H-likelihood opens a new way of estimating genetic values using genome-wide dense marker maps

Xia Shen*, Lars Rönnegård, Örjan Carlborg

*The Linnaeus Centre for Bioinformatics, Uppsala University, Sweden

QTLMAS 2010

Marker analysis v.s. Interval mapping

- Single marker analysis.
- Interval mapping.
- Multiple interval mapping.
- All-marker analysis.

Marker analysis v.s. Interval mapping

- Single marker analysis.
- Interval mapping.
- Multiple interval mapping.
- All-marker analysis.

All-marker analysis

- model-selection-*free*.
- *violates* the usual rule of model dimensionality.
- *shrinks* marker effects with zero values.
- Zeng (1993) *Proc. Natl. Acad. Sci. USA.*
- Meuwissen, et al. (2001) *Genetics.*
- Xu (2003) *Genetics.*
- Xu (2007) *Biometrics.*

All-marker analysis

- model-selection-*free*.
- *violates* the usual rule of model dimensionality.
- *shrinks* marker effects with zero values.
- Zeng (1993) *Proc. Natl. Acad. Sci. USA*.
- Meuwissen, et al. (2001) *Genetics*.
- Xu (2003) *Genetics*.
- Xu (2007) *Biometrics*.

Bayesian analysis v.s. Hierarchical likelihood

- Bayesian
  - flexible.
  - priors are required.
  - time-consuming.
- *h*-likelihood
  - a *unified* and *direct* method for random effect models.
  - hierarchical generalized linear models (HGLM, Lee & Nelder 1996)
  - double HGLM (DHGLM, Lee & Nelder 2006).
  - can be estimated by iterating GLMs.
  - *No prior* specification is required.
  - The computation is *fast*.

Bayesian analysis v.s. Hierarchical likelihood

- Bayesian
  - flexible.
  - priors are required.
  - time-consuming.
- *h*-likelihood
  - a *unified* and *direct* method for random effect models.
  - hierarchical generalized linear models (HGLM, Lee & Nelder 1996)
  - double HGLM (DHGLM, Lee & Nelder 2006).
  - can be estimated by iterating GLMs.
  - *No prior* specification is required.
  - The computation is *fast*.

Double HGLM

The phenotype of individual i is postulated as a random effect model

$$y_i = \sum_k x_{ik}\beta_k + \sum_j z_{ij}\, g_j + e_i$$

with $g_j \sim N(0, \lambda_j)$ for marker j and residual $e_i \sim N(0, \sigma^2)$. The variance of marker effect, λ_j, is modeled as

$$\lambda_j = a + b_j$$

with intercept a and normally distributed random effect b_j.

Double HGLM

The phenotype of individual i is postulated as a random effect model

$$y_i = \sum_k x_{ik}\beta_k + \sum_j z_{ij}\, g_j + e_i$$

with $g_j \sim N(0, \lambda_j)$ for marker j and residual $e_i \sim N(0, \sigma^2)$. The variance of marker effect, λ_j, is modeled as

