Gene-Environment Correlations and Assortative Mating

ESSGN 2026 Workshop

Hans Fredrik Sunde & Espen Eilertsen

2026-03-25

Lecture Outline

> **The unifying theme**
>
> Different sources of confounding ultimately comes down to the same thing: **unmodelled correlation structure**.

Part	Focus
Part 1: Local linkage disequilibrium (LD)	Why nearby correlated SNPs both distort and inform marginal effect estimates
Part 2: Assortative mating and long-range LD	How non-random mating affects correlations between distant loci
Part 3: Gene-environment correlations	How correlations between genetic and environmental factors affect marginal associations and variance decompositions
Part 4: What to do? Use families!	What family-based methods can and cannot solve
Part 5: Practical	GREML and maternal indirect genetic effects with M-GCTA

Before we begin...

In session 1, we assumed that loci were independent from each other and from environmental effects.

$$\underbrace{\text{Cov}(x_j, e) = 0}_{\text{No rGE}} \quad \text{and} \quad \underbrace{\text{Cov}(x_j, x_k) = 0 \text{ for } j \neq k}_{\text{No LD}}$$

This is often not the case.

This session: What happens when those assumptions are violated?

This lecture is mostly **theoretical**.
The main takeaway is not a ready-to-use practical method, but a clearer picture of **what happens to *target* parameters and *estimated* parameters**.

Direct effect *vs* marginal association?

As a reminder, the infinitesimal model says that the phenotype is a sum of many small genetic effects plus a residual (i.e., an environmental effect):

$$y_i = \sum_{j=1}^{M} \beta_j x_{ij} + e_i$$

In a GWAS, each SNP j is tested separately in a linear model:

$$y_i = \alpha_j + \tilde{\beta_j} x_{ij} + \cdots + \varepsilon_{ij}$$

$\tilde{\beta_j}$ is the **marginal coefficient** measuring the ***association*** between x_{ij} and y_i:

$$\tilde{\beta_j} = \underbrace{\beta_j}_{\text{True Effect}} + \underbrace{\sum_{k \neq j}^{M} \beta_k \frac{\text{Cov}(x_j, x_k)}{\text{Var}(x_j)}}_{\text{Linkage Disequilibrium}} + \underbrace{\frac{\text{Cov}(x_j, e)}{\text{Var}(x_j)}}_{\text{rGE}}$$

Here we ignore sampling error, imperfect genetic overlap between GWAS and target phenotypes, and heterogeneity across GWAS samples

Part 1:
Local Linkage Disequilibrium

Nearby SNPs correlate

Recombination shuffles **chromosome segments**, not isolated SNPs. Loci that are close together are less likely to be separated by recombination and therefore tend to be inherited together.

Haplotype 1	1	0	1	1	0	1	0	0	1	1
Haplotype 2	0	1	0	0	1	0	1	1	0	0
Transmitted	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**0**

This makes nearby SNPs correlated in the population: $\mathrm{Cov}(x_j, x_k) \neq 0$ for nearby j, k

This is called local **linkage disequilibrium (LD)**.

> **Note**
>
> If two nearby SNPs are often inherited together, they are statistically *dependent*:
> Once you know the allele at one SNP, you can make a good guess about the allele at the other.

A two-loci toy example

Assume the true model is:

$$y_i = \beta_1 x_{i1} + e_i$$

where x_1 is causal and is in LD with a non-causal locus, x_2:

$$\beta_1 \neq 0, \qquad \beta_2 = 0, \qquad \mathrm{Cov}(x_1, x_2) \neq 0$$

A GWAS tests each locus separately. For x_2, the marginal coefficient is:

$$\tilde{\beta_2} = \beta_2 + \beta_1 \frac{\mathrm{Cov}(x_2, x_1)}{\mathrm{Var}(x_2)}$$

In other words, x_2 will be associated with the phenotype even though it has **no direct causal effect**.

Generalizing to many loci

Reminder: the marginal association at locus j is:

$$\tilde{\beta_j} = \underbrace{\beta_j}_{\text{Direct Effect}} + \underbrace{\sum_{k \neq j}^{M} \beta_k \frac{\text{Cov}(x_j, x_k)}{\text{Var}(x_j)}}_{\text{Linkage Disequilibrium}} + \underbrace{\frac{\text{Cov}(x_j, e)}{\text{Var}(x_j)}}_{\text{rGE}}$$

If we ignore rGE and standardize all SNPs ($\text{Var}(x_j) = 1$, meaning $\text{Cov}(x_j, x_k) = r_{jk}$), the marginal coefficient at SNP j becomes the weighted sum of direct effects at correlated SNPs, including itself:

$$\tilde{\beta_j} = \sum_{k=1}^{M} r_{jk} \beta_k$$

A SNP that is correlated with **many** nearby variants can therefore tag more causal variation in the region.

We can exploit this structure

If a trait is heritable, then SNPs correlated with many nearby variants should, on average, have more inflated test statistics.

Quantifying local LD

For SNP j, the **LD score** is the **sum of squared correlations** with nearby SNPs:

$$\ell_j = \sum_{k \in \square(j)} r_{jk}^2$$

where $\square(j)$ is a local window around SNP j.

Interpretation

A high LD score means the SNP is highly correlated with many nearby variants, and is more likely to tag causal variation in the region.

