Using Mixed Models in GWA studies

Yurii Aulchenko

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

March 27, 2012

Outline

1. **Mixed Models (MM)**
   - Intro to MM for GWAS
   - Use of MM in population-based studies
2. **Genome-wide feasible MM**
   - FASTA-like tests
   - GRAMMAR-like tests
   - Comparison
3. **Optimal algebraic kernels and implementation**
4. **Conclusions**

Contents

1. **Mixed Models (MM)**
   - Intro to MM for GWAS
   - Use of MM in population-based studies
2. **Genome-wide feasible MM**
   - FASTA-like tests
   - GRAMMAR-like tests
   - Comparison
3. **Optimal algebraic kernels and implementation**
4. **Conclusions**

Mixed Models – an outline

- In the framework of Mixed Models, the expectation of the outcome y is modeled using sum of *fixed* and *random* effects.
- Fixed effects are these familiar from the standard linear models – factors, which we can measure directly and include in the model
- Random effects are not directly measured, but we know (assume) their distributed
- In fact, standard linear regression model

$$y_i = \mu + \beta \cdot g_i + \epsilon_i$$

does contain random effect – residual error ϵ_i, which is not measurable, but is assumed to come from Normal distribution with mean zero and some variance σ^2

Correlations between phenotypes of relatives

- In a sample of related individuals, the assumption of independence between measurements (outcome y) does not hold because phenotypes of relatives are correlated (because traits are controlled by genome, and related individuals share genomes!)
- The strength of control of the trait by genome can be characterized by heritability, h^2
- The relationship between a couple of relatives i and j is characterized by coefficient of relationship ϕ_{ij}, which is (the expected) proportion of the genome shared identical-by-descent

Correlations between phenotypes of relatives – continued

- Correlation of phenotypes of relatives i and j depends on the degree of relatedness ϕ_{ij}, and the heritability of the trait h^2: $\rho_{ij} = \phi_{ij} \cdot h^2$
- For example, if $h^2 = 0.9$ (e.g. height) the correlation between the phenotypes of sibs is expected to be $0.5 \cdot 0.9 = 0.45$ and correlation between phenotypes of uncle and niece would be $0.25 \cdot 0.9 = 0.225$
- To do correct association analysis **we need to account for this correlation structure in association model**
- This can be done by introducing *random polygenic effect* with correlation matrix whith elements $\rho_{ij} = \phi_{ij} \cdot h^2$

Accounting for relationship in MM

- More formally, we can describe the distribution of phenotypes y in a sample of related individuals with MM

  $$y_i = \mu + \beta \cdot g_i + G_i + \epsilon_i,$$

  where G is distributed as multivariate Normal with variance-covariance matrix proportional to the relationship matrix
- This model is described by 4 parameters: $\{\mu, \beta, h^2, \sigma^2\}$
- Standard way to test significance of β would be to estimate this model and compare it with the model restraining β to zero
- The problem is that already for sample size of about 1000, testing single SNP may take about 15 minutes! (2007)

Estimation of relationship matrix Φ

- If pedigree is know, Φ can be easily estimated from these data
- However, genome-wide information provides means to do so in absence of pedigree information as well:

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

- It seems that using the ”genomic kinship” provides is better than pedigree-based kinship, at least when working with human data

Structure of NFBC66 sample (Kang etl., 2010)

