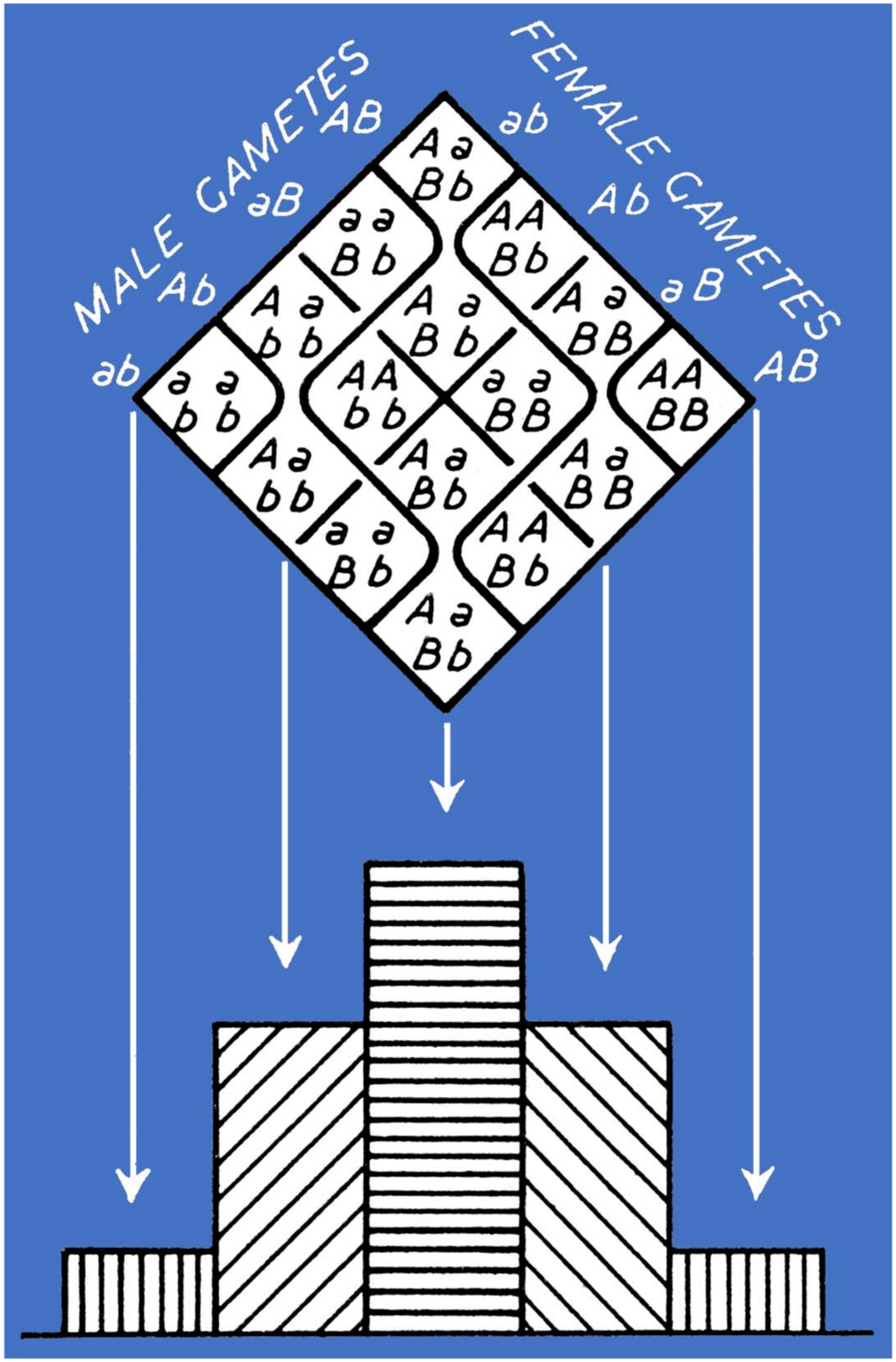
Family based association

Dorret Boomsma, Mike Neale, Conor Dolan, Jouke Jan Hottenga, Jenny van Dongen

International Statistical Genetics Workshop, Boulder Colorado, 2023

Talk about designs that explicitly take into account that data (phenotypes & genotypes) were collected in families / clusters (e.g., pedigrees, families, twin pairs).

NB all papers are in folder Papers

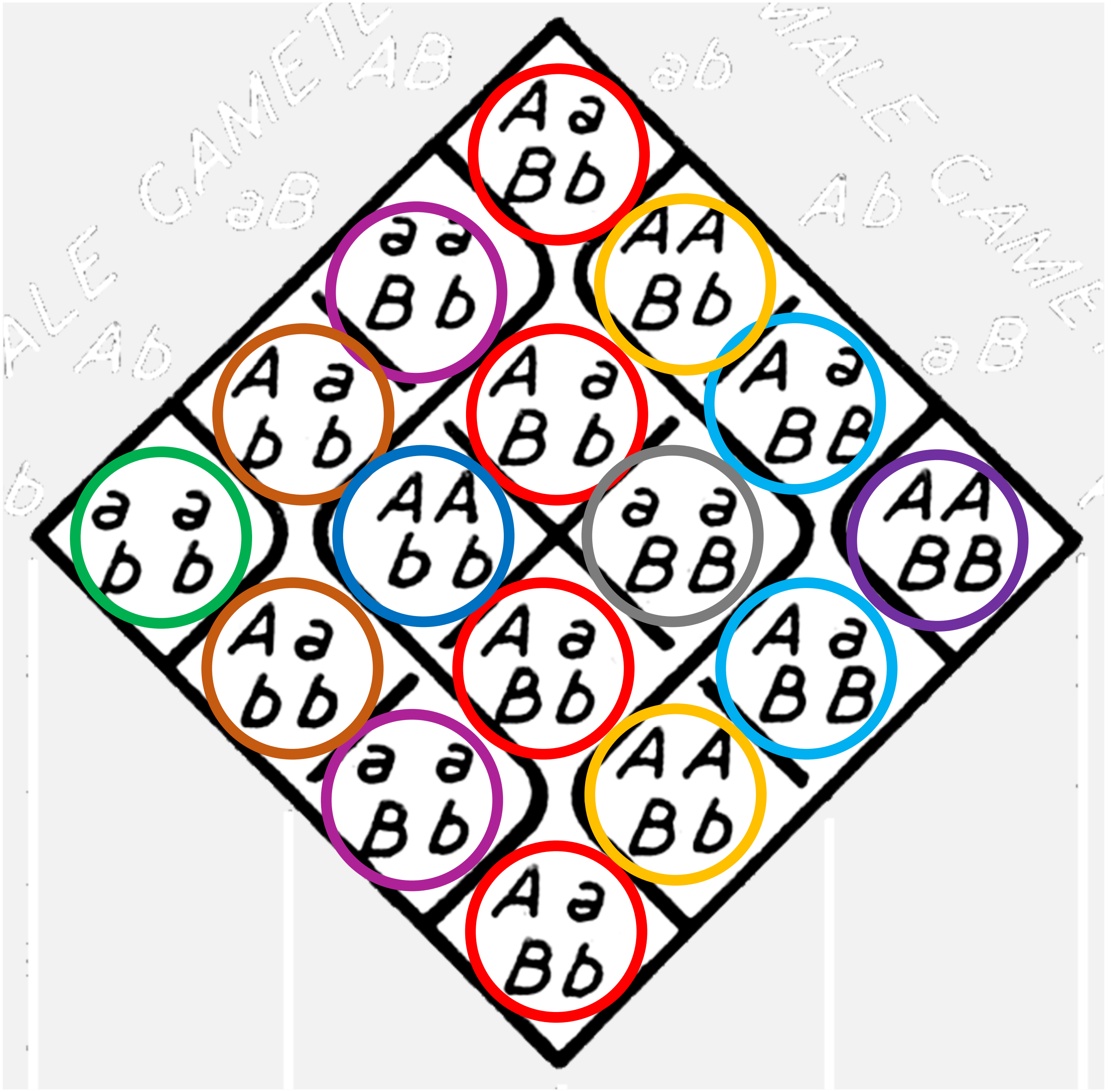


Punnett square

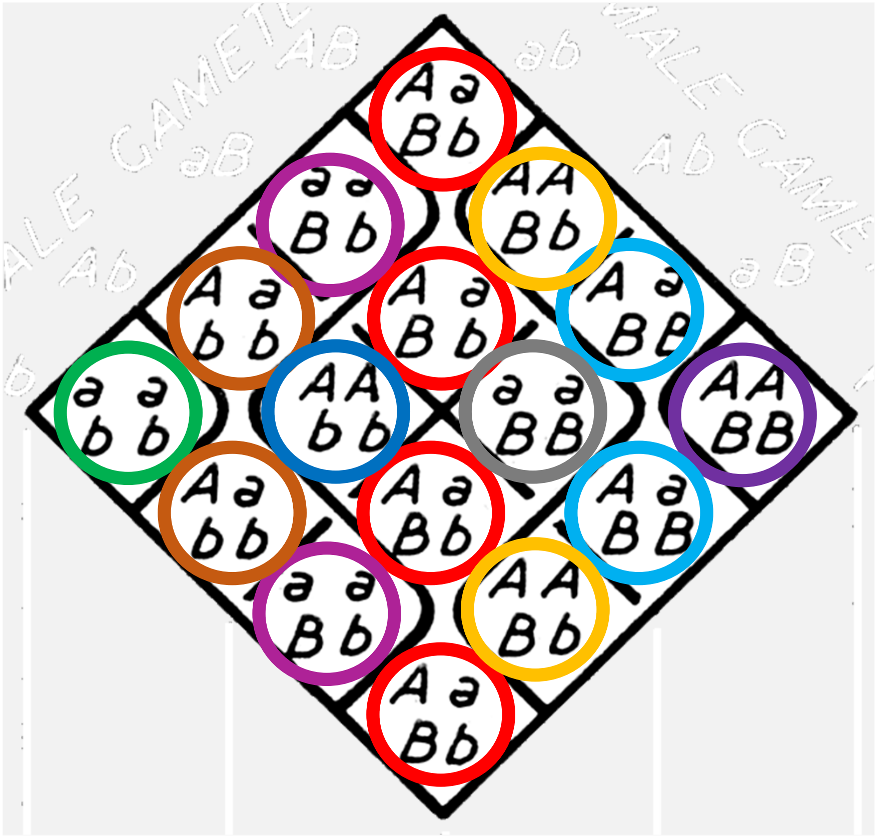
Two genes A and B. Parents are both heterozygotes (AaBb).

Their offspring may have different genotypes.

K Mather, Biometrical Genetics, Dover Publ, 1949



	ab	Ab	aB	AB
AB	AaBb	AABb	AaBB	AABB
aB	aaBb	AaBb	aaBB	AaBB
Ab	Aabb	AAbb	AaBb	AABb
ab	aabb	Aabb	aaBb	AaBb



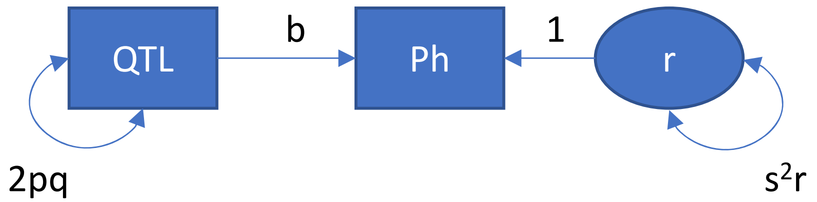
In the population traits of e.g. ab/ab individuals differ from the phenotypes of AB/AB individuals.

Do we see the same differences if these two individuals are siblings?

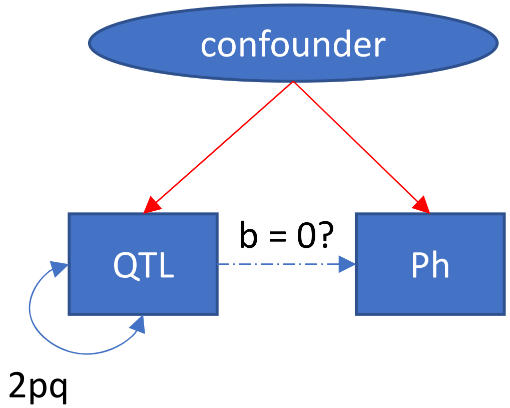
I.e., is variation within families equal to variation between families?

