Genome-wide association studies (GWAS) - Part 2

More advanced topics: Linear Mixed Models and G×G or G×E interactions

Heather J. Cordell

Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

heather.cordell@ncl.ac.uk

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories

Linear Mixed Models (LMMs)

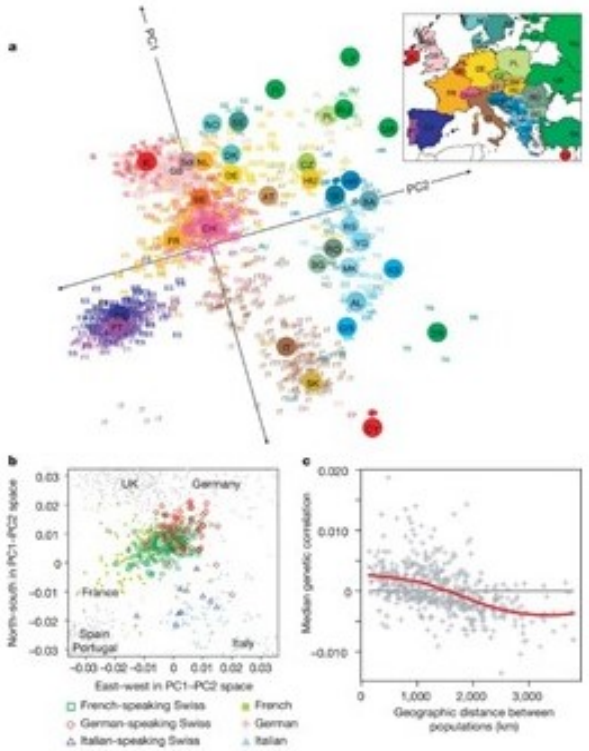
- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories
  - Investigating genetic correlations between different traits

Linear Mixed Models (LMMs)

- Linear Mixed Models have been used for many years in the plant and animal breeding communities
- In the mid 1990s they became popular in the human genetics field, mostly for performing **linkage analysis** and estimating **heritability**
  - Using family (pedigree) data i.e. related individuals
- In recent years they have become popular in the **genetic association** studies field for:
  - Testing for association while accounting for varying degrees of relatedness
    - Close family relationships
    - Distant relationships and population stratification/substructure
  - Estimating the heritability accounted for various partitions of SNPs:
    - All SNPs typed on a GWAS panel
    - All typed SNPs and others in LD with them
    - Partitions of SNPs in various functional categories
  - Investigating genetic correlations between different traits
  - Predicting trait values in a new individual

Population stratification and relatedness

Genes mirror geography within Europe



J Novembre *et al.* (2008) *Nature* **456(7218):98-101**, doi:10.1038/nature07331

Linear Mixed Models (LMMs)

- A linear mixed model is a statistical model in which the dependent variable is a linear function of both fixed and random independent variables
  - Known respectively as fixed and random effects
  - Fixed effects are considered 'fixed' at their measured values
  - Random effects are considered to be sampled from a distribution

Linear Mixed Models (LMMs)

- A linear mixed model is a statistical model in which the dependent variable is a linear function of both fixed and random independent variables
  - Known respectively as fixed and random effects
  - Fixed effects are considered 'fixed' at their measured values
  - Random effects are considered to be sampled from a distribution

