Statistical Modeling 2

Linear models in genomics

Linear models

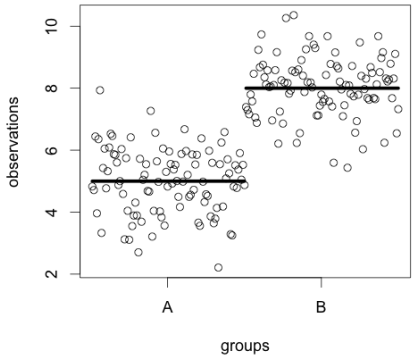
t-test

linear model

[RI]

Two groups (t-test)

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

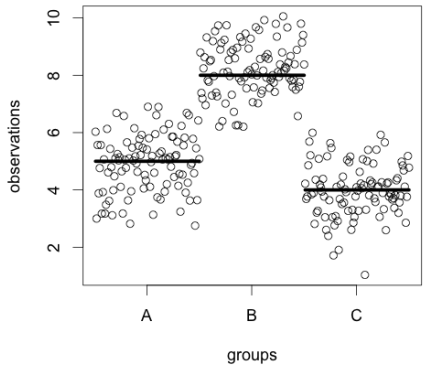


$$Y = X\beta + \epsilon$$

$$\longrightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{and} \quad \text{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Three groups

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$



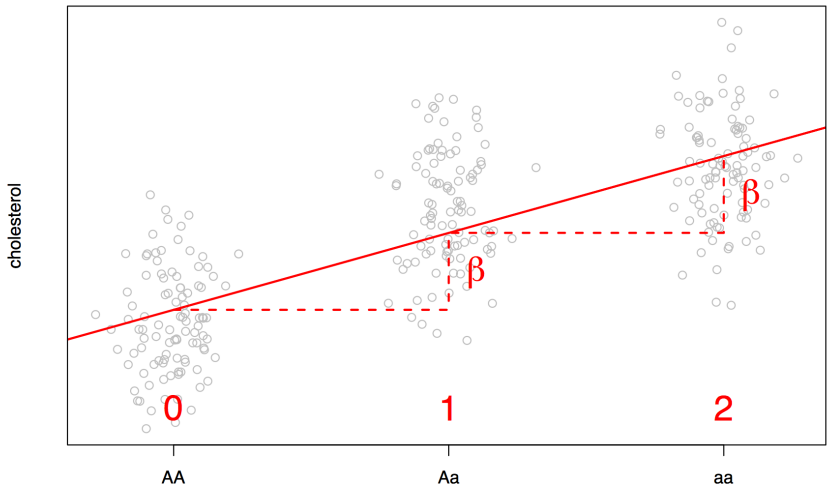
$$Y = X\beta + \epsilon$$

$$\longrightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{and} \quad \text{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Linear regression with SNPs

Many analyses fit the 'additive model'

$$y = \beta_0 + \beta \times \#\text{minor alleles}$$

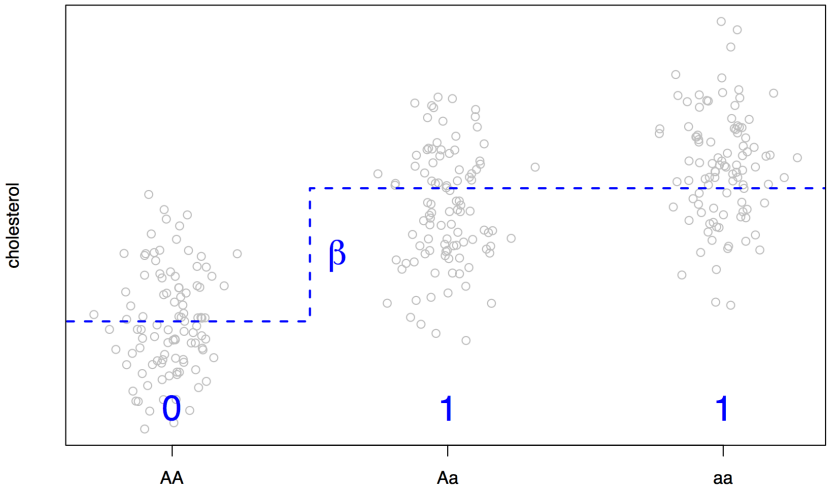


[Thomas Lumley, Ken Rice]

Linear regression with SNPs

An alternative is the 'dominant model';

$$y = \beta_0 + \beta \times (G \neq AA)$$

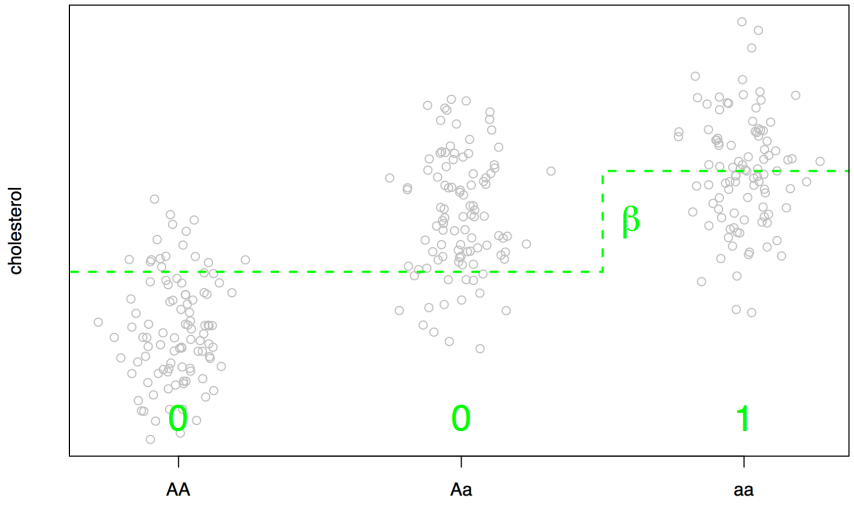


[Thomas Lumley, Ken Rice]

Linear regression with SNPs

or the 'recessive model';

$$y = \beta_0 + \beta \times (G == AA)$$

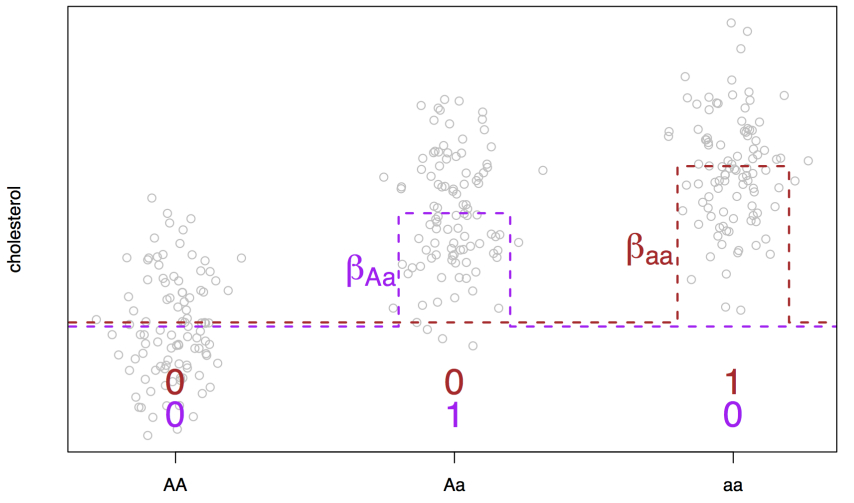


[Thomas Lumley, Ken Rice]

