Meetings this summer

- June 3-6: Behavior Genetics Association (Amsterdam, The Netherlands, see: www.bga.org)

- June 8-10: Int. Society Twin Studies (Ghent, Belgium, see: www.twins2007.be)

Introduction to multivariate QTL

- Theory
- Genetic analysis of lipid data (3 traits)
- QTL analysis of uni- / multivariate data
- Display multivariate linkage results

Dorret Boomsma, Meike Bartels, Jouke Jan Hottenga, Sarah Medland

Directories:	dorret\lipid2007 univariate jobs
	dorret\lipid2007 multivariate jobs
	sarah\graphing

Multivariate approaches

- Principal component analysis (Cholesky)
- Exploratory factor analysis (Spss, SAS)
- Path analysis (S Wright)
- Confirmatory factor analysis (Lisrel, Mx)
- Structural equation models (Joreskog, Neale)

These techniques are used to analyze multivariate data that have been collected in *non-experimental* designs and often involve *latent constructs* that are not directly observed.

These latent constructs underlie the observed variables and account for correlations between variables.

Example: depression

- I feel lonely
- I feel confused or in a fog
- I cry a lot
- I worry about my future.
- I am afraid I might think or do something bad
- I feel that I have to be perfect
- I feel that no one loves me
- I feel worthless or inferior
- I am nervous or tense
- I lack self confidence I am too fearful or anxious
- I feel too guilty
- I am self-conscious or easily embarrassed
- I am unhappy, sad or depressed
- I worry a lot
- I am too concerned about how I look
- I worry about my relations with the opposite sex

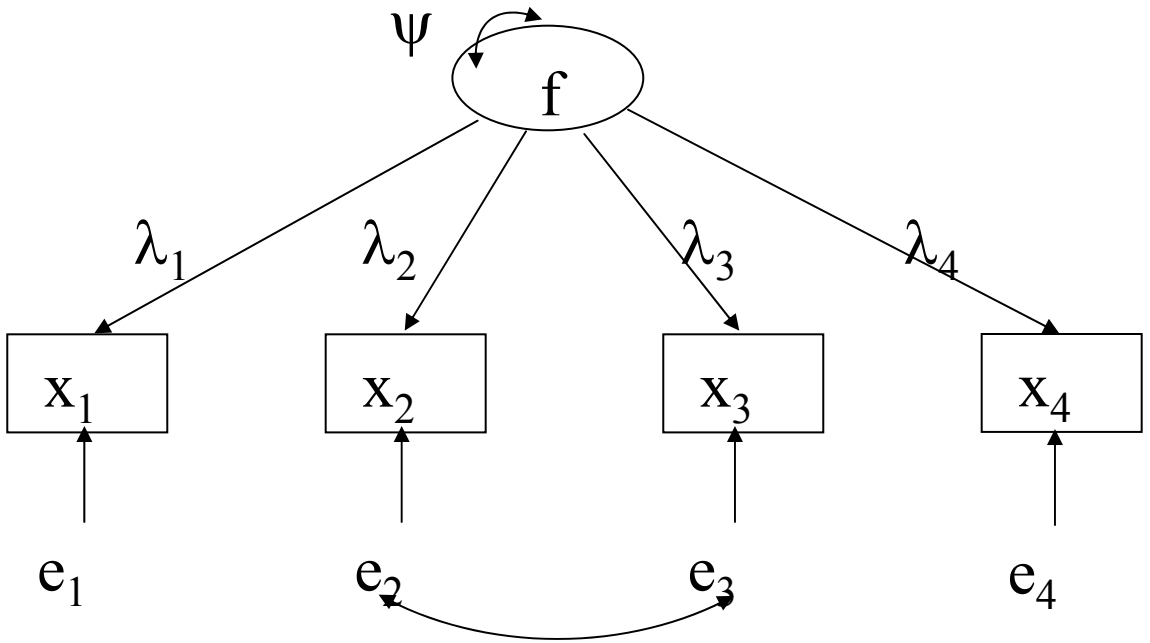
Are these items "indicators" of a trait that we call depression?

Is there a latent construct that underlies the observed items and that accounts for the inter-correlations between variables?

The covariance between item x1 and x4 is:

$\text{cov}(x_1, x_4) = \lambda_1 \lambda_4 \psi = \text{cov}(\lambda_1 f + e_1, \lambda_4 f + e_4)$

where ψ is the variance of f and e_1 and e_4 are uncorrelated



Sometimes $x = \Lambda f + e$ is referred to as the measurement model.

The part of the model that specifies relations among latent factors is the covariance structure model, or the structural equation model

Symbols used in path analysis

square box:observed variable (x)

circle: latent (unobserved) variable (f, G, E)

unenclosed variable: innovation / disturbance term (error) in equation (ζ) or measurement error (e)

straight arrow: causal relation (λ)

curved two-headed arrow: association (r)

two straight arrows: feedback loop

Tracing rules of path analysis

The associations between variables in a path diagram is derived by tracing all connecting paths between variables:

1 trace backward along an arrow, then forward
- never forward and then back;
- never through adjacent arrow heads

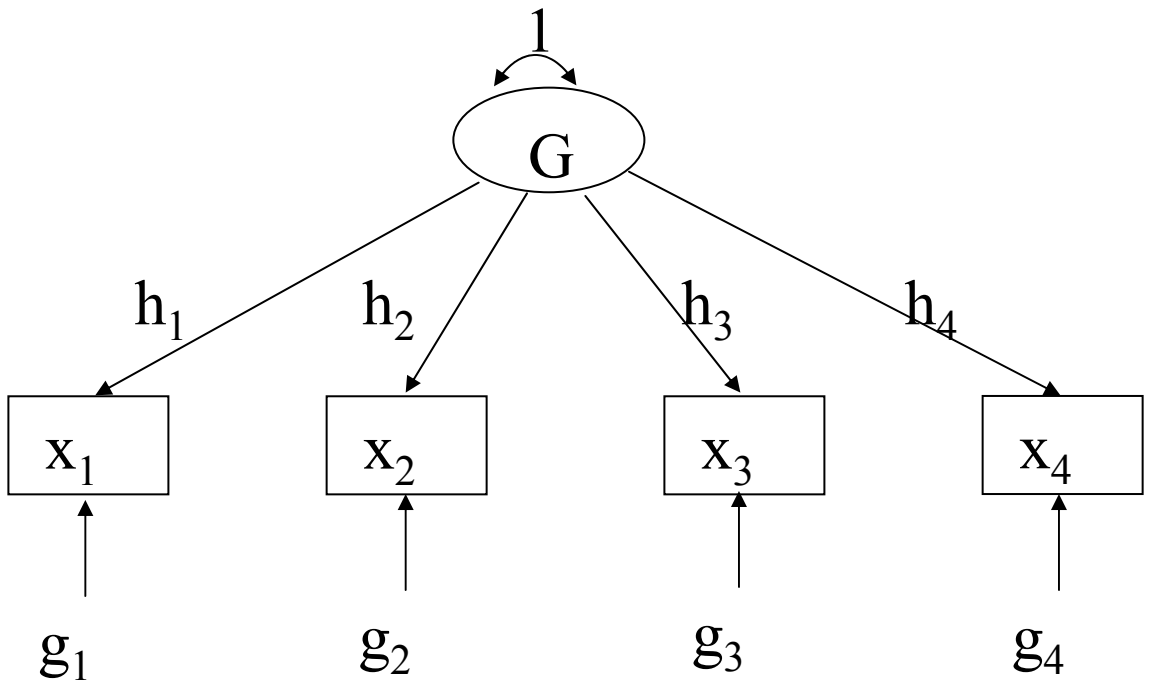
2 pass through each variable only once

3 trace through at most one two-way arrow

The expected correlation/covariance between two variables is the product of all coefficients in a chain and summing over all possible chains (assuming no feedback loops)

$\text{cov}(x_1, x_4) = h_1 h_4$

$\text{Var}(x1) = h^2_1 + \text{var}(g_1)$



Genetic Structural Equation Models

Measurement model / Confirmatory factor model: $x = \Lambda f + e$,

- x = observed variables
- f = (unobserved) factor scores
- e = unique factor / error
- Λ = matrix of factor loadings

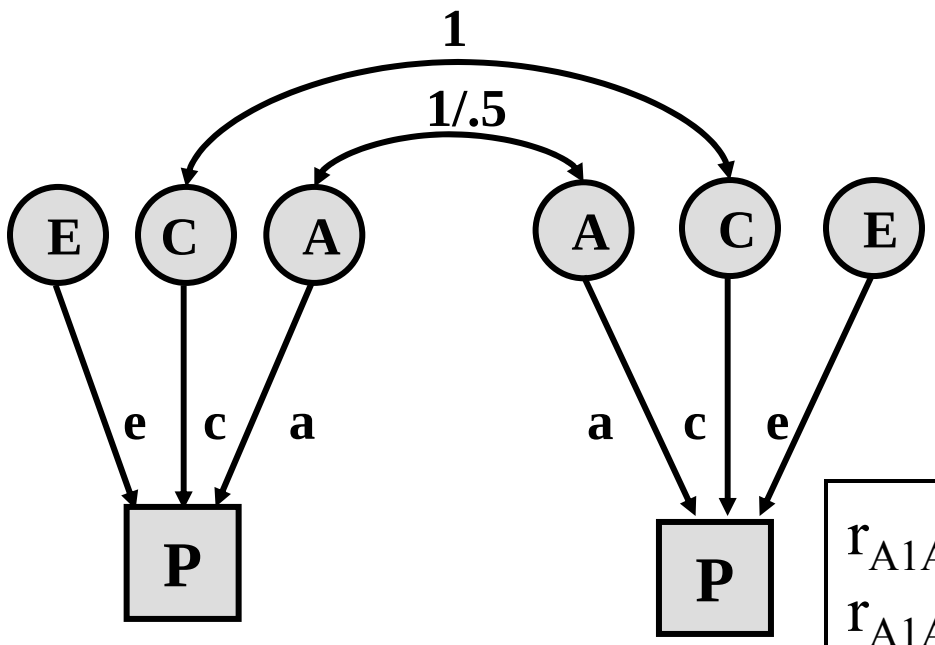
"Univariate" genetic factor model

$$P_j = hG_j + e\,E_j + c\,C_j , \qquad j = 1, \ldots, n \text{ (subjects)}$$

where P = measured phenotype

- G = unmeasured genotypic value
- C = unmeasured environment common to family members
- E = unmeasured unique environment
- Λ = h, c, e (factor loadings/path coefficients)

