# Incorporating relatedness in the study of molecular phenotypes for genomic epidemiology

**Chris Holmes, Ingileif B. Hallgrímsdóttir, George Nicholson**

**Oxford Centre for Gene Function,**

**Department of Statistics,**

**University of Oxford**

## Overview

- MolPAGE study in genomic epidemiology
- Use of measures of relatedness on individuals for estimating genetic and technical variability in molecular phenotypes
- Bayesian Variance Components Models
  - illustrated for Spectral data: ClinProt MALDI-TOF data
- Co-variance components models

## MolPAGE

- MolPAGE stands for Molecular Phenotyping to Accelerate Genomic Epidemiology. Funded under EU FP6.

- Take a common biological sample (fat, urine and plasma) from a set of relateds (twins) and molecular phenotype them on a number of platforms

  (i) Epigenomics (genome-wide methylation profiles)

  (ii) Gene expression (Affy)

  (iii) Proteomics (ClinProt; peptidomics; antibody arrays)

  (iv) Metabon/lomics (NMR, LCMS)

- The first phase is on quantifying the genetic (heritable) components of variation of the molecular traits *and experimental variation* (robustness) inherent in the measurement of the molecular phenotypes.

○ Second stage is in integrative genomics

## Biomarkers

- The motivation for MolPAGE and a major research pursuit in genomic epidemiology is the search for molecular biomarkers of human disease
- Biomarker:

  “*A measurable biological trait (molecular or physiological) which associates with the onset or progression of disease*”
- Traditional biomarkers include,
  - Cholesterol, blood preasure, BMI
  - ER status (breast cancer)

## Uses of Biomarkers

There are three major uses for biomarkers

○ Profiling patients with increased disease risk

“*It is much more important to know the kind of patient that has a disease than to know the kind of disease a patient has*” - $\pi(x|y)$ vrs. $\pi(y|x)$

- Cholesterol (heart disease)

○ Prognosis – more accurate prediction of disease progression

- Number lymph nodes positive (breast cancer)

○ Subtyping – towards “personalised medicine”

- Oestrogen receptor (ER) status (breast cancer)

## Features of a Good Biomarker

A number of features affect the utility of a biomarker (over and above prediction accuracy)

- Stability
  - both in variation of the biomarker trait over time and, as important,
  - sample storage
- Generality - coverage
- Ease of measurement: stability and accuracy of the measurement platform
- Non-invasive
- Cheap (relatively)

## Genomics and Biomarkers

- ○ Genomic technologies have opened up the prospect for finding new molecular markers for familial and non-familial genetic disease

## MolPAGE Study design

○ At the first stage we are performing a twin study to analyse biological and technical variation

○ Twins provide a powerful design for inferring genetic effects

- blocked for in utero, dietary and socio-economic effects due to upbringing
- known amount of genetic sharing between identical (MZ) and non-identical (DZ) twins

## Twin Study

- Twins were contacted from St. Thomas' UK Adult Twin Registry of 10,000 twins
- The initial study has 77 twin pairs
  - 56 MZ (identical) twin pairs (31 twin pairs gave samples twice to capture longitudinal effects)
  - 21 DZ (fraternal) twin pairs
  - Fat, Urine and Plasma samples are taken
  - In total 215 samples from the 154 individuals (split into two aliquots, 430 aliquots)

- The same biological samples are shipped to each technological partner for molecular phenotyping (to allow for direct comparison and integrative genomics), at least 3 technical replicates per aliquot.

- We will denote a generic molecular phenotype measurement as

$$Y_{ijkl}$$

for twin pair $i \in \{1, ..., 77\}$, twin $j \in \{1, 2\}$, visit $k \in \{1, 2\}$, aliquot $l \in \{1, 2\}$.

## Statistical Model

- We analyse many different molecular phenotypes
- Useful to have a common statistical structure for the model

## Twin Model

$$Y_{ijkl} = \mu + a_{ij} + d_{ij} + c_i + e_{ij} + v_{ijk} + l_{ijkl} + b_{B(i,j,k,l)} + \epsilon_{ijkl}$$

| | | |
|---|---|---|
| $\mu$ | : | overall mean |
| $a_{ij}$ | : | additive genetic effect |
| $d_{ij}$ | : | dominant genetic effect |
| $c_i$ | : | common environmental effect |
| $e_{ij}$ | : | individual environmental effect |
| $v_{ijk}$ | : | individual visit effect |
| $l_{ijkl}$ | : | aliquot effect |
| $b_{B(i,j,k,l)}$ | : | batch effect |
| $\epsilon_{ijkl}$ | : | residual error |

## Covariance of genetic components

- Measures of expected relatedness allow us to estimate the genetic (heritable) components of variation
- Since MZ twins are genetically identical $a_{i1} = a_{i2}$ and $d_{i1} = d_{i2}$ if twin pair $i$ is MZ.
- DZ twins share on average half of their genetic material and

$$\begin{aligned}\text{Corr}(a_{i1}, a_{i2}) &= \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \\ \text{Corr}(d_{i1}, d_{i2}) &= \begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}\end{aligned}$$

## Twin Model

- The goal of our analysis is to partition the variability in the phenotype value into that attributable to different sources,
  - genetic ($a_{ij}$ and $d_{ij}$)
  - environmental ($c_i$, $e_{ij}$, $v_{ijk}$)
  - technical/experimental ($l_{ijkl}$, $b_{B(i,j,l,k)}$)

- The genetic components, $a_{ij}$ and $d_{ij}$ and the common environment $c_i$ are not identifiable in the likelihood and so we typically are interested in the proportion of variance attributable to

$$\text{familiality} = [a_{ij} + d_{ij} + c_i]$$

## Bayesian Model

- The effects are unique to an individual therefore we seek to model them using hierarchical structure, for example,

$$\{a_{i1}, a_{i2}\} \sim MVN(0, \sigma_a^2 \Sigma)$$

- where $\Sigma$ is the correlation structure and we adopt a prior

$$\sigma_a^2 \sim \pi(\cdot)$$

- and interest is on the posterior distribution $\pi(\sigma_a^2|Y)$ which can be obtained using MCMC (with analytic integration of the actual effects, $a_{ij}$ etc)

## Variance Decomposition

The total phenotypic variance is

$$\sigma_Y^2 \quad = \quad \sigma_a^2 + \sigma_d^2 + \sigma_c^2 + \sigma_e^2 + \sigma_v^2 + \sigma_l^2 + \sigma_b^2 + \sigma_\epsilon^2$$

and the familiality is

$$f^2 = \frac{\sigma_a^2 + \sigma_d^2 + \sigma_c^2}{\sigma_Y^2}$$

## Gibbs Sampling

- The joint distribution of the variance components is not known explicitly and we cannot sample from it directly.
- However, we can sample from the conditional distributions of each component (conditioned on all the others).
- The joint distribution is stationary w.r.t. the transition rule determined by the conditional distributions. Sequential draws from the conditional distributions are thus a sample from a Markov chain whose stationary distribution is the joint.

## Choice of prior distributions

- We consider the following priors:
  - Gamma distribution on the precision, $1/\sigma_\star^2 \sim$ Gamma$(\epsilon, \epsilon)$.
  - Uniform distribution on the standard deviation, $\sigma_\star \sim$ U$(0, C)$.
  - Half-Cauchy distribution on the standard dev., $\sigma_\star \sim$ hC$(s)$.
- When the number of random effects that share a variance component is large (e.g. there are 154 $e_{ij} \sim N(0, \sigma_e^2)$ the choice of prior does not affect much the posterior distribution.
- For all variance components except $\sigma_b^2$ we choose a uniform prior, but since there are only 5 batches more care needs to be taken in choosing the prior.

## Identifiability

- The parameters $a, d$ and $c$ are not all identifiable in the likelihood.
- One of the benefits of working in a Bayesian framework is that we define the model to match the underlying structure, regardless of identifiability.
- The joint posterior distributions provide us with important insight into how variance can be “transferred” between the variance components $a, d$ and $c$, giving us information about equally valid parameter values

## Identifiability

- From a Bayesian perspective:

  What you do or do not observe should not influence the model you adopt for the underlying process

- That is, you write down the model you believe underlies the data generating process and then condition on the data to hand

