Mixed models and family-based methods for genetic association analysis

Heather J. Cordell

Institute of Genetic Medicine
Newcastle University, UK
heather.cordell@ncl.ac.uk
Consider a trait, $x$, influenced by a single genetic locus.
- e.g. height, weight, blood pressure etc.

Not all of its variation is attributable to the locus — there is variation in $x$ even amongst subjects with identical genotypes at the locus.
Genetic and environmental variance

- Variance of $x$ about overall mean $\mu$, may be split into 2 parts:
  - the genetic variance $\sigma^2_g$ or $\sigma^2_G$ or $V_G$: the variance that is accounted for by genetic effects
  - the environmental variance $\sigma^2_e$ or $\sigma^2_E$ or $V_E$: the variance not accounted for by genetic factors

- The genetic variance may be further partitioned into:
  - An additive component $\sigma^2_a$ or $\sigma^2_A$ or $V_A$
    - Representing variation due to simple additive effects of alleles
  - A dominance component $\sigma^2_d$ or $\sigma^2_D$ or $V_D$
    - Representing more complicated, non-additive genetic effects
    - e.g. in a recessive model, the alleles are not just acting additively
Genetic and environmental variance

- Total variance:
  \[ \text{Var}(x) = \sigma_e^2 + \sigma_g^2 \]

- Total genetic variance:
  \[ \sigma_g^2 = \sigma_a^2 + \sigma_d^2 \]
Genetic and environmental variance

- Genetic variance represents ‘how important’ genetic factors are in accounting for trait variation.

- Environmental variance represents ‘how important’ other (non-genetic) factors are in accounting for trait variation.

- Note that if many \((= m)\) loci contribute additively to a trait, then the additive and dominance variances (and thus the total genetic variance) are just sums of effects due to each locus:

\[
\sigma^2_a = \sum_{l=1}^{m} \sigma^2_{a_l} \quad \sigma^2_d = \sum_{l=1}^{m} \sigma^2_{d_l} \quad \sigma^2_g = \sum_{l=1}^{m} \sigma^2_{g_l}
\]

- Here \(\sigma^2_{g_l}\) represents ‘how important’ a particular genetic locus \(l\) is.
Heritability

- **Heritability (broad-sense)** is defined as the proportion of the total variance that is genetic:
  \[ H^2 = \frac{\sigma^2_g}{\sigma^2_e + \sigma^2_g} \]
  
  - A highly heritable trait (e.g. height) might have heritability ≈ 80%
  - However, we don’t know whether this is due to one genetic factor, or many

- A more commonly-used measure is the **narrow-sense heritability**:
  \[ h^2 = \frac{\sigma^2_a}{\sigma^2_e + \sigma^2_g} \]
  
  - Corresponds to the proportion of the total variance that is explained by additive effects of alleles

- Heritability varies between populations due to variation in \(\sigma^2_e\)

- Also varies according to the **scale** on which the trait is measured
Suppose we have two individuals, measured for some quantitative trait \( x \) (e.g. height, weight, immune response)

It can be shown that, if the trait \( x \) is influenced by a single gene, the covariance (or correlation) between trait values for two individuals depends on:

- The genetic (additive and dominance) variances
  - The higher these are, the higher the correlation in trait values

- The kinship coefficient \( \Phi \) between the individuals
  - The higher this is, the higher the correlation in trait values
Kinship

- Kinship coefficients are defined as half the expected proportion of alleles shared identical by descent (IBD)
  - i.e. inherited IBD from the most recent common ancestor
- Kinship coefficients represent the ‘degree of relatedness’ between pairs of individuals

<table>
<thead>
<tr>
<th>Relationship</th>
<th>$\Phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated</td>
<td>0</td>
</tr>
<tr>
<td>Self, monozygotic twins</td>
<td>1/2</td>
</tr>
<tr>
<td>Siblings, dizygotic twins, parent–offspring</td>
<td>1/4</td>
</tr>
<tr>
<td>Half siblings, double first cousins, uncle–nephew</td>
<td>1/8</td>
</tr>
<tr>
<td>Grandchild–grandparent</td>
<td>1/8</td>
</tr>
<tr>
<td>First cousins</td>
<td>1/16</td>
</tr>
<tr>
<td>Second cousins</td>
<td>1/64</td>
</tr>
</tbody>
</table>