$$\lambda_j = a + b_j$$

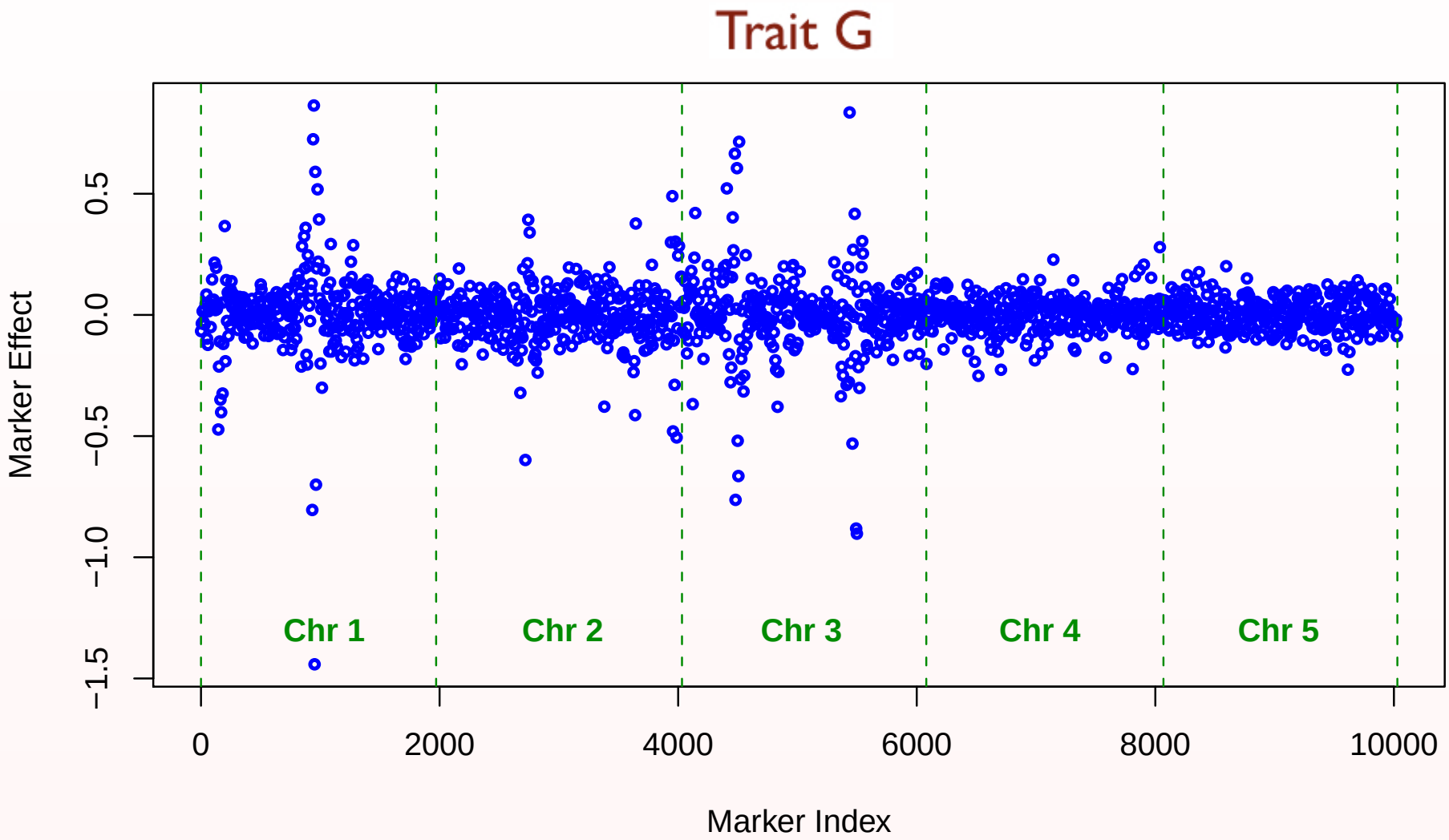
with intercept a and normally distributed random effect b_j.