Univariate ACE Model for a Twin Pair



$$\Sigma_{MZ} = \begin{bmatrix} a^2 + c^2 + e^2 & a^2 + c^2 \\ a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}$$

$$\Sigma_{DZ} = \begin{bmatrix} a^2 + c^2 + e^2 & .5\,a^2 + c^2 \\ .5\,a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}$$

$r_{A1A2} = 1$ for MZ
$r_{A1A2} = 0.5$ for DZ

Covariance (P1, P2)
$= a\, r_{A1A2}\, a + c^2$

$r_{MZ} = a^2 + c^2$

$r_{DZ} = 0.5\, a^2 + c^2$

$2(r_{MZ} - r_{DZ}) = a^2$

Genetic Structural Equation Models

$P_j = hG_j + e\,E_j + c\,C_j\,, \qquad j = 1, \ldots, n$ (subjects)

Can be very easily generalized to multivariate data, where for example P is 2 x 1 (or p x 1) and the dimensions of the other matrices change accordingly.

With covariance matrix: $\Sigma = \Lambda\Psi\Lambda' + \Theta$

Where Σ is pxp and the dimensions of other matrices depend on the model that is evaluated (Λ is the matrix of factor loading; Ψ has the correlations among factor scores and Θ has the error variances (usually a diagonal matrix)

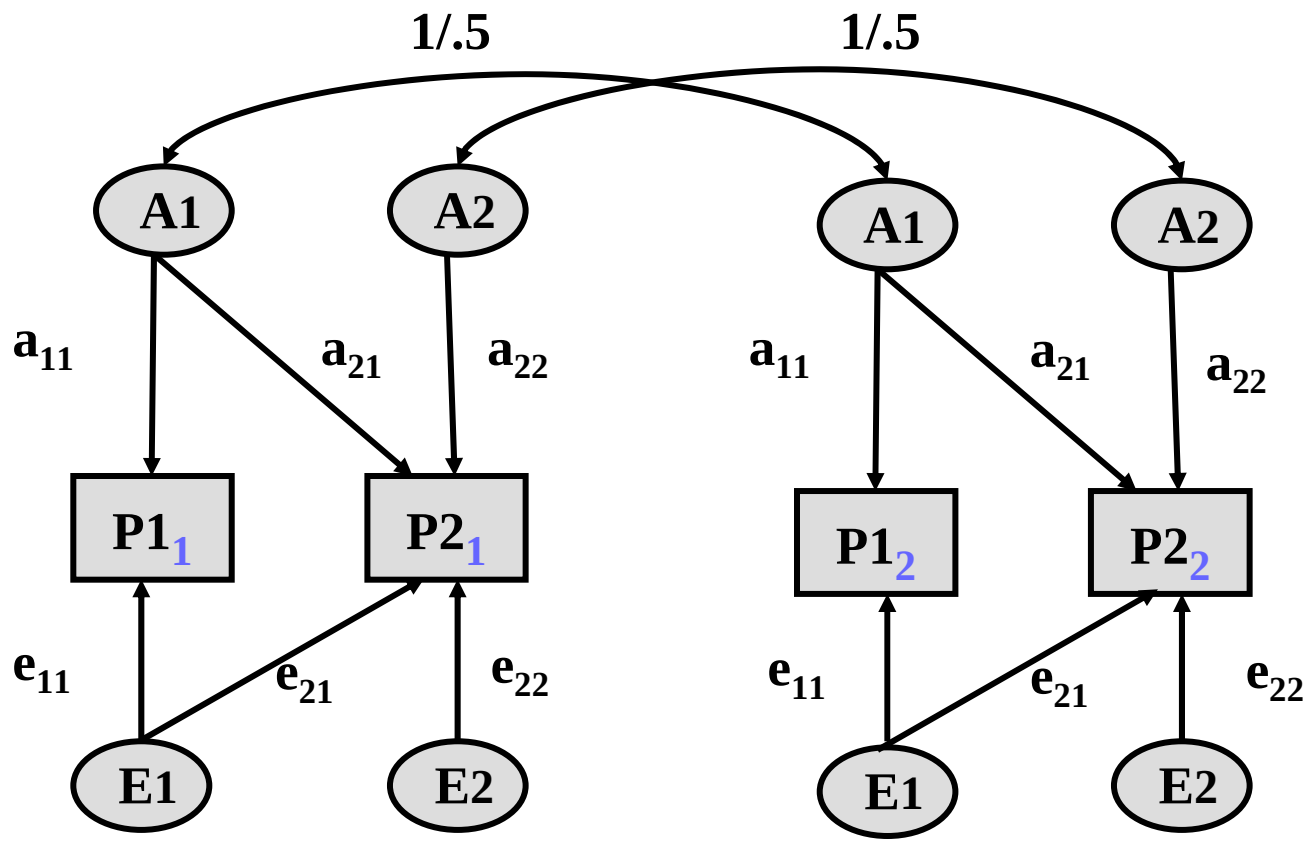
Models in non-experimental research

All models specify a covariance matrix Σ and means vector μ:

$$\Sigma = \Lambda\Psi\Lambda^t + \Theta$$

total covariance matrix [Σ] =
factor variance [$\Lambda\Psi\Lambda^t$] + residual variance [Θ]

means vector μ can be modeled as a function of other (measured) traits e.g. sex, age, cohort, SES



Bivariate twin model:

The first (latent) additive genetic factor influences P1 and P2;
The second additive genetic factor influences P2 only.
A1 in twin 1 and A1 twin 2 are correlated; A2 in twin 1 and A2 in twin 2 are correlated (A1 and A2 are uncorrelated)

- S (pxp) would be 2x2 for 1 person: 4x4 for twin or sib pairs; what we usually do in Mx:

A and E are 2x2 and have the following form:

- $\begin{pmatrix} \mathbf{a11} & \\ \mathbf{a21} & \mathbf{a22} \end{pmatrix}$ $\begin{pmatrix} \mathbf{e11} & \\ \mathbf{e21} & \mathbf{e22} \end{pmatrix}$

And then S is: $A*A' + E*E' \mid r_a A*A'$

$r_a A*A' \mid A*A' + E*E'$

(where r_a is the genetic correlation in MZ/DZ twins and A and E are lower triangular matrices)

Implied covariance structure: A (DZ twins)

(text in red indicates the "within person", text in blue indicates the "between person"- statistics)

A --- DZ twins	X-twin1	Y-twin1	X-twin2	Y-twin2
X-twin1	*Variance X* $a11^2$	equal to row 2, column 1	equal to row 3, column 1	equal to row 4, column 1
Y-twin1	Covariance within person, between variables a11 * a21	*Variance Y* $a21^2 + a22^2$	equal to row 3, column 2	equal to row 4, column 2
X-twin2	Covariance between persons, within variables a11 * .5 * a11	Covariance between persons, between variables a11 * .5 * a21	*Variance X* $a11^2$	equal to row 4, column 3
Y-twin2	Covariance between persons, between variables a11 * .5 * a21	Covariance between persons, within variables a21 * .5 * a21 + a22 * .5 * a22	Covariance within person, between variables a11 * a21	*Variance Y* $a21^2 + a22^2$

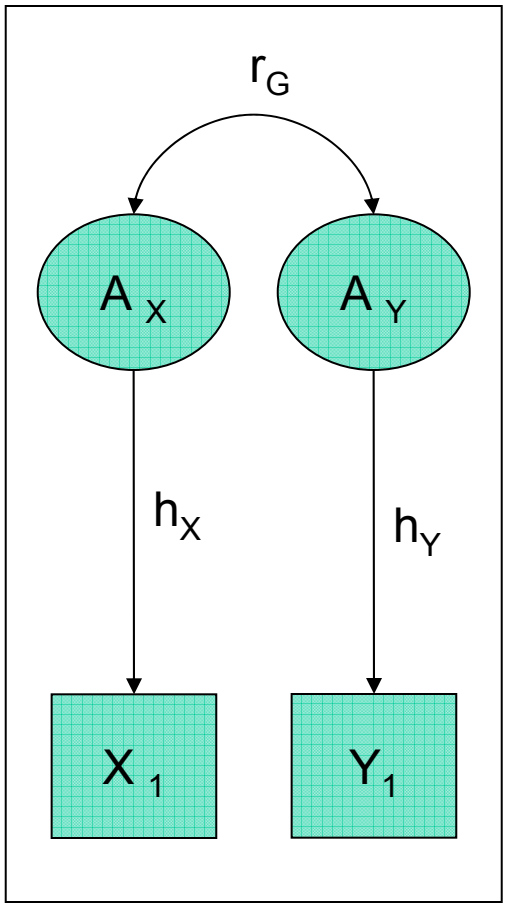
Implied covariance structure: C (MZ and DZ twins)

C --- MZ & DZ twins	X-twin1	Y-twin1	X-twin2	Y-twin2
X-twin1	*Variance X* $c11^2$	equal to row 2, column 1	equal to row 3, column 1	equal to row 4, column 1
Y-twin1	Covariance within person, between variables c11 * c21	*Variance Y* $c21^2 + c22^2$	equal to row 3, column 2	equal to row 4, column 2
X-twin2	Covariance between persons, within variables c11 * c11	Covariance between persons, between variables c11 * c21	*Variance X* $c11^2$	equal to row 4, column 3
Y-twin2	Covariance between persons, between variables c11 * c21	Covariance between persons, within variables c21 * c21 + c22 * c22	Covariance within person, between variables c11 * c21	*Variance Y* $c21^2 + c22^2$

