HOS 6236 Molecular Marker Assisted Plant Breeding

Fall 2017

Last Class:

GWAS - Example 1 on R

Today's Class:

GWAS

Another example – Population structure

	Sample 1 Americans		
	$\chi^2=0$	$p=1$	
	Use of Chopsticks		
A	Yes	No	Total
A_1	320	320	640
A_2	80	80	160
Total	400	400	800

Another example – Population structure

	Sample 2 Chinese		
	$\chi^2=0$	$p=1$	
	Use of Chopsticks		
A	Yes	No	Total
A_1	320	20	340
A_2	320	20	340
Total	640	40	680

Another example – Population structure

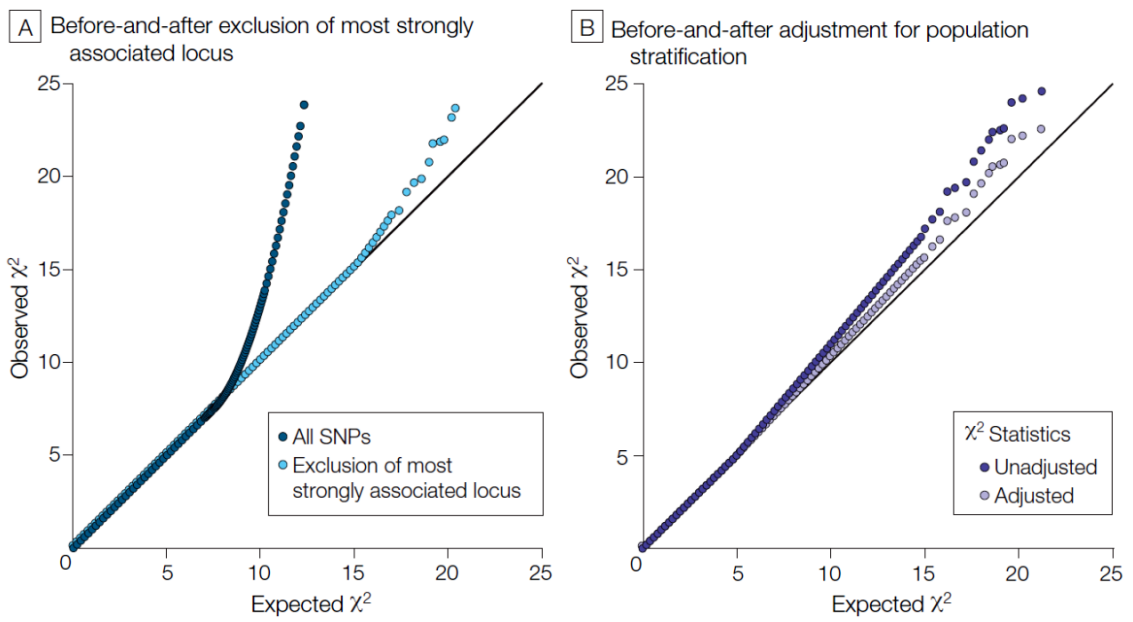
	Sample 3 Americans + Chinese		
	$\chi^2=34.2$	$p=4.9x10^{-9}$	
	Use of Chopsticks		
A	Yes	No	Total
A_1	640	340	980
A_2	400	100	500
Total	1040	440	1480

Quantile-Quantile Plots (Q-Q Plots)

- An essential tool for detecting the problems in a GWAS is a Quantile-Quantile (QQ) plot
- quantile - regular, equally spaced intervals of a random variable that divide the random variable into units of equal distribution
- A Quantile-Quantile (QQ) plot (in general) plots the observed quantiles of one distribution versus another OR plots the observed quantiles of a distribution versus the quantiles of the ideal distribution
- In GWAS we use a QQ plot to plot our the quantile distribution of observed p-values (on the y-axis) versus the quantile distribution of expected p-values (what distribution is this!?)