Linear regression with SNPs

Finally, the 'two degrees of freedom model';

$$y = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$



[Thomas Lumley, Ken Rice]

Linear regression with SNPs

TESTS FOR LINEAR TRENDS IN PROPORTIONS AND FREQUENCIES

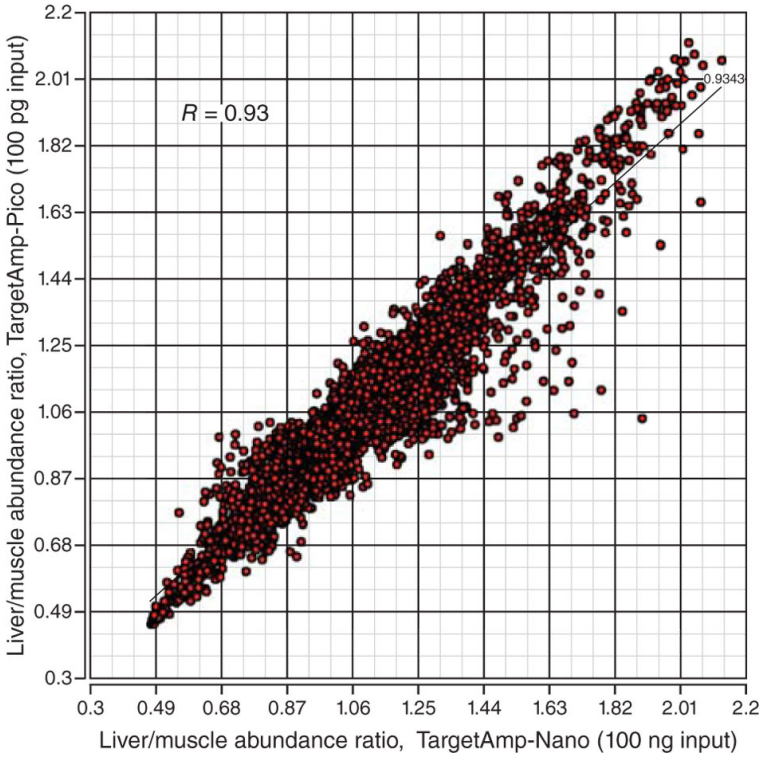
P. ARMITAGE

Statistical Research Unit of the Medical Research Council,
London School of Hygiene and Tropical Medicine

1. *Introduction*

One frequently encounters data consisting of a series of proportions, occurring in groups which fall into some natural order. The question usually asked is then not so much whether the proportions differ significantly, but whether they show a significant trend, upwards or downwards, with the ordering of the groups. In the data shown in Table 1, for instance, the usual test for a 2×3 contingency table yields a χ^2 equal to 7.89 on 2 degrees of freedom, corresponding to a probability of about 0.02.

Source: *Biometrics*, Vol. 11, No. 3 (Sep., 1955), pp. 375-386
Published by: International Biometric Society
Stable URL: http://www.jstor.org/stable/3001775



Jim Pease Nat Meth (advertisement)

Correlation and regression

- In a correlation setting we try to determine whether two random variables vary together (covary).
- There is no ordering between those variables, and we do not try to explain one of the variables as a function of the other.

- In regression settings we describe the dependence of one variable on the other variable.
- There is an ordering of the variables, often called the dependent variable and the independent variable.

Correlation

Let X and Y be random variables with

$$\mu_X = E(X),\ \mu_Y = E(Y),\ \sigma_X = SD(X),\ \sigma_Y = SD(Y)$$

Covariance	Correlation
$cov(X,Y) = E\{(X - \mu_X)\ (Y - \mu_Y)\}$	$cor(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$
$\longrightarrow$ cov(X,Y) can be any real number	$\longrightarrow\ -1 \leq cor(X,Y) \leq 1$

Correlation

Consider n pairs of data: $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$

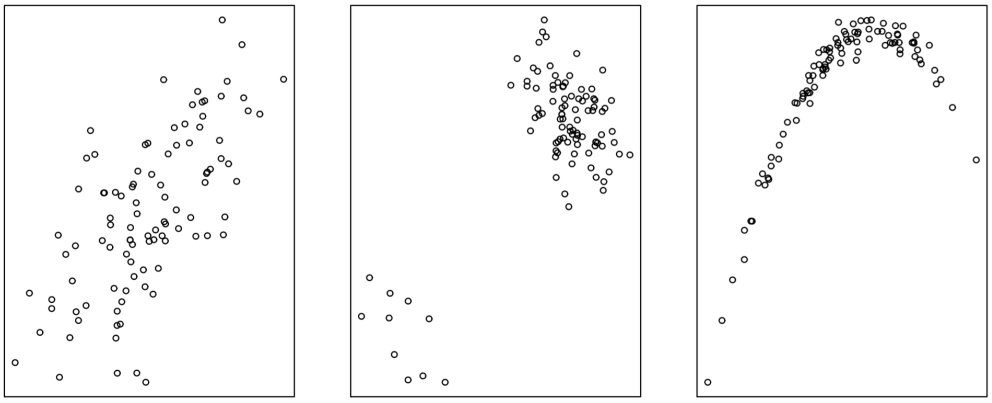
We consider these as independent draws from some bivariate distribution.

We estimate the correlation in the underlying distribution by:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

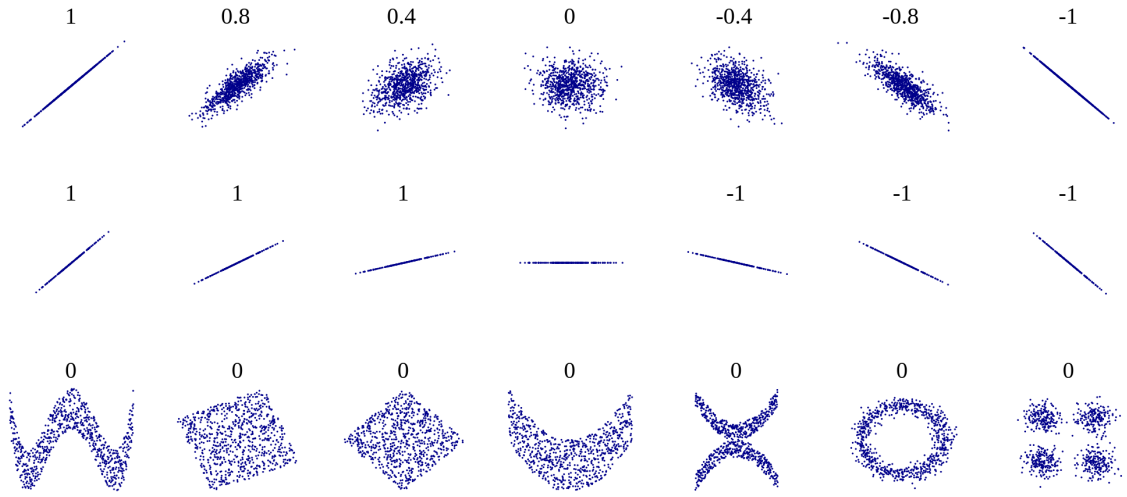
This is sometimes called the correlation coefficient.

