From correlation coefficients to variance components

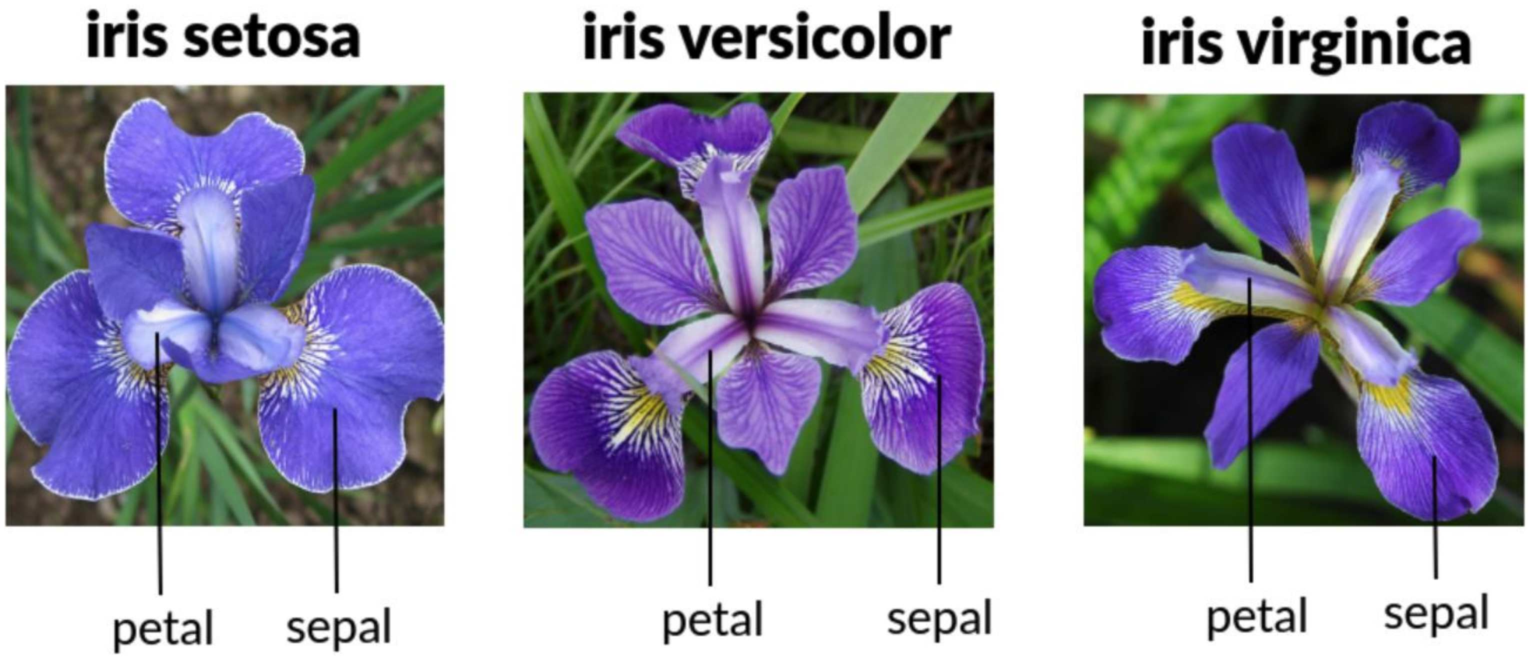
Baptiste Couvy-Duchesne

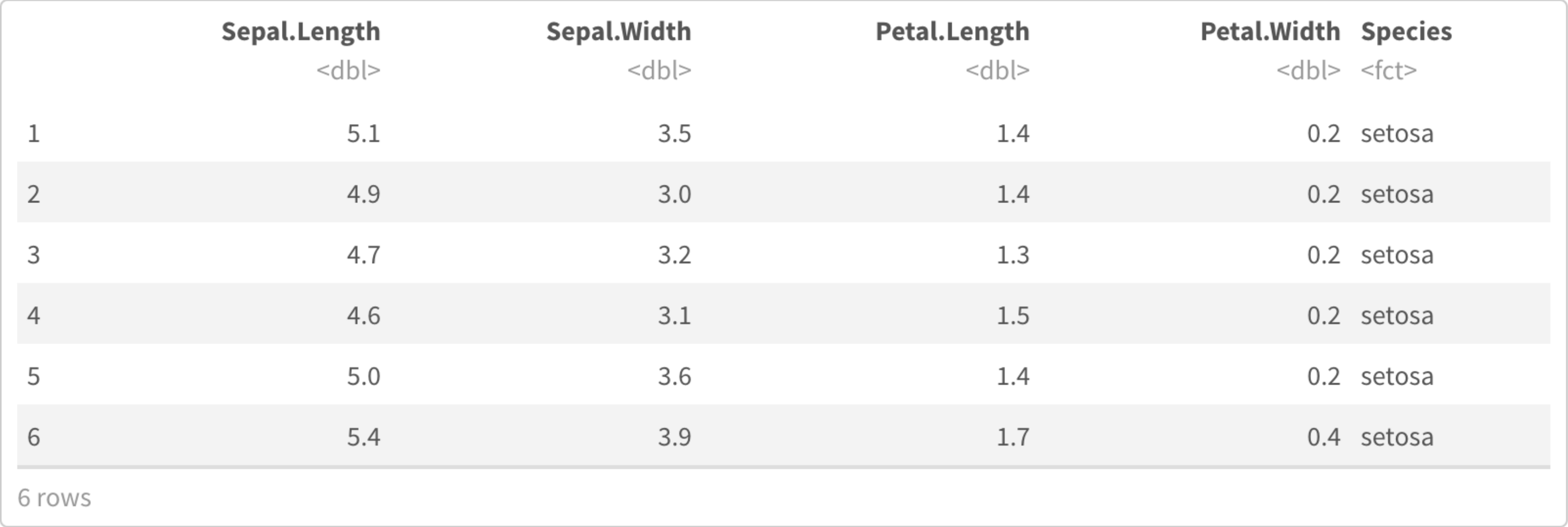
baptiste.couvyduchesne@uq.edu.au

baptiste.couvy@icm-institute.org

https://github.com/baptisteCD

@BaptisteCouvy





	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

6 rows

https://en.wikipedia.org/wiki/Iris_flower_data_set

Correlation and linear model

- **Y** = **X** b + **e**

- Y for example BMI
- X age
- b association
- Maximum Likelihood estimation

$$\begin{bmatrix} BMI_1 \\ \ldots \\ BMI_N \end{bmatrix} = \begin{bmatrix} Age_1 \\ \ldots \\ Age_N \end{bmatrix} . b + \begin{bmatrix} e_1 \\ \ldots \\ e_N \end{bmatrix}$$

$$b = \frac{cov(X,Y)}{var(X)}$$

$$r = \frac{cov(X,Y)}{sd(X)sd(Y)}$$

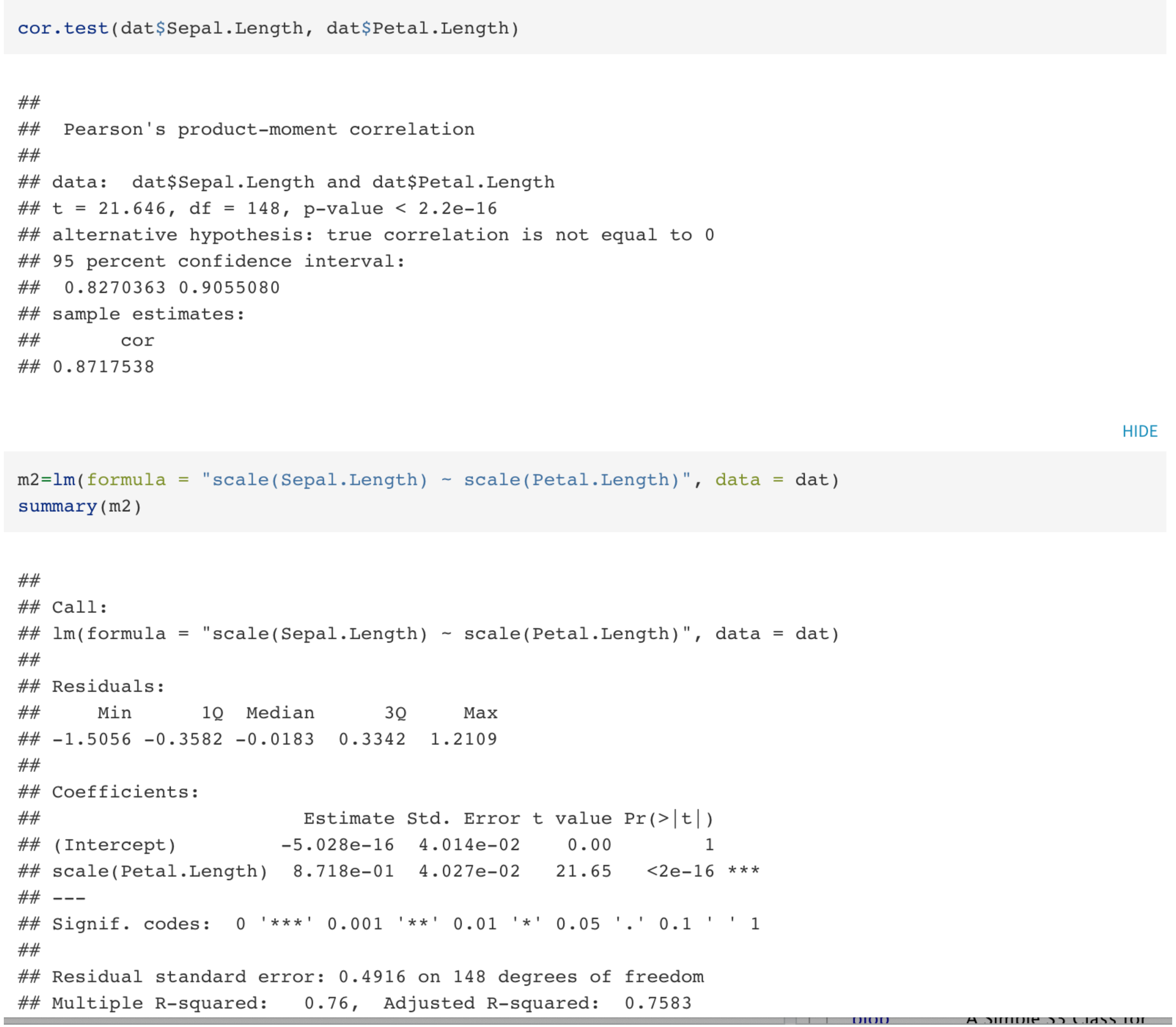
$$\boldsymbol{r} = \frac{\boldsymbol{b} \, . \, \boldsymbol{sd}(\boldsymbol{Y})}{\boldsymbol{sd}(\boldsymbol{X})}$$

- https://en.wikipedia.org/wiki/Linear_regression
- https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

```r
dat=datasets::iris

m1=lm(formula = "Sepal.Length ~ Petal.Length", data = dat)
summary(m1)
```

```
##
## Call:
## lm(formula = "Sepal.Length ~ Petal.Length", data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24675 -0.29657 -0.01515  0.27676  1.00269
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.30660    0.07839   54.94   <2e-16 ***
## Petal.Length  0.40892    0.01889   21.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4071 on 148 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7583
## F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```



```r
cor.test(dat$Sepal.Length, dat$Petal.Length)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dat$Sepal.Length and dat$Petal.Length
## t = 21.646, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8270363 0.9055080
## sample estimates:
##       cor
## 0.8717538
```

```r
m2=lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length)", data = dat)
summary(m2)
```

```
##
## Call:
## lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length)", data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5056 -0.3582 -0.0183  0.3342  1.2109
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -5.028e-16  4.014e-02    0.00        1
## scale(Petal.Length)  8.718e-01  4.027e-02   21.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4916 on 148 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7583
```

Multiple regression

- **Y** = **X1** b1 **+ X2** b2 + **e**
- **Y** = **X b** + **e**

- Y for example BMI
- X age, sex, …
- b = (b1, b2) marginal associations

- Partial correlations, multiple correlations

- Global measure of fit e.g. R2

$$\begin{bmatrix} BMI_1 \\ \dots \\ BMI_N \end{bmatrix} = \begin{bmatrix} Age_1 & Sex_1 & PC_1 \\ \dots & \dots & \dots \\ Age_N & Sex_N & PC_N \end{bmatrix} . \begin{bmatrix} b_{Age} & b_{Sex} & b_{PC} \end{bmatrix} + e$$

```r
# LM with covariates
m3=lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length) + scale(Petal.Width) + factor(Species)", d
summary(m3)
```

```
##
## Call:
## lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length) + scale(Petal.Width) + factor(Species)",

##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90860 -0.27883 -0.00255  0.27896  1.24517
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.493845   0.205568   7.267 2.09e-11 ***
## scale(Petal.Length)        1.931325   0.158419  12.191  < 2e-16 ***
## scale(Petal.Width)        -0.005519   0.143838  -0.038    0.969
## factor(Species)versicolor -1.930234   0.248418  -7.770 1.32e-12 ***
## factor(Species)virginica  -2.551302   0.367150  -6.949 1.16e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4096 on 145 degrees of freedom
## Multiple R-squared:  0.8367, Adjusted R-squared:  0.8322
## F-statistic: 185.8 on 4 and 145 DF,  p-value: < 2.2e-16
```

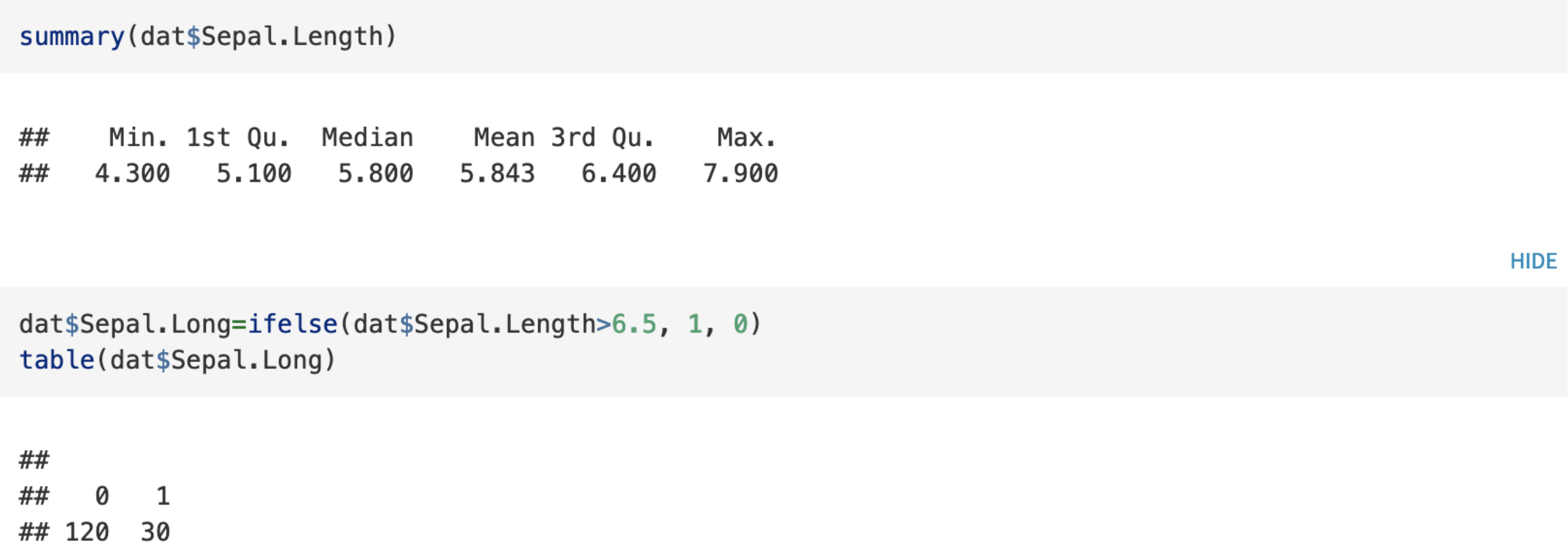
Generalised linear model – for binary or count variabless

- **Y** = **X b** + **e**
- Y : Disease status, number of symptoms
- X age, sex
- b = (b1, b2) associations

Maximum Likelihood estimation

Gaussian, Binomial, poisson, gamma, inverse gaussian distributions

$$\begin{bmatrix} SCZ_1 \\ \dots \\ SCZ_N \end{bmatrix} = \begin{bmatrix} Age_1 & Sex_1 & PC_1 \\ \dots & \dots & \dots \\ Age_N & Sex_N & PC_N \end{bmatrix} . \begin{bmatrix} b_{Age} & b_{Sex} & b_{PC} \end{bmatrix} + e$$



```r
summary(dat$Sepal.Length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.300   5.100   5.800   5.843   6.400   7.900 
```

```r
dat$Sepal.Long=ifelse(dat$Sepal.Length>6.5, 1, 0)
table(dat$Sepal.Long)
```

```
## 
##   0   1 
## 120  30 
```

```r
m4=glm(formula = "Sepal.Long ~ scale(Petal.Length) + factor(Species)", data = dat, family = binomial()
summary(m4)
```