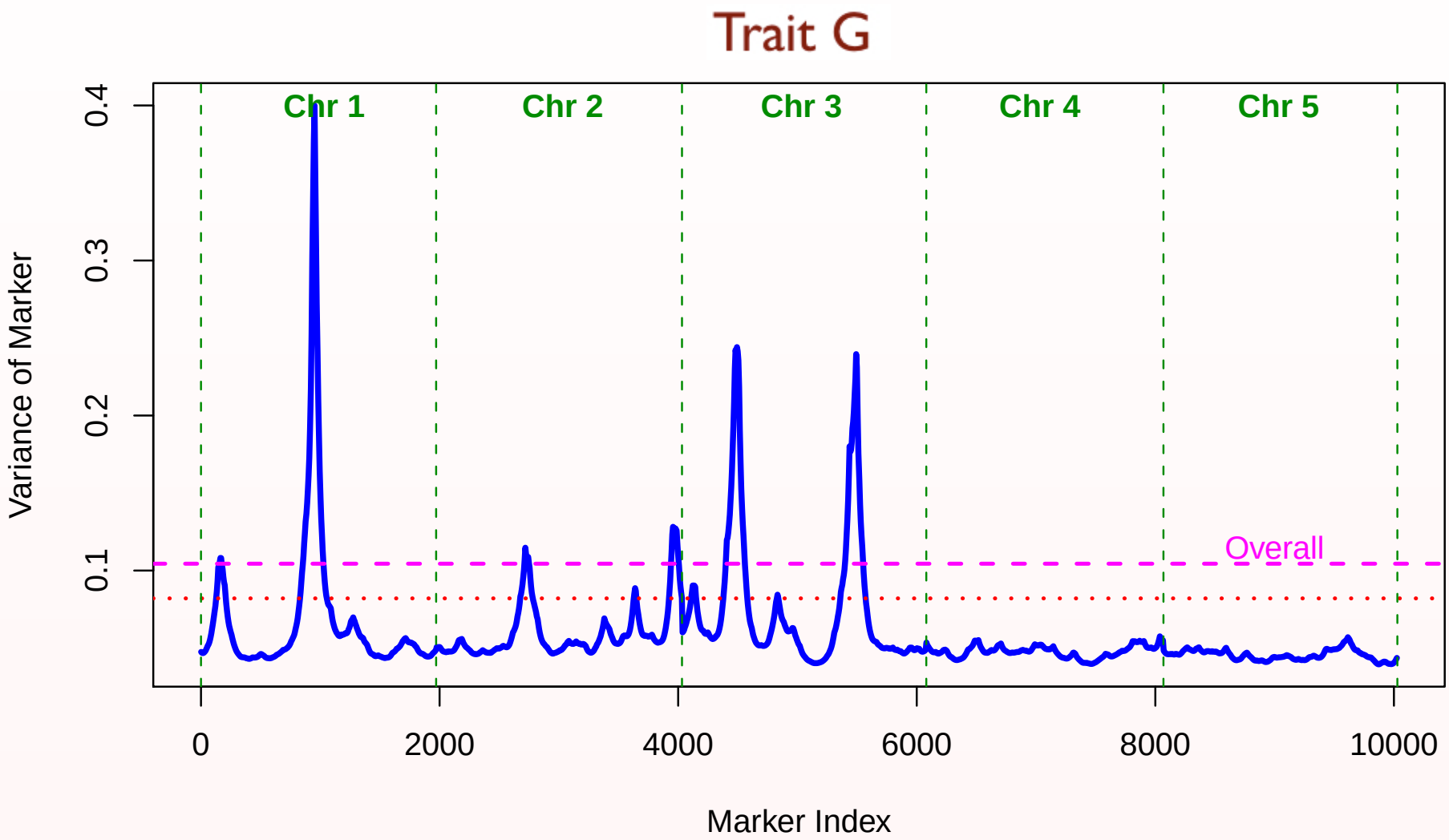
Spatial Correlation

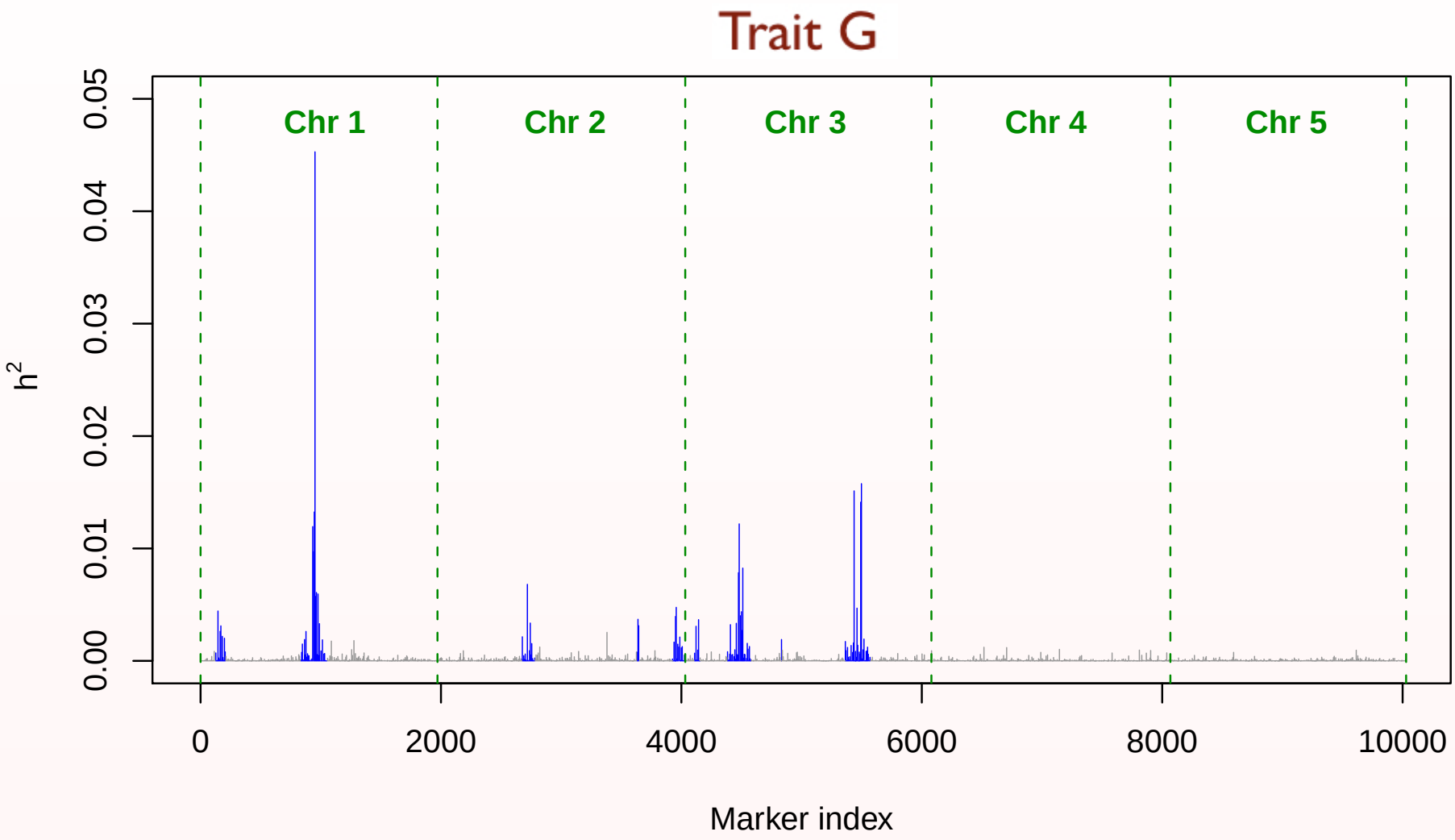
The correlated random effects of the marker-specific variance, b_j's, has a variance-covariance matrix

