Family-based GWAS

(clustered data)

Camelia Minica

Conor Dolan /Jacqueline Vink

Dorret Boomsma

Sarah Medland (prac on Thursday morning; see also BGA 2010)

VU VRIJE UNIVERSITEIT AMSTERDAM

Faculteit der Psychologie en Pedagogiek

European Journal of Human Genetics (2014), 1–7
© 2014 Macmillan Publishers Limited All rights reserved 1018-4813/14
www.nature.com/ejhg

ARTICLE

Sandwich corrected standard errors in family-based genome-wide association studies

Camelia C Minică*[,1], Conor V Dolan[1], Maarten MD Kampert[2], Dorret I Boomsma[1] and Jacqueline M Vink[1]

Molecular Psychiatry (2014), 1–2
© 2014 Macmillan Publishers Limited All rights reserved 1359-4184/14
www.nature.com/mp

LETTER TO THE EDITOR

MZ twin pairs or MZ singletons in population family-based GWAS? More power in pairs

Why is this important?

Ignoring clustering in the data may lead to wrong conclusions (point estimates of effects OK, but SE too small)

The focus is on family-based Genome-Wide Association Studies.
However the analytic strategies to be discussed are regression based approaches, hence, relevant for any analysis involving family-data ; that is, the predictor can be a Genetic Variant, a polygenic score, or any other covariate one might be interested in.

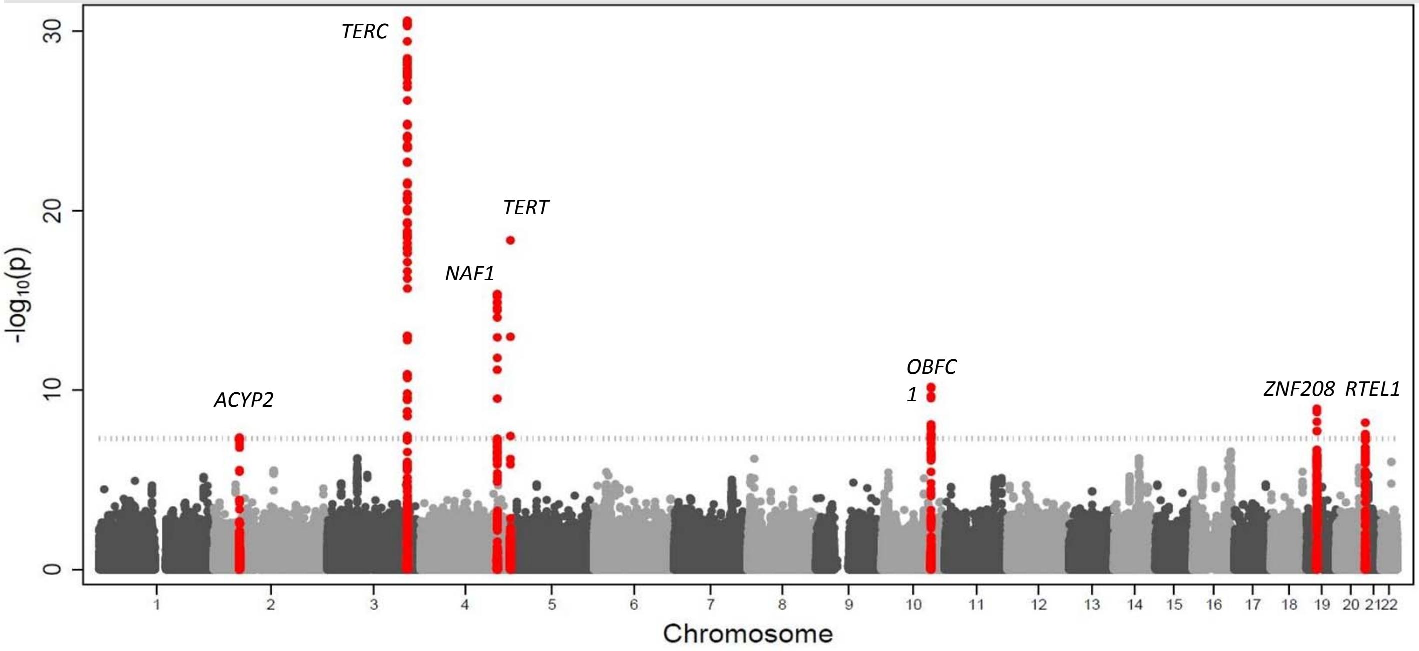
Why is this important?

- Many GWA meta-analyses rely heavily on twin registries
- Twin registries have data collected in families readily available

Identification of seven loci affecting mean TELOMERE length and their association with disease