```
##
## Call:
## glm(formula = "Sepal.Long ~ scale(Petal.Length) + factor(Species)",
##     family = binomial(), data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.92687  -0.39418  -0.00012  -0.00005   2.16753
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -9.742   1459.343  -0.007    0.995
## scale(Petal.Length)          7.719      1.765   4.374 1.22e-05 ***
## factor(Species)versicolor    4.911   1459.344   0.003    0.997
## factor(Species)virginica     1.626   1459.346   0.001    0.999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 150.121  on 149  degrees of freedom
## Residual deviance:  72.838  on 146  degrees of freedom
## AIC: 80.838
##
## Number of Fisher Scoring iterations: 18
```

R2 – variance explained

- **Y** = **X** b + **e**

Var(Y) = b2 var(X) + var(e)

R2 = b2 var(X) / var(Y)

coefficient of determination

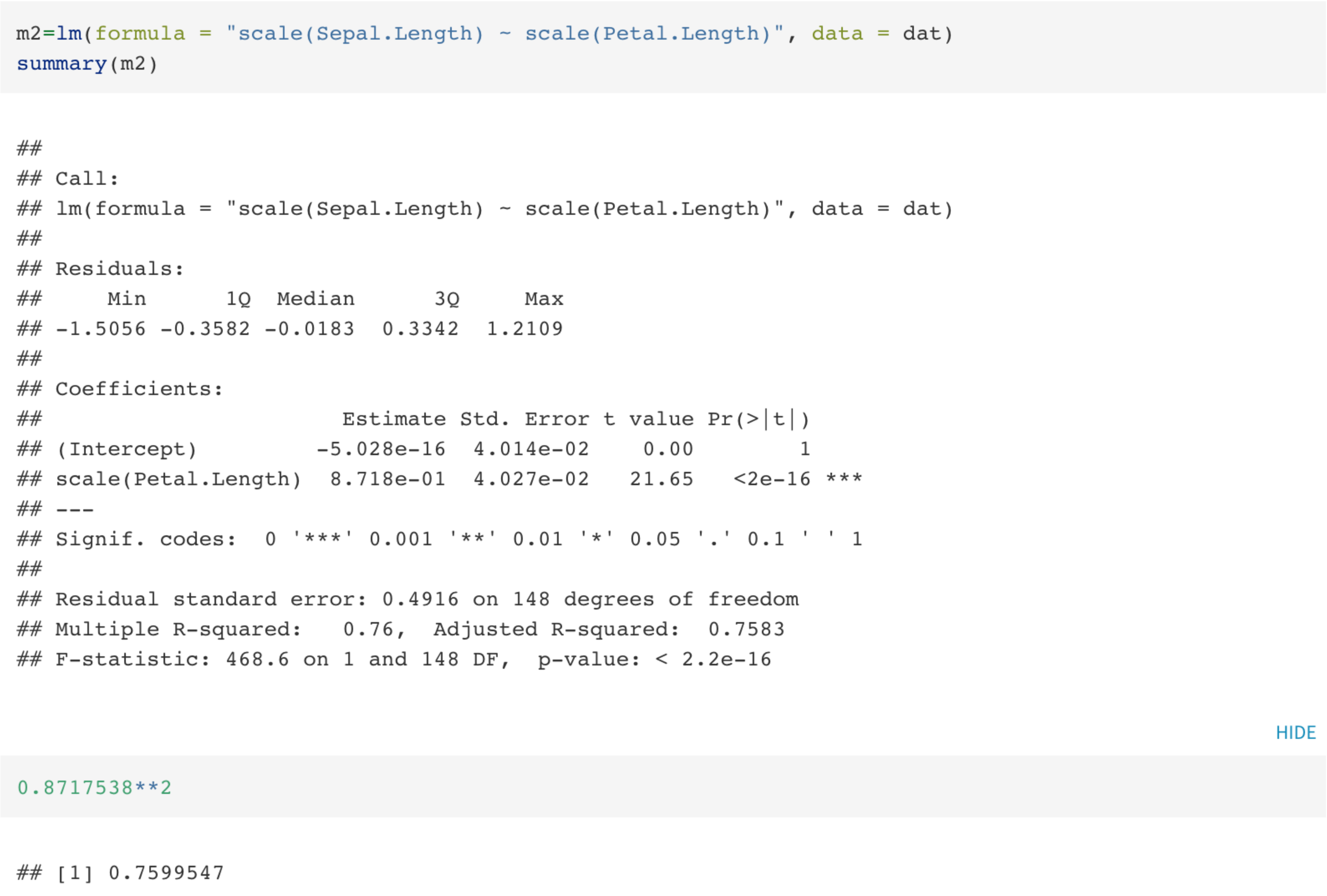
- **Y** = b1 **X1 +** b2**X2** + **e**

R2= 1 – var(**e**) / var(**Y**)

- R2 = Combined association of one or several variables
- pseudo R2 – for logistic / Poisson regression
- R2 expressed in % variance of Y

https://en.wikipedia.org/wiki/Coefficient_of_determination

https://en.wikipedia.org/wiki/Logistic_regression

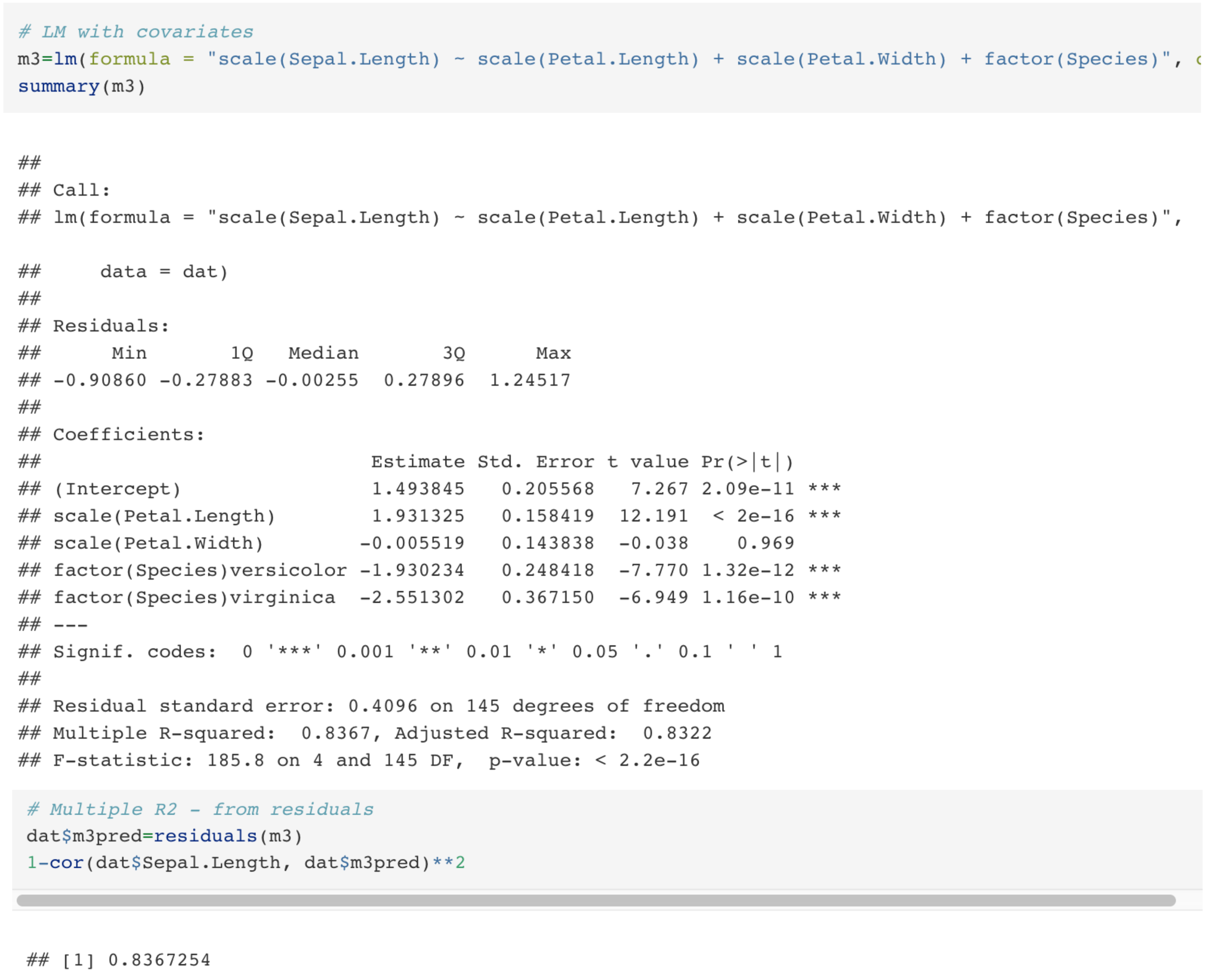


```r
m2=lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length)", data = dat)
summary(m2)
```

```
## 
## Call:
## lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length)", data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.5056 -0.3582 -0.0183  0.3342  1.2109 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        -5.028e-16  4.014e-02    0.00        1    
## scale(Petal.Length)  8.718e-01  4.027e-02   21.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4916 on 148 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7583 
## F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

```r
0.8717538**2
```

```
## [1] 0.7599547
```



```
# LM with covariates
m3=lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length) + scale(Petal.Width) + factor(Species)", 
summary(m3)
```

```
##
## Call:
## lm(formula = "scale(Sepal.Length) ~ scale(Petal.Length) + scale(Petal.Width) + factor(Species)",

##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90860 -0.27883 -0.00255  0.27896  1.24517
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.493845   0.205568   7.267 2.09e-11 ***
## scale(Petal.Length)        1.931325   0.158419  12.191  < 2e-16 ***
## scale(Petal.Width)        -0.005519   0.143838  -0.038    0.969
## factor(Species)versicolor -1.930234   0.248418  -7.770 1.32e-12 ***
## factor(Species)virginica  -2.551302   0.367150  -6.949 1.16e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4096 on 145 degrees of freedom
## Multiple R-squared:  0.8367, Adjusted R-squared:  0.8322
## F-statistic: 185.8 on 4 and 145 DF,  p-value: < 2.2e-16
```

```
# Multiple R2 - from residuals
dat$m3pred=residuals(m3)
1-cor(dat$Sepal.Length, dat$m3pred)**2
```

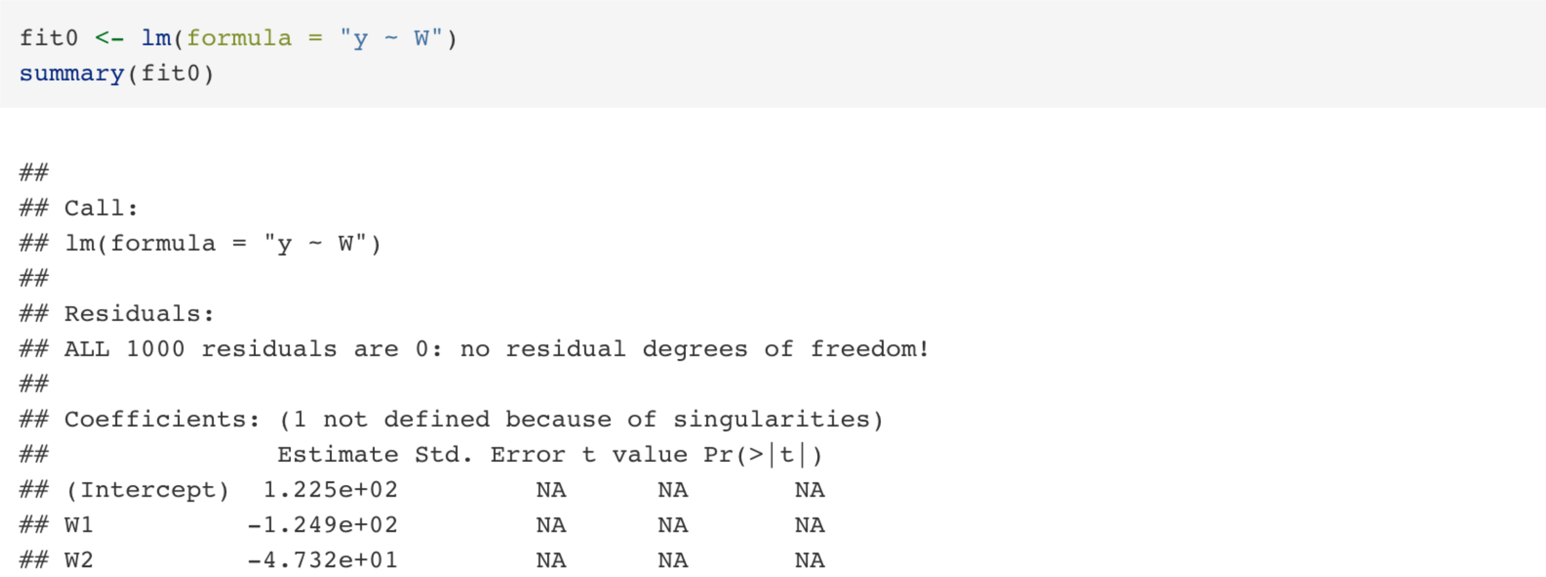
```
## [1] 0.8367254
```

Multiple regression with many more SNPs

Y = X b + e

X full matrix of SNPs; N individuals ~450K, m SNPs ~16 Millions

$$\begin{bmatrix} BMI_1 \\ \dots \\ BMI_N \end{bmatrix} = \begin{bmatrix} rs001_1 & rs\ \dots_1 & rs9999_1 \\ \dots & \dots & \dots \\ rs001_N & rs\ \dots_N & rs9999_N \end{bmatrix} . \begin{bmatrix} b_{rs0011} & b_{rs\dots} & b_{rs9999} \end{bmatrix} + \begin{bmatrix} e_1 \\ \dots \\ e_N \end{bmatrix}$$



```r
fit0 <- lm(formula = "y ~ W")
summary(fit0)
```

```
##
## Call:
## lm(formula = "y ~ W")
##
## Residuals:
## ALL 1000 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.225e+02         NA      NA       NA
## W1          -1.249e+02         NA      NA       NA
## W2          -4.732e+01         NA      NA       NA
```

Random effect model – can handle high dimension

- $\mathbf{Y} = \mathbf{X}\ \mathbf{b} + \mathbf{e}$
- $b \sim N(0,\ I\ .\ sG2\ /\ m\ )$
- $Var(Y) = m\ SG2/m + V(e)$

- X full matrix of SNPs - normalised
- b marginal SNP association coefficients

- Constraint / hypothesis : b normally distributed (unimodal)

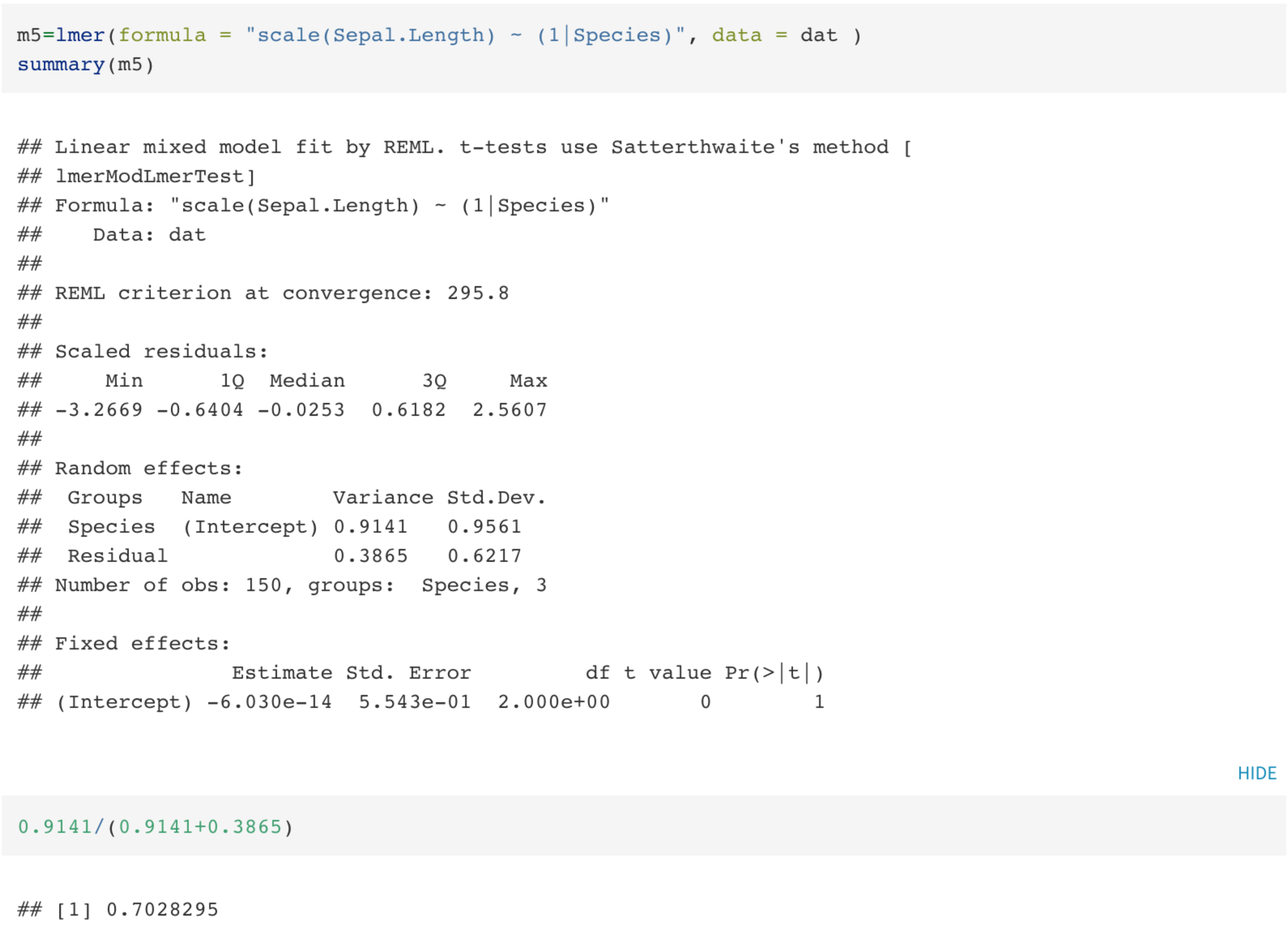
- R2 of random effect only requires estimating sG2 and var(Y)

$R2 = sG2\ /\ var(Y)$

```r
library(lmerTest)
library(lme4)

m5base=lm(formula = "scale(Sepal.Length) ~ factor(Species)", data = dat )
summary(m5base)
```

```
## 
## Call:
## lm(formula = "scale(Sepal.Length) ~ factor(Species)", data = dat)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.03848 -0.39671 -0.00725  0.37678  1.58441 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                -1.01119    0.08792 -11.501  < 2e-16 ***
## factor(Species)versicolor   1.12310    0.12434   9.033 8.77e-16 ***
## factor(Species)virginica    1.91048    0.12434  15.366  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6217 on 147 degrees of freedom
## Multiple R-squared:  0.6187, Adjusted R-squared:  0.6135 
## F-statistic: 119.3 on 2 and 147 DF,  p-value: < 2.2e-16
```



