha}_l \\ \hat{\beta}_l \end{bmatrix}$$

, with sampling variance-covariance matrix $BVar(\hat{\theta}_l)B^T$

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0.5 & 0.5 \\ 1 & 0.5 & 0.5 \end{bmatrix}$$

Tan et al., (2025) - Differences in SNP heritability

	Median effective *N* (HapMap3)		SNP heritability			
Phenotype	**DGE**	**Pop.**	**DGE**	**S.E.**	**Pop.**	**S.E.**
Height	105993	182202	0.352	0.020	0.413	0.021
BMI	81870	178153	0.212	0.013	0.216	0.012
Educational attainment (EA)	47387	91221	0.072	0.008	0.143	0.007
ADHD	44748	102327	0.005	0.014	0.003	0.007
Non-HDL cholesterol	42160	90474	0.168	0.023	0.179	0.021
Number of children	41589	102329	0.041	0.009	0.039	0.004
HDL cholesterol	40029	79576	0.191	0.035	0.181	0.024
Age at first birth (women)	35982	87944	0.044	0.012	0.097	0.008
Self-rated health	35443	83433	0.043	0.013	0.062	0.008
Blood pressure (systolic)	32532	72193	0.097	0.016	0.109	0.010
Blood pressure (diastolic)	32530	71625	0.102	0.017	0.111	0.011
Neuroticism	31649	75046	0.084	0.013	0.075	0.007
Depressive symptoms	31132	75497	0.060	0.015	0.035	0.007
Subjective well-being	28232	65930	0.026	0.016	0.048	0.009
Migraine	25907	67816	0.055	0.021	0.069	0.009
Drinks per week	22137	50345	0.027	0.022	0.029	0.012
Allergic rhinitis	21247	50657	0.086	0.026	0.082	0.015
Age-at-menarche	19678	45504	0.177	0.029	0.216	0.021
FEV1	18645	45121	0.167	0.022	0.138	0.012
Cigarettes per day	16121	37207	0.014	0.022	0.063	0.015
Ever-smoker	14935	34550	0.356	0.029	0.463	0.022
Morning person	13347	36632	0.081	0.042	0.109	0.018
Household income	12884	31956	0.045	0.038	0.107	0.016
Cognitive performance	12361	26345	0.188	0.027	0.186	0.016
Depression	12216	31531	0.025	0.015	0.082	0.010
Hypertension	7506	18771	0.397	0.091	0.372	0.046
Asthma	6549	16229	0.360	0.077	0.378	0.048
Eczema	6326	16139	0.134	0.072	0.169	0.036
Myopia	5498	13859	0.526	0.110	0.517	0.050
Individual income	5489	14742	0.024	0.032	0.041	0.013

Genetic correlation: LDSC approach

Cross-trait LD Score Regression (LDSC)

$$E[z_{\beta,l} \cdot z_{\delta,l}] = \frac{\sqrt{N_\beta N_\delta}}{M} \mathrm{Cov}_g(\delta, \beta)\, \ell_l + c$$

- Estimate:

$$r_g(\delta, \beta) = \frac{\mathrm{Cov}_g(\delta, \beta)}{\sqrt{h^2_\delta h^2_\beta}}$$

- Interpretation:
  - $r_g < 1$: evidence of confounding in population GWAS
- LDSC targets estimates r_g after correcting for population stratification (via LD sructure)
- **Tends to underestimate confounding** in β_l

Reason: Stratification effects may be partially removed by LDSC

Genetic correlation: SNIPAR-based approach

Idea: Moment-based estimator that removes sampling error

For variant l, let

$$\hat{\delta}_l = \delta_l + \varepsilon_{\delta l}, \qquad \hat{\beta}_l = \beta_l + \varepsilon_{\beta l}.$$

Sampling-error covariance:

$$\mathrm{Var}\begin{pmatrix}\varepsilon_{\delta l}\\ \varepsilon_{\beta l}\end{pmatrix} = \begin{bmatrix}\sigma^2_{\delta l} & r_l\sigma_{\delta l}\sigma_{\beta l}\\ r_l\sigma_{\delta l}\sigma_{\beta l} & \sigma^2_{\beta l}\end{bmatrix}.$$

Key moment (law of total variance):

$$c_{\delta\beta} = \mathrm{Cov}(\delta_l, \beta_l) = \mathrm{Cov}(\hat{\delta}_l, \hat{\beta}_l) - \mathbb{E}[\mathrm{Cov}(\varepsilon_{\delta l}, \varepsilon_{\beta l})].$$

