Family-based association methods ->

Observations in the study are not independent

Jenny van Dongen, Dorret Boomsma, Camelia Minica, Conor Dolan, Joshua Pritikin

This session

1. Family based association analyses: intro (*Dorret Boomsma*)

2. Genetic association tests: Plink and R
3. Apply the GWAS results (*Jenny van Dongen*)

Data collected in families

Some methods require twin / family data:

-> heritability

-> linkage

However, when looking at association, we need to adjust for clustering in family data.

Focus: family-based Genome-Wide Association Studies

These are regression based approaches, with outcomes (phenotypes) and predictors (Genetic Variants (GV), or polygenic scores, other and covariates)

Ignoring clustering in family data may lead to wrong conclusions: point estimates of effects OK, but SE too small!

Mixed models

What does ‘mixed model’ mean?

What does ‘mixed model’ mean

Mixed models contain fixed effects (parameters) and random effects

The elementary mixed model is a two-way ANOVA model:

One fixed factor, one random factor

One-way fixed effect ANOVA

Evaluate the effect of a factor with a limited number of levels.

-Level effect definition: mean score

-Obtained in the population of subjects

	FACTOR LEVELS		
	1	2	3
	Y11	**y21**	**y31**
	y12	**y22**	**y32**
	y13	**y23**	**y33**
	y14	**y24**	**y34**
	y15	**y25**	**y35**
means	**m1**	**m2**	**m3**

(the factor could represent 3 genotypes: e.g. AA, Aa and aa).

Procedure

1. randomly sample subjects from the population.

2. randomly “assign” them to the levels of the factor.

3. compute the sample means for each level.

4. “decide” if observed differences between the means are sufficient evidence for differences between the levels in the population.

The population is not observed and observed sample differences might change if the experiment is repeated **with the same factor levels** but **with a different random sample**.

How to conclude that observed differences between the means can be taken as evidence for differences in the population?

We need a yardstick to measure the size of the differences between levels.

How to conclude that differences between level means are evidence for differences in the population?

We need a yardstick to measure the size of the differences -> the variance of the scores within the levels of the factor in the population (i.e. based on the differences of the individual scores to the level means).

Larger variance increases likelihood of observing differences in the sample if the levels have no effect. The probability refers to repetitions of the experiment.

In this simple design, the **levels of the factor** remain the same over repeated experiments. The level effects are therefore **fixed effects** or parameters.

The **individual scores** change over experiments and are therefore called **random effects** (defined as deviations from the level effects).

The linear model: $y_{ij} = m + b_i + e_{ij}$

m = general constant; fixed effect

$\mathbf{b_i}$ **=** effect of factor level *i*; **fixed** effect; same levels in repeated experiments

$\mathbf{e_{ij}} = \mathbf{y_{ij} - (m + b_i)}$ **=** residual: **random** effect (different subjects in repeated experiments)

Statistical properties of $\mathbf{e_{ij}}$: mean zero and variance σ_e^2

Models such as these, where $\mathbf{e_{ij}}$ is the only random effect, are called **fixed effects models**.

A change in the design: Ss are still sampled randomly from the population, but each subject is observed at each factor level.

FACTOR LEVELS

	1	2	3
	y11	**y21**	**y31**
	y12	**y22**	**y32**
	y13	**y23**	**y33**
	y14	**y24**	**y34**
	y15	**y25**	**y35**
means	**m1**	**m2**	**m3**

(the factor cannot represent 3 genotypes for the same person (AA, Aa and aa) but if we have family members they can be observed at each level of the factor).

Row scores come from the same subject / family. In the previous design, y_{11} and y_{12} referred to different Ss, who were sampled ***independently.*** **We now have within subject (or within family) correlation or covariance among y11, y12, y13**.

The linear model now is: $\mathbf{y_{ij} = m + a_i + b_j + e_{ij}}$

m = general constant; fixed effect
$\mathbf{a_i}$ = effect of subject *i*; **random** effect, since $\mathbf{a_i}$ changes over repeated experiments; mean zero; variance σ_a^2
$\mathbf{b_j}$ = effect of factor level *j*; **fixed** effect; same levels in repeated experiments

$\mathbf{e_{ij}} = \mathbf{y_{ij} - (m + a_i + b_j)}$ residual or error; **random** effect;

Model contains fixed factor effects and **one random effect besides the residual**. Such models are called **mixed models**.