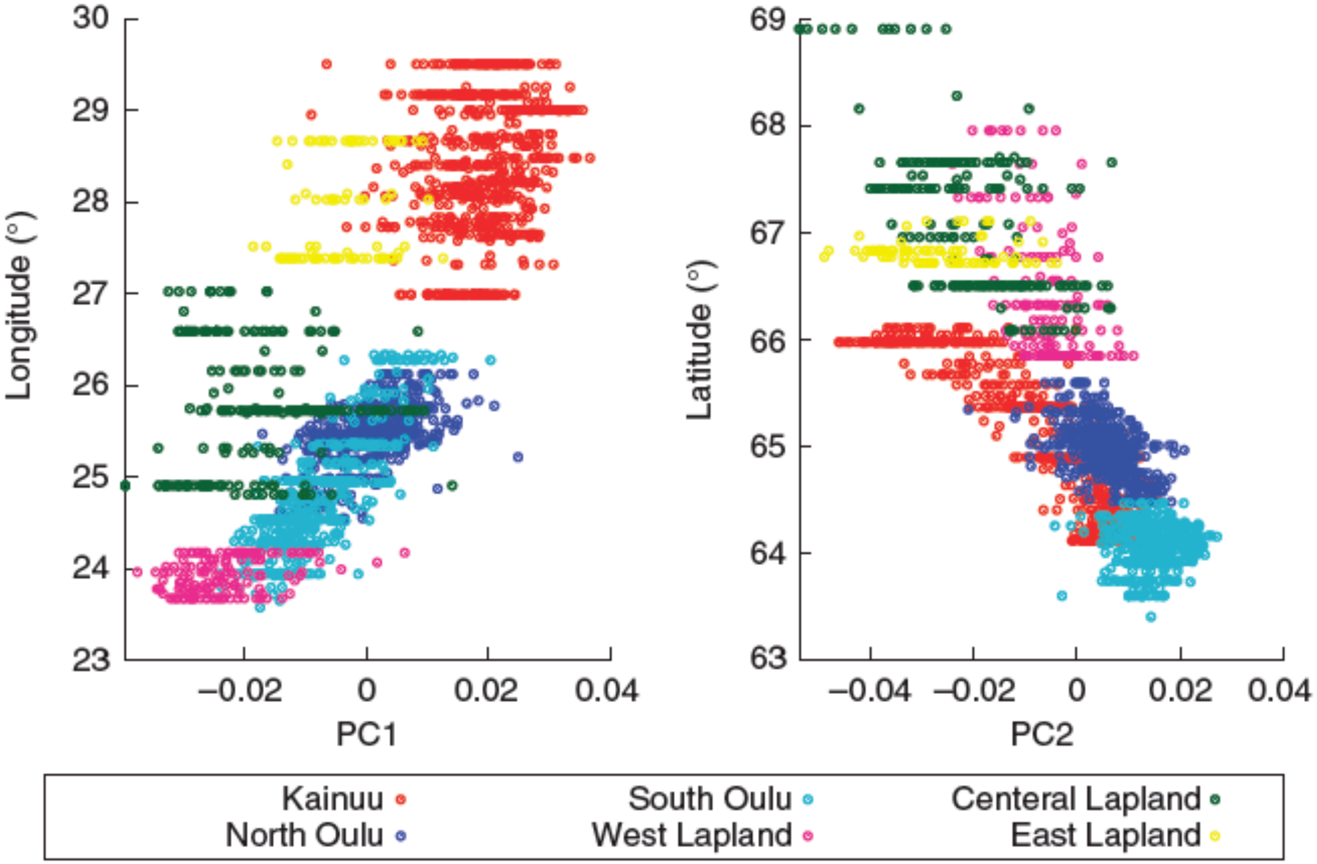


Figure 1 Scatter plots of the first two principal components against latitude and longitude. Only individuals of known ancestry are included in the plot. Latitude and longitude are defined as the average latitude and longitude of the parents' birthplaces. Colors indicate linguistic or geographic subgroups.