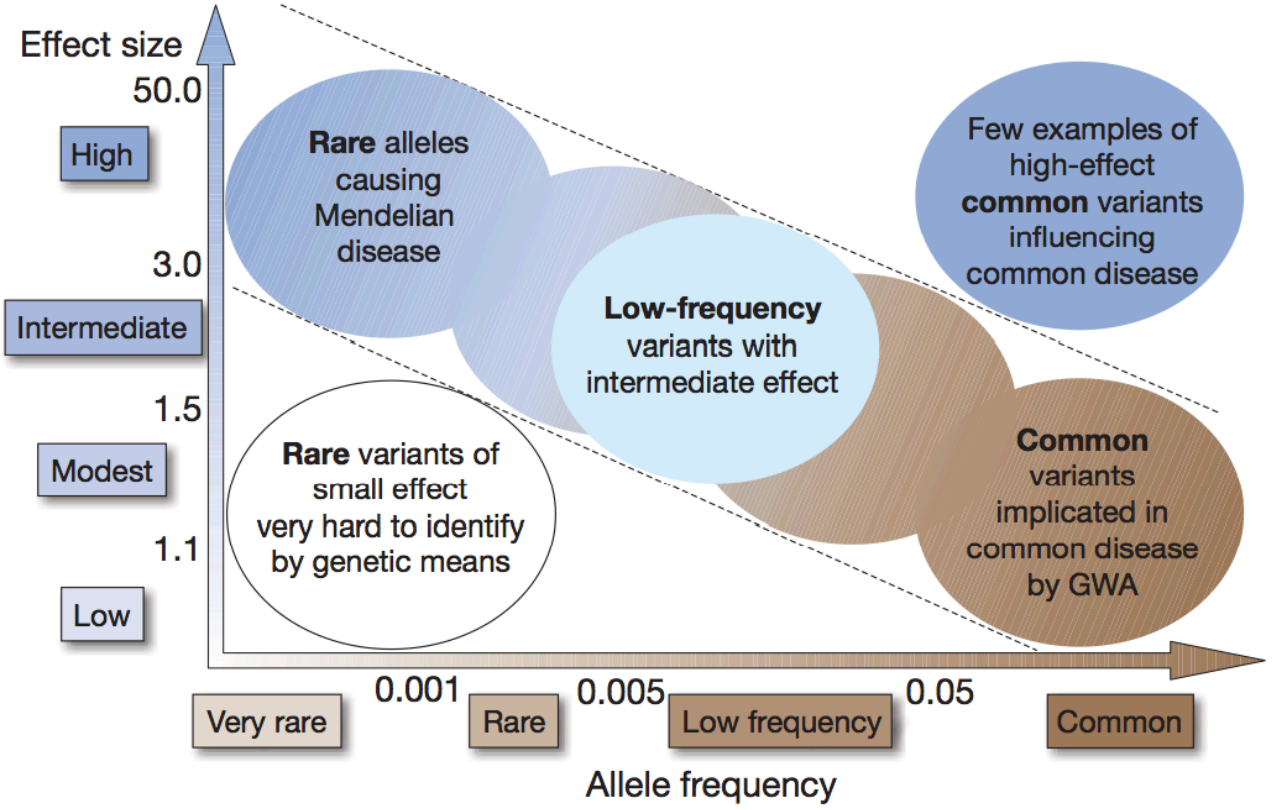
Quantile-Quantile Plots (Q-Q Plots)

Figure 1. Hypothetical Quantile-Quantile Plots in Genome-wide Association Studies



Pearson & Manolio. 2008. JAMA 299:1335-1344

Genetic architecture of complex traits



Software – GCTA (Yang *et al.* 2011)

- Genome-wide complex trait analysis: estimate proportion of phenotype variance explained by genotypes (collection of SNPs)
- Intuition: if a trait is genetically influenced, then individuals who are more genetically similar should be more phenotypically similar