Consequences of the model

Expected variance within the level of a factor is $\sigma_a^2 + \sigma_e^2$

Covariance of observations at two different levels of the factor is σ_a^2

General representation of mixed models as matrix equation:

y = Xb + Za + e

b = [m, b1, b2, b3]'
a = [a1, a2]'

For the first two subjects:

	X				**Z**		**e**
	m	**b1**	**b2**	**b3**	**a1**	**a2**	
y11 =	**1**	**1**	**0**	**0**	**1**	**0**	e_{11}
y12 =	**1**	**0**	**1**	**0**	**1**	**0**	e_{12}
y13 =	**1**	**0**	**0**	**1**	**1**	**0**	e_{13}
y21 =	**1**	**1**	**0**	**0**	**0**	**1**	e_{21}
y22 =	**1**	**0**	**1**	**0**	**0**	**1**	e_{22}
y23 =	**1**	**0**	**0**	**1**	**0**	**1**	e_{23}

Specification of the covariance structure

Expected covariance matrix of the random effects: **V**a:

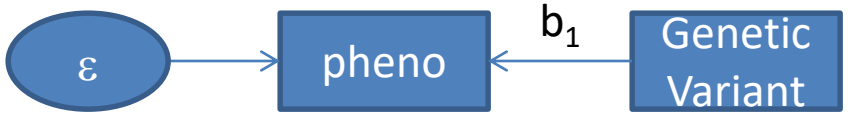
	a1	**a2**
a1	σ_a^2	**0**
a2	**0**	σ_a^2

For the residuals: $\mathbf{R} = \sigma_e^2 \mathbf{I}$ (i.e. a diagonal matrix)

Linear regression (discarding covariates) in unrelated Subjects (j=1...N)

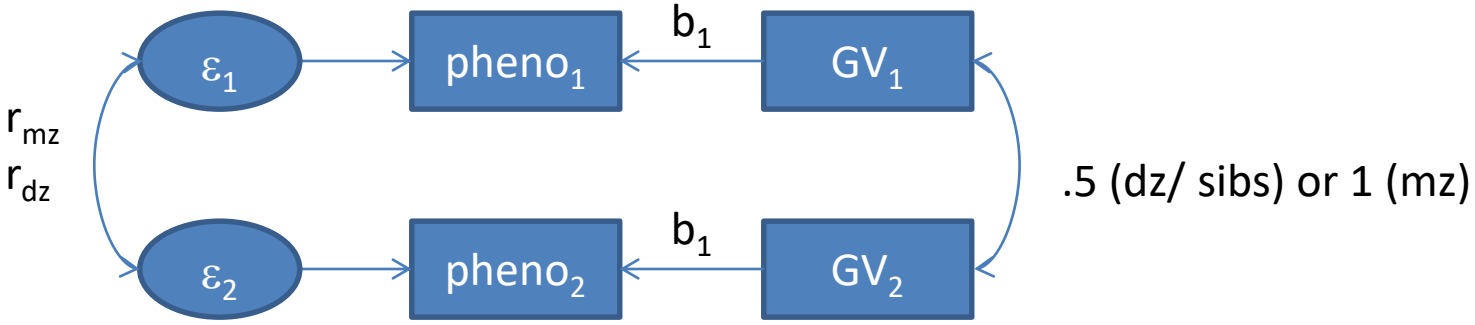
$pheno_j = b_0 + b_1 * GV_j + e_j$ $(y = XB + e)$

Wald test $b_1/s.e.(b_1)$ s.e. (b_1) from $(X^t V^{-1} X)^{-1}$, V is cov matrix of e.



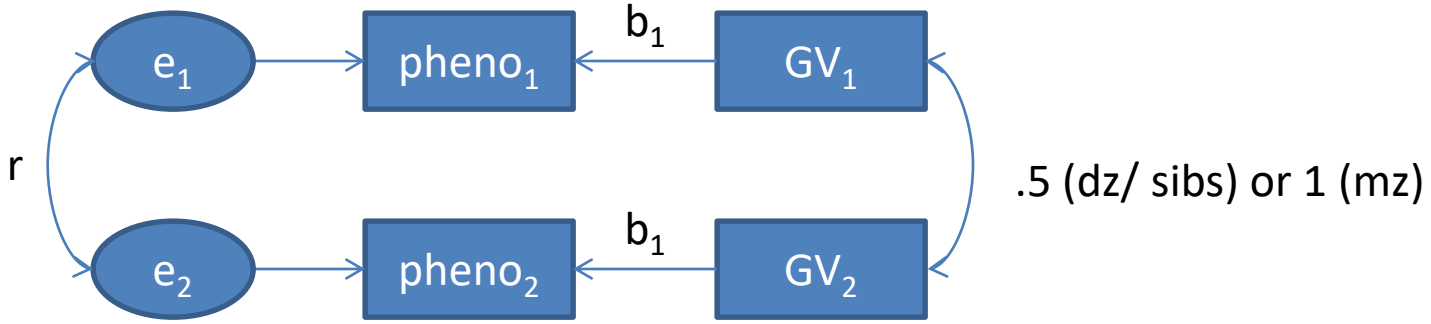
linear regression (discarding covariates) twins / sibs (i=1...Nmz (Ndz) and j=1,2)