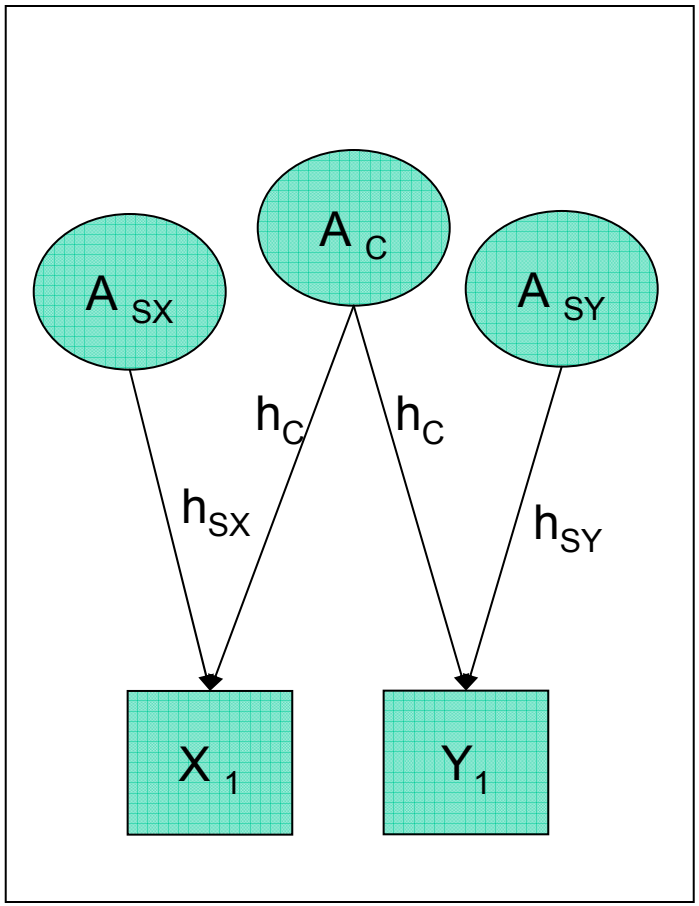
Implied covariance structure: E (MZ and DZ twins)

E --- MZ & DZ twins	**X-twin1**	**Y-twin1**	**X-twin2**	**Y-twin2**
X-twin1	*Variance X* $e11^2$	equal to row 2, column 1	equal to row 3, column 1	equal to row 4, column 1
Y-twin1	Covariance within person, between variables e11 * e21	*Variance Y* $e21^2 + e22^2$	equal to row 3, column 2	equal to row 4, column 2
X-twin2	Covariance between persons, within variables 0	Covariance between persons, between variables 0	*Variance X* $e11^2$	equal to row 4, column 3
Y-twin2	Covariance between persons, between variables 0	Covariance between persons, within variables 0	Covariance within person, between variables e11 * e21	*Variance Y* $e21^2 + e22^2$

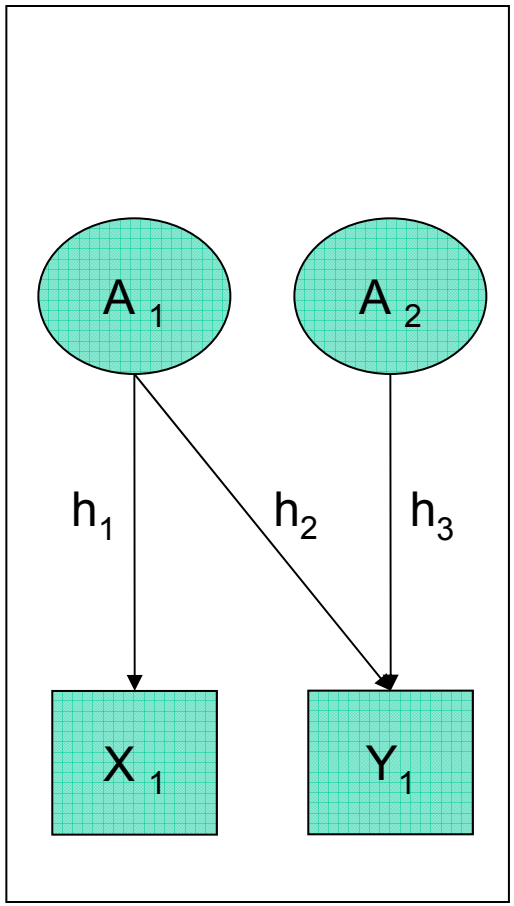
Bivariate Phenotypes



Correlation

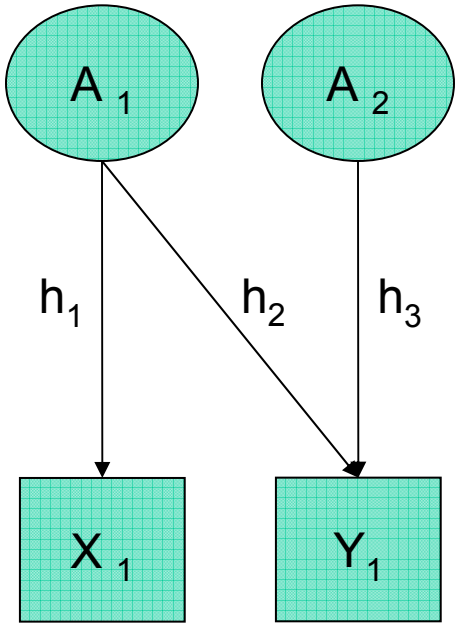


Common factor



Cholesky decomposition

Cholesky decomposition

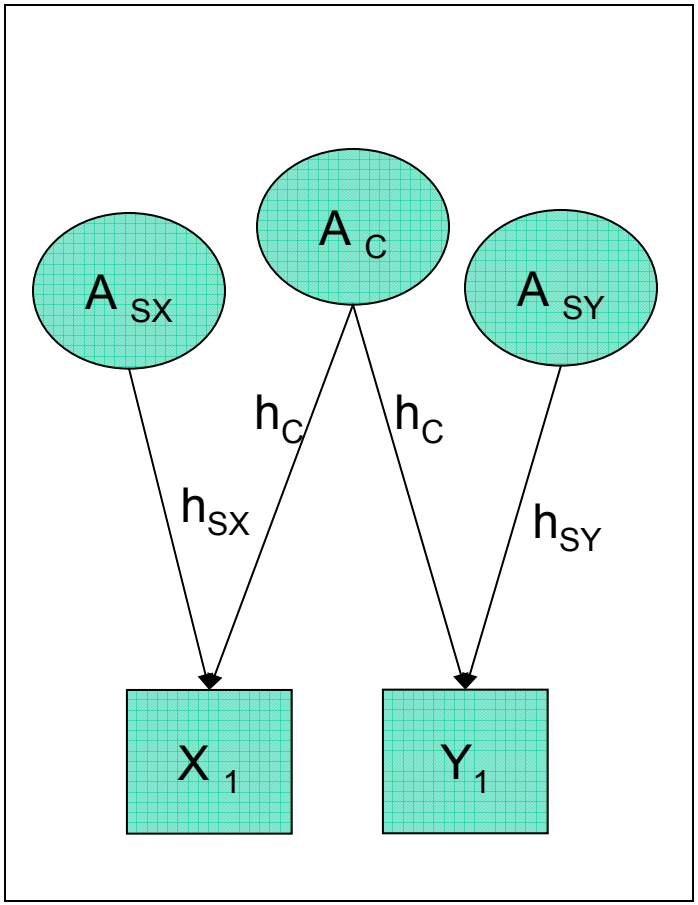


- If $h_3 = 0$: no genetic influences specific to Y
- If $h_2 = 0$: no genetic covariance

- The genetic correlation between X and Y =

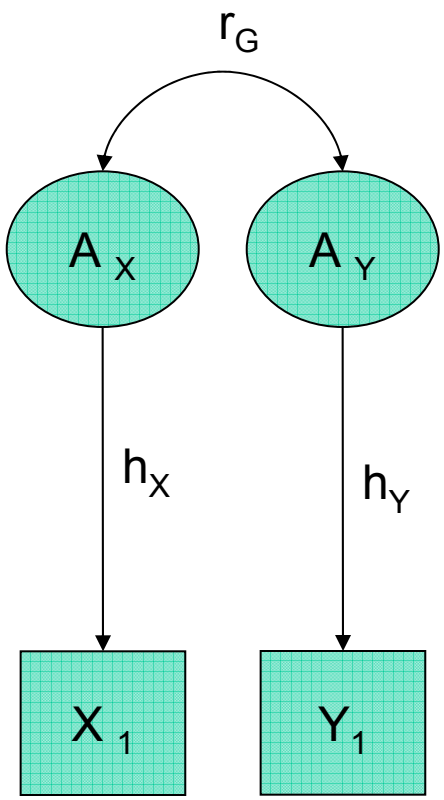
covariance X,Y / SD(X)*SD(Y)

Common factor model



A common factor influences both traits (a constraint on the factor loadings is needed to make this model identified).

Correlated factors



- Genetic correlation r_G
- Component of phenotypic covariance

  $r_{XY} = h_X r_G h_Y [+ c_X r_C c_Y + e_X r_E e_Y]$

Phenotypic correlations can arise, broadly speaking, from two distinct causes (we do not consider other explanations such as phenotypic causation or reciprocal interaction).

The same environmental factors may operate within individuals, leading to within-individual environmental correlations. Secondly, genetic correlations between traits may lead to correlated phenotypes.

The basis for genetic correlations between traits may lie in pleiotropic effects of genes, or in linkage or non-random mating. However, these last two effects are expected to be less permanent and consequently less important (Hazel, 1943).