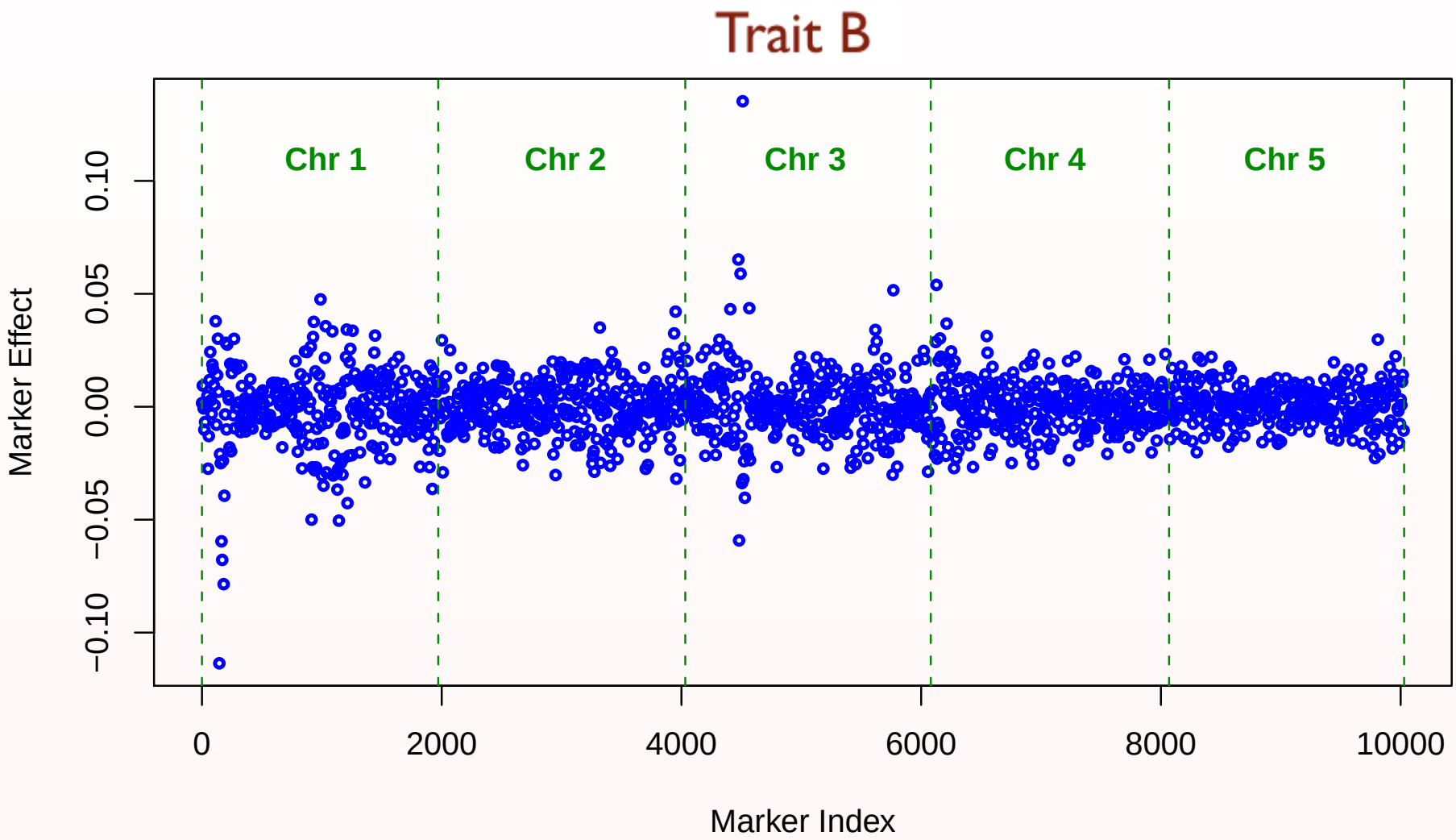
$$\sigma_b^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{q-1} \\ \rho & 1 & \rho & \cdots & \vdots \\ \rho^2 & \rho & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho^{q-1} & \cdots & \cdots & \rho & 1 \end{pmatrix}$$

