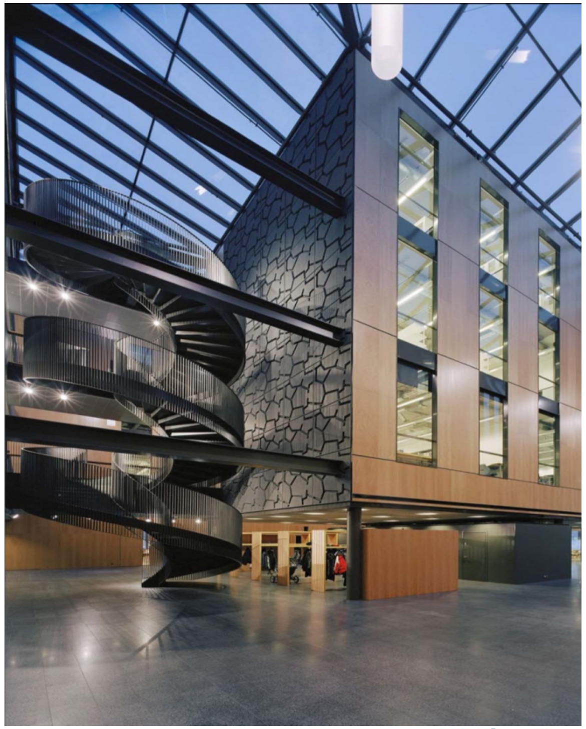
EAAP, Copenhagen, August 25-29th, 2014

Statistical Tools to Dissect the Genetic Architecture of Longitudinal Data



Mikko J. Sillanpää

Department of Mathematical Sciences

Department of Biology,

University of Oulu

Biocenter Oulu

UNIVERSITY of OULU

BIOCENTER OULU

Background

- Dynamic traits or longitudinal traits:

-Change over time during developmental process of life

-Examples: growth traits (*e.g.*, height, boby size), milk production, drug responses

-Phenotype measurements at different time points are often correlated

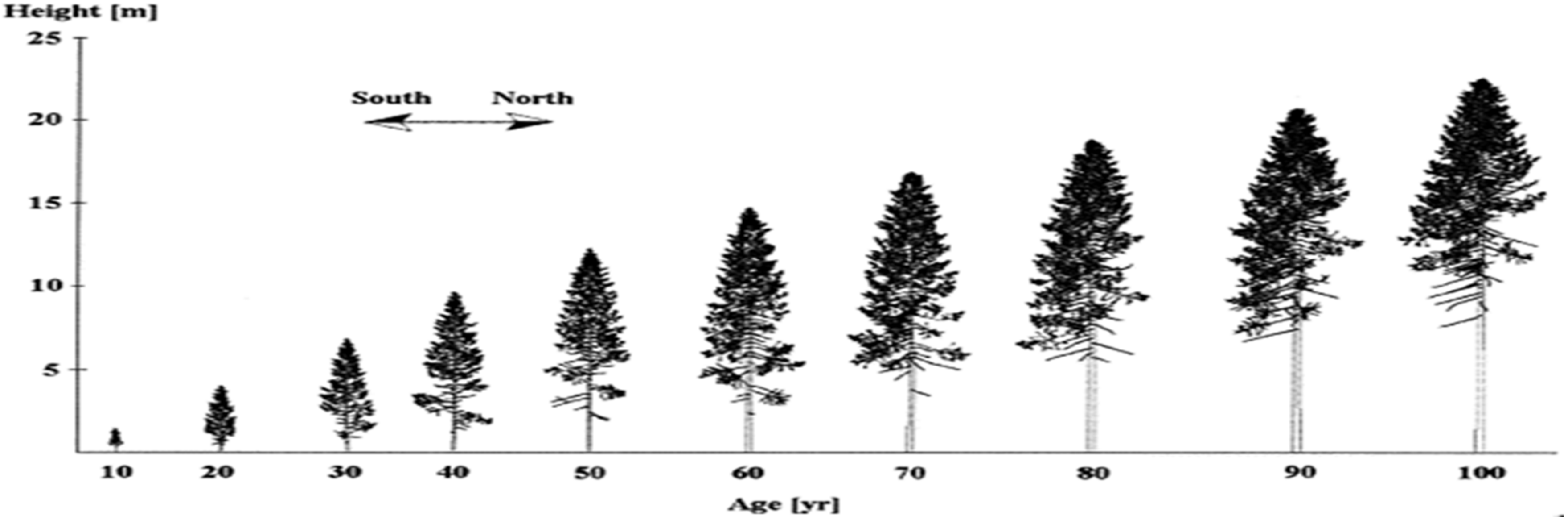
- New automatic phenotyping -> phenotypic measurements with more time-points

