Using Mixed Models in GWA studies (slides available at http://tinyurl.com/ge052013)

Yurii Aulchenko

yurii [dot] aulchenko [at] gmail [dot] com

February 14, 2013

Outline

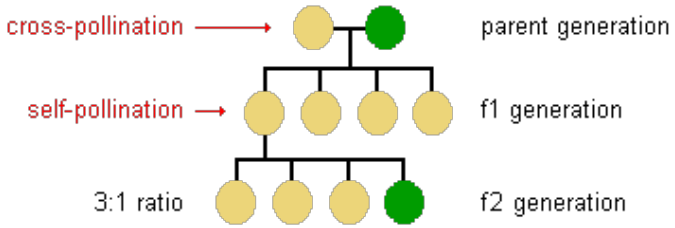
1. **Introduction to MM for GWAS**
   - Fisher's polygenic model
   - Hypothesis testing for polygenic model
2. **Fast approximate MMs**
   - FASTA-like tests
   - GRAMMAR-like tests
3. **Misc**
   - MLMM
   - Using MM's in structured populations
   - Estimation of Φ
4. **Summary**

Contents

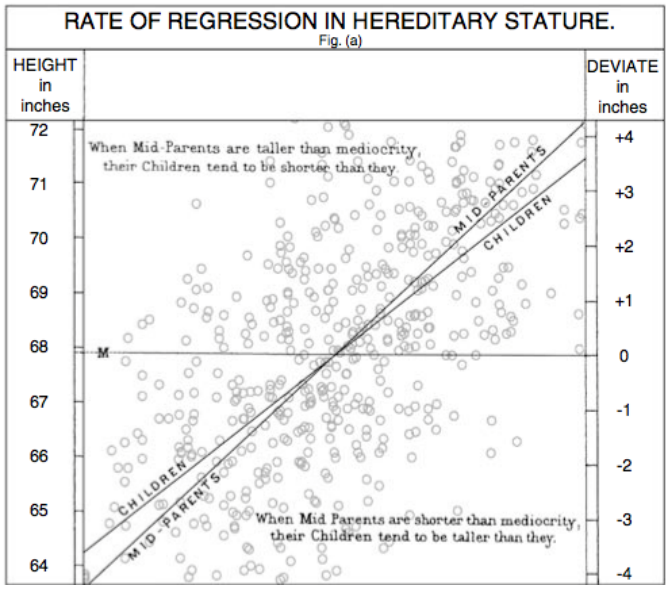
1. **Introduction to MM for GWAS**
   - Fisher's polygenic model
   - Hypothesis testing for polygenic model
2. **Fast approximate MMs**
   - FASTA-like tests
   - GRAMMAR-like tests
3. **Misc**
   - MLMM
   - Using MM's in structured populations
   - Estimation of Φ
4. **Summary**

Controversy between the Mendelian and Galtonian genetics

Mendelian: discrete traits, discrete genetic 'factors'



