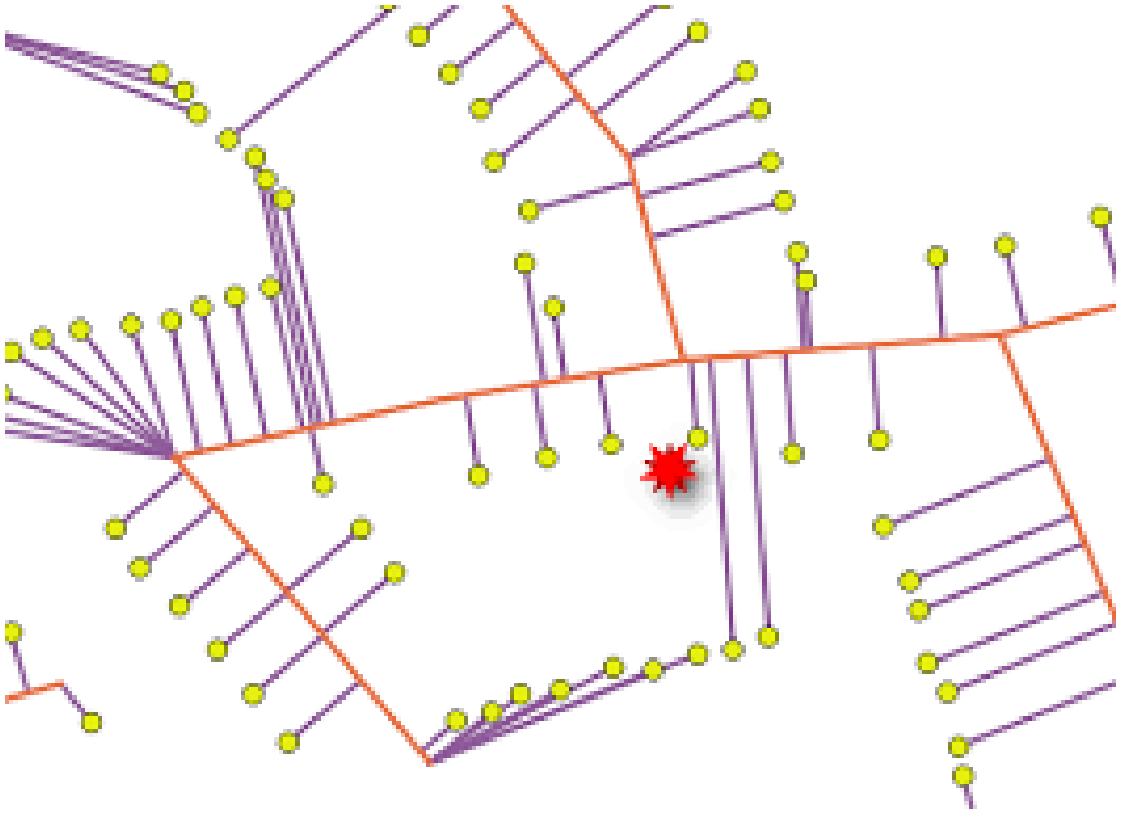
Trio GCTA and Within-family GWAS

Ziada Ayorech, PhD
PROMENTA Research Center
University of Oslo
ziada.ayorech@psykologi.uio.no

Perline Demange, PhD
PROMENTA Research Center
University of Oslo
perlined@uio.no

Day 4 ISPG

Trajectories to ISPG



How to estimate direct and indirect genetic effects?

- Individual trait model:

indirect effects on the phenotype of the index person (e.g child) mediated by phenotype of other individuals

- Limited to the phenotype and traits correlated with the phenotype
- In the case of PGS inherits the issues with PGS and the GWAS weights

Other option?

How to estimate direct and indirect genetic effects?

- Individual trait model:

indirect effects on the phenotype of the index person (e.g child) mediated by phenotype of other individuals

- Variance decomposition model:

avoid specification of the phenotypes that underlie the indirect genetic effects, instead quantifying the total contributions from these effects while being agnostic as to the underlying mechanisms

How to estimate direct and indirect genetic effects?

- Individual trait model:
  - Based on PGS:
    - PGS-SEM
    - Trios/Non-transmitted
    - Adoption
    - Within-sibling

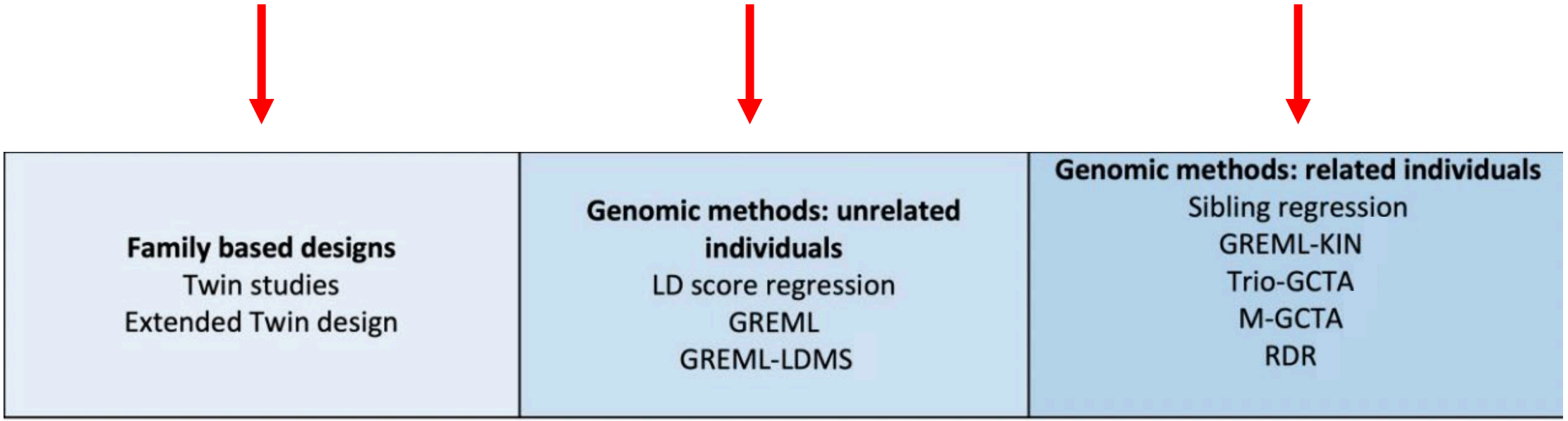
- Variance decomposition model:
  - Twin models (+ PGS)
  - GCTA models:
    - RDR
    - MGCTA
    - TrioGCTA
  - GWAS of direct and indirect genetic effects
    - Within-sibling GWAS
    - with SEM modeling

How to estimate direct and indirect genetic effects?