THE GENETIC BASIS FOR CONSTRUCTING SELECTION INDEXES[1]

L. N. HAZEL[2]

Genetics, 28, 476-490, 1943

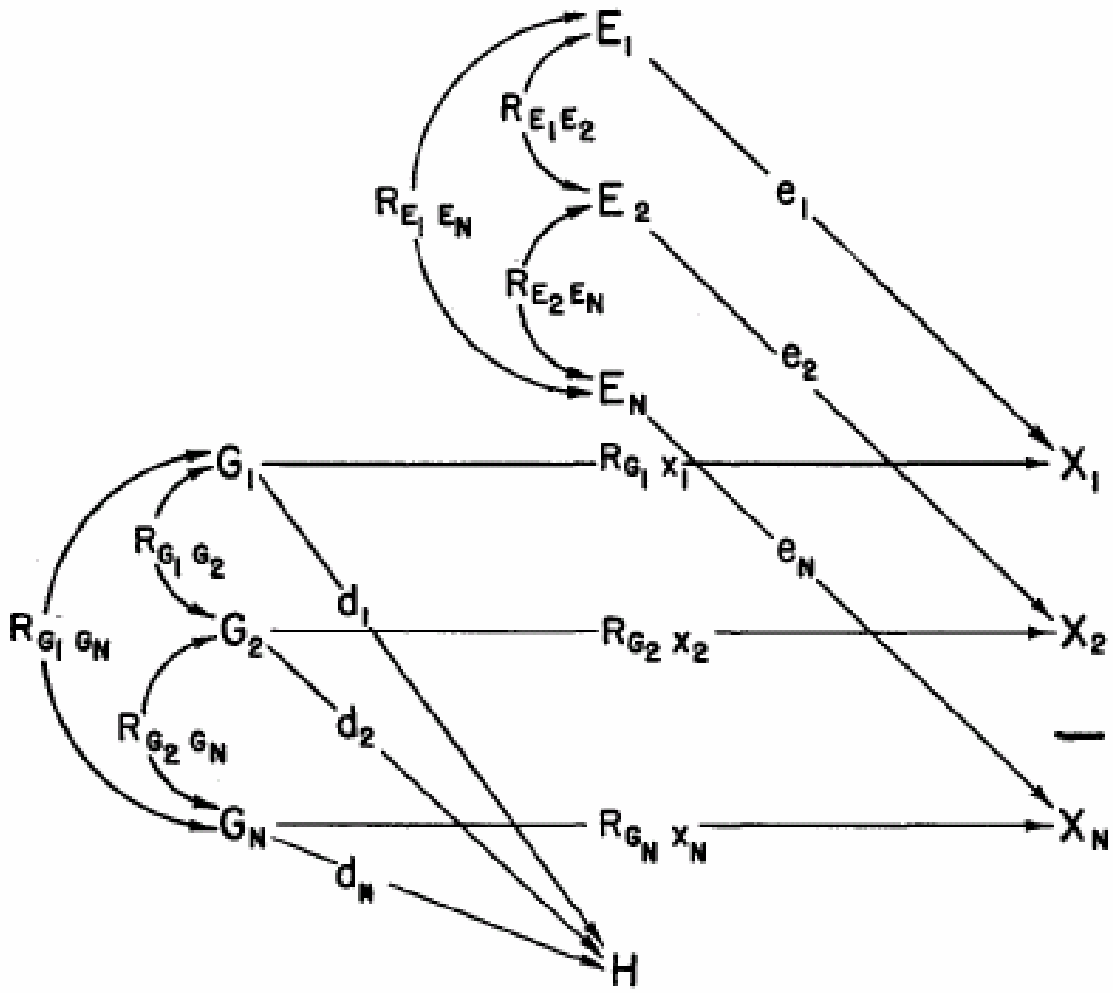


FIGURE 1. Path coefficient diagram showing the relation between phenotypic measurements (X_i) and the aggregate genotype (H). For further explanation of symbols see text.

Both PCA and Cholesky decomposition "rewrite" the data

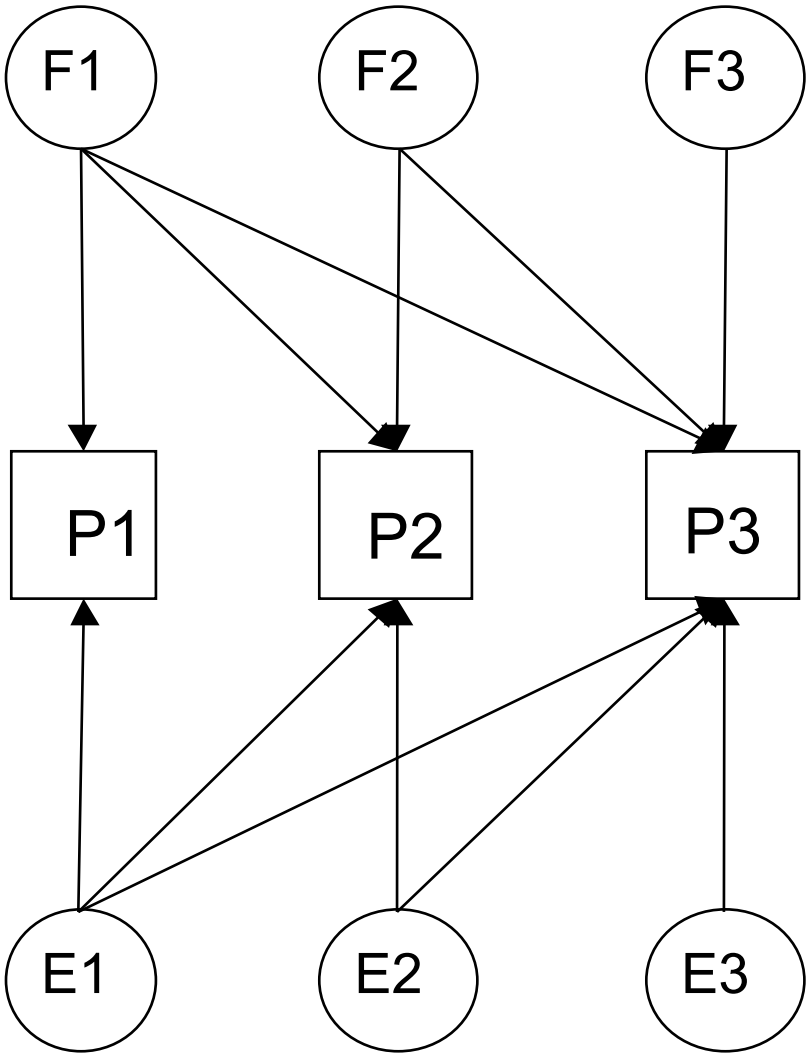
Principal components analysis (PCA): S = P D P' = P* P*'

where S = observed covariance matrix
P'P = I (eigenvectors)
D = diagonal matrix (containing eigenvalues)
P* = P (D1/2)

The first principal component: $y_1 = p_{11}x_1 + p_{12}x_2 + \ldots + p_{1q}x_q$
second principal component: $y_2 = p_{21}x_1 + p_{22}x_2 + \ldots + p_{2q}x_q$
etc.

[p11, p12, … , p1q] is the first eigenvector
d11 is the first eigenvalue (variance associated with y1)

Familial model for 3 variables (can be generalized to p traits)



F: Is there familial (G or C) transmission?

E: Is there transmission of non-familial influences?

Both PCA and Cholesky decomposition "rewrite" the data

Cholesky decomposition: S = F F'
where F = lower diagonal (triangular)

For example, if S is 3 x 3, then F looks like:

f11	0	0
f21	f22	0
f31	f32	f33

And P3 = f31*F1 + f32*F2 + f33*F3

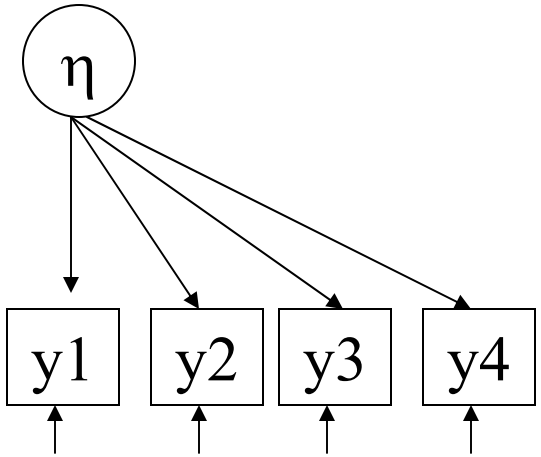
If # factors = # variables, F may be rotated to P*. Both approaches give a transformation of S. Both are completely determinate.

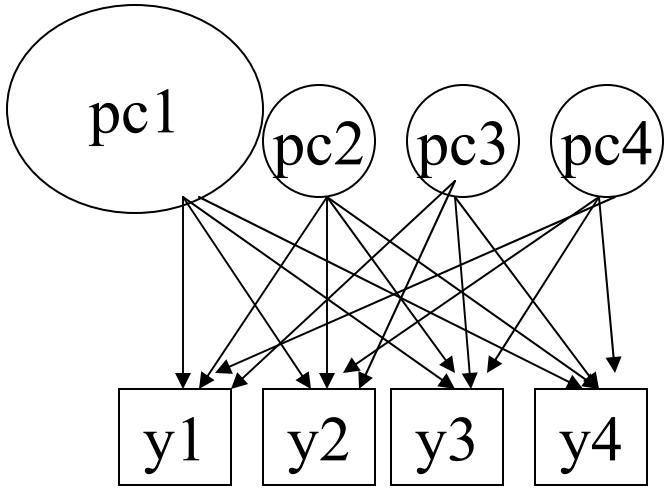
Multivariate phenotypes & multiple QTL effects

For the QTL effect, multiple orthogonal factors can be defined (Cholesky decompostion or triangular matrix).