A low LD score means the SNP is more isolated and less likely to tag nearby causal effects.

LD score regression

If SNPs with higher LD scores tag more nearby causal variation, then their test statistics should be larger on average. **This relationship is proportional to the heritability.**

This allows us to estimate SNP heritability from GWAS summary statistics, using a method called **LD score regression (LDSC)**.

$$E[\chi_j^2 \mid \ell_j] = 1 + h^2 \ell_j \frac{N}{M} + c$$

where $\chi_j^2 = z_j^2$, and $z_j = \hat{\beta}_j / \mathrm{SE}(\hat{\beta}_j)$ for SNP j.

High-level interpretation

- Regress GWAS test statistics on LD scores
- The **slope** tracks polygenic signal, but it must be rescaled to get heritability: $h^2 = \text{slope} \times M/N$
- The **intercept**, c, tracks confounding or other inflation

Other considerations include appropriate regression weights to handle LD-dependent redundancy, plus a well-matched LD reference panel, careful GWAS QC, and checking for sample overlap or ancestry mismatch.

Genetic correlation from cross-trait LDSC

For two traits, the same idea extends to the product of the two marginal z-statistics:

$$E[z_{1j}z_{2j} \mid \ell_j] = 0 + r_g\,\ell_j\left(\frac{\sqrt{N_1N_2}}{M}\,h_1h_2\right) + c_{12}$$

where r_g is the **genetic correlation**, summarizing how strongly the two traits share genetic signal.

High-level interpretation

- Regress cross-trait z-score products on LD scores
- The slope tracks **shared genetic signal** across the two traits, but it first has to be rescaled to get the genetic covariance: $\mathrm{Cov}(g_1, g_2) = \text{slope} \times M/\sqrt{N_1N_2}$
- Then standardize that covariance to get the genetic correlation: $r_g = \frac{\mathrm{Cov}(g_1,g_2)}{h_1h_2}$
- Under random mating and i.i.d. SNP effects, this is often interpreted as overlap of additive effects; in Part 2 we will see how assortative mating can weaken that interpretation.

Local LD is also useful

Local LD can blur causal interpretation, but because it is local, measurable, and predictable, we can often model it and use it.

- **LD-score regression** exploits link between high LD and inflated test statistics to estimate heritability
- Nearby SNPs can **tag** causal variants, so we do not need to observe every causal SNP directly
- Local LD allows **imputation** of untyped variants from typed ones
- Tools like **LDpred** can adjust SNP weights to account for LD structure, improving polygenic prediction
- More broadly, local LD is often a **manageable nuisance**, not an arbitrary source of noise

Summary: local LD both confounds and helps

Step	Main lesson
Nearby SNPs correlate	Recombination makes nearby loci travel together
Marginal coefficients absorb tagged effects	So $\tilde{\beta_j}$ need not equal β_j
Local LD is also useful	Tagging, imputation, and LD-aware methods exploit nearby structure
LD scores quantify local tagging	They summarize how much nearby variation a SNP captures
LD score regression exploits that structure	The same LD that complicates interpretation can be modeled

Reminder:

$$\tilde{\beta_j} = \underbrace{\beta_j}_{\text{True Effect}} + \underbrace{\sum_{k \neq j}^{M} \beta_k \frac{\text{Cov}(x_j, x_k)}{\text{Var}(x_j)}}_{\text{Linkage Disequilibrium}} + \underbrace{\frac{\text{Cov}(x_j, e)}{\text{Var}(x_j)}}_{\text{rGE}}$$

Part 2: Assortative Mating and Long-Range LD

What is assortative mating?

⚠ **Partner similarity, $\mathrm{Cov}(y_m, y_p) > 0$, can arise for different reasons.**

- **Assortative mating**: matching induces partner covariance
- **Convergence**: partners become more similar after pairing
- **Stratification**: partners are similar because both are shaped by the same background factors

Here, we assume that partner covariance is due to assortative mating, and we focus on the genetic consequences of that process.

Assortative mating (AM) means that the **matching process itself** induces covariance between partners, rather than that covariance reflecting a shared cause.

ⓘ **Disassortative mating?**

Technically, **assortative mating can be negative:** ($\mathrm{Cov}(y_m, y_p) < 0$). However, this is rare in human populations. We therefore focus on positive assortative mating.

The consequences of negative assortative mating are generally the same, but with opposite signs.

Consequences of assortative mating

Key intuition: Assortative mating induces correlations among causes by matching on the effects.

Ordinary family resemblance:
Covariance starts in their **genetic values** (and other causes) and is carried forward into their phenotypes

> Correlation flows from $g \rightarrow y$.

Assortative mating:
Covariance starts in their **phenotypes**, and is carried back into their genetic values (and other causes).

> Correlation flows from $y \rightarrow g$.

Assortative mating therefore induces genetic similarity between partners.

Predicted genetic value given phenotype

To quantify the induced genetic similarity, we **start at the phenotype** and ask:
if we observe someone's phenotype y, what genetic value g do we expect?