- Individual trait model:
  - Based on PGS:
    - PGS-SEM
    - Trios/Non-transmitted
    - Adoption
    - Within-sibling

- Variance decomposition model:
  - Twin models (+ PGS)
  - GCTA models:
    - RDR
    - MGCTA
    - TrioGCTA
  - GWAS of direct and indirect genetic effects
    - Within-sibling GWAS
    - with SEM modeling

The big picture of heritability estimation



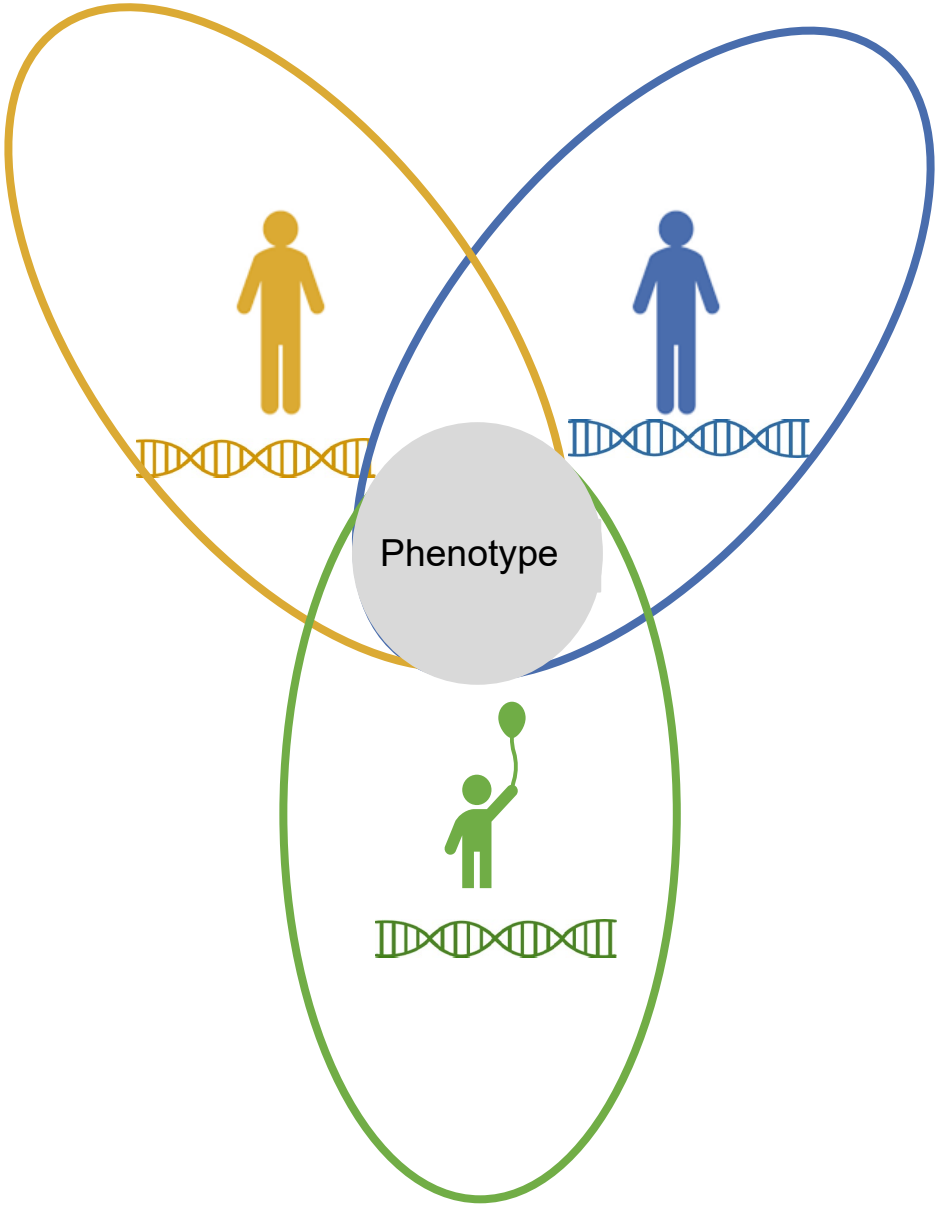
A shared limitation of genomic methods applied to unrelated individuals is an inability to account for environmental confounding

Adapted from Barry, C. J. S., Walker, V. M., Cheesman, R., Davey Smith, G., Morris, T. T., & Davies, N. M. (2023).. *International Journal of Epidemiology*, *52*(2), 624-632.

How to estimate direct and indirect genetic effects?

each a product of our genes and our environments

our environments include the relatives around us



TrioGCTA

Eilertsen, E. et al., (2021). *Behavior Genetics*, *51*, 154-161.

How did we get here

- GREML GCTA on unrelated individuals
- MGCTA
- Trio GCTA

GREML GCTA

In 2010, Yang et al. introduced the G-REML framework:

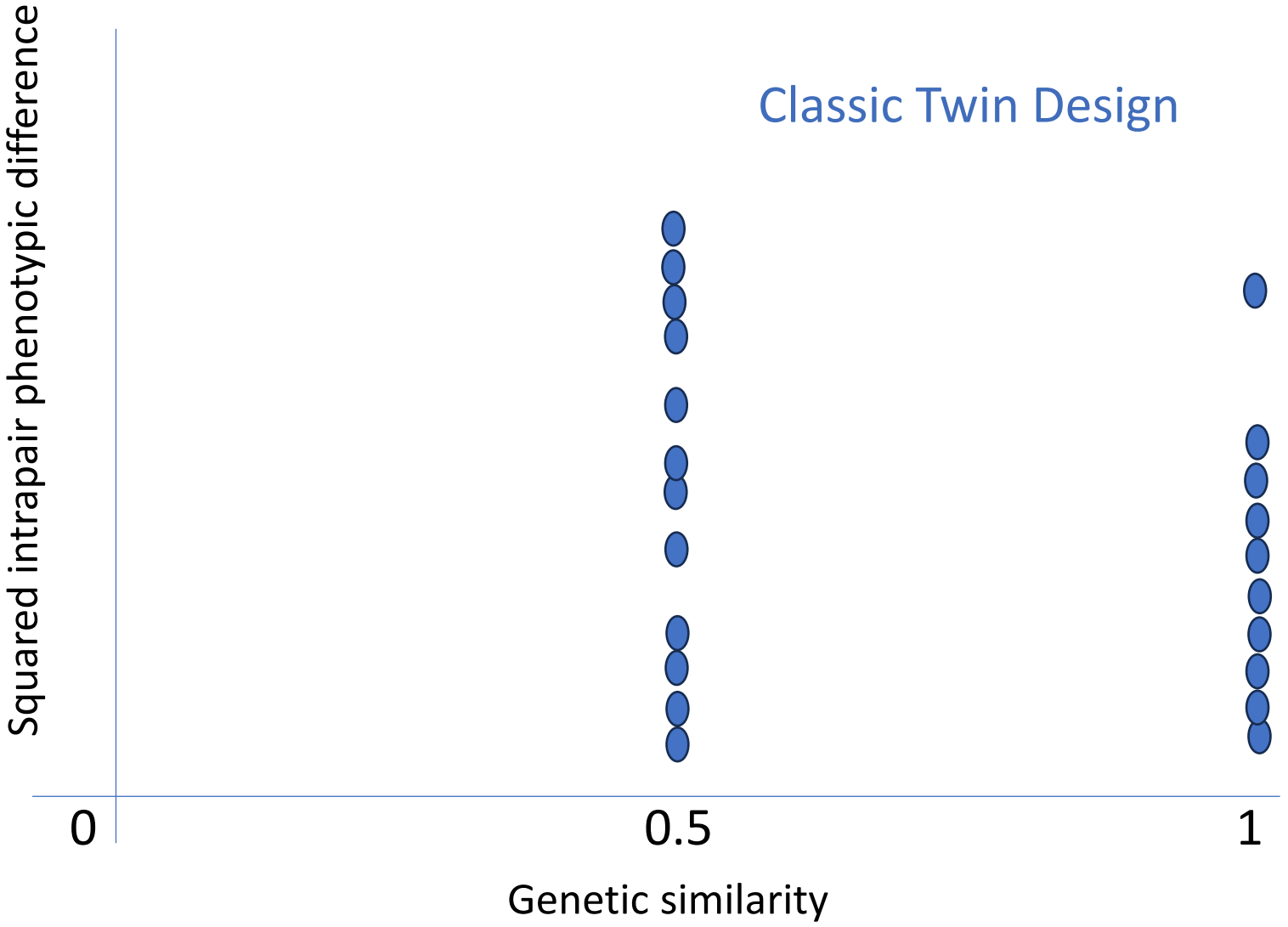
genome-wide genetic similarity between unrelated individuals is used to partition phenotypic variance into genetic components of variation explained by tagged SNPs and residual sources of variation