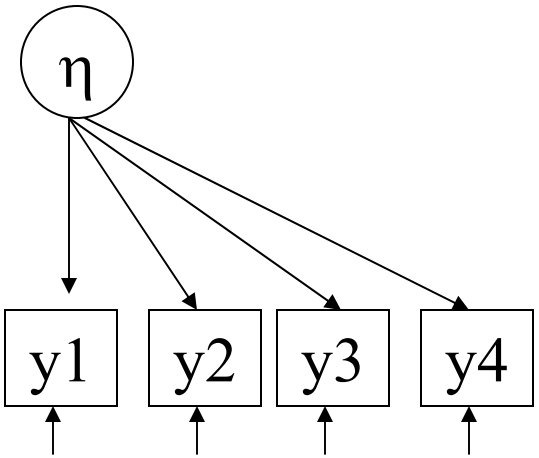
By permitting the maximum number of factors that can be resolved by the data, it is theoretically possible to detect effects of multiple QTLs that are linked to a marker (Vogler et al. Genet Epid 1997)

From multiple latent factors (Cholesky / PCA) to 1 common factor

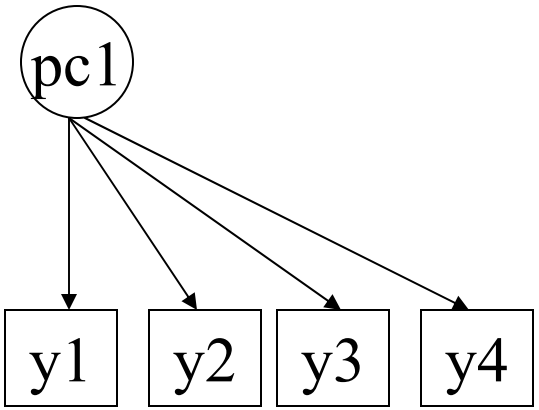




If pc1 is large, in the sense that it account for much variance



=>



Then it resembles the common factor model (without unique variances)

Multivariate QTL effects

Martin N, Boomsma DI, Machin G, A twin-pronged attack on complex traits, Nature Genet, 17, 1997

See: www.tweelingenregister.org

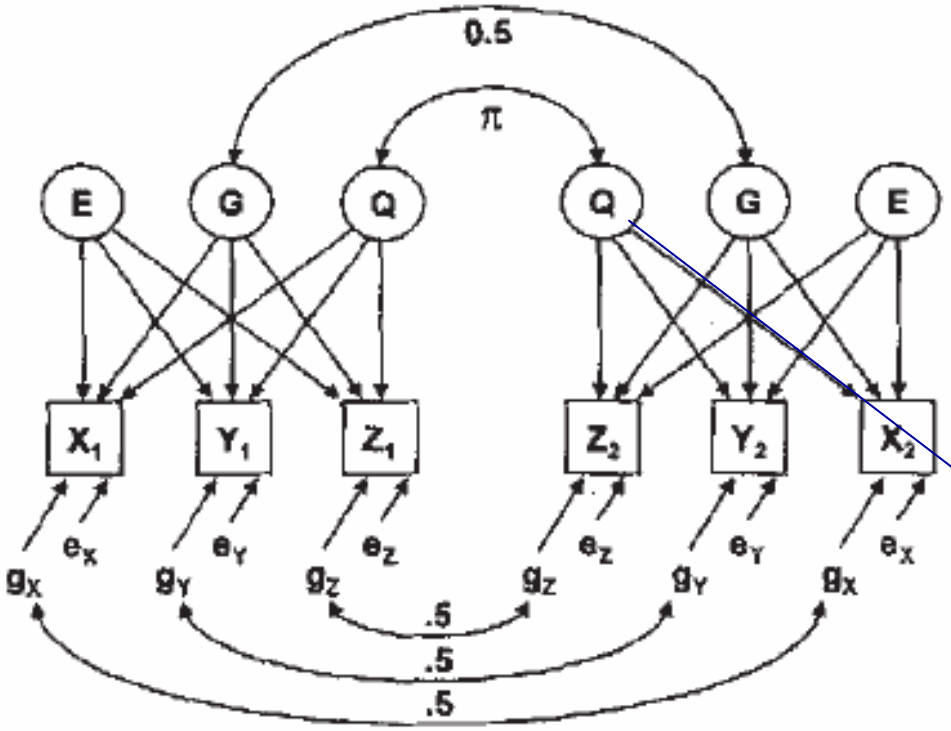


Fig.1 Multivariate path model showing quantitative trait locus (Q), genetic background (G) and environmental factors (E) common to three phenotypes (X, Y and Z) plus genetic (g) and environmental (e) factors unique to each trait. Traits are measured in two siblings, or DZ twins.

QTL modeled as a common factor

Multivariate QTL analysis

- Insight into etiology of genetic associations (pathways)
- Practical considerations (e.g. longitudinal data: use all info)
- Increase in statistical power:

Boomsma DI, Dolan CV, A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor-scores, Behav Genet, 28, 329-340, 1998

Evans DM. The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. Am J Hum Genet. 2002, 1599-602

Marlow et al. Use of multivariate linkage analysis for dissection of a complex cognitive trait. Am J Hum Genet. 2003, 561-70 (see next slide)

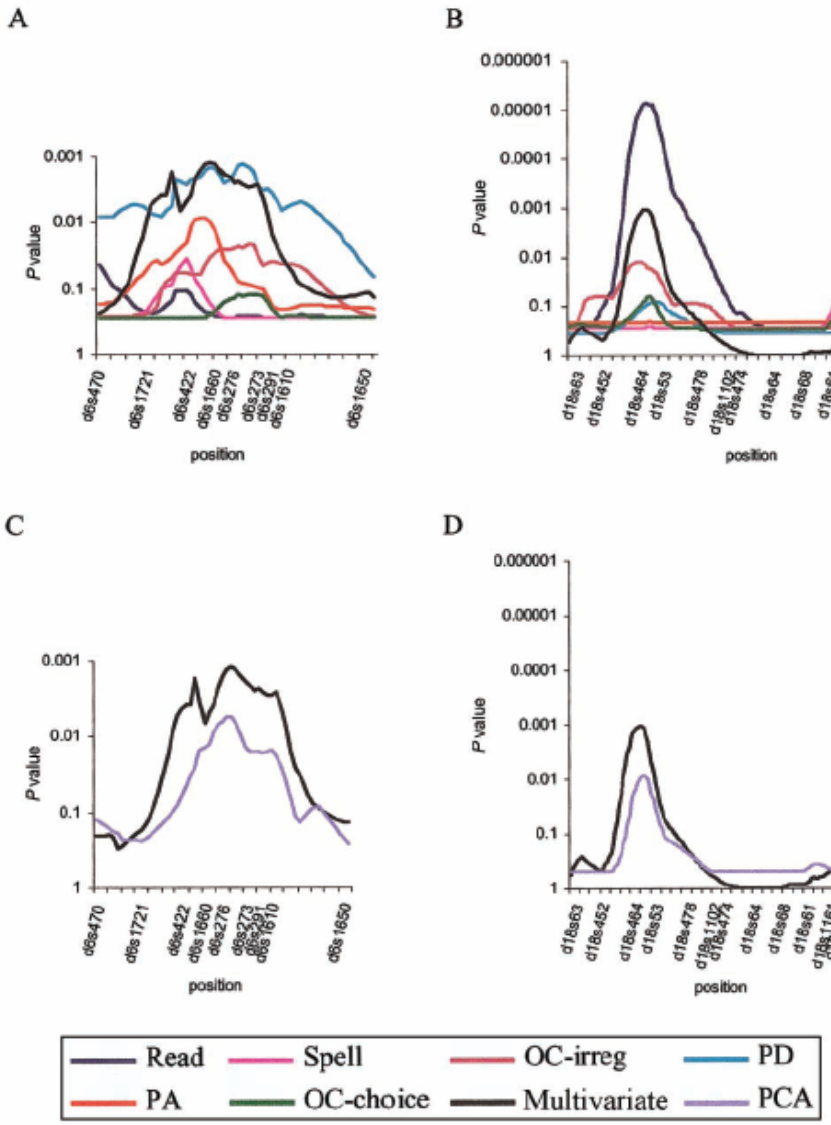


Figure 1 Multivariate and univariate linkage analysis of the six reading-related measures—on a 54-cM region of chromosome 6p (*A*) and a 137-cM region spanning the whole of chromosome 18 (*B*)—and comparison of multivariate linkage and use of the first factor from a PCA approach as the phenotypic measure for linkage analysis, on chromosomes 6p (*C*) and 18 (*D*). A subset of the markers are shown on the graphs. The significance of the linkage results are reported in all cases as *P* values. For univariate measures, the *P* values are empirically derived as described elsewhere (Fisher et al. 2002*a*); for multivariate and PCA results, the *P* values are asymptotic, as described in the text.

Analysis of LDL (low-density lipoprotein), APOB (apo-lipoprotein-B) and APOE (apo-lipoprotein E) levels

- phenotypic correlations
- MZ and DZ correlations
- first (univariate) QTL analysis: partitioned twin analysis (PTA)
- generalize PTA to trivariate data
- multivariate (no QTL model)
- multivariate (QTL)

Multivariate analysis of LDL, APOB and APOE

Phenotypic Correlations			
	LDL	APOB	APOE
LDL	1.00		
APOB	0.88	1.00	
APOE	0.27	0.24	1.00

Multivariate analysis of LDL, APOB and APOE

MZ Correlations			
	LDL TW1	APOB TW1	APOE TW1
LDL TW2	0.75	0.76	0.41
APOB TW2	0.68	0.77	0.37
APOE TW2	0.32	0.31	0.88

DZ Correlations			
	LDL TW1	APOB TW1	APOE TW1
LDL TW2	0.45	0.47	-0.04
APOB TW2	0.36	0.44	-0.06
APOE TW2	0.09	0.06	0.51

Genome-wide scan in DZ twins : lipids

Genotyping in the 117 DZ twin pairs was done for markers with an average spacing of 8 cM on chromosome 19 (see Beekman et al.).