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

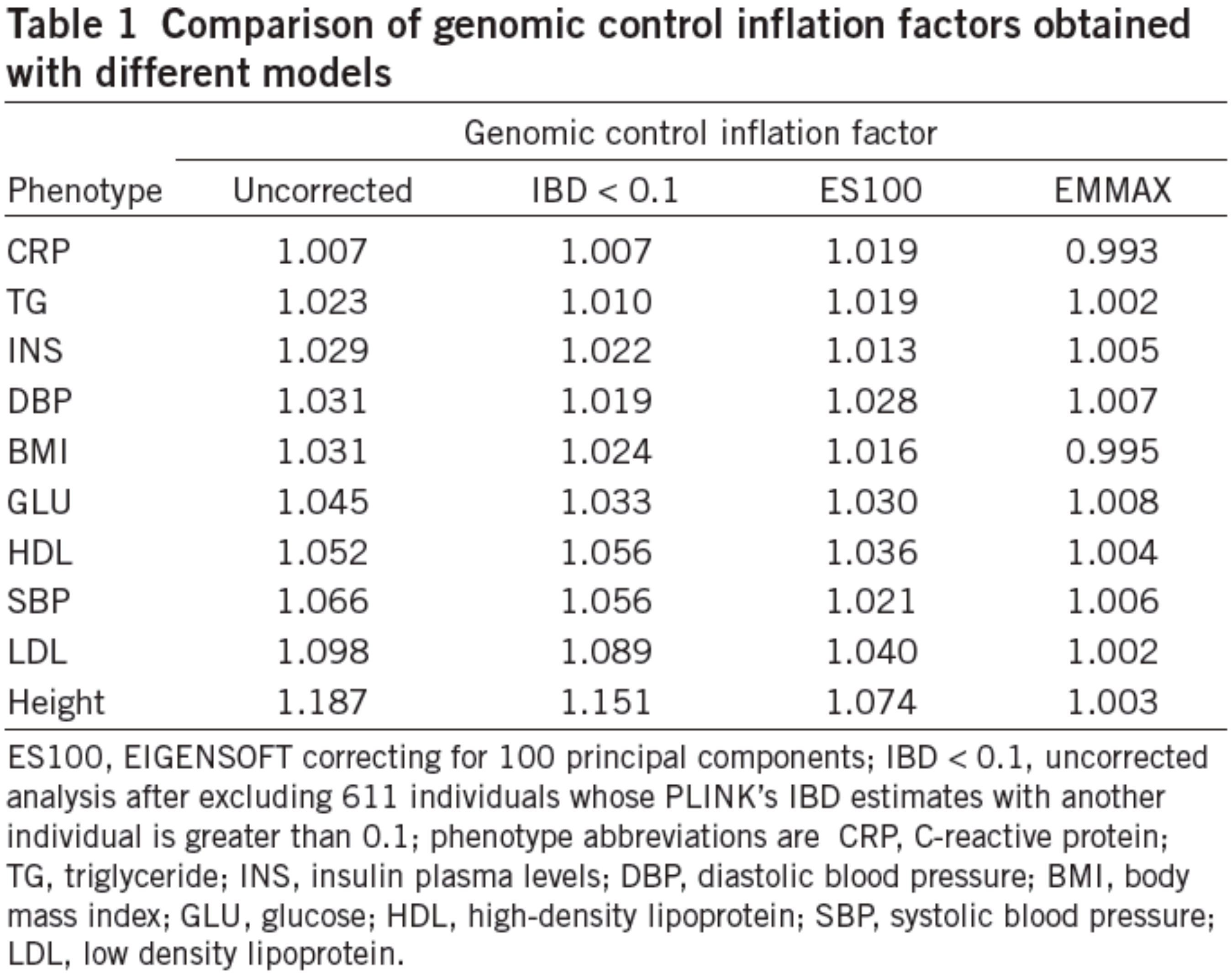


Table 1 Comparison of genomic control inflation factors obtained with different models

	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.

Contents

1. Mixed Models (MM)
   - Intro to MM for GWAS
   - Use of MM in population-based studies
2. Genome-wide feasible MM
   - FASTA-like tests
   - GRAMMAR-like tests
   - Comparison
3. Optimal algebraic kernels and implementation
4. Conclusions

Two-step estimation

- The main problem is estimation of h^2 each time we introduce new SNP into the model
- 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: $y_i = \mu + G_i + \epsilon_i$
  - Use the same estimate $\hat{h}^2$ to correct the test of association for every SNP genome-wide

FASTA (Chen and Abecasis, 2007)

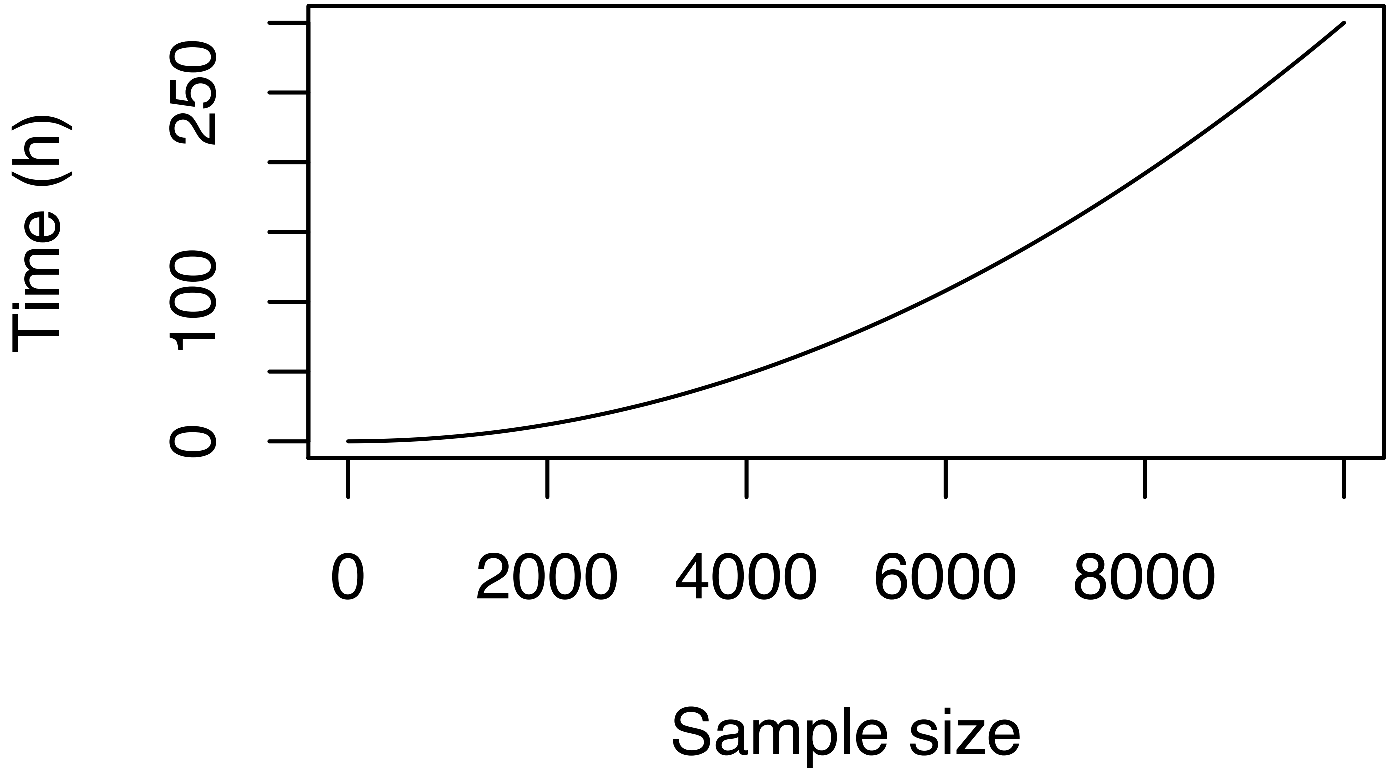
- The obtained estimates are used to construct the variance-covariance matrix for the data, $\hat{\Omega}$
- Score test is constructed accounting for $\hat{\Omega}$:

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

Adjustment for covariates in two-step procedures

- With original FASTA (Merlin, GenABEL::mmscore), adjustment for covariates is done during the first step, and adjusted residuals are used in the second step
- This 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 $(\lambda < 1)$
- ProbABEL::mmscore (Aulchenko et al., 2010), MixABEL::GWFGLS and EMMAX (Kang et al, 2010) allow to keep covariates in both steps

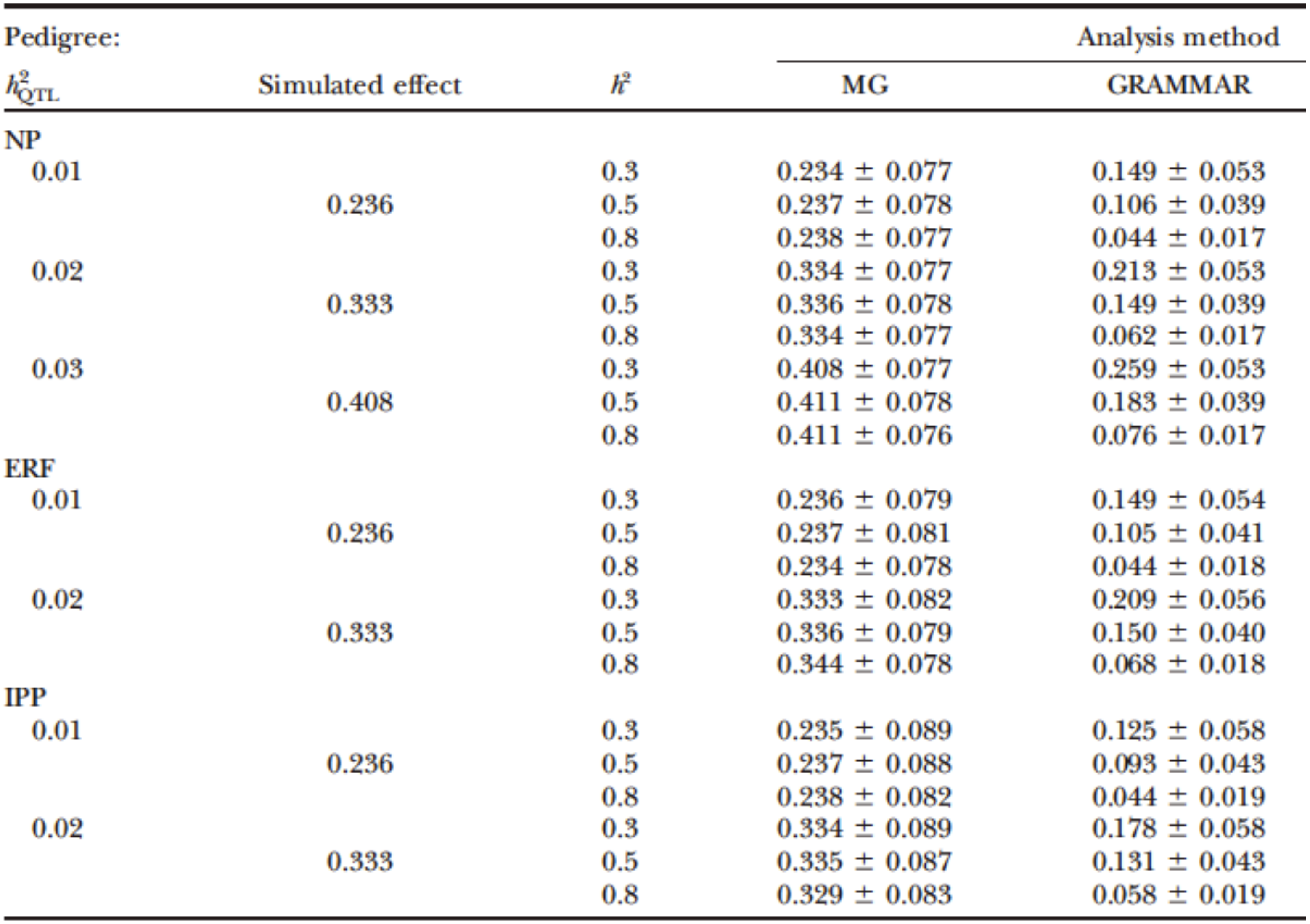
Running time for FASTA (10M SNPs)