Veryan Codd et al. (ENGAGE consortium) *Nature Genetics*, 2013

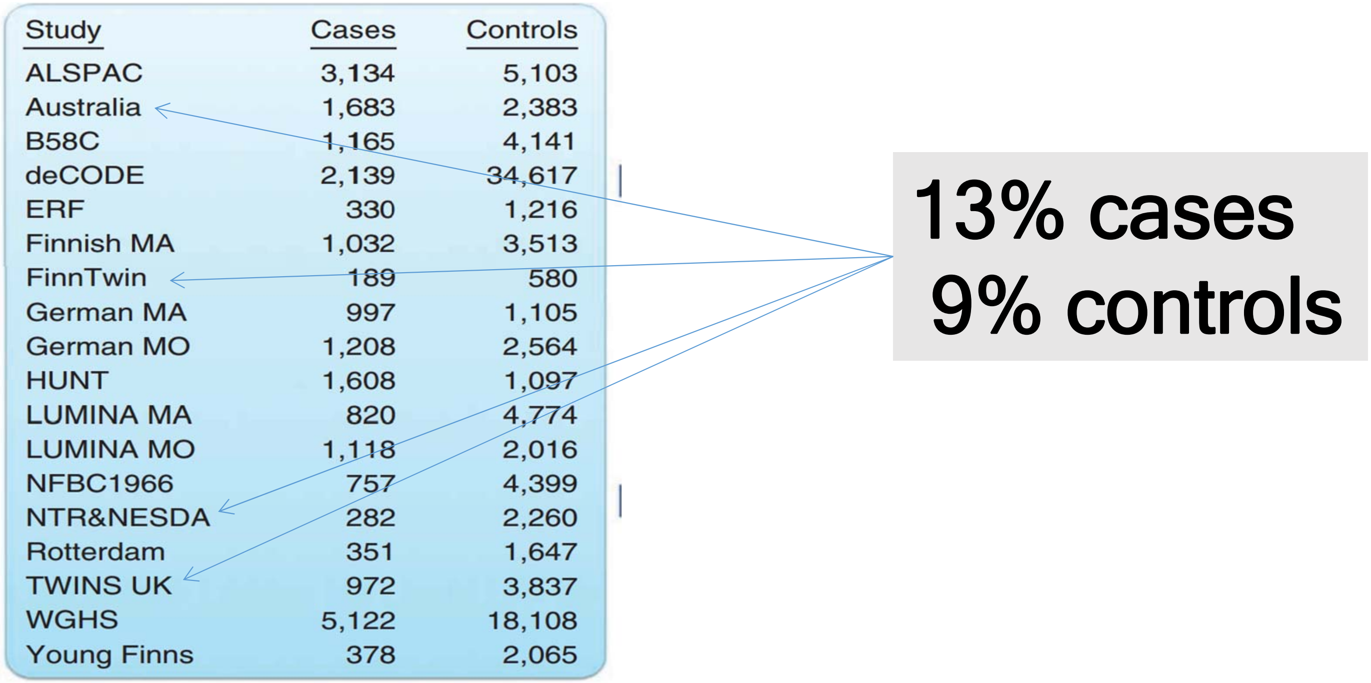
Twin registries supplied 34% of samples



Nature genetics

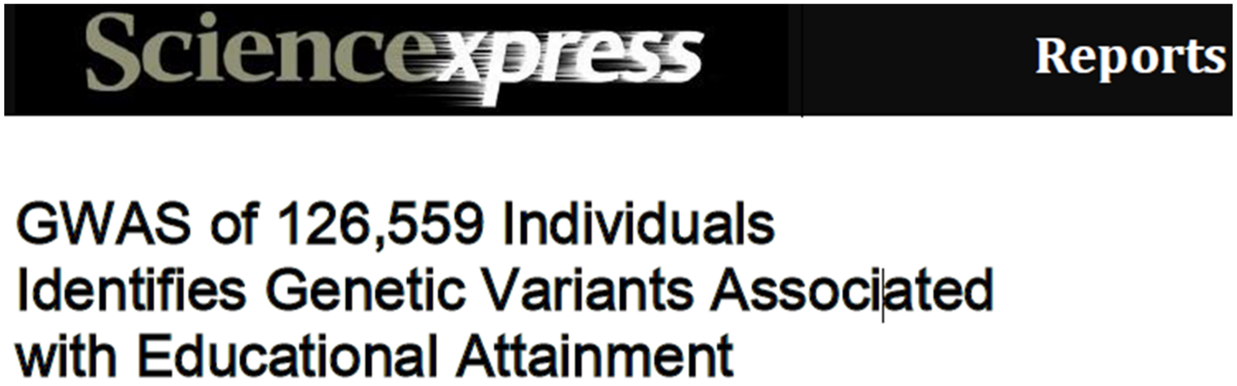
Genome-wide meta-analysis identifies new susceptibility loci for migraine

Verneri Anttila, Bendik S. Winsvold, [...], and Aarno Palotie



Study	Cases	Controls
ALSPAC	3,134	5,103
Australia	1,683	2,383
B58C	1,165	4,141
deCODE	2,139	34,617
ERF	330	1,216
Finnish MA	1,032	3,513
FinnTwin	189	580
German MA	997	1,105
German MO	1,208	2,564
HUNT	1,608	1,097
LUMINA MA	820	4,774
LUMINA MO	1,118	2,016
NFBC1966	757	4,399
NTR&NESDA	282	2,260
Rotterdam	351	1,647
TWINS UK	972	3,837
WGHS	5,122	18,108
Young Finns	378	2,065

13% cases
9% controls



Sciencexpress Reports

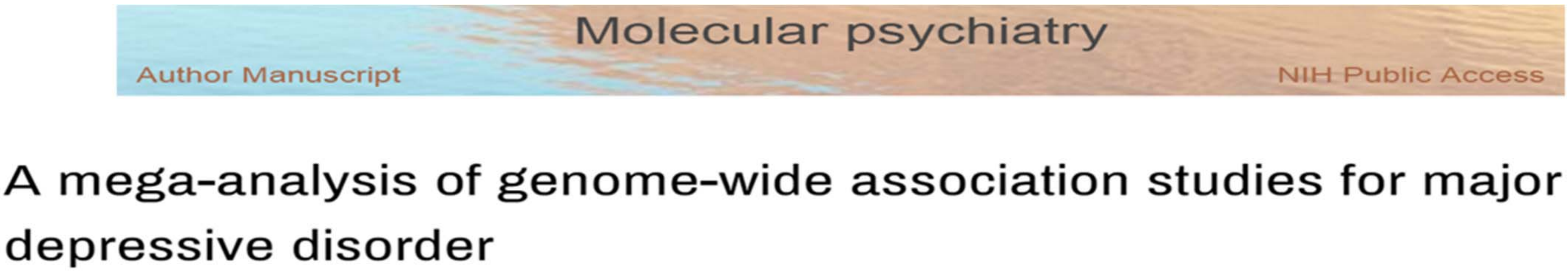
GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

There are 6 twin cohorts and total of 52 cohorts (11%)

- ***Finnish twin cohort***
- ***Netherlands twin register***
- ***QIMR (Queensland twin register)***
- ***Swedish twin register***
- ***TwinsUK***
- ***Minnesota Twin – family study***

Twin registries supplied > 35% of total sample size

Some consortia protocols require discarding family members



Molecular psychiatry

Author Manuscript

NIH Public Access

A mega-analysis of genome-wide association studies for major depressive disorder

Twin registries supplied 31% cases and 19% controls

UNRELATEDS

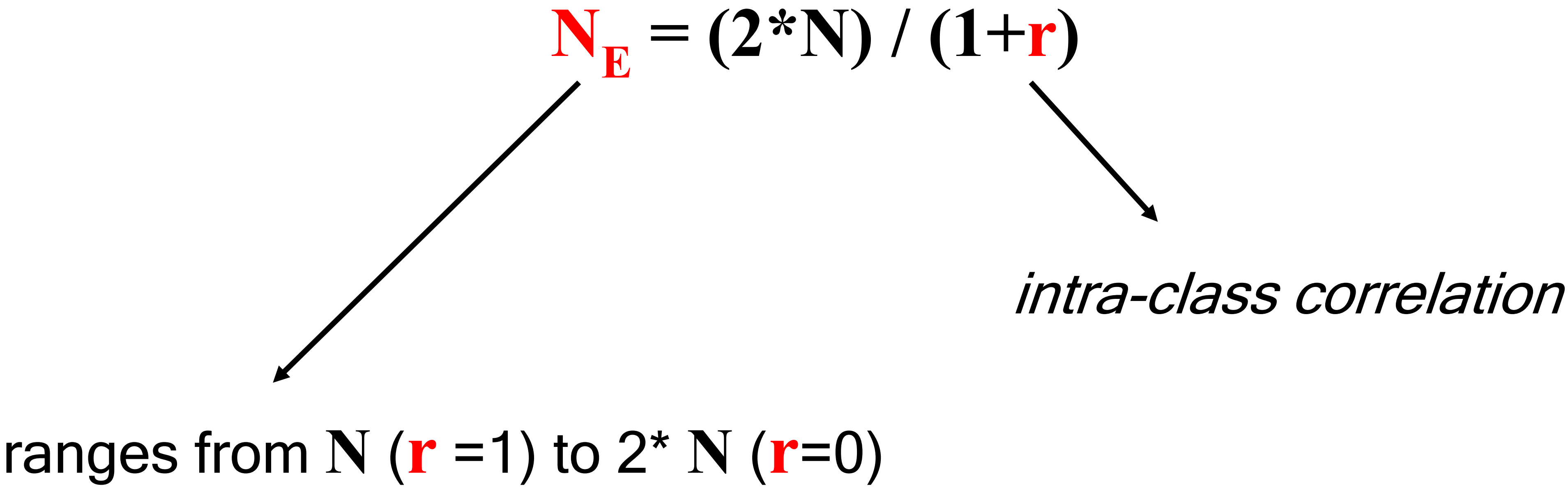
MZ pairs

or

MZ singletons?