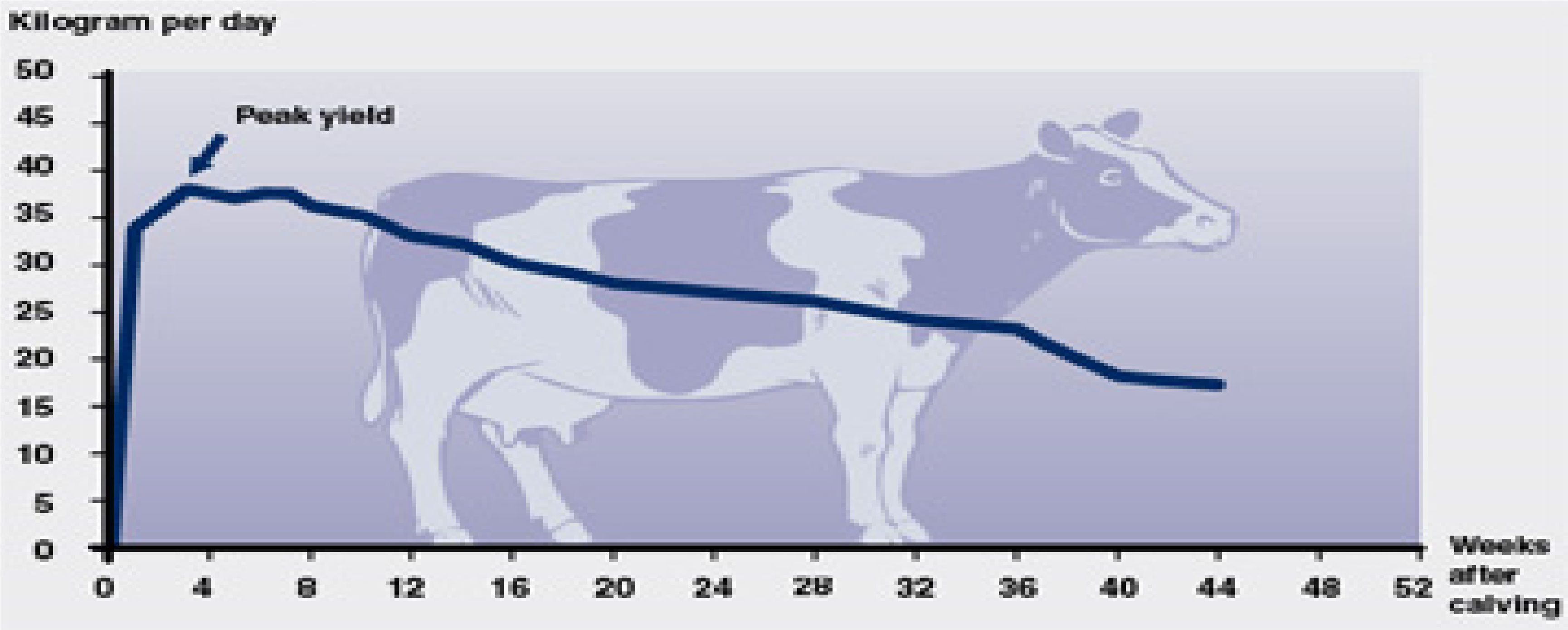
This presentation is based on articles:

Li Z, Hallingbäck HR, Abrahamsson S, Fries A, Anderson B, Sillanpää MJ, Garcia-Gil MR (2014) Functional multi-locus QTL mapping of temporal trends in Scots pine wood traits. (Submitted for publication)

Li Z, Sillanpää MJ (2013) A Bayesian non-parametric approach for mapping dynamic quantitative traits. **Genetics 194: 997-1016.**