- Recall the usual linear regression model

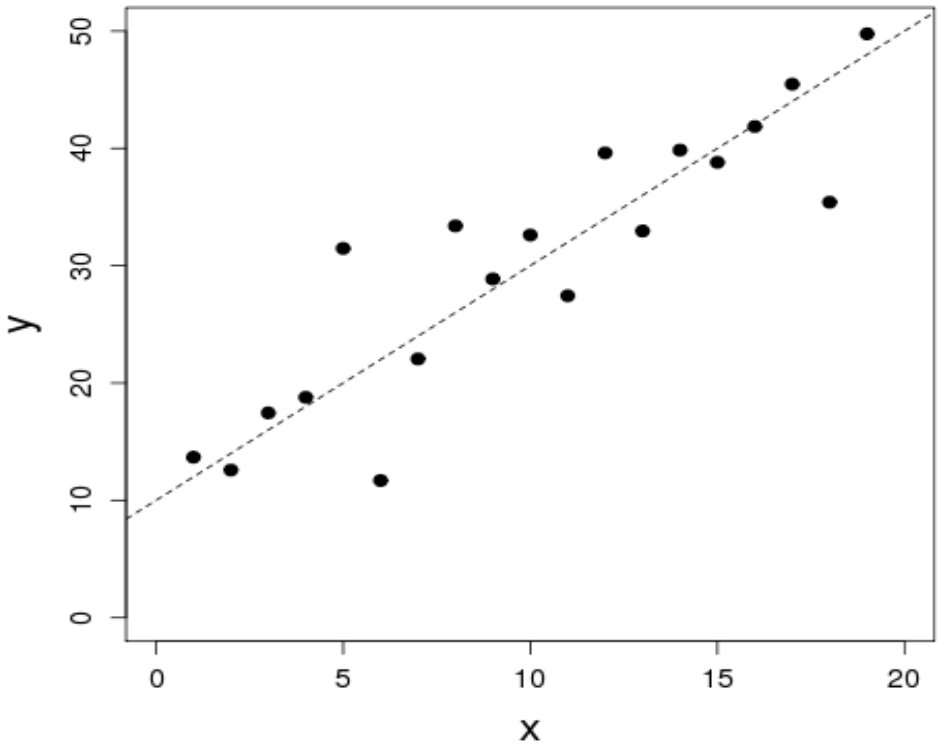
$$y = mx + c \qquad \text{or} \qquad y = \beta_0 + \beta_1 x$$

- This model may also be written

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  - y_i refers to the trait value of person i
  - x_i refers to the measured value of person i's predictor variable
  - ϵ_i refers to the displacement from the regression line
    - i.e. the discrepency between the observed and the predicted y value

Linear Regression



Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
  - x_i is the 'loading' of the fixed effect that someone has (based on their genotype)
- In matrix notation we can write this model:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

Linear Mixed Models (LMMs)

- In linear regression we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Here β_0 and β_1 are fixed effects while ϵ_i is a random error
  - x_i is the 'loading' of the fixed effect that someone has (based on their genotype)
- In matrix notation we can write this model:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- A LMM takes the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$
  - where $\mathbf{u}$ corresponds to a vector of random effects
    - with loadings specified in $\mathbf{Z}$

Linear Mixed Models (LMMs)

- E.g. suppose 2 fixed effects β_1 and β_2, and 3 random effects (plus n random errors)
- Then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$ corresponds to:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ . & . \\ . & . \\ x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ . & . & . \\ . & . & . \\ z_{n1} & z_{n2} & z_{n3} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

  - or $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + u_1 z_{i1} + u_2 z_{i2} + u_3 z_{i3} + \epsilon_i$

LMMs in genetics

- In genetics we generally work with two equivalent forms of LMM
- One is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$
  - The random effect u_l corresponds to a scaled additive effect of causal variant (locus) l
    - We assume there are many (m) such causal variants all across the genome
    - Considering it to be a random effect (within a population of interest) could be thought of as taking a Bayesian perspective

LMMs in genetics

- In genetics we generally work with two equivalent forms of LMM
- One is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$
  - The random effect u_l corresponds to a scaled additive effect of causal variant (locus) l
    - We assume there are many (m) such causal variants all across the genome
    - Considering it to be a random effect (within a population of interest) could be thought of as taking a Bayesian perspective
  - $\mathbf{Z}$ is a standardized genotype matrix i.e. z_{il} takes value

$$\left(\frac{-2f_l}{\sqrt{2f_l(1-f_l)}}, \quad \frac{(1-2f_l)}{\sqrt{2f_l(1-f_l)}}, \quad \frac{2(1-f_l)}{\sqrt{2f_l(1-f_l)}} \right)$$

    if individual i has genotype (qq, Qq, QQ)
    - where f_l is the frequency of allele Q at locus l