$pheno_{ij} = b_0 + b_1 * GV_{ij} + e_{ij}$ $(y = XB + e)$



linear regression (discarding covariates) in twins / sibs

$pheno_{ij} = b_0 + b_1 * GV_{ij} + e_{ij}$ (i=1...Nmz (Ndz) and j=1,2)



Analysis options:

A. ignore relatedness

B. model correlated background

C. discard 1 twin member (e.g., occasionally: drop 1 MZ twin)

D. GEE regression (GEE = Generalized Estimating Equations) -> prac

options: A~~. ignore relatedness~~
B. model correlated residual (background)
~~C. discard 1 twin member~~
D. GEE regression

A. **BAD** – results in downward bias in s.e. (b1) and increase in type I error rate (false positives!)
B. Good, linear mixed modeling or OpenMx
C. **BAD** – loss of power
D. Good, corrects s.e. (b1) for correlated residuals

Computational Burden:

B. 1. Genetic covariance structure modeling (ACE / ADE) in OpenMx or linear mixed modeling (SPSS, R: nlme, R: lmer) – heavy, unwieldy
B. 2. Based on genetic relatedness matrix OK: GCTA, Fast-LLM (any pedigree structure)
D.1. GEE (Generalized Estimating Equations) regression – light, simple OK in the case of nuclear family data (what about extended pedigrees?).

Both GEE and mixed model are suitable when independent errors' assumption is violated.
GEE takes into account the within-cluster correlations by using an empirical covariance matrix (*sandwich*). It can really only account for one source of clustering at a time. In a GEE we cannot put any structure the correlation pattern.
A mixed model accounts for correlated outcomes by using random effects for each cluster variable. So mixed models are more versatile .

Benefits of family data (in genetic association studies)

Control for factors that can spuriously influence association tests (e.g. population stratification).

Base estimates of association effects on within family tests.

QC: Can test for Mendelian transmission errors.

Can obtain estimates of transmitted and un-transmitted PRS (if family design involves parents and offspring).

Can estimate heritability from pedigree (check on phenotype data).

May be easier to recruit large numbers by targeting families.