GRAMMAR (Aulchenko et al, 2007)

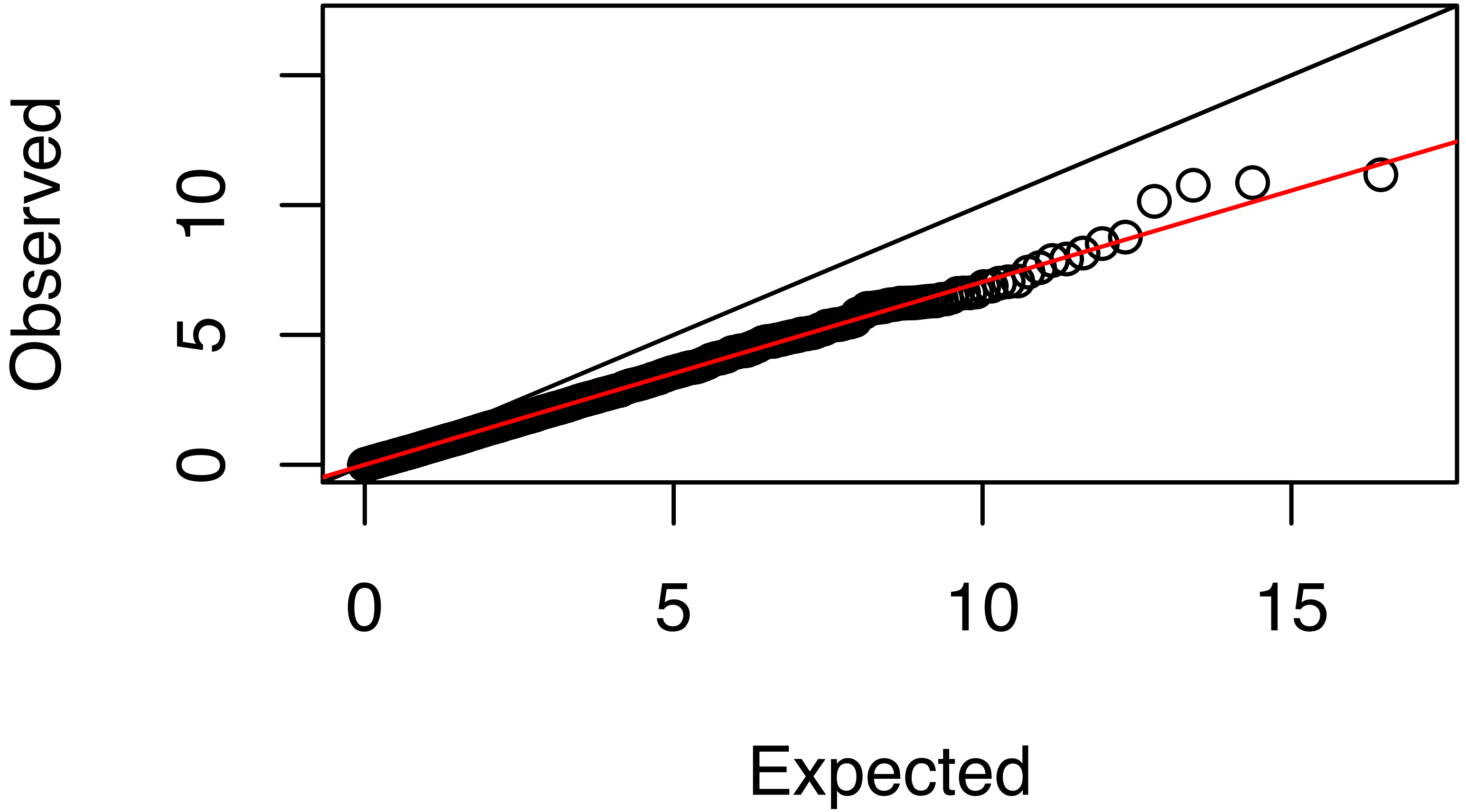
- The obtained estimates are used to compute *environmental residuals*, $\hat{\epsilon}_i$
- These residuals are not correlated between relatives, and thus any standard association method can be used for analysis
- Advantage of this method is that analysis of transformed trait is very fast (much faster 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	0.236	0.3	0.234 ± 0.077	0.149 ± 0.053
		0.5	0.237 ± 0.078	0.106 ± 0.039
		0.8	0.238 ± 0.077	0.044 ± 0.017
0.02	0.333	0.3	0.334 ± 0.077	0.213 ± 0.053
		0.5	0.336 ± 0.078	0.149 ± 0.039
		0.8	0.334 ± 0.077	0.062 ± 0.017
0.03	0.408	0.3	0.408 ± 0.077	0.259 ± 0.053
		0.5	0.411 ± 0.078	0.183 ± 0.039
		0.8	0.411 ± 0.076	0.076 ± 0.017
ERF				
0.01	0.236	0.3	0.236 ± 0.079	0.149 ± 0.054
		0.5	0.237 ± 0.081	0.105 ± 0.041
		0.8	0.234 ± 0.078	0.044 ± 0.018
0.02	0.333	0.3	0.333 ± 0.082	0.209 ± 0.056
		0.5	0.336 ± 0.079	0.150 ± 0.040
		0.8	0.344 ± 0.078	0.068 ± 0.018
IPP				
0.01	0.236	0.3	0.235 ± 0.089	0.125 ± 0.058
		0.5	0.237 ± 0.088	0.093 ± 0.043
		0.8	0.238 ± 0.082	0.044 ± 0.019
0.02	0.333	0.3	0.334 ± 0.089	0.178 ± 0.058
		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. $\hat{\lambda} = \frac{median(T_1^2, T_2^2, \ldots)}{0.455}$)
- Derive GRAMMAR-GC test statistics as $T_{i,GC}^2 = \frac{T_i^2}{\hat{\lambda}}$
- Solves the conservatively of the test