Galtonian: continuous traits, correlations between relatives



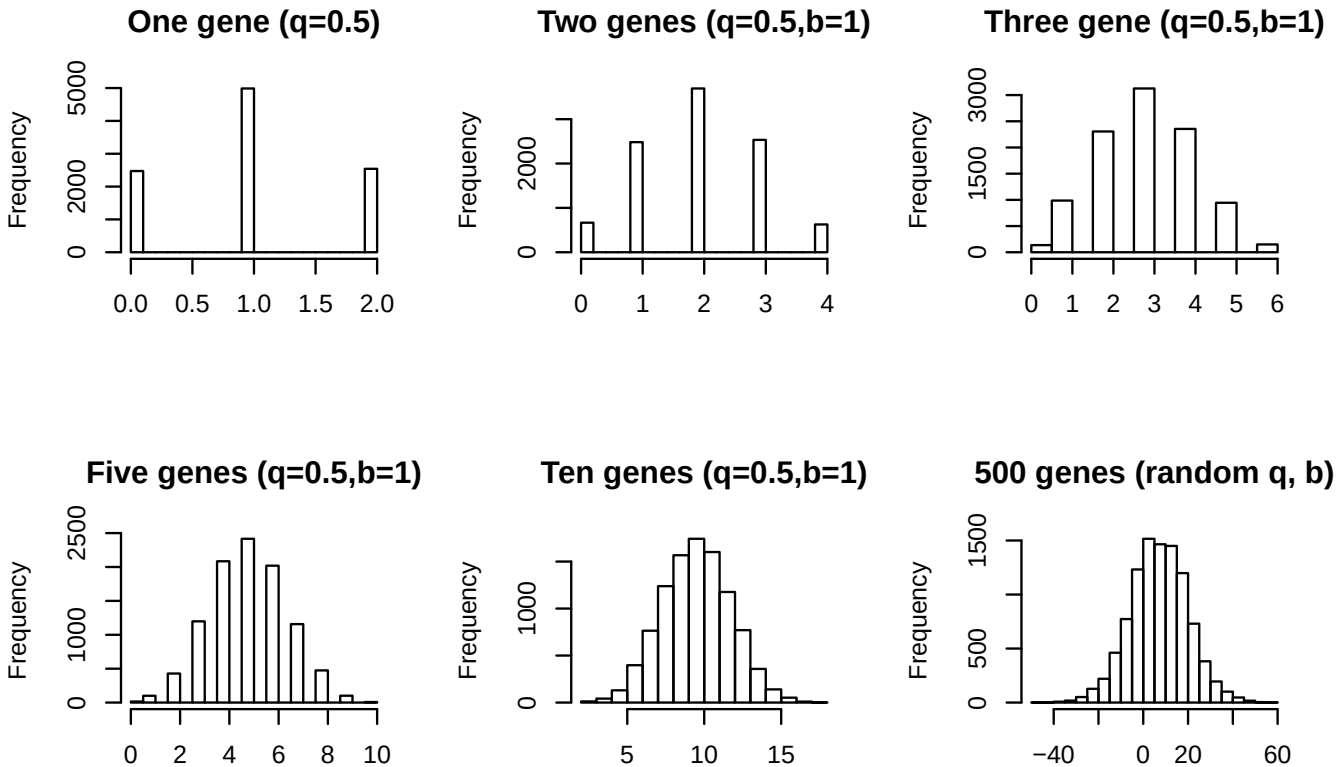
Fisher's (1918) polygenic model

Assumption: the value of a quantitative trait is determined by *additive effects of many* Mendelian loci and the effects of the environment. The value of the trait of i-th individual is

$$y_i = \mu + \sum_{j=1}^{M} g_{ij} b_j + \epsilon_i$$

where μ is a constant (intercept), g_{ij} is a genotype of the i-th individual at j-th locus, b_j is an effect of j-th locus (M loci in total), and ϵ_i is a random error term (assumed to be distributed normally with mean zero and some variance σ_e^2). Such model can describe continuous distribution and correlation between the phenotypes of relatives observed.

Fisher's model with different M



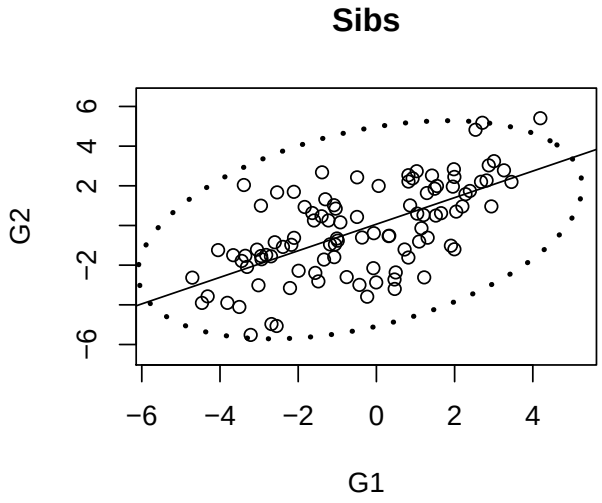
Correlations between relatives

- When the number of loci M is large, the distribution of the genetic effects $G_i = \sum_{j=1}^{M} g_{ij} b_j$ can be approximated by the normal distribution
- The relatives share large proportions of their genomes IBD (and hence IBS), and therefore the phenotypes of relatives are correlated
- To describe the joint distribution of phenotypes of relatives, we can make use of *multivariate normal distribution*

Distribution of polygene G for pairs of relatives

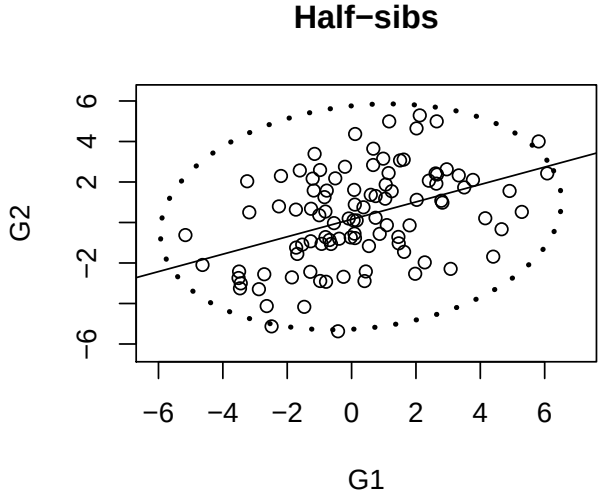
Sibs share 50% of their genome IBD, and therefore the joint distribution is described by bi-variate normal with correlation matrix

$$\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$



Half-sibs share 25% of their genomes IBD, and hence the correlation matrix is

$$\begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}$$



Distribution of G for a sample of related individuals

- In general, the distribution of the polygene G for N individuals is described by N-variate normal with NxN correlation matrix Φ
- The elements of Φ consist of *relationship coefficients* for pairs of relatives, ϕ_{kl} - the expected proportions of genome k and l share IBD

Fisher's model (1)

The model

$$y_i = \mu + \sum_{j=1}^{M} g_{ij} b_j + \epsilon_i$$

when M is large can be approximated with

$$y = \mu + G + \epsilon$$

where G comes from multivariate normal with mean zero and VC-matrix $\sigma_G^2 \Phi$ and ϵ from a multivariate normal with mean zero and VC-matrix $\sigma_e^2 I$.