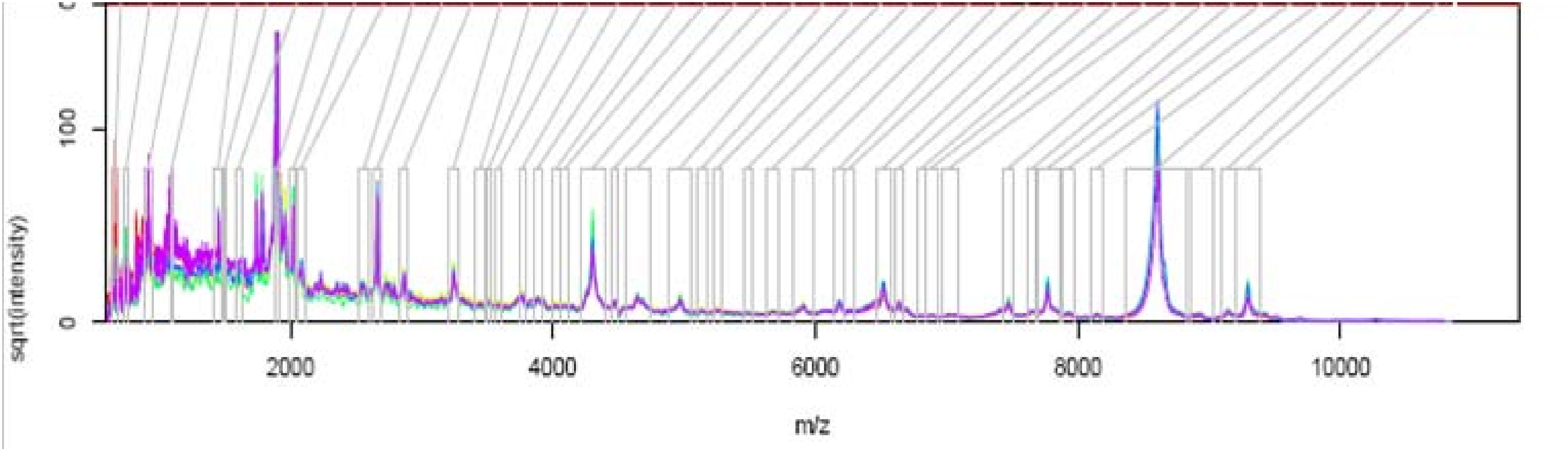
# Case study using ClinProt proteomics

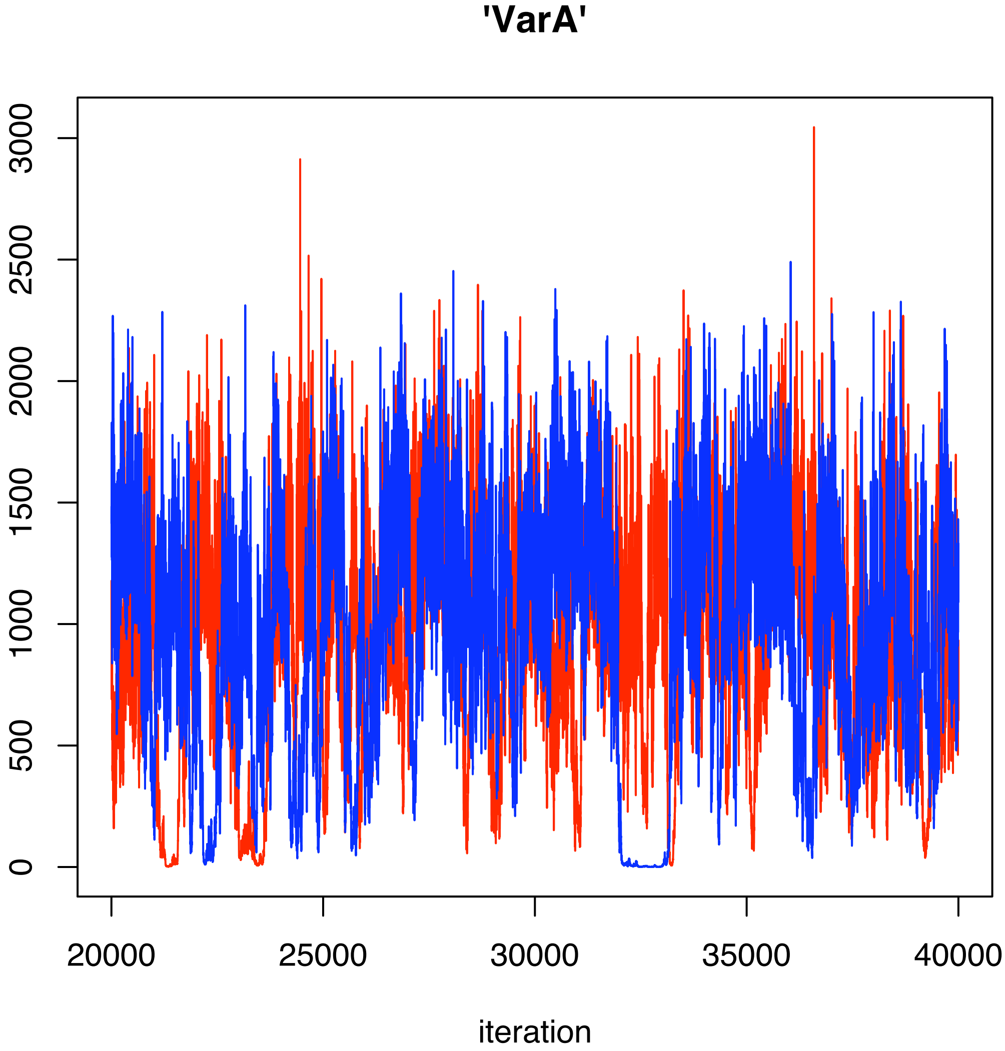
## ClinProt MALDI-TOF Data

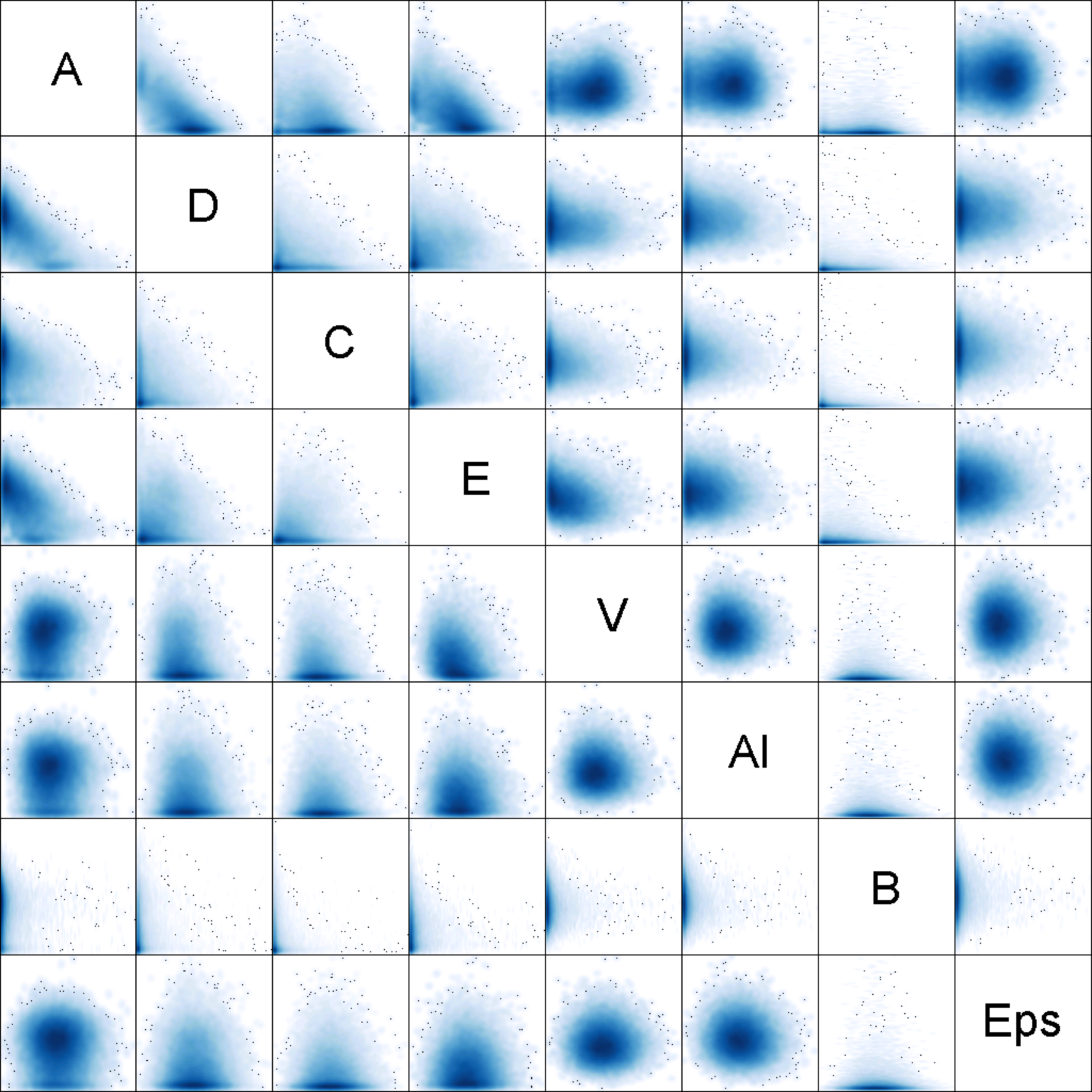
- Magnetic beads with functional surfaces are used to bind proteins and peptides from a sample (plasma, serum or urine).
- After elution the captured proteins and peptides are analysed in a MALDI-TOF mass spectrometer.
- We perform pre-processing of the spectra, including denoising, baseline subtraction, normalisation, alignment and peak extraction.