Correlation measures linear dependency

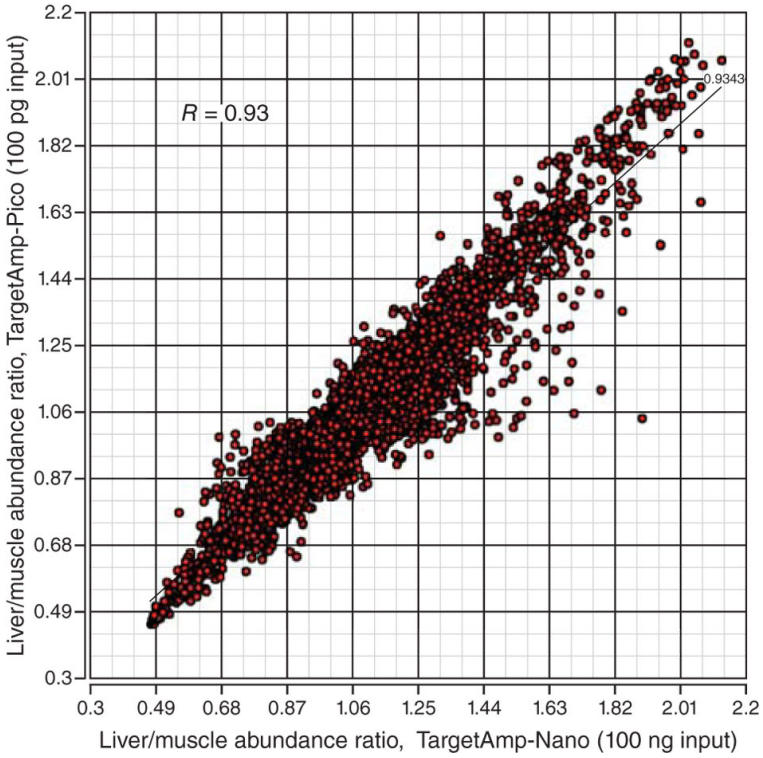


$\longrightarrow$ All three plots have correlation ≈ 0.7!

Correlation measures linear dependency



www.wikipedia.org



Jim Pease Nat Meth (advertisement)

The correlation coefficient of two jointly distributed random variables X and Y is defined as

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\text{cov}(X,Y)$ is the covariance between X and Y, and σ_X and σ_Y are their respective standard deviations.

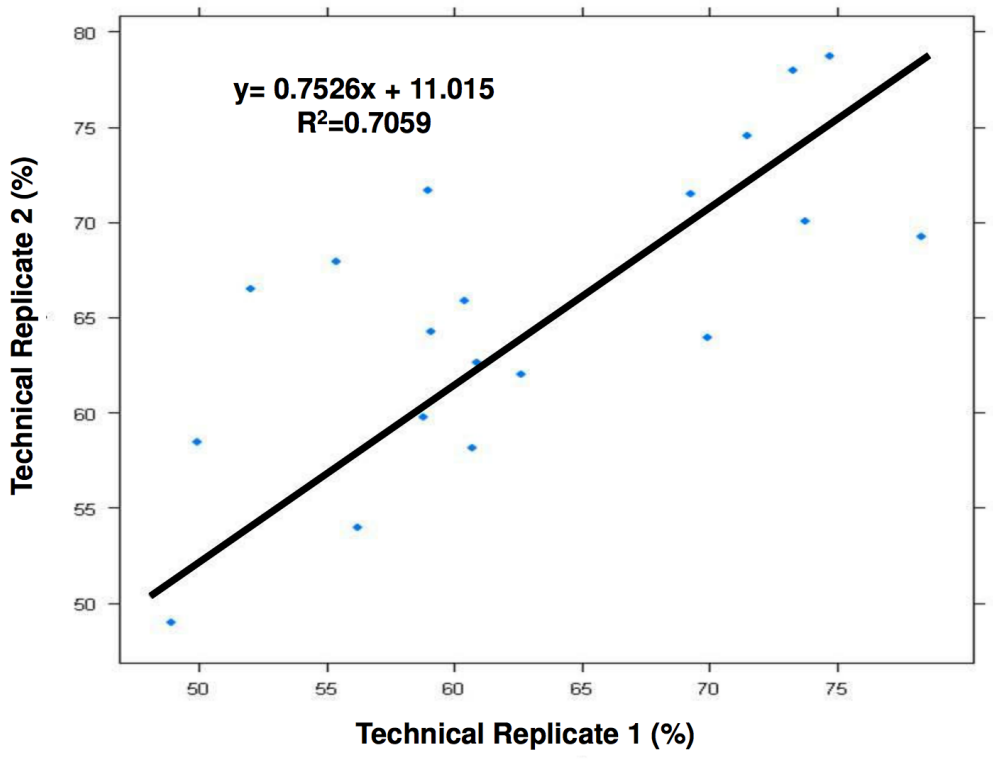
If X and Y follow a bivariate normal distribution with correlation ρ

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right)$$

then

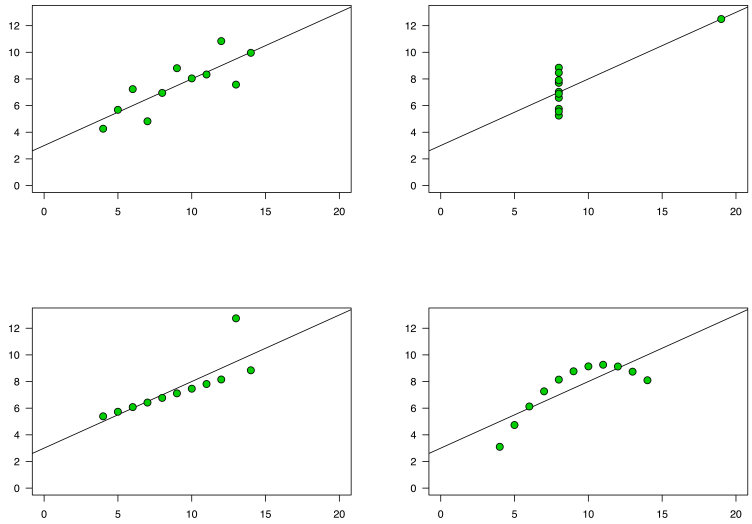
$$y_i | x_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

where $\beta_0 = \mu_Y - \beta_1 \mu_X$, $\beta_1 = \rho\sigma_Y/\sigma_X$, and $\sigma^2 = \sigma_Y^2(1-\rho^2)$.