LMMs in genetics

- The other form is: $\mathbf{y} = \mathbf{X}\beta + \mathbf{g} + \epsilon$
  - Where $g_i = \sum_{l=1}^{m} z_{il} u_l$ is the total genetic effect in individual i, summed over all the causal loci
- In this form, g_i **can be considered as a random effect operating in individual** i
  - The vector of random effects $\mathbf{g}$ takes distribution $\mathbf{g} \sim N(0, \mathbf{G}\sigma_a^2)$
    - Where $\mathbf{G}$ is the genetic relationship matrix (GRM) between individuals – i.e. their IBD sharing at the causal loci
    - $\sigma_a^2 = m\sigma_u^2$ is the total additive genetic variance
    - $\mathbf{G}=\mathbf{ZZ'}/m$

LMMs in genetics

- The other form is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}$
  - Where $g_i = \sum_{l=1}^{m} z_{il} u_l$ is the total genetic effect in individual i, summed over all the causal loci
- In this form, g_i **can be considered as a random effect operating in individual** i
  - The vector of random effects $\mathbf{g}$ takes distribution $\mathbf{g} \sim N(0, \mathbf{G}\sigma_a^2)$
    - Where $\mathbf{G}$ is the genetic relationship matrix (GRM) between individuals – i.e. their IBD sharing at the causal loci
    - $\sigma_a^2 = m\sigma_u^2$ is the total additive genetic variance
    - $\mathbf{G}=\mathbf{ZZ'}/m$
- For family data (close relatives), the expected values of the elements of $\mathbf{G}$ equal the expected IBD sharing
  - i.e. twice the kinship coefficients
  - Thus $\mathbf{G}$ is just equal to twice the kinship matrix
    - Models their expected relatedness at the causal loci (and elsewhere)

Use of LMMs in genetics

- The formulation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}$ is known as the **Animal Model** and has been used extensively in plant and animal breeding
  - Mostly to predict the *breeding values* g_i in order to inform breeding strategies
    - E.g. to increase milk yield, meat production etc. etc.
  - Similar approaches could be used for *prediction* of trait values given genotype data
- In the mid 1990s it became popular in human genetics as the backbone of **variance components linkage analysis**
- Now commonly used in association analysis (GWAS)
  - To correct for relatedness, when testing for association

Testing for association using LMMs

- Idea is to test a fixed SNP effect β_1
  - While including a random effect γ_i that models relatedness
- Fit regression model: $y_i = \beta_0 + \beta_1 x_i + \gamma_i$
  - y is the trait value
  - x is a variable coding for genotype at the test SNP (e.g. an allele count, coded 0, 1, 2 for genotypes 1/1, 1/2, 2/2)
  - $\gamma_i = g_i + \epsilon_i$

Testing for association using LMMs

- Idea is to test a fixed SNP effect β_1
  - While including a random effect γ_i that models relatedness
- Fit regression model: $y_i = \beta_0 + \beta_1 x_i + \gamma_i$
  - y is the trait value
  - x is a variable coding for genotype at the test SNP (e.g. an allele count, coded 0, 1, 2 for genotypes 1/1, 1/2, 2/2)
  - $\gamma_i = g_i + \epsilon_i$
  - We assume $\boldsymbol{\gamma} \sim MVN(0, \mathbf{V})$ where variance/covariance matrix $\mathbf{V}$ follows standard variance components model
    - Variance/covariance matrix structured as:

$$\begin{aligned} V_{ij} &= \sigma_a^2 + \sigma_e^2 \quad (i = j) \\ V_{ij} &= 2\Phi_{ij}\sigma_a^2 \quad (i \neq j) \end{aligned}$$

    - σ_a^2, σ_e^2 represent the additive polygenic variance (due to all loci) and the environmental (=error) variance, respectively