IBD probabilities were obtained from Merlin 1.0 and was calculated as 0.5 x IBD1 + 1.0 x IBD2 for every 2 cM on chromosome 19.

Beekman M, et al. Combined association and linkage analysis applied to the APOE locus. Genet Epidemiol. 2004, 26:328-37.

Beekman M et al. Evidence for a QTL on chromosome 19 influencing LDL cholesterol levels in the general population. Eur J Hum Genet. 2003, 11:845-50

Genome-wide scan in DZ twins

- Marker-data: calculate proportion alleles shared identical-by-decent ($\hat{\pi}$)

- $\hat{\pi} = \pi_1/2 + \pi_2$

- IBD estimates obtained from Merlin
- Decode genetic map

Quality controls:
- MZ twins tested
- Check relationships (GRR)
- Mendel checks (Pedstats / Unknown)
- Unlikely double recombinants (Merlin)

Partitioned twin analysis:

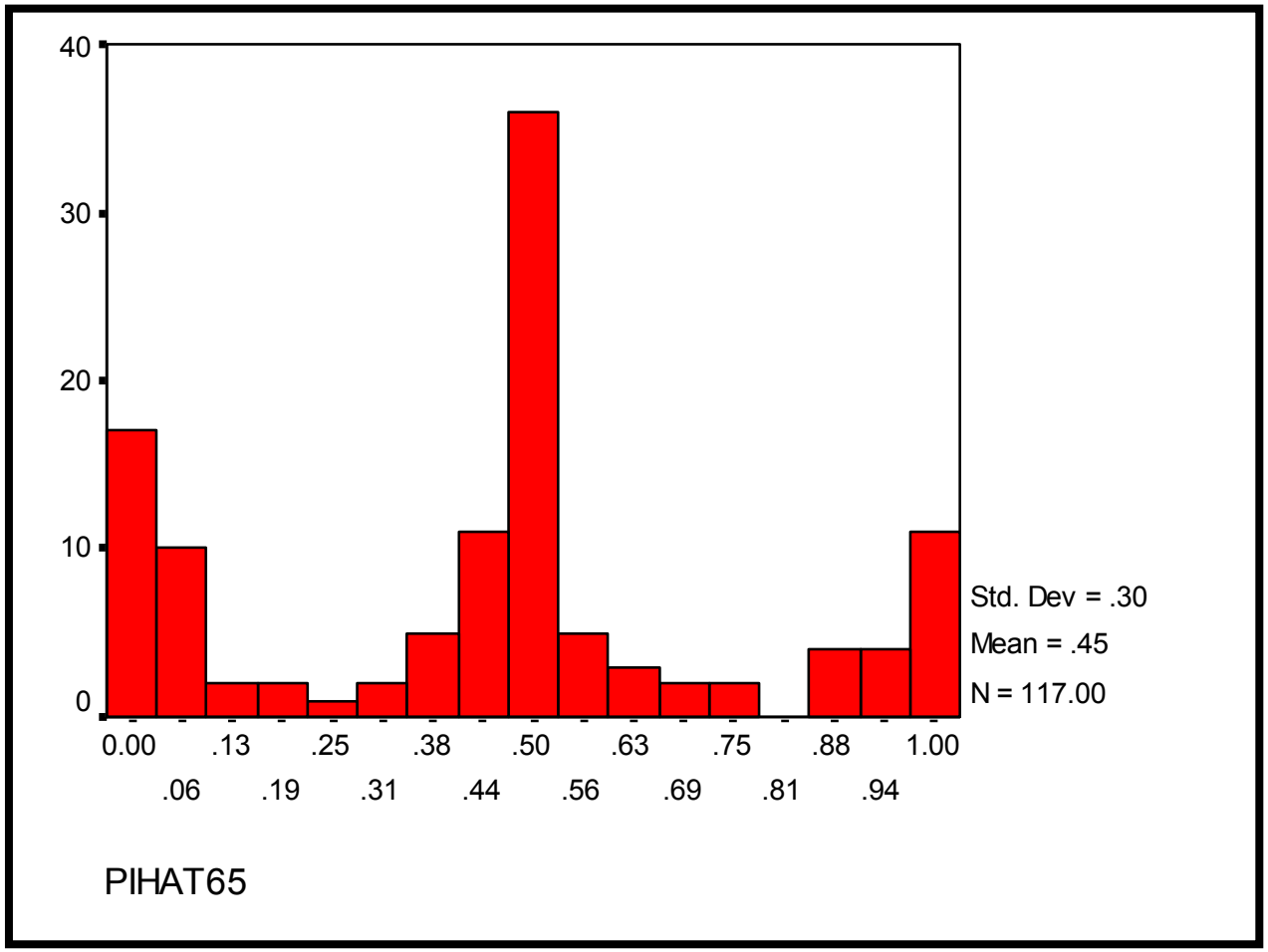
Can resemblance (correlations) between sib pairs / DZ twins, be modeled as a function of DNA marker sharing at a particular chromosomal location? (3 groups)

IBD = 2 (all markers identical by descent)
IBD = 1
IBD = 0

Are the correlations (in lipid levels) different for the 3 groups?

Adult Dutch DZ pairs: distribution pi-hat (π) at 65 cM (chromosome 19).

π = IBD/2; all pairs with π <0.25 have been assigned to IBD=0 group; all pairs with π > 0.75 to IBD=2 group; others to the IBD=1 group.



Exercise

- Model DZ correlation in LDL as a function of IBD
- Test if the 3 correlations are the same
- Add data of MZ twins
- Test if the correlation in the DZ group with IBD = 2 is the same as the MZ correlation
- Repeat for apoB and ln(apoE) levels

- Do cross-correlations (across twins/across traits) differ as a function of IBD? (trivariate analysis)

Basic scripts & data (LDL, apoB, apoE)

- Correlation estimation in DZ: BasicCorrelationsDZ(ibd).mx
- Complete (MZ + DZ + tests) job: AllCorrelations(ibd).mx

- Information on data: datainfo.doc
- Datafiles: DZ: partionedAdultDutch3.dat
  MZ: AdultDutchMZ3.dat

Correlations as a function of IBD

	IBD2	IBD1	IBD0	MZ
LDL	0.81	0.49	-0.21	0.78
ApoB	0.64	0.50	0.02	0.79
lnApoE	0.83	0.55	0.14	0.89

Evidence for linkage?
Evidence for other QTLs?

Correlations as a function of IBD

chi-squared tests

	all DZ equal	DZ(ibd2)=MZ
LDL	21.77	0.0975
apoB	7.98	1.53
apoE	12.45	0.576
	(df=2)	(df=1)
	NO	YES

Linkage analysis in DZ / MZ twin pairs

3 DZ groups: IBD=2,1,0 (π=1, 0.5, 0)

Model the covariance as a function of IBD

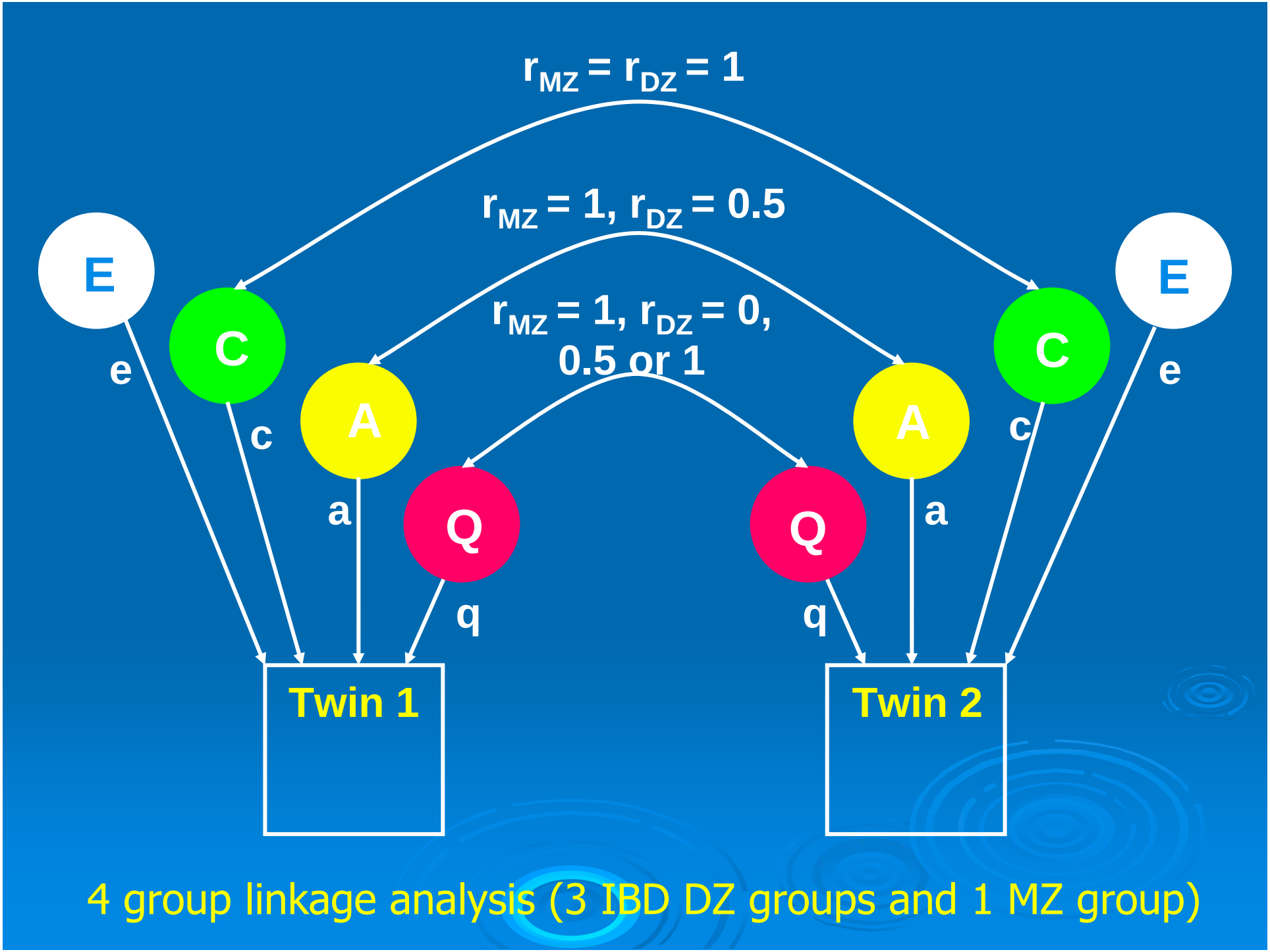
Allow for background familial variance

Total variance also includes E

Covariance = πQ + F + E

Variance = Q + F + E

MZ pairs: Covariance = Q + F + E



4 group linkage analysis (3 IBD DZ groups and 1 MZ group)

Exercise

- Fit FQE model to DZ data (i.e. F=familial, Q=QTL effect, E=unique environment)
- Fit FE model to DZ lipid data (drop Q)
- Is the QTL effect significant?