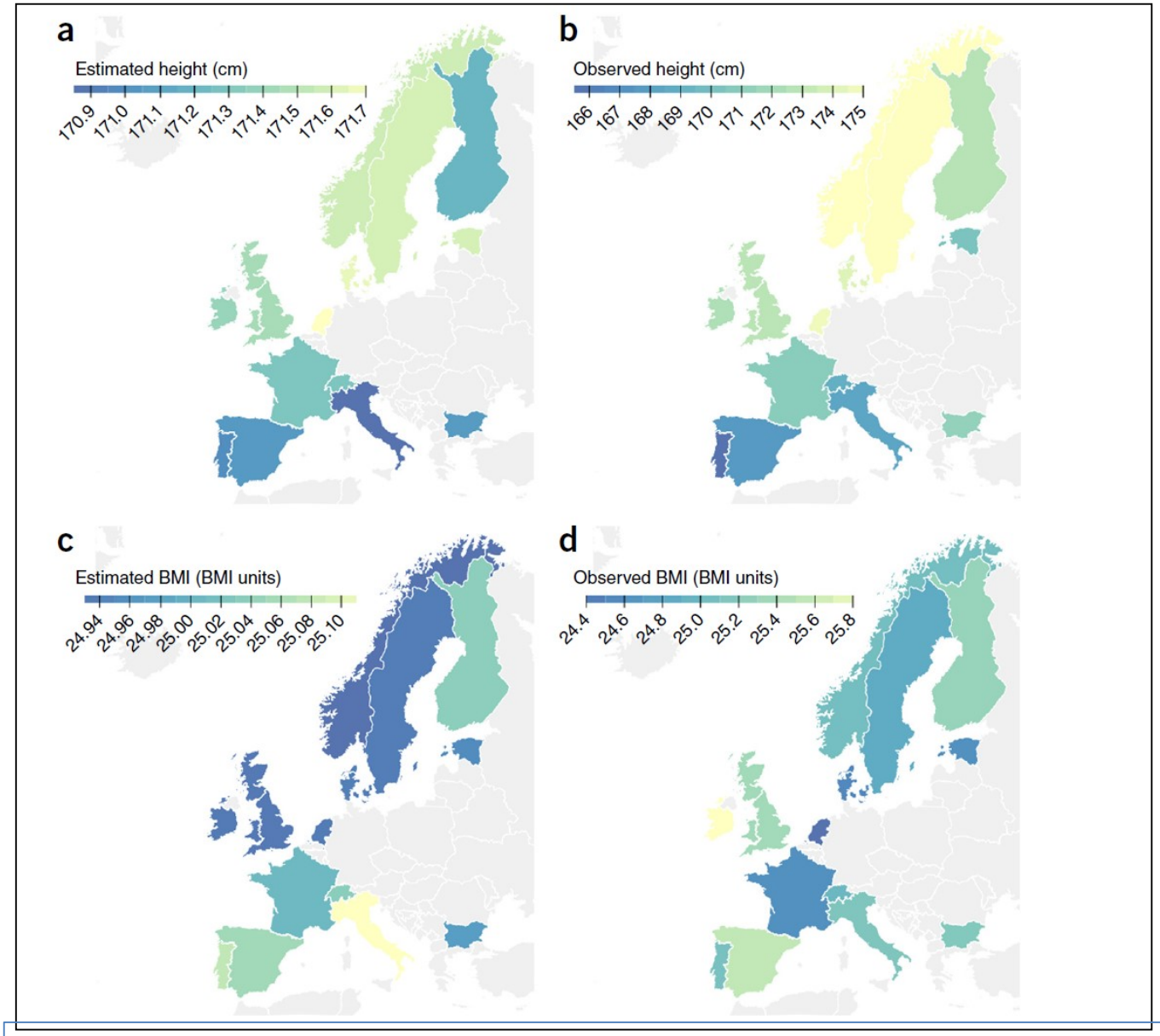
nature genetics

Within family tests

Population genetic differentiation of height and body mass index across Europe

Matthew R Robinson[1], Gibran Hemani[1], Carolina Medina-Gomez[2], Massimo Mezzavilla[3,4], Tonu Esko[5–8], Konstantin Shakhbazov[1], Joseph E Powell[1,9], Anna Vinkhuyzen[1], Sonja I Berndt[10], Stefan Gustafsson[11], Anne E Justice[12], Bratati Kahali[13,14], Adam E Locke[15], Tune H Pers[6–8,16], Sailaja Vedantam[6,7], Andrew R Wood[17], Wouter van Rheenen[18], Ole A Andreassen[19], Paolo Gasparini[3,4], Andres Metspalu[5], Leonard H van den Berg[18], Jan H Veldink[18], Fernando Rivadeneira[2], Thomas M Werge[20–22], Goncalo R Abecasis[15], Dorret I Boomsma[23–25], Daniel I Chasman[8,26], Eco J C de Geus[23–25], Timothy M Frayling[17], Joel N Hirschhorn[5–8], Jouke Jan Hottenga[23–25], Erik Ingelsson[11,27], Ruth J F Loos[28–31], Patrik K E Magnusson[32], Nicholas G Martin[33], Grant W Montgomery[33], Kari E North[13,14,34], Nancy L Pedersen[32], Timothy D Spector[35], Elizabeth K Speliotes[15], Michael E Goddard[36,37], Jian Yang[1,9] & Peter M Visscher[1,9]

GWAS meta-analyses for height / BMI in a Europeans (~250,000 Ss for height and ~350,000 for BMI). **We re-estimated the effects of each SNP in a within-family design, which is unbiased by population stratification, and used these effect sizes to create a genetic predictor for both phenotypes** (also termed ‘polygenic score’).