PMID 20021669

R^2 does not assess whether the model fits



$\hat{\beta}_0=3.0$ $\hat{\beta}_1=0.5$ $\hat{\sigma}^2=13.75$ $R^2=0.667$

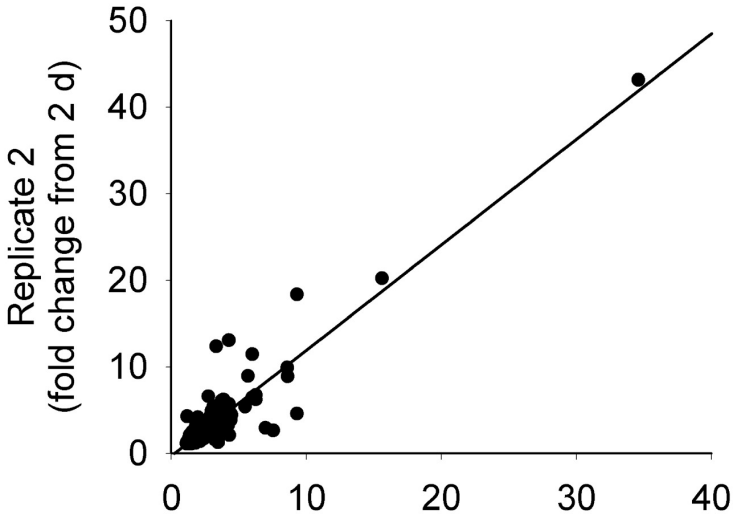
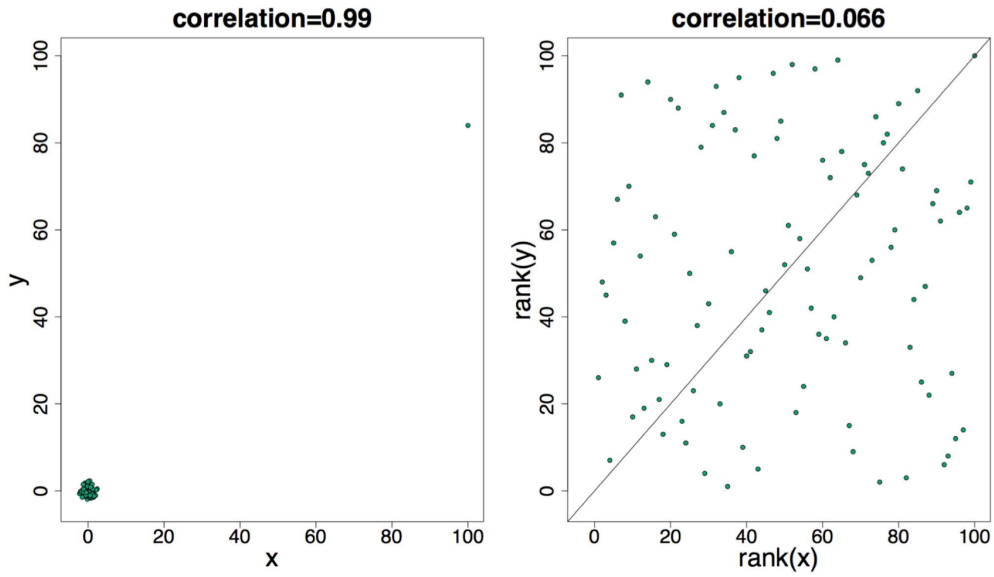
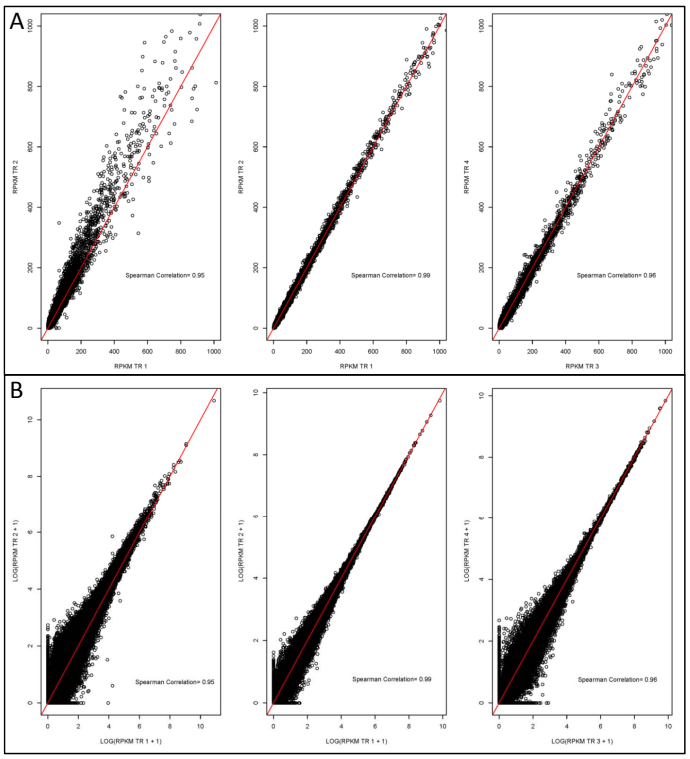


Fig. 1. Duplicate occurrences of genes are highly correlated. Seventy-four genes were found to be represented more than once in the list of differentially expressed genes we identified. A fold change relative to the 2-day value was determined for these genes at 8 and 15 days (higher/lower value to give a value >1), and these fold changes were compared across the replicate samples. A simple linear correlation was calculated for 71 of these genes. The regression line was defined by the following equation: *replicate 2* = 1.22(*replicate 1*) − 0.3; $r^2 = 0.84$.