Assume centered variables and the simple additive model:

$$y_i = g_i + e_i; \quad \mathrm{Cov}(g, e) = 0; \quad E[g_i] = E[e_i] = 0 \quad \Rightarrow \quad \mathrm{Cov}(g, y) = \mathrm{Var}(g) = \sigma_g^2$$

Here, the regression slope of g on y is

$$b_{g|y} = \frac{\mathrm{Cov}(g, y)}{\mathrm{Var}(y)} = \frac{\sigma_g^2}{\sigma_y^2} = h^2$$

Meaning the corresponding expected genetic value given phenotype is

$$E[g_i \mid y_i] = h^2 y_i$$

⚠ $\mathrm{Cov}(g, y)/\mathrm{Var}(y)$ **is still the relevant slope even if** $\mathrm{Cov}(g, e) \neq 0$**, but it no longer equals** h^2**.**

▶ Show the algebra

Quantifying genetic similarity

Applying $E[g \mid y] = h^2 y$ to both mates, the induced covariance between their genetic values is:

$$\mathrm{Cov}(g_m, g_p) = \mathrm{Cov}(h^2 y_m, h^2 y_p) = h^4 \mathrm{Cov}(y_m, y_p)$$

So the induced genotypic correlation between mates is

$$r_g^{\text{mates}} = \frac{(\sigma_g^2/\sigma_y^2)^2 \mathrm{Cov}(y_m, y_p)}{\sigma_g^2} = \frac{\sigma_g^2}{\sigma_y^2} \frac{\mathrm{Cov}(y_m, y_p)}{\sigma_y^2} = h^2 r_y^{\text{mates}}$$

> **Note**
>
> The h^4 appears on the covariance scale because the phenotype-to-genotype regression slope ($\sigma_g^2/\sigma_y^2 = h^2$) is applied **once for each mate**. When we convert covariance to correlation, one power of σ_g^2 is divided back out, leaving the simpler result $r_g^{\text{mates}} = h^2 r_y^{\text{mates}}$.

Direct versus indirect assortative mating

The preceding (and following) equations assume that partners match directly on the **focal phenotype** y.

In reality, they may match on a related signal or bundle of traits s that correlates with y.

Case	What partners match on	Induced genotypic correlation between partners
Direct assortment	The focal phenotype itself: $s = y$	$r_g^{\text{mates}} = \text{Corr}(g, y)^2 r_y^{\text{mates}} = h^2 r_y^{\text{mates}}$
Indirect assortment	A related signal or bundle of traits: $s \neq y$, but $\text{Corr}(s, y) \neq 0$	$r_g^{\text{mates}} = \text{Corr}(g, s)^2 r_s^{\text{mates}}$

The sorting factor

The genotypic correlation between partners depends on **how strongly the genetic value for the focal phenotype correlates with** s. Indirect assortative mating can imply **more or less** genetic resemblance between partners than you would otherwise expect.

In this lecture, we assume direct assortment where $s = y$. However, note that assortment is indirect for many social traits (e.g., educational attainment).

Genetic similarity spreads across loci

Rememeber, the genetic value is a weighted sum of SNPs: $g_i = \sum_{j=1}^{M} \beta_j x_{ij}$.

Covariance between partners' genetic values is therefore the sum of all SNP-level covariances:

$$\mathrm{Cov}(g_m, g_p) = \mathrm{Cov}\left(\underbrace{\sum_{j=1}^{M} \beta_j x_{mj}}_{g_m} , \underbrace{\sum_{k=1}^{M} \beta_k x_{pk}}_{g_p} \right) = \sum_{j=1}^{M} \sum_{k=1}^{M} \beta_j \beta_k \, \mathrm{Cov}(x_{mj}, x_{pk})$$

Importantly, all these covariances **have the same sign** (i.e., they are all positive) for trait-associated SNPs, because the matching process aligns the genetic values of the partners.

SNP-level covariance

To calculate $\mathrm{Cov}(x_{mj}, x_{pk})$, we can use the same regression logic as before where we regress the two SNPs on their phenotype and combine that with the phenotypic covariance: $\frac{\mathrm{Cov}(x_{mj}, y_m)}{\mathrm{Var}(y_m)} \mathrm{Cov}(y_m, y_p) \frac{\mathrm{Cov}(y_p, x_{pk})}{\mathrm{Var}(y_p)}$.

Why is this a problem?

Once partner genomes become correlated, transmission and recombination **embeds the correlation structure** in the next following generation.

Assortative mating across generations **induces a signed correlation structure between trait-associated alleles** (i.e., linkage disequilibrium), and that correlation structure can increase marginal SNP effects and family resemblance.

Now we will unpack how that happens.

What happens after *one* generation of AM

First, we define the genetic values separately for the maternal and paternal haplotypes:

$$g_c^m = \sum_{j=1}^{M} \beta_j x_j^m, \qquad g_c^p = \sum_{j=1}^{M} \beta_j x_j^p, \qquad g_c = g_c^m + g_c^p$$

where x_j^m and x_j^p are the maternally and paternally transmitted alleles at SNP j, respectively.