- People who are more distantly related tend to have lower kinship values
  - And thus have less correlation in their trait values
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.
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

\[ y = mx + c \quad \text{or} \quad y = \beta_0 + \beta_1 x \]

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
- \( \epsilon_i \) refers to the displacement from the regression line
  - i.e. the discrepancy 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 $\beta_0$ and $\beta_1$ are fixed effects while $\epsilon_i$ is a random error

- In matrix notation we can write this model:
  \[
  \begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
  \end{bmatrix} = \begin{bmatrix}
  1 & x_1 \\
  1 & x_2 \\
  \vdots & \vdots \\
  1 & x_n
  \end{bmatrix} \begin{bmatrix}
  \beta_0 \\
  \beta_1
  \end{bmatrix} + \begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_n
  \end{bmatrix}
  \]
  or $y = X\beta + \epsilon$

- A LMM takes the form $y = X\beta + Zu + \epsilon$
  - where $u$ corresponds to a vector of random effects
E.g. suppose 2 fixed effects $\beta_1$ and $\beta_2$, and 3 random effects

Then $y = X\beta + Zu + \epsilon$ corresponds to:

$$
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} =
\begin{bmatrix}
x_{11} & x_{12} \\
x_{21} & x_{22} \\
\vdots & \vdots \\
x_{n1} & x_{n2}
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2
\end{bmatrix}
+ 
\begin{bmatrix}
z_{11} & z_{12} & z_{13} \\
z_{21} & z_{22} & z_{23} \\
\vdots & \vdots & \vdots \\
z_{n1} & z_{n2} & z_{n3}
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2 \\
u_3 \\
\epsilon_1 & \epsilon_2 & \epsilon_3
\end{bmatrix}
$$

or

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + u_1 z_{i1} + u_2 z_{i2} + u_3 z_{i3} + \epsilon_i$$
In genetics we generally work with two equivalent forms of LMM

One is: \[ y = X\beta + Zu + \epsilon \]

- The random effect \( u_l \) corresponds to a scaled additive effect of causal variant \( l \)
  - Assuming many \((m)\) such causal variants all across the genome

- The random effects \( u_l \) all have variance \( \sigma_u^2 \) and are uncorrelated with each other
  - So \( u = (u_1, u_2, \ldots, u_m) \sim N(0, I\sigma_u^2) \)

- \( Z \) is a standardized genotype matrix i.e. \( z_{il} \) takes value

\[
\begin{pmatrix}
-2f_l \\
\sqrt{2f_l(1-f_l)} \\
(1-2f_l) \\
\sqrt{2f_l(1-f_l)} \\
2(1-f_l) \\
\sqrt{2f_l(1-f_l)}
\end{pmatrix}
\]

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: \( y = X\beta + g + \epsilon, \) or

\[
y_i = \sum_k \beta_k x_{ik} + g_i + \epsilon_i
\]

- 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 \( g \) takes distribution \( g \sim N(0, G\sigma_a^2) \)

  - Where \( G = ZZ' / m \) is the genetic relationship matrix (GRM) between individuals at the causal loci

  - \( \sigma_a^2 = m\sigma_u^2 \) is the total additive genetic variance

- For family data (close relatives), the expected values of the elements of \( G \) are just equal to the kinship coefficients i.e. \( G \) is just equal to the kinship matrix

  - Models their relatedness at the causal loci (and elsewhere)
The formulation $y = X\beta + g + \epsilon$ 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**.
Variance components linkage analysis

- Given a set of family data genotyped at a test genomic region, we calculate the likelihood of observed data based on a multivariate normal distribution:

\[ y_i = \beta_0 + \sum_k \beta_k x_{ik} + \gamma_i \]

- Here \( \beta \) represent fixed effects and \( \gamma_i \) corresponds to \( g_i + \epsilon_i \) in the previous formulation.