PMID 12644633

Pearson and Spearman





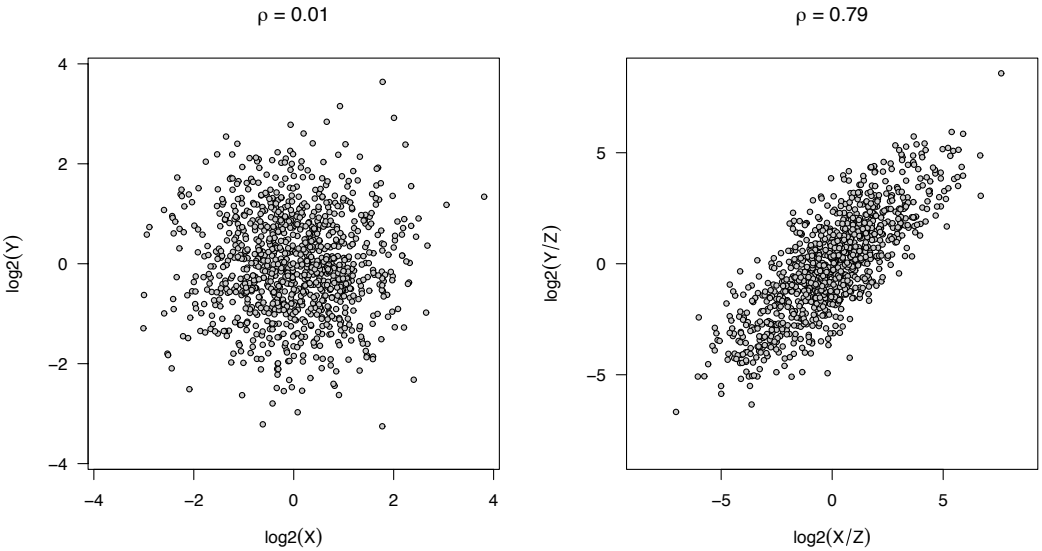
PMID 21645359

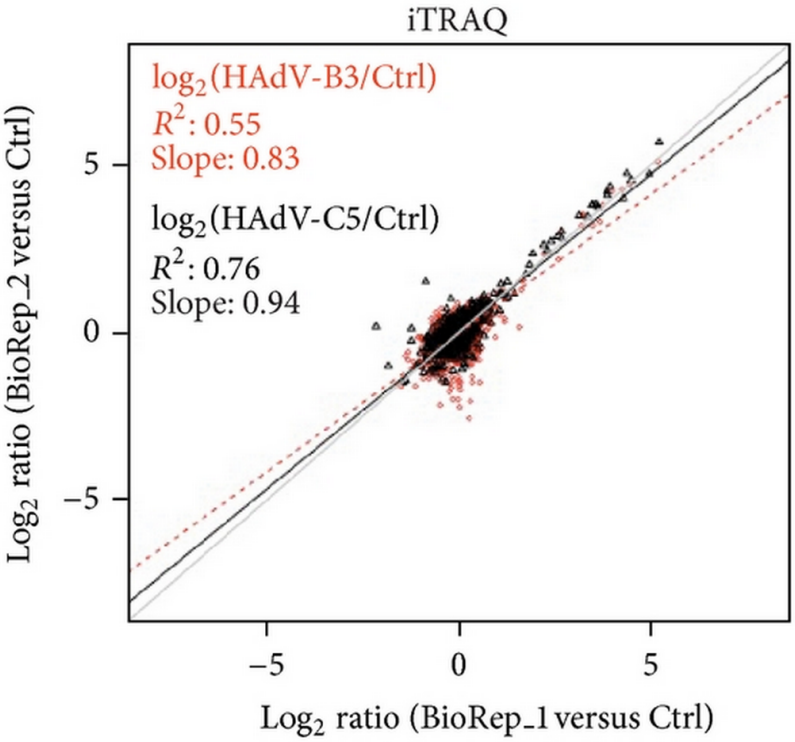
If you want to show that two sets of measurements are alike (such as gene expression from two technical replicates of the same sample) use the concordance correlation coefficient.

The concordance correlation between two random variables X and Y is defined as

$$\rho_{CC}(X,Y) = \frac{2 \times \mathrm{cov}(X,Y)}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}.$$

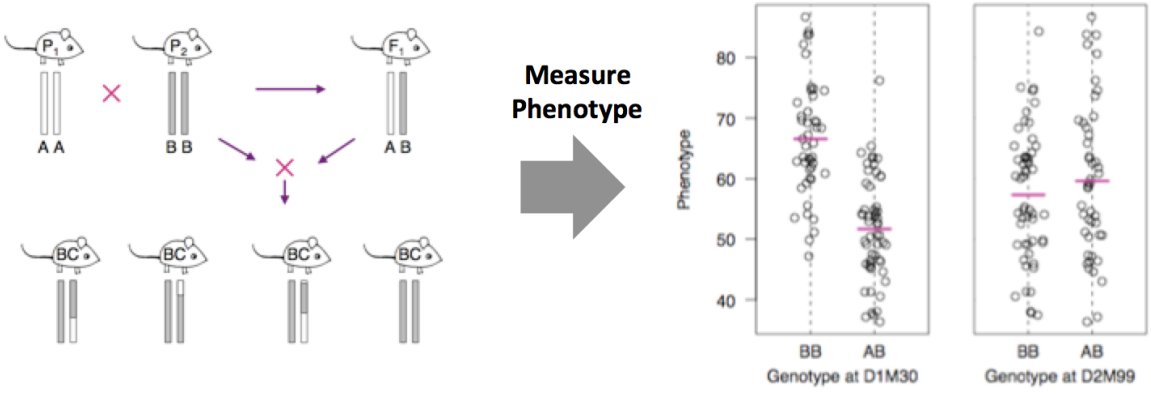
Unlike the Pearson correlation coefficient, the concordance correlation is not invariant to changes in location and scale, and assesses the actual agreement between X and Y, rather then their correlation alone.





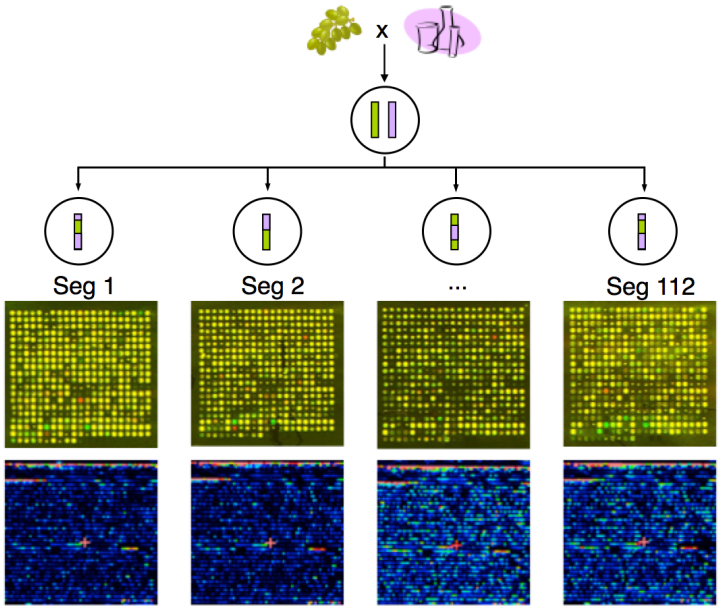
PMID 23555056

Tradtional QTL mapping



[KB]