First new quantitative genetic analytic method in 100 year's

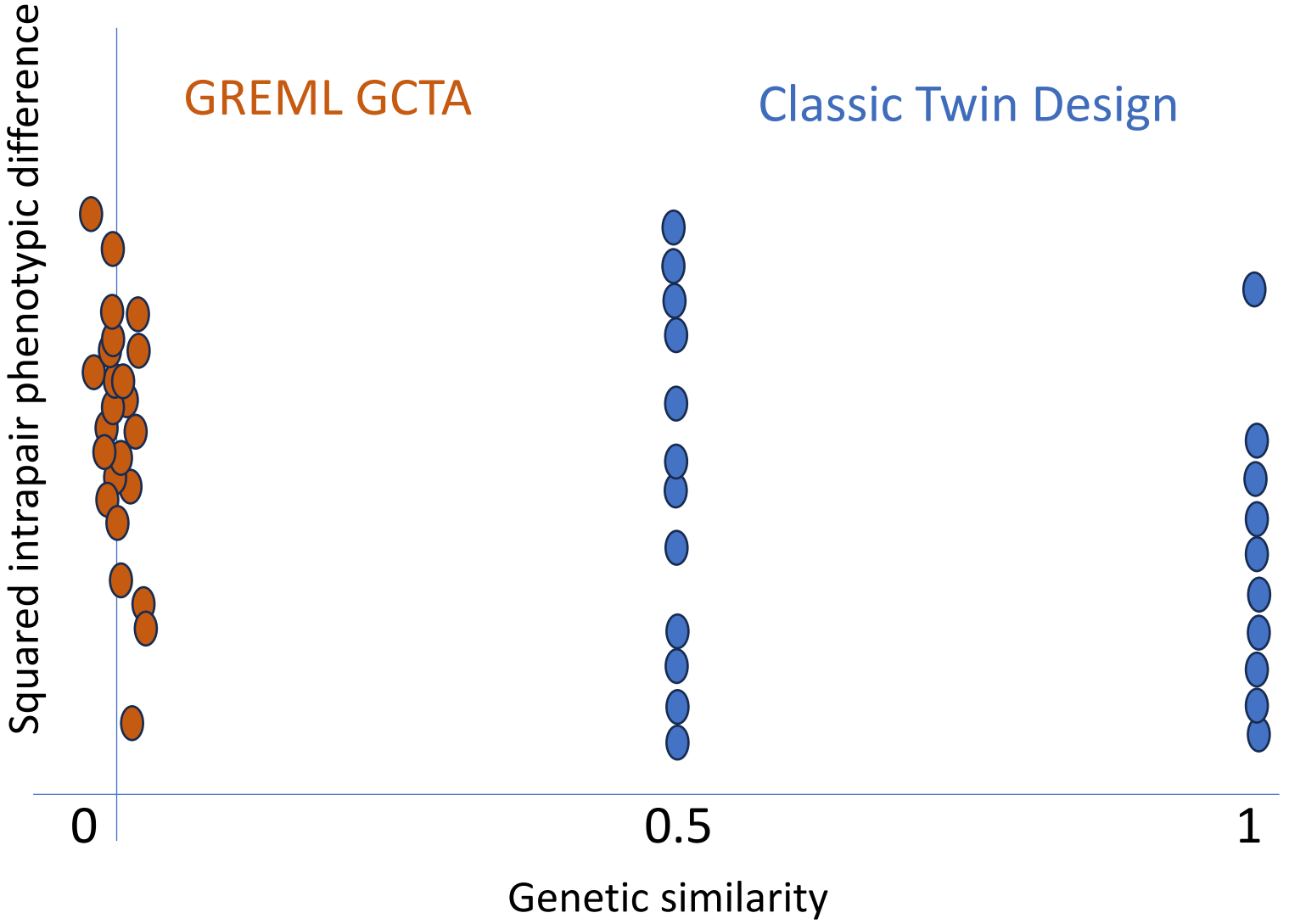
Yang J, et al., (2010). Nat Genet 42:565–569

GREML GCTA intuition

uses unrelated individuals who vary randomly in genetic similarity by chance



Adapted from Dave Evans M-GCTA ISPG slides



Builds intuition on why the sample size requirements for twin versus greml GCTA models is so different

Adapted from Dave Evans M-GCTA ISPG slides

GREML GCTA

Practically, this method involves two steps:

1) Estimating a genetic relatedness matrix:
2) Estimate variance components

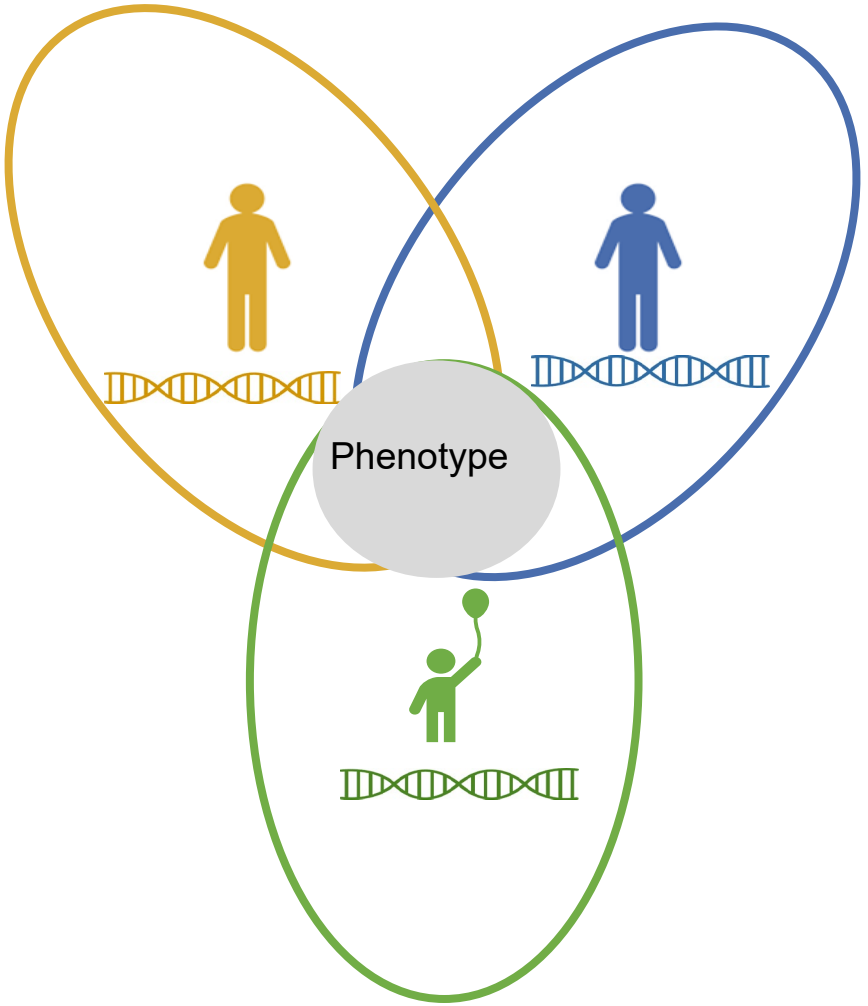
Partitions phenotypic variance into proportion explained by tagged SNPs and residual variance