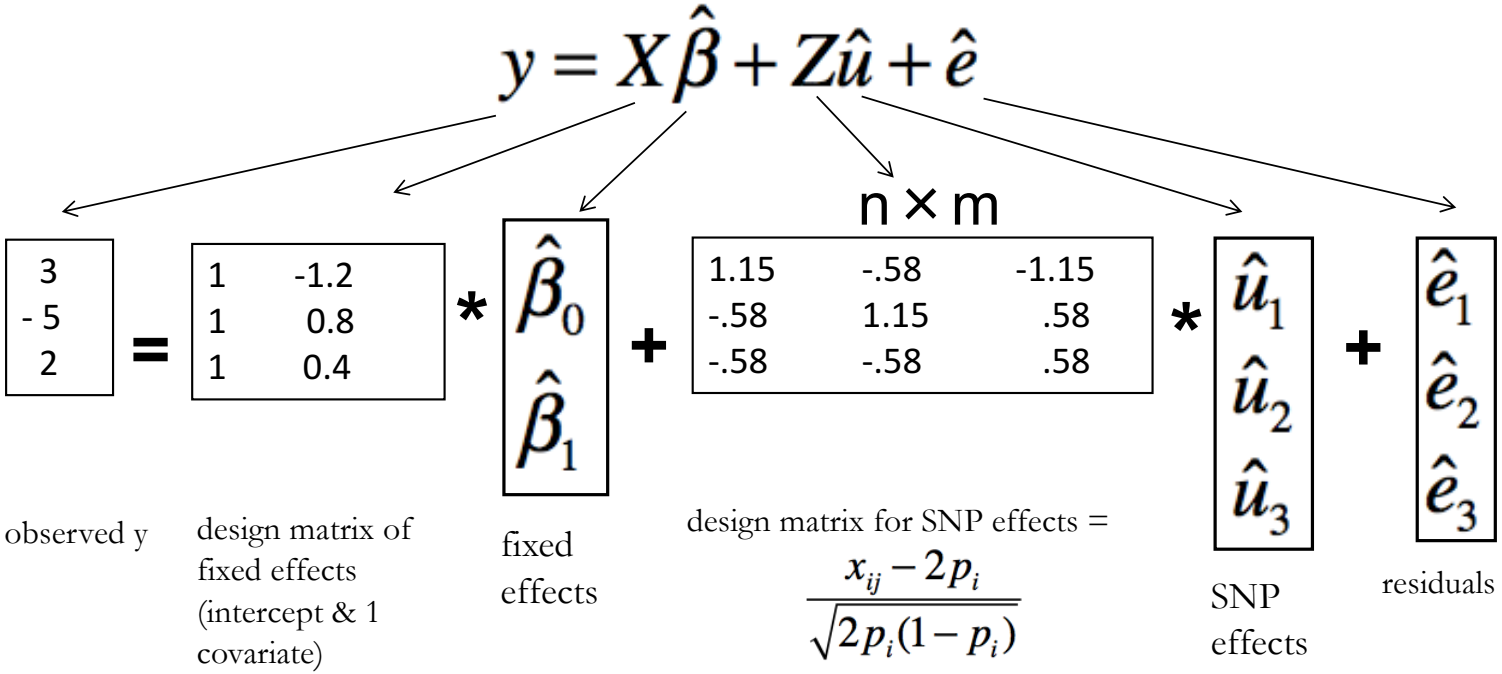
GCTA (Yang *et al.* 2011) – Mixed Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\underset{(n \times 1)}{\begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}} = \underset{(n \times m)}{\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}} \underset{(m \times 1)}{\begin{bmatrix} \beta_1 \\ \cdots \\ \beta_m \end{bmatrix}} + \underset{(n \times k)}{\begin{bmatrix} w_{11} & \cdots & w_{1k} \\ \cdots & \cdots & \cdots \\ w_{n1} & \cdots & w_{nk} \end{bmatrix}} \underset{(k \times 1)}{\begin{bmatrix} u_1 \\ \cdots \\ u_k \end{bmatrix}} + \underset{(n \times 1)}{\begin{bmatrix} \varepsilon_1 \\ \cdots \\ \varepsilon_k \end{bmatrix}}$$