Testing for association using LMMs

- LMMs were first (?) applied in human genetics by Boerwinkle et al. (1986) and Abney et al. (2002)
- Chen and Abecasis (2007) implemented them via the "FAmily based Score Test Approximation" (FASTA) in the MERLIN software package
  - Closely related to earlier QTDT method (Abecasis et al. 2000a;b) which implements a slightly more general/complex model
  - FASTA was also implemented in GenABEL, along with a similar test called GRAMMAR (Aulchenko et al. 2007)

Estimating the genetic relationship matrix

- These early implementations calculated the kinship matrix $\mathbf{\Phi}$ on the basis of known (theoretical) kinships constructed from known pedigree relationships
- Amin et al. (2007) proposed instead *estimating* the kinships based on genome-wide SNP data
  - Ideally we want to use $\mathbf{G}=\mathbf{ZZ'}/m$, the genetic relationship matrix (GRM) between individuals at the causal loci
  - Since we don't know the causal loci, we approximate $\mathbf{G}$ by $\mathbf{A}$, the overall GRM between individuals
    - Various different ways to estimate this, usually based on scaled (by allele frequency) matrix of *identity-by-state* (IBS) sharing

Estimating the genetic relationship matrix

- Once you move to estimating the GRM, you are no longer limited to using family data
- Kang et al. (2010) and Zhang et al. (2010) suggested applying the approach to apparently unrelated individuals
  - As a way of accounting for population substructure/stratification
  - Also proposed applying to binary traits (case/control coded 1/0)
  - Implemented in EMMAX and TASSEL software, respectively

Estimating the genetic relationship matrix

- Once you move to estimating the GRM, you are no longer limited to using family data
- Kang et al. (2010) and Zhang et al. (2010) suggested applying the approach to apparently unrelated individuals
  - As a way of accounting for population substructure/stratification
  - Also proposed applying to binary traits (case/control coded 1/0)
  - Implemented in EMMAX and TASSEL software, respectively
- Subsequently a number of other publications/software packages have implemented essentially the same model
  - FaST-LMM (Lippert et al. 2011)
  - GEMMA (Zhou and Stephens 2012)
  - GenABEL (GRAMMAR-Gamma) (Svishcheva et al. 2012)
  - MMM (Pirinen et al. 2013)
  - MENDEL (Zhou et al. 2014)
  - RAREMETALWORKER
  - GCTA
  - DISSECT

Software implementations

- Main difference between them is the precise computational tricks used to speed up the calculations
  - And the convenience/ease of use
    - See comparison in Eu-Ahsunthornwattana et al. (2014) PLoS Genetics 10(7):e1004445

Software implementations

- Main difference between them is the precise computational tricks used to speed up the calculations
  - And the convenience/ease of use
    - See comparison in Eu-Ahsunthornwattana et al. (2014) PLoS Genetics 10(7):e1004445
- BOLT-LMM (Loh et al. 2016) uses a slightly different approach, based on a Bayesian implementation of LMM formulation 1:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$$

  - One of the first mixed model packages that worked for really large-scale (e.g. UK Biobank) datasets
  - Now potentially (?) superseded by fastGWA module in GCTA
  - And by REGENIE (https://doi.org/10.1038/s41588-021-00870-7), which uses a slightly different formulation based on analysing the residuals following a whole-genome blockwise ridge regression
    - Again based on LMM formulation 1: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$

Binary traits

- For binary traits, coding cases and controls as a 1/0 quantitative trait is not optimal
  - Though in practice it seems to work reasonably well
- LTMLM (Hayeck et al. 2015) and LEAP (Weissbrod et al. 2015) instead use an underlying *liability model* to improve power
  - Assuming known disease prevalence

Binary traits

- For binary traits, coding cases and controls as a 1/0 quantitative trait is not optimal
  - Though in practice it seems to work reasonably well
- LTMLM (Hayeck et al. 2015) and LEAP (Weissbrod et al. 2015) instead use an underlying *liability model* to improve power
  - Assuming known disease prevalence
- Chen et al. (2016) showed that high levels of population stratification can invalidate the analysis, when applied to a case/control sample
  - Resulting in a mixture of **inflated** and **deflated** test statistics
  - Developed GMMAT software to address this problem
  - See also CARAT software (Jiang et al. 2016, AJHG 98:243-55)