eQTL mapping

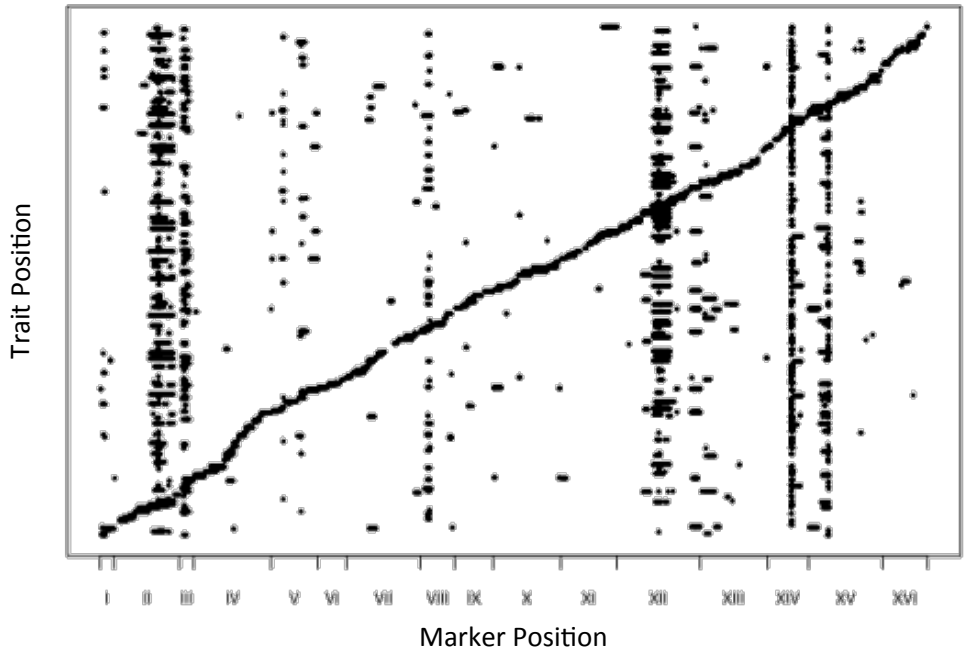


mRNA Expression
6,000 phenotypes

Oligo Genotyping
3,000 markers

PMID 11923494

eQTL mapping



eQTL mapping

For eQTL mapping (gene g and marker m):

$$Y_g = X_m \beta_{\text{eQTL}} + \epsilon_{\text{eQTL}}$$

$$\hat{\beta}_{\text{eQTL}} = (X_m^T X_m)^{-1} X_m^T Y_g \quad \text{and} \quad \text{cov}(\hat{\beta}_{\text{eQTL}}) = \sigma^2_{\text{eQTL}} (X_m^T X_m)^{-1}$$

For rapid computations in eQTL mapping, store the terms

$$(X_m^T X_m)^{-1} X_m^T \quad \text{and} \quad (X_m^T X_m)^{-1}.$$

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 10 2012, pages 1353–1358
doi:10.1093/bioinformatics/bts163

Gene expression Advance Access publication April 6, 2012

Matrix eQTL: ultra fast eQTL analysis via large matrix operations

Andrey A. Shabalin
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Expression quantitative trait loci (eQTL) analysis links variations in gene expression levels to genotypes. For modern datasets, eQTL analysis is a computationally intensive task as it involves testing for association of billions of transcript-SNP (single-nucleotide polymorphism) pair. The heavy computational burden makes eQTL analysis less popular and sometimes forces analysts to restrict their attention to just a small subset of transcript-SNP pairs. As more transcripts and SNPs get interrogated over a growing number of samples, the demand for faster tools for eQTL analysis grows stronger.

Results: We have developed a new software for computationally efficient eQTL analysis called Matrix eQTL. In tests on large datasets, it was 2–3 orders of magnitude faster than existing popular tools for QTL/eQTL analysis, while finding the same eQTLs. The fast performance is achieved by special preprocessing and expressing the most computationally intensive part of the algorithm in terms of large matrix operations. Matrix eQTL supports additive linear and ANOVA models with covariates, including models with correlated and heteroskedastic errors. The issue of multiple testing is addressed by calculating false discovery rate; this can be done separately for cis- and trans-eQTLs.

Availability: Matlab and R implementations are available for free at http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL

Contact: shabalin@email.unc.edu

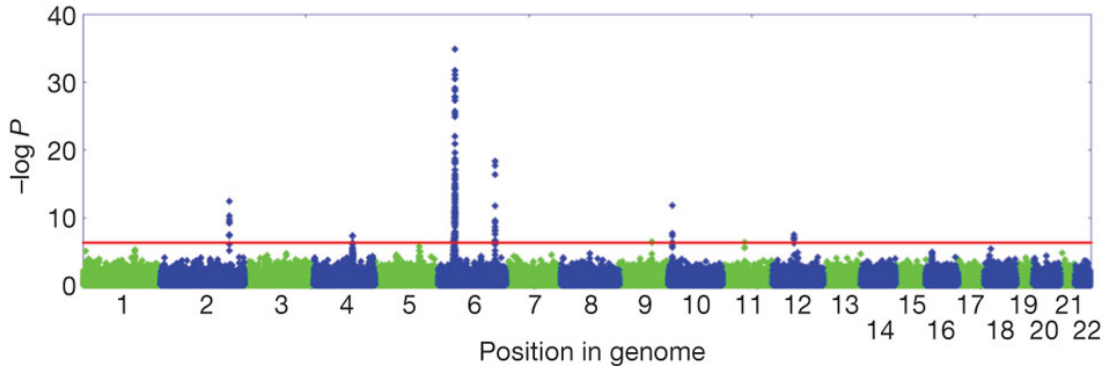
Table 1. Estimated performance of various eQTL software on the CF dataset

Method\No. of covariates	Zero	Ten	
Plink	9.4	583.3	days
Merlin	19.6	20.0	days
R/qtl (Revolution R)	1.0	4.7	days
snpMatrix	3.2	5.1	days
eMap	17.8	N/A	days
FastMap	10.3	N/A	**hours**
Matrix eQTL (Matlab)	11.8	11.8	**minutes**
Matrix eQTL (Revolution R)	14.6	14.6	**minutes**
Matrix eQTL (R, Goto BLAS)	19.4	19.4	**minutes**

The time for all methods is projected from tests on a dataset with 2201 genes and 57 333 SNPs. The timings projections for Matrix eQTL implementations were refined by applying them to the complete dataset.

PMID 22492648

GWAs permutation tests



For rapid permutations in a GWAs, store the terms

$(X_m^T X_m)^{-1} X_m^T$ and $(X_m^T X_m)^{-1}$.

Linear model theory

(and what it means)

Theorem: If $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}(= \mathbf{A}')$ and $\mathbf{B}$ are constant matrices, then $\mathbf{X}'\mathbf{A}\mathbf{X}$ and $\mathbf{B}\mathbf{X}$ are independently distributed iff $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$.

For normally distributed data, the sample mean and the sample variance are independent.