Quantifies the pairwise genetic relatedness between pairs of individuals in the sample

Key assumptions of GREML GCTA on unrelated

1. All genetic effects are direct
2. Individuals do not share environmental influences
3. GREML only captures direct, additive effects of common SNPs
4. random mating
5. strong assumptions about genetic architecture
6. true genetic effect variance–covariance structure between pairs of individuals is known

the same genetic variants can have both direct and indirect effects, failing to account for the indirect genetic effects of relatives when attempting to measure heritability can result in misleading quantifications of the importance of direct genetic effects

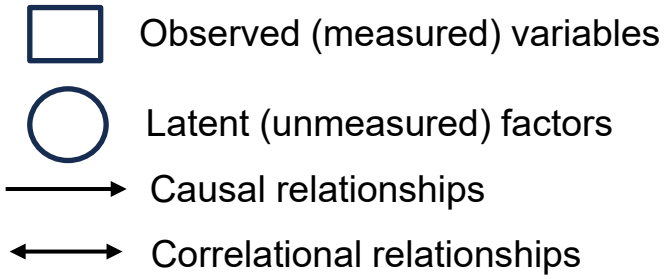


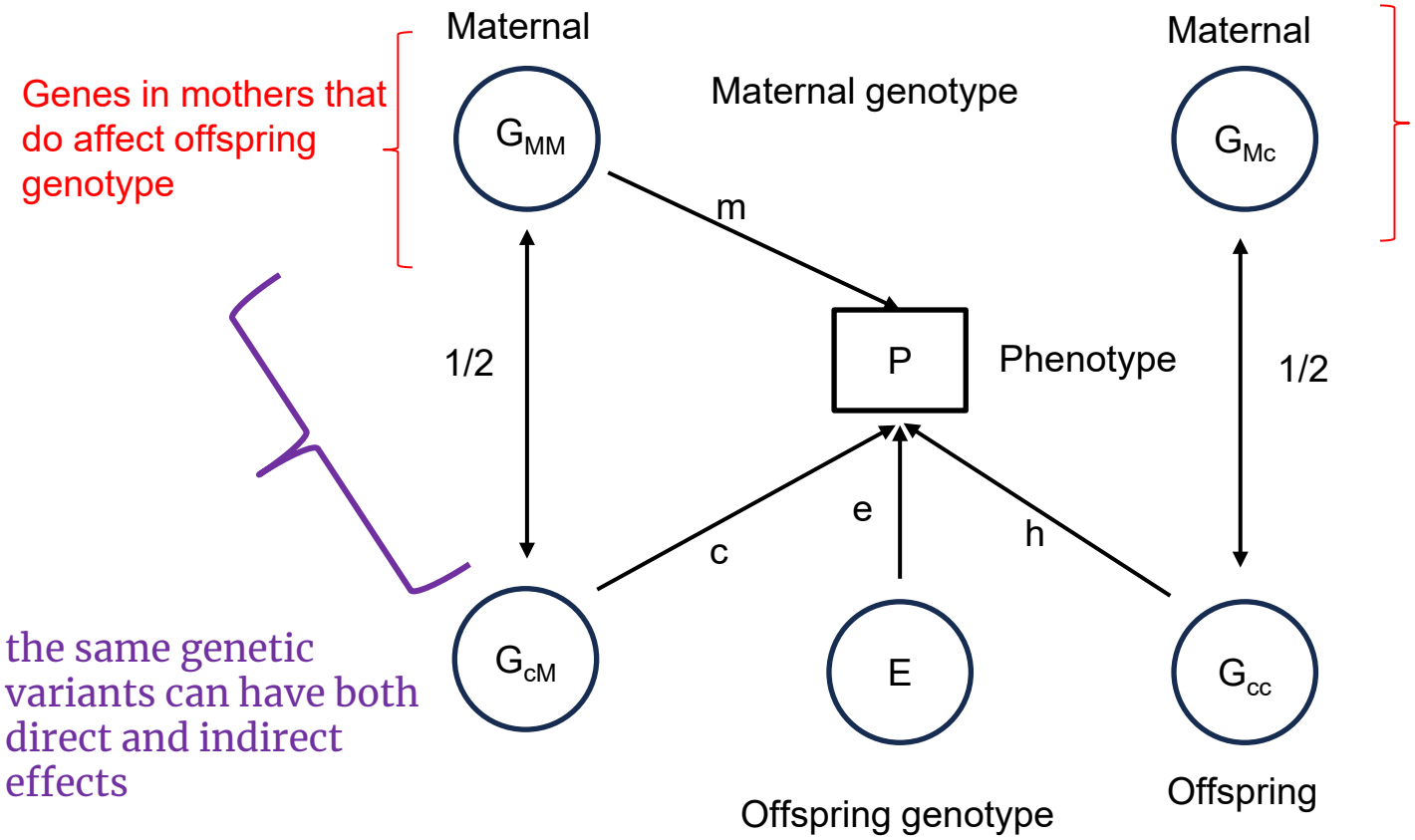
MGCTA

Extends GREML GCTA to estimate the proportion of the child's phenotypic variance explained by the maternal (or paternal) and offspring genotype

Eaves LJ, Pourcain BS, Smith GD, York TP, Evans DM (2014). Behav Genet 44:445–455
Qiao, Z., Zheng, J., Helgeland, Ø., Vaudel, M., Johansson, S., Njølstad, P. R., ... & Evans, D. M. (2020).. *Behavior Genetics*, *50*, 51-66.