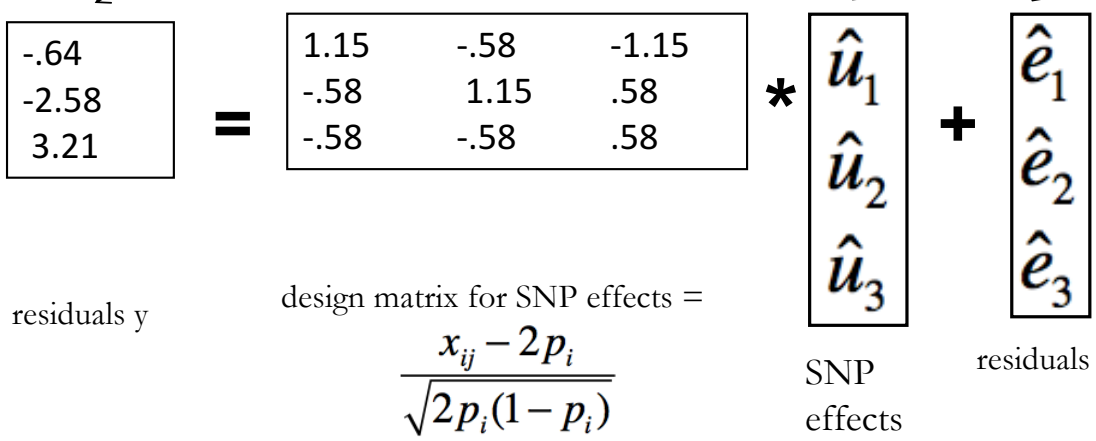
- **y:** phenotypes
- **x:** covariates (sex, age, etc)
- ***β:*** fixed effects regression coefficients
- **W:** genotype dosages
- **u:** random effects coefficients
- k: number of SNPs
- m: number of covariates
- n: number of individuals

GREML Model
(here, n=3, q=2 fixed effects, m=3 SNPs)



GREML Model
(after removing fixed effects on y)

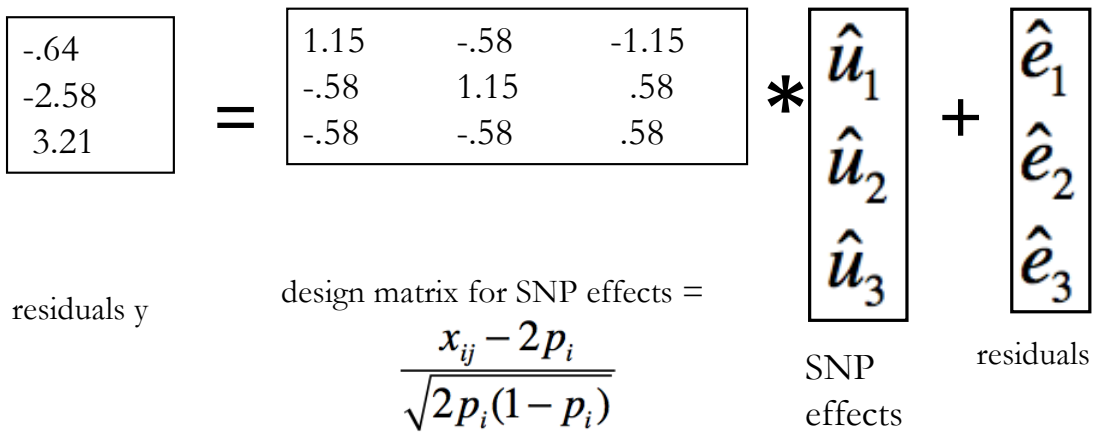
$$y_{.X} = Z\hat{u} + \hat{e}$$



We aren't interested in estimating each u_i because $m >> n$ usually, and because such individual estimates would be unreliable.
Instead, estimate the variance of u_i.

GREML Model
(after removing fixed effects on y)

$$y_{.X} = Z\hat{u} + \hat{e}$$

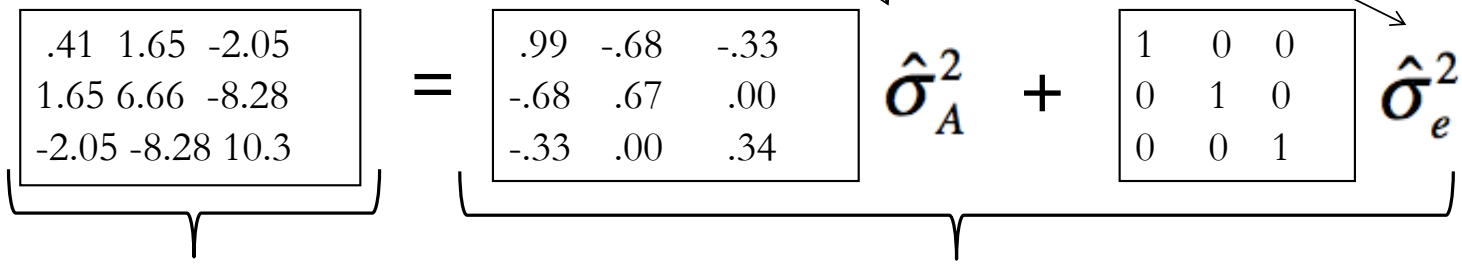


We assume $u \sim N(0, \sigma_u^2)$

and therefore $\sigma_A^2 = \sum_{i=1}^{m} \sigma_u^2 = m\sigma_u^2$

GREML Model
(after removing fixed effects on y)

$$\text{var}(y_{.X}) = ZZ'\sigma_u^2 + I\sigma_e^2$$
$$= ZZ'(\sigma_A^2 / m) + I\sigma_e^2$$
$$= G\sigma_A^2 + I\sigma_e^2$$