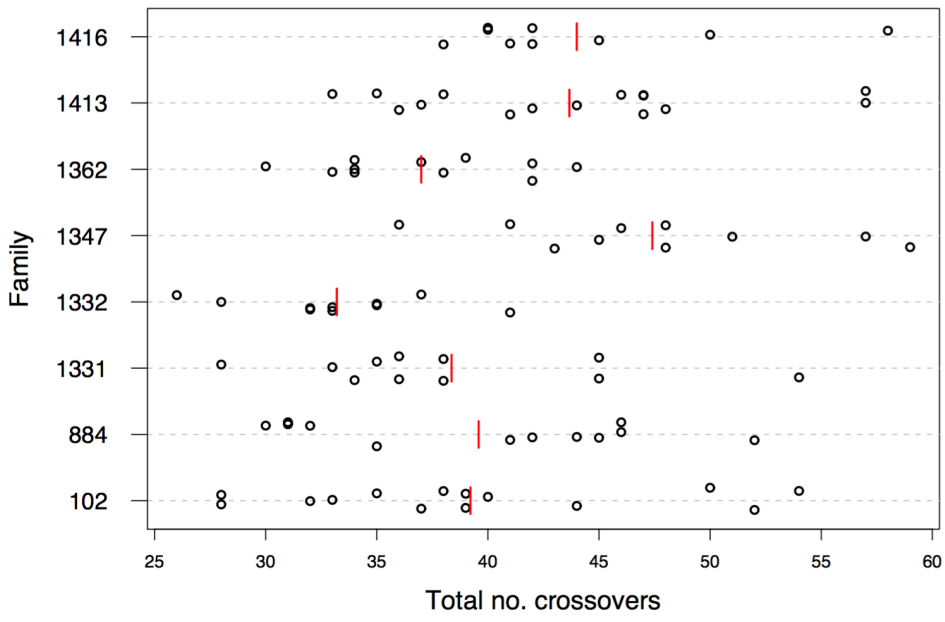
Theorem: If $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{C}_{p\times n}$ is a constant matrix of rank p, then $\mathbf{C}\mathbf{Y} \sim N_p(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$.

Linearly transformed normal data (including the sample mean) remain normal.

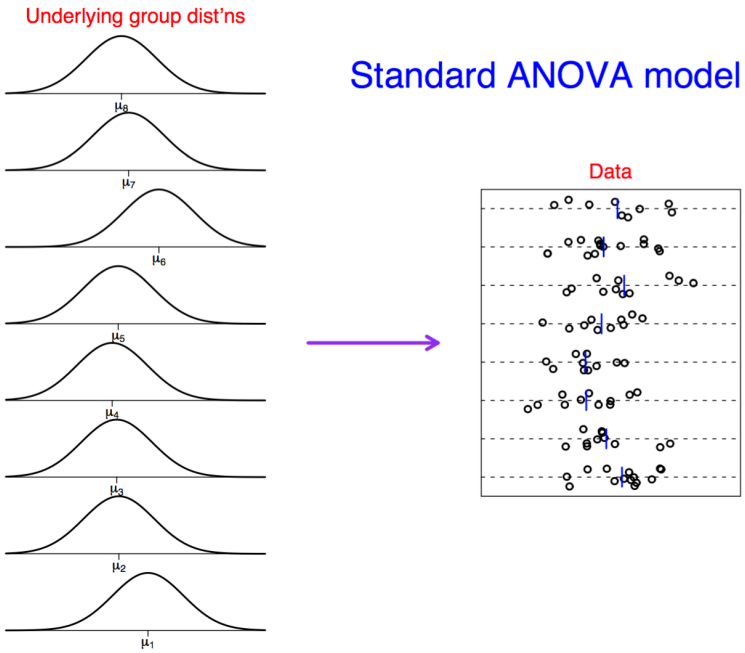
Effects of model violations

	Effect of Underfitting	Effect of Overfitting
$\hat{\boldsymbol{\beta}}$	biased	unbiased
$\hat{\mathbf{Y}}$	biased	unbiased
S^2	biased upward	unbiased
$\text{cov}(\hat{\boldsymbol{\beta}})$	still $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	$>$ than necessary

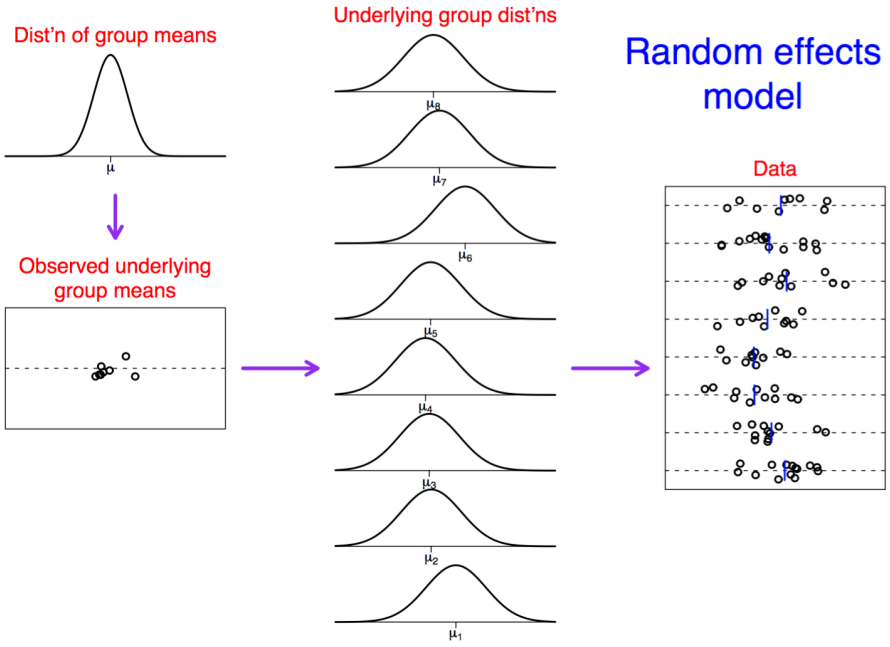
Random effects



Fixed effects



Random effects



GENOME-WIDE ASSOCIATION STUDIES

Meta-analysis methods for genome-wide association studies and beyond

Evangelos Evangelou[1] and John P. A. Ioannidis[2,3]

Abstract | Meta-analysis of genome-wide association studies (GWASs) has become a popular method for discovering genetic risk variants. Here, we overview both widely applied and newer statistical methods for GWAS meta-analysis, including issues of interpretation and assessment of sources of heterogeneity. We also discuss extensions of these meta-analysis methods to complex data. Where possible, we provide guidelines for researchers who are planning to use these methods. Furthermore, we address special issues that may arise for meta-analysis of sequencing data and rare variants. Finally, we discuss challenges and solutions surrounding the goals of making meta-analysis data publicly available and building powerful consortia.