- \( \gamma \sim MVN(0, V) \) where the variance/covariance matrix \( V \) can be written:

\[ V_{ii} = \sigma^2_T + \sigma^2_{a-T} + \sigma^2_e \]
\[ V_{ij} = \pi_{ij} \sigma^2_T + 2\Phi_{ij} \sigma^2_{a-T} \]

- \( \sigma^2_T, \sigma^2_{a-T}, \sigma^2_e \) are population parameters corresponding to:
  - the additive genetic variance at the locus you are testing
  - the polygenic additive variance (due to all other loci)
  - the environmental (=error) variance

- \( \pi_{ij} \) is the IBD sharing by individuals \( i \) and \( j \) (at the test locus)
- \( 2\Phi_{ij} \) is their expected IBD sharing (= twice their kinship coefficient)
Find set of parameters $\beta_0$, $\sigma_T^2$, $\sigma_{a-T}^2$, $\sigma_e^2$ that maximises likelihood

Test null hypothesis by comparing model where $\sigma_T^2$ is estimated and model where $\sigma_T^2 = 0$ (via likelihood ratio test)

So method comes down to testing the variance of a random effect

Popular method, but suffers from general problem of low power of linkage approaches for complex traits
Testing for association using LMMs

- The same variance components model can be used to correct for relatedness, when testing for association
  - Via testing a fixed SNP effect $\beta_1$

- 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 (e.g. 0, 1, 2 for genotypes 1/1, 1/2, 2/2)
  - $\gamma \sim MVN(0, \mathbf{V})$ where variance/covariance matrix $\mathbf{V}$ follows standard variance components model

- Variance/covariance matrix structured as:
  $$V_{ij} = \begin{cases} \sigma_a^2 + \sigma_e^2 & (i = j) \\ 2\Phi_{ij}\sigma_a^2 & (i \neq j) \end{cases}$$
  - $\sigma_a^2$, $\sigma_e^2$ represent the additive polygenic variance (due to all loci) and the environmental (=error) variance, respectively
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 $\Phi_{ij}$ 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 $G = ZZ' / m$, the genetic relationship matrix (GRM) between individuals at the causal loci.
- Since we don’t know the causal loci, we approximate $G$ by $A$, 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.
Traditional family-based association tests

- Traditional family-based tests look at transmission of particular ‘high-risk’ alleles within pedigrees
  - In an approach conceptually similar to linkage analysis
  - Gives robustness to population stratification
    - Non-transmitted alleles (or all alleles that could have been transmitted) act as ‘control’ for transmitted allele

- Simplest example is the Transmission/Disequilibrium Test (TDT)
  - Based on sampling case/parent trios
  - Test for non-random transmission of ‘high-risk’ alleles to the case from heterozygous parents
  - Has been extended to larger families (FBAT, PBAT, PDT)
    - Similar principle to TDT: examine increased transmission of particular allele to set of affected individuals in pedigree
    - Or relate the transmission probabilities to the value of some quantitative trait
Traditional family-based tests are very robust, but tend to have lower power

- Owing to lower effective sample size
- E.g. in Fakiola et al. (2013) we analysed data from Brazilian families with visceral leishmaniasis
  - 308 families; 3626 individuals; 1970 genotyped and phenotyped individuals (of which 357 were affected)
  - FBAT uses 357 cases compared to 357 ‘pseudo’ controls
  - LMMs use 357 cases compared to 1613 genuine controls
- Still worth considering if they match your ascertainment scheme better (e.g. case/parent trios)

There are alternative (non-LMM based) approaches that try to make use of all available data in large pedigrees

- MQLS, ROADTRIPS, MASTOR, GTAM
  - Try to correct the ‘usual’ association test statistics for the expected inflation due to relatedness
- Eu-Ahsunthornwattana et al. (2014) found them to perform quite similarly to LMMs
Manhattan plots (real VL data)
Manhattan plots (simulated data)
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
  (http://genome.sph.umich.edu/wiki/RAREMETALWORKER)
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

Association testing also implemented in some more general packages
- GCTA
- DISSECT
- EPACTS

BOLT-LMM (Loh et al. 2016) uses a slightly different approach, based on a Bayesian implementation of LMM formulation 1:

\[ y = X\beta + Zu + \epsilon \]

Model can also be extended to bivariate traits (Korte et al. 2012, Nat Genet 44:1066-1071), implemented in MTMM/ASREML and DISSECT
Power of MTMM (2 correlated phenotypes)