MZ pairs or MZ singletons?

- Compute effective sample size:



$$N_E = (2*N) / (1+r)$$

intra-class correlation

ranges from **N** (**r** =1) to 2* **N** (**r**=0)

Assume a study of depression including MZ pairs: take pairs or MZ singletons?

Study 1 ~1890 +
Study 2 ~786 +
Study 3 ~300

= 2976 pairs -> 2976 individuals analyzed

r (MDD) = .35 -> **N_E** =4409

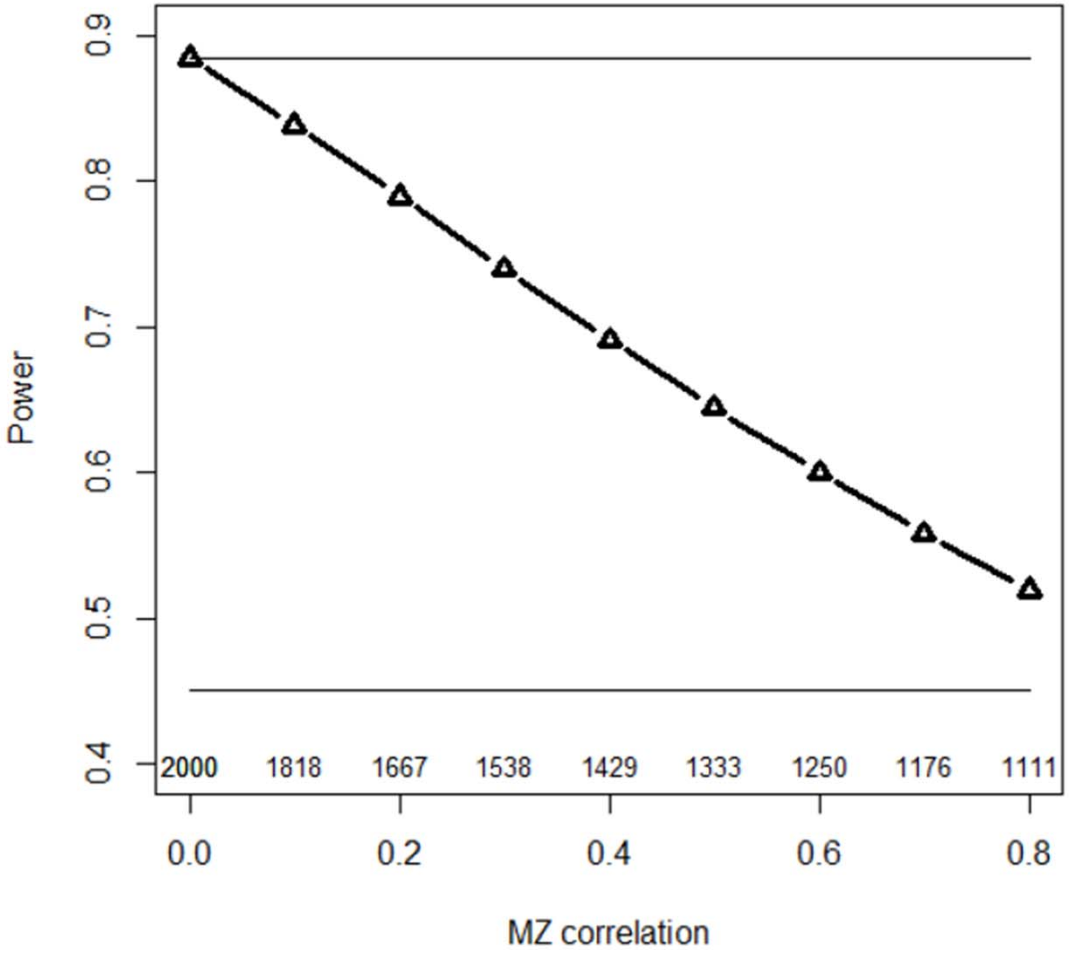
MZ pairs or MZ singletons?

Study 1 ~1890 +
Study 2 ~786 +
Study 3 ~300
= 2976 MZ twin pairs

DISCARDED:

4409-2976 = 1433 ‘unrelated’ -> by restricting the sample to MZ singletons, the equivalent of 1433 unrelated individuals was discarded.

MZ pairs ~~or MZ singletons?~~



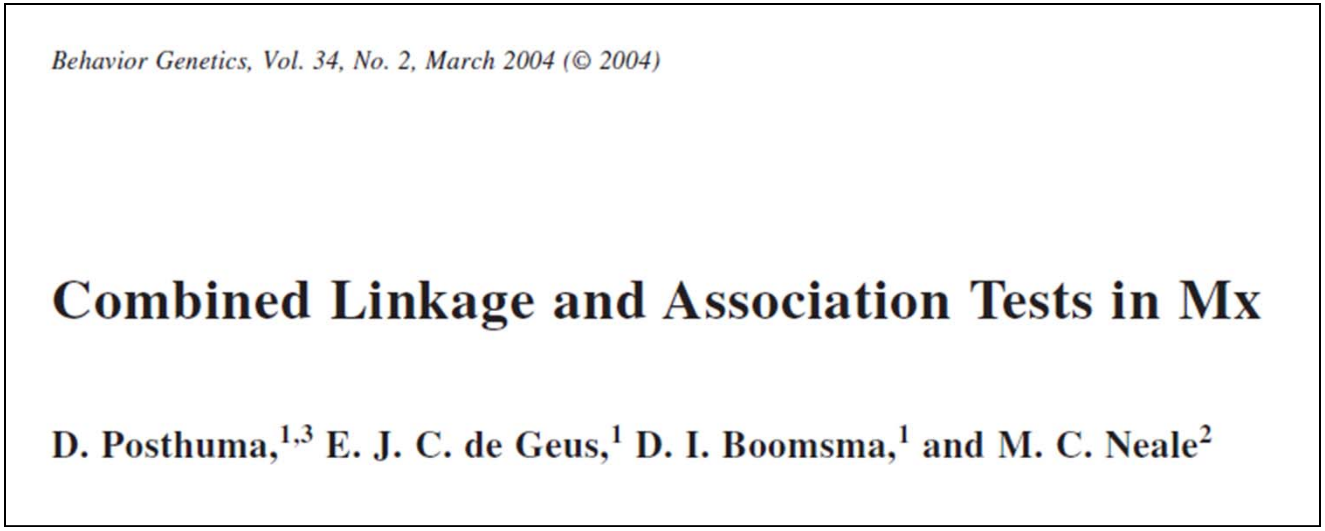
The figure illustrates how power varies as a function of the MZ correlation. The figure also includes the corresponding gains in effective sample size estimated for a sample of 1000 MZ twin pairs and a gene explaining 1% of the phenotypic variance.