REML find values of $\hat{\sigma}_A^2$ & $\hat{\sigma}_e^2$ that maximizes the likelihood of the observed data.

Missing Heritability

- Single genetic variations cannot account for much of the heritability of diseases, behaviors, and other phenotypes.

Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits

Trait or Disease	h^2 Pedigree Studies	h^2 GWAS Hits[a]	h^2 All GWAS SNPs[b]
Type 1 diabetes	0.9[98]	0.6[99,c]	0.3[12]
Type 2 diabetes	0.3–0.6[100]	0.05-0.10[34]	
Obesity (BMI)	0.4–0.6[101,102]	0.01-0.02[36]	0.2[14]
Crohn's disease	0.6–0.8[103]	0.1[11]	0.4[12]
Ulcerative colitis	0.5[103]	0.05[12]	
Multiple sclerosis	0.3–0.8[104]	0.1[45]	
Ankylosing spondylitis	>0.90[105]	0.2[106]	
Rheumatoid arthritis	0.6[107]		
Schizophrenia	0.7–0.8[108]	0.01[79]	0.3[109]
Bipolar disorder	0.6–0.7[108]	0.02[79]	0.4[12]
Breast cancer	0.3[110]	0.08[111]	
Von Willebrand factor	0.66–0.75[112,113]	0.13[114]	0.25[14]
Height	0.8[115,116]	0.1[13]	0.5[13,14]
Bone mineral density	0.6-0.8[117]	0.05[118]	
QT interval	0.37–0.60[119,120]	0.07[121]	0.2[14]
HDL cholesterol	0.5[122]	0.1[57]	
Platelet count	0.8[123]	0.05–0.1[58]	

[a] Proportion of phenotypic variance or variance in liability explained by genome-wide-significant and validated SNPs. For a number of diseases, other parameters were reported, and these were converted and approximated to the scale of total variation explained. Blank cells indicate that these parameters have not been reported in the literature.

[b] Proportion of phenotypic variance or variance in liability explained when all GWAS SNPs are considered simultaneously. Blank cell indicate that these parameters have not been reported in the literature.

[c] Includes pre-GWAS loci with large effects.

Visscher, AJHG 2011

Mixed Linear Model (MLM)

- The test of association is performed in the fixed effects part of the model ("model for the means")
- "Relatedness" between individuals (due to both population structure and cryptic relatedness) is captured in the modelling of the covariance between individuals
- Can increase power by implicitly conditioning on associated loci other than the candidate locus (quantitative traits)
- Variety of software packages (e.g. GCTA, GEMMA, LMM-BOLT)

Mixed Linear Model (MLM)

- Q-K method (structured association)

$$y = X\beta + S\alpha + Qv + Zu + e$$

Fixed effects:

β	Vector of fixed effects
α	Vector of SNPs effects
v	Vector of subpopulation effects

Random effects:

u	Vector of kinship effects
e	Residuals

Q	Matrix of population association (STRUCTURE)
X, S, Z	Incidence Matrices

$$y = X\beta + S\alpha + Qv + Zu + e$$

	Location ID		SNP ID	Population ID		Genotype ID			
Trait	L1	L2	SNP1	P1	P2	G1	G2	G3	G4
y_1	1	0	1	1	0	1	0	0	0
y_2	1	0	1	0	1	0	0	1	0
y_3	1	0	1	0	1	0	0	1	0
y_4	1	0	0	0	1	0	0	1	0
y_5	0	1	0	1	0	0	1	0	0
y_6	0	1	0	0	1	0	0	0	1
y_7	0	1	1	1	0	0	1	0	0
y_8	0	1	1	0	1	0	0	0	1