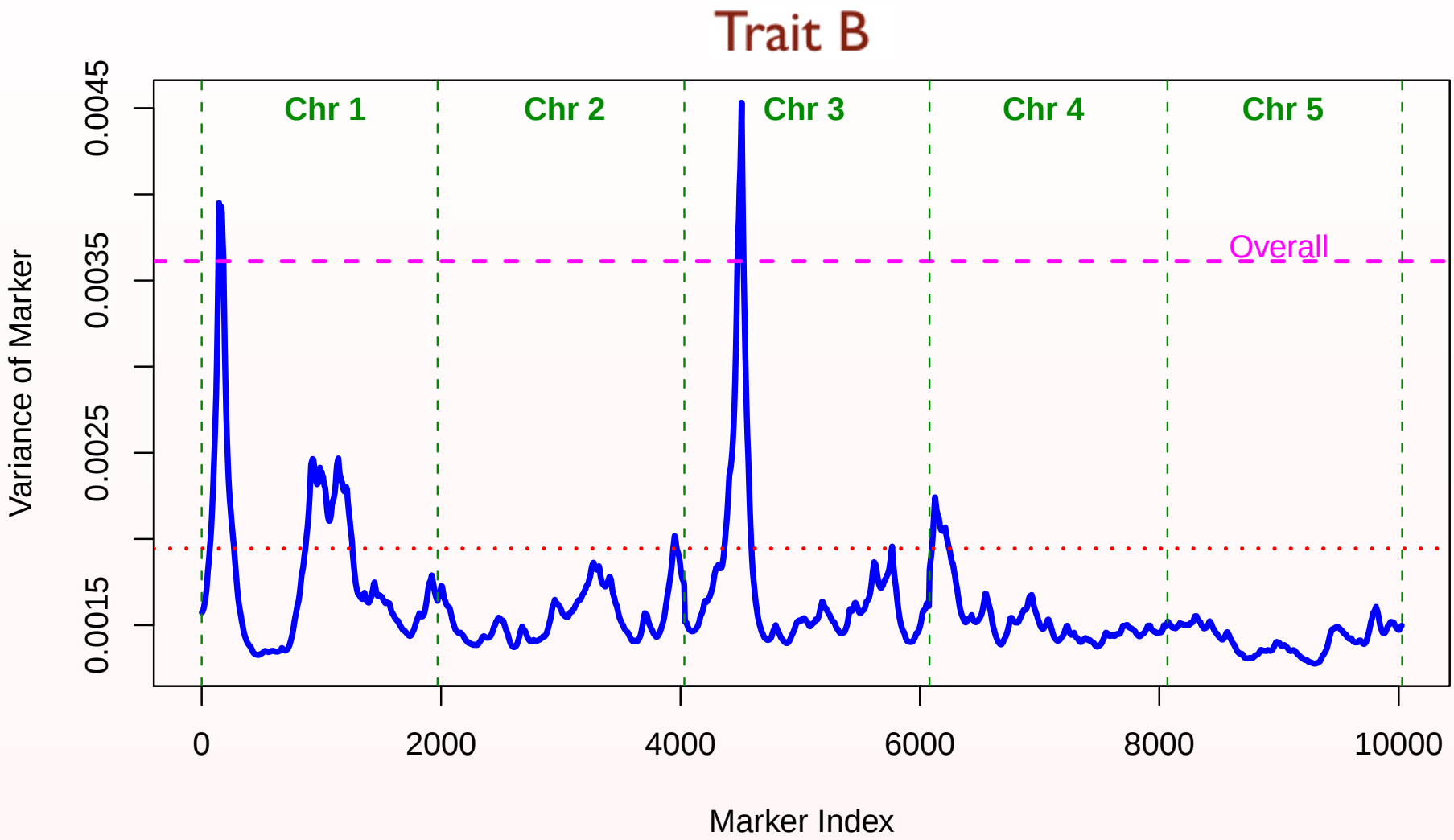
where q is number of markers. This is a *spatial correlation* defined for the second level of DHGLM.