Conclusion: Retaining both MZ twins of a pair almost always increases power. Type I error rate is not affected. Reducing MZ twin pairs to singletons is not justified.

FAMILY-BASED GWAS: using efficiently correlated observations

Given the availability of family-based samples, an important questions is: what is the most efficient analytic strategy in the context of GWAS?
We will see that the choice of the model for the familial covariance matrix – the V matrix is central to the efficiency of family-based analyses.

Not about partitioning the association effects into a *between family* component and a *within family* component.
The *between family* effects reflect both the genuine and the possible spurious association between locus alleles and a trait (or allelic association between locus alleles and trait locus alleles).
The *within family* effects reflect only the genuine association.



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]

Family-based GWAS: regression model

$$\mathbf{y}_{\mathrm{ij}} = \mathbf{b}_0 + \mathbf{b}_1 * \mathbf{g}_{\mathrm{ij}} + \boldsymbol{\varepsilon}_{\mathrm{ij}}$$

i is indicator of family (i=1..Nfam) and j is subjects (j=1..N), **y**, **b** and **ε** are vectors, **X** is the matrix of observed predictors – the genotypes, **b** is the vector of parameter for the observed predictor and **y** is the vector of observed phenotypes

$$\mathbf{X} = \begin{pmatrix} 1 & g_1 \\ 1 & g_2 \\ \vdots & \vdots \\ 1 & g_N \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} \mathbf{b}_0 \\ \mathbf{b}_1 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} ph_1 \\ ph_2 \\ \vdots \\ ph_N \end{pmatrix}$$

Family-based GWAS

(model in matrix notation)

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{V})$$

vector of residuals is normally distributed with zero mean and the familial covariance matrix V.

THE FAMILIAL COVARIANCE MATRIX $\mathbf{V}$

The power of the family-based analyses depends on the model one choses for the familial covariance matrix.

V has block diagonal form, with the diagonal blocks V1…VNfam representing the residual covariance matrix for each family. **O** denotes a matrix of zeros - observations on individuals belonging to different families are independent.

$$\boldsymbol{\varepsilon} | \mathbf{X} \sim \mathrm{N}(0, \mathbf{V})$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{O} & & \mathbf{O} \\ \mathbf{O} & \mathbf{V}_2 & & \mathbf{O} \\ & & \ddots & \\ \mathbf{O} & \mathbf{O} & & \mathbf{V}_N \end{bmatrix}_{\text{fam}}$$

The conditional (on the tested SNP) familial covariance matrix can be modeled to accommodate the effects of additive genetic factors, shared and unshared environment; these collectively contribute to the trait variance and are known as variance components (one can also consider dominance components).

$$\varepsilon|X \sim N(0, V)$$

$$V(\Theta)$$

$$\Theta=[\sigma^2{}_A, \sigma^2{}_C, \sigma^2{}_E]$$

$\mathbf{V}$ modeled as an ACE

$$\mathbf{V}(\Theta) = A \otimes \sigma^2_A + C \otimes \sigma^2_C + I \otimes \sigma^2_E$$

$$A = \begin{bmatrix} 1 & 0 & .5 & .5 & 0 & 0 & 0 & 0 \\ 0 & 1 & .5 & .5 & 0 & 0 & 0 & 0 \\ .5 & .5 & 1 & .5 & 0 & 0 & 0 & 0 \\ .5 & .5 & .5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & .5 & .5 \\ 0 & 0 & 0 & 0 & 0 & 1 & .5 & .5 \\ 0 & 0 & 0 & 0 & .5 & .5 & 1 & .5 \\ 0 & 0 & 0 & 0 & .5 & .5 & .5 & 1 \end{bmatrix}$$

Expected (==OBSERVED) proportion of the genome shared IBD

Genetic Relationship Matrix

e.g., 2 parents + 2 DZ twins

Equivalently, estimation may be based on the observed proportion of the genome shared between relatives. Estimation based on expected and observed genetic relationship should give equivalent results (Visscher et al 2006).
Note that the genetic relationships within the family cluster need not be the same (e.g., parental correlation vs. parent-offspring correlation); furthermore, clusters might be more complicated genetically, containing more than one type of sibling (e.g., MZ, DZ, half siblings, with or without parents).

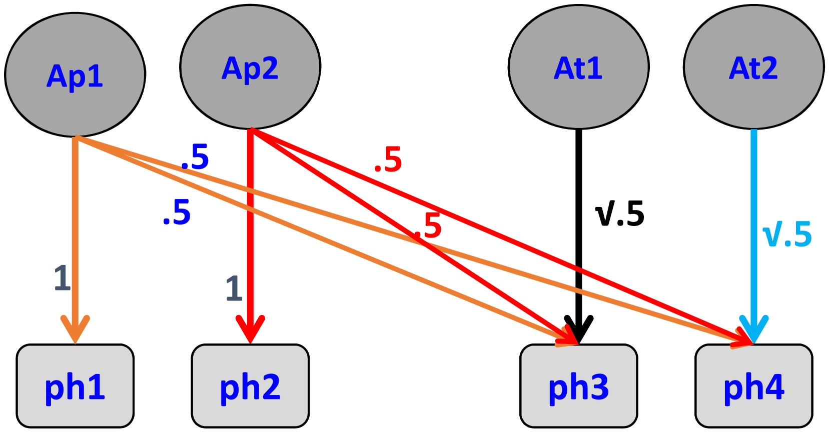
$\mathbf{V}$ modeled as an ACE

$$\mathbf{V}(\boldsymbol{\Theta}) = A \otimes \sigma^2_A + C \otimes \sigma^2_C + I \otimes \sigma^2_E$$