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} * \begin{pmatrix} a_1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{pmatrix}$$

$$y_i = X\beta + S\alpha + Qv + Zu + e_i$$

$$y_3 = b_1 + a_1 + v_2 + u_3 + e_3$$

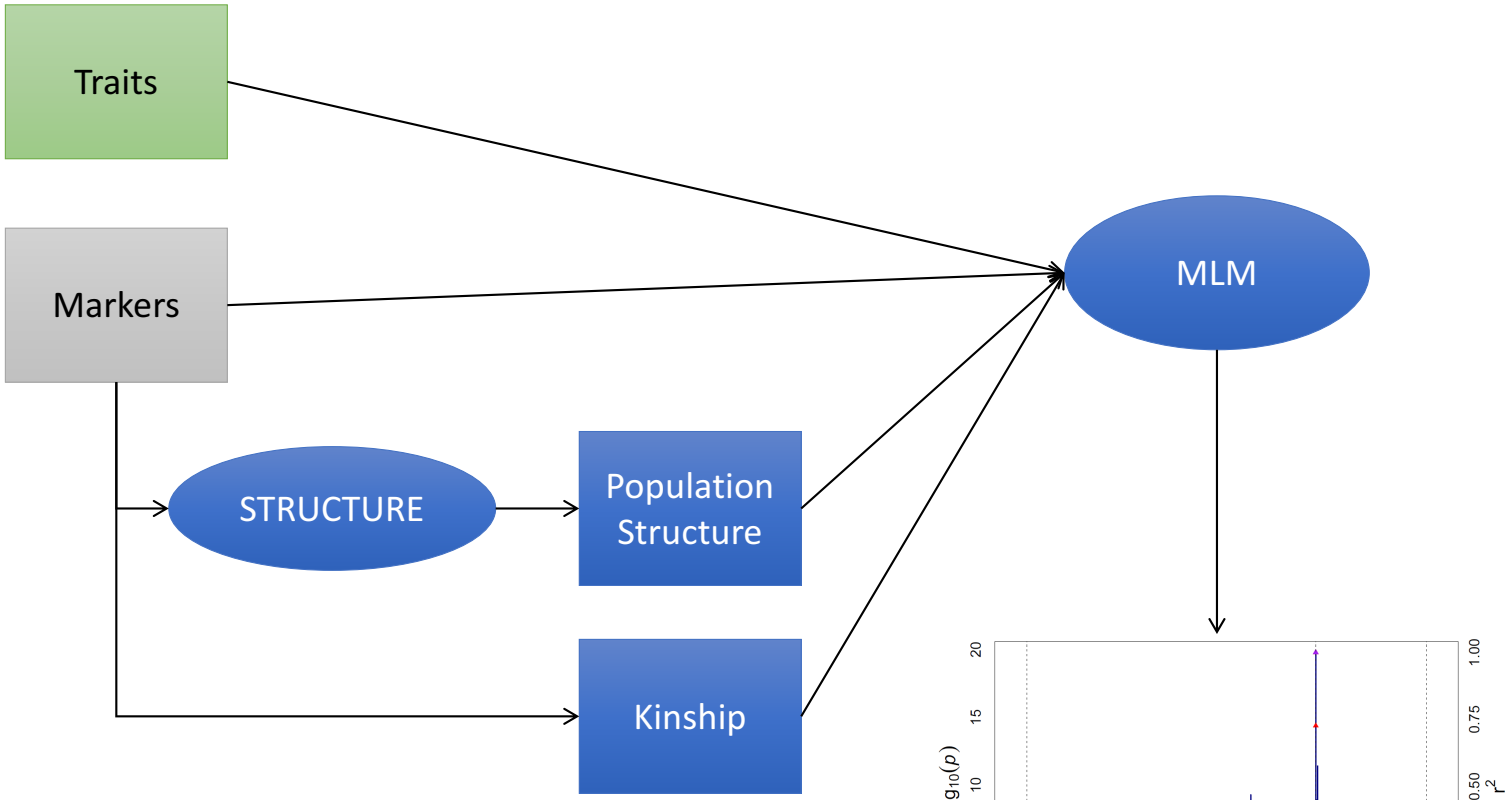
= Measured trait

= Fixed effects (BLUE = Best Linear Unbiased Esitimates)

= Random effects (BLUP = Best Linear Unbiased Predictions)