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. $\hat{\lambda} = \frac{median(T_1^2, T_2^2, \ldots)}{0.455}$)
- 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

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. $\hat{\lambda} = \frac{median(T_1^2, T_2^2, \ldots)}{0.455}$)
- Derive GRAMMAR-GC test statistics as $T_{i,GC}^2 = \frac{T_i^2}{\hat{\lambda}}$

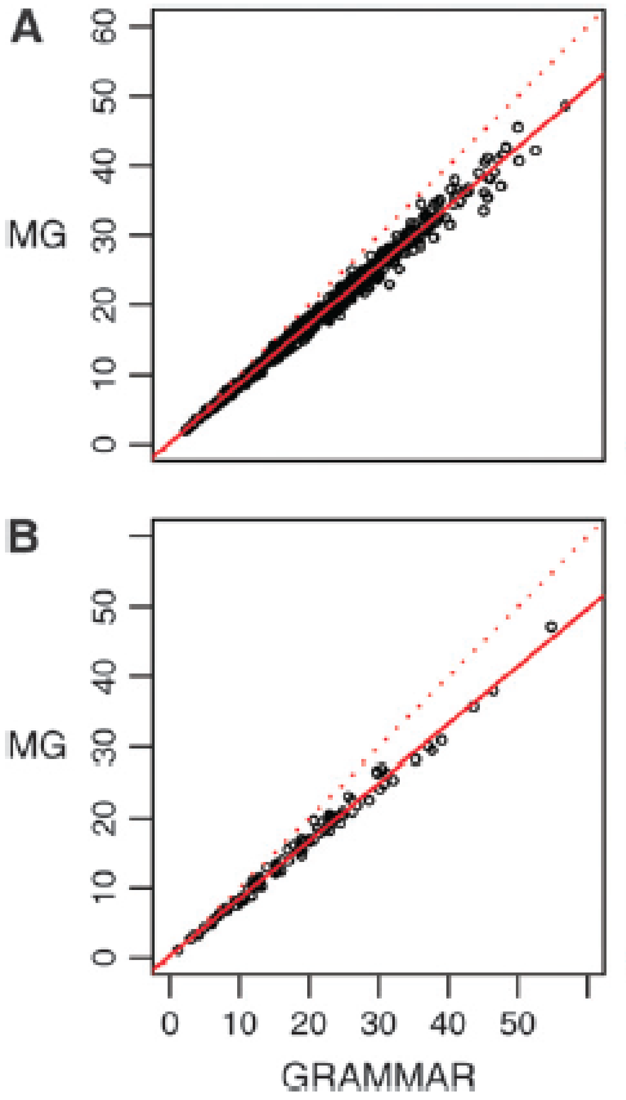
- Solves the conservatively of the test
- Does not solve the problem of the effect under-estimation
- Does not leave means to judge if MM-correction was adequate for the data (λ is 1 by definition of GRAMMAR-GC!)