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

V modeled as an AE

$$\mathbf{V}(\Theta) = A \otimes \sigma^2_A + \mathbf{I} \otimes \sigma^2_E$$

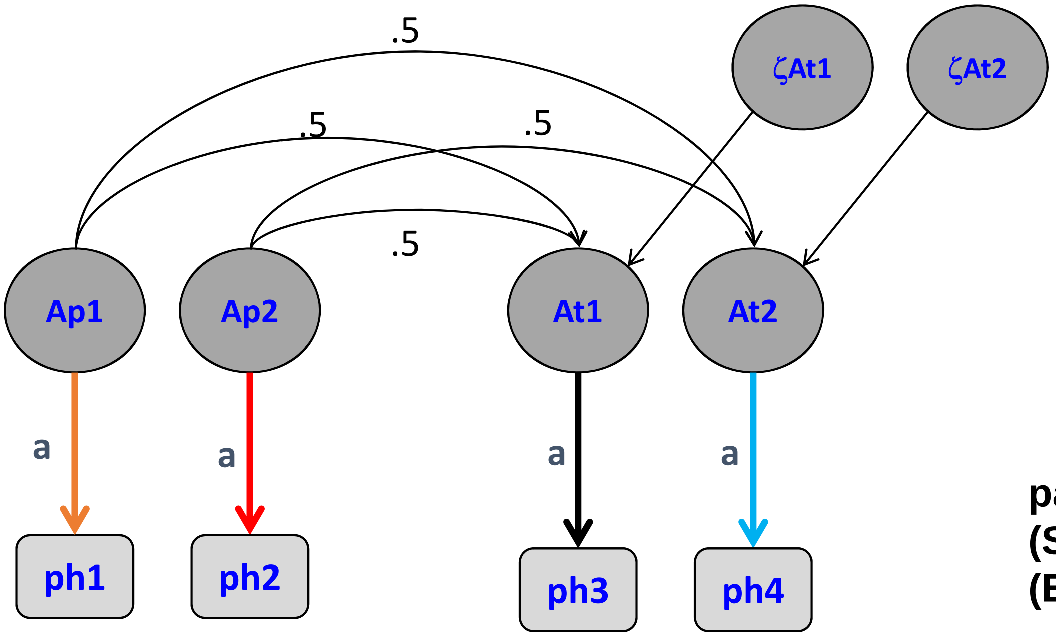


Circles =additive genetic effects (random effects in the linear mixed approach, latent variable in SEM), squares are observed phenotypes.
The weights associated with the random effects are fixed at values that indicate the proportion of additive genetic variance shared between relatives.
The implied covariance matrix: parents are uncorrelated, and each pair within the cluster share half their genome.

This is the additive genetic specification in linear mixed, where
var(Ap1) = var(Ap2) = var(At1) = var(At2) = var(A)
It may look unfamiliar but it is equivalent to the more familiar path diagram

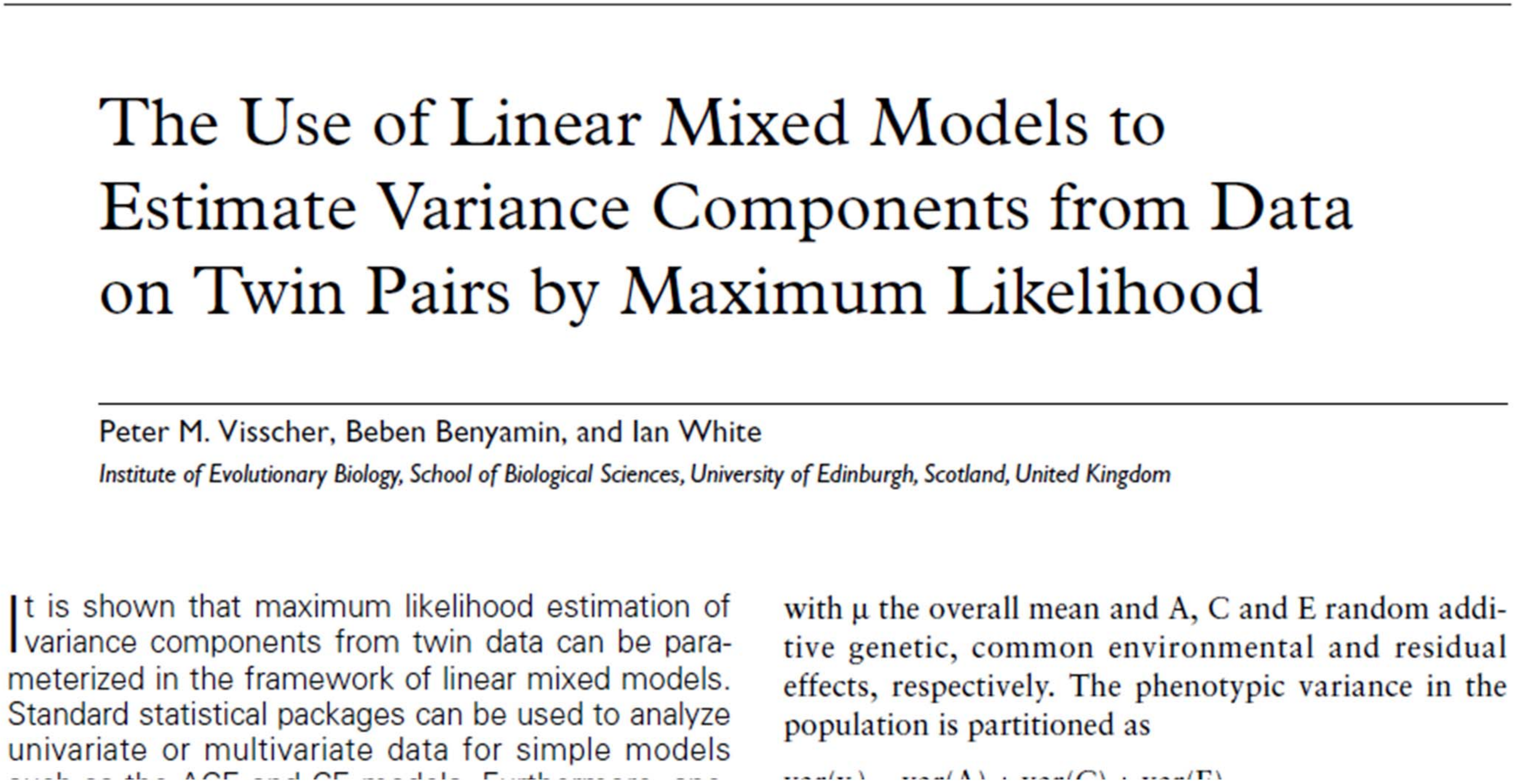
V modeled as an AE

$$\mathbf{V}(\Theta) = A \otimes \sigma^2_A + \mathbf{I} \otimes \sigma^2_E$$



path diagram
(SEM specification)
(E excluded)

Twin Res 2004



The Use of Linear Mixed Models to Estimate Variance Components from Data on Twin Pairs by Maximum Likelihood