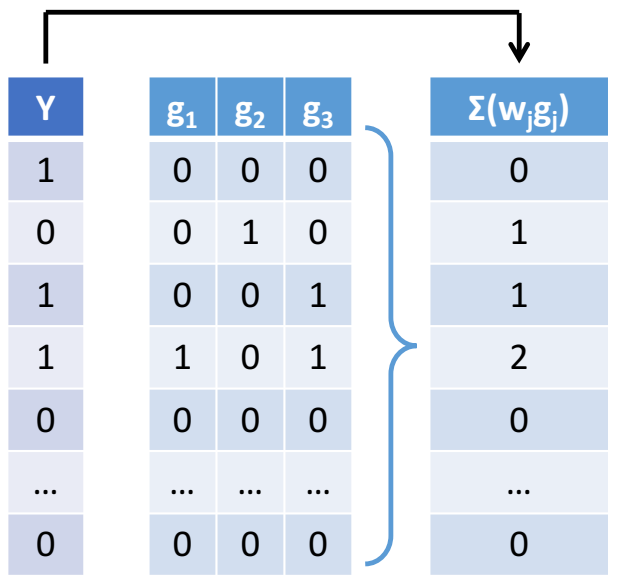
Yu et al. 2006

Mixed Linear Model (MLM)



Gene-based tests

- Gene-based tests jointly analyze multiple rare variants in genetic region (e.g. gene)
- Increases power by:
  - Combining information across rare variants
  - Requiring less stringent α, e.g. $\alpha = 2.5 \times 10^{-6}$ for 20K genes



Burden Test
(with weights w_j)

(Madsen & Browning, *PLoS Genet.*, 2009)

Selecting variants for gene-based tests

- If include variants of all frequencies, non-causal and common variants will dilute signal
- Commonly used filters or "masks":
  - Include variants MAF $\leq$ 0.05 or 0.01
  - Weight variants by MAF
    - E.g. $w_j \sim$ Beta(MAF, 1, 25)
  - Select variants based on functional annotation:
    - E.g. Protein Truncating Variants only, nonsynonymous, missense, etc.
- If mask is too restrictive, will reduce to single variant test, and no gain in power

Categories of aggregation tests

- **Burden tests** test association between (weighted) sum of rare alleles with disease or QT
  - CMC (Li & Leal, 2008), WSS (Madsen & Browning, 2009)

- **Dispersion tests** measure deviations from expected distribution
  - SKAT (Wu et al., 2011), C-alpha (Neale et al., 2011)

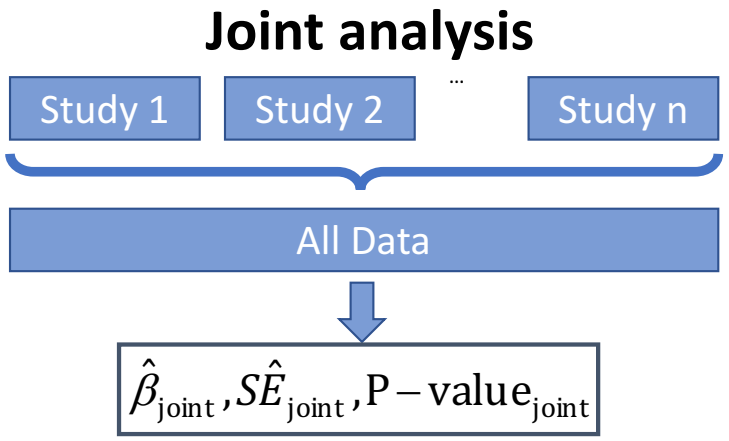
- **Combined tests** combine strengths of burden and dispersion tests
  - SKAT-O (Lee et al., 2012)