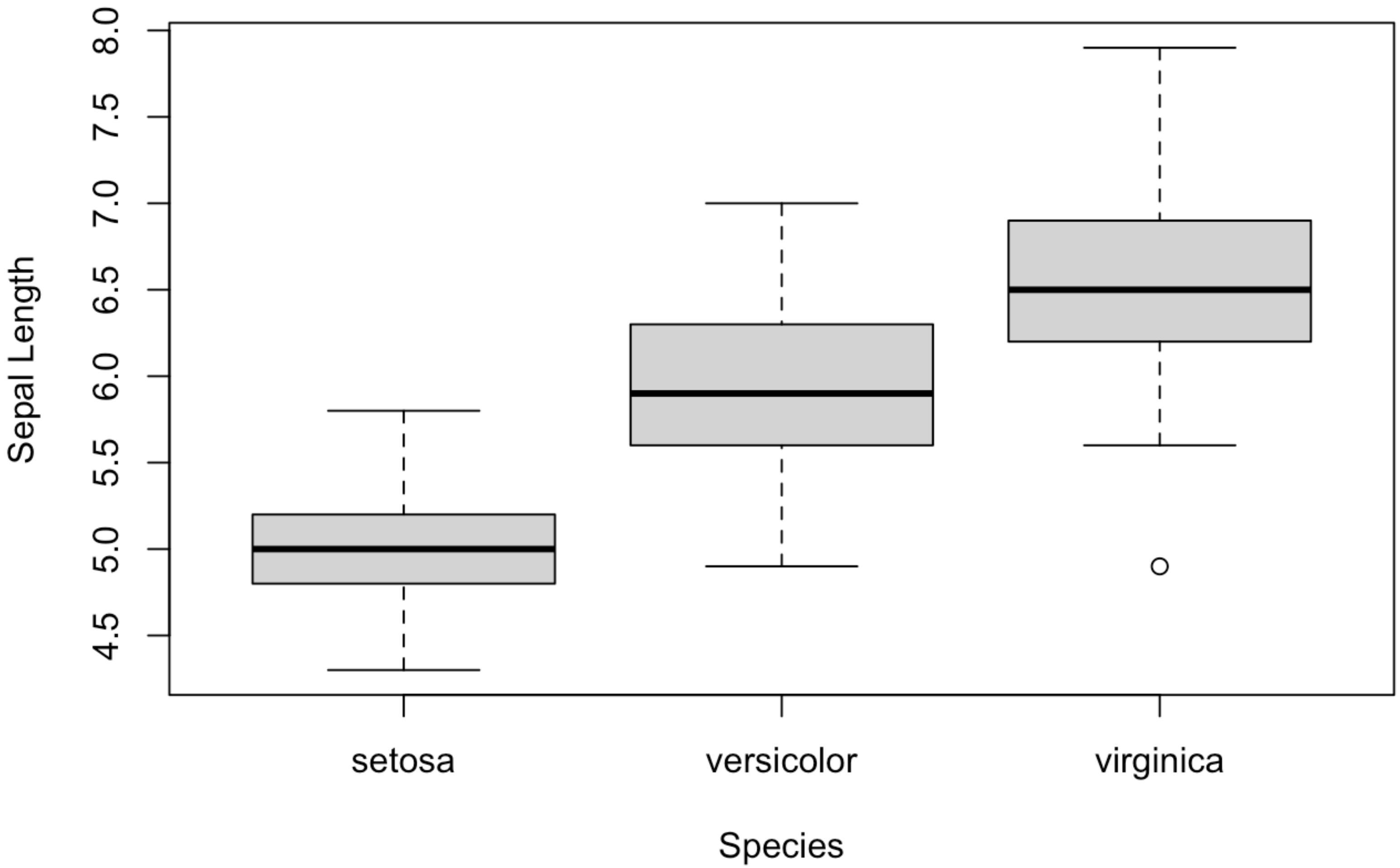
```r
m5=lmer(formula = "scale(Sepal.Length) ~ (1|Species)", data = dat )
summary(m5)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: "scale(Sepal.Length) ~ (1|Species)"
##    Data: dat
##
## REML criterion at convergence: 295.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2669 -0.6404 -0.0253  0.6182  2.5607
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Species  (Intercept) 0.9141   0.9561
##  Residual             0.3865   0.6217
## Number of obs: 150, groups:  Species, 3
##
## Fixed effects:
##               Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) -6.030e-14  5.543e-01  2.000e+00       0        1
```

```r
0.9141/(0.9141+0.3865)
```

```
## [1] 0.7028295
```

```
boxplot(dat$Sepal.Length ~ factor(dat$Species), xlab = "Species", ylab="Sepal Length")
```



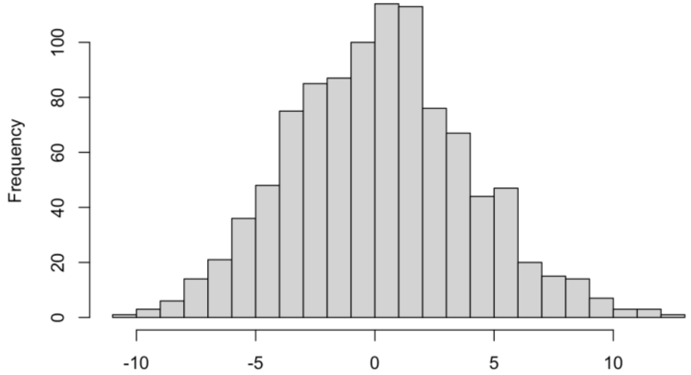
Random effect reformulation

- **Y** = **g** + **e**

- g ~ N(0, GRM . sG^2)
- GRM = XX’/m

- Var(Y)=sG^2 + V(e)
- X full matrix of SNPs centered

$$\begin{bmatrix} BMI_1 \\ \dots \\ BMI_N \end{bmatrix} = \begin{bmatrix} g_1 \\ \dots \\ g_N \end{bmatrix} + \begin{bmatrix} e_1 \\ \dots \\ e_N \end{bmatrix}$$



May start to resemble models seen before

Two formulas one model

Y = X b + e	**Y = g + e**
b ~ N(0, I . sG^2/m)	g ~ N(0, GRM . sG^2)
X full matrix of SNPs	GRM = XX'/m
Var(Y)=p. sG^2 / m + V(e)	Var(Y)=sG^2 + V(e)
High dimensional	Latent factor

Linear mixed model – fixed and random effect

Y = Z b + g + e

g~ N(0, GRM . sG2)

GRM = X’X/m

Z covariates

$$\begin{bmatrix} BMI_1 \\ \dots \\ BMI_N \end{bmatrix} = \begin{bmatrix} Age_1 & Sex_1 & PC_1 \\ \dots & \dots & \dots \\ Age_N & Sex_N & PC_N \end{bmatrix} . \begin{bmatrix} b_{Age} & b_{Sex} & b_{PC} \end{bmatrix} + \begin{bmatrix} rs001_1 & rs\ \dots_1 & rs9999_1 \\ \dots & \dots & \dots \\ rs001_N & rs\ \dots_N & rs9999_N \end{bmatrix} . \begin{bmatrix} b_{rs0011} & b_{rs\dots} & b_{rs9999} \end{bmatrix} + e$$

Focus on GRM

GRM = XX’/m

$$\boldsymbol{SNP} = [\boldsymbol{rs001} \quad \boldsymbol{rs002} \quad \boldsymbol{rs\ldots} \quad \boldsymbol{rs9998} \quad \boldsymbol{rs9999}]$$

X standardised SNP matrix

SNP – raw SNP matrix

$$\boldsymbol{X} = \left[\frac{\boldsymbol{rs001} - mean(rs001)}{sd\,(rs001)} \quad \ldots \quad \frac{\boldsymbol{rs9999} - mean(rs9999)}{sd\,(rs9999)}\right]$$

p_{rsm}: frequency of reference allele for SNP m

$$mean(rs001) = 2p_{rs001}$$

$$sd(rs001) = \sqrt{2p_{rs001}(1 - p_{rs001})}$$

https://en.wikipedia.org/wiki/Binomial_distribution

$\boldsymbol{rsm}$: dosage {0, 1, 2} of SNP m

$$\boxed{\boldsymbol{X} = \left[\frac{\boldsymbol{rs001} - 2p_{rs001}}{\sqrt{2p_{rs001}(1 - p_{rs001})}} \quad \ldots \quad \frac{\boldsymbol{rs9999} - 2p_{rs9999}}{\sqrt{2p_{rs9999}(1 - p_{rs9999})}}\right]}$$

Focus on GRM

GRM = XX’/m

$$\begin{bmatrix} X1_1 & X1_2 & X1_N \\ X2_1 & X2_2 & X2_N \\ X \ldots_1 & X \ldots_2 & X \ldots_N \\ X9998_1 & X9998_2 & X9998_N \\ X9999_1 & X9999_2 & X9999_N \end{bmatrix}$$

$$\begin{bmatrix} X1_1 & X2_1 & X \ldots_1 & X9998_1 & X9999_1 \\ X1_2 & X2_2 & X \ldots_2 & X9998_2 & X9999_2 \\ X1_N & X2_N & \ldots_N & X9998_N & X9999_N \end{bmatrix}$$

$$\begin{bmatrix} \boxed{var(i1)} & cov(i1, i2) & cov(i1, iN) \\ cov(i1, i2) & var(i2) & cov(i1, iN) \\ cov(i1, iN) & cov(i2, iN) & var(indN) \end{bmatrix}$$

Focus on GRM

GRM = XX’/m

$$\begin{bmatrix} X1_1 & X1_2 & X1_N \\ X2_1 & X2_2 & X2_N \\ X\ \ldots_1 & X\ \ldots_2 & X\ \ldots_N \\ X9998_1 & X9998_2 & X9998_N \\ X9999_1 & X9999_2 & X9999_N \end{bmatrix}$$

$$\begin{bmatrix} X1_1 & X2_1 & X\ \ldots_1 & X9998_1 & X9999_1 \\ X1_2 & X2_2 & X\ \ldots_2 & X9998_2 & X9999_2 \\ X1_N & X2_N & \ldots_N & X9998_N & X9999_N \end{bmatrix}$$

$$\begin{bmatrix} var(i1) & \boxed{cov(i1,i2)} & cov(i1,iN) \\ cov(i1,i2) & var(i2) & cov(i1,iN) \\ cov(i1,iN) & cov(i2,iN) & var(indN) \end{bmatrix}$$

Focus on GRM

GRM = XX’/m

$$GRM_{ij} = \frac{1}{m}\sum\nolimits_{m} Xm_i \,.\, Xm_j$$

$$GRM_{ij} = \frac{1}{m}\sum\nolimits_{m} \frac{rsm_i - 2p_{rsm}}{\sqrt{2p_{rsm}(1-p_{rsm})}} \cdot \frac{rsm_j - 2p_{rsm}}{\sqrt{2p_{rsm}(1-p_{rsm})}}$$

$$\begin{bmatrix} X1_1 & X1_2 & X1_N \\ X2_1 & X2_2 & X2_N \\ X\,\ldots_1 & X\,\ldots_2 & X\,\ldots_N \\ X9998_1 & X9998_2 & X9998_N \\ X9999_1 & X9999_2 & X9999_N \end{bmatrix}$$

$$\begin{bmatrix} X1_1 & X2_1 & X\,\ldots_1 & X9998_1 & X9999_1 \\ X1_2 & X2_2 & X\,\ldots_2 & X9998_2 & X9999_2 \\ X1_N & X2_N & \ldots_N & X9998_N & X9999_N \end{bmatrix}$$

$$\begin{bmatrix} var(i1) & \boxed{cov(i1,i2)} & cov(i1,iN) \\ cov(i1,i2) & var(i2) & cov(i1,iN) \\ cov(i1,iN) & cov(i2,iN) & var(indN) \end{bmatrix}$$

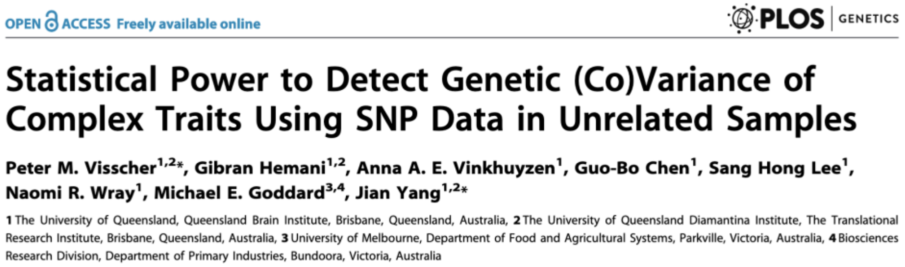
Statistical power and GRM

Power depends on the variance of GRM off-diagonal elements

https://shiny.cnsgenomics.com/TwinPower/

https://shiny.cnsgenomics.com/gctaPower/

Statistical power : probability of detecting a true association. Function of: type of effect (fixed, random), hypothesised effect size r, sample size N, risk alpha (5%).

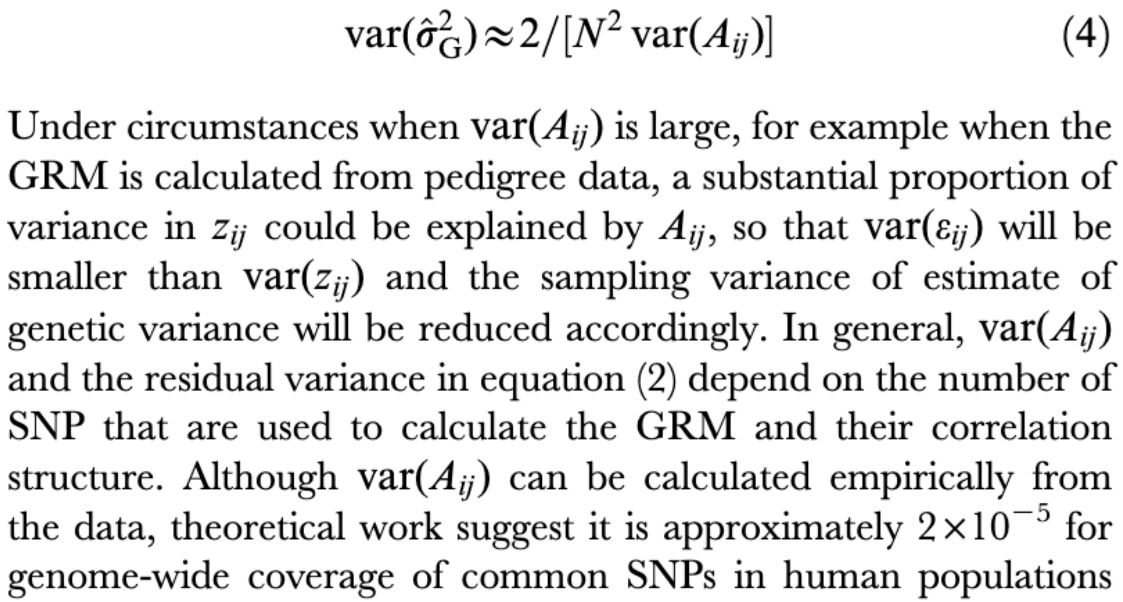


OPEN ACCESS Freely available online PLOS GENETICS

Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples

Peter M. Visscher[1,2]*, Gibran Hemani[1,2], Anna A. E. Vinkhuyzen[1], Guo-Bo Chen[1], Sang Hong Lee[1], Naomi R. Wray[1], Michael E. Goddard[3,4], Jian Yang[1,2]*

1 The University of Queensland, Queensland Brain Institute, Brisbane, Queensland, Australia, 2 The University of Queensland Diamantina Institute, The Translational Research Institute, Brisbane, Queensland, Australia, 3 University of Melbourne, Department of Food and Agricultural Systems, Parkville, Victoria, Australia, 4 Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria, Australia



$$\mathrm{var}(\hat{\sigma}^2_G) \approx 2/[N^2\,\mathrm{var}(A_{ij})] \qquad (4)$$

Under circumstances when $\mathrm{var}(A_{ij})$ is large, for example when the GRM is calculated from pedigree data, a substantial proportion of variance in z_{ij} could be explained by A_{ij}, so that $\mathrm{var}(\varepsilon_{ij})$ will be smaller than $\mathrm{var}(z_{ij})$ and the sampling variance of estimate of genetic variance will be reduced accordingly. In general, $\mathrm{var}(A_{ij})$ and the residual variance in equation (2) depend on the number of SNP that are used to calculate the GRM and their correlation structure. Although $\mathrm{var}(A_{ij})$ can be calculated empirically from the data, theoretical work suggest it is approximately 2×10^{-5} for genome-wide coverage of common SNPs in human populations

Association testing - pvalues

Nested models
compare two models, one being subset of other (in terms of variables)

Y = Z b + g + e **Full**

Y = Z b + e **Nested**

Difference between models where 1 or several parameters of interest are dropped

$$\lambda_{\text{LR}} = -2 \left[\ell(\theta_0) - \ell(\hat{\theta}) \right]$$

Follows a chi2(z) distribution
Z : difference of parameters between full and nested models

https://en.wikipedia.org/wiki/Likelihood-ratio_test

Summary

LMM are extensions of GLM

R2 useful to measure association with several variables

GRM = XX’/p – NxN variance covariance matrix of random effect

GRM sufficient to describe random effect – easier to manipulate than full matrix of SNPs X

GRM contains information about sample composition (familial, cryptic, statistical power..)

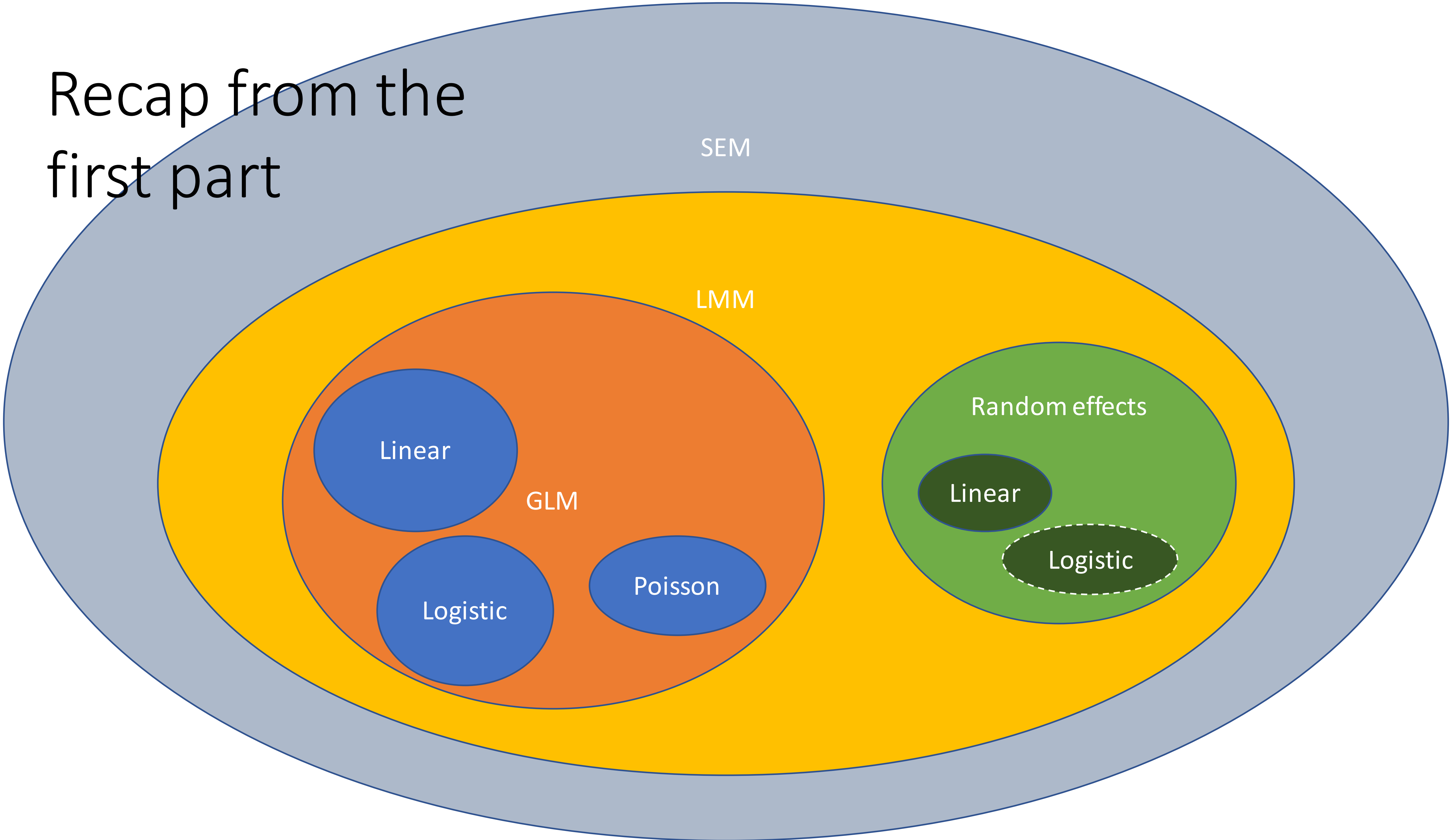
General GLM & LMM model formulation and estimation via likelihood