Figure 1  Simulation results. (a–d) Scenarios simulated—with positive pleiotropy, alternative common effect across environments (a); positive pleiotropy, alternative common effect across environments, with size of effect differing between traits and environments (b); effect only on one trait, alternative only in one environment (c); and negative pleiotropy, alternative opposite effect across environments (d). (e) Estimated relationship between power and false discovery rate (FDR) using six different statistical tests for the scenario in a. (f) Estimated relationship between power and FDR for the scenario in b. (g) Estimated relationship between power and FDR for the scenario in c. (h) Estimated relationship between power and FDR for the scenario in d. Dots on curves denote nominal Bonferroni-corrected 5% significance thresholds. Both power and FDR were calculated with respect to the single focal locus only.
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) [AJHG 98:653-66] 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
  - CARAT software (Jiang et al. 2016, AJHG 98:243-55) also appears to address this problem effectively
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’
- So modelling effects at all genotyped SNPs explained the ‘known’ heritability (≈ 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
Basic idea is to use formulation

\[ y = X\beta + g + \epsilon \]

with \( g \sim N(0, A\sigma^2_a) \) and \( \epsilon \sim N(0, I\sigma^2_e) \) so \( V = A\sigma^2_a + I\sigma^2_e \)

- \( A \) is the GRM between individuals, estimated using all genotyped SNPs
- \( \sigma^2_a \) and \( \sigma^2_e \) estimated using REML (or MLE)
- Thus we can estimate heritability accounted for by the genotyped SNPs as \( \sigma^2_a/(\sigma^2_a + \sigma^2_e) \)

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
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

$$y = X\beta + \sum_{c=1}^{22} g_c + \epsilon$$

with

$$V = \sum_{c=1}^{22} A_c \sigma^2_c + I \sigma^2_e,$$

where $g_c$ is a vector of effects attributed to the $c$th chromosome, and $A_c$ is the GRM estimated from SNPs on the $c$th 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:

- Davis et al. (2013) partitioned SNPs for Tourette Syndrome and OCD based on chromosome, MAF bin and functional annotation:
  - genic or intergenic (using ANNOVAR)
  - eQTL (using various databases)

- Gusev et al. (2014) [AJHG 95:535-552] partitioned SNPs based on functional and regulatory annotation, for 11 different diseases:
  - coding
  - UTR
  - promoter
  - DNaseI hypersensitivity site (DHS) in any of 217 cell types
  - intronic
  - intergenic
Prediction

- Another benefit of LMMs is that you can transform between the formulations \( y = X\beta + g + \epsilon \) and \( y = X\beta + Zu + \epsilon \) to perform prediction
  - Fit model using formulation \( y = X\beta + g + \epsilon \)
  - Calculate the best linear unbiased prediction (BLUP) of the total genetic effect (the breeding value, \( g_i \)) for all individuals:
    \[
    \hat{g} = A\sigma_a^2 V^{-1} (y - X\hat{\beta})
    \]
  - The BLUP of \( g \) can be transformed to the BLUP of \( u \) by:
    \[
    \hat{u} = Z'A^{-1}\hat{g}/N \quad (\text{where } N \text{ is the number of genotyped SNPs})
    \]
  - We can predict genetic values of individuals in a new data set as:
    \( g_{\text{new}} = Z_{\text{new}} \hat{u} \)
    - And thus predict \( y_{\text{new}} \)
  - Conceptually similar to the idea of polygenic risk scores, except that we use all of the SNPs for prediction
Conclusions and recent developments

- LMM approach has proved highly useful and flexible for a variety of different purposes
- However, the approach can be computationally intensive
  - And requires access to individual-level data
- For association testing, summary statistics from LMMs can be combined by meta-analysis in the usual way
  - Provided you have an effect estimate (regression coefficient or log OR) and its standard error
Conclusions and recent developments

Some recent work has focussed on estimating (a) heritability explained by sets of SNPs, and (b) genetic correlations across traits, using summary statistics only

- 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’

Another recent approach uses multi-variate mixed models to impute missing phenotypes

- Dahl et al. (2016) [Nat Genet 48:466-472]
- Implemented in PHENIX software