Multiple genetic association studies

- Most associated common variants have small effect sizes

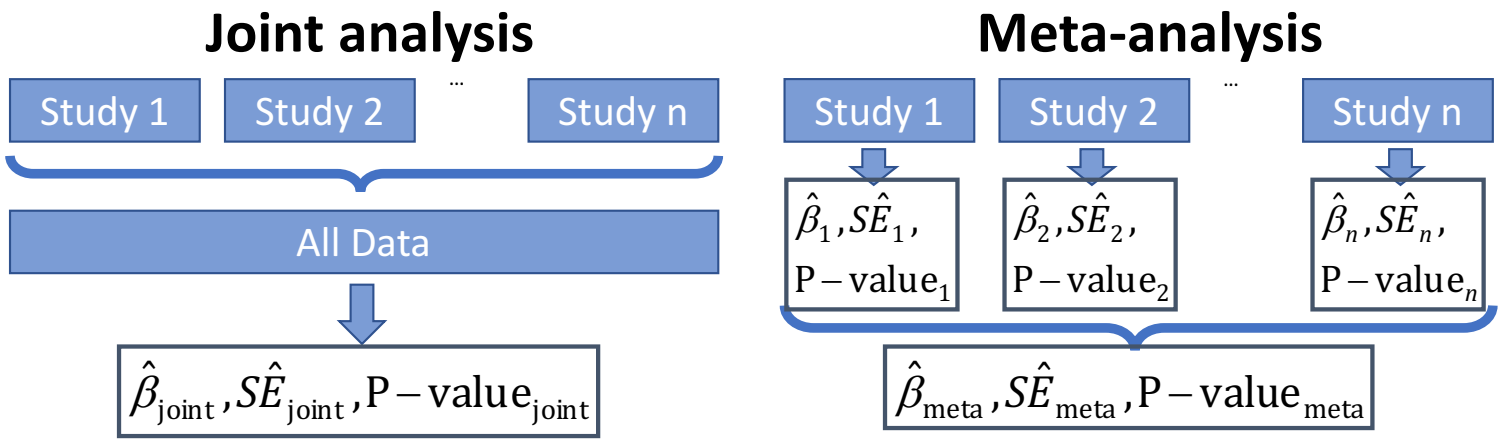
- To increase power to detect small genetic effect sizes, combine information across studies using
  - Meta-analysis of study-level association results
  - Joint analysis of all individual-level data

Multiple genetic association studies



- Combine individual-level data and analyze jointly

Multiple genetic association studies



- Combine study-level association results using:
  - Inverse-variance weights
  - Sample-size weights

Joint vs. meta-analysis

- For common variants, both joint and meta-analysis are both well-calibrated, and have near-equivalent power
- Meta-analysis is more commonly used
  - Sharing individual-level data is difficult due to logistical and ethical restrictions
- Combining multiple studies is critical to increase power to detect small effect sizes

(Lin & Zeng, *Genet. Epidemiol.*, 2010)

Linear Regression Including Dominance

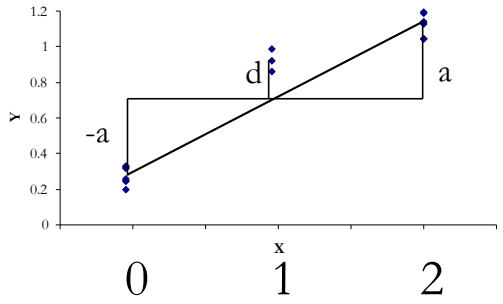
$$Y_i = a + b_x X_i + b_z Zi + e_i$$

where

Y_i = trait value for individual i

X_i = **1** if individual i has genotype 'AA'
0 if individual i has genotype 'Aa'
-1 if individual i has genotype 'aa'

Z_i = **0** for 'AA'
1 for 'Aa'
0 for 'aa'



Haplotypes?

- We may wish to consider more than one SNP at a time in the linear regression.
  - More information in a set of close SNPs
  - May wish to study a set of SNPs to see if one explains the phenotypic difference, i.e., does the evidence for one SNP disappear when controlling for other SNPs.

Haplotypes?

- Zaykin et al (2002) Hum Hered 53:79-91
- Use haplotypes in logistic regression
- For a pair of SNPs, there are 4 haplotypes, so there will be 3 "dummy" variables
- Assume pair of haplotypes in an individual are "additive", so only need 3 regression coefficients
- If haplotypes are known with certainty, then:

Haplotypes?

Haplotype	X1	X2	X3
h_1/ h_1	2	0	0
h_1/ h_2	1	1	0
h_1/ h_3	1	0	1
h_1/ h_4	1	0	0
h_2/ h_2	0	2	0
h_2/ h_3	0	1	1
h_2/ h_4	0	1	0
h_3/ h_3	0	0	2
h_3/ h_4	0	0	1
h_4/ h_4	0	0	0