Binary traits

- SAIGE software (Zhou et al. 2018, AJHG 50(9):1335-1341) implements a mixed model test that deals with large case-control imbalance, as you might see (for example) in UK Biobank
- REGENIE also implements this same saddle point approximation (SPA) test
  - Along with an approximate Firth penalized likelihood-ratio test

Elucidating genetic architecture

- Seminal paper by Yang et al. (2010) [Nat Genet 42(7):565-9]
- Showed that by framing the relationship between height and genetic factors as an LMM, 45% of variance could be explained by considering 294,831 SNPs simultaneously
  - So-called 'SNP heritability' or 'chip heritability'
  - Demonstrated that modelling effects at all genotyped SNPs explained the 'known' heritability ($\approx 80\%$) much better than just the top SNPs from GWAS
- Moreover, if you estimate effects of additional SNPs in LD with the genotyped SNPS, the variance explained goes up to 84% (s.e. 16%), consistent with 'known' value
- Subsequently many papers have shown similar results for a variety of complex traits

Elucidating genetic architecture

- Basic idea is to use formulation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}$$

with $\mathbf{g} \sim N(0, \mathbf{A}\sigma_a^2)$ and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_e^2)$ so $\mathbf{V} = \mathbf{A}\sigma_a^2 + \mathbf{I}\sigma_e^2$

  - $\mathbf{A}$ is the GRM between individuals, estimated using all genotyped SNPs
  - σ_a^2 and σ_e^2 estimated using REML (or MLE)
  - Thus we can estimate heritability accounted for by the genotyped SNPs as $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$

- Implemented in several software packages including GCTA and DISSECT
  - ALBI software (Schweiger et al. 2016, AJHG 98:1181-1192) can then be used to construct accurate confidence intervals for the heritability

Partitioning variance

- The same formulation can be used to partition the variance explained by different subsets of SNPs
  - Yang et al. (2010) partitioned variance onto each of the 22 autosomes using formulation

$$\mathbf{y} = \mathbf{X}\beta + \sum_{c=1}^{22} \mathbf{g_c} + \epsilon \qquad \text{with } \mathbf{V} = \sum_{c=1}^{22} \mathbf{A_c}\sigma_c^2 + \mathbf{I}\sigma_e^2,$$

    where $\mathbf{g_c}$ is a vector of effects attributed to the cth chromosome, and $\mathbf{A_c}$ is the GRM estimated from SNPs on the cth chromosome
    - Slight adjustment is needed for estimating variance explained by SNPs on chromosome X
- Similar partitioning can be used to examine subsets of SNPs defined in other ways e.g. according to MAF or functional annotation

Other approaches

- Some recent work has focussed on achieving similar ends
  - i.e. estimating
    - heritability explained by sets of SNPs
    - genetic correlations across traits

  using summary statistics only

  - Bulik-Sullivan et al. (2015) [Nat Genet 47:291-295]
  - Bulik-Sullivan et al. (2015) [Nat Genet 47:1236-1241]
    - Clever idea that allows the variance component parameters to be estimated via a simple regression on 'LD Scores'

Short break

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198
- Effects operating through interactions may not be visible unless you stratify by or take account of the interacting genetic (or environmental) factors
  - By modelling interactions, we hope to increase our power to detect loci with weak marginal effects

Gene-gene (and gene-environment) interactions

- GWAS have been extraordinarily successful at detecting genetic locations harboring genes associated with complex disease
  - But the SNPs identified do not account for the known (estimated) heritability for most disorders
  - Could G×G and G×E effects account for part of the 'missing heritability'?
    - Zuk et al. (2012) PNAS 109:1193-1198
- Effects operating through interactions may not be visible unless you stratify by or take account of the interacting genetic (or environmental) factors
  - By modelling interactions, we hope to increase our power to detect loci with weak marginal effects
- Phenomenon of biological interest?
  - Identifying genes that interact to cause disease could help us understand the mechanisms and pathways in disease progression