Fisher's model (2)

- This equation describes a multivariate normal with expectation $E[y] = \mu$ and VC-matrix Ω defined by $var_i = \sigma^2_G + \sigma^2_e$ and $cov_{ij} = \sigma^2_G \phi_{ij}$
- Denote total variance $\sigma^2_G + \sigma^2_e$ as σ^2 and $h^2 = \sigma^2_G / \sigma^2$ (the coefficient of heritability: proportion of total variance explained by additive genetic effects)
- Then, the correlation between the phenotypes of a pair of relatives i and j is

$$\rho_{ij} = \frac{cov_{ij}}{\sqrt{var_i var_j}} = \frac{\sigma^2_G \phi_{ij}}{\sigma^2} = h^2 \phi_{ij}$$

- Hence h^2 has clear relation to expected correlation between the phenotypes of relatives

Introducing fixed effects in polygenic model

- Fixed effects can be easily introduced in the polygenic model (Boerwinkle, 1986), e.g.

$$y_i = \mu + \beta_{sex} S_i + \beta_g g_i + G_i + \epsilon_i$$

  describes a model with fixed effects of sex S and genotype at some specific locus, g
- Introducing fixed effects does not change the multi-variate's normal VC-matrix, but re-defines the expectation as

$$E[y_i] = \mu + \beta_{sex} S_i + \beta_g g_i$$

  (instead of $E[y_i] = \mu$ in previous model)

Relation to mixed models

- The Fisher's polygenic model with fixed effects is a particular type of what is known in statistics as *mixed models* - the models including both fixed (in our example, g) and random (G) effects
- In mixed models, the *fixed effects* are these familiar from the standard linear models – factors, which we can measure directly and include in the model
- The *random effects* are not directly measured, but we know (assume) their distributed

Log-Likelihood under polygenic model

The log-likelihood function for model

$$y_i = \mu + \beta_g g_i + G_i + \epsilon_i$$

is defined by the multivariate normal distribution and can be written as

$$lnL(y|\theta) \propto -\frac{1}{2}\left(ln|\Omega| + (y - E[y])^T \Omega^{-1}(y - E[y])\right)$$

This function can be maximized over the θ giving Maximum Likelihood (ML) $\hat{L}$ and ML estimates (MLE) $\hat{\theta}$.

Likelihood ratio test (LRT) for fixed effect

- The hypotheses concerning model's parameters can be tested in hierarchical manner using the Likelihood Ratio Test (LRT)
- For example, testing the hypothesis that $\beta_g = \hat{\beta}_g$ versus the null hypothesis of $\beta_g = 0$ can be done with

$$LRT = 2ln\frac{L(y|\hat{\theta}_1)}{L(y|\hat{\theta}_0)} = 2[lnL(y|\hat{\theta}_1) - lnL(y|\hat{\theta}_0)]$$

  where $\hat{\theta}_1 = \{\hat{\mu}, \hat{\beta}_g, \hat{h^2}, \hat{\sigma^2}\}$ and $\hat{\theta}_0 = \{\hat{\mu}, \beta_g = 0, \hat{h^2}, \hat{\sigma^2}\}$
- Under the null hypothesis, the LRT is distributed as χ^2 with the number of degrees for freedom equal to the number of parameters constrained (χ^2_1 in the above case)

GWAS using Mixed Models

- The polygenic model with fixed effects (mixed model)

$$y_i = \mu + \beta_g g_i + G_i + \epsilon_i$$

allows for GWAS using polygenic (mixed) model by testing the hypothesis $\beta_g = \hat{\beta}_g$ vs. $\beta_g = 0$ for each SNP in GWAS in turn

- **PS: Why do we bother?** – when the sample consist of related individuals, not accounting for correlations between the phenotypes of relatives leads to inflation of the test statistic (false positives). Also, in general, the model best describing the nature of the data is the most powerful

Using ML and LRT in GWAS

- When using ML/LRT as described above the computational complexity may be an issue: for every SNP tested, we need to estimate $\hat{\theta}_1$ and $\hat{\theta}_0$ - and this is quite laborious procedure
- In 2007 (Aulchenko et al.), for sample size of about 1000, testing single SNP was taking about 15 minutes
- Remarkable progress was achieved in last few years by using smarter algorithms (FMM of Astle and Balding, 2010; FaST-LMM of Lippert et al., 2011)
- Still, when it comes to millions of markers and analyses of multiple traits, the ML/LRT may be an expensive option

Contents

1. Introduction to MM for GWAS
   - Fisher's polygenic model
   - Hypothesis testing for polygenic model
2. Fast approximate MMs
   - FASTA-like tests
   - GRAMMAR-like tests
3. Misc
   - MLMM
   - Using MM's in structured populations
   - Estimation of Φ
4. Summary

Two-step estimation

- The main problem is estimation of h^2 each time we introduce new SNP into the model: if we knew h^2, the estimation of other parameters would be straightforward
- If we assume that a SNP has small effect on the trait, then its inclusion into the model should not change the estimate of h^2 much
- Therefore two-step estimation approach can be used:

First , estimate h^2 using MM without SNP: $y_i = \mu + G_i + \epsilon_i$

Second , use the same estimate $\hat{h}^2$ to correct the test of association for every SNP genome-wide

FASTA (Chen and Abecasis, 2007)

- The obtained estimates are used to construct the variance-covariance matrix for the data, $\hat{\Omega}$
- Score test is constructed accounting for $\hat{\Omega}$:

$$T_i^2 = \frac{(\bar{g_i}^T \hat{\Omega}^{-1} \bar{Y})^2}{\bar{g_i}^T \hat{\Omega}^{-1} \bar{g_i}}$$

  is distributed as χ_1^2 under the null hypothesis
- Similar ideas were suggested and implemented in the 'mmscore' of the GenABEL project (2008), P3D/Tassel (Zhang et al., 2010) and EMMAX (Kang et al., 2010)