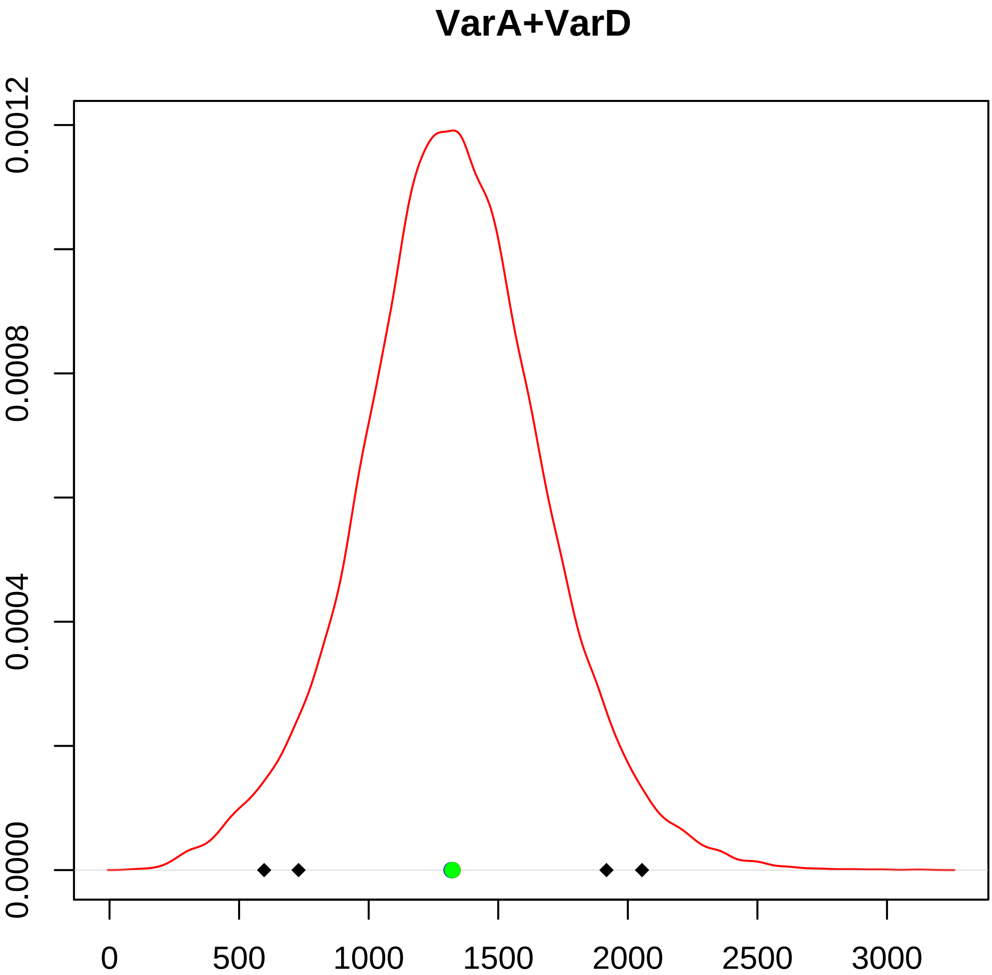
## Analysis of Peaks

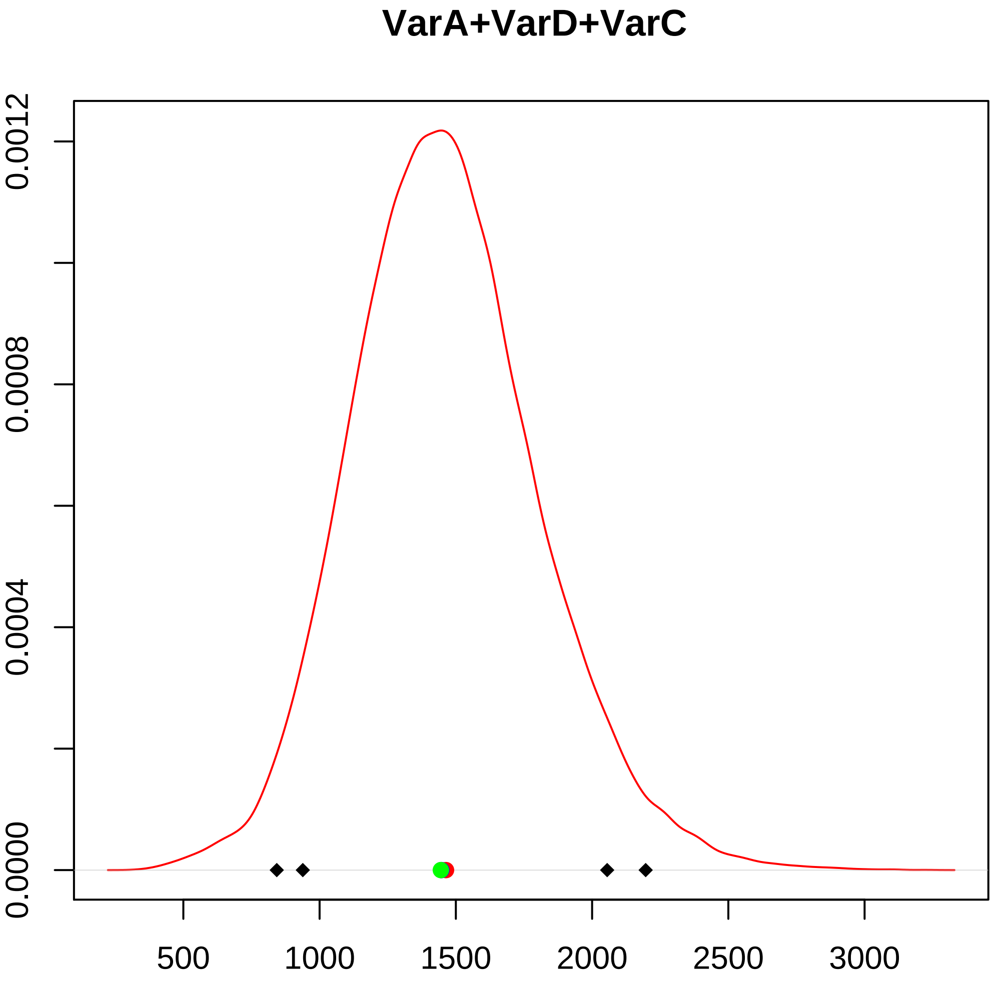
- Peak areas are extracted and area under the peak is treated as phenotype $Y_{ijkl}$
- Initially we treat each peak independently
- This is equivalent to stating that (initially) we allow for interactions between the peptides and the genetic effects
- We run our MCMC simulations and report posterior distributions on the variance components
- For example, the output for peptide abundance under one peak would look like