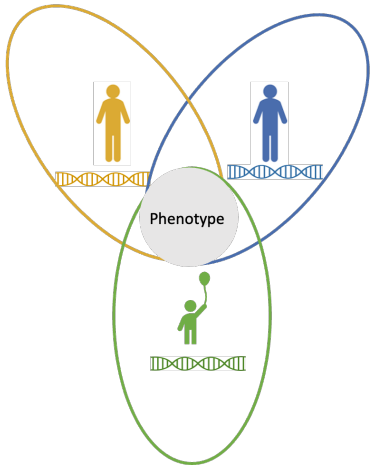
MGCTA





These genes don't have affect on offspring phenotype when in mother

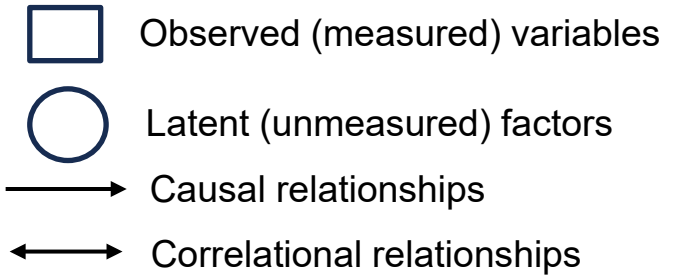
One parent at a time

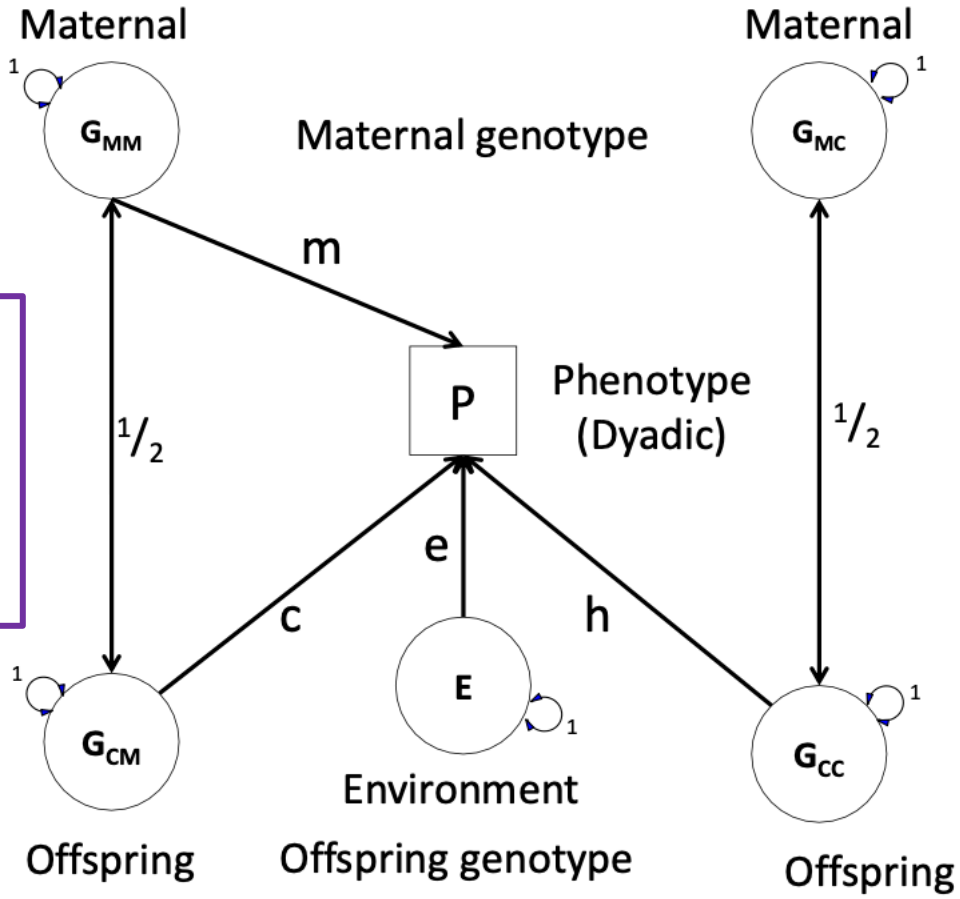


Eaves LJ, Pourcain BS, Smith GD, York TP, Evans DM (2014). Behav Genet 44:445–455

Qiao, Z., Zheng, J., Helgeland, Ø., Vaudel, M., Johansson, S., Njølstad, P. R., ... & Evans, D. M. (2020).. *Behavior Genetics*, *50*, 51-66.

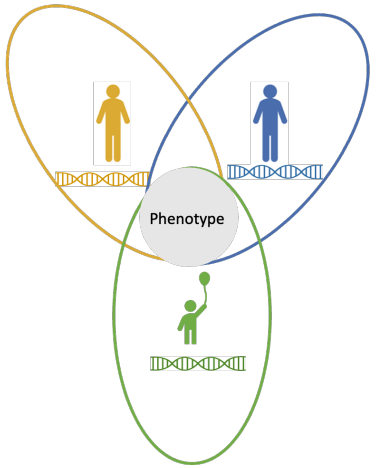
MGCTA





Double headed arrows around latent factors to emphasize every latent factor is standardized to unit variance

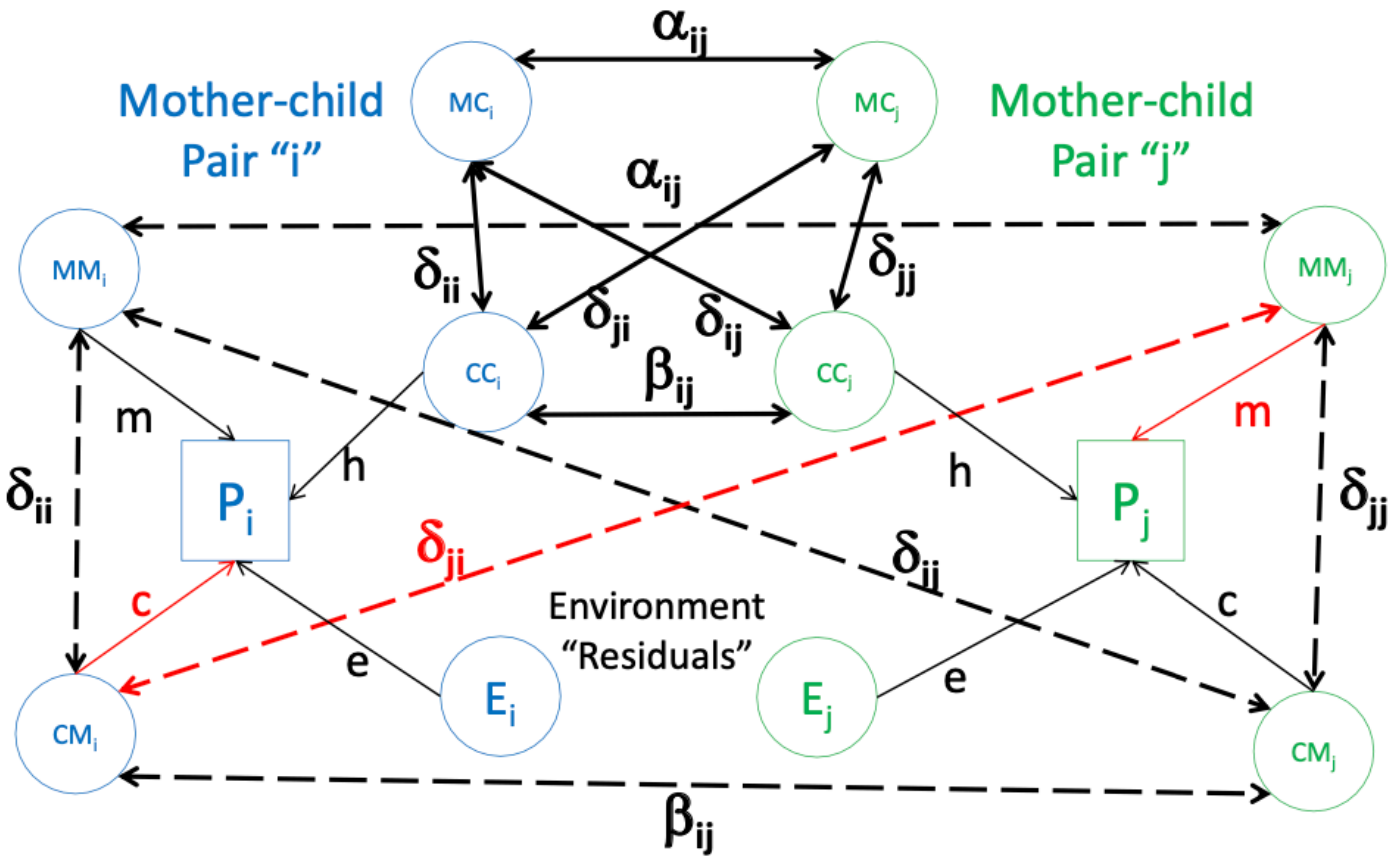
One parent at a time



Eaves LJ, Pourcain BS, Smith GD, York TP, Evans DM (2014). Behav Genet 44:445–455

Qiao, Z., Zheng, J., Helgeland, Ø., Vaudel, M., Johansson, S., Njølstad, P. R., ... & Evans, D. M. (2020).. *Behavior Genetics*, *50*, 51-66.