Adjustment for additional covariates in two-step procedures

- With original FASTA method, the adjustment for covariates is done during the first step only, and adjusted residuals are used in the second step
- This is fine as far as there is no covariance between covariates and genotypes (most situations)
- If covariance is present (e.g. covariates are "genetic strata" or inherited traits) above approach may lead to conservative test ($\lambda < 1$)
- The Generalized Least Squares procedure, which allows keeping covariates at both steps, is recommended (implemented in ProbABEL::mmscore (Aulchenko et al., 2010), MixABEL::GWFGLS and EMMAX)

Running time for FASTA type of methods

- Note that term $\hat{\Omega}^{-1}\bar{g}_i$ implies multiplication of a vector of length N by an NxN matrix for every SNP. Thus the computational time is $O(N^2)$
- While this is usually not a big problem for most studies, it acn become a problem for larger (say, $N > 5000$) studies and/or analysis of multiple traits (e.g. exploration of 'omics' space)

GRAMMAR (Aulchenko et al, 2007)

- Another type of two-step approach

First , estimate h^2 using MM without SNP: $y_i = \mu + G_i + \epsilon_i$

Second , the obtained estimates are used to compute *environmental residuals*, $y^* = \hat{\epsilon}_i$

- These residuals are not correlated between relatives, and thus any standard association method can be used for analysis, e.g. the score test

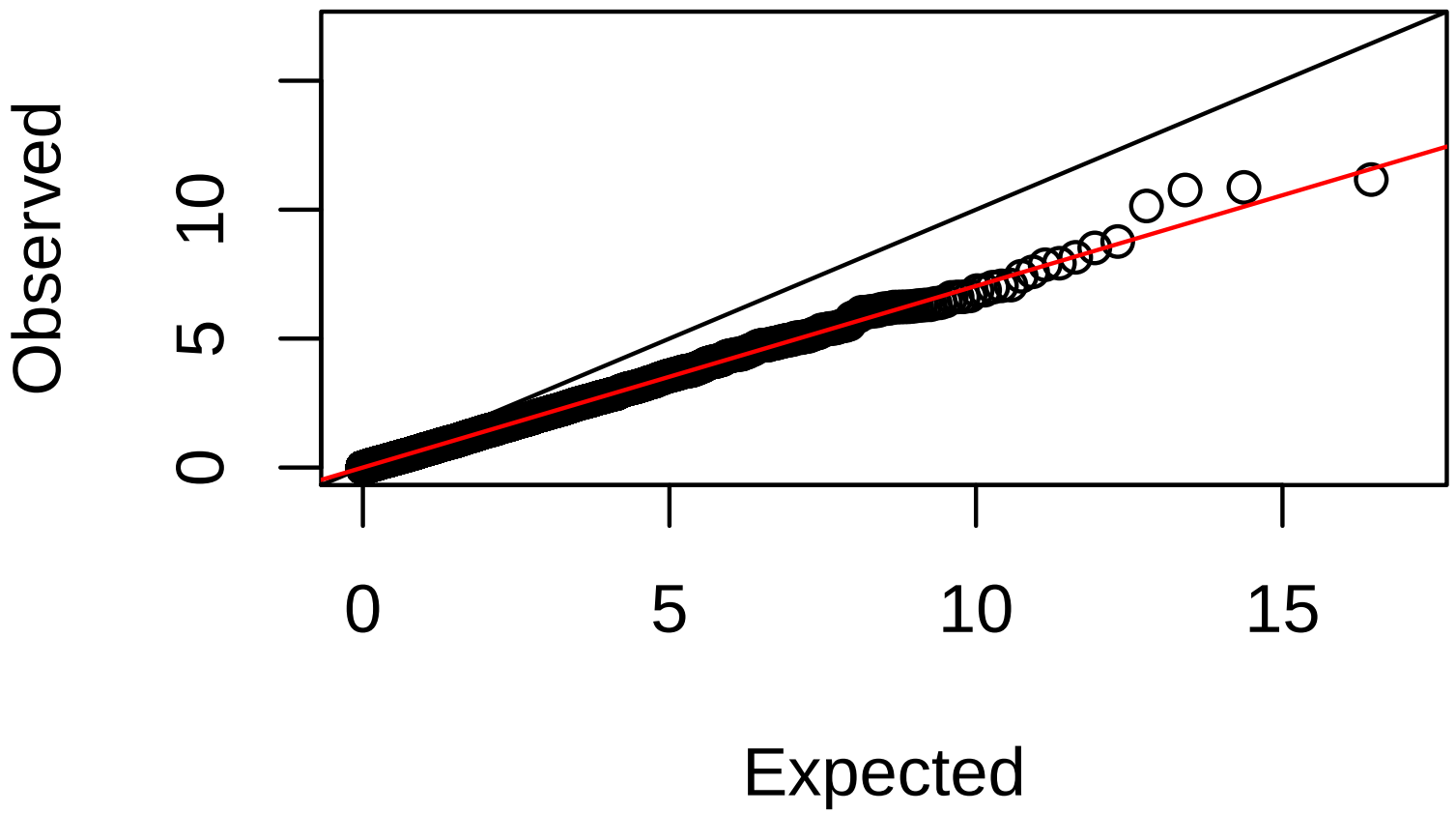
$$T_i^2 = \frac{(\bar{g_i}^T y^*)^2}{\bar{g_i}^T \bar{g_i}}$$