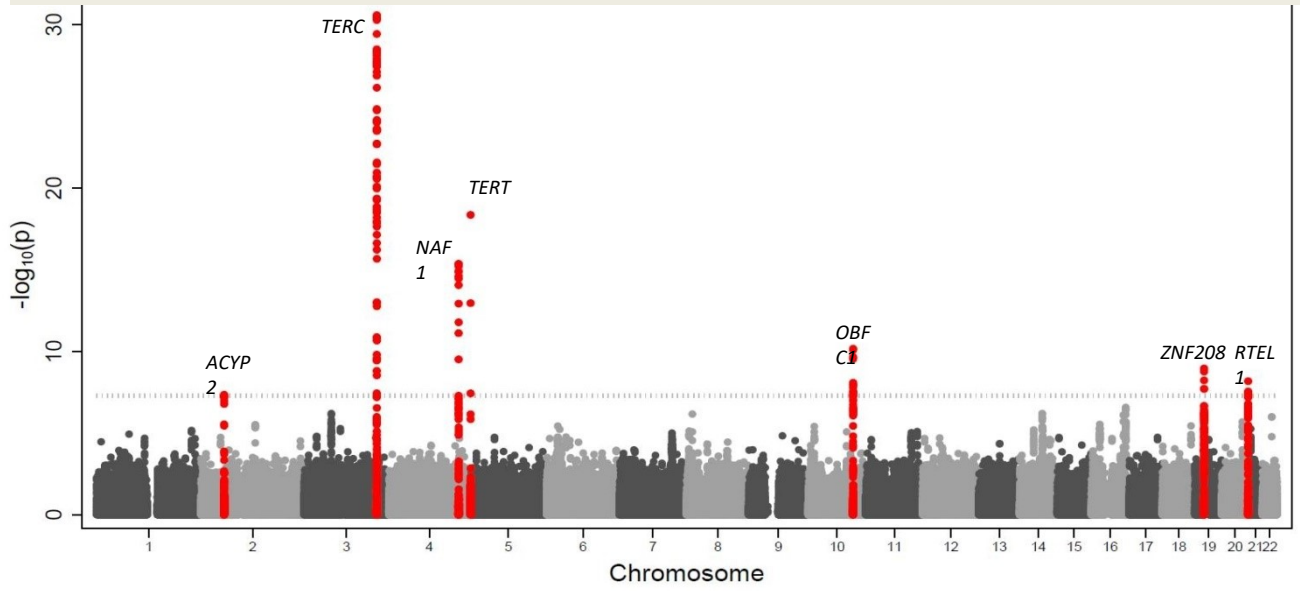


Predicted genetic means (**a**,**c**) and observed means (**b**,**d**) for height and BMI for 14 Eu nations. From recently published data, we estimated national differences in mean height and BMI, with a European average height of 171.1 cm and an average BMI of 25.0 for males.

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
Author Manuscript
Europe PMC Funders

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

Other considerations

Does not control cryptic relatedness.

Family sizes differ (not everyone can participate with their family).

May decrease statistical power.

Power calculation (see R script (provided by Conor) in faculty folder)

Suppose you have N unrelated Ss and you want to calculate power?
Simple: use Gpower, R libraries (pow), or dedicated (genetics) software

Suppose you have N_{mz} and N_{dz} twin pairs and N unrelated Ss and want to calculate power?
Simple: calculate effective sample size and use standard software

N_{mz} pairs is effectively $N_{1mz} = (N_{mz}*2) / (1 + r_{mz})$
N_{dz} pairs is effectively $N_{1dz} = (N_{dz}*2) / (1 + r_{dz})$
N unrelateds $N=N_{1u}$

r_{mz} and r_{dz} are phenotypic (intraclass) correlation coefficients.

Total effective sample size $N = N_{1u} + N_{1mz} + N_{1dz}$.

General equation is NE = (K*M) / (1 + (M-1)*r)

	Applied to N_{mz} pairs
NE = effective sample size	
K = number of clusters	N_{mz} pairs
M = number of members per cluster	M=2 (for pairs!)
r = intra-class correlation	rmz

Suppose we have MZ pairs, with and without siblings and
DZ pairs with and without siblings

Rough and ready: suppose we have

300 MZ + 0 sibs, i.e., 600 individuals
200 MZ + 1 sibs, i.e., 400 + 200 = 600 individuals
150 MZ + 2 sibs, i.e., 300 + 300 = 600 individuals