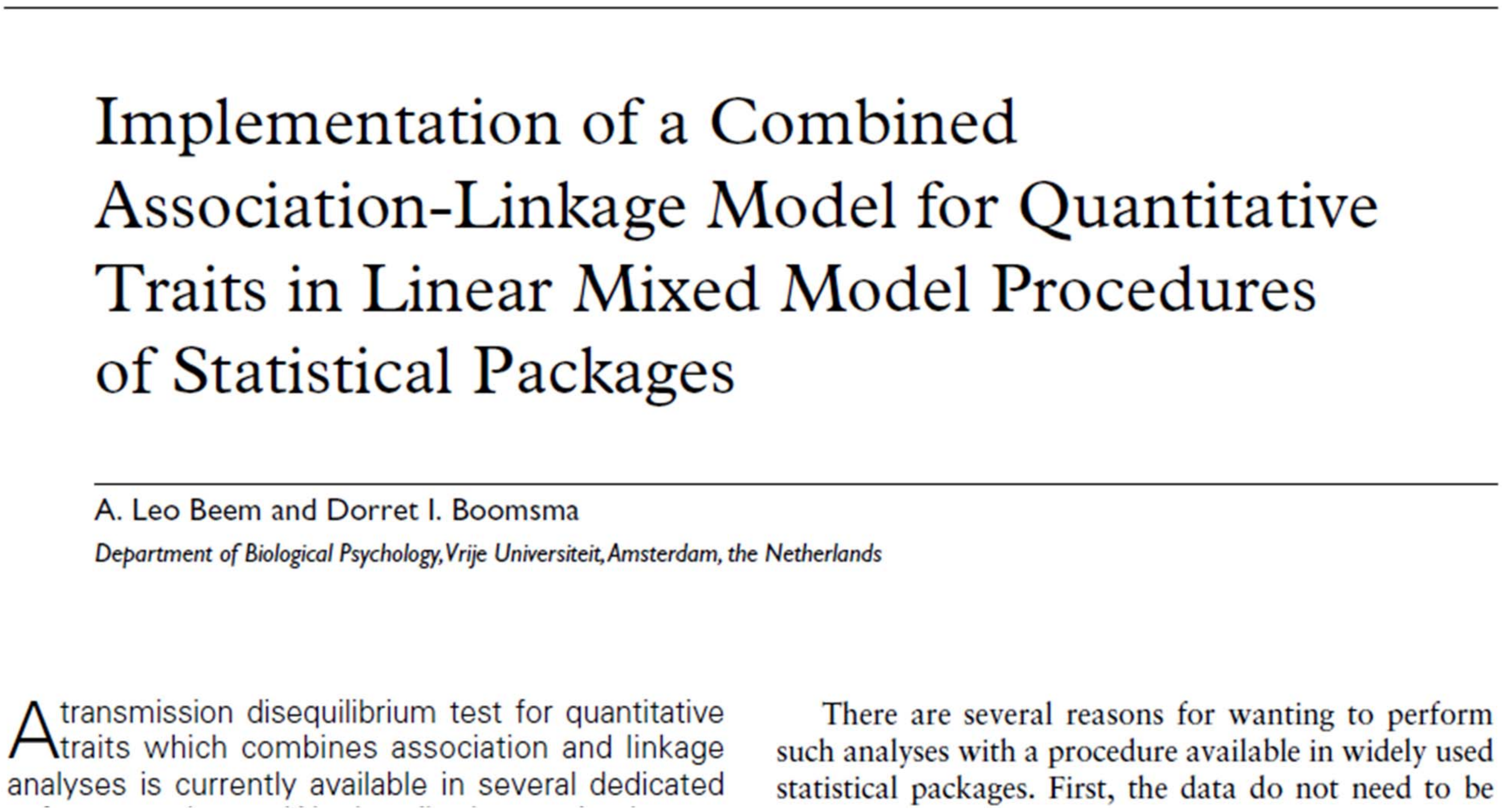
Peter M. Visscher, Beben Benyamin, and Ian White

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Scotland, United Kingdom

It is shown that maximum likelihood estimation of variance components from twin data can be parameterized in the framework of linear mixed models. Standard statistical packages can be used to analyze univariate or multivariate data for simple models

with μ the overall mean and A, C and E random additive genetic, common environmental and residual effects, respectively. The phenotypic variance in the population is partitioned as

Twin Res Hum Genet 2006



Implementation of a Combined Association-Linkage Model for Quantitative Traits in Linear Mixed Model Procedures of Statistical Packages

A. Leo Beem and Dorret I. Boomsma

Department of Biological Psychology, Vrije Universiteit, Amsterdam, the Netherlands

A transmission disequilibrium test for quantitative traits which combines association and linkage analyses is currently available in several dedicated

There are several reasons for wanting to perform such analyses with a procedure available in widely used statistical packages. First, the data do not need to be

$\mathbf{V}$ modeled as an AE

$$\mathbf{V}(\Theta) = \mathrm{A} \otimes \sigma^2_{\mathrm{A}} + \mathbf{I} \otimes \sigma^2_{\mathrm{E}}$$

$$\mathrm{A} = \begin{bmatrix} 1 & 0 & .5 & .5 & 0 & 0 & 0 & 0 \\ 0 & 1 & .5 & .5 & 0 & 0 & 0 & 0 \\ .5 & .5 & 1 & .5 & 0 & 0 & 0 & 0 \\ .5 & .5 & .5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & .5 & .5 \\ 0 & 0 & 0 & 0 & 0 & 1 & .5 & .5 \\ 0 & 0 & 0 & 0 & .5 & .5 & 1 & .5 \\ 0 & 0 & 0 & 0 & .5 & .5 & .5 & 1 \end{bmatrix}$$

Block diagonal structure can be relaxed to ACCOMMODATE DISTANT RELATEDNESS

(e.g., fastLMM, GCTA)

ESTIMATION?

Maximum Likelihood

$$\hat{\mathbf{b}}_{\mathrm{ML}} = \left(\mathbf{X}^{t} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^{t} \mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1} \mathbf{y}$$

$$\mathrm{var}(\hat{\mathbf{b}}_{\mathrm{ML}}) = (\mathbf{X}^{t}\,\mathbf{V}(\widehat{\boldsymbol{\Theta}})^{-1}\mathbf{X})^{-1}$$

The familial covariance matrix is used in the estimation of the parameter of interest (i.e., the SNP effect) and of its variance.
Estimation can be performed by maximum likelihood and involves maximizing the log likelihood function (i.e., finding the values for the parameters that render the derivatives of the likelihood function to equal zero).

Usually V is not known but it has to be estimated from the data by an iterative process. This involves:
a.starting with V, as an identity matrix
b. computing an estimate of beta
c. Use this estimate to form the residuals (y-Xb) and
d. update V
e. The updated V is then used to get a new estimate of beta
And so on, until convergence is achieved.

Maximum Likelihood

$$\hat{\mathbf{b}}_{\text{ML}} = \left(\mathbf{X}^{t}\mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{t}\mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1}\mathbf{y}$$

correct model

$$\text{var}(\hat{\mathbf{b}}_{\text{ML}}) = (\mathbf{X}^{t}\,\mathbf{V}(\widehat{\boldsymbol{\Theta}})^{-1}\mathbf{X})^{-1}$$

One can arrive at a correct estimate of the variance of beta (of the SNP effect) under the condition that the model imposed on the familial covariance matrix V is correctly specified (true).
A correct estimate of the variance is needed as this is used to assess the statistical significance of the SNP effect (e.g. by the means of a Wald test (b/sqrt(varb)).
A correct estimates ensures the conclusions drawn based on the data at hand are correct.
If variance is underestimated, this will inflate the type I error rate (false positives). Overestimation will result in a power loss (false negatives).