- Advantage of this method is that analysis of transformed trait is very fast (much faster than FASTA/mmscore/EMMAX), and wide variety of methods developed for population-based studies can be used
- Disadvantage of this method is that it results in biased estimates of β and conservative test statistics (false negatives

GRAMMAR estimates are biased

Pedigree:			Analysis method	
h^2_{QTL}	Simulated effect	h^2	MG	GRAMMAR
NP				
0.01		0.3	0.234 ± 0.077	0.149 ± 0.053
	0.236	0.5	0.237 ± 0.078	0.106 ± 0.039
		0.8	0.238 ± 0.077	0.044 ± 0.017
0.02		0.3	0.334 ± 0.077	0.213 ± 0.053
	0.333	0.5	0.336 ± 0.078	0.149 ± 0.039
		0.8	0.334 ± 0.077	0.062 ± 0.017
0.03		0.3	0.408 ± 0.077	0.259 ± 0.053
	0.408	0.5	0.411 ± 0.078	0.183 ± 0.039
		0.8	0.411 ± 0.076	0.076 ± 0.017
ERF				
0.01		0.3	0.236 ± 0.079	0.149 ± 0.054
	0.236	0.5	0.237 ± 0.081	0.105 ± 0.041
		0.8	0.234 ± 0.078	0.044 ± 0.018
0.02		0.3	0.333 ± 0.082	0.209 ± 0.056
	0.333	0.5	0.336 ± 0.079	0.150 ± 0.040
		0.8	0.344 ± 0.078	0.068 ± 0.018
IPP				
0.01		0.3	0.235 ± 0.089	0.125 ± 0.058
	0.236	0.5	0.237 ± 0.088	0.093 ± 0.043
		0.8	0.238 ± 0.082	0.044 ± 0.019
0.02		0.3	0.334 ± 0.089	0.178 ± 0.058
	0.333	0.5	0.335 ± 0.087	0.131 ± 0.043
		0.8	0.329 ± 0.083	0.058 ± 0.019

... so are the test statistics values

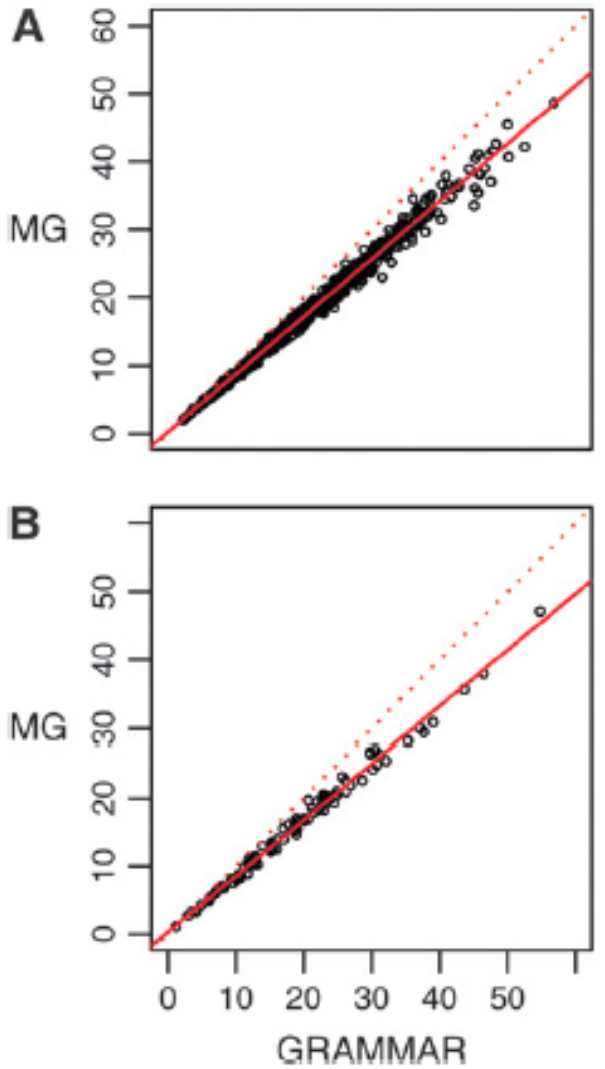


GRAMMAR + reverse Genomic Control (Amin et al, 2007)

- Estimate polygenic model and compute GRAMMAR test statistics T_i^2 genome-wide
- Estimate GC λ (< 1) in usual manner (e.g. $\hat{\lambda} = \frac{median(T_1^2, T_2^2, ...)}{0.455}$)
- Derive GRAMMAR-GC test statistics as $T_{i,GC}^2 = \frac{T_i^2}{\hat{\lambda}}$

- Solves the conservatively of the test
- Does not solve the problem of the effect under-estimation
- Does not leave means to judge if MM-correction was adequate for the data (λ is 1 by definition of GRAMMAR-GC!)
- Is an approximation: but how we judge if it works well for this particular data?