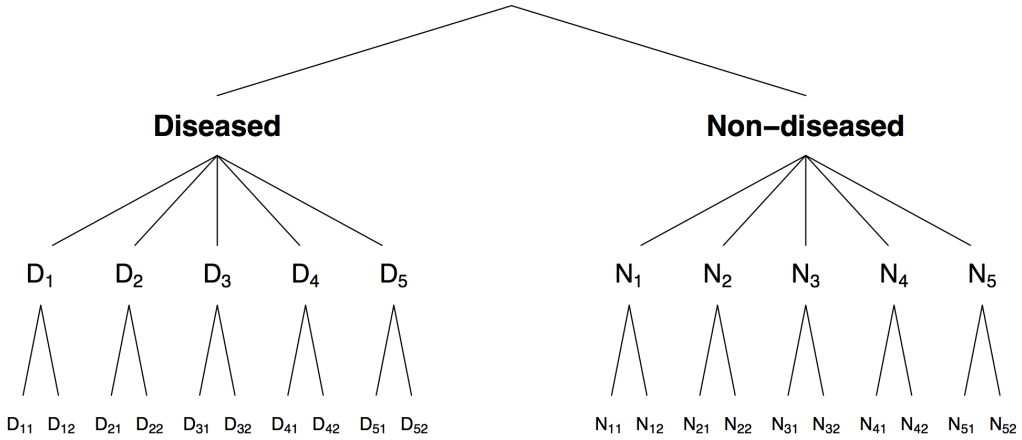
Table 2 | **Comparison of meta-analysis software packages**

	METAL	GWAMA	MetABEL	PLINK	R packages
Ability to process files from GWAS analysis tools; software used	No	Yes; SNPTEST, PLINK	Yes; ABEL	Yes; PLINK	No
Fixed effects implemented?	Yes	Yes	Yes	Yes	Yes
Random effects implemented?	No	Yes	No	No	Yes
Heterogeneity metrics generated	Q, I^2	Q, I^2	Q, I^2	Q, I^2	Q, I^2
Graphical illustration of meta-analysis results	No	Manhattan and QQ plots	Forest plots	No	Yes

GWAS, genome-wide association study.

PMID 23657481

Fixed and random effects



Analysis of variance

Nested ANOVA

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$$

Mixed effects model

α_i fixed; $\sum \alpha_i = 0$

$\beta_{ij} \sim \text{Normal}(0, \sigma^2_{B|A})$

$\epsilon_{ijk} \sim \text{Normal}(0, \sigma^2)$

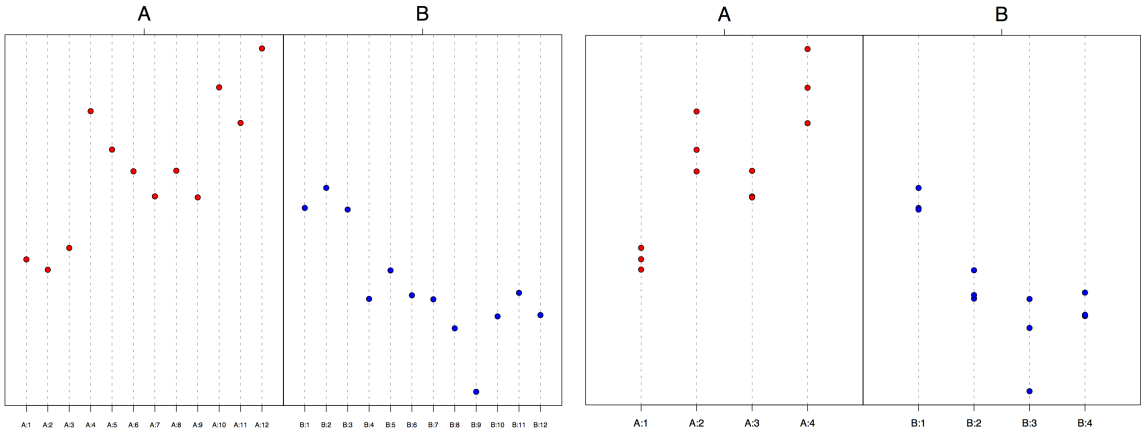
The expected mean squares are $\sigma^2 + n\,\sigma^2_{B|A} + n\,b\,\frac{\sum \alpha^2}{a-1}$

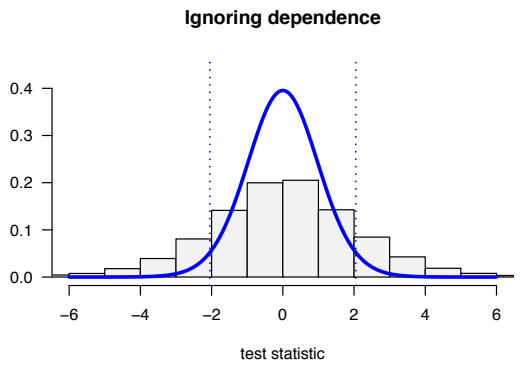
$\sigma^2 + n\,\sigma^2_{B|A}$

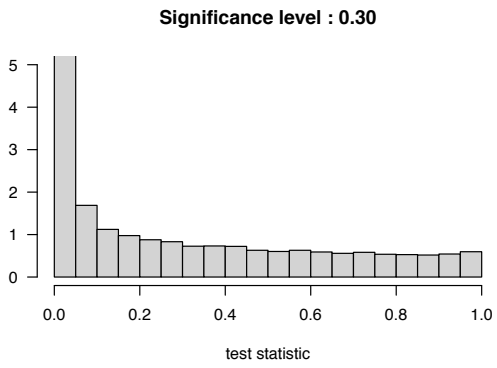
σ^2

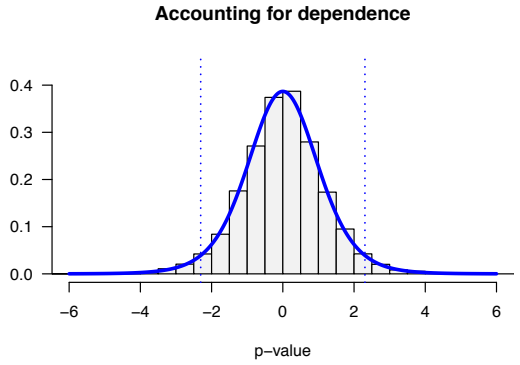
[140.615]

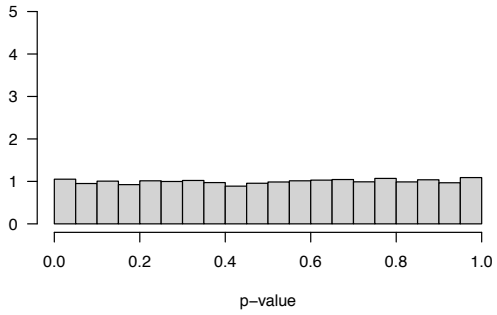
Technical replicates











5 biological replicates per group, with 3 technical replicates each.
Biological variability (SD) ten times larger than technical variability.