Definition of (pairwise) interaction

- Statistical interaction most easily described in terms a of (logistic) regression framework
  - Supppose x_1 and x_2 are binary factors whose presence/absence (coded 1/0) may be associated with a disease outcome
  - Logistic regression models their effect on the log odds of disease as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

Marginal effect of factor 1

$$\log \frac{p}{1-p} = \beta_0 + \beta_2 x_2$$

Marginal effect of factor 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Main effects of factors 1 and 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Main effects and interaction term

  - For quantitative traits, use linear regression (replace $\log \frac{p}{1-p}$ with y)
  - For modelling as an LMM, add in a random effect γ

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'
    - The 'effect' of factor 2 is different in the presence/absence of factor 1

Interaction

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value
      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'
    - The 'effect' of factor 2 is different in the presence/absence of factor 1

- Suppose no main effects ($\beta_1 = \beta_2 = 0$)

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_{12}$	β_0
0	β_0	β_0

  - Trait value only differs from baseline if both factors present

Gene-gene interaction (epistasis)

- However SNPs are not binary, but rather take 3 levels according to the number of copies (0,1,2) of the susceptibility allele possessed
- Most general 'saturated' (9 parameter) genotype model allows all 9 penetrances to take different values
  - Via modelling log odds in terms of:
    - A baseline effect (β_0)
    - Main effects of locus G (β_{G_1}, β_{G_2})
    - Main effects of locus H (β_{H_1}, β_{H_2})
    - 4 interaction terms

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G_2}+\beta_{H_2}+\beta_{22}$	$\beta_0+\beta_{G_2}+\beta_{H_1}+\beta_{21}$	$\beta_0+\beta_{G_2}$
1	$\beta_0+\beta_{G_1}+\beta_{H_2}+\beta_{12}$	$\beta_0+\beta_{G_1}+\beta_{H_1}+\beta_{11}$	$\beta_0+\beta_{G_1}$
0	$\beta_0+\beta_{H_2}$	$\beta_0+\beta_{H_1}$	β_0

  - Corresponds in statistical analysis packages to coding x_1, x_2 (0,1,2) as a "factor"

Gene-gene interaction

- Alternatively we can assume additive effects of each allele at each locus:
  - Corresponds to fitting

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

with x_1, x_2 coded (0,1,2)

	Locus H		
Locus G	2	1	0
2	$\beta_0 + 2\beta_G + 2\beta_H + 4\beta_{GH}$	$\beta_0 + 2\beta_G + \beta_H + 2\beta_{GH}$	$\beta_0 + 2\beta_G$
1	$\beta_0 + \beta_G + 2\beta_H + 2\beta_{GH}$	$\beta_0 + \beta_G + \beta_H + \beta_{GH}$	$\beta_0 + \beta_G$
0	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	β_0

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way

Change of scale

- Transformations of outcome variable y can change whether or not the predictor variables interact
  - Due to definition of interaction as departure from a linear model for the effects of x_1 and x_2, for predicting y
    - Two SNPs that interact on the log odds scale may not interact on the penetrance scale (and vice versa)
    - Makes biological interpretation of resulting interaction model difficult
- Much discussion in the literature
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387; Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171; Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159; Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540; Wang, Elston and Zhu (2010) Hum Hered 70:269-277
- Bottom line is, little direct correspondence between statistical interaction and biological interaction
  - In terms of whether, for example, gene products physically interact
- However, existence of statistical interaction does imply both loci are "involved" in disease in some way
  - Good starting point for further investigation of their (joint) action

Gene-environment (G×E) interactions

- The same regression model

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

can be used to model interaction between a genetic factor G and an environmental factor H

  - With the environmental variable x_2 coded in binary fashion (e.g. smoking) or quantitatively (e.g. age)