Sillanpää MJ, Pikkuhookana P, Abrahamsson S, Knürr T, Fries A, Lerceteau E, Waldmann P, Garcia-Gil MR (2012) Simultaneous estimation of multiple quantitative trait loci and growth curve parameters through hierarchical Bayesian modeling**. Heredity 108: 134-146.**


Height [m]
25
20
15
10
5
South
North
10
20
30
40
50
60
70
80
90
100
Age [yr]


Kilogram per day
50
45
40
35
30
25
20
15
10
5
0
Peak yield
0
4
8
12
16
20
24
28
32
36
40
44
48
52
Weeks after calving

QTL analysis of dynamic traits

- Traditional approach: single trait mapping

  - analyze single time point

  - find loci affecting the trait at a particular developmental stage

- Newer approach: (Multiple-trait) functional mapping

  - Ma *et al*. (2002, Genetics), Wu and Lin (2006, Nat. Rev. Genet.)

  - jointly analyze all repeated measurements of traits

  -understand how loci are influencing the whole developmental process

  - take the temporal correlation among data into account

Multi-trait= effect for each trait (time-point)

Adjacent time points should be more similar = smoothing

Fit smooth curve

1) to phenotypes ?

or

2) to QTL effects ?

1) To fit curve to the phenotypes

First fit curve to the phenotypes over time

- **Then** treat curve parameters as ”traits” in QTL mapping:
- (Heuven and Janss, 2010; BMC Proc; Sillanpää *et al.*, 2012, Heredity)

$$y_i(t_r) = \left\{ \frac{a_i}{1+b_i \exp(c_i t_r)} \right\}_{r=1}^{k}$$

a_i, b_i and c_i are considered as three seperate latent traits

or

even simpler curve

$$y_i(t_r) = a_i + b_i t_r + c_i t_r^{\,2} + e_{i,r}$$

Two-step approach

- 1) Fit simple curve over phenotypes (time points)

  -> own curve for each individual

- 2) Treat curve parameter as ”phenotype” in your favorite QTL mapping method (LASSO, BLASSO, SSVS, PLINK, EMMAX,…)

- 1 & 2 were done simultaneously in Sillanpää et al. (2012; Heredity).

Data are represented as

-**Phenotypes** : y_{ik} , for individual i=1,…,n, and repeated measurements k=1,…,m_i.

-**Time** (hour, day, age…): t_{ik}

-**Genotypes** : x_{ij}, for individual i=1,…,n, and locus j=1,…,p.

Multi-level model

- **Level 1:** Estimate the (linear) temporal trend among the phenotypes

$$y_{ik} = \mu_{i0} + \mu_{i1} t_{ik} + \varepsilon_{ik}, \qquad \varepsilon_{ik} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_{i0}^2)$$

- **Level 2:** Map the trend parameters to genotypes

$$\begin{cases} \mu_{i0} = \alpha_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \alpha_{i0}, & \alpha_{i0} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_0^2) \\ \mu_{i1} = \alpha_1 + \sum_{j=1}^{p} x_{ij}\gamma_j + \alpha_{i1}, & \alpha_{i1} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_1^2) \end{cases}$$

Linear mixed effect model for longitudinal data

- We may use a two-step approach to separately estimate the equations in level 1, and 2.
- Alternatively, it is possible to combine them in one linear mixed effects model (LMM):

$$
\begin{aligned}
y_{ik} &= \mu_{i0} + \mu_{i1} t_{ik} + \varepsilon_{ik} \\
&= (\alpha_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \alpha_{i0}) + (\alpha_1 + \sum_{j=1}^{p} x_{ij}\gamma_j + \alpha_{i1}) t_{ik} + \varepsilon_{ik} \\
&= \underbrace{\alpha_0 + \alpha_1 t_{ik}}_{\text{Fixed intercept and slope terms}} + \underbrace{\alpha_{i0} + \alpha_{i1} t_{ik}}_{\text{Random intercept and slope terms}} + \underbrace{\sum_{j=1}^{p} x_{ij}\beta_j}_{\text{marker effects (stable over time)}} + \underbrace{\sum_{j=1}^{p} x_{ij} t_{ik} \gamma_j}_{\text{Marker-time interaction}} + \varepsilon_{ik}, \qquad \varepsilon_{ik} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_0^2)
\end{aligned}
$$