If yes: “true” genetic association
If no: ? (confounding)

Regression model: Phenotype = a + b*QTL + residual



What can go wrong? Stratification



Lindon Eaves
(e.g. Inferring the Causes of Human Variation, 1977)

The genetic and environmental variation is partitioned into within and between family components.

G1=*within* - family genetic component

G2=*between* - family genetic component

E1=*within*-family environment ("E")

E2=*between*-family environment ("C")

In the absence of GE interaction or GE correlation total variance is partitioned into: $\sigma^2 t = \sigma^2 w + \sigma^2 b$, and familial resemblance is: $ICC = \sigma^2 b / (\sigma^2 w + \sigma^2 b)$

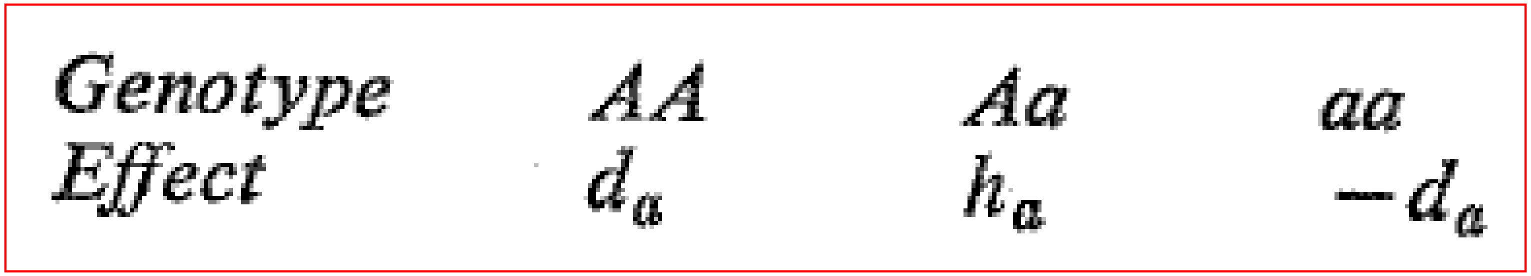
G = ½ Dr + ¼ Hr

(Vg = Va + Vd)

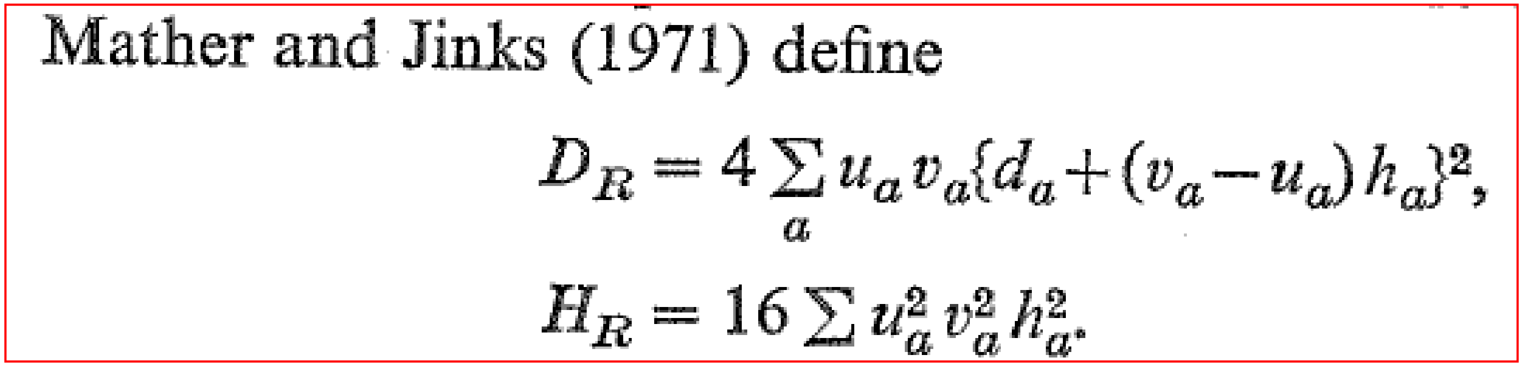
E1 + E2 = E

G1 + G2 = G, and

G1 = G2 if there is random mating / no dominance



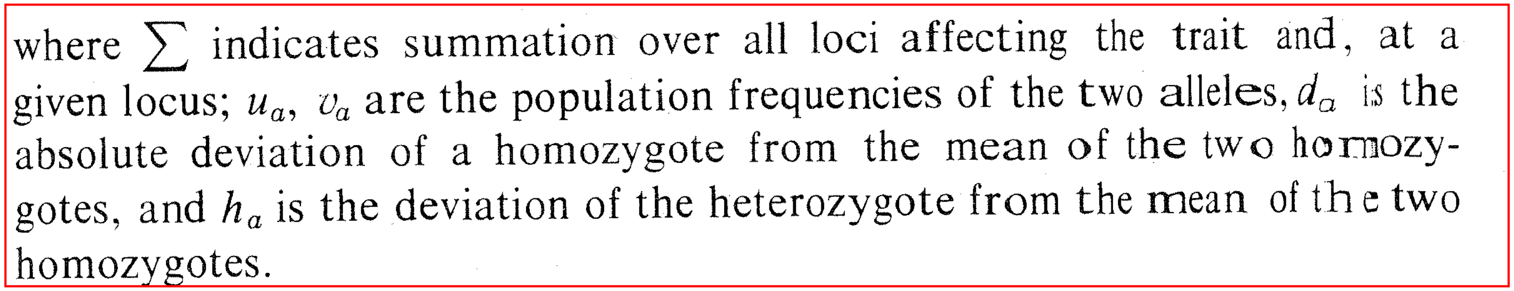
Genotype	*AA*	*Aa*	*aa*
Effect	d_a	h_a	$-d_a$



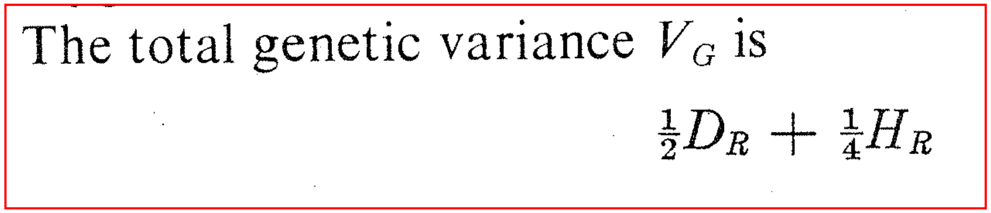
Mather and Jinks (1971) define

$$D_R = 4\sum_a u_a v_a \{d_a + (v_a - u_a) h_a\}^2,$$

$$H_R = 16 \sum u_a^2 v_a^2 h_a^2.$$



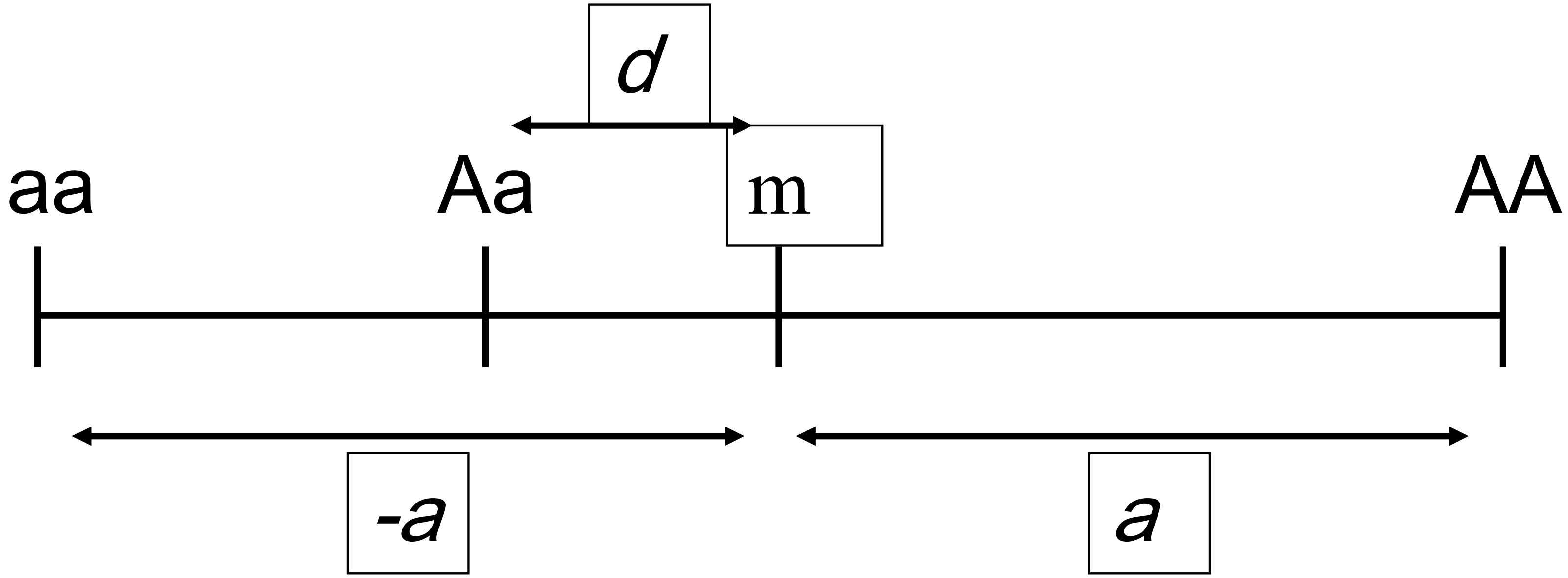
where $\sum$ indicates summation over all loci affecting the trait and, at a given locus; u_a, v_a are the population frequencies of the two alleles, d_a is the absolute deviation of a homozygote from the mean of the two homozygotes, and h_a is the deviation of the heterozygote from the mean of the two homozygotes.



The total genetic variance V_G is

$$\tfrac{1}{2}D_R + \tfrac{1}{4}H_R$$

One locus model: gene with 2 alleles A and a and 3 genotypes AA, Aa and aa



The deviation from m (middle) of the heterozygote Aa is d

$$V_g = V_a + V_d = \quad 2pq[a+d(q-p)]^2 + (2pqd)^2$$

Punnett square: Within family genetic differences

Haseman-Elston: Sib-pair analysis (linkage analysis based on IBD)

Fulker / Posthuma / Neale: combined linkage & B-W association

Selzam et al.: Within- & Between-Family Polygenic Score Prediction

Howe et al.: within sib-pair GWAS

Van Dongen et al.: BMI discordant twins (*practical*)

Family based association. Will not talk about analyses that estimate population association (given clustered data). In such cases:

***Ignore clustering**

***Run analyses per group**

***Robust standard errors (sandwich)**

***GEE (generalized estimating equations)**

***Mixed models / multi-level**

***SEM (structural equation model; ‘case = family’)**

-Hippocrates (5th century BCE) attributed different diseases in twins to different material circumstances.
-Posidonius (1st century BCE) attributed similarities to shared astrological circumstances.
-Augustinus (354-430): If twins are born at the same time how can some twins become so different from each other?