Gene-environment (G×E) interactions

- The same regression model

$$\log \frac{p}{1-p} = \beta_0 + \beta_G x_1 + \beta_H x_2 + \beta_{GH} x_1 x_2$$

  can be used to model interaction between a genetic factor G and an environmental factor H
  - With the environmental variable x_2 coded in binary fashion (e.g. smoking) or quantitatively (e.g. age)
- Focus of analysis is often risk estimation
  - Estimating genetic risks in particular environments
  - Estimating effect of environmental factor on particular genetic background
    - Important for treatment/screening strategies and public health interventions
- For G×G, focus of interest is more related to
  - Increasing power to detect an effect (by taking into account the effects of other genetic loci)
  - Modelling the biology, especially related to the joint action of the loci

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone

- Depending on circumstances, any of these tests may be a sensible option

Testing association and/or interaction

- Go back to binary coding of genetic (and/or environmental) factors

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - 3df test of $\beta_1 = \beta_2 = \beta_{12} = 0$ tests for association at both loci (or both variables), allowing for their possible interaction
  - 2df test of $\beta_2 = \beta_{12} = 0$ tests for association at locus 2, while allowing for possible interaction with locus (or variable) 1
  - 1df test of $\beta_{12} = 0$ tests the interaction term alone
- Depending on circumstances, any of these tests may be a sensible option
- Most tests of interaction/joint action can be thought of as a version of one or other of these tests
  - Although different tests vary in their precise details
  - And their relationship to the logistic regression formulation not always clearly described
    - See Howey and Cordell (2017) https://pubmed.ncbi.nlm.nih.gov/28852712/

G×G versus G×E in the context of GWAS

- Typically GWAS measure thousands if not millions of genetic variants
  - But only a few (tens or at most 100s) of environmental factors
- Feasible to consider all G×E combinations
- All pairwise G×G combinations possible, but much more time consuming
  - And leads to greater multiplicity of tests
  - Also, why stop at 2-way interactions?
    - Could look at all 3 way, 4 way etc. combinations
    - Scale of problem quickly gets out of hand
    - Less obvious reason to do this for G×E...

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502
    "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"
- Or have proposed filtering based on single-locus significance or other (biological or statistical) considerations
  - Reduces multiple testing burden, improves interpretability

G×G in the context of GWAS

- Many recent publications have focussed on finding clever computational tricks to speed up exhaustive search procedure
  - BOOST (Wan et al. (2010) AJHG 87:325-340)
  - SIXPAC (Prabhu and Pe'er (2012) Genome Res 22:2230-2240)
  - Kam-Thong et al. (2012) Hum Hered 73:220-236 (GPUs)
  - Fråanberg et al. (2015) PLOS Genetics 11(9):e1005502 "Discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests"
- Or have proposed filtering based on single-locus significance or other (biological or statistical) considerations
  - Reduces multiple testing burden, improves interpretability
- Or have proposed testing at the gene level rather than the SNP level
  - Ma et al. (2013) PLoS Genet 9(2): e1003321
    - Compared 4 different tests that combine P values from pairwise (SNP x SNP) interaction tests
    - Showed that the truncated tests did best
    - Presented an application only considering gene pairs known to exhibit protein-protein interactions

Case-only analysis

- Piergorsh et al. 1994; Yang et al. 1999; Weinberg and Umbach 2000
- Several authors have shown that, for binary predictor variables, a test of the interaction term β_{12} in the logistic regresssion model

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  can be obtained by testing for correlation (association) between the genotypes at two separate loci, within the sample of cases
- Gains power from making assumption that genotypes (alleles) at the two loci are uncorrelated in the population
  - So only really suitable for unlinked or loosely linked loci (since closely linked loci are likely to be in LD)
- Alternatively contrast the genotype correlations in cases with those seen in controls (`--fast-epistasis` in PLINK)

Testing correlation between loci

- A similar idea is implemented in EPIBLASTER (Kam-Thong et al. 2011; EJHG 19:465-571)
- Wu et al. (2010) (PLoS Genet 6:e1001131) also proposed a similar approach – though needs adjustment to give correct type I error rates
- See also Joint Effects (JE) statistics (Ueki and Cordell 2012; PLoS Genetics 8(4):e1002625)
- All these methods test whether correlation exists (case-only) or is different in cases and controls (case/control)
  - Via testing a log OR for association between two loci
  - However, the log OR for association (λ) encapsulates a slightly different quantity between the different methods
- All implemented (along with standard logistic and linear regression) in CASSI
  - http://www.staff.ncl.ac.uk/richard.howey/cassi/