What if my model for V is misspecified?

e.g.: model = ACE , but ignore C

the variance/standard errors are no longer correct if you ignore the family clustering due to common environment. The effect of misspecification on the estimate will depend on the contribution of the common environment to the trait variance: the larger the C variance component, the more likely the standard errors will be underestimated.

So what to do? This is where the sandwich comes in. You can use a sandwich to arrive at correct standard errors.

Maximum Likelihood

$$\hat{\mathbf{b}}_{\mathrm{ML}} = \left(\mathbf{X}^{t}\mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{t}\mathbf{V}\left(\hat{\boldsymbol{\Theta}}\right)^{-1}\mathbf{y}$$

SANDWICH correction

misspecification?

$$\mathbf{V}\left(\widehat{\Theta}\right) = [\sigma^{2}_{A},\ \sigma^{2}_{E}]$$

$$\mathrm{var}\left(\hat{\mathbf{b}}_{\mathrm{R-ML}}\right) = \left(\mathbf{X}^{t}\mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{t}\mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})(\mathbf{y}-\mathbf{X}\mathbf{b})^{t}\,\mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1}\mathbf{X}\left(\mathbf{X}^{t}\mathbf{V}\left(\hat{\boldsymbol{\Theta}}_{\mathrm{m}}\right)^{-1}\mathbf{X}\right)^{-1}$$

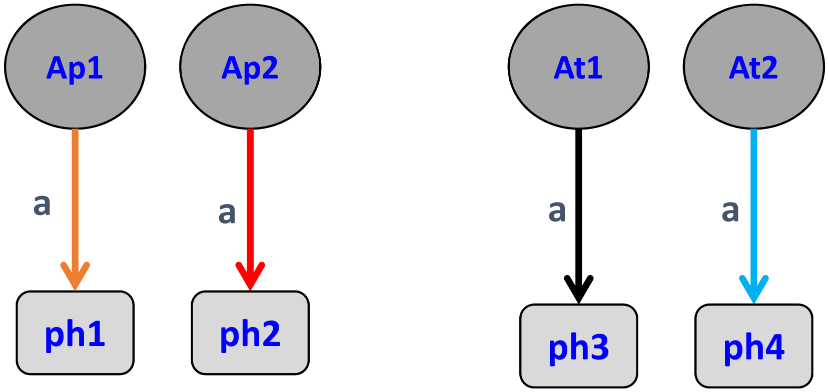
What if the degree of misspecification is even larger?

e.g.: model an ACE trait but ignore AC

V modeled as an E

$$\mathbf{V}(\hat{\Theta}) = \mathbf{I} \otimes \sigma^2_E$$

You assume there is no significant covariance between family members.



$\mathbf{V}$ modeled as an E

$$\mathbf{V}(\hat{\Theta}) = \mathbf{I} \otimes \sigma^2_E$$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

ESTIMATION?

Estimation can also be performed by Least Squares. Least squares is fast as it is a non-iterative procedure. This aspect is very important in the context of GWAS given that you perform millions of association tests.

Under IID conditions (data are identically and independently distributed) the estimates obtained by using ML and the Unweighted Least Squares function are identical.

Unweighted Least Squares

$$\mathbf{b}_{\mathrm{ULS}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$\mathrm{var}(\mathbf{b})_{\mathrm{ULS}} = (\mathbf{X'X})^{-1}\ \hat{\sigma}^{2}{}_{\mathrm{E}}$$

$$\mathbf{V}(\widehat{\Theta}) = \hat{\sigma}^{2}{}_{\mathrm{E}}\mathbf{I}$$

Unweighted Least Squares

$$\mathbf{b}_{\mathrm{ULS}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$\mathrm{var}(\mathbf{b})_{\mathrm{ULS}} = (\mathbf{X'X})^{-1}\, \hat{\sigma}^2{}_{\mathrm{E}}$$

$$\mathbf{V}(\hat{\Theta}) = \hat{\sigma}^2{}_{\mathrm{E}}\mathbf{I}$$

Misspecification:

standard errors are likely underestimated (too small).

What to do? The least squares estimates of beta are still good, because least squares is still consistent even if the residuals are not independently, identically and normally distributed. The ULS estimates are consistent in the sense that as N is larger, the estimates tend to their true population values.

However, the standard errors are incorrect; they are too small (the type I error rate will be inflated). A sandwich correction is equally applicable here, in order to arrive at correct standard errors.

Unweighted Least Squares

$$\mathbf{b}_{\mathrm{ULS}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$\mathrm{var}(\mathbf{b})_{\mathrm{ULS}} = (\mathbf{X'X})^{-1}\ \hat{\sigma}^2{}_{E}$$

SANDWICH correction

$$\mathbf{V}(\hat{\Theta}) = \hat{\sigma}^2{}_{E}\mathbf{I}$$

$$\mathrm{var}\left(\hat{\mathbf{b}}_{\mathrm{R-ULS}}\right) = \left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{t}}\left(\mathbf{y}-\mathbf{Xb}\right)\left(\mathbf{y}-\mathbf{Xb}\right)^{\mathrm{t}}\mathbf{X}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}$$

Least squares – implemented in plink
ML – implemented by LMM (fastLMM)

LEAST SQUARES: - non-iterative, very fast;

- correct standard errors;
- **E** model for the covariance matrix

misspecification

ML : - iterative;

- fast;
- **AE** model for the covariance matrix

misspecification for ACE traits

SIMULATION: ACE trait
Aim: compare the power of ULS & ML

N=4000

σ^2_A= .6

σ^2_c=.2

σ^2_E=1-σ^2_C-σ^2_A

maf=.5	# minor allele frequency
effsize=1%	# SNP effect size

alpha=10^-7

SIMULATION: ACE trait
Aim: compare the power of ULS & ML

Family structure		ML standard ACE model (true)	Sandwich corrected ML false: AE model	Sandwich corrected ML false: CE model	Sandwich corrected ULS false: E model
4 sibs	mean(b1)	-0.142	-0.142	-0.142	-0.142
	mean (st.err.)	0.023	0.024	0.024	0.031
	mean (t-value)	-6.03	-5.98	-5.98	-4.65
	power	**75.7**	**74.2**	**74.2**	**25.1**

The loss in power incurred by the sandwich corrected ULS procedure is large. The degree of model misspecification is extreme in this case.

Power of the sandwich corrected ML procedure with an AE or a CE model for the familial covariance matrix is appreciably larger;

Note also that the ML procedure with a misspecified model preserves the power to the level afforded by full correct modeling (ML with an ACE model for the V, the true model used for simulation).

Specifying a simpler model for the familial covariance matrix - an AE or a CE model (equivalent to using an exchangeable model for the working correlation matrix in the R-package GEE) in combination with a sandwich correction suffices to maintain the power almost equal to that conferred by correct full modelling (which can be computationally intensive).