In Li et al. (2014) , we compared

- Two-step approach: Multi-level LASSO

•1) Fit simple curve to phenotypes

$$y_i(t_r) = a_i + b_i t_r + e_{i,r}$$

•2) Use **LASSO** to map QTLs influencing to the intercept and slope of the curve

- Single-step approach: Bayesian linear mixed effect model

  - all parameters are estimated simultaneosly

LASSO-regression

- *Tibshirani R (1996) Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B. 58: 267-288*

$$\overset{\wedge}{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\},$$

where λ>0 is a tuning parameter.

- LASSO-solution is in optimum when the most regression coefficients in the penalty function $\sum_{j=1}^{p} \left| \beta_j \right|$ are β_j=0.

Bayesian inference

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)\, p(\theta)}{p(\text{data})}$$

- p(data|θ) is a likelihood
- p(θ) is a prior density
- p(data) is a normalizing constant

Priors for intercept and slope parameters

- Flat uniform priors for fixed random intercept and slope parameters:

$$\alpha_0 \sim U(-\infty, \infty),\ \alpha_1 \sim U(-\infty, \infty)$$

- Normal priors for random intercept and slope parameters:

$$[\alpha_{i0}, \alpha_{i1}]^T \,\big|\, \mathbf{\Sigma}_{2\times 2} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma}_{2\times 2}),$$

$$\mathbf{\Sigma}_{2\times 2} \sim W^{-1}(\mathbf{\Psi}_{2\times 2}, v),$$

$$\mathbf{\Psi}_{2\times 2} = \mathbf{I}_{2\times 2}, v = 1.$$

Priors for marker effects β (or γ)

- Spike and slab prior (a mixture of a normal and point mass at zero)

$$\beta_j \left| r_j \sim (1 - r_j)\mathrm{I}_{\left\{\beta_j = 0\right\}} + r_j N(0, \sigma_j^2), \qquad (r_j = 0, 1)\right.$$

$$p(r_j \mid w) = w^{r_j}(1 - w)^{1 - r_j},$$

$$\sigma_j^2 \sim \text{Inv-Gamma}(0.1, 0.1).$$

Computation and posterior inference

- MCMC (Gibbs) sampling is used to evaluate the posterior distribution.

- From the MCMC samples, A Bayesian false discovery rate (BFDR) type of decision rule was derived to identify QTLs (e.g. Ventrucci and Scott 2011)

LASSO: uncertainty measure

- Stability selection (Meinshausen and Bühlmann 2010)

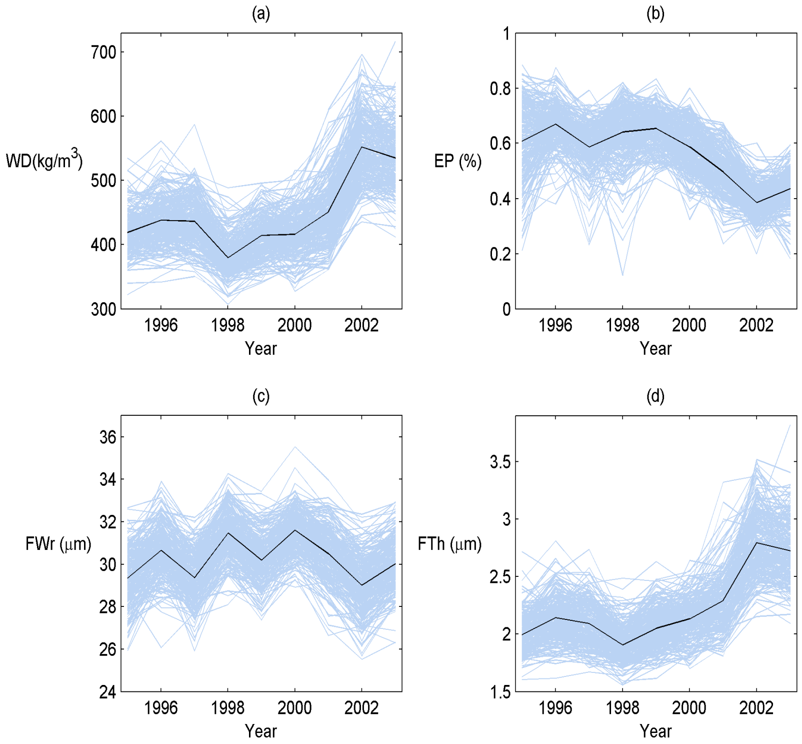
  -closely related to boostraping and false discovery rate control