MGCTA



Observed (measured) variables

Latent (unmeasured) factors

Causal relationships

Correlational relationships

$$\text{cov}(P_i, P_j) = \alpha_{ij}m^2 + \beta_{ij}(c^2 + h^2) + (\delta_{ij} + \delta_{ji})mc$$

$$\text{VAR}(P) = m^2 + c^2 + h^2 + mc + e^2$$

The maths and path tracing are described beautifully here

https://www.colorado.edu/ibg/international-workshop/2022-international-statistical-genetics-workshop/syllabus/m-gcta-trio-gcta

Eaves LJ, Pourcain BS, Smith GD, York TP, Evans DM (2014). Behav Genet 44:445–455

Qiao, Z., Zheng, J., Helgeland, Ø., Vaudel, M., Johansson, S., Njølstad, P. R., ... & Evans, D. M. (2020).. *Behavior Genetics*, *50*, 51-66.

TRIO GCTA

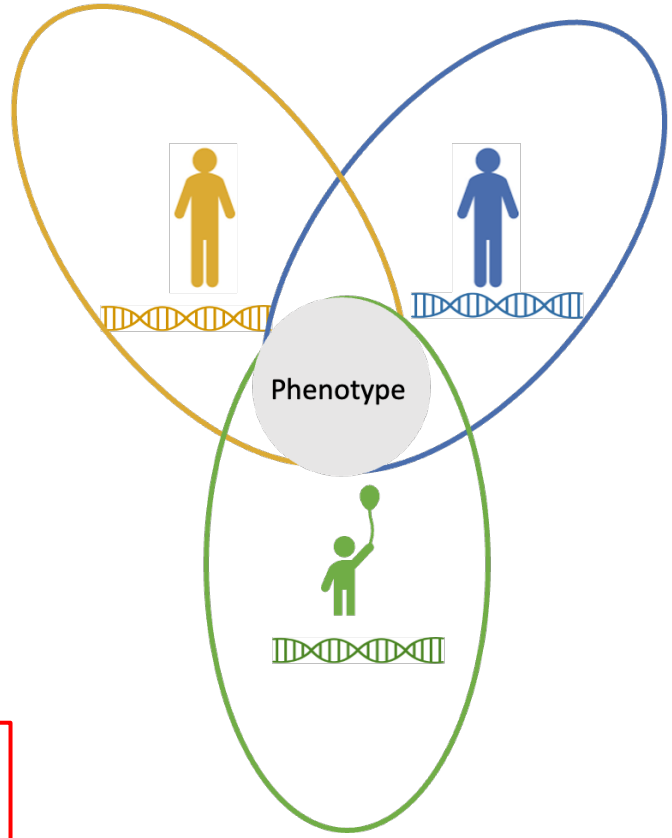
Very similar to MGCTA except can simultaneously model both parents

Direct and Indirect Effects of Maternal, Paternal, and Offspring Genotypes: Trio–GCTA

Espen Moen Eilertsen [1 2 3], Eshim Shahid Jami [4], Tom A McAdams [5 6], Laurie J Hannigan [7], Alexandra S Havdahl [8 7 9 6], Per Magnus [10], David M Evans [9 11], Eivind Ystrom [8 6 12]

Affiliations + expand

PMID: 33387132 DOI: 10.1007/s10519-020-10036-6



Model is applicable to any family member being the ‘focal individual’

Trio GCTA Step 1: estimate GRM

X is now the dosage (0,1,2) of a SNP

Mean is now the average genotype

$$A_{jk} = \frac{1}{M}\sum_{i=1}^{M}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}.$$

$$r = \frac{1}{N}\frac{\sum(X-\bar{X})(Y-\bar{Y})}{SD_X SD_Y}$$

A_{jk} is the genomic relatedness between individuals j and k

x_{ij} is the dosage $\{x_{ij} = 0,1,2\}$ of the reference allele for SNP i for individual j

x_{ik} is the dosage $\{x_{ik} = 0,1,2\}$ of the reference allele for SNP i for individual k

p_i is the frequency of the reference allele for SNP i

M is the number of markers across the genome

The formula for the GRM is analogous to the formula for a pearson correlation

Think of A_{jk} as like an average genetic correlation across the genome for individuals j and k

The GRM formula is described beautifully here

https://www.colorado.edu/ibg/isg-workshop/online-lectures-and-practicals/g2-gcta-greml-m-gcta-part-2-gcta-genetic-relationship

Trio GCTA Step 1: estimate GRM

n

n

	ID1	ID2	ID3	ID4	ID5	ID6
ID1	1.000000000	0.064615312	0.01451343	0.007633084	0.05972810	0.025441527
ID2	0.07329026	1.000000000	0.06902873	0.002975003	0.02708290	0.002633122
ID3	0.03380476	0.075339151	1.00000000	0.078313322	0.02402007	0.029352703
ID4	0.09942796	0.090551806	0.02848736	1.000000000	0.09833939	0.025873929
ID5	0.05864953	0.004237787	0.06911177	0.074792206	1.00000000	0.098110961
ID6	0.00537572	0.013444390	0.07103385	0.065866721	0.03474527	1.000000000

n x n matrix capturing the degree of relatedness between every pair of individuals at every SNP location

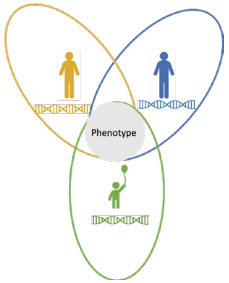
Crucially!! Your trio data are organised with all mothers, stacked on top of all fathers, who are stacked on top of all offspring

Trio GCTA additional Step 1: subset GRM into blocks

	ID1	ID2
ID1	1.00000000	0.06461531
ID2	0.07329026	1.00000000

	ID3	ID4
ID3	1.00000000	0.07831332
ID4	0.02848736	1.00000000

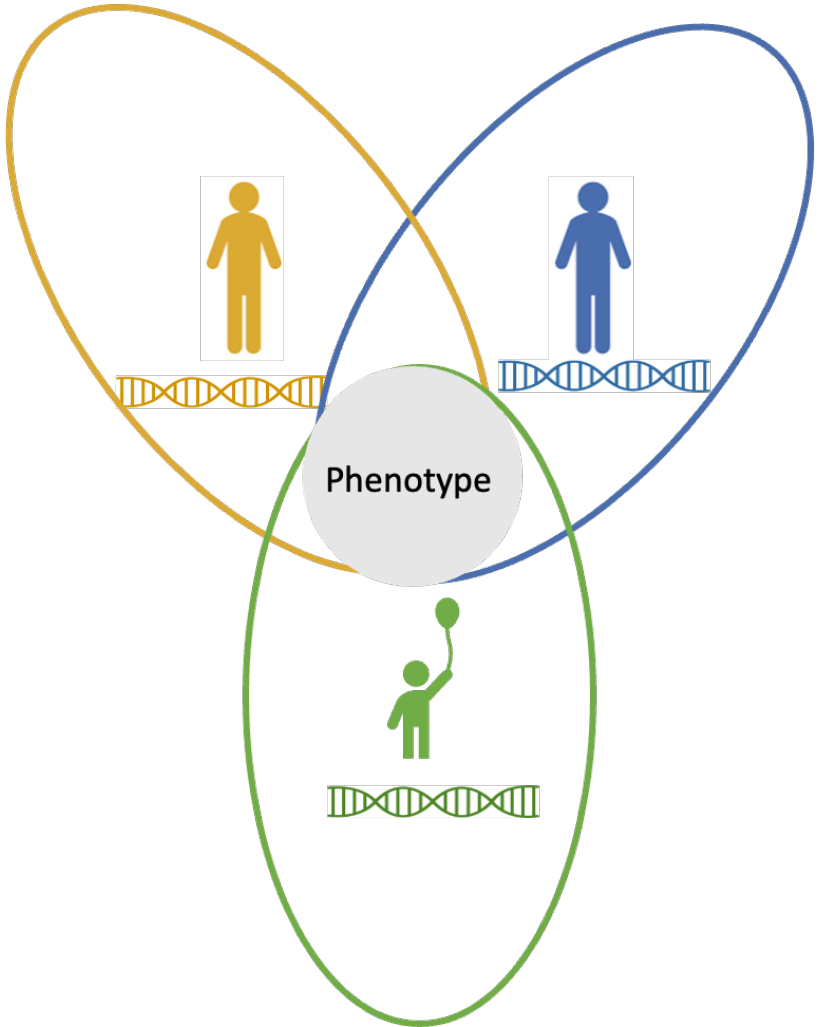
	ID5	ID6
ID5	1.00000000	0.09811096
ID6	0.03474527	1.00000000