Part 2 – these models look familiar

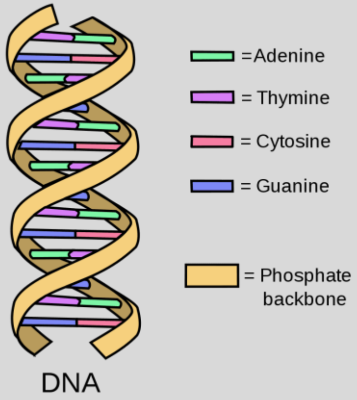
Baptiste Couvy-Duchesne

baptiste.couvyduchesne@uq.edu.au
baptiste.couvy@icm-institute.org
https://github.com/baptisteCD
@BaptisteCouvy

Recap from the first part



SNP heritability



Source: wikicommons

LMM

$\mathbf{Y} = \mathbf{Z}\,\mathbf{a} + \mathbf{X}\,b + \mathbf{e}$

- b~ N(0, I . sG2/m)
- Z : age, sex, site …

- Formulation emphasises X (Large matrix of SNPs)
- sG2/var(Y) = SNP heritability
- In practice LMM with covariates (age, sex, site…)
- Estimated from individual level data using GCTA

ACE model

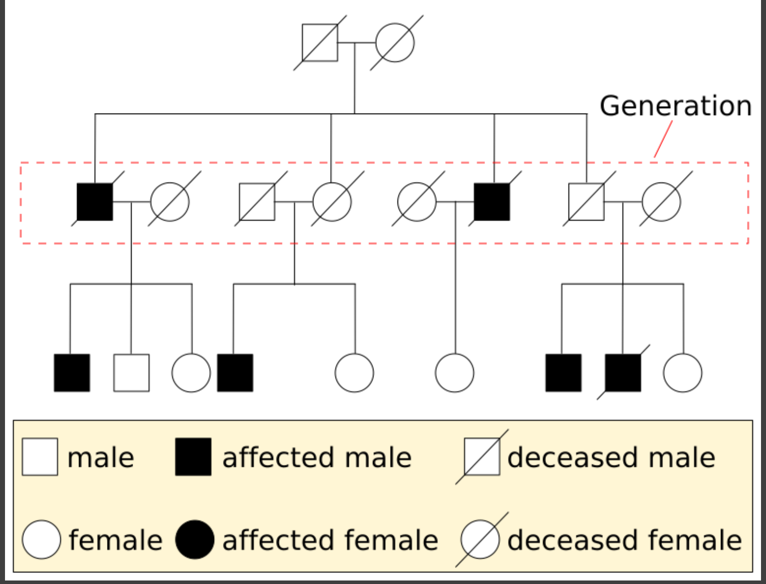
LMM

- **Y = Z a + g + c + e**
- g ~ N(0, GRM . sG2)
- c ~ N(0, CRM . CG2)

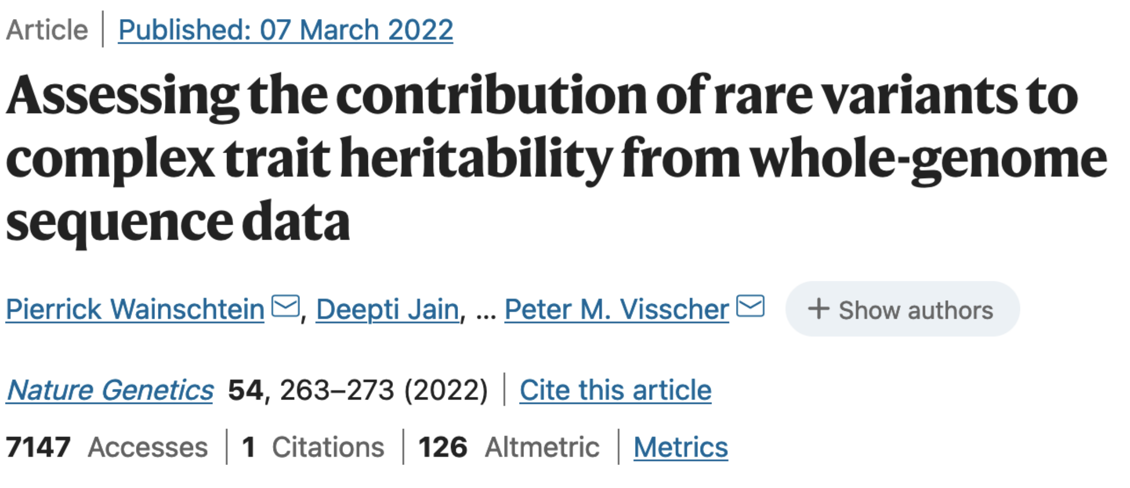
- sG2/var(Y) = heritability
- sC2/var(Y) = shared E

- SNPs not observed, no direct measurement of C

- Pedigree: approximation of GRM calculated from common and rare SNPs
- Assumed variance covariance of shared environment



Heritability from Whole Genome Sequencing



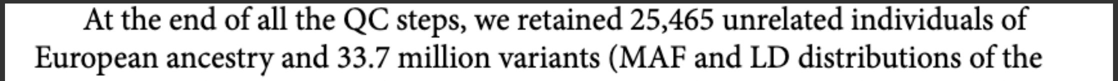
Article | Published: 07 March 2022

Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data

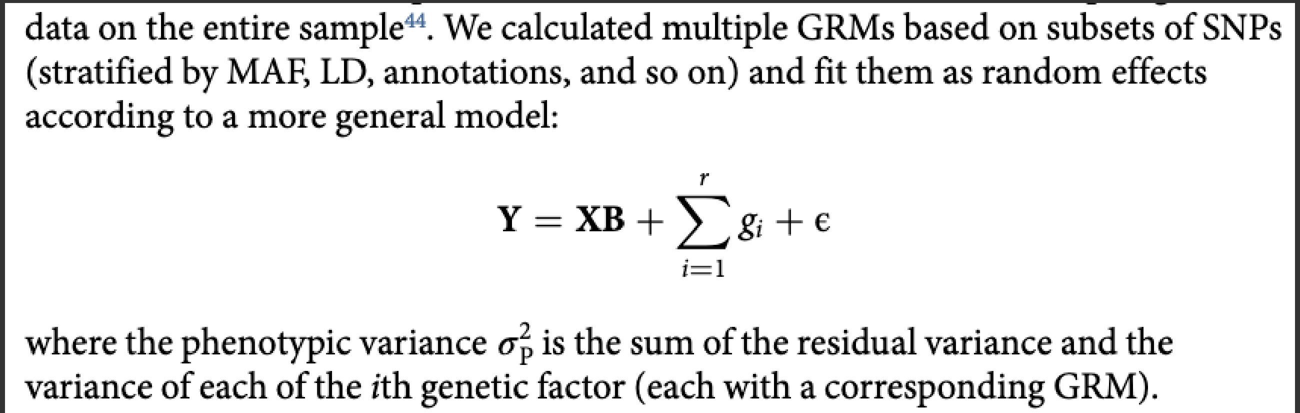
Pierrick Wainschtein, Deepti Jain, ... Peter M. Visscher   + Show authors

Nature Genetics **54**, 263–273 (2022) | Cite this article

7147 Accesses | **1** Citations | **126** Altmetric | Metrics



At the end of all the QC steps, we retained 25,465 unrelated individuals of European ancestry and 33.7 million variants (MAF and LD distributions of the



data on the entire sample[44]. We calculated multiple GRMs based on subsets of SNPs (stratified by MAF, LD, annotations, and so on) and fit them as random effects according to a more general model:

$$\mathbf{Y} = \mathbf{XB} + \sum_{i=1}^{r} g_i + \epsilon$$

where the phenotypic variance σ_P^2 is the sum of the residual variance and the variance of each of the ith genetic factor (each with a corresponding GRM).

- Variance partitioning – rare and common SNPs

- Slightly different kinship metric for complex population structure (see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6220858/)

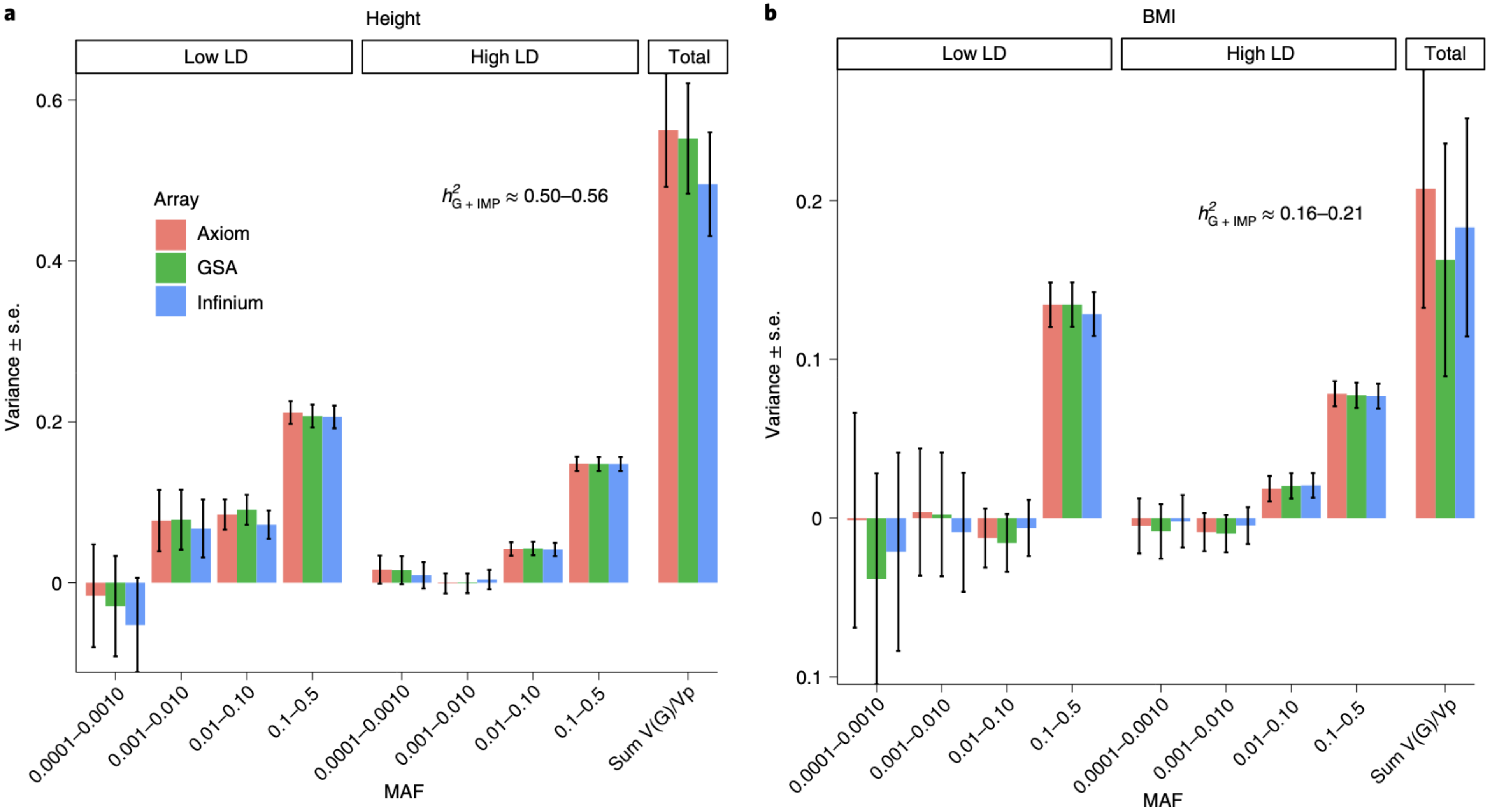
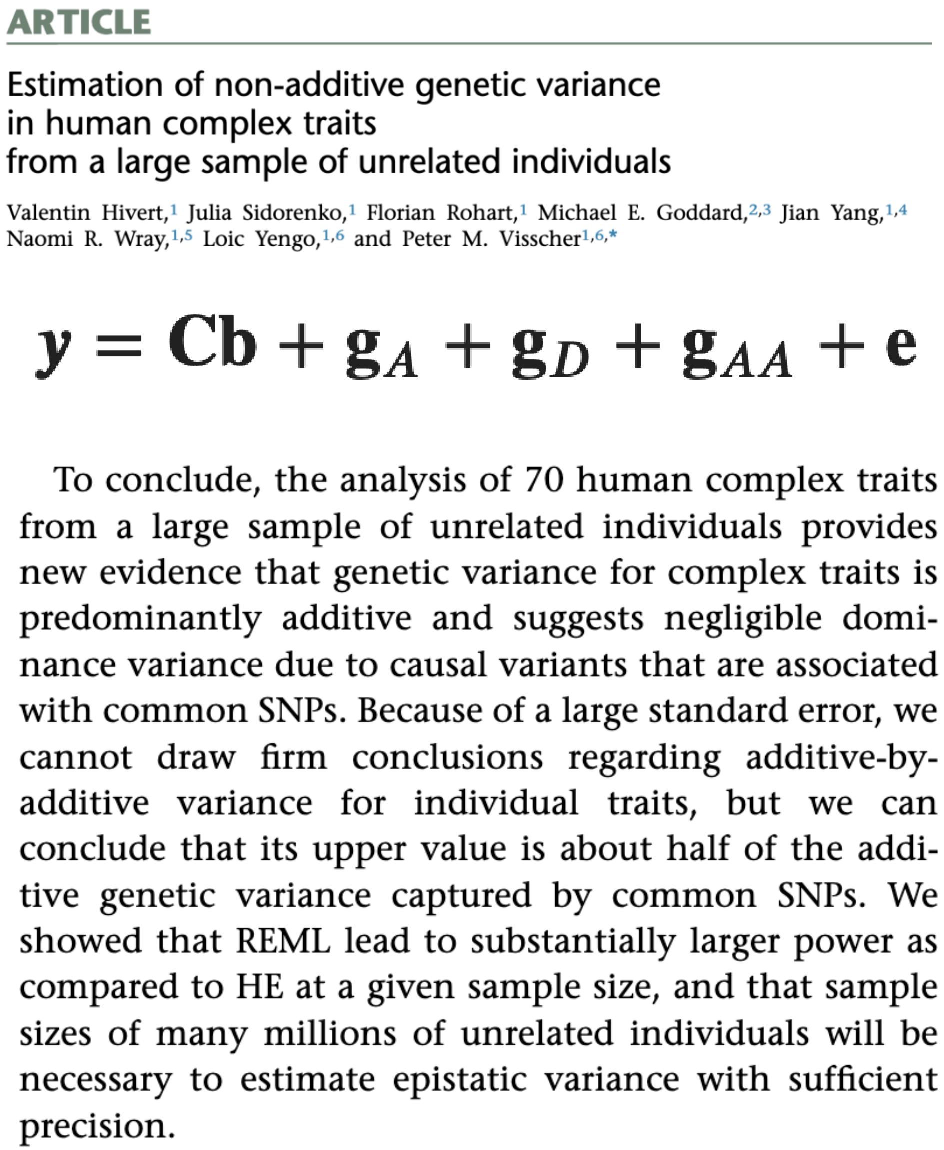


Fig. 1 | GREML-LDMS estimates with 8 bins (2 LD bins for each of the 4 MAF bins) correcting for 20 PCs (calculated from LD-pruned HM3 SNPs) after imputing SNPs from Illumina InfiniumCore24, GSA 24 and Affymetrix Axiom arrays using HRC reference panels for $n = 25,465$ samples. a, Estimates of h^2_{G+IMP} for height are between 0.50 and 0.56 (s.e. = 0.06–0.07). **b**, Estimates for BMI are between 0.16 and 0.21 (s.e. = 0.07). The large s.e. values of

Additive, Dominance and Additive by Additive variance

Interaction between random effects



ARTICLE

Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals

Valentin Hivert,[1] Julia Sidorenko,[1] Florian Rohart,[1] Michael E. Goddard,[2,3] Jian Yang,[1,4] Naomi R. Wray,[1,5] Loic Yengo,[1,6] and Peter M. Visscher[1,6,*]

$$\boldsymbol{y} = \mathbf{Cb} + \mathbf{g}_A + \mathbf{g}_D + \mathbf{g}_{AA} + \mathbf{e}$$

To conclude, the analysis of 70 human complex traits from a large sample of unrelated individuals provides new evidence that genetic variance for complex traits is predominantly additive and suggests negligible dominance variance due to causal variants that are associated with common SNPs. Because of a large standard error, we cannot draw firm conclusions regarding additive-by-additive variance for individual traits, but we can conclude that its upper value is about half of the additive genetic variance captured by common SNPs. We showed that REML lead to substantially larger power as compared to HE at a given sample size, and that sample sizes of many millions of unrelated individuals will be necessary to estimate epistatic variance with sufficient precision.