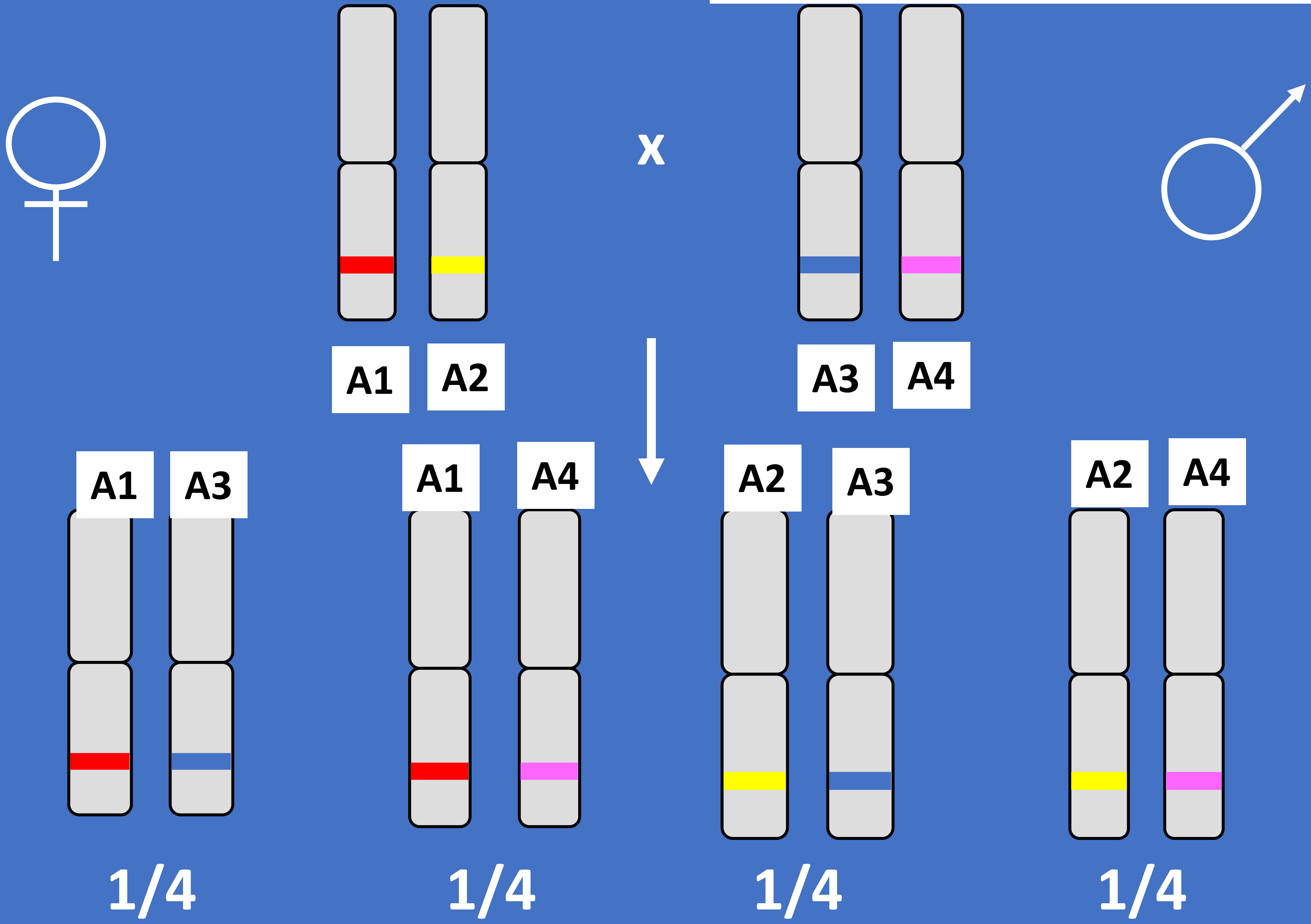
DISCORDANT MZ TWIN DESIGN

Gustav III, King of Sweden: study of dangers of tea and coffee consumption. He commuted death sentences of a pair of twin murderers if they participated in a trial.

They spent the rest of their lives in prison: **one twin drank 3 pots of coffee and the other 3 pots of tea each day.** The tea drinking twin died first at the age of 83, long after Gustav III, who was assassinated in 1792. The age of death of the coffee-drinking twin is not known, as both doctors assigned by the king to monitor this study predeceased him.

1746 – 1792

Color coding identifies origin of alleles



IDENTITY BY DESCENT (IBD)

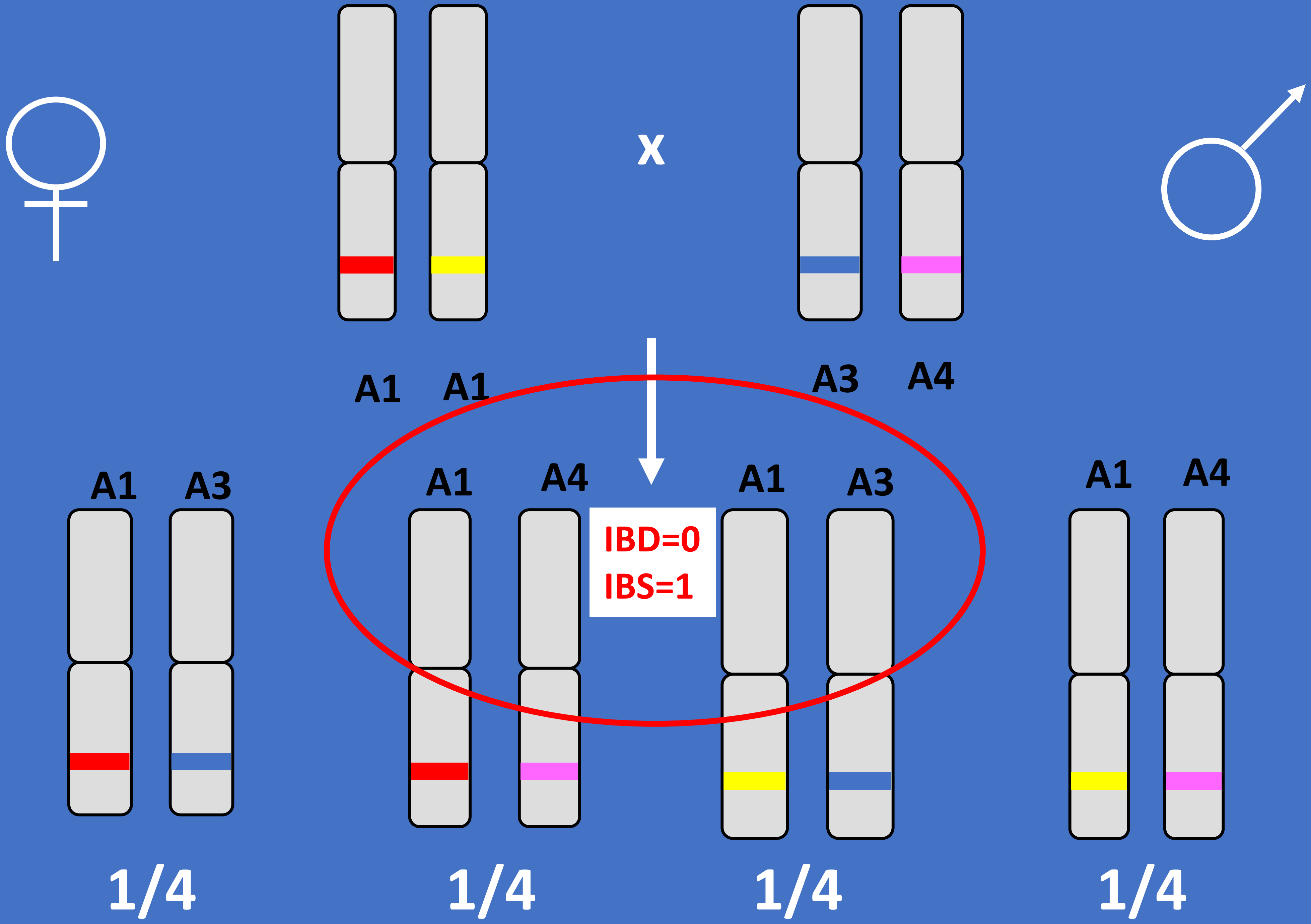
		Sib 1			
		A1 A3	A1 A4	A2 A3	A2 A4
Sib 2	A1 A3	2	1	1	0
	A1 A4	1	2	0	1
	A2 A3	1	0	2	1
	A2 A4	0	1	1	2

■ 4/16 = 1/4 sibs share BOTH parental alleles IBD = 2

▒ 8/16 = 1/2 sibs share ONE parental allele IBD = 1

□ 4/16 = 1/4 sibs share NO parental alleles IBD = 0

Color = IBD, number= IBS



IBD=0

IBS=1

IBD mapping: Sib-pair design to localize QTLs

(QTL = Quantitative Trait Locus)

- Multiple ‘families’ of two (or more) siblings
- Phenotypes on siblings
- Marker genotypes on sibs (& parents)

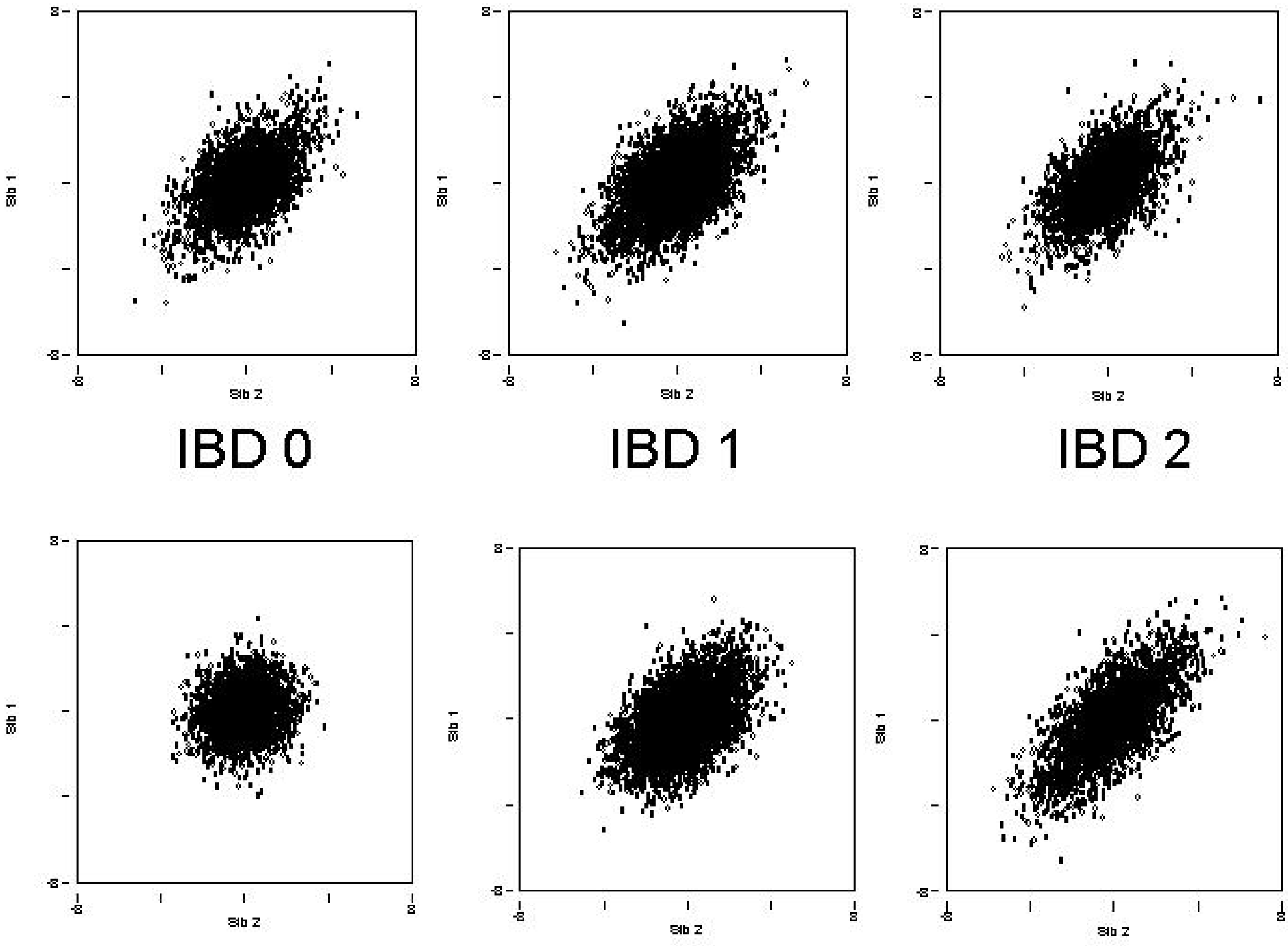
Haseman-Elston regression (1972)

The more alleles pairs of relatives share at a QTL, the greater their phenotypic similarity (IBD 2 more similar than IBD 1 or 0).

Or

The more alleles they share IBD, the smaller the difference in their phenotype.