- Add MZ data: ACQE model (A= additive genetic effects, C=common environment), does this change the estimate / significance of QTL?

Basic script and data (LDL, apoB, apoE)

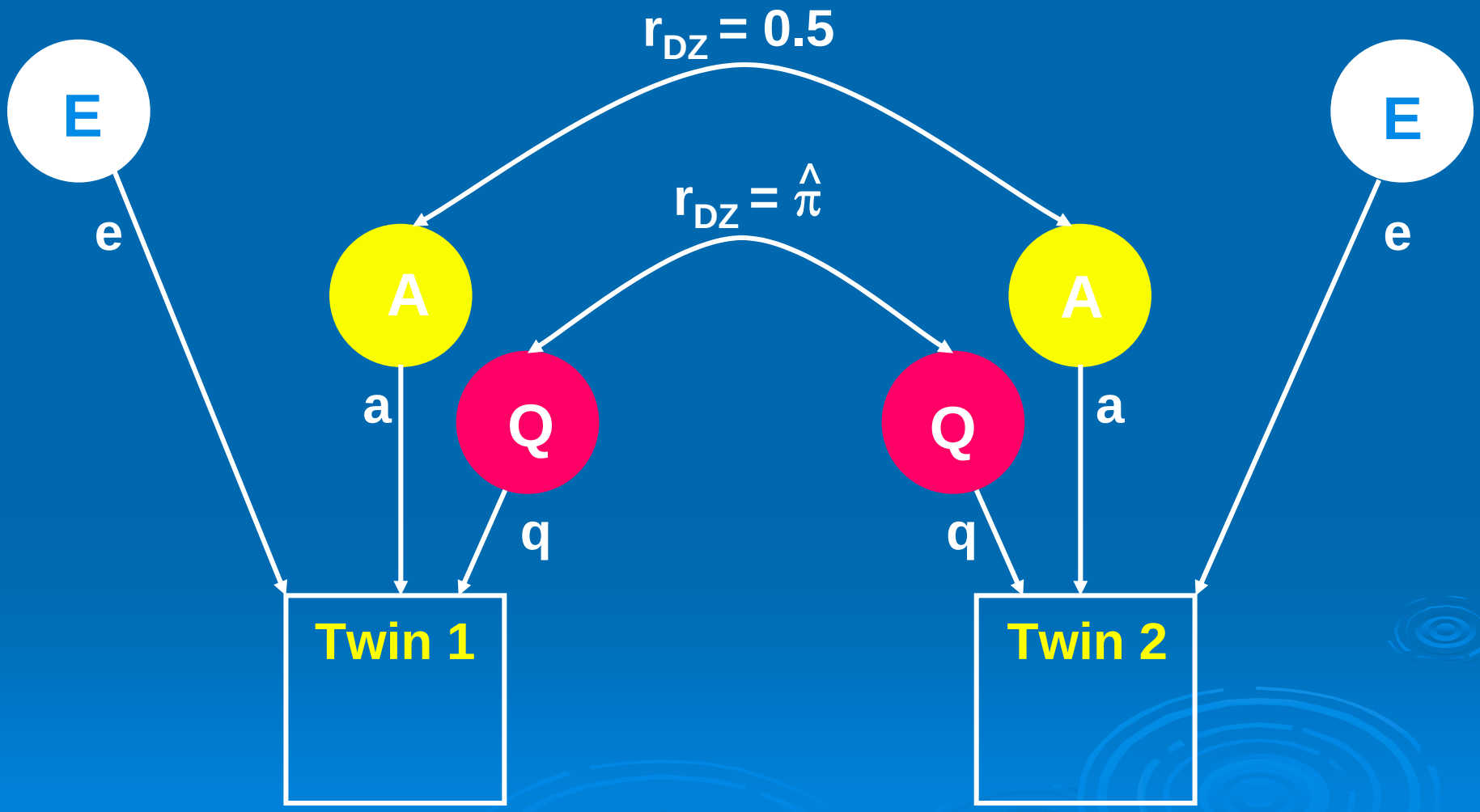
- FQE model in DZ twins: FQEmodel-DZ.mx
- Complete (MZ data + DZ data + tests) job: ACEQ-mzdz.mx

- Information on data: datainfo.doc
- Datafiles: DZ: partionedAdultDutch3.dat
  MZ: AdultDutchMZ3.dat

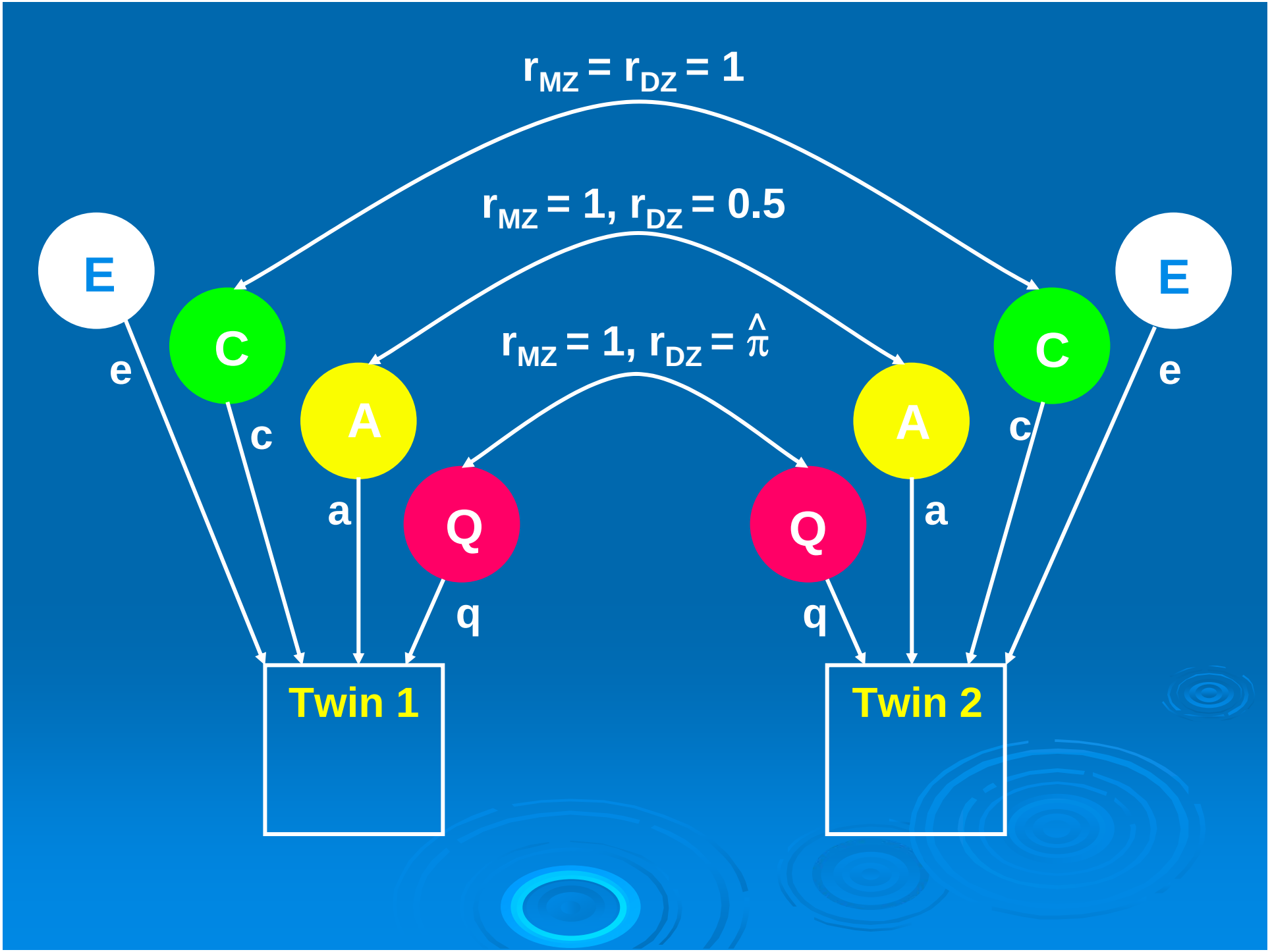
Test of the QTL: chi-squared test (df = 1)

	DZ pairs	DZ+MZ pairs
LDL	12.247	12.561
apoB	1.945	2.128
apoE	12.448	12.292

Use pi-hat: single group analysis (DZ only)



Exercise: PiHatModelDZ.mx


$r_{MZ} = r_{DZ} = 1$
$r_{MZ} = 1, r_{DZ} = 0.5$
$r_{MZ} = 1, r_{DZ} = \hat{\pi}$
E
C
A
Q
E
C
A
Q
e
c
a
q
q
a
c
e
Twin 1
Twin 2