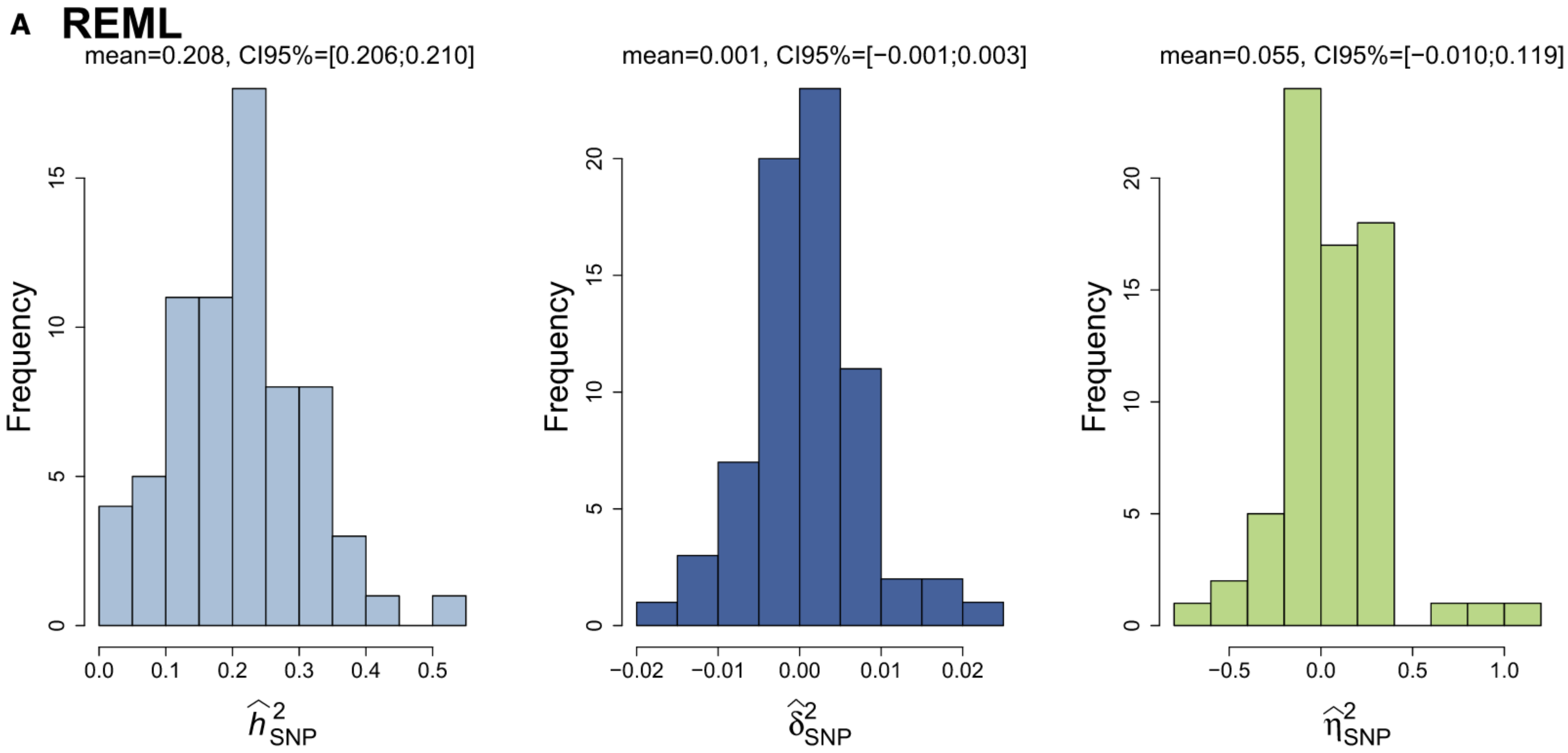


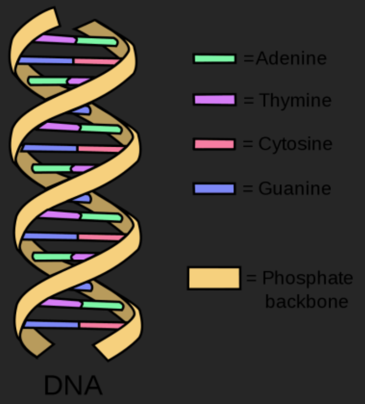
Figure 4. Distributions of the REML and HE estimates of SNP-based h^2_{SNP}, δ^2_{SNP} and η^2_{SNP} for 70 continuous traits in the UK Biobank
For each distribution of variance components estimates—REML (A) and HE (B)—we indicate the mean estimate as well as the 95% confidence interval (CI95%).

GWAS in PLINK

Y = X b + a **SNP**i + **e**

X : age, sex, site, genetic PCs

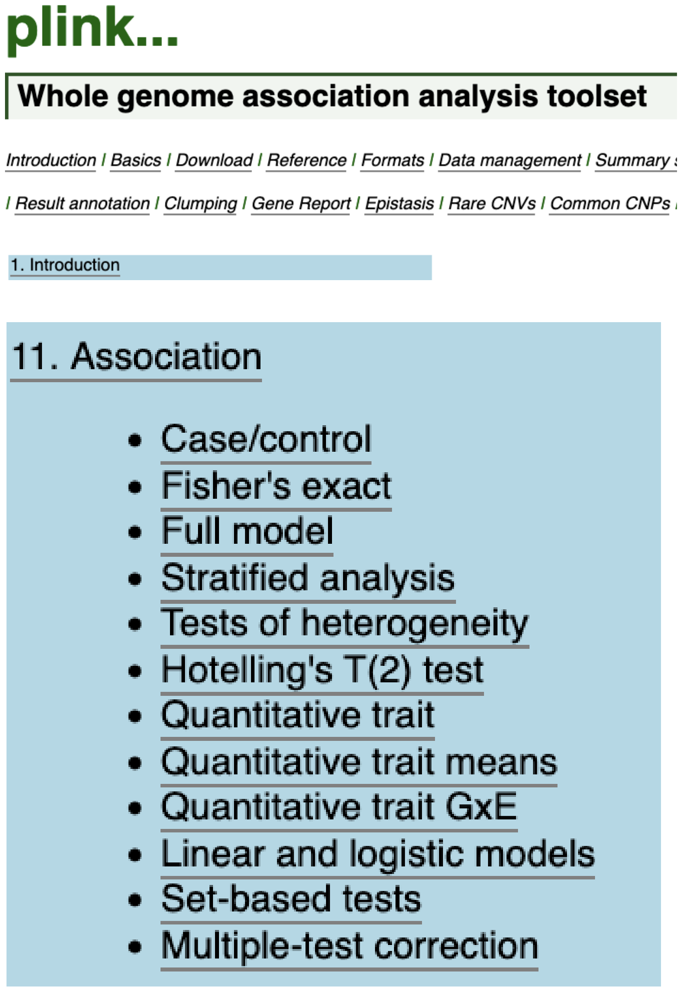
https://zzz.bwh.harvard.edu/plink/



Source: wikicommons

Linear

Logistic



plink...

Whole genome association analysis toolset

Introduction | *Basics* | *Download* | *Reference* | *Formats* | *Data management* | *Summary* :

| *Result annotation* | *Clumping* | *Gene Report* | *Epistasis* | *Rare CNVs* | *Common CNPs* :

1. Introduction

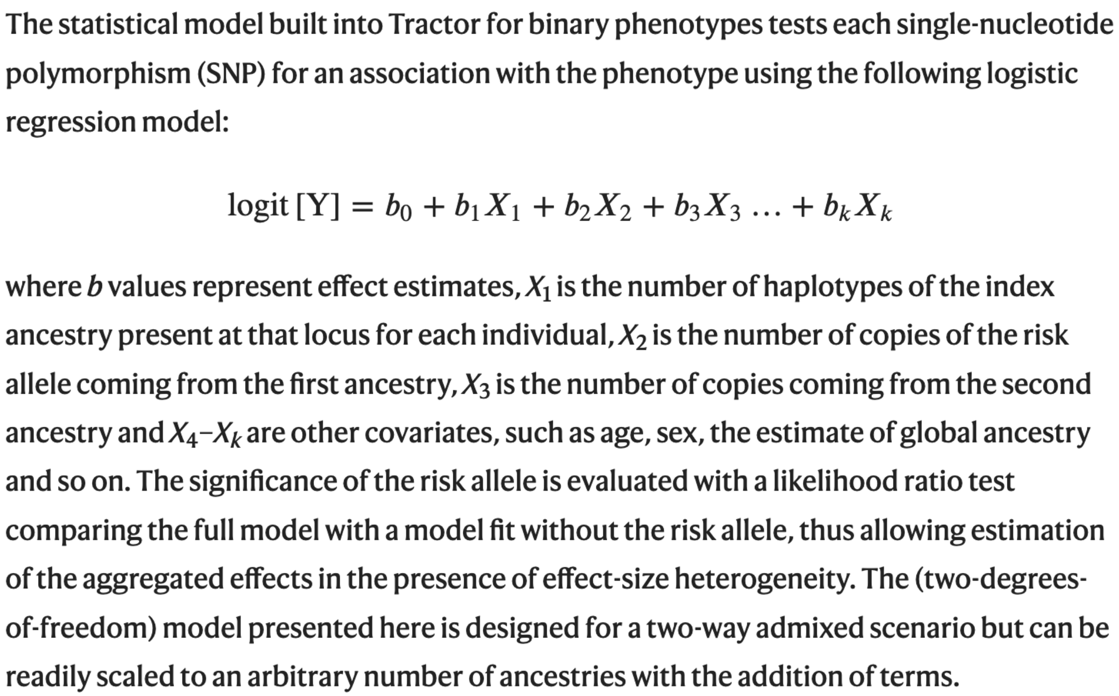
11. Association

- Case/control
- Fisher's exact
- Full model
- Stratified analysis
- Tests of heterogeneity
- Hotelling's T(2) test
- Quantitative trait
- Quantitative trait means
- Quantitative trait GxE
- Linear and logistic models
- Set-based tests
- Multiple-test correction

GWAS in TRACTOR

Linear

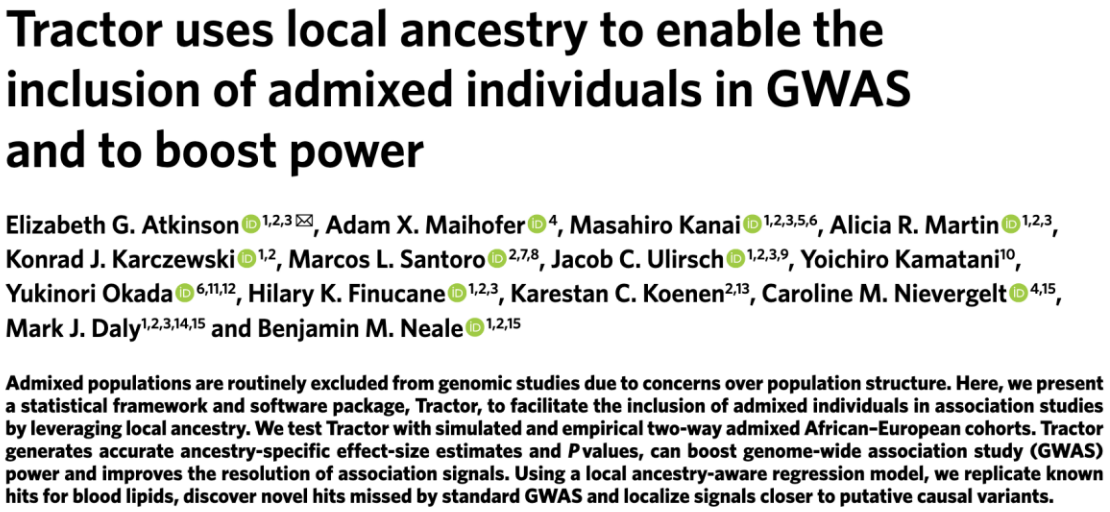
Logistic



The statistical model built into Tractor for binary phenotypes tests each single-nucleotide polymorphism (SNP) for an association with the phenotype using the following logistic regression model:

$$\text{logit}\,[\text{Y}] = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \ldots + b_kX_k$$

where b values represent effect estimates, X_1 is the number of haplotypes of the index ancestry present at that locus for each individual, X_2 is the number of copies of the risk allele coming from the first ancestry, X_3 is the number of copies coming from the second ancestry and X_4–X_k are other covariates, such as age, sex, the estimate of global ancestry and so on. The significance of the risk allele is evaluated with a likelihood ratio test comparing the full model with a model fit without the risk allele, thus allowing estimation of the aggregated effects in the presence of effect-size heterogeneity. The (two-degrees-of-freedom) model presented here is designed for a two-way admixed scenario but can be readily scaled to an arbitrary number of ancestries with the addition of terms.

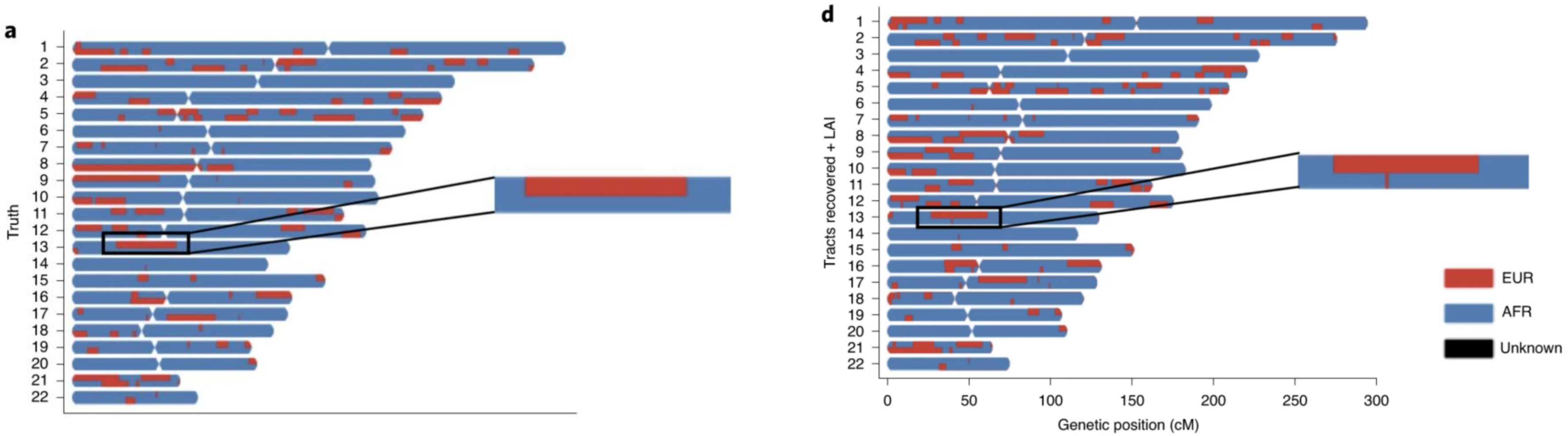


Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power

Elizabeth G. Atkinson[1,2,3]✉, Adam X. Maihofer[4], Masahiro Kanai[1,2,3,5,6], Alicia R. Martin[1,2,3], Konrad J. Karczewski[1,2], Marcos L. Santoro[2,7,8], Jacob C. Ulirsch[1,2,3,9], Yoichiro Kamatani[10], Yukinori Okada[6,11,12], Hilary K. Finucane[1,2,3], Karestan C. Koenen[2,13], Caroline M. Nievergelt[4,15], Mark J. Daly[1,2,3,14,15] and Benjamin M. Neale[1,2,15]

Admixed populations are routinely excluded from genomic studies due to concerns over population structure. Here, we present a statistical framework and software package, Tractor, to facilitate the inclusion of admixed individuals in association studies by leveraging local ancestry. We test Tractor with simulated and empirical two-way admixed African–European cohorts. Tractor generates accurate ancestry-specific effect-size estimates and P values, can boost genome-wide association study (GWAS) power and improves the resolution of association signals. Using a local ancestry-aware regression model, we replicate known hits for blood lipids, discover novel hits missed by standard GWAS and localize signals closer to putative causal variants.

Here, we have developed a scalable framework that allows for the incorporation of admixed individuals into large-scale genomics efforts by using local ancestry inference (LAI) (Fig. 1). Our framework, distributed as a software package named Tractor, generates ancestry dosages at each site from LAI calls, extracts painted haplotype segments to correct population structure at the genotype level, and runs a local ancestry-aware regression model, producing ancestry-specific effect-size estimates and *P* values.

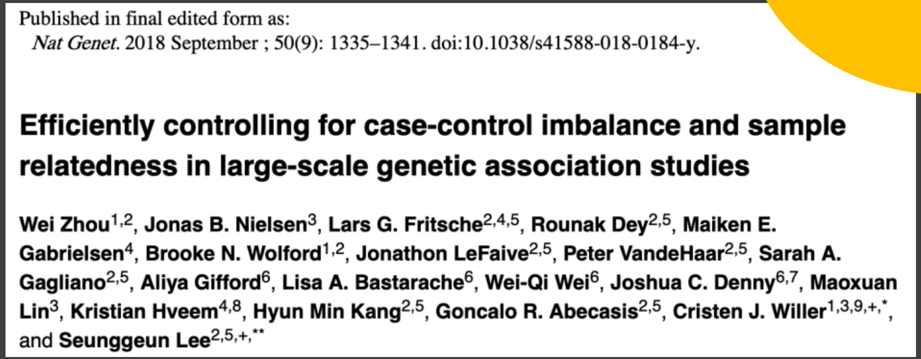


a, Truth results for an example individual in our simulated African American cohort. **b**, Results for the person after statistical phasing. Note the disruption of long haplotypes resulting from phasing switch errors. **c**, Recovery of tracts broken by switch errors in phasing. **d**, Smoothing and further improvement of tracts acquired through an additional round of LAI. The same section of chromosome 13, showing an example tract at higher resolution, is pictured on the right of each panel to highlight tract recovery. Local EUR and AFR ancestral tracts are shown in red and blue, respectively.

GWAS in SAIGE, gcta, bolt-LMM, fastLMM…

LMM

- **Y** = **Z a** + **X** b + a SNPi + **e**
- b~ N(0, I . sG2 / m)
- Z : age, sex, site …
- X : can be leave one chromosome out set of SNP

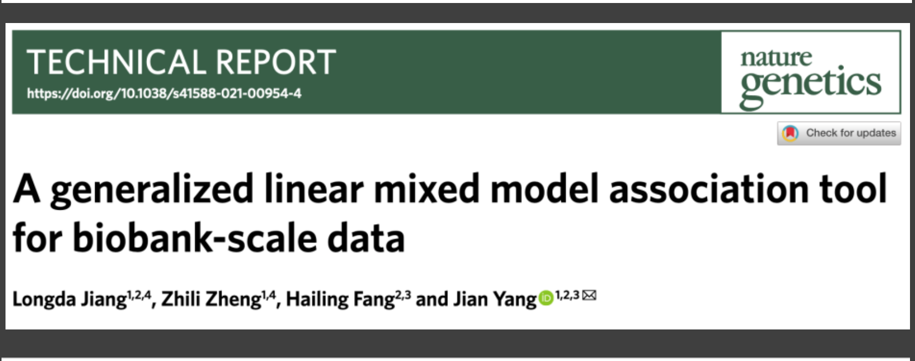


Published in final edited form as:
Nat Genet. 2018 September ; 50(9): 1335–1341. doi:10.1038/s41588-018-0184-y.

Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies

Wei Zhou[1,2], **Jonas B. Nielsen**[3], **Lars G. Fritsche**[2,4,5], **Rounak Dey**[2,5], **Maiken E. Gabrielsen**[4], **Brooke N. Wolford**[1,2], **Jonathon LeFaive**[2,5], **Peter VandeHaar**[2,5], **Sarah A. Gagliano**[2,5], **Aliya Gifford**[6], **Lisa A. Bastarache**[6], **Wei-Qi Wei**[6], **Joshua C. Denny**[6,7], **Maoxuan Lin**[3], **Kristian Hveem**[4,8], **Hyun Min Kang**[2,5], **Goncalo R. Abecasis**[2,5], **Cristen J. Willer**[1,3,9,+,*], and **Seunggeun Lee**[2,5,+,**]

SAIGE



TECHNICAL REPORT
https://doi.org/10.1038/s41588-021-00954-4

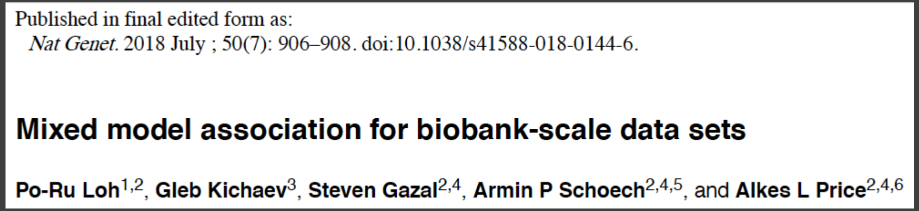
nature genetics

Check for updates

A generalized linear mixed model association tool for biobank-scale data

Longda Jiang[1,2,4], **Zhili Zheng**[1,4], **Hailing Fang**[2,3] and **Jian Yang**[1,2,3]

Fast-GWAS GLMM



Published in final edited form as:
Nat Genet. 2018 July ; 50(7): 906–908. doi:10.1038/s41588-018-0144-6.