Assuming no relevant local LD, the transmitted alleles are independent **within** haplotypes:

$$\mathrm{Cov}(x_j^m, x_k^m) = 0, \qquad \mathrm{Cov}(x_j^p, x_k^p) = 0 \qquad \text{for } j \neq k$$

However, because the parents had correlated genotypes, the transmitted alleles are correlated **across** haplotypes:

$$\mathrm{Cov}(x_j^m, x_k^p) > 0 \text{ for trait-associated } j, k$$

Genetic variance after one generation of AM

Because the transmitted alleles are still independent within haplotypes, the variance of each haplotype contribution is just a sum of single-locus terms:

$$\mathrm{Var}(g_c^m) = \sum_{j=1}^{M} \beta_j^2 \mathrm{Var}(x_j^m), \qquad \mathrm{Var}(g_c^p) = \sum_{j=1}^{M} \beta_j^2 \mathrm{Var}(x_j^p)$$

However, because the transmitted alleles from either parent are correlated, the total variance of the genetic value includes a positive covariance term:

$$\mathrm{Var}(g_c) = \mathrm{Var}(g_c^m) + \mathrm{Var}(g_c^p) + 2\mathrm{Cov}(g_c^m, g_c^p)$$

$$\text{where} \quad \mathrm{Cov}(g_c^m, g_c^p) = \sum_{j=1}^{M} \sum_{k=1}^{M} \beta_j \beta_k \, \mathrm{Cov}(x_j^m, x_k^p)$$

Terminology: trans vs cis covariance

Covariance **across** the maternal and paternal haplotypes is called **trans** covariance.

Covariance **within** the maternal or paternal haplotype is called **cis** covariance.

What happens after *several* generations of AM

Repeated assortment plus recombination gradually converts part of the inherited **trans** covariance into **cis** covariance.

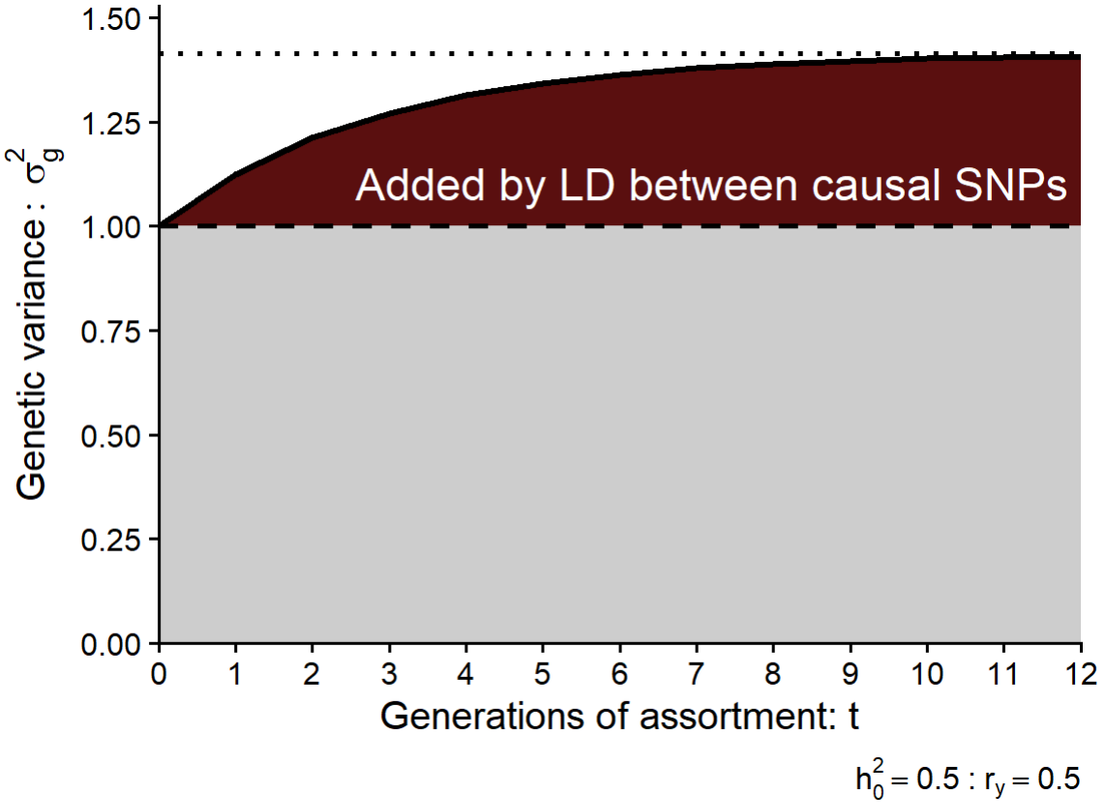
Written out fully, the three variance components become:

$$\mathrm{Var}(g_c) = \underbrace{\sum_{j=1}^{M}\sum_{k=1}^{M}\beta_j\beta_k\,\mathrm{Cov}(x_j^m, x_k^m)}_{\mathrm{Var}(g_c^m)\ \text{(cis)}} + \underbrace{\sum_{j=1}^{M}\sum_{k=1}^{M}\beta_j\beta_k\,\mathrm{Cov}(x_j^p, x_k^p)}_{\mathrm{Var}\left(g_c^p\right)\ \text{(cis)}} + \underbrace{2\sum_{j=1}^{M}\sum_{k=1}^{M}\beta_j\beta_k\,\mathrm{Cov}(x_j^m, x_k^p)}_{2\mathrm{Cov}\left(g_c^m, g_c^p\right)\ \text{(trans)}}$$

When	Where	Interpretation
1 generation	Across haplotypes (trans)	Maternal and paternal contributions become aligned
2+ generations	Within haplotypes (cis) also builds up	Aligned alleles begin to co-occur within haplotypes

Genetic variance across generations

Genetic variance increases across generations to an equilibrium::



Heritability changes too

Changes in genetic variance changes the ratio of genetic to phenotypic variance:

$$h^2_\infty = \sigma^2_{g,\infty} / (\sigma^2_{g,\infty} + \sigma^2_e)$$

Assortative mating therefore changes the heritability.