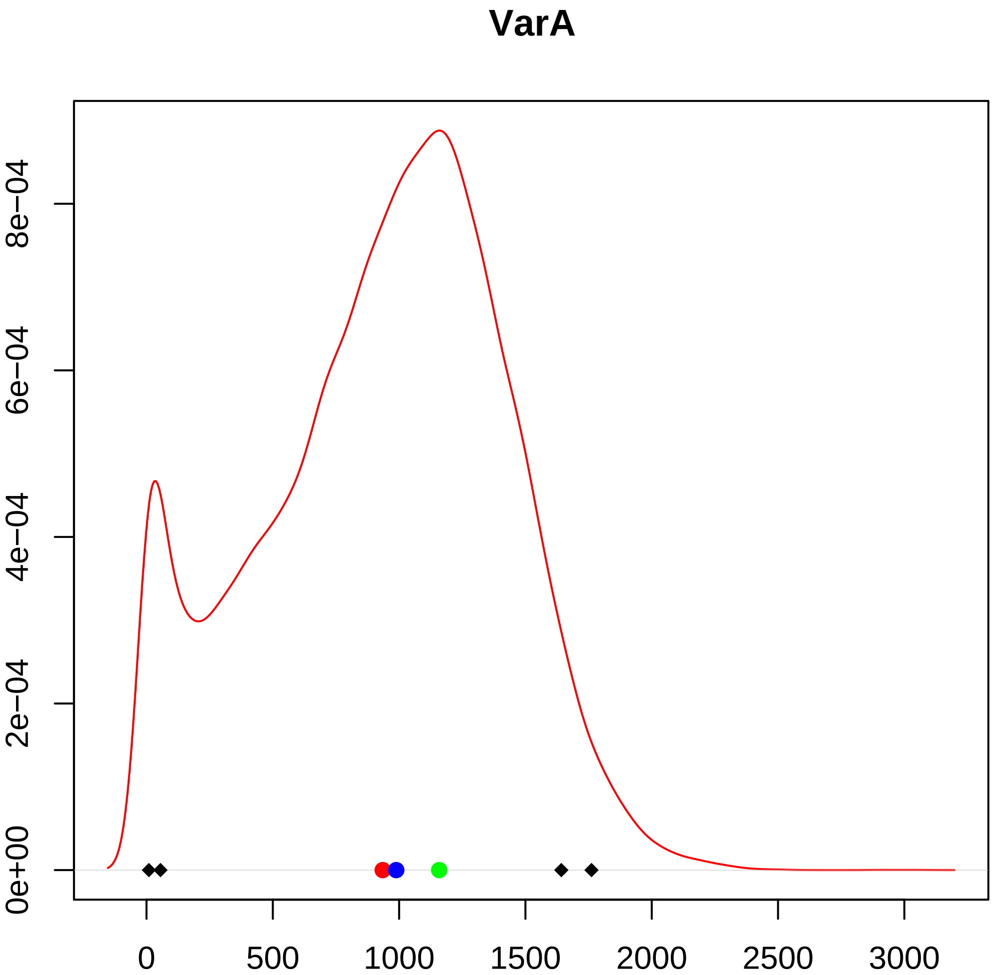

sqrt(intensity)
0
100
2000
4000
6000
8000
10000
m/z


'VarA'
0
500
1000
1500
2000
2500
3000
20000
25000
30000
35000
40000
iteration

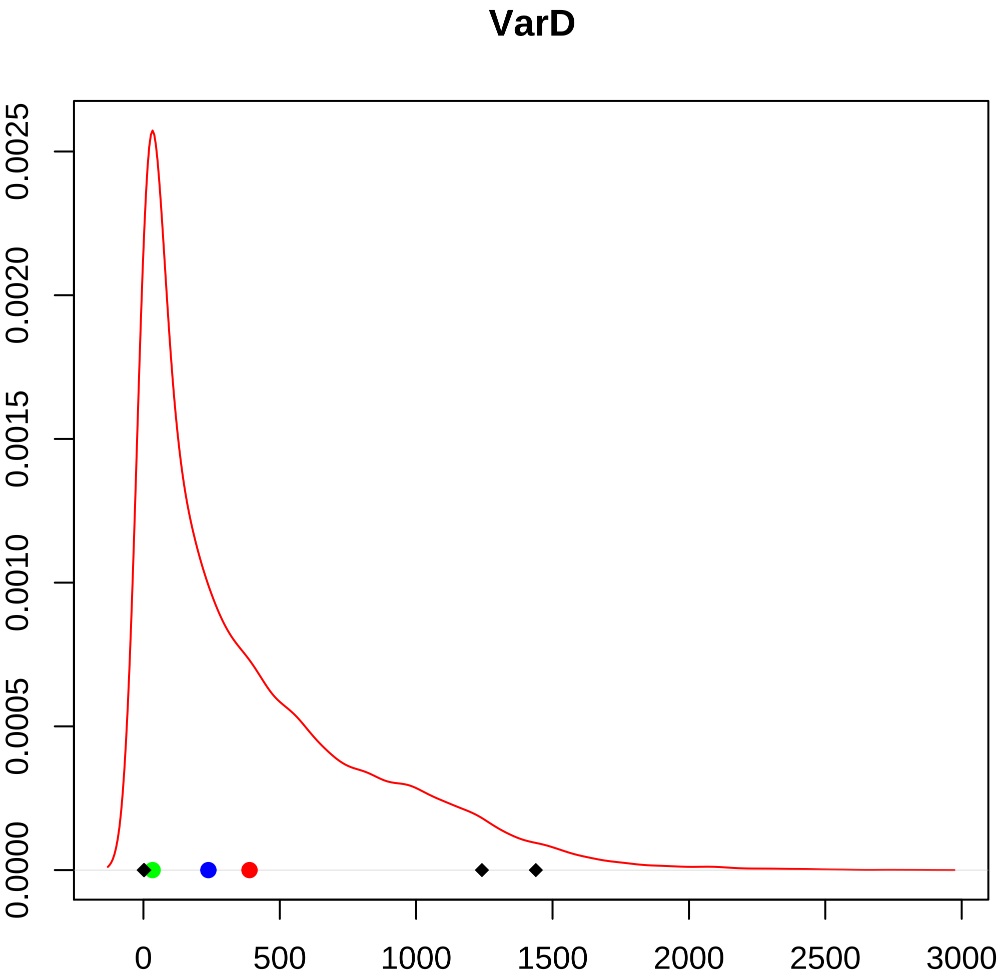

A
D
C
E
V
Al
B
Eps

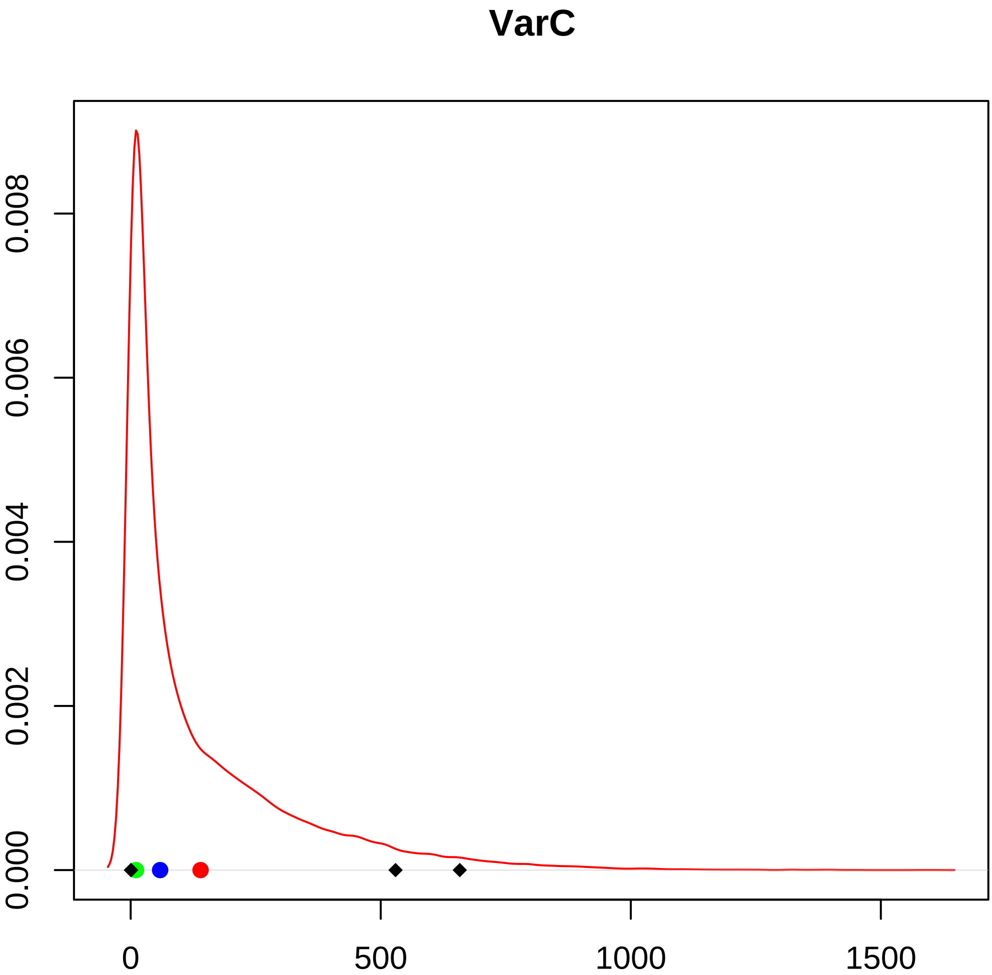

VarA+VarD
0.0000
0.0004
0.0008
0.0012
0
500
1000
1500
2000
2500
3000

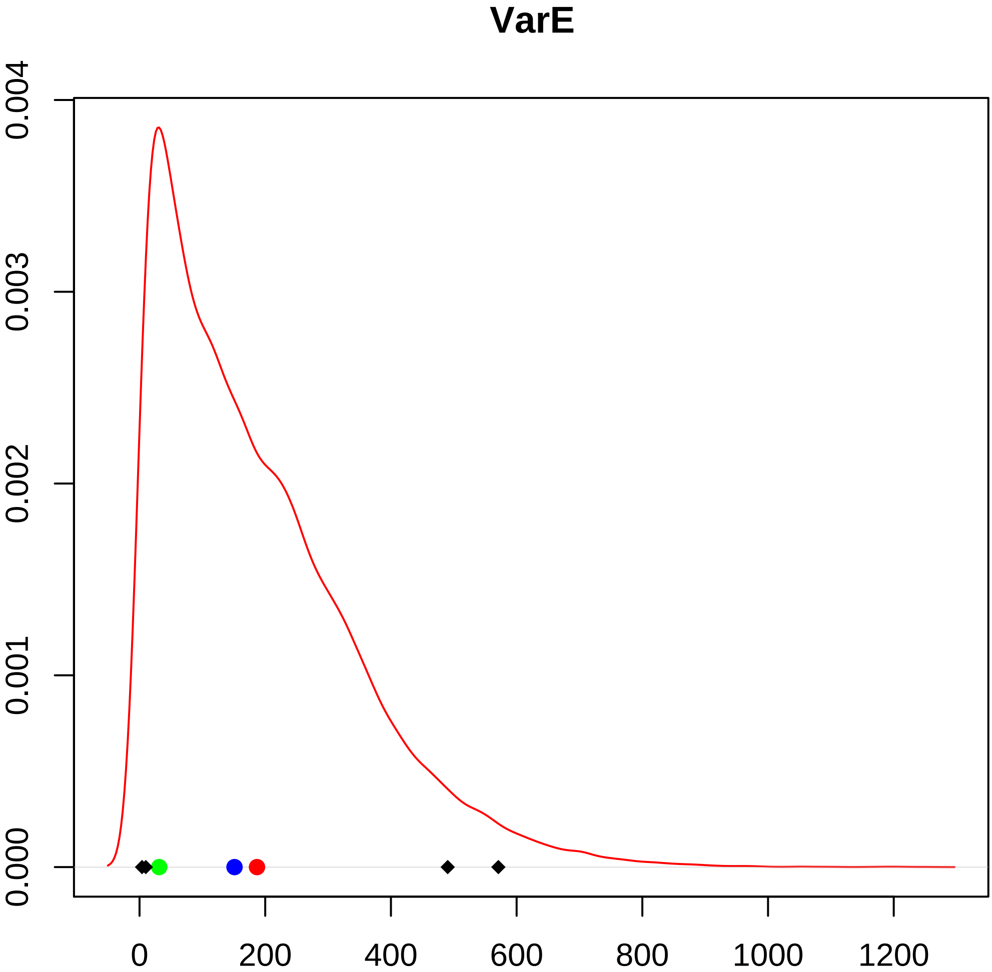