Mixed model association for biobank-scale data sets

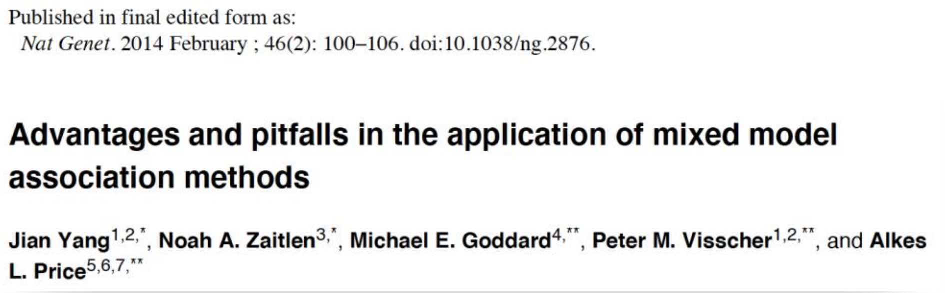
Po-Ru Loh[1,2], **Gleb Kichaev**[3], **Steven Gazal**[2,4], **Armin P Schoech**[2,4,5], and **Alkes L Price**[2,4,6]

Bolt-LMM

Advantages and pitfalls of LMM (in GWAS)

- Power increase and lower false positive rate
- Better control of population structure
- Greater power by conditioning on other hits

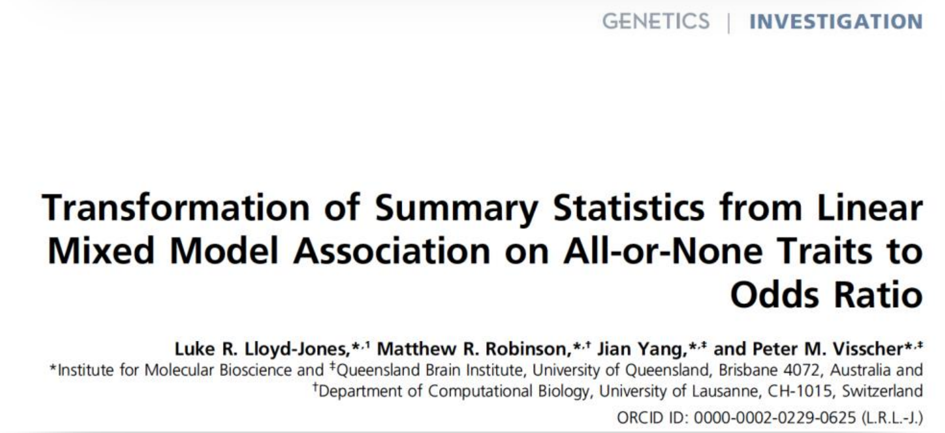
- Higher computational burden than GLM
- Weary of not enough SNPs in GRM
- Double fitting in GWAS
- Logistic LMM not always implemented



Published in final edited form as:
Nat Genet. 2014 February ; 46(2): 100–106. doi:10.1038/ng.2876.

Advantages and pitfalls in the application of mixed model association methods

Jian Yang[1,2,*], **Noah A. Zaitlen**[3,*], **Michael E. Goddard**[4,**], **Peter M. Visscher**[1,2,**], and **Alkes L. Price**[5,6,7,**]



GENETICS | INVESTIGATION

Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio

Luke R. Lloyd-Jones,[*,1] **Matthew R. Robinson,**[*,†] **Jian Yang,**[*,‡] **and Peter M. Visscher**[*,‡]
[*]Institute for Molecular Bioscience and [‡]Queensland Brain Institute, University of Queensland, Brisbane 4072, Australia and [†]Department of Computational Biology, University of Lausanne, CH-1015, Switzerland
ORCID ID: 0000-0002-0229-0625 (L.R.L.-J.)

Part 3 – We can write a lot more models

Baptiste Couvy-Duchesne

baptiste.couvyduchesne@uq.edu.au
baptiste.couvy@icm-institute.org
https://github.com/baptisteCD
@BaptisteCouvy

International Statistical Genetics Workshop • Boulder, Colorado •
IBG

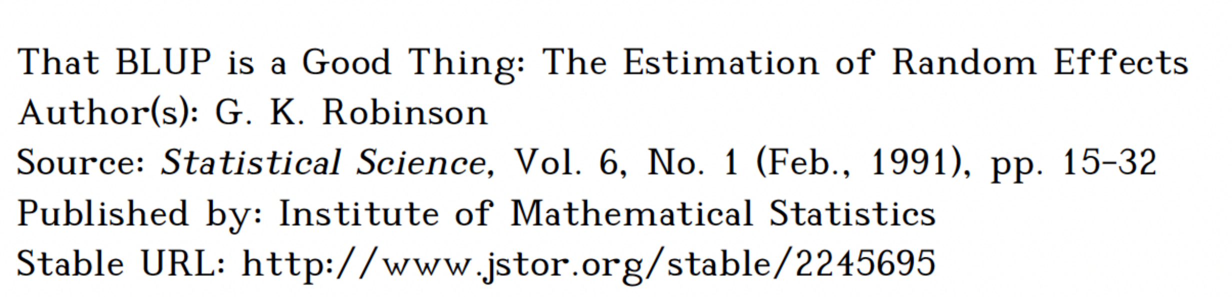
Best linear Unbiased prediction (BLUP)

LMM

Y = X b + e

b~ N(0, I. sG2 / m)

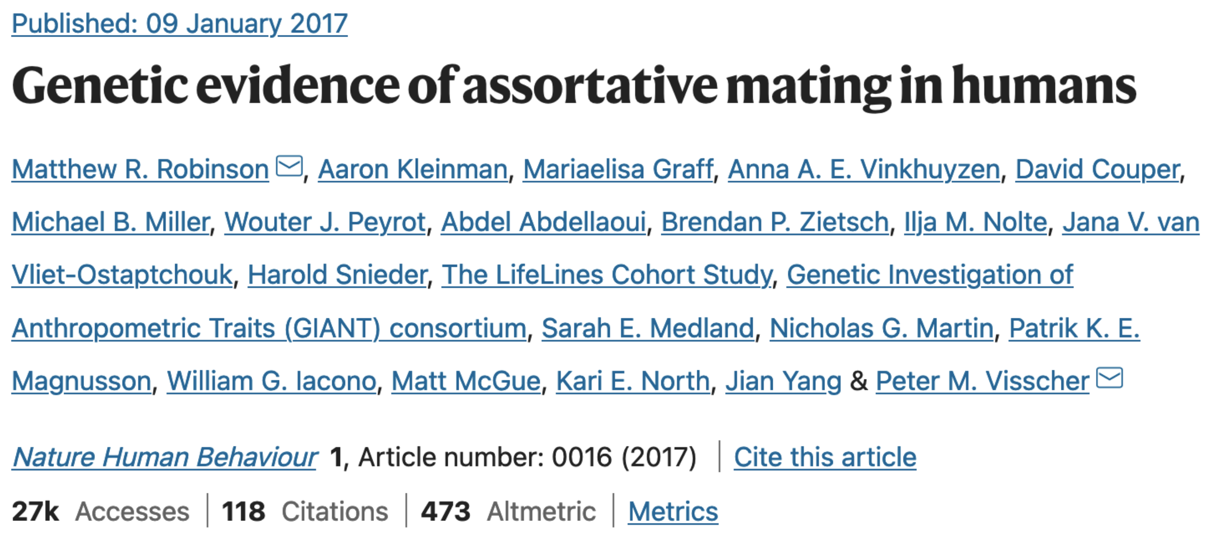
$X\,\hat{b}$ – in a new sample: prediction from SNPs



That BLUP is a Good Thing: The Estimation of Random Effects
Author(s): G. K. Robinson
Source: *Statistical Science*, Vol. 6, No. 1 (Feb., 1991), pp. 15-32
Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/2245695

<u>In Practice:</u>

sBLUP – summary statistics – beta from GWAS and transform effects into marginal effects using reference LD matrix



Published: 09 January 2017

Genetic evidence of assortative mating in humans

Matthew R. Robinson, Aaron Kleinman, Mariaelisa Graff, Anna A. E. Vinkhuyzen, David Couper, Michael B. Miller, Wouter J. Peyrot, Abdel Abdellaoui, Brendan P. Zietsch, Ilja M. Nolte, Jana V. van Vliet-Ostaptchouk, Harold Snieder, The LifeLines Cohort Study, Genetic Investigation of Anthropometric Traits (GIANT) consortium, Sarah E. Medland, Nicholas G. Martin, Patrik K. E. Magnusson, William G. Iacono, Matt McGue, Kari E. North, Jian Yang & Peter M. Visscher

Nature Human Behaviour **1**, Article number: 0016 (2017) | Cite this article

27k Accesses | **118** Citations | **473** Altmetric | Metrics

(s)BayesR predictor

LMM

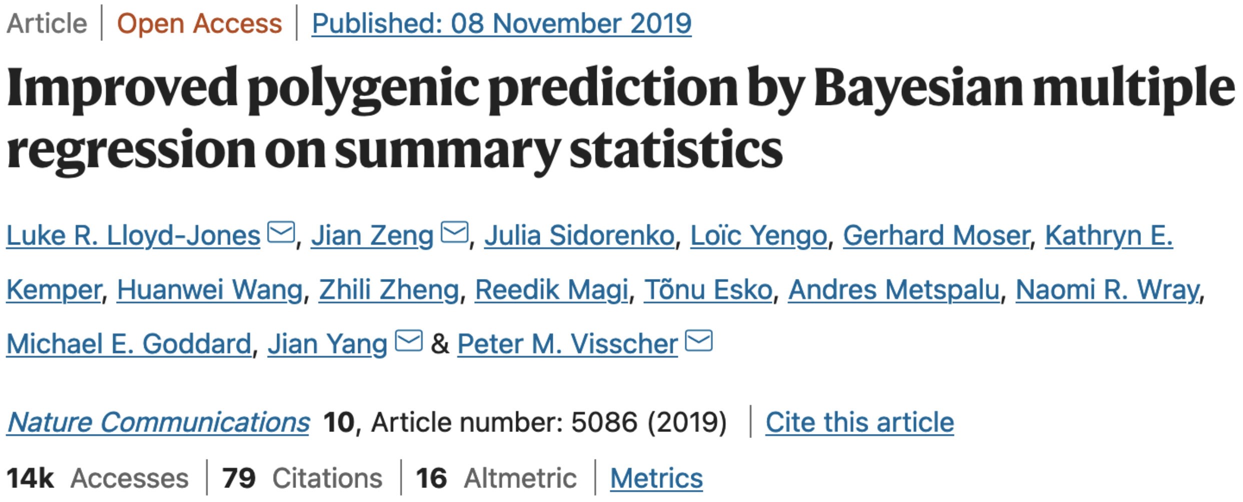
Y = **X1 b1** + **X2 b2** + **X3 b3** + **e**

bi ~ N(0, I . sGi2)

- Mixture of components

 (SNPs stratified by MAF)

- Relax hypothesis of single distribution



Article | Open Access | Published: 08 November 2019

Improved polygenic prediction by Bayesian multiple regression on summary statistics

Luke R. Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R. Wray, Michael E. Goddard, Jian Yang & Peter M. Visscher

Nature Communications **10**, Article number: 5086 (2019) | Cite this article

14k Accesses | **79** Citations | **16** Altmetric | Metrics

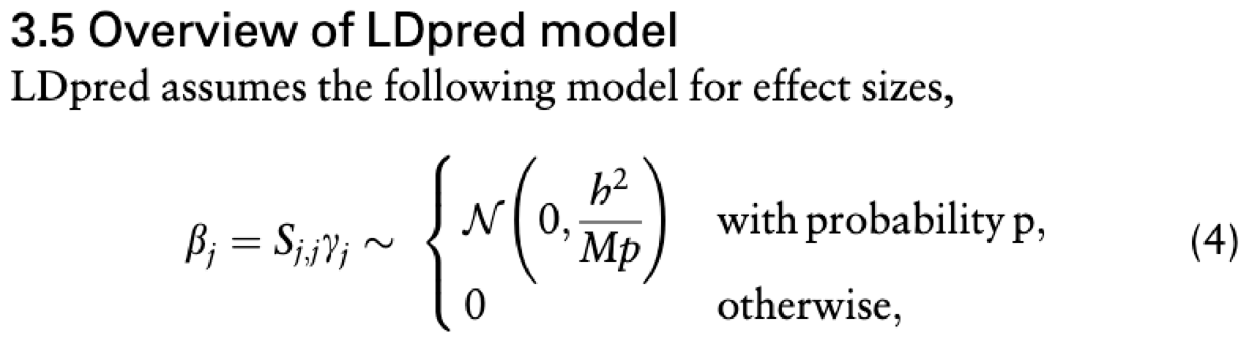
Estimated using GCTB (summary statistics)
https://cnsgenomics.com/software/gctb/#Overview

LDPRED & LDPRED2

LMM

- **Y = X b + e**
- b ~ N(0, I . sG2 / pm) with proba p
- 0 otherwise

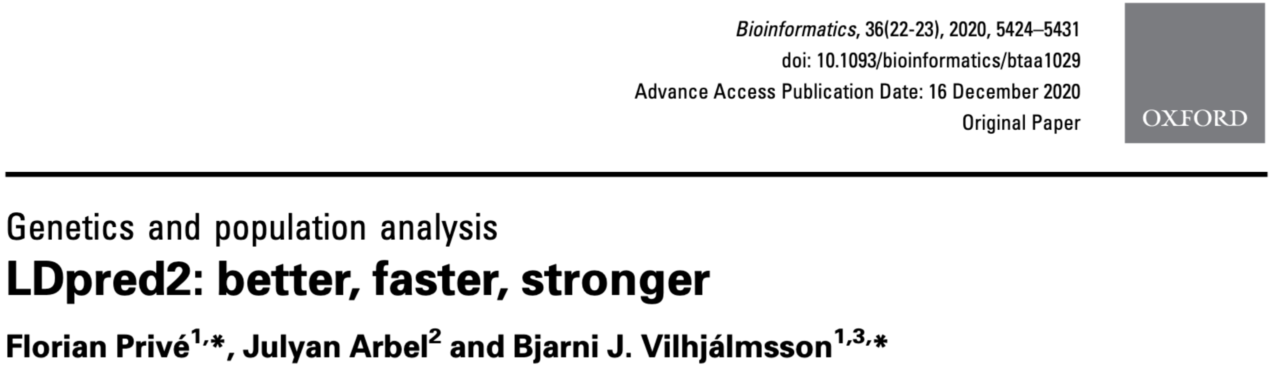
- Mixture of components
- Extra parameter p (estimated automatically in LDpred2)



3.5 Overview of LDpred model

LDpred assumes the following model for effect sizes,

$$\beta_j = S_{j,j}\gamma_j \sim \begin{cases} \mathcal{N}\left(0, \dfrac{h^2}{Mp}\right) & \text{with probability p,} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$



Bioinformatics, 36(22-23), 2020, 5424–5431
doi: 10.1093/bioinformatics/btaa1029
Advance Access Publication Date: 16 December 2020
Original Paper

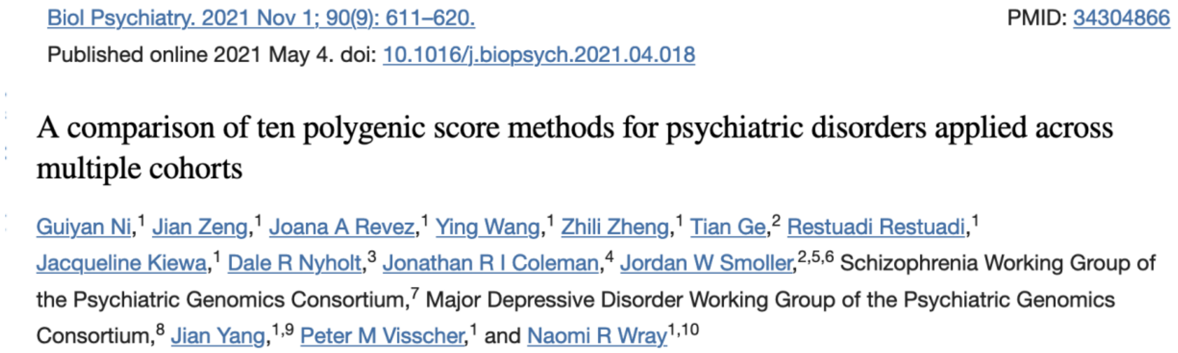
OXFORD

Genetics and population analysis

LDpred2: better, faster, stronger

Florian Privé[1,*], Julyan Arbel[2] and Bjarni J. Vilhjálmsson[1,3,*]

Polygenic risk scores

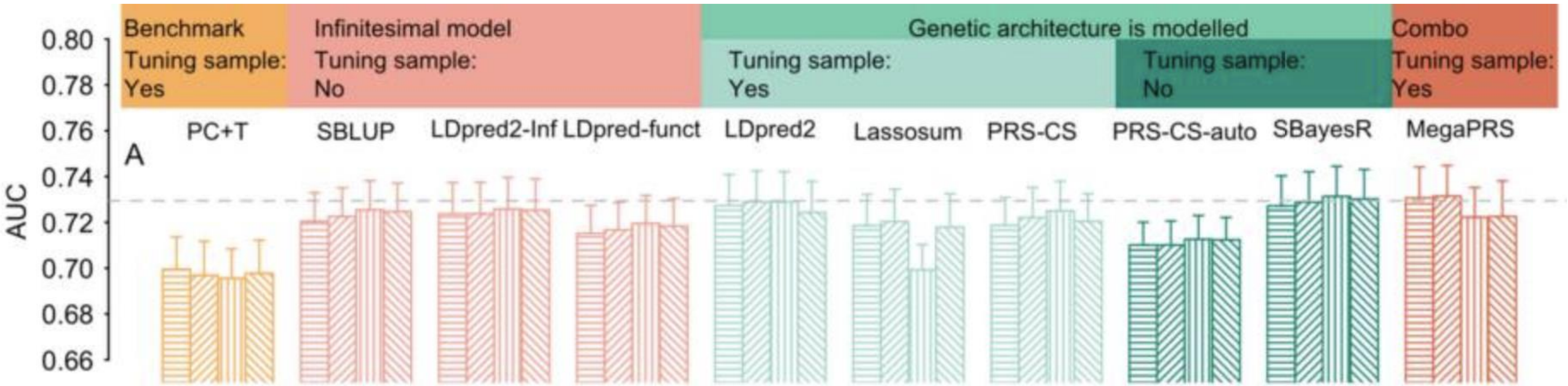


Biol Psychiatry. 2021 Nov 1; 90(9): 611–620. PMID: 34304866

Published online 2021 May 4. doi: 10.1016/j.biopsych.2021.04.018

A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts

Guiyan Ni,[1] Jian Zeng,[1] Joana A Revez,[1] Ying Wang,[1] Zhili Zheng,[1] Tian Ge,[2] Restuadi Restuadi,[1] Jacqueline Kiewa,[1] Dale R Nyholt,[3] Jonathan R I Coleman,[4] Jordan W Smoller,[2,5,6] Schizophrenia Working Group of the Psychiatric Genomics Consortium,[7] Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium,[8] Jian Yang,[1,9] Peter M Visscher,[1] and Naomi R Wray[1,10]