  -provides a selection probability for each marker (if probabilily is close to 1, then we say that the marker is likely to be a QTL)

- Note that the stability selection is more liberal compared to some mutiple hypothesis testing methods achieving familiy wise error control such as bonferroni correction

Phenotype Data

- The studied field test: Flurkmark (S23F881485), located 25 km north of Umeå in northern Sweden (lat. 64°02'N, long 20°30'E, alt 115 m a sl)
- 286 trees were selected for wood sampling. They were located together in order to minimize the environmental variation.
- Wood property traits: wood density (WD), radial and tangential fiberwidth (FWr & FWt), fiberwall thickness (FTh), microfibril angle (MFA), dynamic modulus of elasticity (MOE), grain angle (GA)
- Repeat measurements over 9 years during 1995-2003 (age: 7-15)



Trajectory of four wood traits including (a) wood density, (b) early wood percentage, (c) radial fiberwidth and (d) fiberwall thickness

Genotype data

- 492 progeny individuals were thus genotyped using 508 AFLP markers

- After pre-processing steps (such as filtering out some markers with low coverage), we eventually obtained:

  - AFLP data set with 273 individuals and 153 AFLP markers
    (expect in GA, 451 individuals)

Results from BLMM analysis

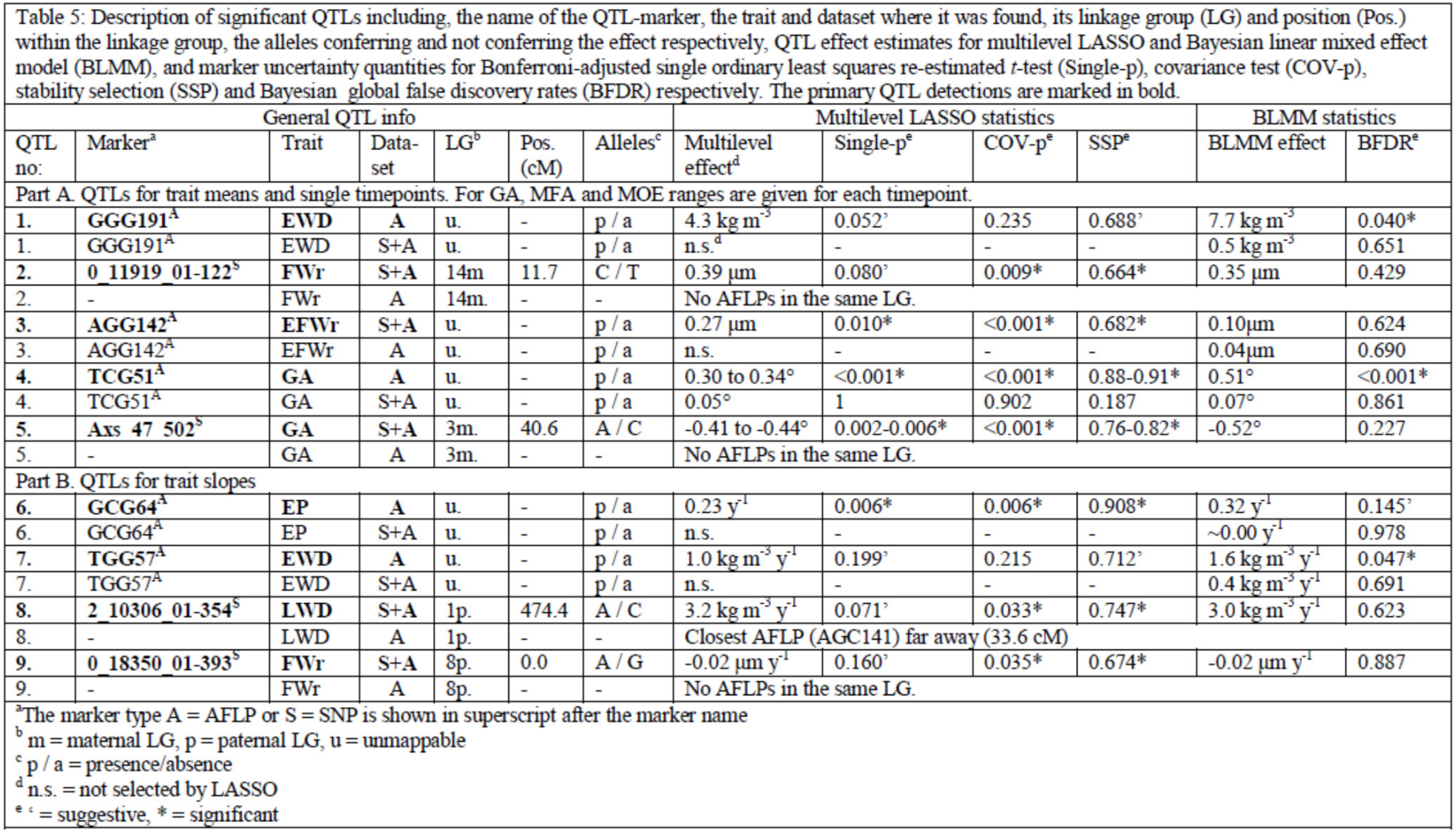


Table 5: Description of significant QTLs including, the name of the QTL-marker, the trait and dataset where it was found, its linkage group (LG) and position (Pos.) within the linkage group, the alleles conferring and not conferring the effect respectively, QTL effect estimates for multilevel LASSO and Bayesian linear mixed effect model (BLMM), and marker uncertainty quantities for Bonferroni-adjusted single ordinary least squares re-estimated *t*-test (Single-p), covariance test (COV-p), stability selection (SSP) and Bayesian global false discovery rates (BFDR) respectively. The primary QTL detections are marked in bold.