Sib1-Sib2 distributions for a quantitative phenotype



No linkage

IBD 0

IBD 1

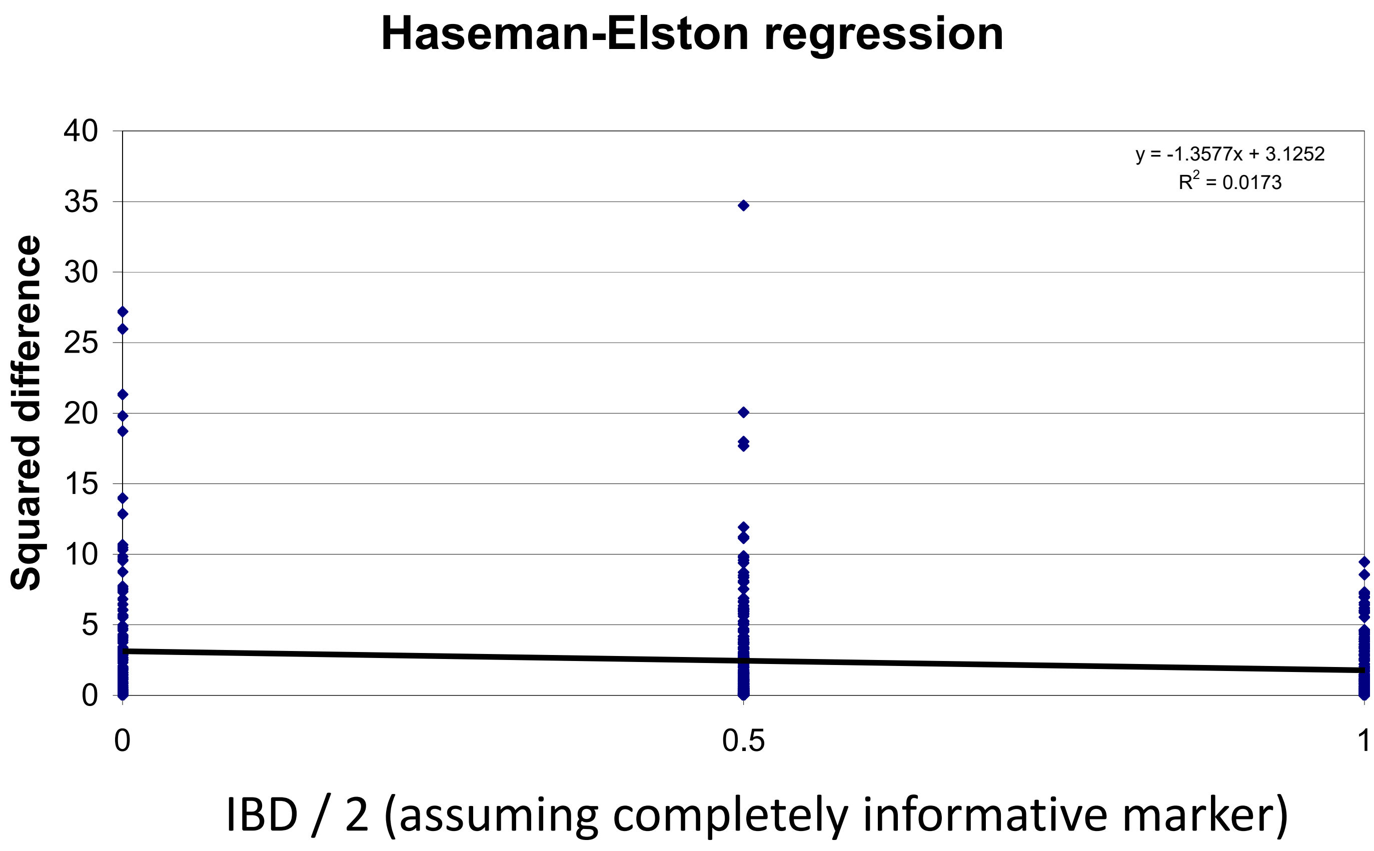
IBD 2

Under linkage

IBD 0

IBD 1

IBD 2



A significant negative slope in a regression analysis (as shown) indicates linkage to a QTL (single marker)

$\beta = \quad -2(1 - 2r)^2\, \sigma_q^{\,2}$ $\qquad\qquad \sigma_q^{\,2}$ = variance due to QTL

$\alpha = \quad 2[1 - 2(1-r)r]\, \sigma_q^{\,2} + \sigma_\varepsilon^{\,2}$

r = recombination fraction between marker & QTL

Test: $\beta < 0$?

QTL (quantitative trait locus) as a random effect

$$\mathbf{y_i \quad = \quad m \quad + \quad Q_i \quad + \quad A_i \quad + \quad E_i}$$

y_i = phenotype (for person i)

m = mean

Q_i = QTL genotype contribution for a chromosomal segment

A_i = Contribution from rest of genome

E_i = residual

$$\mathbf{var(y) \quad = \quad s_q^2 \quad + \quad s_a^2 \quad + \quad s_e^2}$$

Genetic covariance between relatives

$$\mathrm{cov}(y_i, y_j) = p_{ij}\, s_q^{\,2} + a_{ij}\, s_a^{\,2}$$

a_{ij} = average prop. of alleles shared in genome (kinship coefficient (e.g. 0.5 for DZ twins))

p_{ij} = proportion of alleles IBD at QTL (0, ½ or 1) (pi-hat)

= Pr(2 alleles IBD) + ½Pr(1 allele IBD)

p_{ij}

= proportion of alleles IBD in non-inbred pedigree

Estimate with genetic markers (advantageous to have parental genotypes)

Am. J. Hum. Genet. 64:259–267, 1999

Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits

D. W. Fulker,[1,2] S. S. Cherny,[1,2] P. C. Sham,[2] and J. K. Hewitt[1]

[1]Institute for Behavioral Genetics, University of Colorado, Boulder; and [2]Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, University of London, London

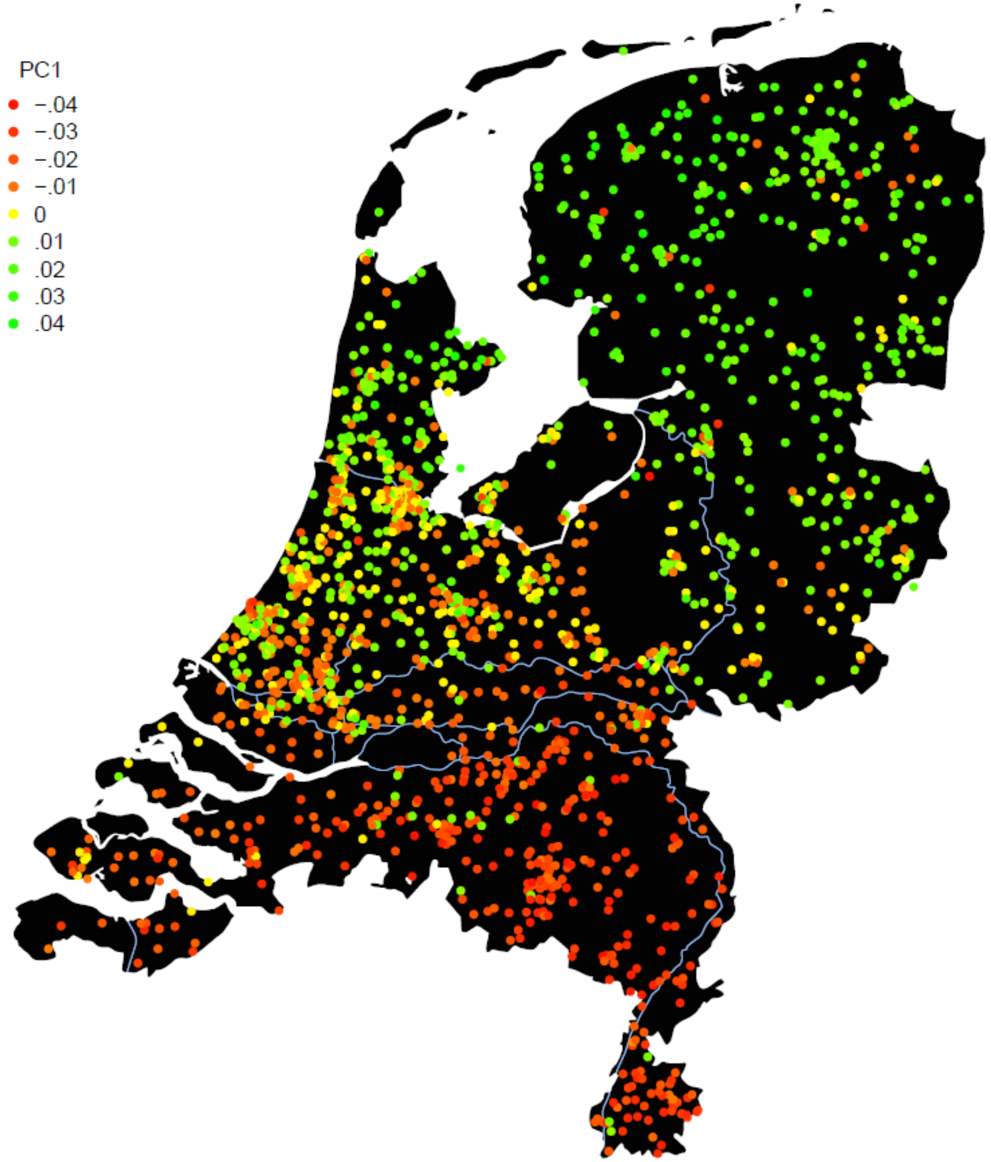
Summary

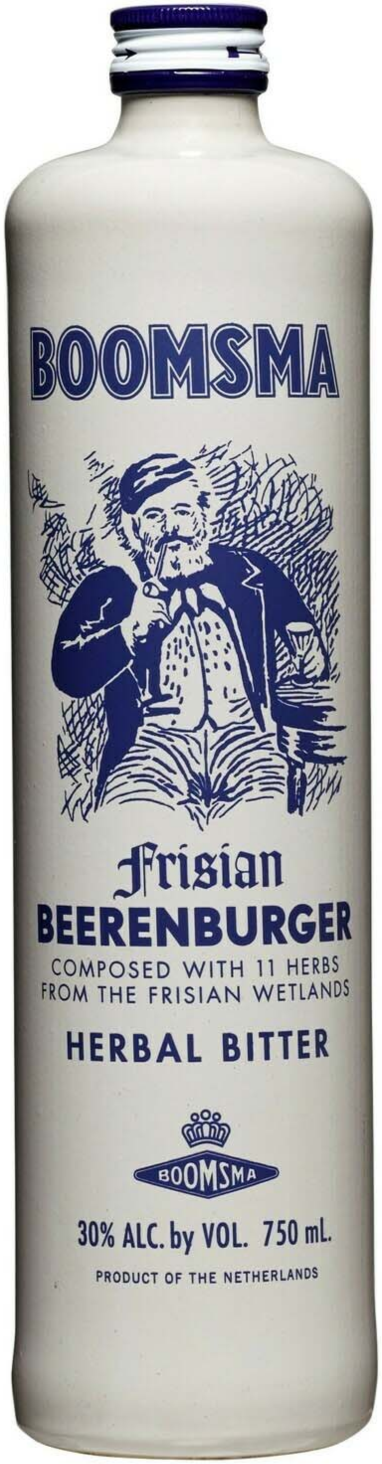
An extension to current maximum-likeliho
components procedures for mapping quan
loci in sib pairs that allows a simultaneous
association is proposed. The method involv
of the allelic means for a test of associati
multaneous modeling of the sib-pair covar
ture for a test of linkage. By partitioning
effect of a locus into between- and within-
ponents, the method controls for spurious
due to population stratification and adm
power and efficacy of the method are illustr
simulation of various models of both real
association.