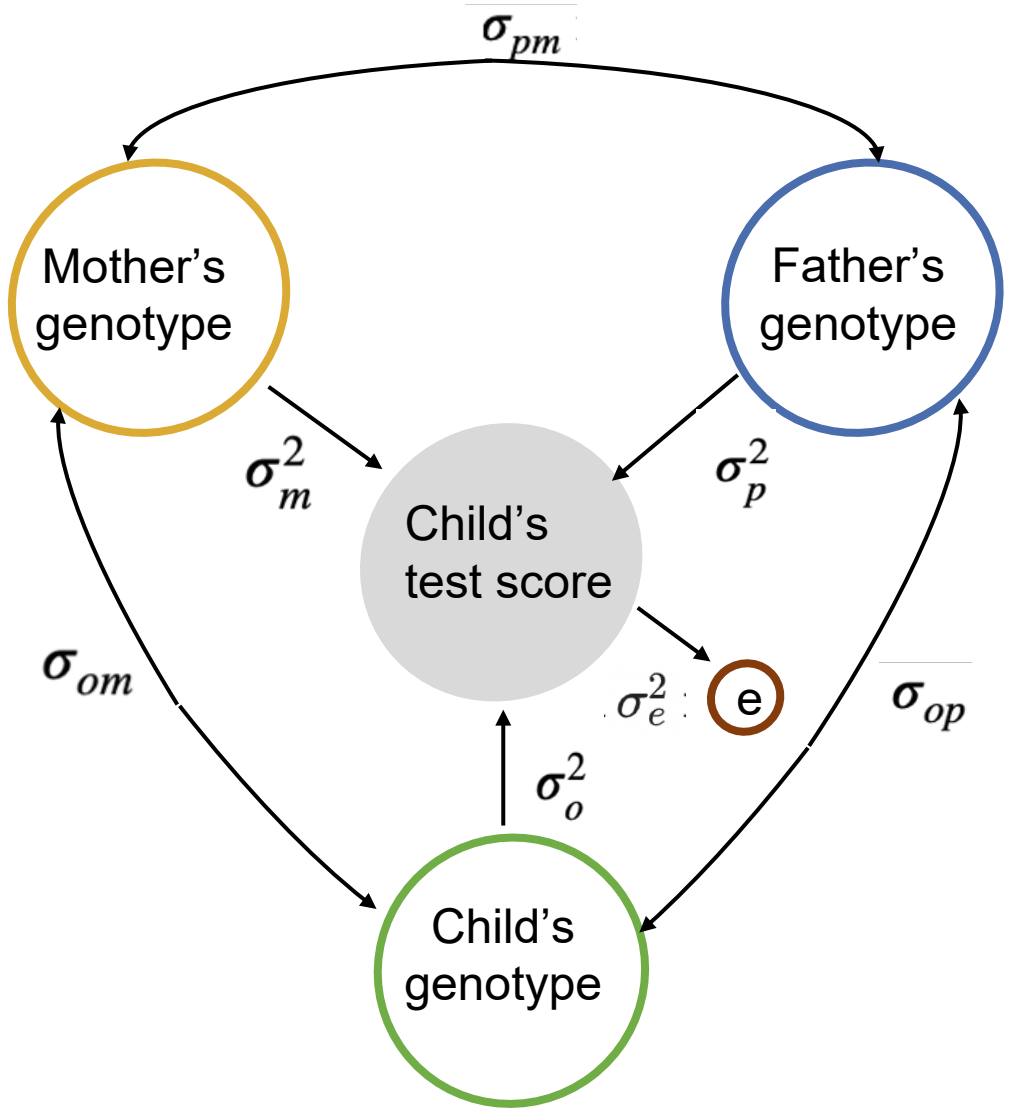


mid = 1:2
pid = 3:4
oid = 5:6
Amm = A[mid, mid]
App = A[pid, pid]
Aoo = A[oid, oid]
Dpm = A[pid, mid] + A[mid, pid
Dom = A[oid, mid] + A[mid, oid]
Dop = A[oid, pid] + A[pid, oid]

These blocks define the correlation structure of each family members effects and will be used to hold their relative direct and indirect effects and their covariances

GREML Step 2 : estimate variance components

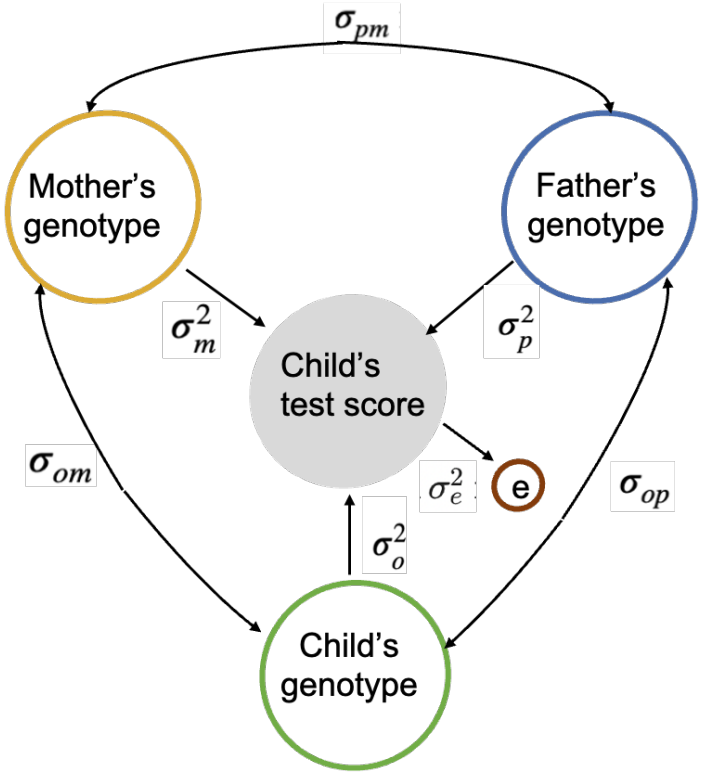




GREML Step 2 : estimate variance components

The extent to which the same SNPs contribute to direct and indirect effects

$$\mathrm{Var}(y_k) = \sigma^2_m + \sigma^2_p + \sigma^2_o + \overbrace{\sigma_{om} + \sigma_{op}} + \sigma^2_e.$$



Direct or Indirect depending on focal individual

we do not include the covariance between the father and mother in the calculation of the total variance. We do this under the assumption that parents are unrelated and have an expected genetic relatedness of 0.