General QTL info							Multilevel LASSO statistics				BLMM statistics	
QTL no:	Marker[a]	Trait	Data-set	LG[b]	Pos. (cM)	Alleles[c]	Multilevel effect[d]	Single-p[e]	COV-p[e]	SSP[e]	BLMM effect	BFDR[e]
Part A. QTLs for trait means and single timepoints. For GA, MFA and MOE ranges are given for each timepoint.												
1.	**GGG191^{A}**	**EWD**	**A**	u.	-	p / a	4.3 kg m^{-3}	0.052’	0.235	0.688’	7.7 kg m^{-3}	0.040*
1.	GGG191^{A}	EWD	S+A	u.	-	p / a	n.s.^{d}	-	-	-	0.5 kg m^{-3}	0.651
2.	**0_11919_01-122^{S}**	**FWr**	**S+A**	14m	11.7	C / T	0.39 µm	0.080’	0.009*	0.664*	0.35 µm	0.429
2.	-	FWr	A	14m.	-	-	No AFLPs in the same LG.					
3.	**AGG142^{A}**	**EFWr**	**S+A**	u.	-	p / a	0.27 µm	0.010*	<0.001*	0.682*	0.10µm	0.624
3.	AGG142^{A}	EFWr	A	u.	-	p / a	n.s.	-	-	-	0.04µm	0.690
4.	**TCG51^{A}**	**GA**	**A**	u.	-	p / a	0.30 to 0.34°	<0.001*	<0.001*	0.88-0.91*	0.51°	<0.001*
4.	TCG51^{A}	GA	S+A	u.	-	p / a	0.05°	1	0.902	0.187	0.07°	0.861
5.	**Axs 47 502^{S}**	**GA**	**S+A**	3m.	40.6	A / C	-0.41 to -0.44°	0.002-0.006*	<0.001*	0.76-0.82*	-0.52°	0.227
5.	-	GA	A	3m.	-	-	No AFLPs in the same LG.					
Part B. QTLs for trait slopes												
6.	**GCG64^{A}**	**EP**	**A**	u.	-	p / a	0.23 y^{-1}	0.006*	0.006*	0.908*	0.32 y^{-1}	0.145’
6.	GCG64^{A}	EP	S+A	u.	-	p / a	n.s.	-	-	-	~0.00 y^{-1}	0.978
7.	**TGG57^{A}**	**EWD**	**A**	u.	-	p / a	1.0 kg m^{-3} y^{-1}	0.199’	0.215	0.712’	1.6 kg m^{-3} y^{-1}	0.047*
7.	TGG57^{A}	EWD	S+A	u.	-	p / a	n.s.	-	-	-	0.4 kg m^{-3} y^{-1}	0.691
8.	**2_10306_01-354^{S}**	**LWD**	**S+A**	1p.	474.4	A / C	3.2 kg m^{-3} y^{-1}	0.071’	0.033*	0.747*	3.0 kg m^{-3} y^{-1}	0.623
8.	-	LWD	A	1p.	-	-	Closest AFLP (AGC141) far away (33.6 cM)					
9.	**0_18350_01-393^{S}**	**FWr**	**S+A**	8p.	0.0	A / G	-0.02 µm y^{-1}	0.160’	0.035*	0.674*	-0.02 µm y^{-1}	0.887
9.	-	FWr	A	8p.	-	-	No AFLPs in the same LG.					