It typically **increases**, but with environmental transmission, the environmental variance changes too and the net effects are more complex:

$$h^2_\infty = \sigma^2_{g,\infty} / (\sigma^2_{g,\infty} + \sigma^2_{e,\infty} + 2\sigma_{g,e,\infty})$$

▶ Show recurrence equation for genetic variance

Why should we care?

Assortative mating in earlier generations means that **trait-associated SNPs are not independent from each other**.

Consequence 1: Marginal associations are inflated

Reminder: the marginal association at locus j is

$$\tilde{\beta_j} = \underbrace{\beta_j}_{\text{Direct Effect}} + \underbrace{\sum_{k \neq j}^{M} \beta_k \frac{\text{Cov}(x_j, x_k)}{\text{Var}(x_j)}}_{\text{Linkage Disequilibrium}} + \underbrace{\frac{\text{Cov}(x_j, e)}{\text{Var}(x_j)}}_{\text{rGE}}$$

Again set the rGE term aside and focus on the LD component. The sum is typically dominated by nearby loci (local LD). Under assortative mating, some of those covariance terms involve **distant** causal variants.

The upward inflation is *specific to causal loci* (and loci that tag causal loci through local LD).

For these loci, the marginal association at SNP j partly reflects genome-wide covariance structure, not only local biology.

Because assortative mating changes marginal associations estimated by GWAS, downstream methods using GWAS summary statistics are affected.

Consequence 2: LDSC can overstate h^2

In Part 1, LDSC exploited that $E[\chi^2]$ at SNP j is proportional to heritability and its LD score, ℓ_j:

$$E[\chi_j^2 \mid \ell_j] = 1 + h^2 \ell_j \frac{N}{M} + c$$

- Under assortative mating, χ_j^2 can be inflated by **long-range LD** between causal loci
- ℓ_j only capture **local LD** around SNP j

> **Important**
>
> Because the causal SNPs are in LD *with each other*, all of their marginal effects are inflated by the same long-range covariance structure. Genetic effects are therefore effectively double-counted, and h^2 is overestimated.

Consequence 3: GREML behaves differently

In Session 1, GREML used the genomic relatedness matrix $\mathbf{G}$ to capture genetic similarity among unrelated people.

- $\mathbf{G}$ is built from **same-SNP, cross-person** similarity
- Assortative mating mainly adds **long-range LD between different SNPs**
- AM-induced extra genetic variance is not well captured by the GRM used by GREML

GREML keeps fitting the Session 1 model $\sigma_g^2\mathbf{G} + \sigma_e^2\mathbf{I}$.

Under assortative mating, this is the wrong model and estimates will be biased.

Important

- In realistic samples, GREML is generally **upwardly biased** under assortative mating
  - Unfortunately, the size of that bias can be hard to quantify and is highly dependent on sample size
- As sample sizes increase, the estimate drifts toward the random-mating heritability, h_0^2
  - In very large samples (N > 200K), GREML can therefore be **downwardly biased** because $h_0^2 < h_\infty^2$

Consequence 4: family resemblance changes too

At equilibrium, the genotypic correlation between d-degree relatives becomes

$$r_{g,d} = \left(\frac{1 + r_g^{\text{mates}}}{2} \right)^d$$

Family-based estimators typically assume $r_{g,d} = (1/2)^d$, but that is only true under random mating

> **Note**
>
> The exact bias is method dependent, but classic twin methods are typically **downwardly biased** under positive assortative mating.

Consequence 5: Cross-trait assortment affects r_g

Reminder: two definitions of genetic correlation

$$\text{Score correlation} = \text{Corr}(g_1, g_2), \qquad \text{Effect correlation} = \text{Corr}(\beta_1, \beta_2)$$

If SNPs were i.i.d., these would be equal. Assortative mating can break this equivalence.

Cross-trait assortative mating (xAM) means that people sort across two traits, or on a bundle of traits related to both y_1 and y_2.

Across generations, this **builds long-range LD between trait-increasing alleles across y_1 and y_2.**

This can change the *score correlation* even when *effect correlation* is zero

Estimators of r_g (e.g., LDSC) can be biased under xAM.

Notably, the resulting estimates do not correspond to either the *score correlation* or the *effect correlation*.