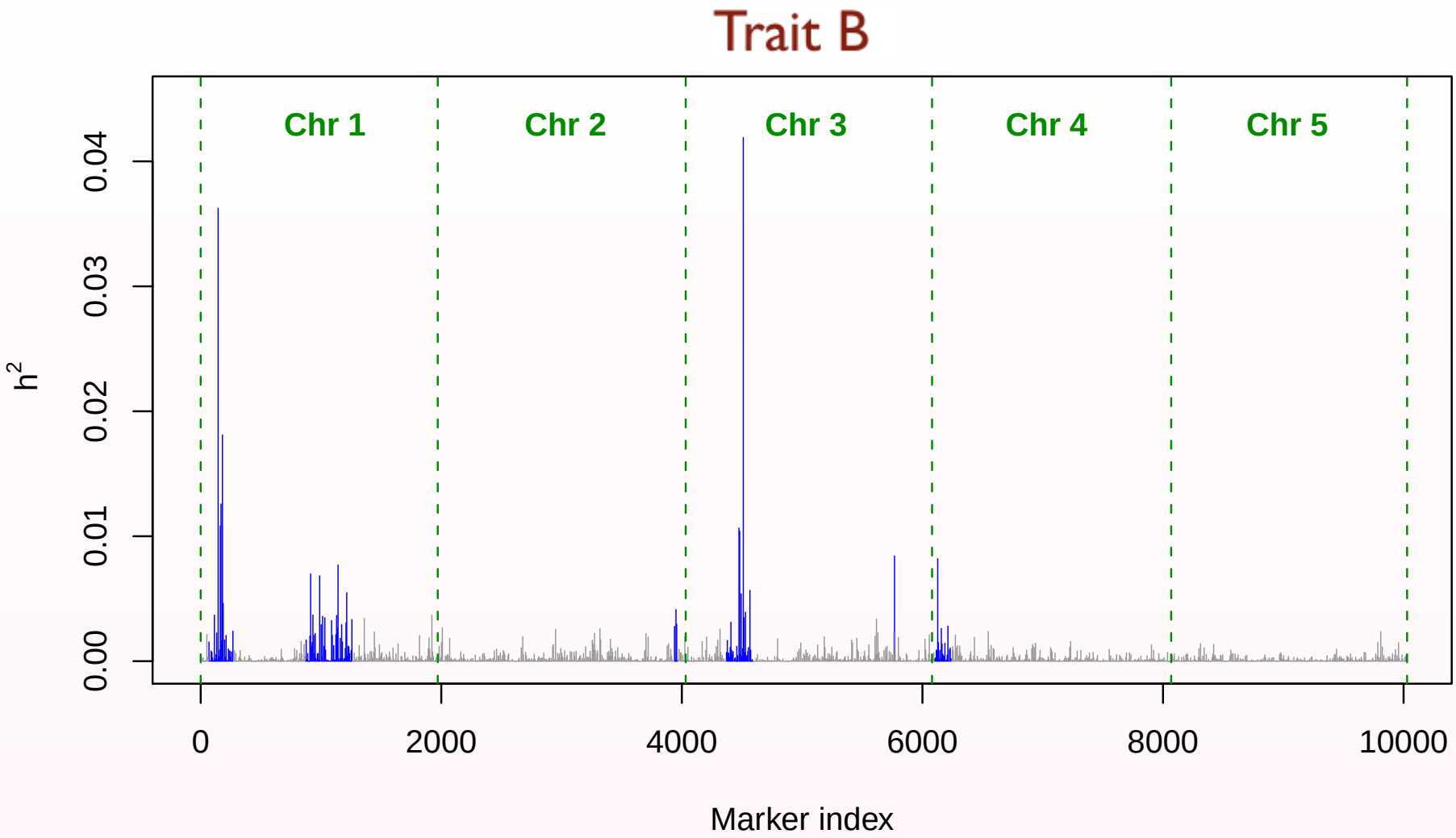

Trait G
Marker Effect
0.5
0.0
−0.5
−1.0
−1.5
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
0
2000
4000
6000
8000
10000
Marker Index