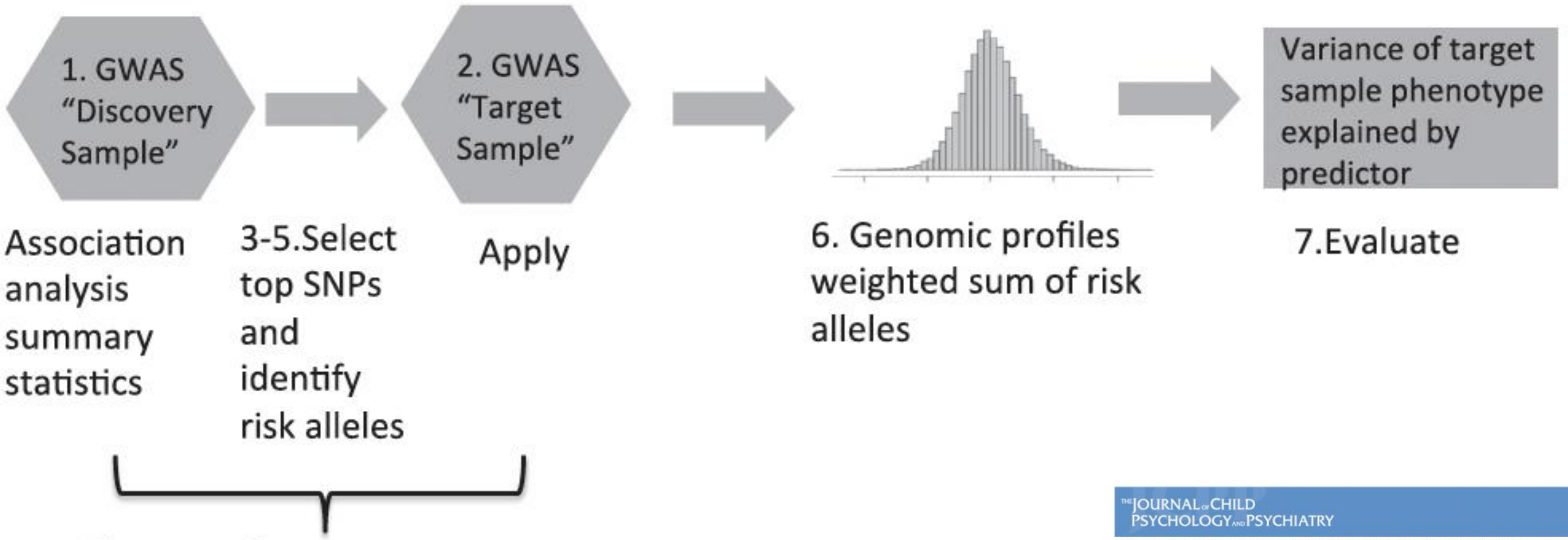
Prediction accuracy of a PRS

$$\mathbf{Y} = \mathbf{Z}\ \mathbf{a} + \mathbf{PRS}\ .\ b + \mathbf{e}$$

Z : age, sex, site, genetic PCs …

Linear

Logistic


1. GWAS "Discovery Sample"
2. GWAS "Target Sample"
Variance of target sample phenotype explained by predictor

Association analysis summary statistics

3-5.Select top SNPs and identify risk alleles

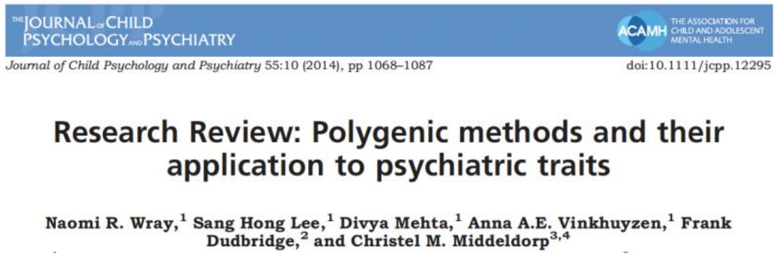
Apply

6. Genomic profiles weighted sum of risk alleles

7.Evaluate

Discovery & Target samples could be:

A. Same Disorder
B. Different disorders
C. Disorder subtypes


THE JOURNAL OF CHILD PSYCHOLOGY AND PSYCHIATRY

ACAMH THE ASSOCIATION FOR CHILD AND ADOLESCENT MENTAL HEALTH

Journal of Child Psychology and Psychiatry 55:10 (2014), pp 1068–1087 doi:10.1111/jcpp.12295

Research Review: Polygenic methods and their application to psychiatric traits

Naomi R. Wray,[1] Sang Hong Lee,[1] Divya Mehta,[1] Anna A.E. Vinkhuyzen,[1] Frank Dudbridge,[2] and Christel M. Middeldorp[3,4]

Prediction accuracy of a PRS in a twin sample

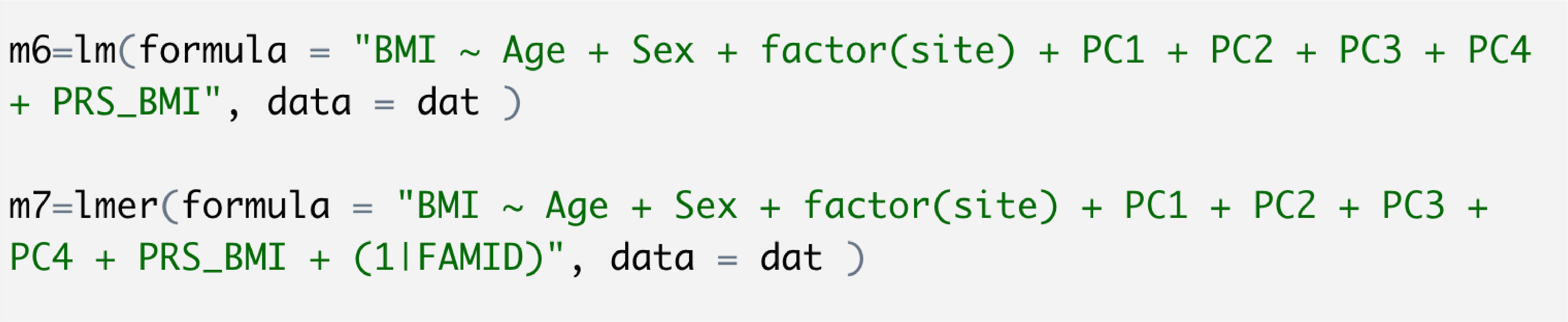
Y = Z a + c + e

c ~ N(0, CRM . sC2)

Z : age, sex, site, genetic PCs, PRS

Control for A and/or C

LMM

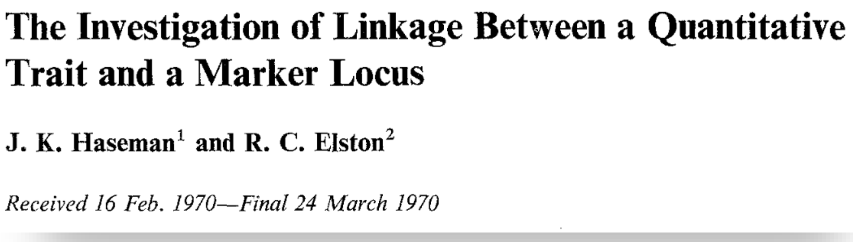


```
m6=lm(formula = "BMI ~ Age + Sex + factor(site) + PC1 + PC2 + PC3 + PC4
+ PRS_BMI", data = dat )

m7=lmer(formula = "BMI ~ Age + Sex + factor(site) + PC1 + PC2 + PC3 +
PC4 + PRS_BMI + (1|FAMID)", data = dat )
```

Haseman-Elston regression

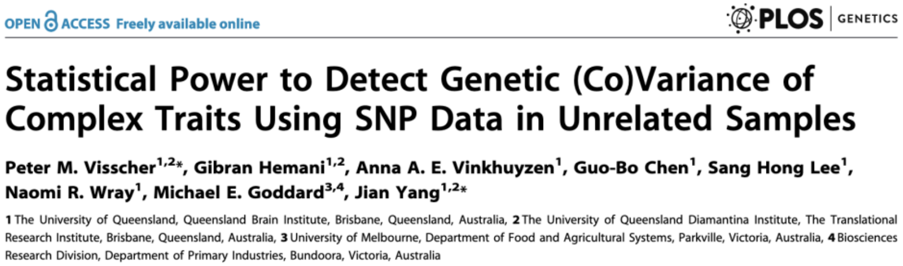
Linear



The Investigation of Linkage Between a Quantitative Trait and a Marker Locus

J. K. Haseman[1] and R. C. Elston[2]

Received 16 Feb. 1970—Final 24 March 1970

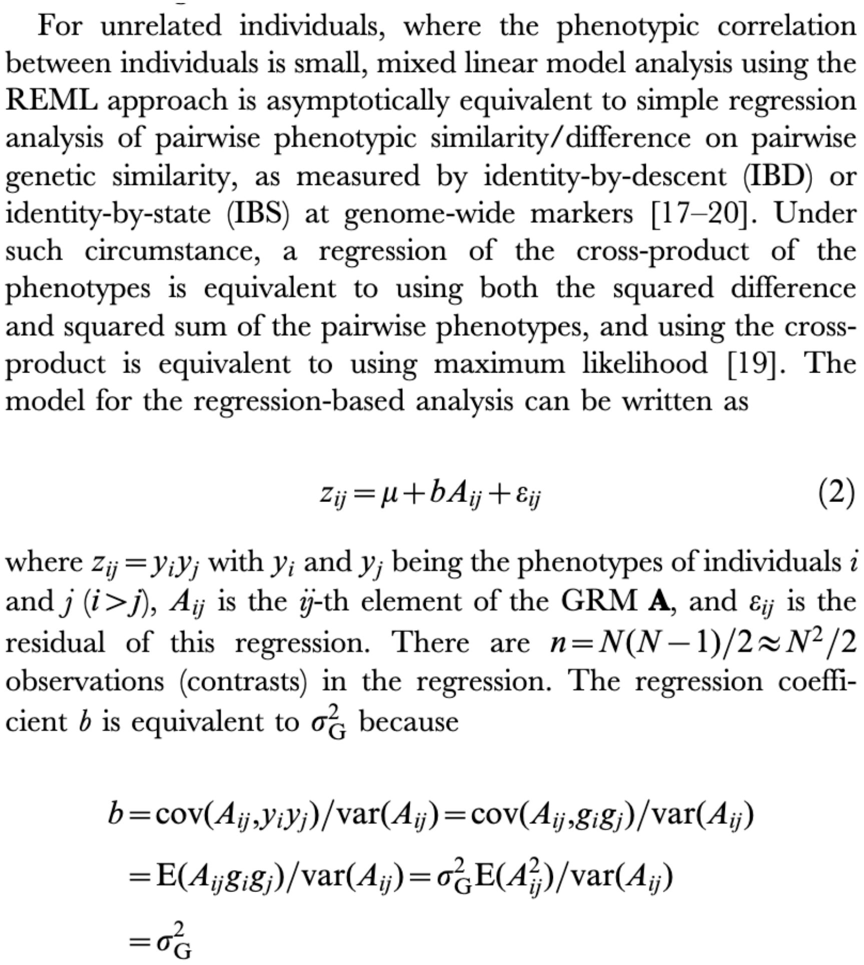


OPEN ACCESS Freely available online PLOS | GENETICS

Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples

Peter M. Visscher[1,2]*, Gibran Hemani[1,2], Anna A. E. Vinkhuyzen[1], Guo-Bo Chen[1], Sang Hong Lee[1], Naomi R. Wray[1], Michael E. Goddard[3,4], Jian Yang[1,2]*

1 The University of Queensland, Queensland Brain Institute, Brisbane, Queensland, Australia, 2 The University of Queensland Diamantina Institute, The Translational Research Institute, Brisbane, Queensland, Australia, 3 University of Melbourne, Department of Food and Agricultural Systems, Parkville, Victoria, Australia, 4 Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria, Australia



For unrelated individuals, where the phenotypic correlation between individuals is small, mixed linear model analysis using the REML approach is asymptotically equivalent to simple regression analysis of pairwise phenotypic similarity/difference on pairwise genetic similarity, as measured by identity-by-descent (IBD) or identity-by-state (IBS) at genome-wide markers [17–20]. Under such circumstance, a regression of the cross-product of the phenotypes is equivalent to using both the squared difference and squared sum of the pairwise phenotypes, and using the cross-product is equivalent to using maximum likelihood [19]. The model for the regression-based analysis can be written as

$$z_{ij} = \mu + bA_{ij} + \varepsilon_{ij} \qquad (2)$$

where $z_{ij} = y_i y_j$ with y_i and y_j being the phenotypes of individuals i and j $(i > j)$, A_{ij} is the ij-th element of the GRM $\mathbf{A}$, and ε_{ij} is the residual of this regression. There are $n = N(N-1)/2 \approx N^2/2$ observations (contrasts) in the regression. The regression coefficient b is equivalent to σ_G^2 because

$$\begin{aligned} b &= \mathrm{cov}(A_{ij}, y_i y_j)/\mathrm{var}(A_{ij}) = \mathrm{cov}(A_{ij}, g_i g_j)/\mathrm{var}(A_{ij}) \\ &= \mathrm{E}(A_{ij} g_i g_j)/\mathrm{var}(A_{ij}) = \sigma_G^2 \mathrm{E}(A_{ij}^2)/\mathrm{var}(A_{ij}) \\ &= \sigma_G^2 \end{aligned}$$

N = 1,000

n = 499,500

LMM

Longitudinal models

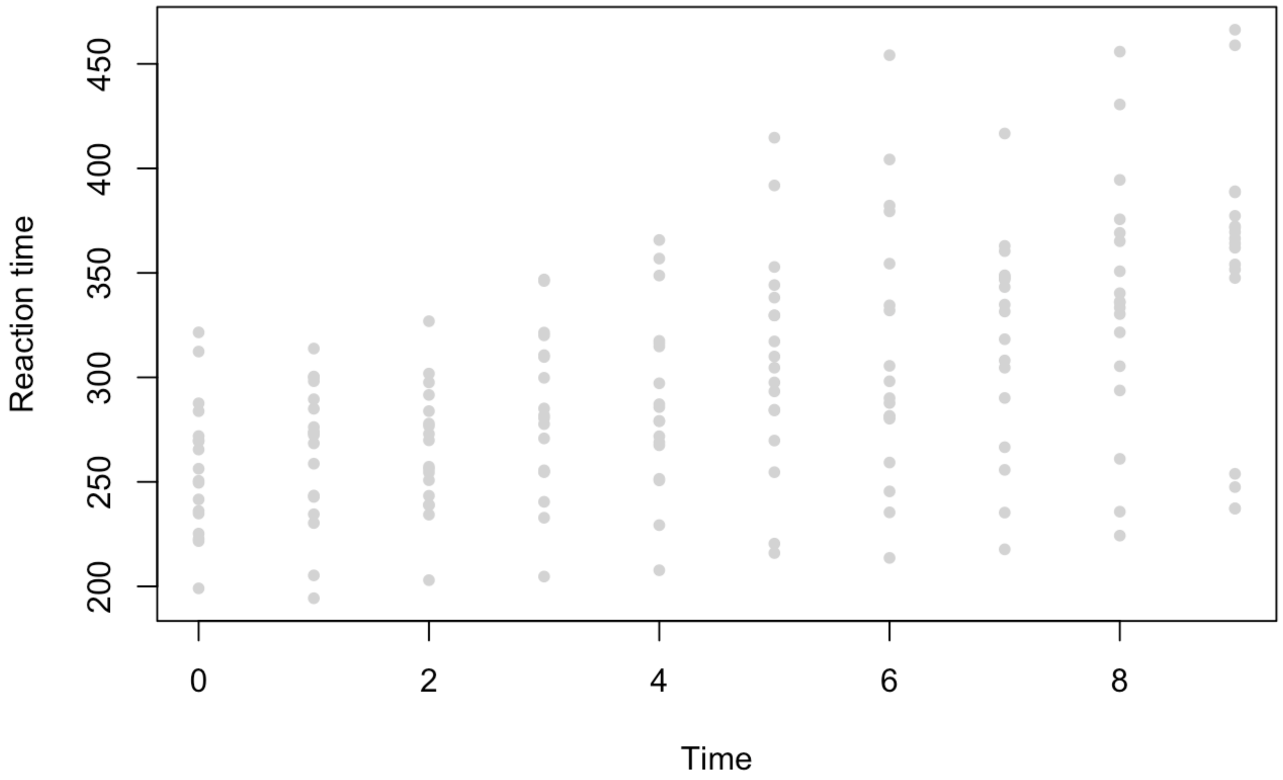
Y = Z a + t + e

t ~ N(0, TRM . sG2)

Z : age, sex, site

TRM: matrix of visit/wave – identifies observations of the same individual

https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf
http://www.bristol.ac.uk/cmm/learning/videos/random-slopes.html