Summary:
assortative mating changes targets and estimators

Step	Main lesson
AM induces genetic similarity	$\mathrm{Cov}(y_m, y_p) > 0 \quad \Rightarrow \quad \mathrm{Cov}(g_m, g_p) > 0$
Repeated AM	Long-range LD builds up over generations
Population target	Heritability itself can change because long-range LD adds genetic variance
GWAS / LDSC	Marginal associations absorb distant tagged signal, pushing $\hat{h}^2_{\mathrm{LDSC}}$ upward
GREML	In realistic sample sizes, heritability estimates are usually upwardly biased, but converges to random-mating heritability as sample size increases
Relatives / family methods	Relatives become more genetically similar, so family-based estimators are affected too; the bias is method dependent
Cross-trait AM	Score correlation can exceed effect correlation under xAM. Estimators of r_g are biased

Part 3:
Gene-Environment Correlations

What is rGE

Gene-environment correlation (rGE) means that genetic differences and environmental differences are statistically correlated, rather than varying independently.

rGE can be on the SNP-level ($\mathrm{Cov}(x_j, e)$) and on the trait-level ($\mathrm{Cov}(g, e)$)

Only passive gene-environment correlation is confounding

- **Passive** r_{ge}: Genotype and environment are correlated because of shared causes (e.g., parents)
- **Active / Evocative** r_{ge}: Genotype and environment are correlated because the focal person selects, creates, or evokes their environment based on their genotype (e.g., peers, partners, neighborhoods) Crucially, **active and evocative gene-environment correlations are not confounding**. They are part of the causal pathway from genotype to phenotype, and are therefore mediators of the genetic effect. This is also important for how "direct" genetic effects are interpreted.

Note: *rGE* and *GxE* are not the same thing

rGE means genes and environments are correlated (for example, $\mathrm{Cov}(x_j, e) \neq 0$), while **GxE** means genetic effects (β_j) varies across environments (an interaction).

Where can $\mathrm{Cov}(x_j, e)$ come from?

Source	Why genotype and environment become correlated
Stratification / population structure	Allele frequencies and environments differ across groups
Indirect genetic effects	Related people's genotypes shape the focal person's environment
Residual confounding	Selection, ascertainment, or measurement structure
Assortative mating	Can increase and even induce covariance between inherited alleles and family environments

Different mechanisms can produce the **same statistical symptom**: $\mathrm{Cov}(x_j, e) \neq 0$.

Across many loci, this can aggregate into $\mathrm{Cov}(g, e) \neq 0$.

Stratification

Suppose allele frequency at locus x_1 differs across some strata $S \in 1, 2$:

$$E[x_1 | S = 1] \neq E[x_1 | S = 2]$$

Suppose further that the strata has an environmentally mediated effect on the phenotype:

$$y_i = \sum_{j=1}^{M} \beta_j x_{ij} + e_i \quad \text{where} \quad e_i = \beta_S S_i + e'_i$$

Under these assumptions, the genotype at locus x_1 will be indirectly correlated with the phenotype via the environment:

$$\text{Cov}(x_1, e) = \text{Cov}(x_1, \beta_S S)$$

Indirect genetic effects, SNP-level

Suppose that a relative ℓ (e.g., a parent) environmentally influence the focal person's phenotype, and that this influence was heritable on behalf of the relative:

$$\underbrace{y_i = \sum_{j=1}^{M} \beta_j x_{ij}}_{\text{additive model}} + \underbrace{e_i}_{\text{environmental effects}} \quad \text{where} \quad e_i = \sum_{\ell} e_{i\ell}^{R} + e_i' \quad \text{and} \quad \underbrace{e_{i\ell}^{R} = \sum_{k=1}^{M} \gamma_{k\ell}\, x_{i\ell k}}_{\text{relative's effect}} + e_{i\ell}$$

Here $e_{i\ell}^{R}$ is the part of the focal person's environment created by a specific relative ℓ.

If SNPs are i.i.d., then the correlation between the focal person's genotype and the environment created by relative ℓ will proportional to the genetic relatedness r_{g}^{ℓ}:

$$\mathrm{Cov}(x_j, e_{\ell}^{R}) = \gamma_{k\ell}\, \mathrm{Cov}(x_j, x_{j\ell}) = r_{\mathrm{g}}^{\ell}\, \gamma_{k\ell}\, Var(x_j)$$

Things can get complicated when there are multiple relatives (e.g., mother, father, *and* siblings) affecting the focal person, because their effects are typically not independent.

Indirect genetic effects, trait-level

Define the focal person's direct genetic value and the relative's environmental genetic value:

$$g_i = \sum_{j=1}^{M} \beta_j x_{ij}, \qquad g^R_{i\ell} = \sum_{j=1}^{M} \gamma_{\ell j} x_{i\ell j}, \qquad \mathrm{Cov}(\mathrm{g}, e) = \sum_{\ell} \mathrm{Cov}(\mathrm{g}, \mathrm{g}^R_\ell).$$

The environmental effects can be conceived as *extended phenotypes* on behalf of the relatives' genes

If SNPs are i.i.d., then each relative-specific covariance term is:

$$\mathrm{Cov}(\mathrm{g}, \mathrm{g}^R_\ell) = \sum_{j=1}^{M} \beta_j \gamma_j \mathrm{Cov}(x_j, x_{j\ell}) \qquad \propto \qquad r^\ell_{\mathrm{g}}\, \rho_{\beta,\gamma_\ell}$$

where ρ_{β,γ_ℓ} is the genetic correlation between direct and indirect genetic effects and r^ℓ_{g} is the genotypic correlation between the focal person and relative ℓ.