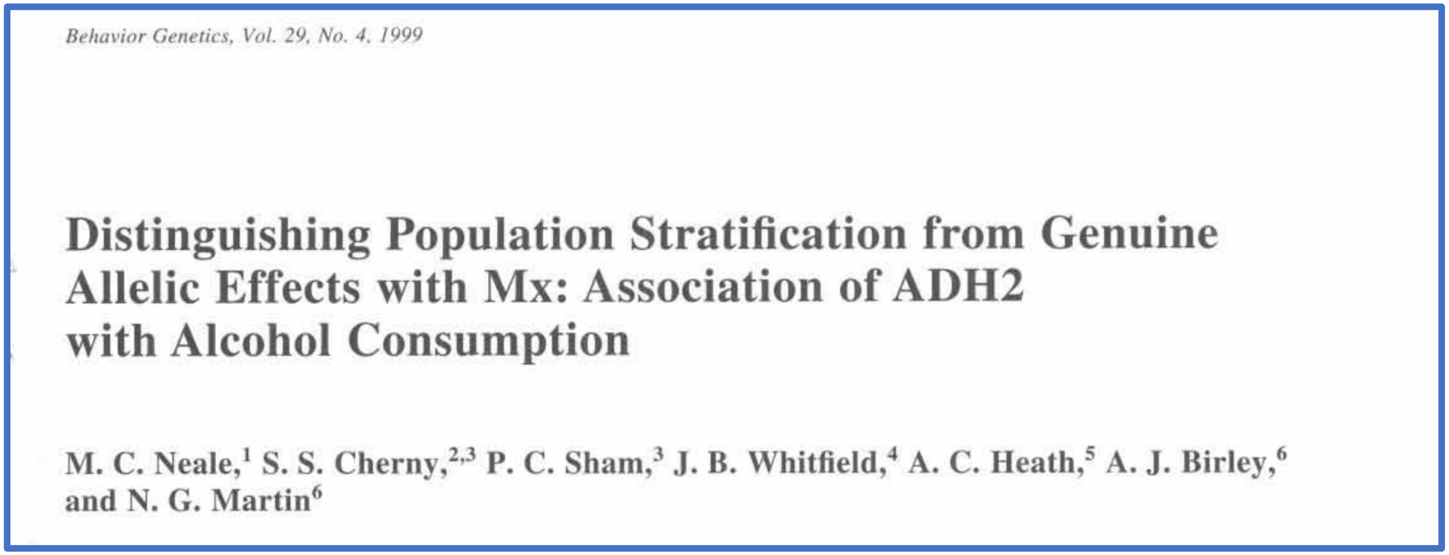
has been due to their perceived importance within the

An extension to current maximum-likelihood variance-components procedures for mapping quantitative-trait loci in sib pairs that allows a simultaneous test of allelic association is proposed. The method involves modeling of the allelic means for a test of association, with simultaneous modeling of the sib-pair covariance structure for a test of linkage. By partitioning of the mean effect of a locus into between- and within-sibship components, the method controls for spurious associations due to population stratification and admixture. The power and efficacy of the method are illustrated through simulation of various models of both real and spurious association.

PC1 (North-South (Abdel, 2013)) / GoNL (2015)



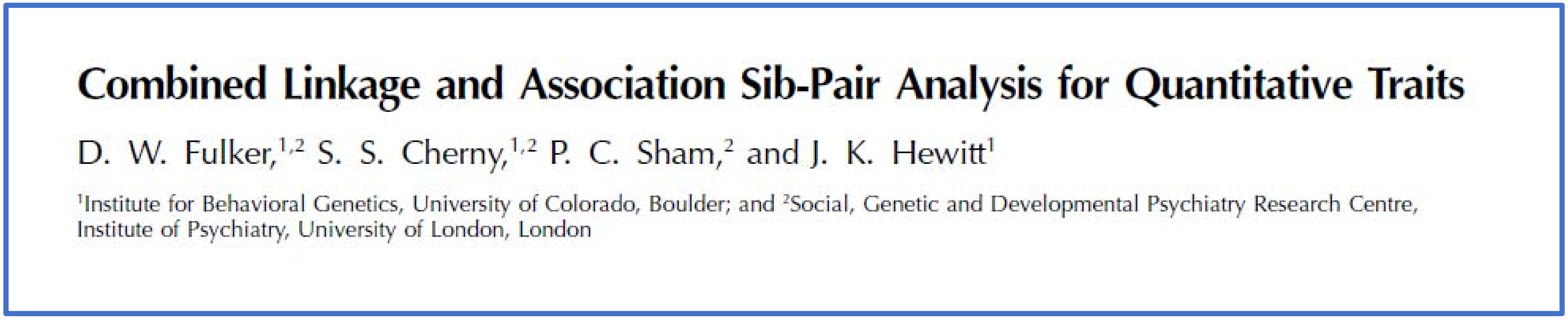




Behavior Genetics, Vol. 29, No. 4, 1999

Distinguishing Population Stratification from Genuine Allelic Effects with Mx: Association of ADH2 with Alcohol Consumption

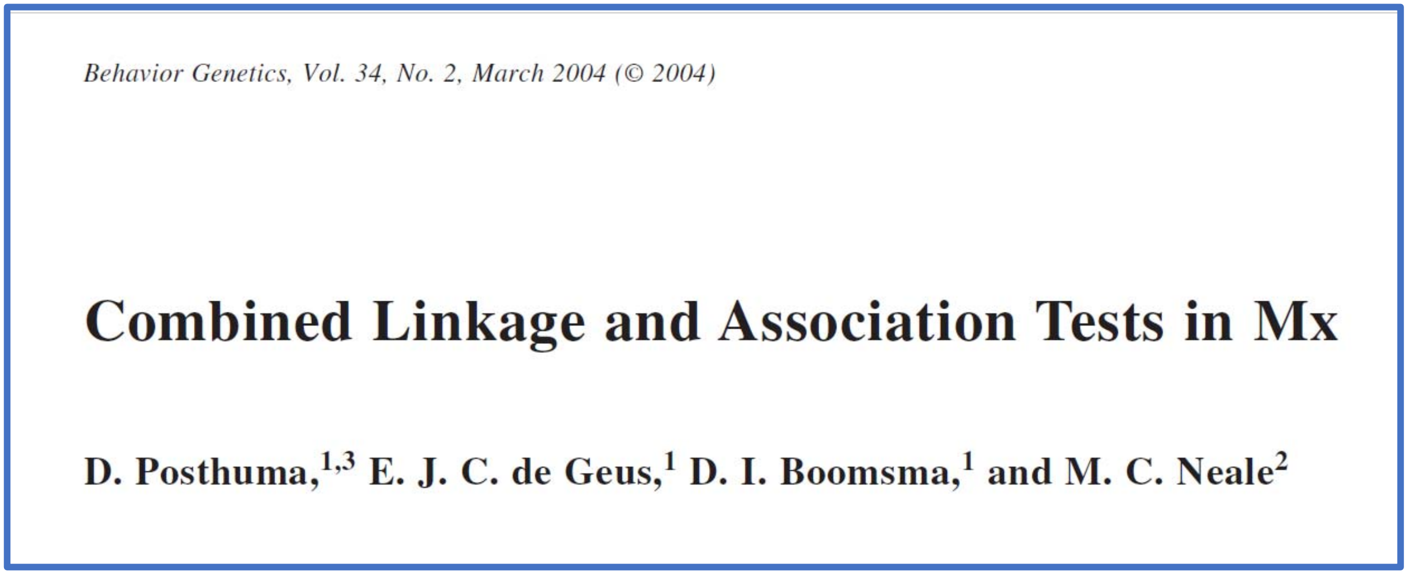
M. C. Neale,[1] S. S. Cherny,[2,3] P. C. Sham,[3] J. B. Whitfield,[4] A. C. Heath,[5] A. J. Birley,[6] and N. G. Martin[6]



Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits

D. W. Fulker,[1,2] S. S. Cherny,[1,2] P. C. Sham,[2] and J. K. Hewitt[1]

[1]Institute for Behavioral Genetics, University of Colorado, Boulder; and [2]Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, University of London, London



Behavior Genetics, Vol. 34, No. 2, March 2004 (© 2004)

Combined Linkage and Association Tests in Mx

D. Posthuma,[1,3] E. J. C. de Geus,[1] D. I. Boomsma,[1] and M. C. Neale[2]

Neale / Fulker / Posthuma: Expected Sib-Pair Means (=between effect) and Differences (=within effect) and Their Frequencies for a Single Additive Two-Allele Locus

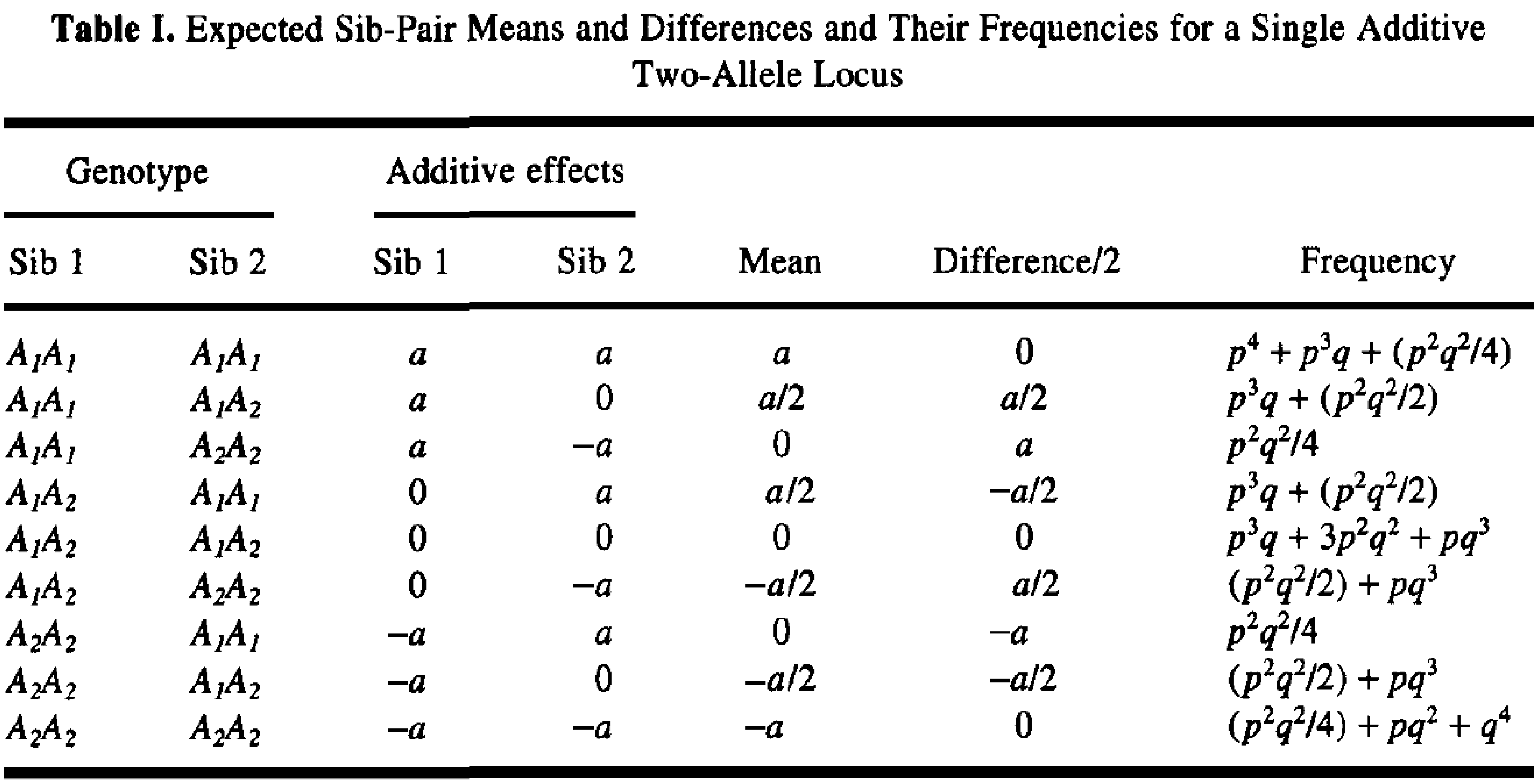
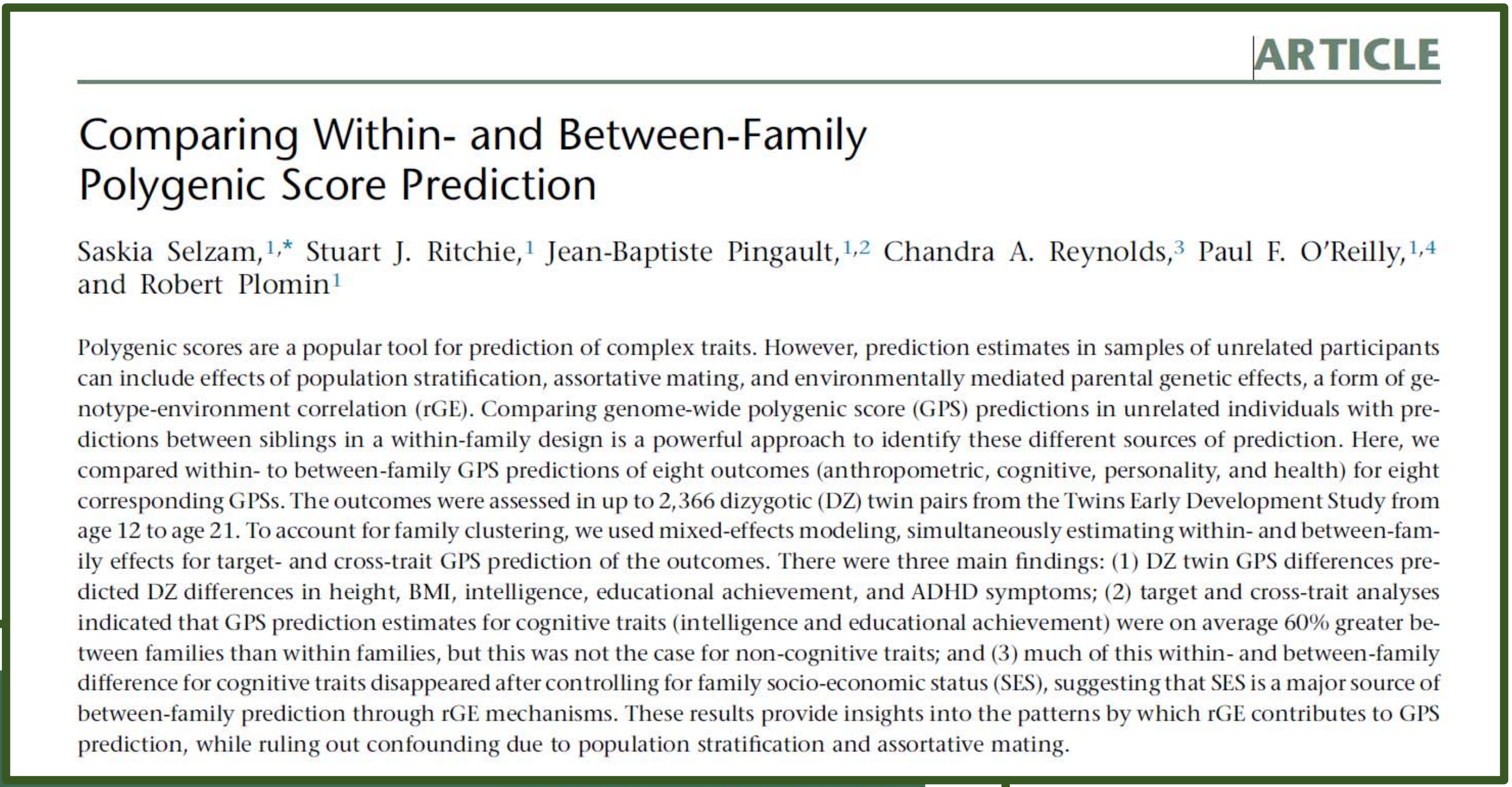