ML sandwich correction has not yet been implemented by any of the current software for GWAS that can handle family data.

CONCLUSIONS

- Model the familial covariance matrix **V** as an **AE** or a **CE** & use a **SANDWICH**

- USE the **GEE** library in R because it \
  --is **fast** \
  --comes with a **sandwich** correction \
  --covers the **generalized linear model** \
  --families supported: **Gaussian , binomial, poisson, Gamma, quasi**\
  --can be accessed from **PLINK \**
  --see EXAMPLE : http://cameliaminica.nl/scripts.php

Families might be highly variable in composition, hence full correct modeling of the conditional covariance matrix can be complicated. Hence choosing a simpler model for the background (AE or CE, but not an E!) and using a sandwich correction is an efficient and computationally feasible strategy. Note that generalized estimating equations (gee) procedure, as implemented in R has four useful aspects.

- a choice of models for the familial covariance matrix, including the independence model (equivalent to the ULS with a sandwich procedure) and exchangeable model (equivalent to the CE model in linear mixed modeling).
- it includes sandwich corrected standard errors of the parameters b (robustness to misspecification of the familial covariance matrix);
- covers generalized linear model (distributions supported in gee : Gaussian , binomial (binary traits), poisson (counts), Gamma, and quasi)
- can be accessed from Plink and so provides a computationally feasible strategy for running genome-wide scans in family data.

USEFUL SOFTWARE:

PLINK1.7 + R-GEE+sandwich (also in PLINK1.9) :

http://pngu.mgh.Harvard.edu/~purcell/plink/rfunc.shtml

https://www.cog-genomics.org/plink2/

see EXAMPLE GEE: http://cameliaminica.nl/scripts.php

MERLIN and MERLIN-offline:

http://genepi.qimr.edu.au/staff/sarahMe/merlin-offline.html

GCTA-MLM-LOCO:

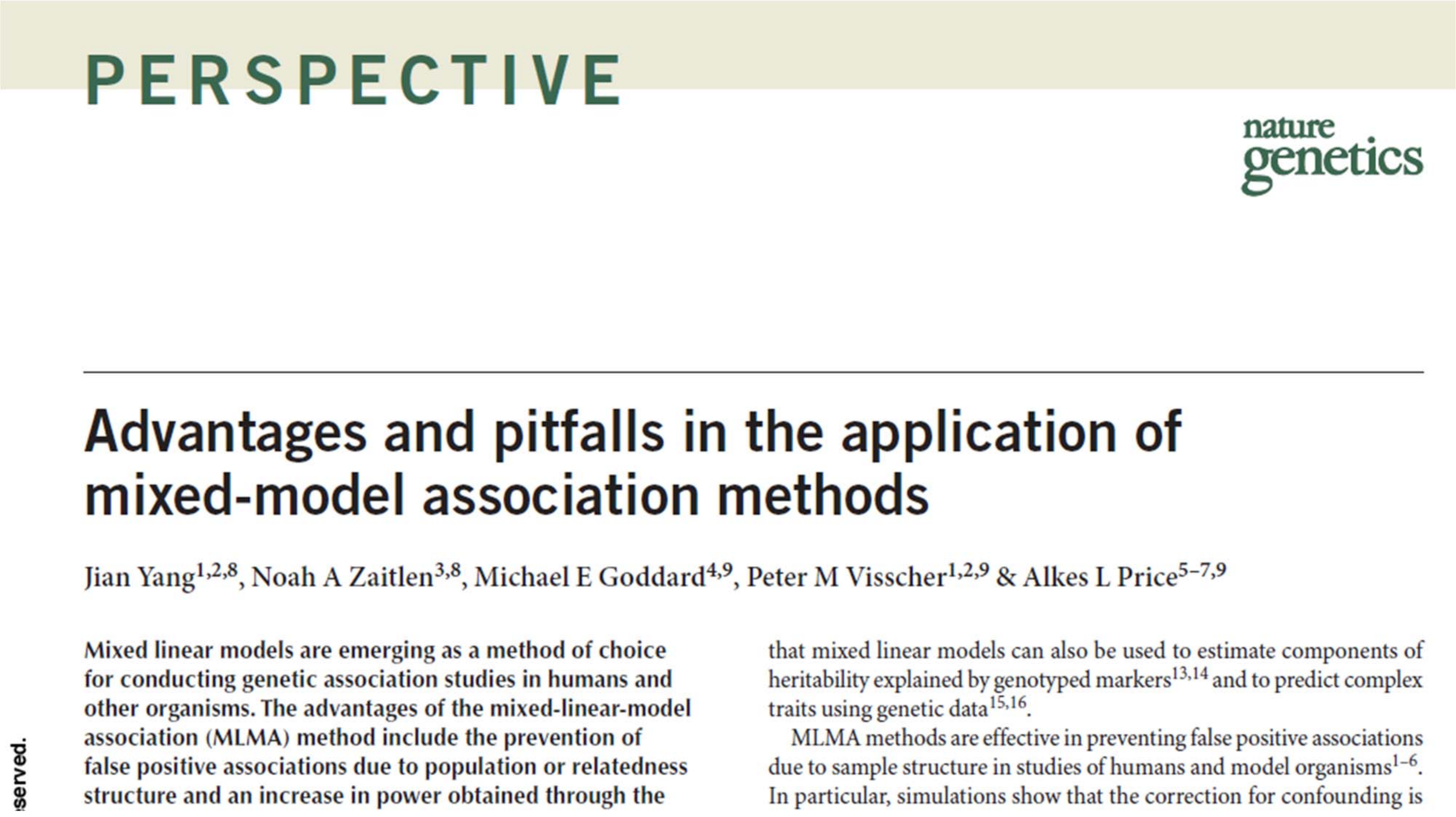
http://www.complextraitgenomics.com/software/gcta/mlmassoc.html

FAST-LMM:https://github.com/MicrosoftGenomics/FaST-LMM

The Gee+Plink approach requires best-guess genotypes (but it comes with a sandwich correction and it can handle continuous, dichotomous traits, counts).
Best-guess data are also required by GCTA and fast-lmm.
Merlin handles dosages.

SNPs should be well imputed.

2014



PERSPECTIVE

nature genetics

Advantages and pitfalls in the application of mixed-model association methods

Jian Yang[1,2,8], Noah A Zaitlen[3,8], Michael E Goddard[4,9], Peter M Visscher[1,2,9] & Alkes L Price[5–7,9]

Mixed linear models are emerging as a method of choice for conducting genetic association studies in humans and other organisms. The advantages of the mixed-linear-model association (MLMA) method include the prevention of false positive associations due to population or relatedness structure and an increase in power obtained through the

that mixed linear models can also be used to estimate components of heritability explained by genotyped markers[13,14] and to predict complex traits using genetic data[15,16].

MLMA methods are effective in preventing false positive associations due to sample structure in studies of humans and model organisms[1–6]. In particular, simulations show that the correction for confounding is