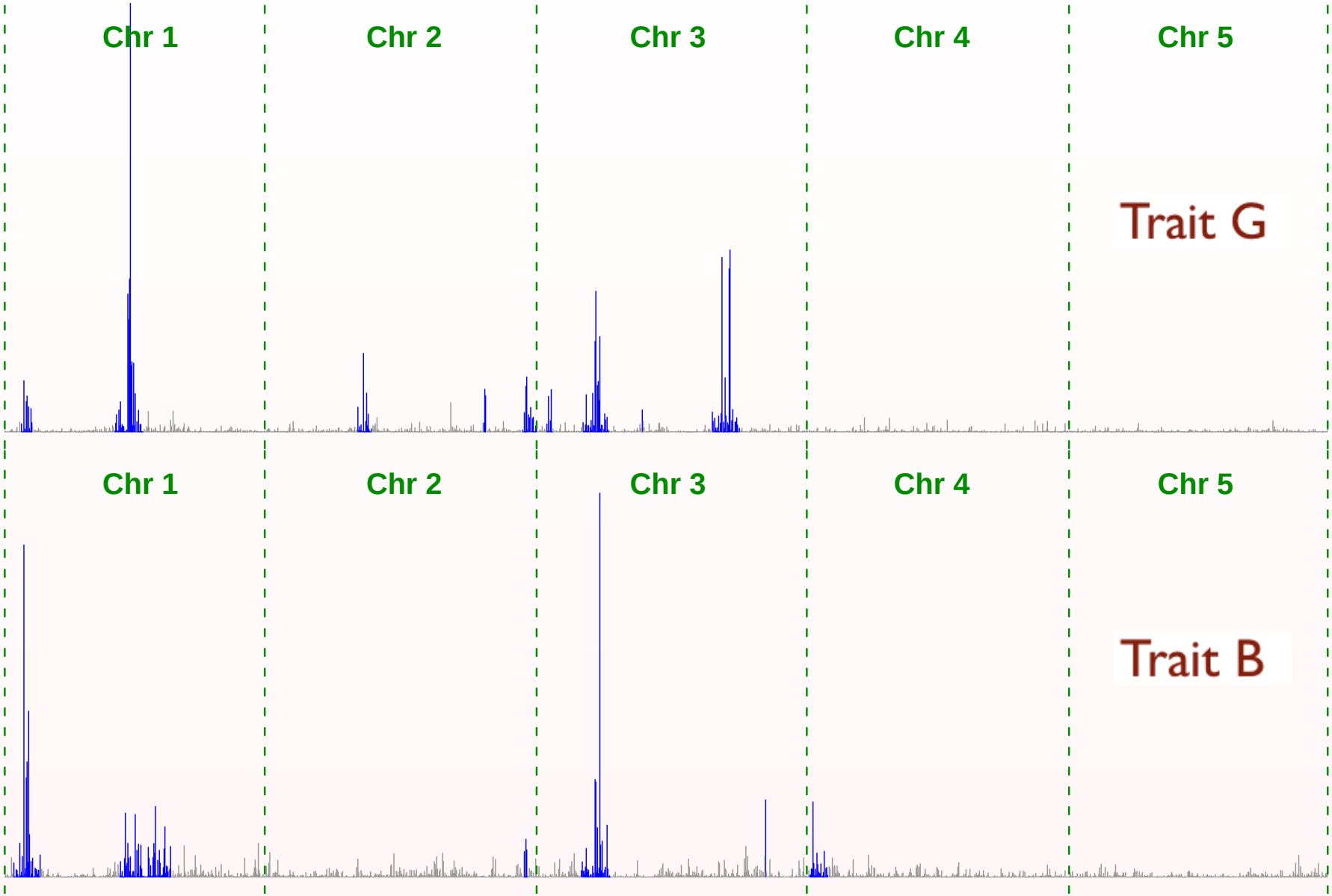

Trait G
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Overall
Variance of Marker
0.1
0.2
0.3
0.4
0
2000
4000
6000
8000
10000
Marker Index


Trait G
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
h2
0.00
0.01
0.02
0.03
0.04
0.05
0
2000
4000
6000
8000
10000
Marker index


Trait B
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Marker Effect
0.10
0.05
0.00
−0.05
−0.10
0
2000
4000
6000
8000
10000
Marker Index


Trait B
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Overall
Variance of Marker
0.0015
0.0025
0.0035
0.0045
0
2000
4000
6000
8000
10000
Marker Index


Trait B
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
h2
0.00
0.01
0.02
0.03
0.04
0
2000
4000
6000
8000
10000
Marker index


Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Trait G
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Trait B

Conclusions & Discussion

- Using h-likelihood, the all-marker shrinkage analysis can be done with a *non-Bayesian* framework.
- The DHGLM algorithm is *fast* and is able to handle *various* distribution families.
- Good starting values lead to faster convergence.

Implementation

- **R** package **hglm** (Rönnegård, Shen & Alam 2010).
- New implementation is in progress...

Acknowledgement

- *Supervisors*
  - Lars Rönnegård and Örjan Carlborg
- *Group members*
  - Lucy Crooks, Weronica Ek, Anna Johansson, Marcin Kierczak, Xidan Li, Stefan Marklund, Ronnie Nelson, and Mats Pettersson.
- *Previous members*
  - François Besnier
- *Collaboration*
  - Carl Nettelblad