Table I. Expected Sib-Pair Means and Differences and Their Frequencies for a Single Additive Two-Allele Locus

Genotype		Additive effects				
Sib 1	Sib 2	Sib 1	Sib 2	Mean	Difference/2	Frequency
A_1A_1	A_1A_1	a	a	a	0	$p^4 + p^3q + (p^2q^2/4)$
A_1A_1	A_1A_2	a	0	$a/2$	$a/2$	$p^3q + (p^2q^2/2)$
A_1A_1	A_2A_2	a	$-a$	0	a	$p^2q^2/4$
A_1A_2	A_1A_1	0	a	$a/2$	$-a/2$	$p^3q + (p^2q^2/2)$
A_1A_2	A_1A_2	0	0	0	0	$p^3q + 3p^2q^2 + pq^3$
A_1A_2	A_2A_2	0	$-a$	$-a/2$	$a/2$	$(p^2q^2/2) + pq^3$
A_2A_2	A_1A_1	$-a$	a	0	$-a$	$p^2q^2/4$
A_2A_2	A_1A_2	$-a$	0	$-a/2$	$-a/2$	$(p^2q^2/2) + pq^3$
A_2A_2	A_2A_2	$-a$	$-a$	$-a$	0	$(p^2q^2/4) + pq^2 + q^4$

For classic Mx script see: Posthuma paper (appendix)

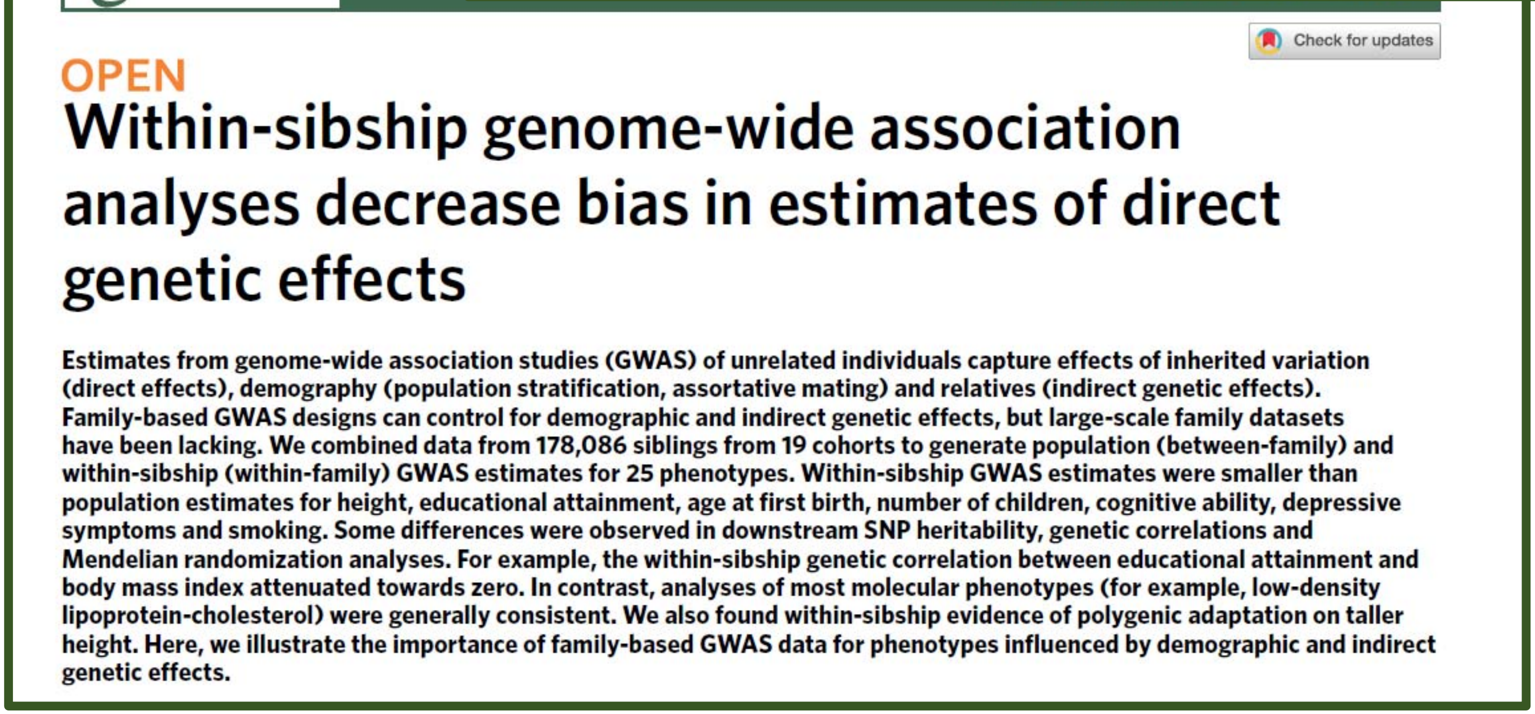


ARTICLE

Comparing Within- and Between-Family Polygenic Score Prediction

Saskia Selzam,[1,*] Stuart J. Ritchie,[1] Jean-Baptiste Pingault,[1,2] Chandra A. Reynolds,[3] Paul F. O'Reilly,[1,4] and Robert Plomin[1]

Polygenic scores are a popular tool for prediction of complex traits. However, prediction estimates in samples of unrelated participants can include effects of population stratification, assortative mating, and environmentally mediated parental genetic effects, a form of genotype-environment correlation (rGE). Comparing genome-wide polygenic score (GPS) predictions in unrelated individuals with predictions between siblings in a within-family design is a powerful approach to identify these different sources of prediction. Here, we compared within- to between-family GPS predictions of eight outcomes (anthropometric, cognitive, personality, and health) for eight corresponding GPSs. The outcomes were assessed in up to 2,366 dizygotic (DZ) twin pairs from the Twins Early Development Study from age 12 to age 21. To account for family clustering, we used mixed-effects modeling, simultaneously estimating within- and between-family effects for target- and cross-trait GPS prediction of the outcomes. There were three main findings: (1) DZ twin GPS differences predicted DZ differences in height, BMI, intelligence, educational achievement, and ADHD symptoms; (2) target and cross-trait analyses indicated that GPS prediction estimates for cognitive traits (intelligence and educational achievement) were on average 60% greater between families than within families, but this was not the case for non-cognitive traits; and (3) much of this within- and between-family difference for cognitive traits disappeared after controlling for family socio-economic status (SES), suggesting that SES is a major source of between-family prediction through rGE mechanisms. These results provide insights into the patterns by which rGE contributes to GPS prediction, while ruling out confounding due to population stratification and assortative mating.



nature genetics

Check for updates

OPEN

Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects

Estimates from genome-wide association studies (GWAS) of unrelated individuals capture effects of inherited variation (direct effects), demography (population stratification, assortative mating) and relatives (indirect genetic effects). Family-based GWAS designs can control for demographic and indirect genetic effects, but large-scale family datasets have been lacking. We combined data from 178,086 siblings from 19 cohorts to generate population (between-family) and within-sibship (within-family) GWAS estimates for 25 phenotypes. Within-sibship GWAS estimates were smaller than population estimates for height, educational attainment, age at first birth, number of children, cognitive ability, depressive symptoms and smoking. Some differences were observed in downstream SNP heritability, genetic correlations and Mendelian randomization analyses. For example, the within-sibship genetic correlation between educational attainment and body mass index attenuated towards zero. In contrast, analyses of most molecular phenotypes (for example, low-density lipoprotein-cholesterol) were generally consistent. We also found within-sibship evidence of polygenic adaptation on taller height. Here, we illustrate the importance of family-based GWAS data for phenotypes influenced by demographic and indirect genetic effects.

$$\mathbf{Y_{ij} = a_0 + b_W (GPS_{ij} - \textit{mean}GPS_j) + b_B\, \textit{mean}GPS_j + g_j + \varepsilon_{ij}}$$

Y = the outcome, GPS = polygenic score, *mean*GPS = mean GPS in family j, i = {1,2} = individual twin within family j

a_0 = intercept
g_j = random effect: change in intercept for twins in family j
ε_{ij} = the independent random error for each individual i in family j

b_B = between-family effect: expected change in Y given one unit change in *mean*GPS

b_W = within-family effect: expected change given one unit change in the difference between the individual GPS and the *mean*GPS.

$$\mathbf{Y_{ij} = a_0 + b_W (GPS_{ij} - \textit{mean}GPS_j) + b_B\ \textit{mean}GPS_j + g_j + \varepsilon_{ij}}$$

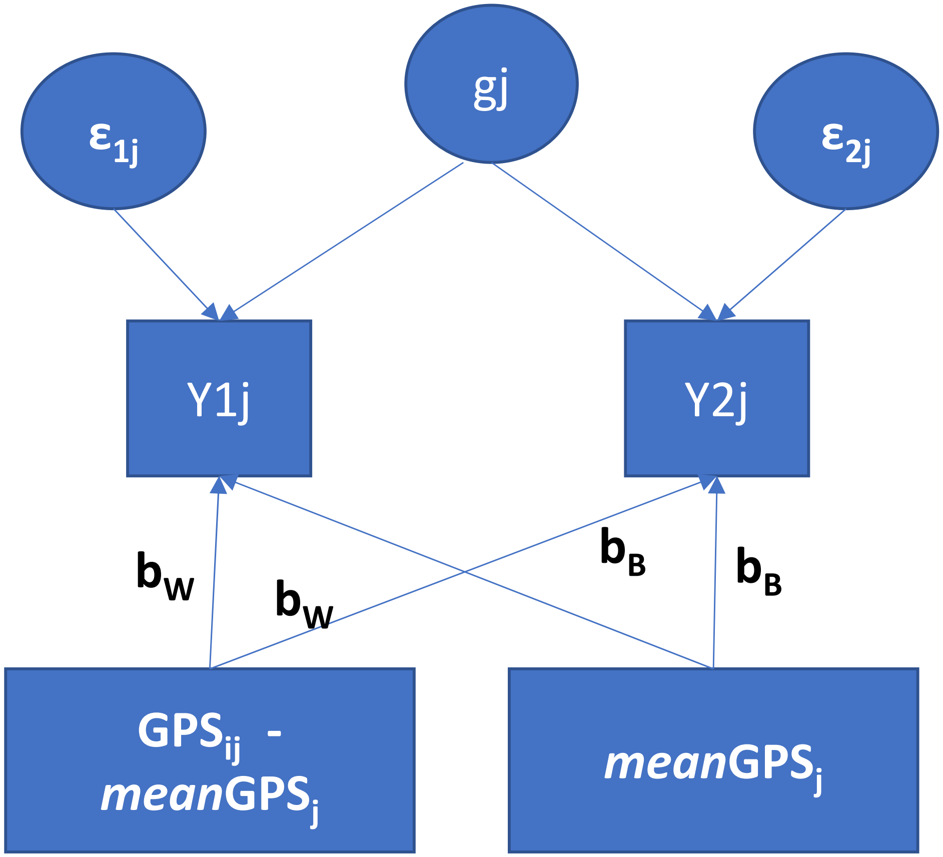
A random effect term $\sigma^2(g)$ estimates the difference between each group intercept g_j and the overall intercept a_0, accounting for residual structure in the data (genetic and environmental) that lead to trait similarity of twins.