Calculate NE for each using the intraclass correlation (average phenotypic relatedness)

rmz = .5
rfs (full sib) = .25

300 MZ + 0 sibs, i.e., 600 individuals

1	.5		
.5	1	r=.5	NE=300*2/(1+.5) = ~400

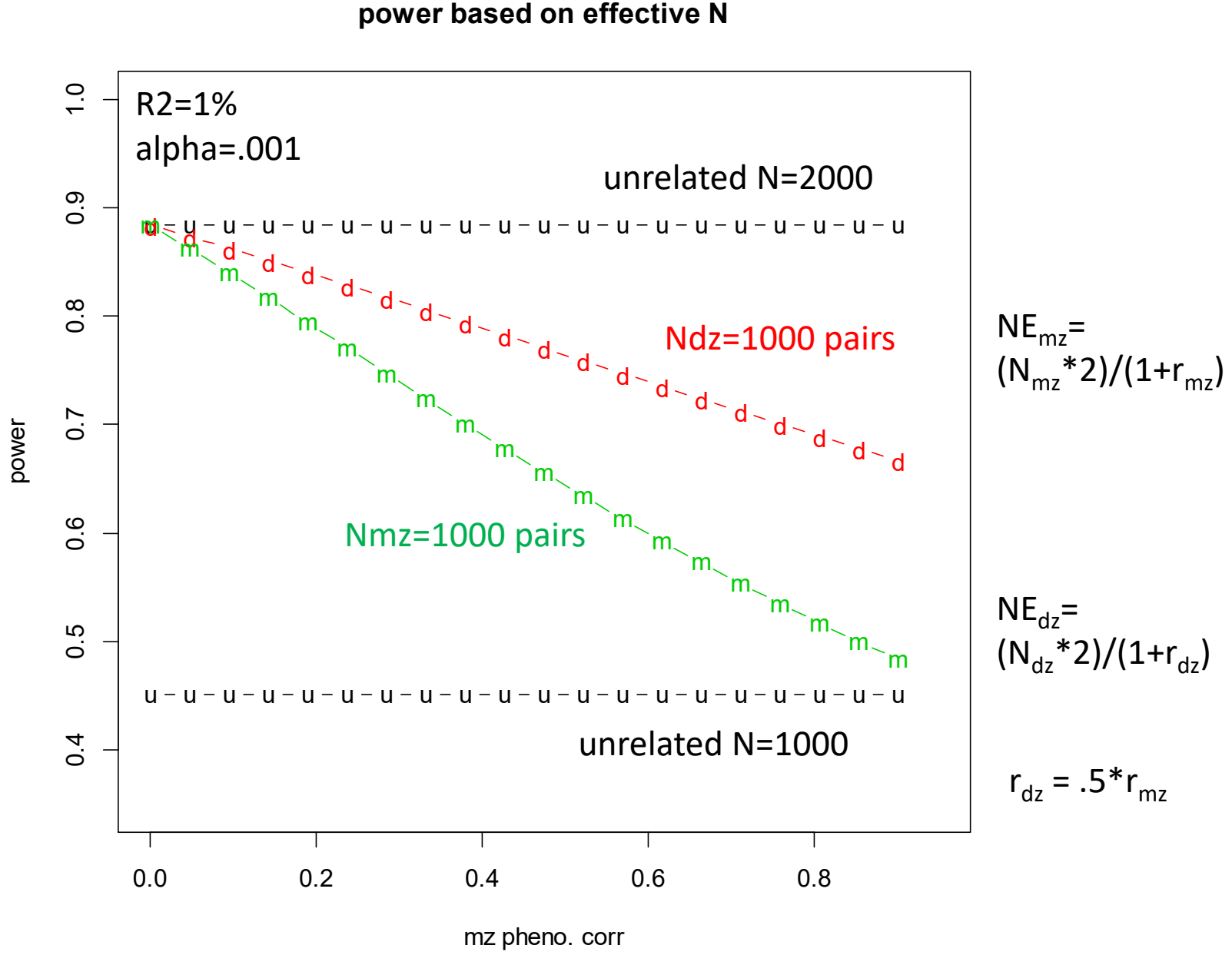
200 MZ + 1 sibs, i.e., 400 + 200 = 600 individuals

1	.5	.25		
.5	1	.25		
.25	.25	1	r =~.333	NE=(200*3)/(1+.333) = ~450

150 MZ + 2 sibs, i.e., 300 + 300 = 600 individuals

1	.5	.25	.25		
.5	1	.25	.25		
.25	.25	1	.25		
.25	.25	.25	.25	r= ~.29	NE=(150*4)/(1+.29) = ~465

Total sample size in individuals = 600+600+600 = 1800
Total effective sample size = 400 +450 + 465 =1315



$NE_{mz} = (N_{mz}*2)/(1+r_{mz})$

$NE_{dz} = (N_{dz}*2)/(1+r_{dz})$

$r_{dz} = .5*r_{mz}$