VarA+VarD+VarC
0.0000
0.0004
0.0008
0.0012
500
1000
1500
2000
2500
3000

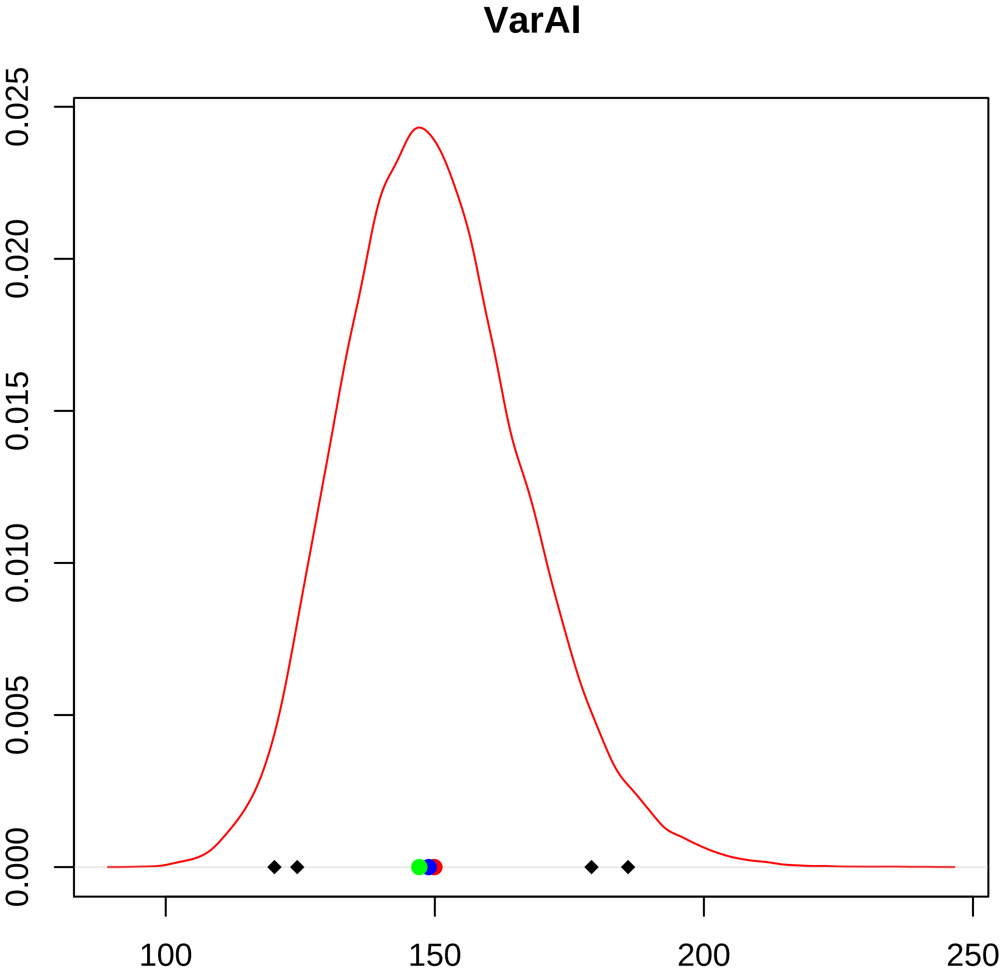


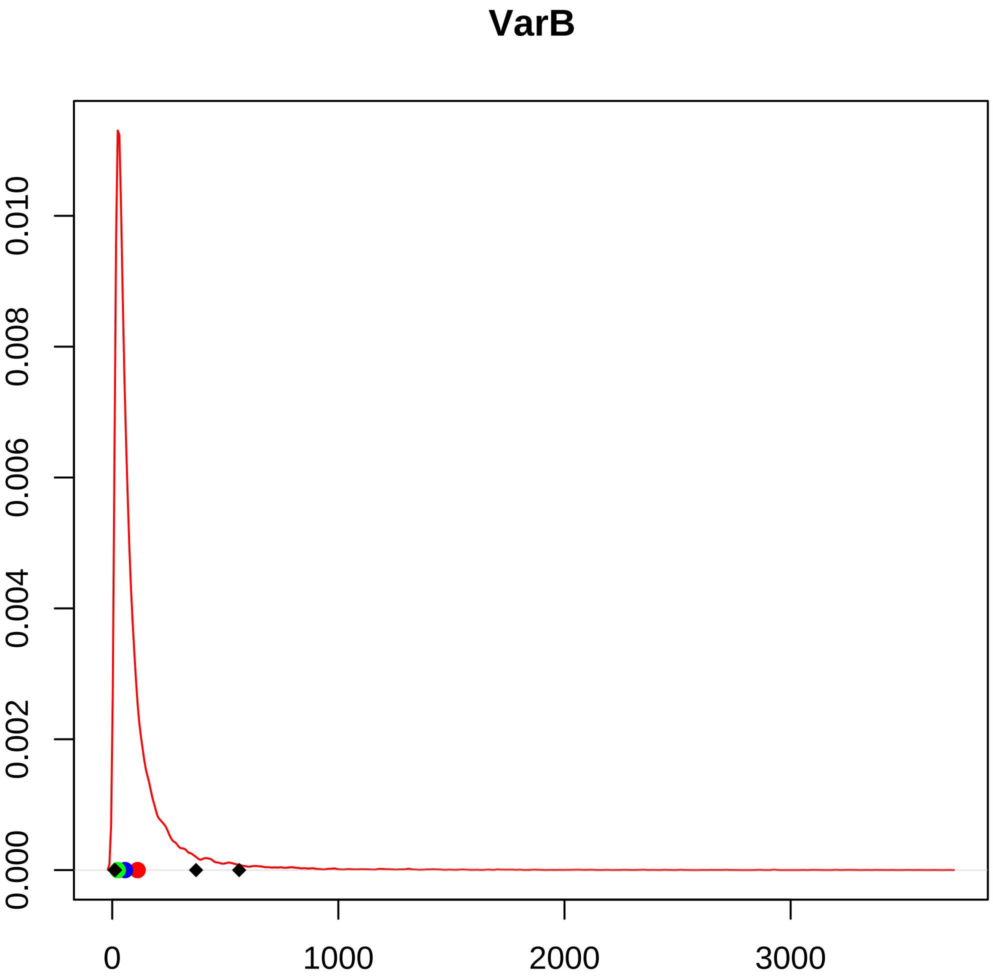
VarA
0e+00
2e−04
4e−04
6e−04
8e−04
0
500
1000
1500
2000
2500
3000

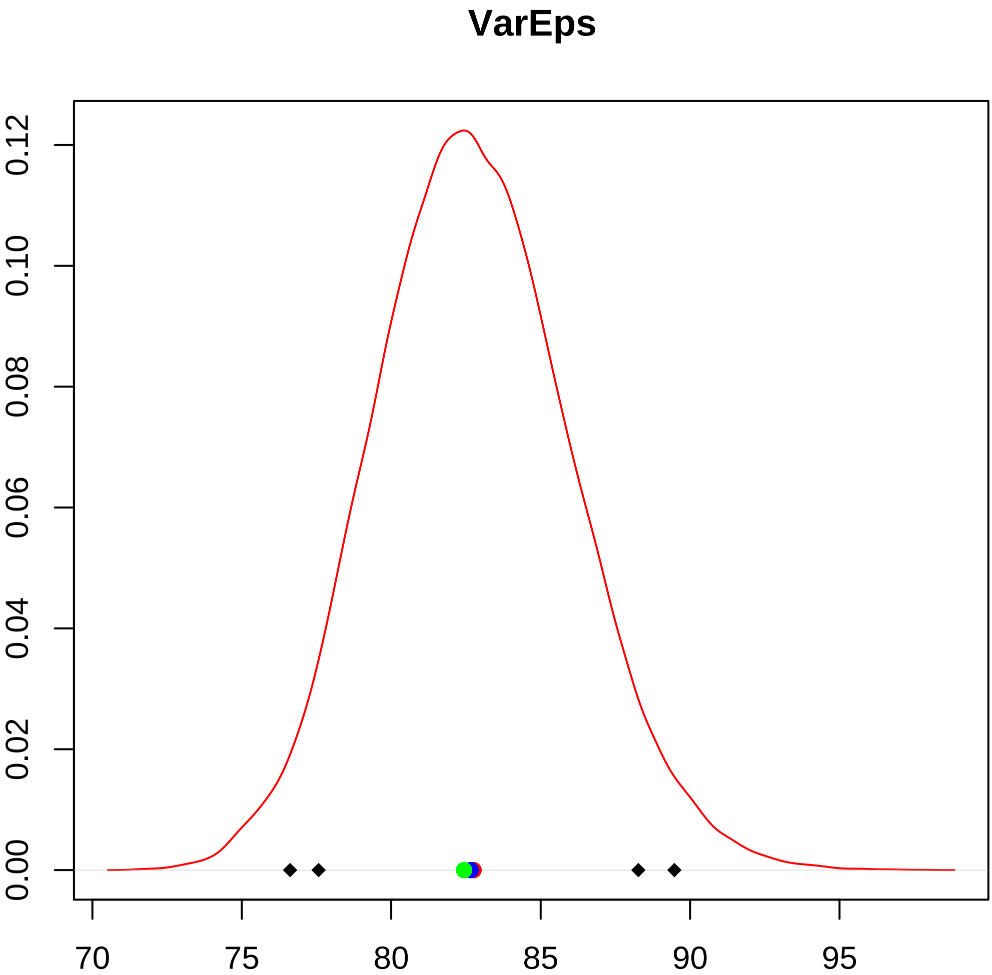

VarD
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0
500
1000
1500
2000
2500
3000