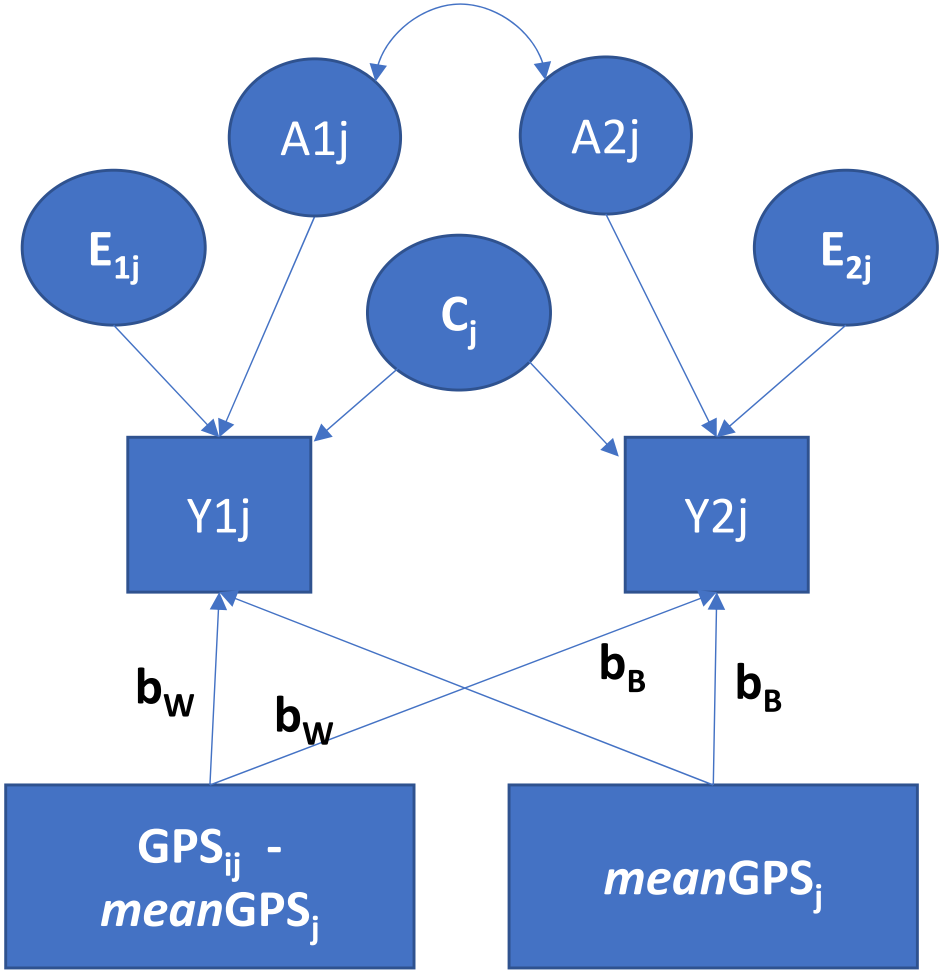
The use of a mixed-effects model is justified if co-twins correlate in the outcome. This can be estimated by ICC (intraclass correlation):

$$\mathbf{ICC = cor\ (Y_{1j}, Y_{2j}) = \sigma^2 g\ /\ (\sigma^2 g + \sigma^2\ \varepsilon)}$$

$$\mathbf{Y_{ij} = a_0 + b_W (GPS_{ij} - \mathit{mean}GPS_j) + b_B\, \mathit{mean}GPS_j + g_j + \varepsilon_{ij}}$$

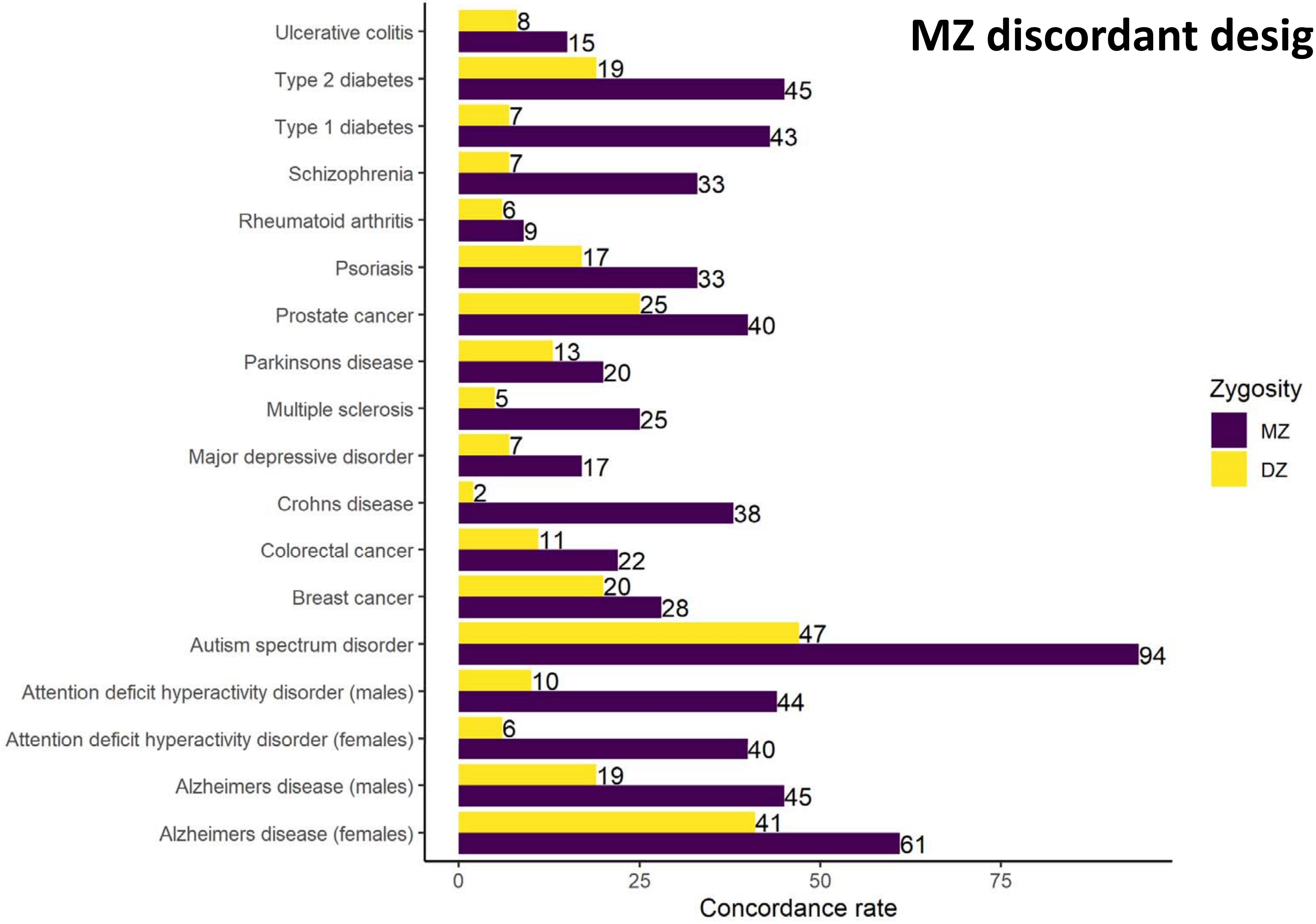
ε_{ij} = the independent random error for each individual i in family j





These are equivalent models.
The model on the left includes epsilon, which includes genetic and environmental effects.
The model right is the traditional twin model where A, C and E feature explicitly.
The terms (left) "random error" is perhaps confusing as the epsilon includes 1) measurement error, 2) unshared environmental effects and genetic effect (mendelian segregation). The term gj (left) can include shared environmental effects in addition to the average breeding value (of parents).

MZ discordant design



Hagenbeek et al.: Twin Studies in Rapidly Changing Times. https://osf.io/rne4s/

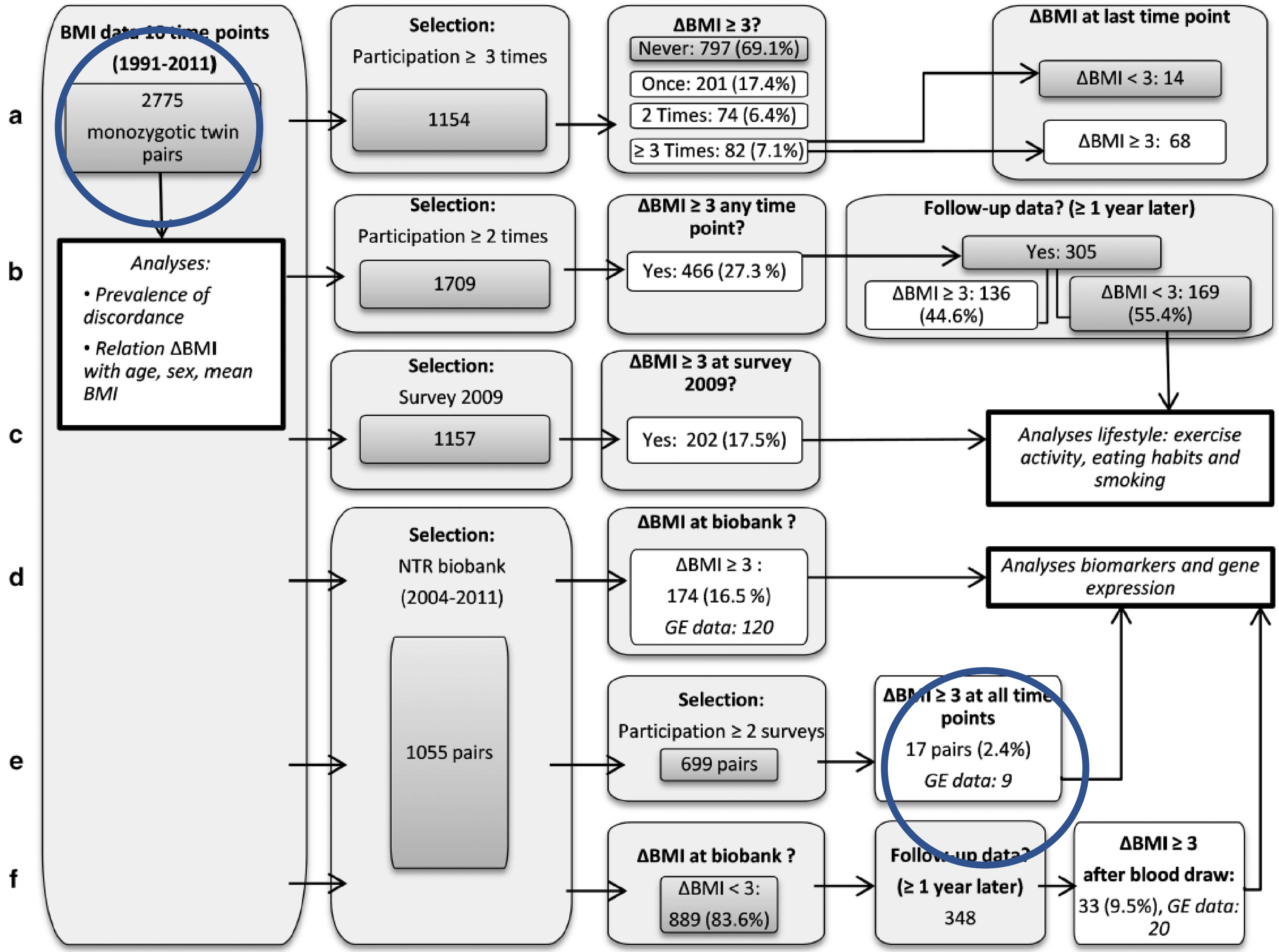
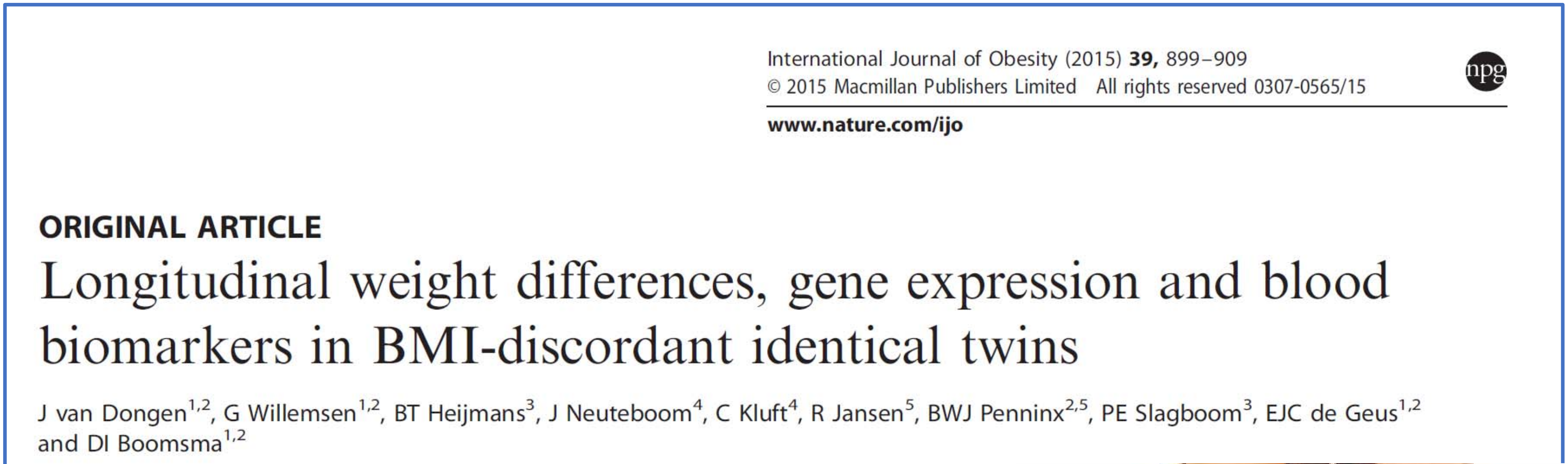