Summary of univariate jobs

- basicCorrelations: DZ (ibd) correlations
- Allcorrelations: plus MZ pairs
- Tricorrelations: trivariate correlation matrix

- FQEmodel-dz.mx
- PIhatModel-dz.mx
- aceq-mzdz.mx

Multivariate analysis of LDL, APOB, and APOE

- use MZ and DZ twin pairs
- fixed effect of age and sex on mean values
- model the effects of additive genes, common and unique environment (ACE model)
- test the significance of common environment (and / or of additive genetic influences)

Multivariate analysis of LDL (low-density lipids), APOB (apo-lipoprotein-B) and APOE (apo-lipoprotein E)

- Cholesky decomposition (obtain the genetic correlations among traits): lipidchol no QTL.mx
- Common factor model (i.e. all correlations of latent factors are unity) :

lipid Common Factor no qtl.mx

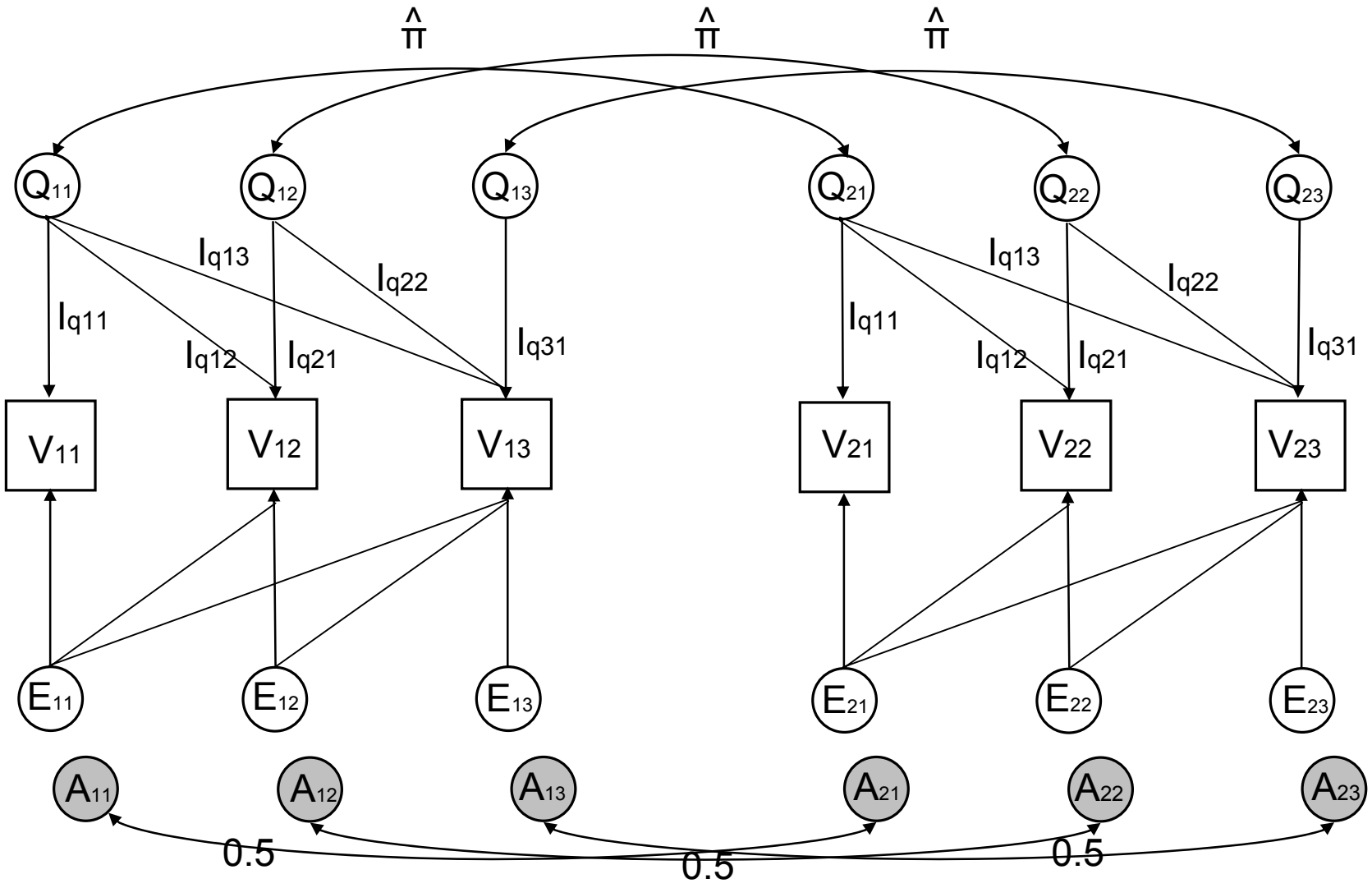
Effect of C not significant

Genetic correlations among LDL, APOB and LNAPOE (Cholesky no QTL)

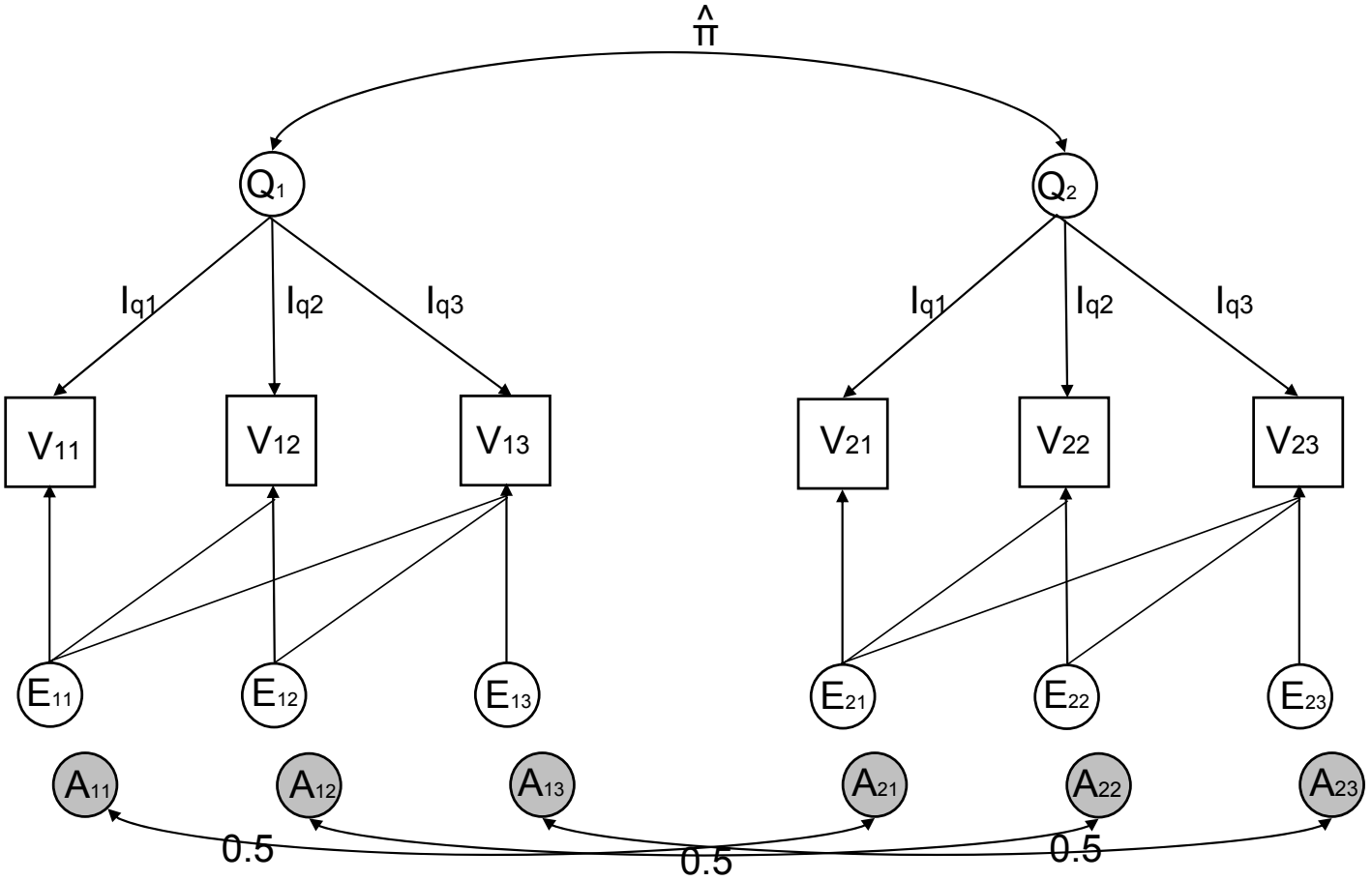
- MATRIX N
- This is a computed FULL matrix of order 3 by 3
- [=\STND(A)]

	1	2	3
1	1.0000	0.9559	0.2157
2	0.9559	1.0000	0.1867
3	0.2157	0.1867	1.0000

Cholesky decomposition: 3 QTL's (latent factors) influencing 3 (observed) lipid traits



QTL as a common factor



A (additive genetic) background and E (unique environment) modeled as Choleky

Tests of multivariate QTL: more than 1 df

- Take the $\chi 2$ distribution with *n* df, where *n* is equal to the difference in number of estimated variance components between the QTL / no QTL models.
- Convert back p-values to a $\chi 2$ value with 1 degree of freedom This $\chi 2$ value can then be divided by 2ln(10) to obtain a LOD score.
- Given that we ignore the mixture distribution problem, the p-values the results will be too conservative (see e.g. Visscher, 2006 in TRHG).

2 jobs for QTL analysis

- Cholesky decomposition for QTL:

lipidchol QTL.mx

- Common factor model for QTL:

lipid Common Factor no qtl.mx

Run the jobs and test for significance of the QTL effect

Include MZ twins (What are the IBD0, IBD1 and IBD2 probabilities?)

Summary: uni- and multivariate