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. $\hat{\lambda} = \frac{median(T_1^2, T_2^2, \ldots)}{0.455}$)
- 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)

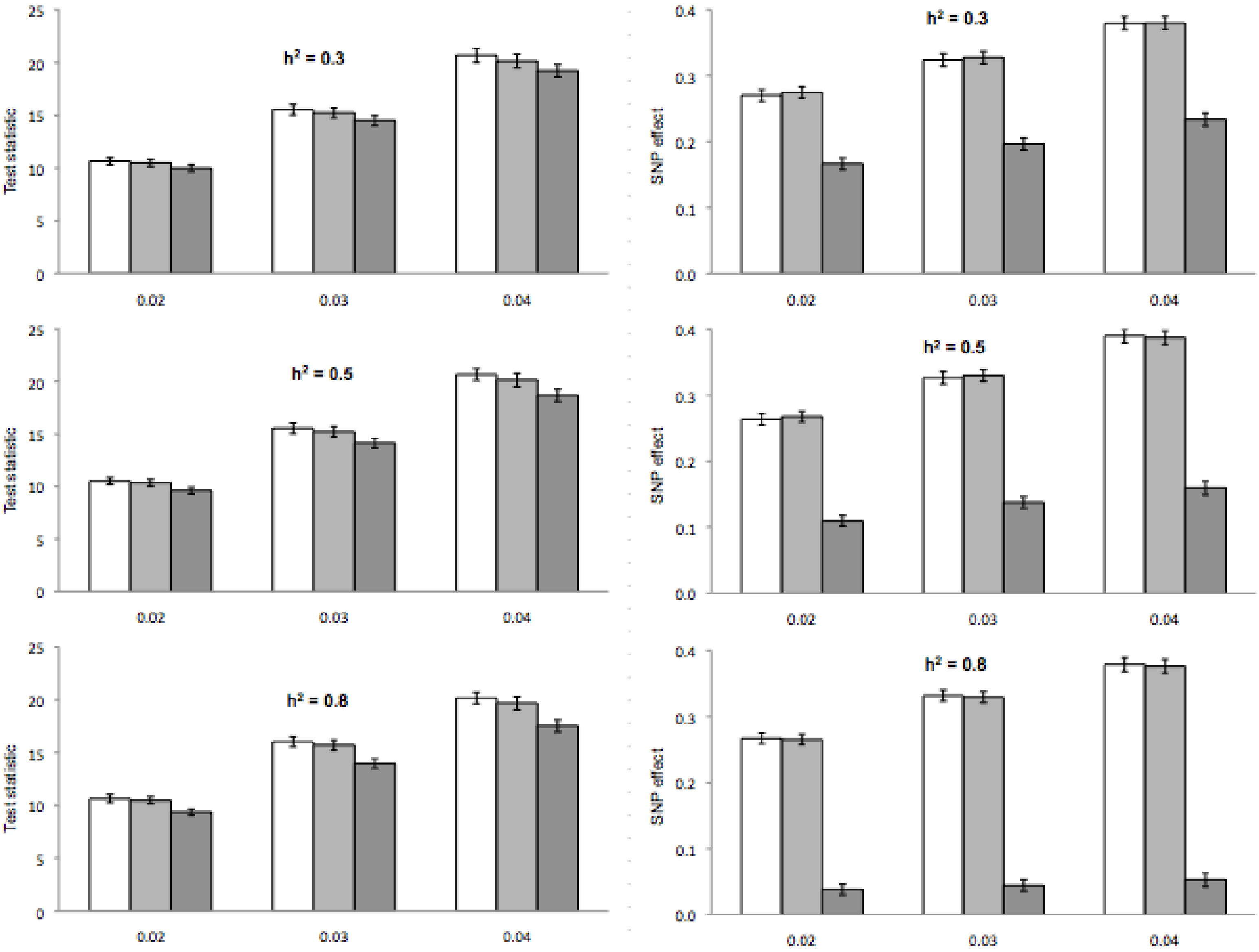


GRAMMAR-γ (work in progress)