There is good correlation between FASTA and GRAMMAR results (Aulchenko et al., 2007)



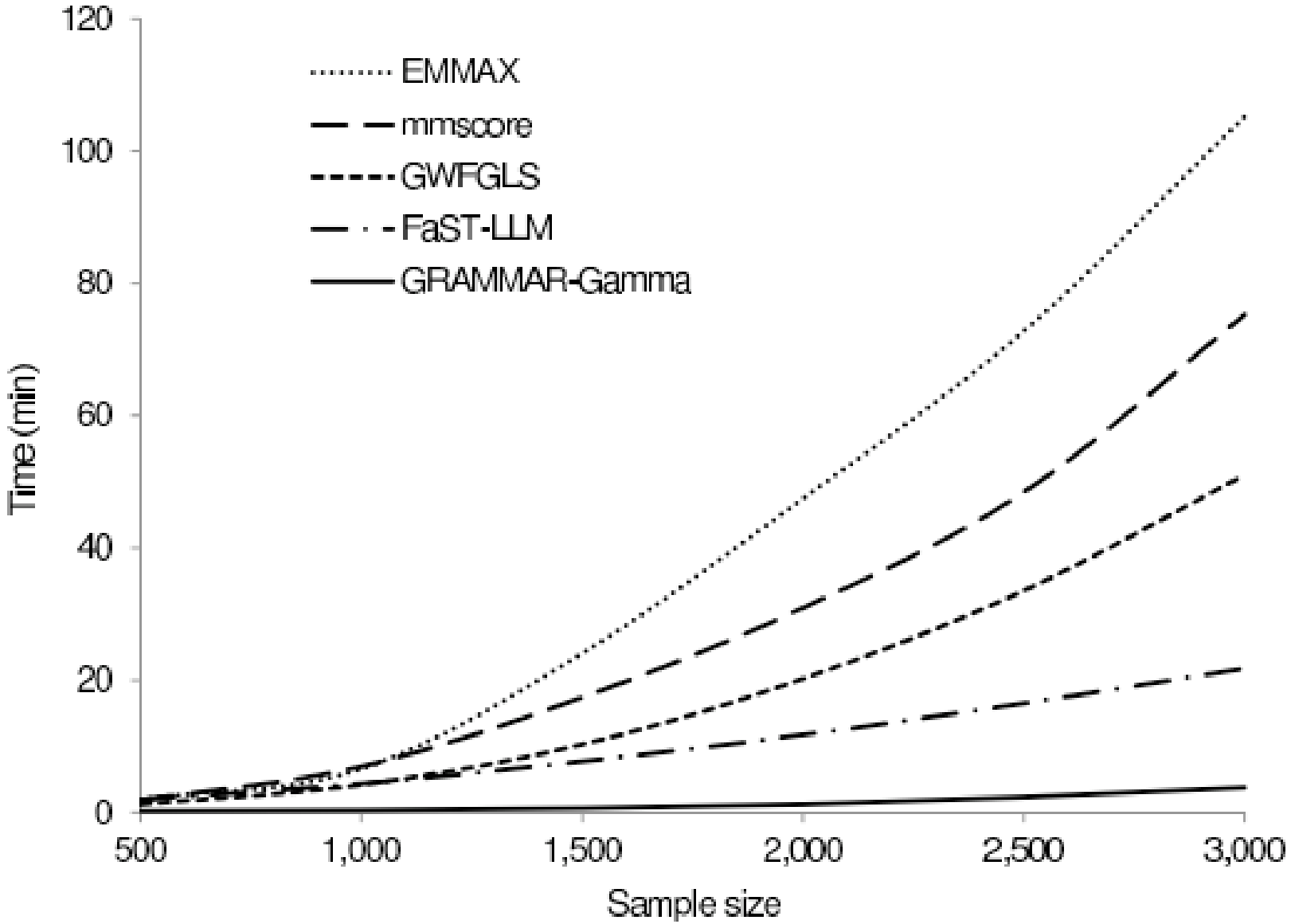
GRAMMAR-γ (Svishcheva et al., 2012)

- The bias in test statistics and effect estimates is proportional to some constant, $\gamma = F(\hat{\Omega})$
- Instead of correction of test statistics only with GC, allows correction of both test statistics and effect estimates