Key point:

Even if many loci have $\mathrm{Cov}(x_j, e) \neq 0$, **the aggregate** $\mathrm{Cov}(g, e)$ **depends on the genetic correlation between direct** (β) **and indirect** (γ_ℓ) **effects**. If $\rho_{\beta,\gamma} = 0$ across relatives, then $\mathrm{Cov}(g, e) = 0$ even if $\mathrm{Cov}(x_j, e) \neq 0$

Indirect genetic effects + assortative mating?

Assortative mating can severely increase gene-environment correlations

1. Assortative mating induces long-range LD, meaning SNPs are not i.i.d.:
   Locus x_j may correlate with loci **across the *relatives'*** genome, not just $x_{\ell j}$.
2. The genotypic correlation between the focal individual and their relatives can increase
3. The effects of different relatives (e.g., mother and father) are likely to be correlated
4. Assortative mating opens weird pathways.
   For example, maternally inherited alleles can be correlated with paternal environmental effects
5. The genetic (score) correlation between direct and indirect genetic effects may change.

Aggregating over relatives and loci:

$$\mathrm{Cov}(g, e) = \sum_{\ell} \sum_{j=1}^{M} \sum_{k=1}^{M} \beta_j \gamma_\ell \ \mathrm{Cov}(x_j, x_{\ell k}).$$

Why should we care?

Consequence 1: Marginal associations include rGE

Reminder: the marginal association at locus j is

[Math Processing Error]

If $\mathrm{Cov}(x_j, e) \neq 0$, the marginal SNP association absorbs rGE and is no longer a pure direct effect.

This consequence happens even if $\mathrm{Cov}(g, e) = 0$

Important

Just like under assortative mating, once GWAS marginal associations are affected, downstream analyses that rely on GWAS summary statistics (e.g., LDSC) are affected too.

Consequence 2: $\mathrm{Cov}(g, y)$ is affected

$$y = g + e \qquad \Longrightarrow \qquad \mathrm{Cov}(g, y) = \mathrm{Cov}(g, g + e) = \mathrm{Var}(g) + \mathrm{Cov}(g, e)$$

With rGE, $\mathrm{Cov}(g, y)$ no longer equals $\mathrm{Var}(g)$.

This matters for polygenic scores

Even with a **perfect** score, where $PGS_i = g_i$, regressing y_i on PGS_i still gives

$$\frac{\mathrm{Cov}(PGS, y)}{\mathrm{Var}(PGS)} = \frac{\mathrm{Cov}(g, y)}{\mathrm{Var}(g)} = 1 + \frac{\mathrm{Cov}(g, e)}{\mathrm{Var}(g)}.$$

Polygenic scores can tag environmental effects **even without biased weights**, because the true genetic value is also tagging environmental effects

Consequence 3: Variance decomposition becomes awkward

If $\mathrm{Cov}(g, e) \neq 0$, the gene-environment correlation has its own variance component

$$\mathrm{Var}(y) = \mathrm{Var}(g) + \mathrm{Var}(e) + 2\mathrm{Cov}(g, e)$$

The sign of the covariance matters:

Sign of $\mathrm{Cov}(g, e)$	**Interpretation**	**Related concept**
Positive	Genes and environments tend to push in **the same direction**	Matthew effects
Negative	Genes and environments tend to push in **opposite directions**	Compensatory effects

(!) $\mathrm{Var}(y) = \mathrm{Var}(g) + \mathrm{Var}(e) + 2\mathrm{Cov}(g, e)$ **is a statistical identity, not a clean causal partition of phenotypic variation.**

If we suppose $\mathrm{Var}(g) = Var(e) = 1$ and $\mathrm{Cov}(g, e) = -0.30$, then $\mathrm{Var}(y) = 1.4$ and $h^2 = \mathrm{Var}(g)/\mathrm{Var}(y) = 0.72$. Does it make sense to say that heritability is 72% in this example? How much variance does environment explain?

In theory, if rGE is negative enough, the environment can work against genetic differences so strongly that $\mathrm{Var}(g) > \mathrm{Var}(y)$.

Summary: gene-environment correlations

Step	Main lesson
Confounding vs mediation	Only passive rGE is confounding; active and evocative rGE are part of the causal pathway from genotype to phenotype
Shared statistical symptom	Stratification, indirect genetic effects, residual confounding, and assortative mating can all induce $\mathrm{Cov}(x_j, e) \neq 0$
SNP level	If $\mathrm{Cov}(x_j, e) \neq 0$, the marginal association $\tilde{\beta_j}$ is not a pure direct effect, and downstream summary-statistic methods inherit that contamination
Trait level	Many nonzero $\mathrm{Cov}(x_j, e)$ terms do not guarantee $\mathrm{Cov}(g, e) \neq 0$; the aggregate depends on alignment between direct and indirect effect architectures
Assortative mating	AM can strengthen rGE by increasing focal-relative genetic correlations and opening extra paths between inherited alleles and family environments
Prediction / PGS	Even a perfect score tracks $\mathrm{Cov}(g, y) = \mathrm{Var}(g) + \mathrm{Cov}(g, e)$, so its slope can contain environmentally correlated signal
Variance decomposition	$\mathrm{Var}(y) = \mathrm{Var}(g) + \mathrm{Var}(e) + 2\mathrm{Cov}(g, e)$ is an identity, not a clean causal partition; *Especially with negative rGE*

Part 4: What to do? Use families!