Figure 1. Flowchart of the selection procedure of MZ twin pairs included in each analysis. All numbers in this figure represent the numbers of MZ twin pairs. GE = gene expression. Each row (**a**–**f**) illustrates the available data and selection criteria for MZ pairs included in a particular analysis. (**a**) Frequency of BMI discordance at one, two or more longitudinal time points in MZ pairs with longitudinal BMI data. (**b**) Number of MZ pairs who are discordant across all projects and number of pairs who are still discordant at the first next available follow-up time point. (**c**) Discordant pairs included in the analyses of lifestyle data. (**d**) MZ pairs who were discordant at blood draw and were included in the analyses of biomarkers and gene expression. (**e**) MZ pairs who were discordant at all time points of participation and were included in the analyses of biomarkers and gene expression. (**f**) MZ pairs who became discordant after blood draw and who were studied to examine biomarkers and gene expression difference before BMI discordance onset.

MZ discordant designs

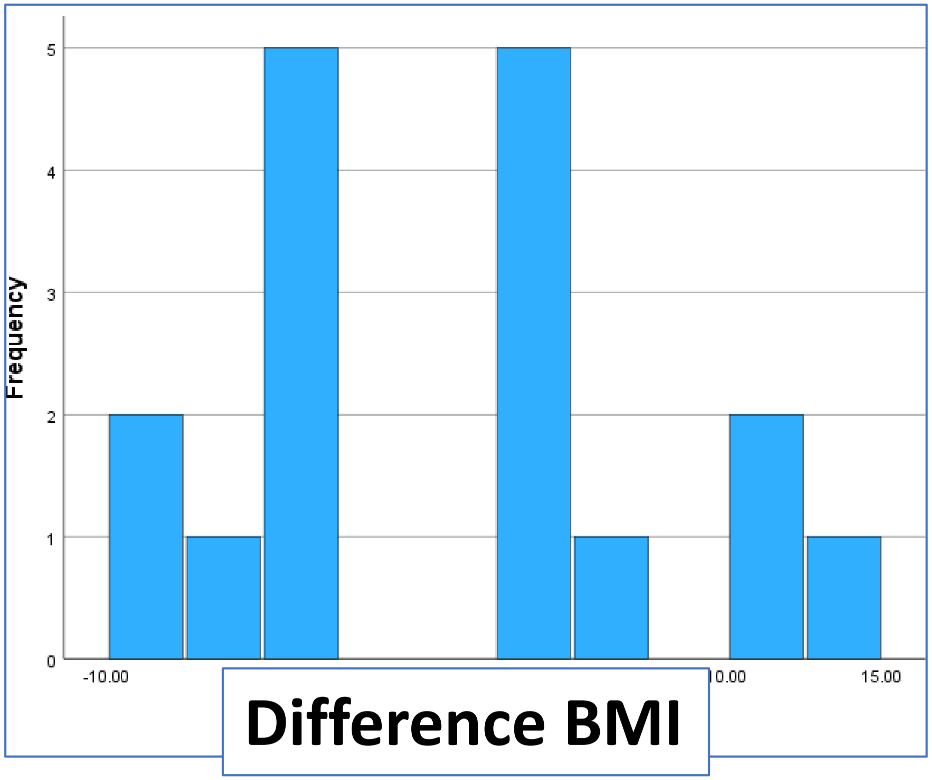


International Journal of Obesity (2015) **39,** 899–909
© 2015 Macmillan Publishers Limited All rights reserved 0307-0565/15
www.nature.com/ijo

ORIGINAL ARTICLE

Longitudinal weight differences, gene expression and blood biomarkers in BMI-discordant identical twins

J van Dongen[1,2], G Willemsen[1,2], BT Heijmans[3], J Neuteboom[4], C Kluft[4], R Jansen[5], BWJ Penninx[2,5], PE Slagboom[3], EJC de Geus[1,2] and DI Boomsma[1,2]



17 longitudinally discordant pairs

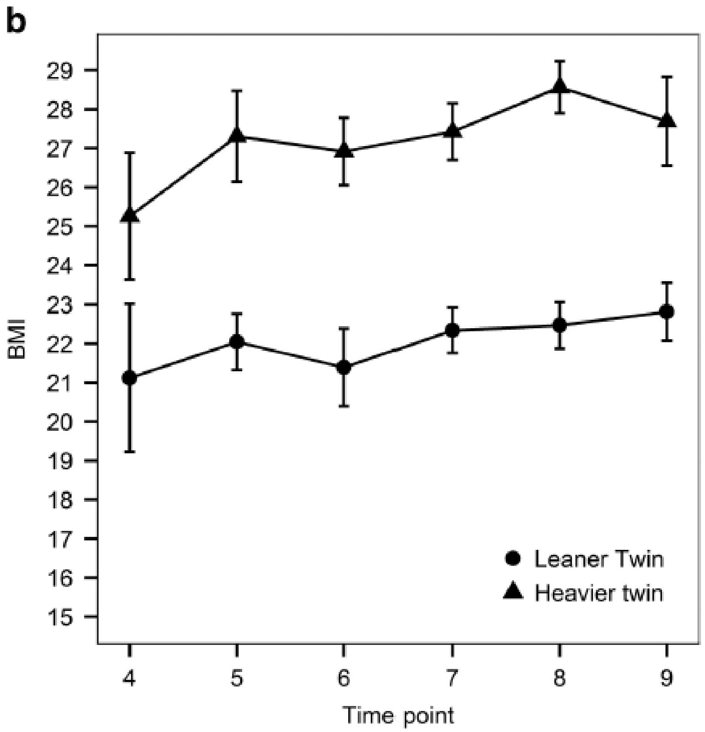
Studying biomarkers in BMI-discordant twins **rules out genetic pleiotropy as an explanation for the association by design**:
If the association would solely exist because genetic variants that predispose to a high BMI also cause changes in biomarkers, MZ twins who are discordant for BMI should have similar biomarker levels because MZ twins have the same genetic vulnerability.

In NTR we had longitudinal BMI data on 2775 MZ pairs, in NTR_biobank there were 1055 pairs with biomarker data.

There were 17 pairs who were longitudinally discordant (> 3 BMI points, differences in body weight 80 vs 63 kg.)

BMI discordant MZ twin pairs had differences in all metabolic markers with the heavier twin having an unfavorable metabolic profile.
The heavier twins also had higher blood levels of IL-6, soluble IL-6 receptor, C-reactive protein and GGT(gamma glutamyl transferase).

	Heavier twin	*Leaner twin*	*Mean difference (heavier – leaner twin)*	P-value
N	17	17		
N, male/female pairs	1/16	1/16		
Age (years)	45.6 (11.6)			
Birth weight (g)	2366 (742)	2538 (752)	– 172	0.27
BMI (kg m^{-2})	28.6 (2.7)	22.5 (2.4)	6.1	5.1×10^{-7}
Weight (kg)	80.1 (9.6)	63.1 (9.2)	17.0	1.9×10^{-7}
Height (cm)	167.4 (5.6)	167.3 (5.3)	0.1	0.84
Waist (cm)	92.2 (10.3)	76.6 (8.5)	15.6	3.9×10^{-8}
Hip (cm)	109.7 (5.8)	98.6 (6.9)	11.1	8.1×10^{-10}
WHR (cm cm^{-1})	0.84 (0.08)	0.78 (0.07)	0.06	6.6×10^{-5}
Glucose (mmol l^{-1})	5.4 (0.6)	5.1 (0.4)	0.3	0.06
Insulin (μIU ml^{-1})	10.1 (6.0)	5.3 (2.6)	4.8	0.03
Total Chol (mmol l^{-1})	5.9 (1.5)	5.2 (1.4)	0.7	0.03
LDL (mmol l^{-1})	3.6 (1.4)	3.2 (1.5)	0.4	0.15
HDL (mmol l^{-1})	1.5 (0.3)	1.6 (0.3)	-0.1	0.20
Triglycerides (mmol l^{-1})	1.5 (0.9)	0.9 (0.4)	0.6	9.2×10^{-4}
CRP (mg l^{-1})	4.4 (4.6)	1.2 (1.4)	3.2	4.9×10^{-4}
TNF-α (pg ml^{-1})	1.95 (3.7)	1.05 (1.3)	0.9	0.86
IL-6 (pg ml^{-1})	2.3 (1.9)	1.4 (0.8)	0.9	0.12
sIL-6R (pg ml^{-1})	42 645 (13 375)	38 672 (13 340)	3973	0.14
Fibrinogen (g l^{-1})	3.3 (0.7)	2.6 (0.6)	0.7	1.3×10^{-5}
AST (U l^{-1})	22.9 (6.9)	22.4 (10.0)	0.5	0.32
ALT (U l^{-1})	12.6 (6.6)	12.0 (5.4)	0.6	0.90
GGT (U l^{-1})	29.4 (18.0)	20.6 (10.6)	8.8	1.9×10^{-3}
N (%) using lipid-lowering medication	1 (5.9%)	1 (5.9%)	0	0.99
N (%) using diabetes medication	0 (0%)	0 (0%)	0	0.99
N (percentage of female twins) menopa	6 (37.5)	6 (37.5%)		0.99



Dataset: 17 twin MZ pairs, longitudinally discordant for BMI

File 1: twin = case (twin 1 / 2 is not leaner/heavier)
file 2: pair = case (one twin (recoded 3 and 4) is the heavier twin.

Variables: Bmi, weight, hip, waist, glucose
Reproduce the results in table 3.