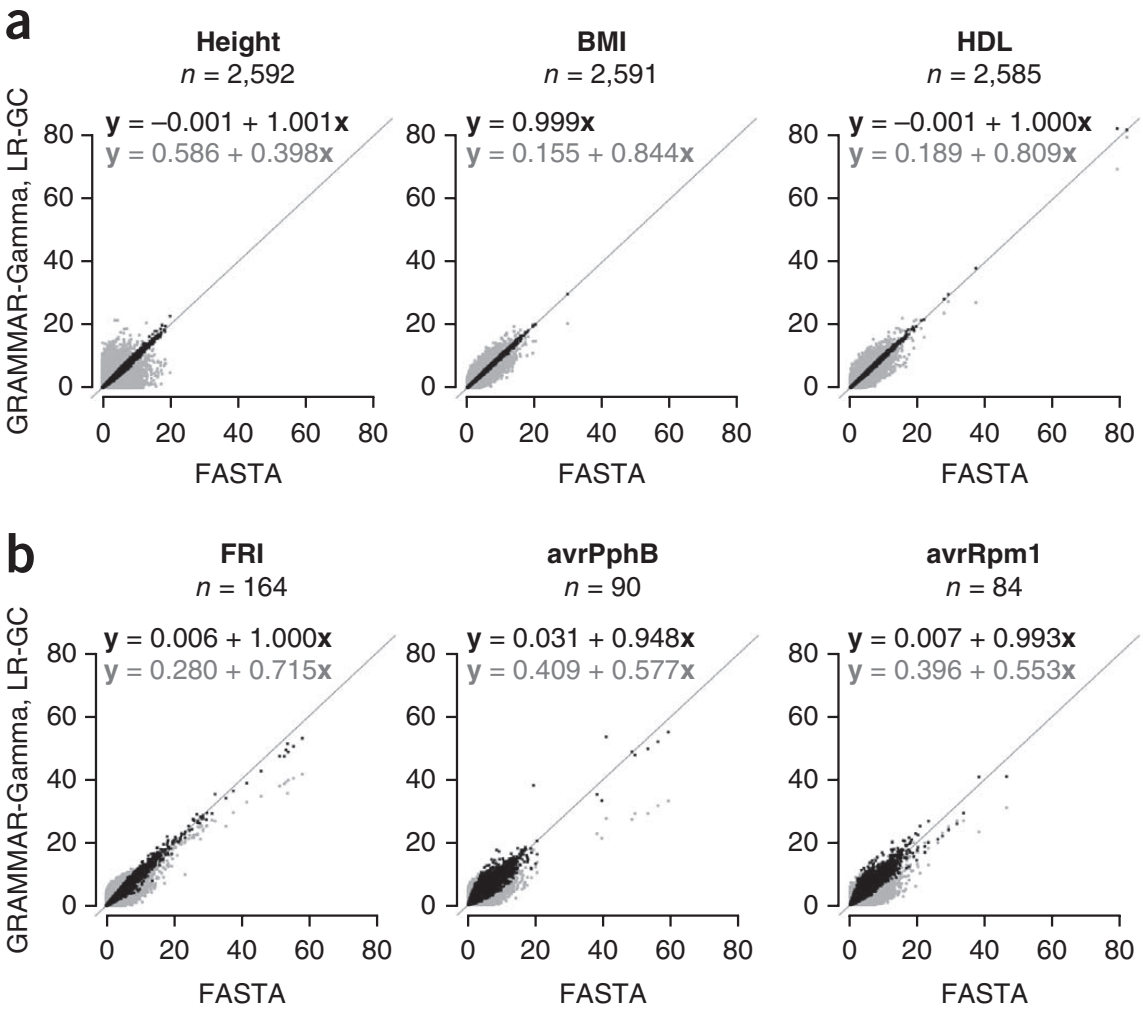
$$T_i^2 = \frac{1}{\gamma} \frac{(\bar{g_i}^T y^*)^2}{\bar{g_i}^T \bar{g_i}}$$

- Solves the conservatively of the test
- Solves the problem of the effect under-estimation
- Provides the means to judge if correction was adequate for the data (how much λ deviate from 1?)
- Provides means to check if approximation was good for particular data

Speed comparison between methods (500k SNPs)



Accuracy of Grammar-γ approximation



Grey dots: FASTA vs LM. Black dots: Grammar-γ vs FASTA. Upper row: human data (γ is almost the same); lower row: *A. thaliana* highly structured data (less accurate) (*Svishcheva et al., Nat Genet, 2012*)

Contents

1. Introduction to MM for GWAS
   - Fisher's polygenic model
   - Hypothesis testing for polygenic model
2. Fast approximate MMs
   - FASTA-like tests
   - GRAMMAR-like tests
3. Misc
   - MLMM
   - Using MM's in structured populations
   - Estimation of Φ
4. Summary

Multi-locus Mixed Models (Segura et al., 2012)

- The idea is to perform iterative GWAS: after first round, the most significant SNP is included into the model as a fixed effect, and the scan is repeated; best SNP included into fixed part of the model again; so on
- The method has better location accuracy: for a particular locus, it is quite typical to see multiple signals because of LD. MLMM can refine these associations and indicate only independent signals
- The method may also have an improved power – in case loci explaining large proportion of trait's variance are present (such as the case in 'omics' studies)

Structure of NFBC66 sample (Kang etl., 2010)