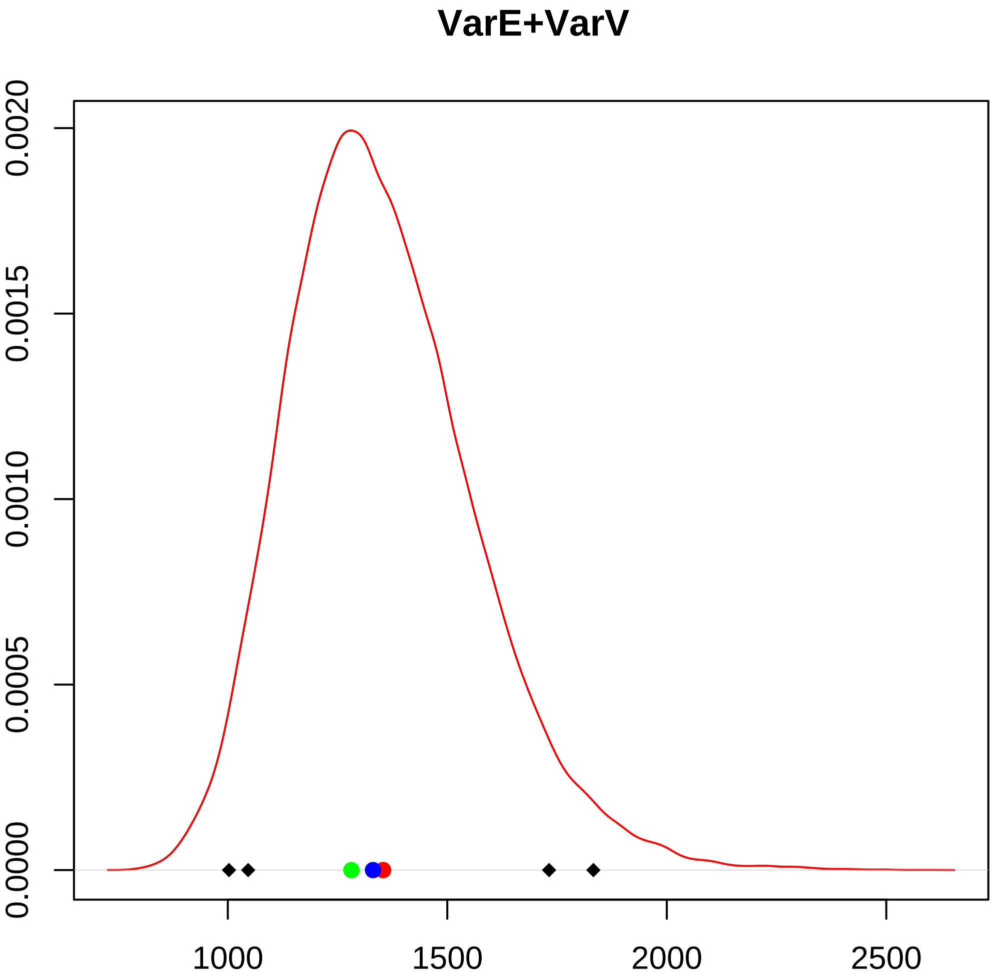

VarC
0.000
0.002
0.004
0.006
0.008
0
500
1000
1500

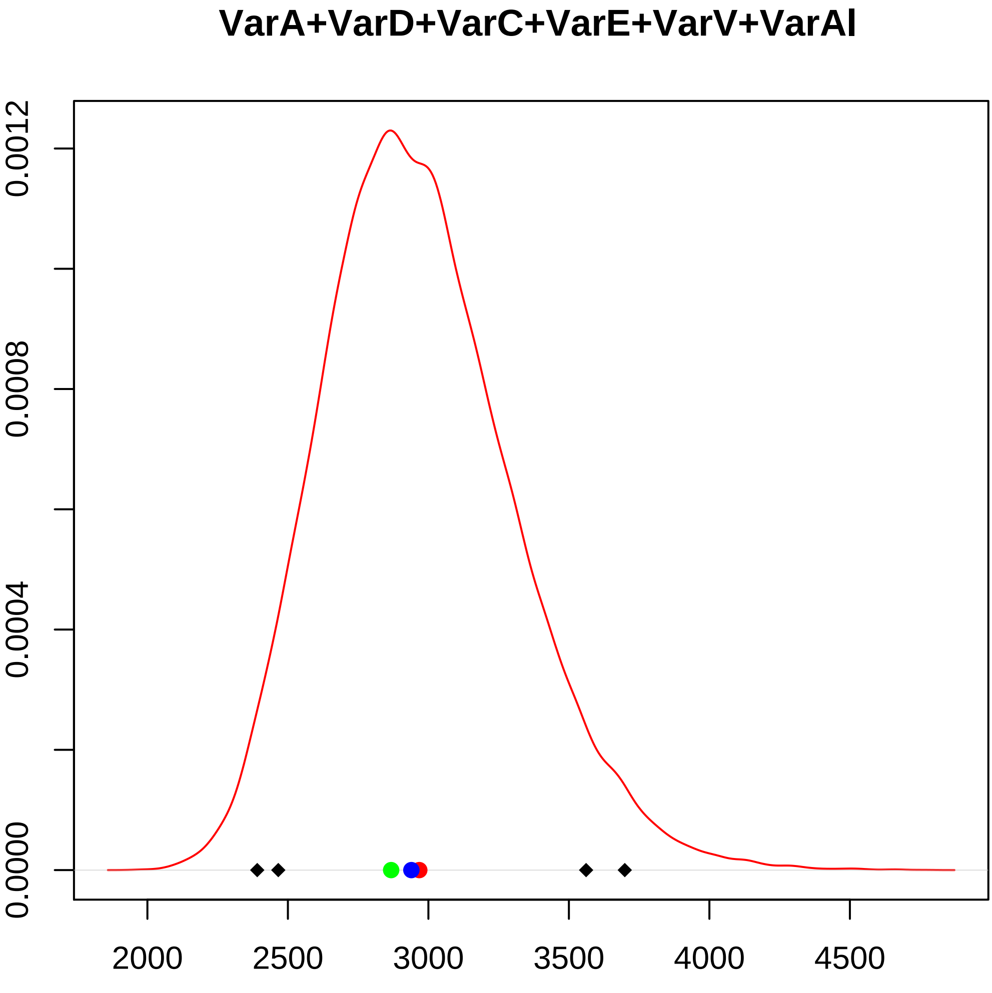

VarE
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
1000
1200

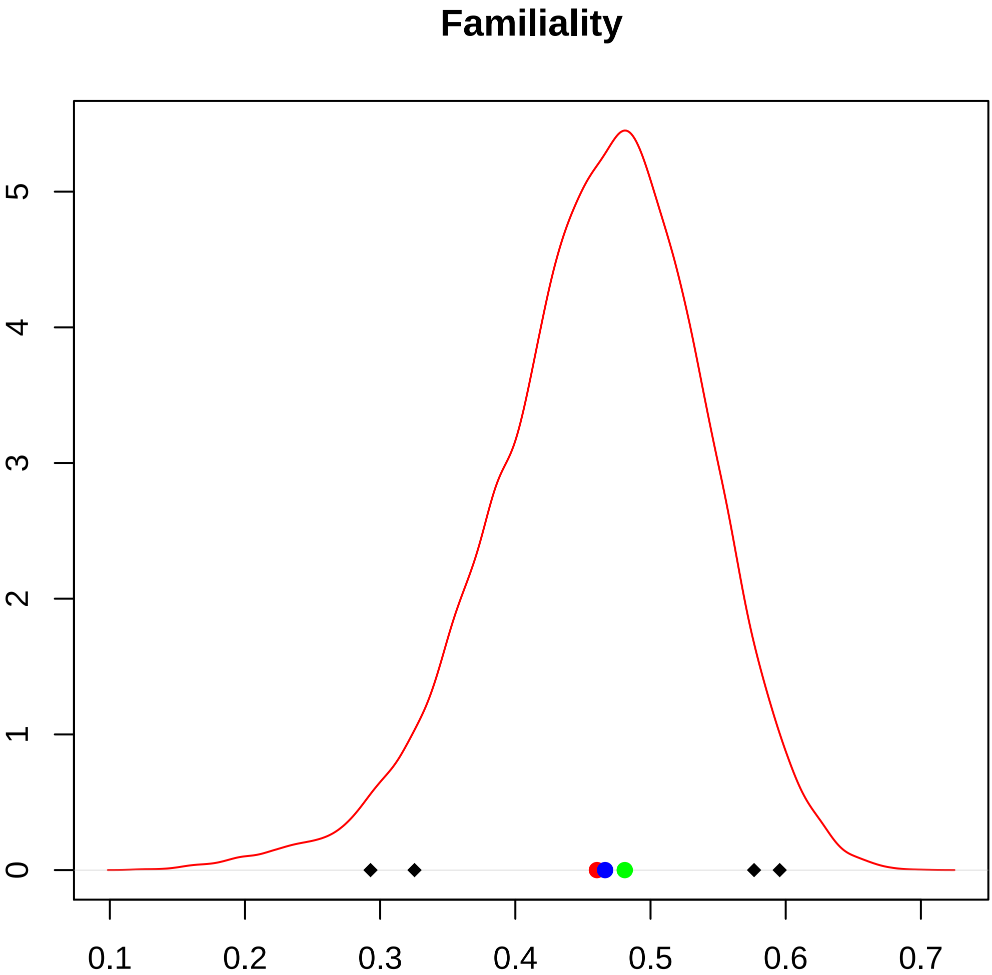

VarAI
0.000
0.005
0.010
0.015
0.020
0.025
100
150
200
250

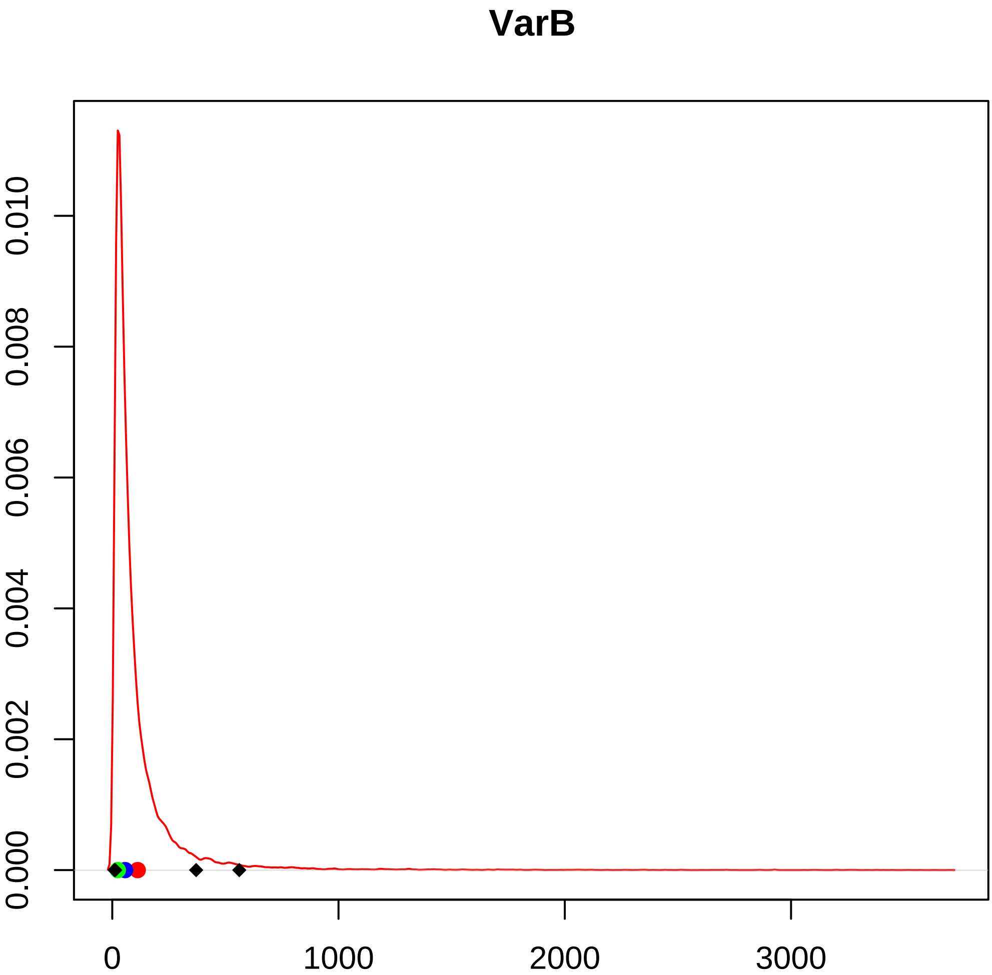

VarB
0.000
0.002
0.004
0.006
0.008
0.010
0
1000
2000
3000

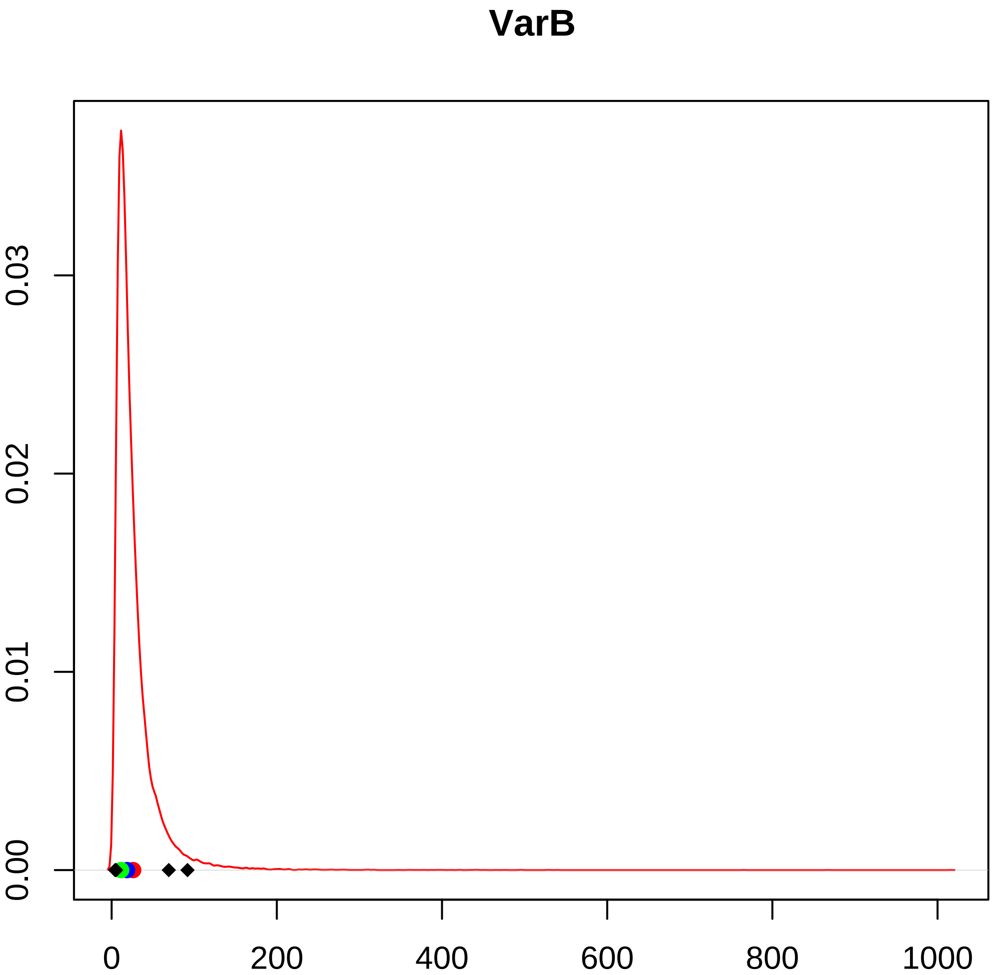

VarEps
0.00
0.02
0.04
0.06
0.08
0.10
0.12
70
75
80
85
90
95

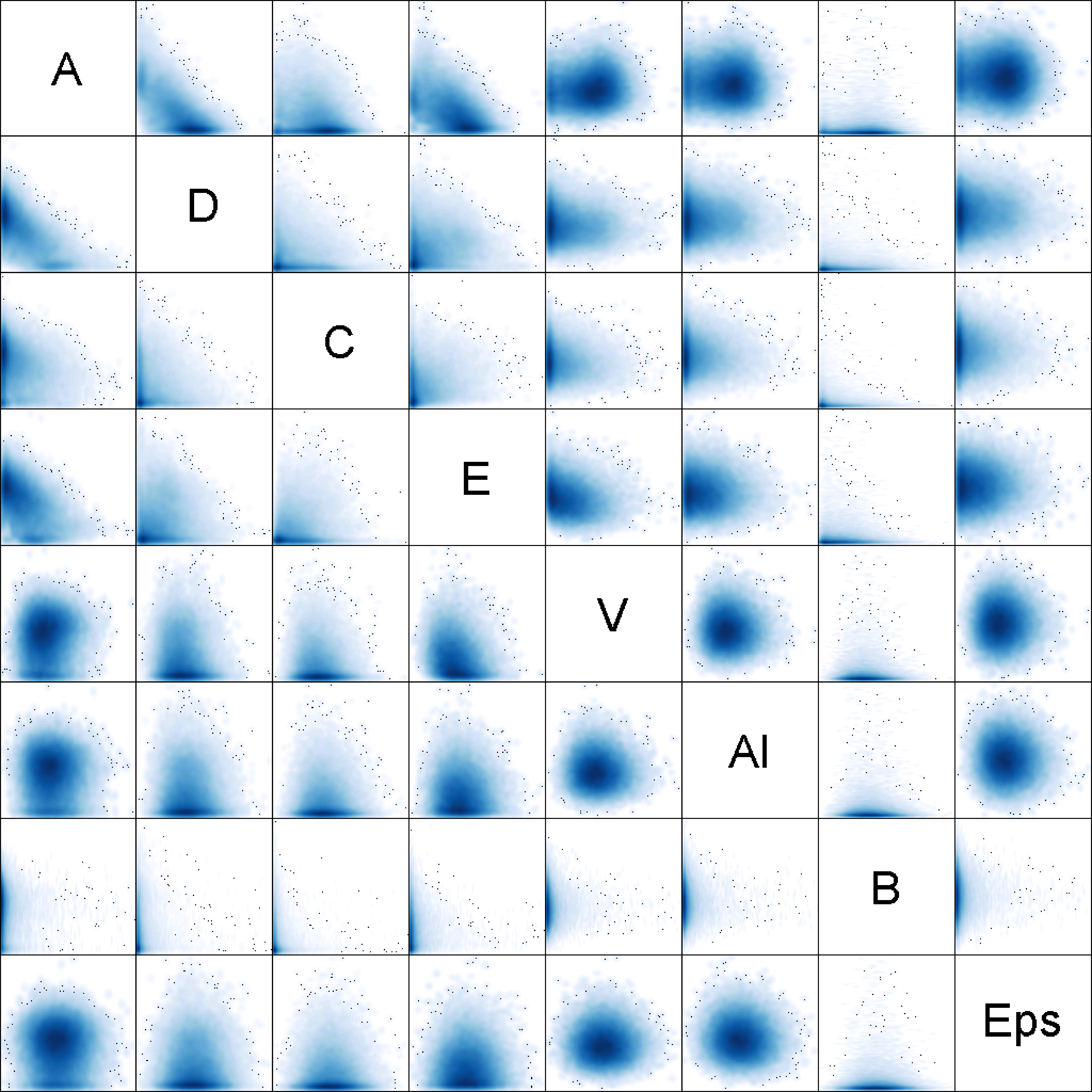

VarE+VarV
0.0000
0.0005
0.0010
0.0015
0.0020
1000