LMM

Longitudinal models

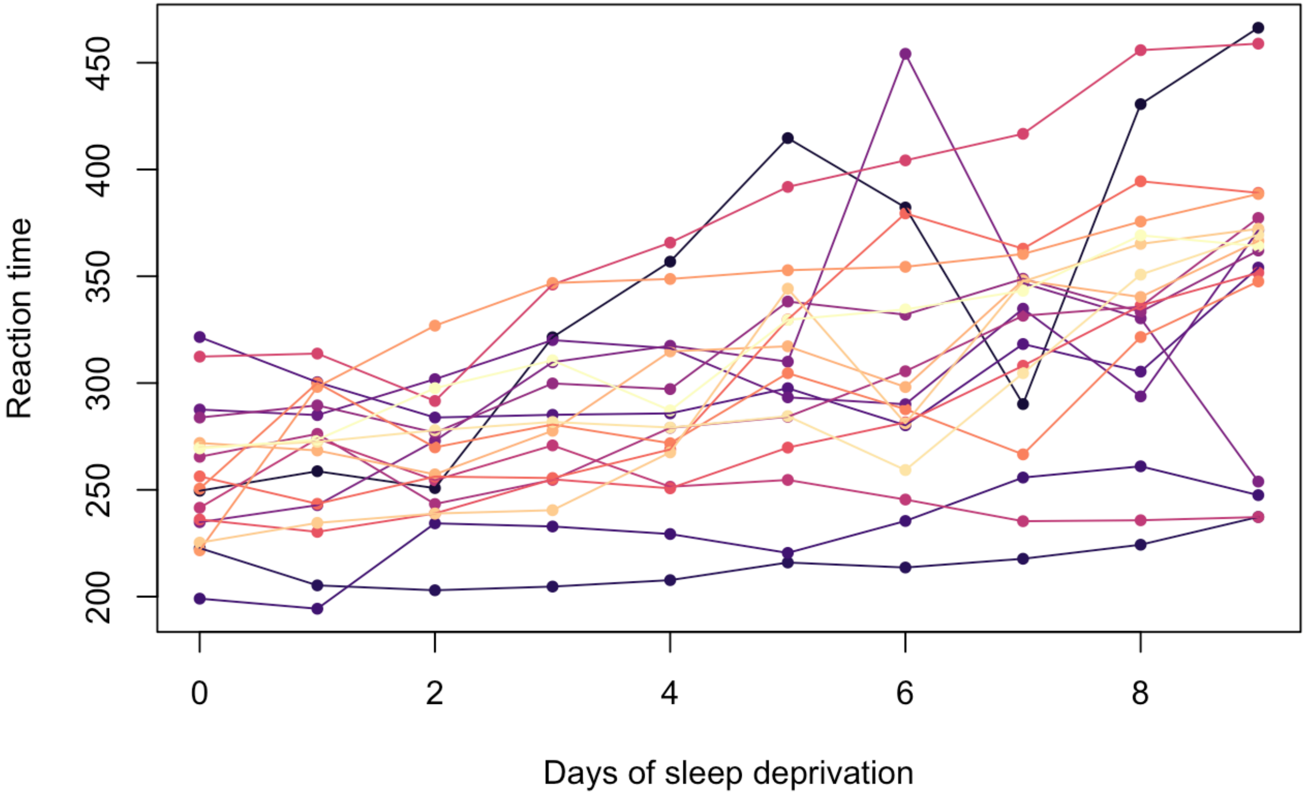
Y = **Z a** + **t** + **e**

t ~ N(0, TRM . sG2)

Z : age, sex, site

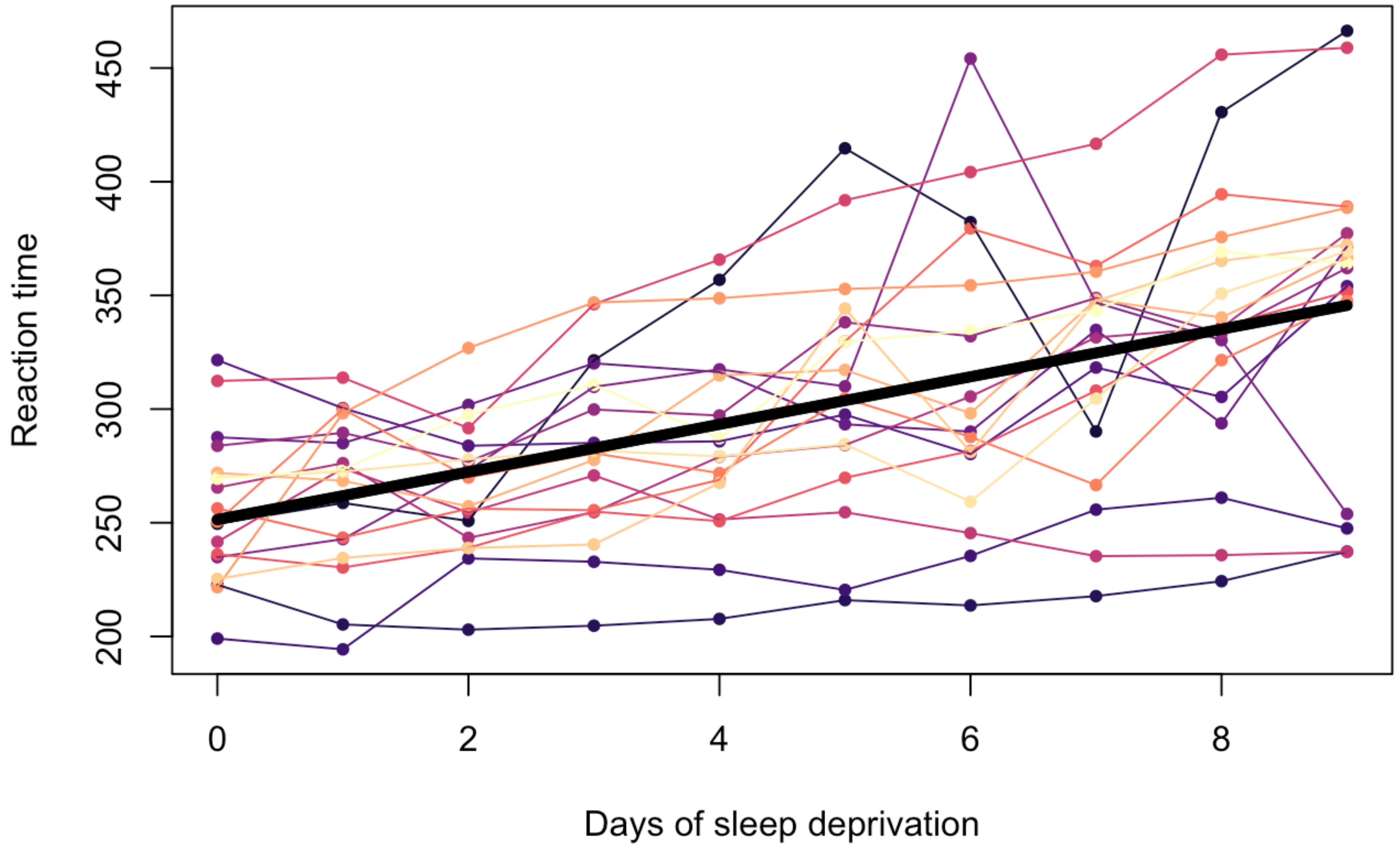
TRM: matrix of visit/wave – identifies observations of the same individual

https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf
http://www.bristol.ac.uk/cmm/learning/videos/random-slopes.html



Fixed slope and intercept model

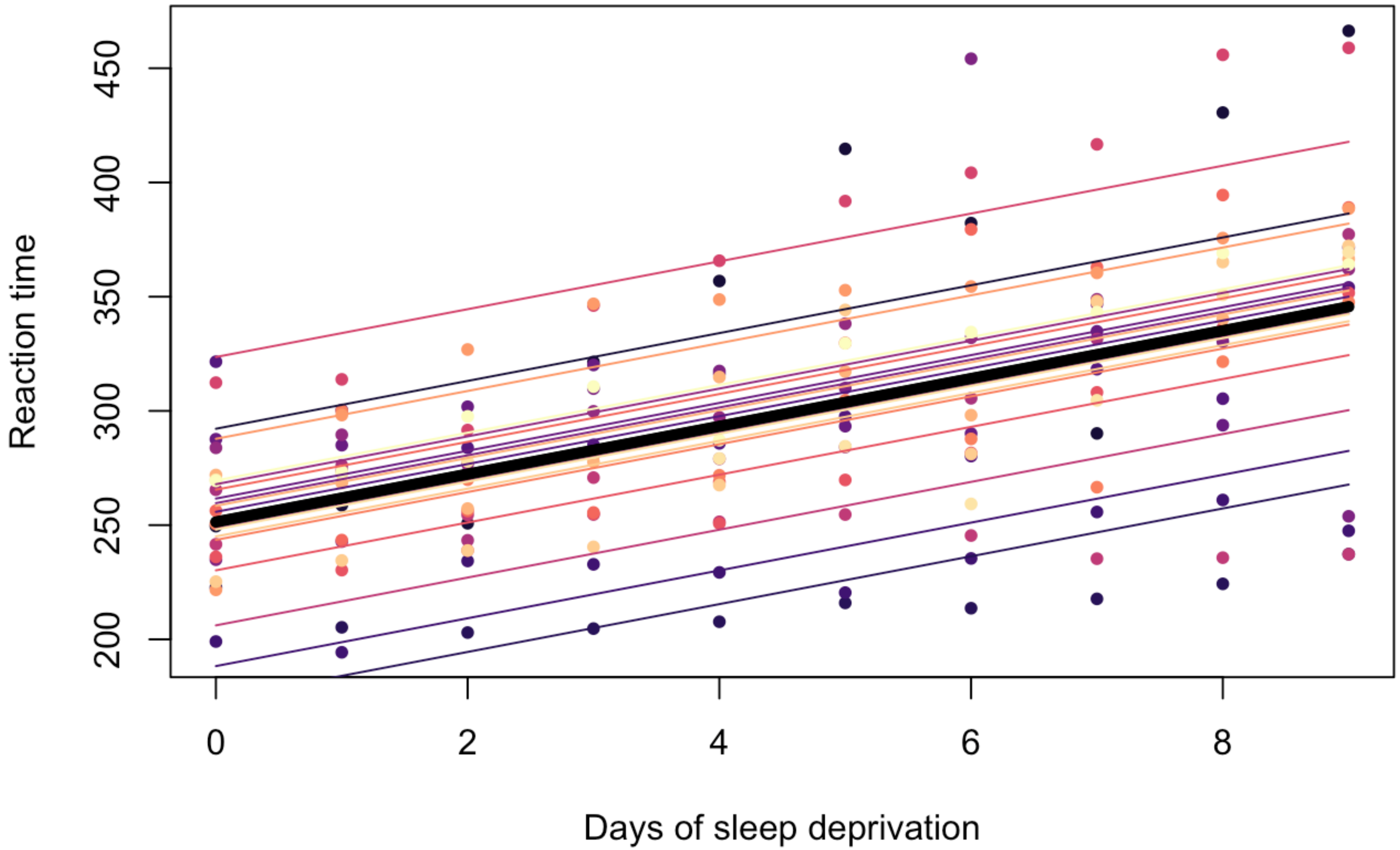
- **Y** = **T** b + **e**
- **T**: days of sleep deprivation
- **GLMs**



```
m00=lm("Reaction ~ Days ", data = sleepstudy)
```

Mixed intercept model

- **Y** = **T** b + **t0** + **e**
- **T**: days of sleep deprivation
- **U0 ~ N(0, S)**
- **LMM**



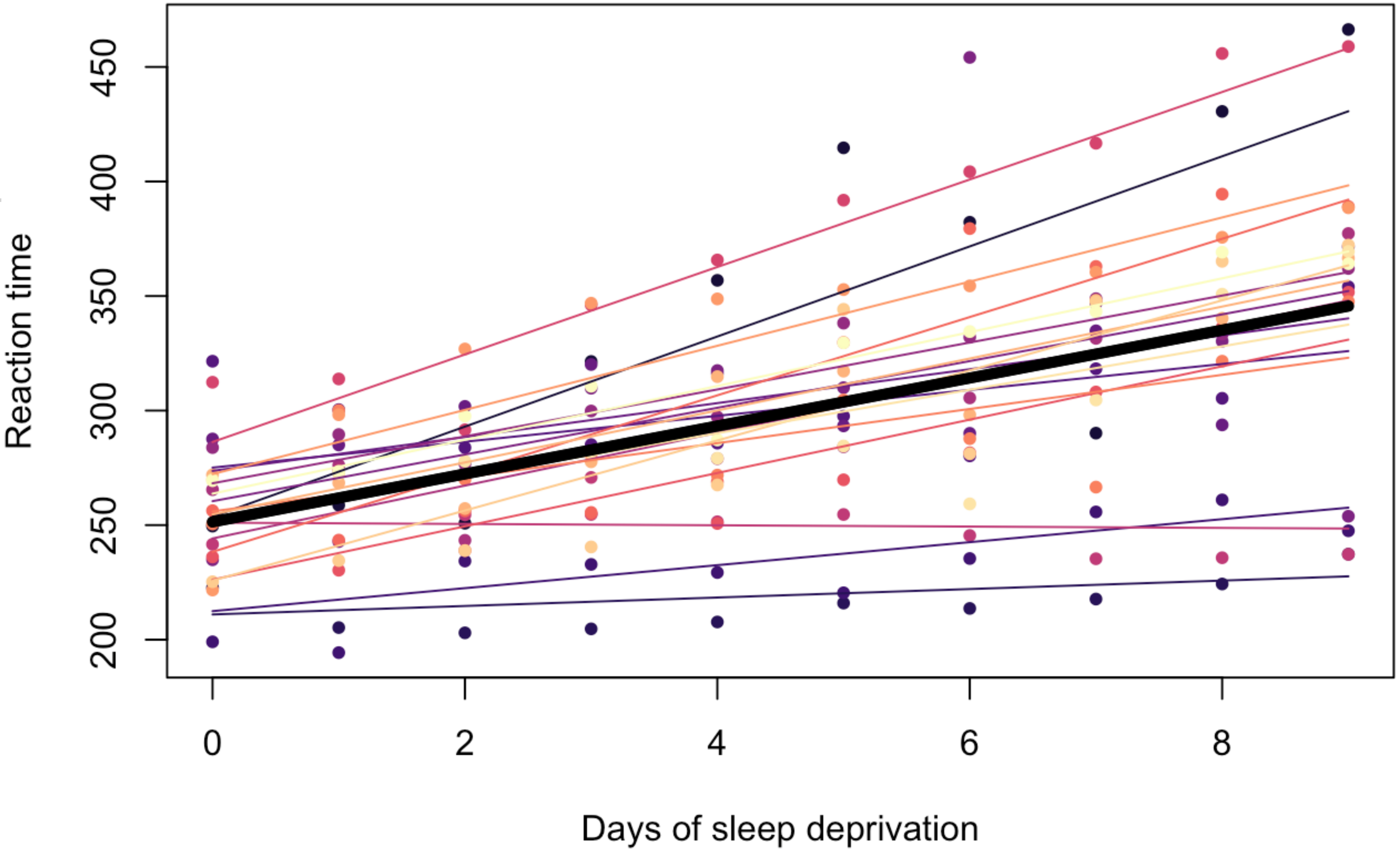
```
m01=lmer("Reaction ~ Days + (1|Subject) ", data = sleepstudy)
```

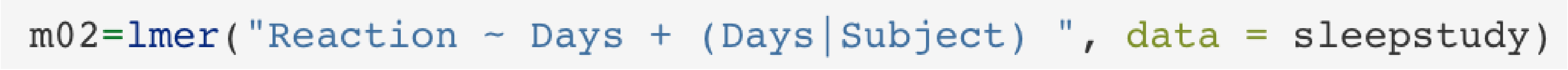
Mixed slope and intercept model

Y = **T .** b + **u0** + **T . u1** + **e**

T: days of sleep deprivation

u0,1 ~ N(0, S); u0 and u1 correlated





```
m02=lmer("Reaction ~ Days + (Days|Subject) ", data = sleepstudy)
```

SEM – multivariate mixed effects

- Set of linear mixed models
- OpenMx

SEM

P1 = A1 + E1

P2 = A1 + E1 + A2 + E2

P3 = ...

A1 , A2 ~ N(0, GRM . sAi)

E1, E2 ~ N(0, I . sEi)

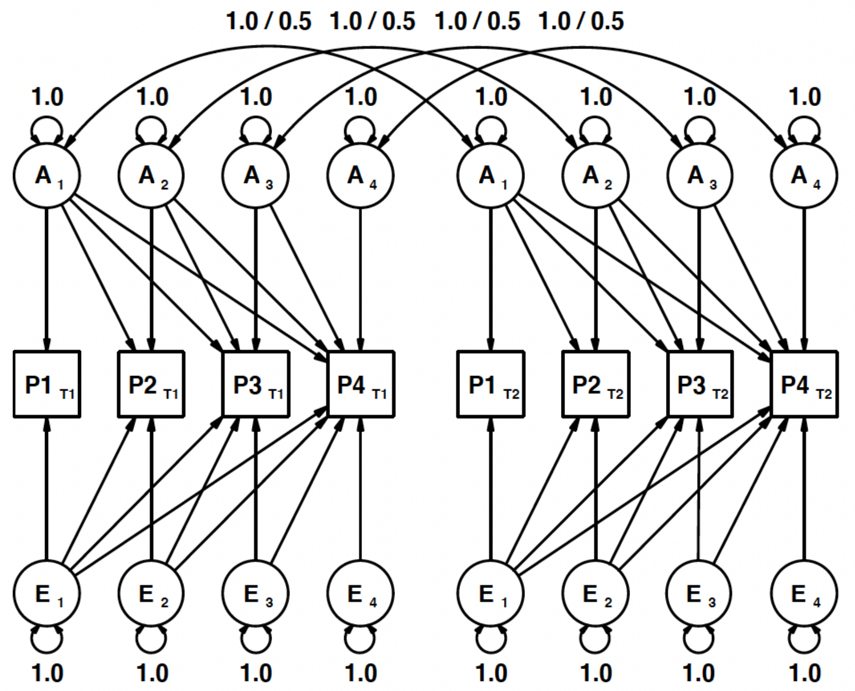


Figure 10.2: Phenotypic Cholesky decomposition model for four variables. All labels for path-coefficients have been omitted.

From Methodology for Genetic Studies of Twins and Families, Neale et al., 1992.

Model	Formula	Statistics of interest	Application	In R

Generalised Linear model	**Y** = **Xb** + **e**	b, correlation
		Odd Ratio, r2
		Partial corretations
			Test association
			(e.g. Prediction accuracy in pop sample)
			Haseman Elston
				Lm()
				Glm()

Random effect model	**Y = Xb + e**
	b~N(0, I. sG2 / 2)
		sG, variance component (R2)	AE or ACE model
			Longitudinal model
			Model site effect
				Lme4()
				openMx()
				heritability()
				qgg()

Linear Mixed Model	**Y = Za + Xb + e**
	b~N(0, I . sG2 / p)
		a : partial & multiple correlations, R2
		sG: variance components (R2)
			ACE with covariates
			SNP h2 with covariates
			Longitudinal with covariates
			Quadratic, interactions
			More…
				Lme4()
				nlme()
				openMx()
				Umx()
				heritability()
				qgg()

Structural equation modelling	Set of GLM or LMM	A :partial correlations, sG : variance components, rG and latent factor model	Complex multivariate LMMs models
			rG
			Comnnon pathway / independent pathway
				lavaan()
				openMx()

Summary

The different models we use a lot of the time share some common theory and concepts (e.g. likelihood)

Random effects can accommodate low and high-dimensional data, observed or latent variables (with known covariance).

Hoping that seeing models in perspective – may give new ideas on how to approach research questions