The overarching problem (and solution)

The unifying theme: Different sources of confounding ultimately comes down to the same thing: **unmodelled correlation structure**.

In theory, this correlation structure could be accounted for by adding enough covariates.
In reality, this is not feasable, even for observed confounders like other SNPs.
Instead, we must deal with the correlation structure by design.

One key idea is to isolate **random segregation variance**.

Remember, the offspring genotype at SNP j can be written as

$$x_{jc} = \frac{1}{2}x_{jp} + \frac{1}{2}x_{jm} + (\epsilon_{jp} + \epsilon_{jm})$$

Key idea:

By conditioning on parents' or siblings' genotype, you remove the variance that is predictable from the parents, and are left with only the random segregeation variance.
This variance is independent from long-range LD and environmental effects.

Where does the problem live?

We must distinguish problems in **marginal SNP effects** and problems in **realized genotypes and true genetic values**.

If the problem is in…	Family-based response
$\tilde{\beta_j}$	Condition on parental genotypes **in the GWAS** so the estimate moves closer to the direct effect β_j
x_j, g, **or a PGS**	If the inherited genotype or true genetic value is itself correlated with family background, then cleaner SNP weights does not solve our issues. The solution depends on the estimand, but to attain direct genetic effects, we would need to condition on **parental genetic values** (or parental PGSs)
Trait-level covariance	If the problem is that direct and indirect pathways are mixed together in $\mathrm{Cov}(g, y)$, use a **family mixed model** such as M-GCTA to explicitly model direct and indirect genetic variance, *as well as their covariance*

Reminder

A perfect polygenic score ($PGS_i = g$) is not automatically unconfounded, because the true genetic value could be correlated with environmental effects.

Mixed-model family designs

One can use use relatives inside a **mixed model** to separate direct and indirect genetic pathways.

Intuition: ordinary GREML vs M-GCTA

Ordinary GREML asks:

- Are focal people who are slightly more genetically similar also slightly more phenotypically similar?

M-GCTA asks:

- Additionally, are focal people whose **mothers** are slightly more genetically similar also slightly more phenotypically similar?

In both cases the logic is covariance matching. If a source of genetic similarity creates phenotype similarity, give the mixed model the corresponding covariance matrix.

The logic can also easily be extended to, say, trios (Trio-GCTA)

The M-GCTA model

In Session 1, GREML used a GRM to model covariance from individuals' own genotypes:

$$\mathbf{y} = \mathbf{g} + \mathbf{e}, \qquad \mathbf{g} \sim \square(\mathbf{0},\ \sigma^2_g \mathbf{G})$$

If a relative's genotype (e.g., the mother, M) also shape that phenotype, then part of what we are calling "environment" has its own genetic covariance structure:

[Math Processing Error]

In this session's practical, the focal phenotype is the offspring and the relevant relative is the mother. So M-GCTA extends the direct-only GREML model by adding:

- a direct component based on $\mathbf{G}_{oo}$ = `Goo`
- a maternal indirect component based on $\mathbf{G}_{mm}$ = `Gmm`
- a covariance component for overlap between direct and maternal effects ($\mathbf{G\_om} + \mathbf{Gmo}$): `Dom`

Part 5: Practical

Practical 1: GREML and M-GCTA with simulated trios

Session 2 continues the same `practical/sim_genomic.R` (or `practical/sim_genomic.jl`) script used in Session 1.

Today's new extension is:

1. Start from the direct-only setup already covered in Session 1
2. Simulate an offspring phenotype with a maternal indirect genetic effect
3. Refit the wrong direct-only GREML model
4. Fit the maternal model (M-GCTA) and compare the variance components

> **Note**
>
> Parts 1-4 of the script are the same setup from Session 1, so they are not repeated on the slides here. If needed, rerun them in the script to recreate `X_tilde`, `yo`, `G`, `Goo`, and the direct-only GREML fit.

Starting point from Session 1

By this point, you should already have the direct-only objects from the Session 1 practical:

- standardized genotypes `X_tilde`
- offspring phenotype `yo`
- genomic relationship matrix `G`
- offspring block `Goo`
- direct-only GREML fit

The following slides focus only on the **new Session 2 extension**.

▶ Show Session 1 code

Part 5: Simulate a maternal indirect genetic effect

R | Julia

```r
Sg = rbind(
  c(0.2, 0.2),
  c(0.2, 0.5)
)
s2_e = 0.1
B = matrix(rnorm(2 * M), M, 2) %*% chol(Sg / M)
g_m = X_tilde %*% B[, 1]
g_d = X_tilde %*% B[, 2]
e_2 = g_m[idm] + rnorm(N, 0, sqrt(s2_e))
yo_2 = g_d[ido] + e_2
```

Part 6: Fit the wrong direct-only model

R | Julia

```r
m_direct_2 = lmm.aireml(yo_2, K = Goo, verbose = FALSE)
m_direct_2$tau
```

```
[1] 0.7340484
```

Part 7: Fit the maternal model (M-GCTA)

R | Julia

```
Gmm = G[idm, idm]
Dom = G[ido, idm] + G[idm, ido]
m_maternal = lmm.aireml(yo_2, K = list(Gmm, Dom, Goo), verbose
m_maternal$tau
```

```
[1] 0.2851037 0.1212508 0.5192185
```