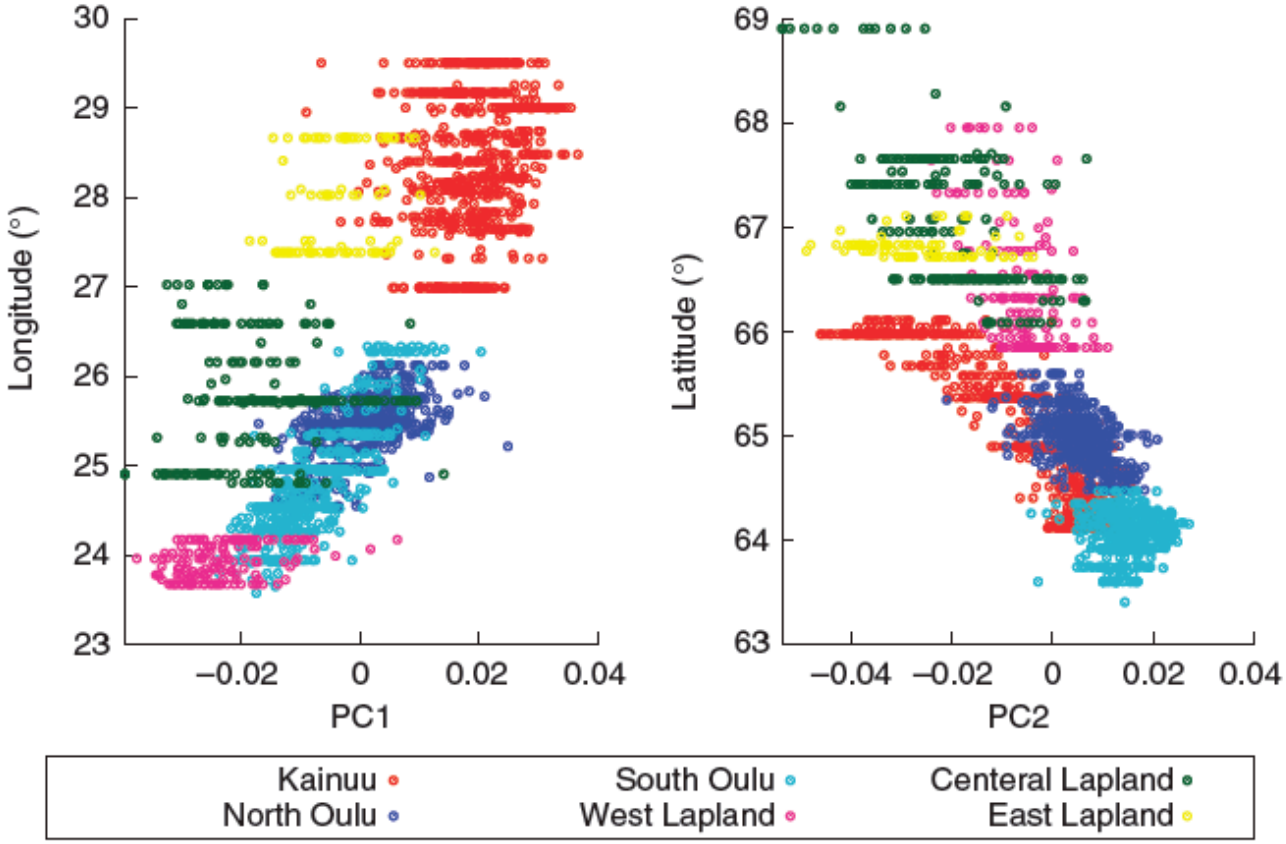


Figure 1 Scatter plots of the first two principal components against latitude and longitude. Only individuals of known ancestry are included in the plot. Latitude and longitude are defined as the average latitude and longitude of the parents' birthplaces. Colors indicate linguistic or geographic subgroups.

Genomic control λ for different methods (Kang etl., 2010)

Table 1 Comparison of genomic control inflation factors obtained with different models

	Genomic control inflation factor			
Phenotype	Uncorrected	IBD < 0.1	ES100	EMMAX
CRP	1.007	1.007	1.019	0.993
TG	1.023	1.010	1.019	1.002
INS	1.029	1.022	1.013	1.005
DBP	1.031	1.019	1.028	1.007
BMI	1.031	1.024	1.016	0.995
GLU	1.045	1.033	1.030	1.008
HDL	1.052	1.056	1.036	1.004
SBP	1.066	1.056	1.021	1.006
LDL	1.098	1.089	1.040	1.002
Height	1.187	1.151	1.074	1.003

ES100, EIGENSOFT correcting for 100 principal components; IBD < 0.1, uncorrected analysis after excluding 611 individuals whose PLINK's IBD estimates with another individual is greater than 0.1; phenotype abbreviations are CRP, C-reactive protein; TG, triglyceride; INS, insulin plasma levels; DBP, diastolic blood pressure; BMI, body mass index; GLU, glucose; HDL, high-density lipoprotein; SBP, systolic blood pressure; LDL, low density lipoprotein.

Estimation of relationship matrix Φ

- If pedigree is know, Φ can be easily estimated from these data
- Genome-wide information provides means to do so in absence of pedigree information as well
- One of the most accepted methods assumes computation of

$$\phi_{ij} = \frac{1}{M} \sum_{k=1}^{M} \frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1 - p_k)}$$

  which was shown to be unbiased estimator of kinship
- In human genetically isolated populations it seems that using the "genomic kinship" provides is better than pedigree-based kinship
- The plain IBS matrix works better in highly structured populations such as plants

Contents

1. Introduction to MM for GWAS
   - Fisher's polygenic model
   - Hypothesis testing for polygenic model
2. Fast approximate MMs
   - FASTA-like tests
   - GRAMMAR-like tests
3. Misc
   - MLMM
   - Using MM's in structured populations
   - Estimation of Φ
4. **Summary**

Summary of MM-based methods

- Fast implementations of LRT-based VC test are available now (MixABEL::FMM of W. Astle, FaST-LMM of Lippert et al.). These are theoretically superior!
- Two-Step FASTA-like approximations are excellent unless SNPs have large effects. Speed-up achieved cf. LRT is 5-20 times. Implementations include ProbABEL, MixABEL::GWFGLS, GenABEL::mmscore, EMMAX, P3D/Tassel, FaST-LMM
- Grammar type of analyses have superior speed and can be use to analyze tens of millions of SNPs in many thousands of individuals. Caution should be exercised when analyzing data with uneven relationship structure (e.g. plants lines/stocks data).

Summary of advantages of use of MM in GWAS

- MM can account well for complicated relationship structure. Such structure is typical for family-based design, genetically isolated populations, outbred animal data, but can also be found in contemporary large "population-based" studies.
- The advantages of use of MM will become more and more visible with increased sample sizes
- MM provides natural means to study complex designs, such as twin data and repeated measurements
- Use of optimal algebraic kernels and effective implementation will be critical for effective analysis of statistically complex problems