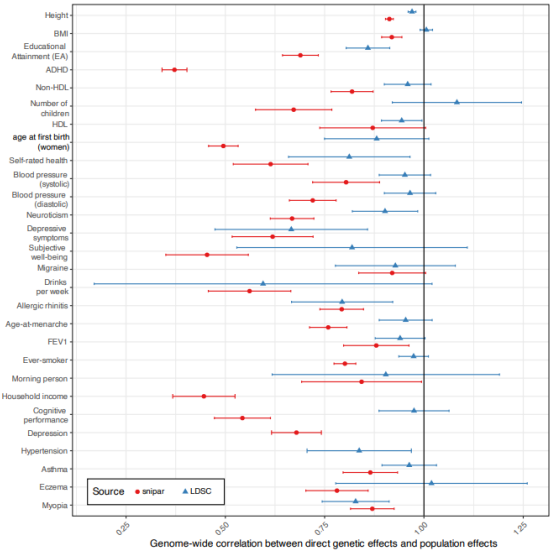
Estimator (across SNPs):

$$\hat{c}_{\delta\beta} = \sum_l w_l\big(\hat{\delta}_l\hat{\beta}_l - r_l\sigma_{\delta l}\sigma_{\beta l}\big).$$

Genetic correlation:

$$r(\delta, \beta) = \frac{c_{\delta\beta}}{\sqrt{v_\delta v_\beta}}.$$

Genetic correlations of population and DGEs



- $r_G < 1$ implies bias in population effects
- If $r_{\text{LDSC}} < 1$, bias likely due to IGEs and/or non-random mating
- If $r_{\text{SNIPAR}} < r_{\text{LDSC}}$, population stratification contributes to the bias

Tan et al., further results

- Overall inflation/deflation estimated by regressing pop effects onto DGEs, accounting for their sampling errors
- Height and systolic blood pressure have inflated pop. effects
- But 10 phenotypes (e.g., ADHD depression, cognitive performance) have deflated pop. effects!
- This could be explained by a (counterintuitive) negative correlation between DGEs and NTCs
- This could be driven by selection bias and non-representativeness of the underlying data

Remaining biases in FGWAS

Population GWAS: Sources of Bias (Veller & Coop, 2024)

The expected population GWAS coefficient at focal locus λ is:

$$\hat{\alpha}_\lambda^{\text{pop}} = \frac{2}{V_\lambda}\left(\sum_{l \in L_{\text{local}}} D_{\lambda l}\alpha_l^{\text{d}} + \underbrace{\sum_{l \in L \setminus L_{\text{local}}} D_{\lambda l}\alpha_l^{\text{d}} + \sum_{l \in L} \tilde{D}_{\lambda l}\alpha_l^{\text{d}}}_{\text{genetic confounds, direct}} + \underbrace{\sum_{l \in L} [D'_{\lambda l} + \tilde{D}'_{\lambda l} + 2\tilde{D}_{\lambda l}]\alpha_l^{\text{i}}}_{\text{genetic confounds, indirect}} + \underbrace{\tfrac{1}{2}\text{Cov}(\mathbf{g}_\lambda, \epsilon)}_{\text{environmental confound}} \right)$$

- First term: **local direct effects** — what we want to estimate
- $D_{\lambda l}$: *direct* LD (*cis-LD*); $\tilde{D}_{\lambda l}$: **indirect** LD (*trans-LD*) across homologous chromosomes
- Genetic confounds arise from LD with *other causal loci*, both on the **same** chromosome (direct LD) and **across** homologs (indirect LD)
- Indirect genetic effects (α^{i}): parental genotypes influencing offspring phenotype via the environment, where D' and $\tilde{D}'$ are the LD-matrices in parents' generation
- Environmental confound: residual gene–environment correlation (e.g. population stratification, assortative mating)

Within-Family GWAS: What Is Removed?

The within-family (sibling-difference) estimator approximately equals:

$$\hat{\alpha}_\lambda^{\mathsf{wf}} \approx \frac{2}{H_\lambda} \left(\sum_{l \in L_{\mathsf{local}}} D'_{\lambda l} \alpha_l^{\mathsf{d}} + \underbrace{\sum_{l \in L \setminus L_{\mathsf{local}}} (1 - 2c_{\lambda l})(D'_{\lambda l} - \tilde{D}'_{\lambda l}) \alpha_l^{\mathsf{d}}}_{\text{genetic confounds, direct}} \right)$$

Some bias remains:

- Direct effects at *local* loci (the target)
- Residual confounding from *distant* loci on the same chromosome, scaled by $(1 - 2c_{\lambda l})$ — attenuated by recombination (c: recombination fraction)
- Confounding persists only for loci that are linked (low $c_{\lambda l}$); for distant loci ($c_{\lambda l} \approx 0.5$), $(1 - 2c_{\lambda l}) \approx 0$ and bias vanishes
- Especially if l and λ are on different chromosomes, $c_{\lambda l} = 0.5$
- Within-family GWAS removes genome-wide confounding, but residual bias remains locally due to incomplete recombination

V_λ vs. H_λ: Rescaling and a LATE Interpretation

Population GWAS normalizes by the population genotypic variance:

$$V_\lambda = 2f_\lambda(1 - f_\lambda)$$

Within-family GWAS normalizes by the fraction of parents heterozygous at λ: H_λ
Under Hardy–Weinberg, $H_\lambda = V_\lambda$. They diverge under assortative mating, population structure, or drift.

$\Rightarrow$ The within-family estimate is a **LATE**, identified from segregating families (heterozygous parnts) only (Veller, Przeworksi & Coop, 2024)

$LATE \neq ATE$ if the DGEs are heterogeneous across subpopulations with different frequencies of heterozygosity

Polygenic indices and within-family data

What does the coefficient on the PGI measure?