Empirical evidence for G×G interactions

- Epistasis among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* in multiple sclerosis (Lincoln et al. 2009 PNAS 106:7542-7547)
- *HLA-C* and *ERAP1* in psoriasis (Strange et al. 2010)
- *HLA-B27* and *ERAP1* in ankylosing spondylitis (Evans et al. 2011)
- *BANK1* and *BLK* in SLE (Castillejo-Lopez et al. 2012)
- Gusareva et al. (2014) found a reasonably convincing (partially replicating) interaction between SNPs on chromosome 6 (*KHDRBS2*) and 13 (*CRYL1*) in Alzheimer's disease
- Dai et al. (2016) [AJHG 99:352-365] identified 3 loci simultaneously interacting with established risk factors gastresophageal reflux, obesity and tobacco smoking, with respect to risk for Barrett's esophagus

Empirical evidence for G×G interactions

- Hemani et al. 2014 (Nature 508:249-253) found 501 instances of epistatic effects on gene expression, of which 30 could be replicated in two independent samples
  - Many SNPs are close together, could represent haplotype effects?
  - Or the effect of a single untyped variant?
  - See caveats in
    - Wood et al. (2014) Nature 514(7520):E3-5. PMID:25279928
    - Fish et al. (2016) Am J Hum Genet 99(4):817830. PMID:27640306
- The Hemani et al. paper was subsequently retracted (https://www.nature.com/articles/s41586-021-03766-y)

Empirical evidence for G×E interactions

- Myers et al. (2014) Hum Mol Genet 23(19): 5251-9 "Genome-wide Interaction Studies Reveal Sex-Specific Asthma Risk Alleles"
- Small et al. (2018) Nat Genet 50(4): 572-580 "Regulatory Variants at KLF14 Influence Type 2 Diabetes Risk via a Female-Specific Effect on Adipocyte Size and Body Composition"
- Sung et al. (2019) Hum Molec Genet 28(15): 2615-2633 "A multi-ancestry genome-wide study incorporating gene-smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure."

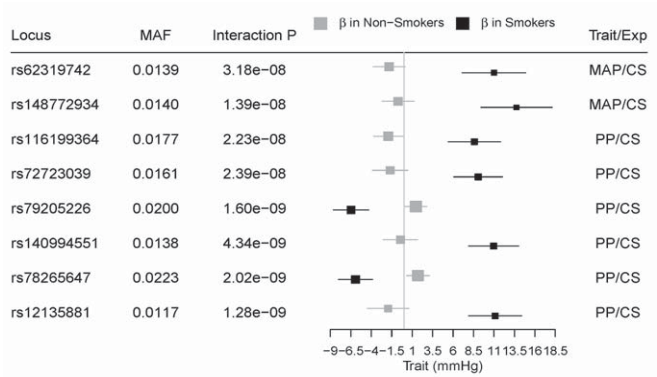


Figure 3. Smoking-specific genetic effect sizes in African ancestry for MAP or PP. Among the 138 loci significantly associated with MAP and/or PP, 8 loci show significant interactions with smoking exposure status in African ancestry. Smoking-specific effect estimates and 95% confidence intervals for variants associated with BP traits are shown as red and blue squares for current-smokers and non-current smokers, respectively. SNP effects between two strata are significantly different (one DF interaction $P < 5 \times 10^{-8}$). These results were based on African-specific results in stage 1. MAP: mean arterial pressure; PP: pulse pressure; CS: current-smoking.

Empirical evidence for G×E interactions

- Favé et al. (2018) Nat Commun 9(1): 827 "Gene-by-environment Interactions in Urban Populations Modulate Risk Phenotypes"

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03202-2 ARTICLE