Expected covariance structure of phenotype across all individuals

M = SNPs available for all individuals

Multiply by transpose*

$$\mathrm{Cov}(\mathbf{y}) = \frac{\sigma^2_m}{M}\mathbf{Z}_m\mathbf{Z}_m^\top + \frac{\sigma^2_p}{M}\mathbf{Z}_p\mathbf{Z}_p^\top + \frac{\sigma^2_o}{M}\mathbf{Z}_o\mathbf{Z}_o^\top + \frac{\sigma_{om}}{M}(\mathbf{Z}_o\mathbf{Z}_m^\top + \mathbf{Z}_m\mathbf{Z}_o^\top) + \frac{\sigma_{op}}{M}(\mathbf{Z}_o\mathbf{Z}_p^\top + \mathbf{Z}_p\mathbf{Z}_o^\top) + \frac{\sigma_{pm}}{M}(\mathbf{Z}_p\mathbf{Z}_m^\top + \mathbf{Z}_m\mathbf{Z}_p^\top) + \sigma^2_e\mathbf{I}.$$

matrices of maternal, paternal and offspring standardized genotype dosages are represented by Z_m, Z_p, Z_o, respectively.

*Matrix algebra to allow us to work with matrices and vectors: we swap rows and columns. If Z_m is a K X M matrix, then Z_m^T is a M X K matrix.

		Maths	Reading	English
Direct genetic effect	σ_o^2	0.32	0.27	0.34
Indirect genetic effect	σ_m^2	0.00	0.00	0.01
	σ_p^2	0.04	0.00	0.05
Covariance	σ_{om}	0.00	0.01	-0.01
	σ_{op}	0.02	0.01	0.01
	σ_{mp}	0.00	0.00	0.02
Residual variance	σ_ϵ^2	0.62	0.71	0.60

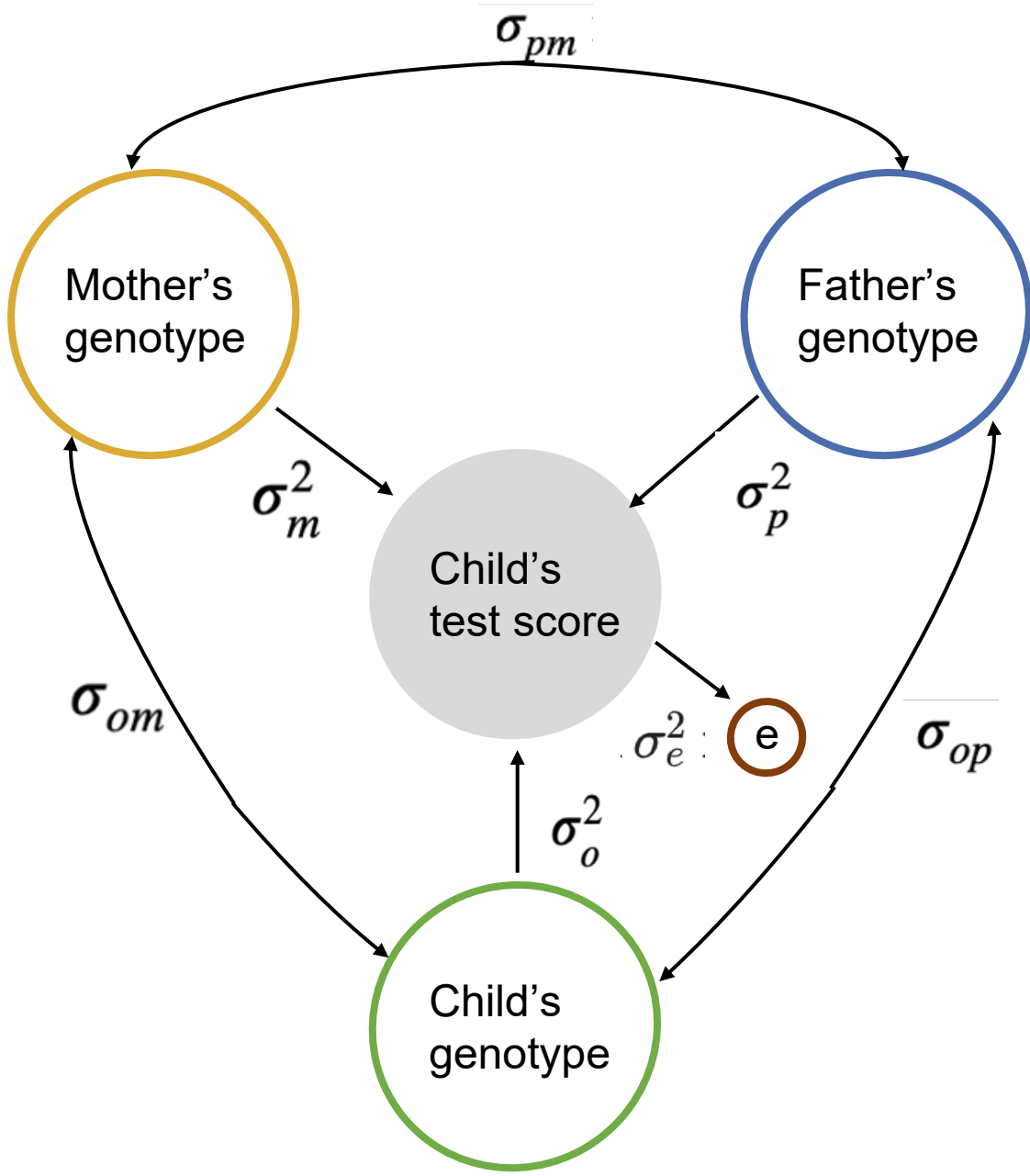


Table 1 A summary of the strengths and limitations of each discussed method

Method	Strengths	Limitations
GREML	Provides a 'lower-bound' estimate for heritability; possible to implement using widely available software; partially accounts for population stratification; can be implemented in large-scale biobanks of unrelated individuals	Cannot distinguish direct and indirect genetic effects, which may inflate estimates; samples restricted to genotyped individuals; estimates inflated by assortative mating; additional assumptions about SNP effect sizes required; not suited to estimate the contribution of rare SNPs; estimates are highly sensitive to LD
M-GCTA[a]	Able to account for indirect genetic effects within heritability estimates; possible to implement using freely available software	Requires genotyped maternal or paternal–offspring pairs; may be biased by assortative mating; large sample size required; cannot simultaneously account for both parental genotypes
Trio-GCTA[a]	Able to account for indirect genetic effects within heritability estimates; assumptions about the structure of LD are not necessary; partially accounts for population stratification	Requires genotyped parent–offspring trios; additional assumptions about the distribution of genetic and residual effects are necessary; assortative mating and epistatic effects will bias estimates

Adapted from Barry, C. J. S., Walker, V. M., Cheesman, R., Davey Smith, G., Morris, T. T., & Davies, N. M. (2023).. *International Journal of Epidemiology*, *52*(2), 624-632.

Now for the practical

https://qimr.az1.qualtrics.com/jfe/form/SV_bjXq3wCxYVR0SEu

```
# Create a directory to hold today's work
mkdir day4_triogcta

# Move into that directory, and then copy over my folder
cd day4_triogcta
cp -r /home/ziada/2024/Day-4/TrioGCTA_practical* ./
```