1500
2000
2500

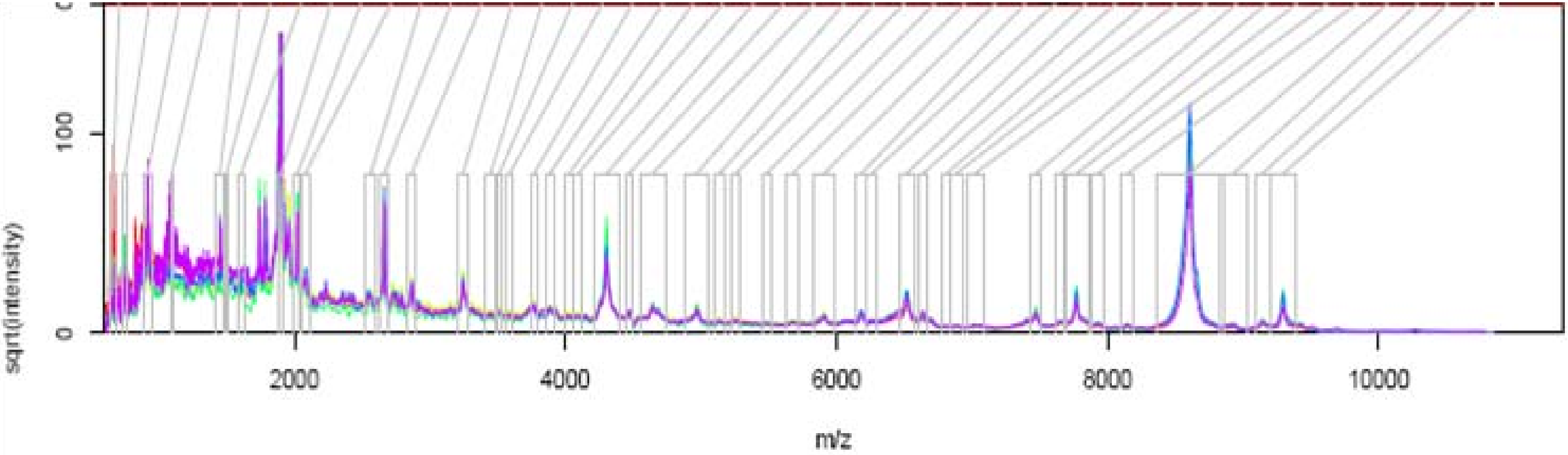

VarA+VarD+VarC+VarE+VarV+VarAl
0.0000
0.0004
0.0008
0.0012
2000
2500
3000
3500
4000
4500


Familiality
0
1
2
3
4
5
0.1
0.2
0.3
0.4
0.5
0.6
0.7


VarB
0.000
0.002
0.004
0.006
0.008
0.010
0
1000
2000
3000


VarB
0.00
0.01
0.02
0.03
0
200
400
600
800
1000

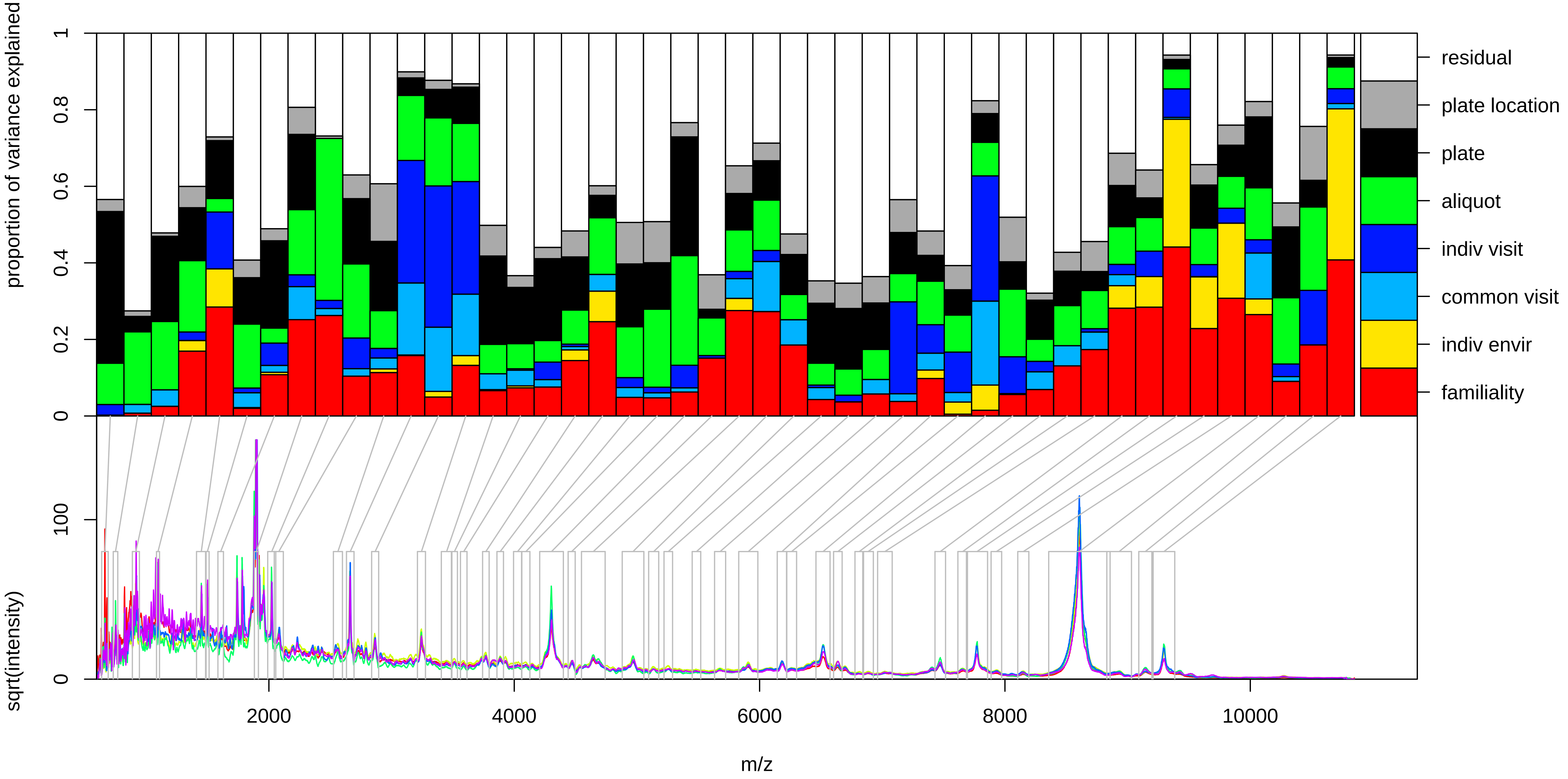

A
D
C
E
V
Al
B
Eps


sqrt(intensity)
0
100
2000
4000
6000
8000
10000
m/z

## Output

- There are typically many peaks per spectra
- Our code does the spectral preprocessing, extracts peaks, runs the mcmc, and then reports posterior summary statistics