Consider the "true" model:

$$y = \eta_0 + x'\eta_1 + x^{p'}\eta_2 + v$$

with x', $x^{p'}$ all SNPs of the child, and parents, respectively

Instead, we estimate:

$$y = \alpha_0 + \alpha_1 x'w + \alpha_2 x^{p'}w + \varepsilon$$

Key result (Benjamin et al., 2024):

$$\alpha_1 = \frac{w' \operatorname{Var}(x)}{\operatorname{Var}(xw)}\eta_1$$

- α_1 is a **weighted average of true causal SNP effects (LATE)**
- Weights depend on w and the variance structure of x, where only heterozygous individuals contribute for each locus
- Interpretation depends on how well $w \approx \eta_1$

Why do PGI weights matter?

In practice, $w \neq \eta_1$ due to:

1. **Sampling error in GWAS** ($N_{\text{GWAS}} < \infty$) (Becker et al., 2021, Van Kippersluis et al., 2023)
2. **Bias in population GWAS:** (Alemu et al., 2025)
   - Indirect genetic effects
   - Assortative mating
   - Population stratification

within-family GWAS performs better on (2), but worse on (1), compared to pop-GWAS

Using population PGIs within-family

Even if w comes from **population GWAS**, we can identify causal effects by conditioning on parental genotype.

Estimate:

$$y = \alpha_0 + \alpha_1 \, \mathsf{PGI}_i + \alpha_2 \, \mathsf{PGI}_{\mathsf{parents}} + \varepsilon$$

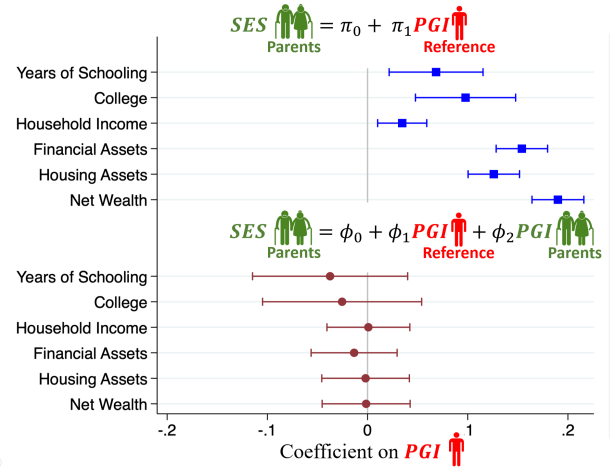
Intuition:

- Parent PGI absorbs:
  - genetic nurture
  - assortative mating components
- Remaining variation in child PGI is:
  - driven by Mendelian segregation
  - $\Rightarrow$ quasi-random within families

Result:

- α_1 recovers a causal effect
- while using **well-powered population weights**

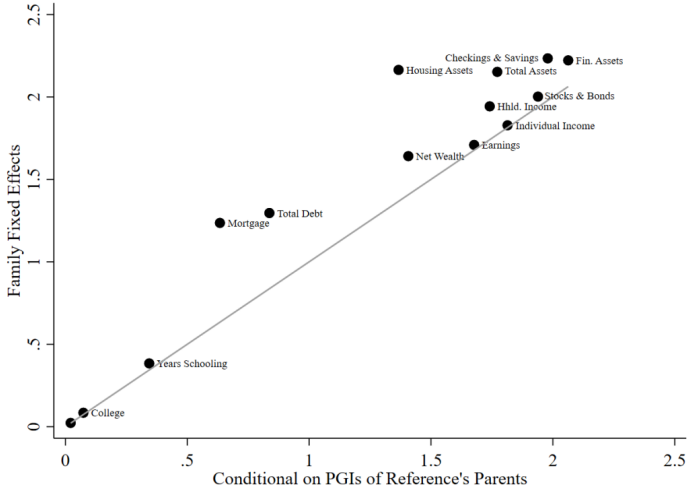
Balance Tests for the EA PGI (pop. GWAS)



Parental PGIs constructed using Mendellian imputation in Lifelines data

van Alten, et al. (2025). A chip off the old block? genetics and the intergenerational transmission of socioeconomic status

Equivalence of PGI effects using MI and Family FE



Concluding Remarks

- Within-family analysis eliminates virtually all biases that plague pop. GWAS
- The cost is a great loss in power
  - Genetic variance within families is only half that of the population
  - Restricting to only siblings greatly reduces sample size
- Comparing estimates of DGEs with estimates of pop. GWAS suggest pop. GWAS for some phenotypes are greatly biased, whereas other phenotypes show little to no evidence of bias
- Because of the bias-variance trade-off, both population GWAS and FGWAS will be of use in the near future
- When using FGWAS summary statistics in a post-GWAS pipeline (e.g., LD-score regression), you might need to account for the loss in power
- PGIs based on pop. GWAS are not biased in within-family analysis, and for now remain more predictive

Thank you!

E-mail:
s.j.d.van.alten@vu.nl
Bluesky:
@sjoerdalten.bsky.social
LinkedIn:
sjoerd-van-alten-711573277