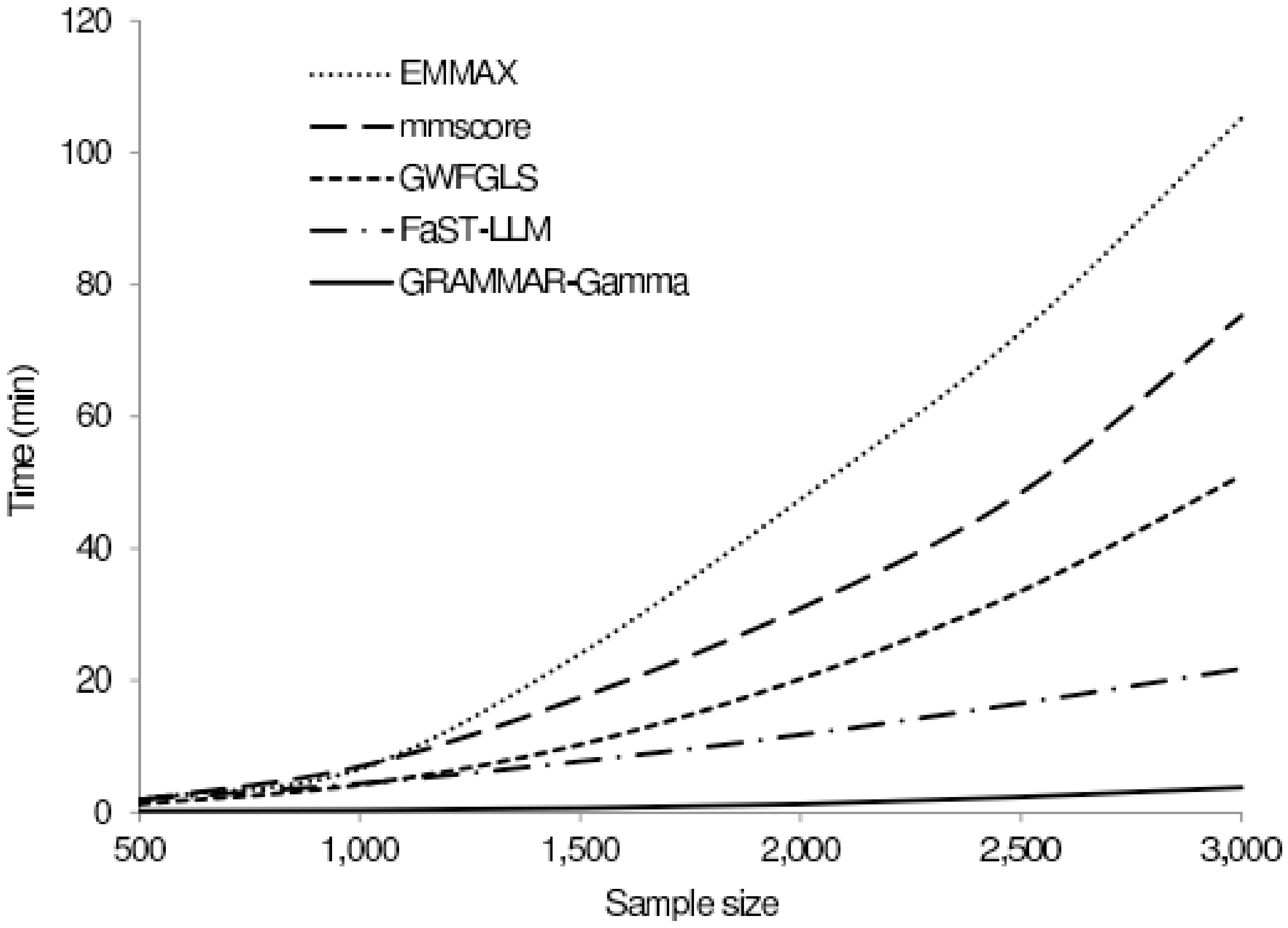
- The bias in test statistics and effect estimates is proportional to some constant, $\gamma = F(\hat{\Omega})$
- Instead of correction of test statistics only with GC, allows correction of both test statistics and effect estimates

- 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

NCP and effect estimates by FMM, Grammar-γ and GRAMMAR-GC



Speed comparison between methods (500k SNPs)



Contents

1. Mixed Models (MM)
   - Intro to MM for GWAS
   - Use of MM in population-based studies
2. Genome-wide feasible MM
   - FASTA-like tests
   - GRAMMAR-like tests
   - Comparison
3. **Optimal algebraic kernels and implementation**
4. Conclusions

Contents

1. Mixed Models (MM)
   - Intro to MM for GWAS
   - Use of MM in population-based studies
2. Genome-wide feasible MM
   - FASTA-like tests
   - GRAMMAR-like tests
   - Comparison
3. Optimal algebraic kernels and implementation
4. **Conclusions**

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 and have reasonable running time on samples $<$3k (? BUT no implementation for imputed data ?)
- Two-Step approximations are excellent unless SNPs have large effects. Speed-up achieved cf. LRT is 5-20 times. Implementations for imputed data available (ProbABEL, MixABEL::GWFGLS).
- Grammar-γ (work in progress) has 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

Acknowledgements

- The GenABEL project (www.genabel.org) contributors (special thanks: mr. Struchalin, dr. Karssen)
- Grammar: prof. de Koning, prof. Haley
- Grammar-GC: dr. Amin, prof. van Duijn
- Grammar-γ: dr. Svisheva, prof. Axenovich, dr. Belonogova
- Optimal kernels and implementation: mr. Fabregat, prof. Bientinesi