**imac bead type**
**top: estimated variance components for each of 46 peaks summarised by sqrt(total)**
**bottom: median spectrum (post−processing) from each plate, with peak summary intervals**



## Covariance Components models of association

- We are interested in associating changes in molecular phenotype levels with changes in a clinical phenotype
- We have developed a new approach for this when we have data on relateds
- Consider a clinical phenotype, $Z$ and molecular phenotype $Y$
- A typical model would consider testing

$$\pi(Z|Y) \neq \pi(Z)$$

- However, it is interesting (we believe) to look for genetic components of association
- That is,

$$\pi(Z_{genetic}|Y_{genetic}) = \pi(Z_{genetic})$$

- We do this by investigating association between the genetical components of variation

- Consider the two phenotypes, one clinical and one molecular

$$\begin{aligned} Y_{ijkl} &= \mu + a_{ij}^{(Y)} + d_{ij}^{(Y)} + c_i^{(Y)} + e_{ij}^{(Y)} + v_{ijk}^{(Y)} + l_{ijkl}^{(Y)} + b_{B(i,j,k,l)}^{(Y)} + \epsilon_{ijkl}^{(Y)} \\ Z_{ijkl} &= \mu + a_{ij}^{(Z)} + d_{ij}^{(Z)} + c_i^{(Z)} + e_{ij}^{(Z)} + v_{ijk}^{(Z)} + l_{ijkl}^{(Z)} + b_{B(i,j,k,l)}^{(Z)} + \epsilon_{ijkl}^{(Z)} \end{aligned}$$

- We can put a joint dependence structure on “interesting” components

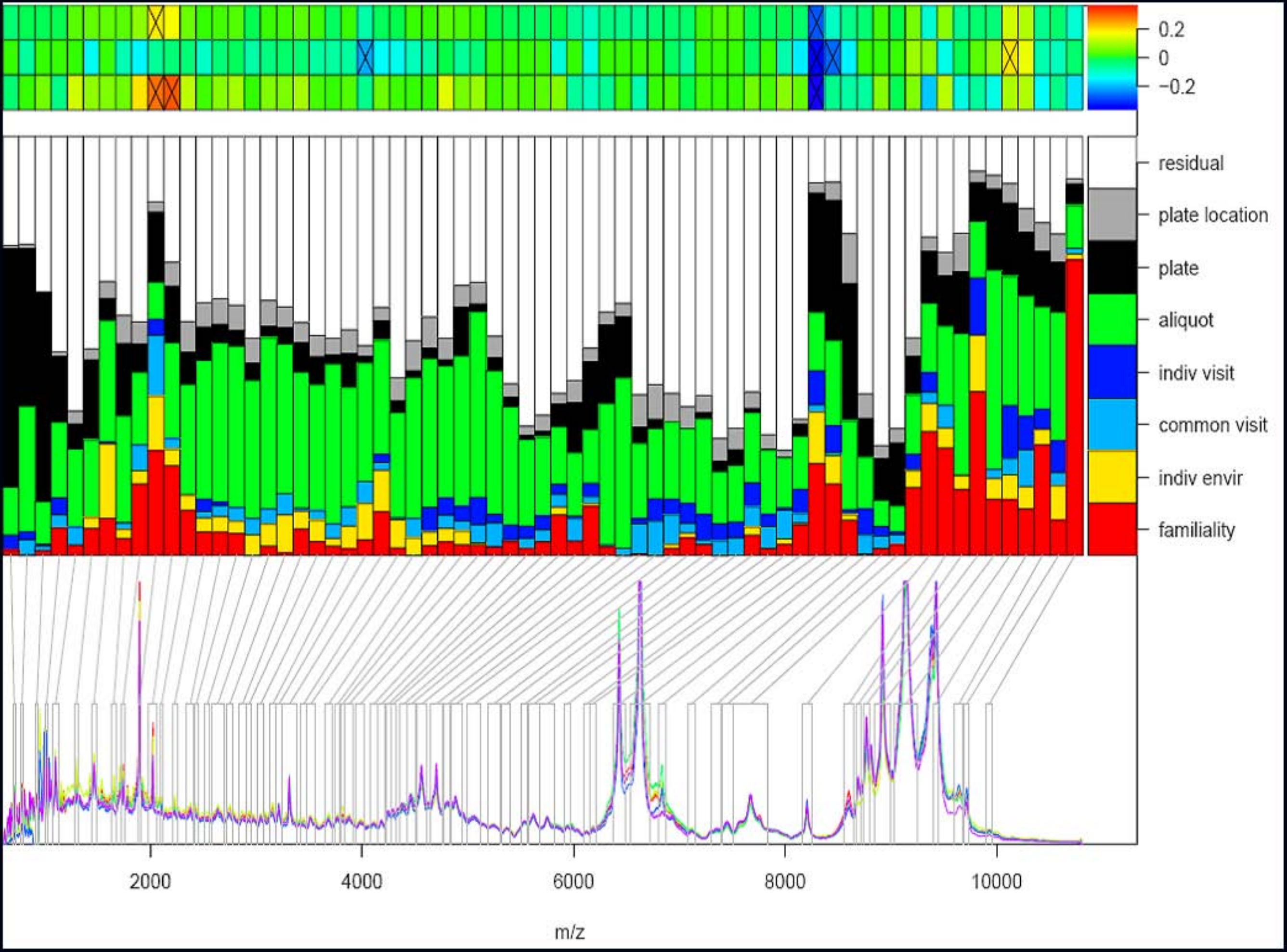
- For example,

$$\begin{aligned}\{a^{(Y)}, a^{(Z)}\} &\sim N(0, \rho_g \sigma_a^{(Y)} \sigma_a^{(Z)}) \\ \{d^{(Y)}, d^{(Z)}\} &\sim N(0, \rho_g \sigma_d^{(Y)} \sigma_d^{(Z)})\end{aligned}$$

- with prior say

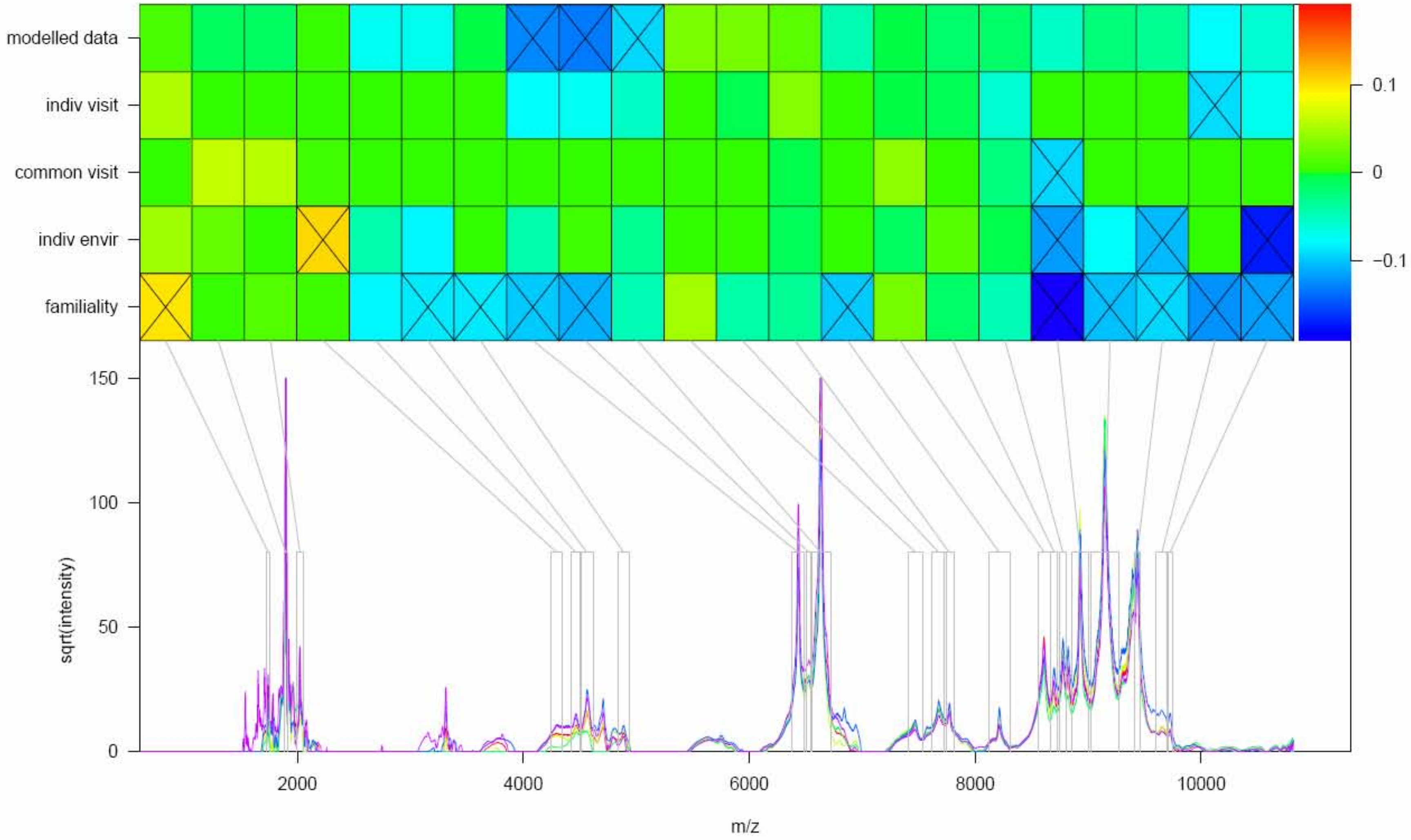
$$\pi(\rho_g) \sim U(-1, 1)$$

- and then investigate $\pi(\rho_g|Y)$
- This looks for association in the genetical axis of variation between $Z$ and $Y$
- That is, in genetical projections orthogonal to that variability spanned by environmental and technical effects
- Summarise posterior mean associations



0.2
0
−0.2
residual
plate location
plate
aliquot
indiv visit
common visit
indiv envir
familiality
2000
4000
6000
8000
10000
m/z

c8 bead type
top: correlation of log(bmi) with each fitted random effect for each of 22 peaks summarised by sqrt(total);
significantly non−zero correlation estimates marked with X; fdr = 0.05
bottom: median spectrum (post−processing) from each plate, with peak summary intervals



## Summary

- Information of relatedness allows us to separate out genetical from environmental factors in molecular phenotypes
- Bayesian framework very useful for what we do
- Covariance components models allow us to explore interesting axes of association
  - interested in extensions in graphical models/networks

Acknowledgements: Medical Research Council and EU.