[a] The marker type A = AFLP or S = SNP is shown in superscript after the marker name

[b] m = maternal LG, p = paternal LG, u = unmappable

[c] p / a = presence/absence

[d] n.s. = not selected by LASSO

[e] ‘ = suggestive, * = significant

Summary

- QTLs were detected in several traits such as early wood density (EWD), and radial and tangential fiberwidth (FWr)

- A few QTLs seem to be biological interpretable (at protein level)

- No previous longitudinal QTL analysis has been performed for wood property traits, and no earlier results available that we can compare with

- The findings are rather hypothetical, and requires further molecular investigations

2) To fit curve to QTL effects

- **In Li and Sillanpää (2013) ,** we fitted smooth curve to QTL effects instead of phenotypes

- This is so called **VARYING COEFFICIENT MODEL** which have own effect coefficient for each trait (time point)

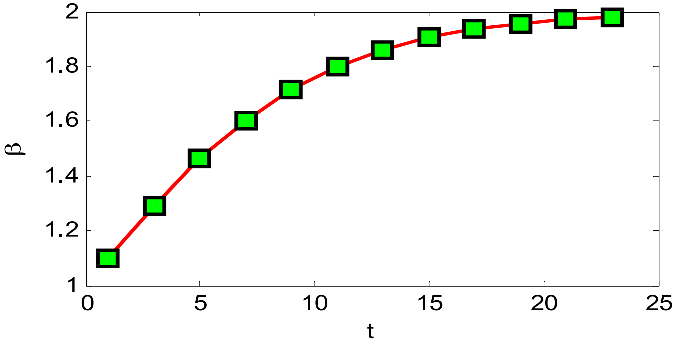
2) To fit curve to QTL effects

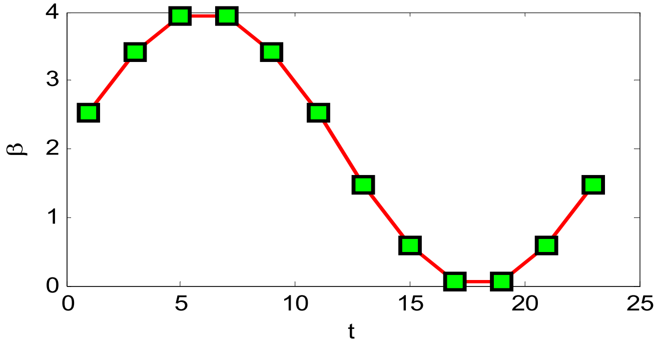
- Phenotype $y_i(t_r)$, Individuals $i=1,\ldots\ldots,n$, time points $t_1,\ldots\ldots,t_k$ (hours, days, years…)
- genotype x_i=1, 0, -1 for AA, AB, BB
- single locus model multiple loci model

$$y_i(t_1) = \beta(t_1)x_i + e_i(t_1)$$
$$y_i(t_2) = \beta(t_2)x_i + e_i(t_2)$$
$$\vdots$$
$$y_i(t_k) = \beta(t_k)x_i + e_i(t_k)$$

$$y_i(t_1) = \sum_{j=1}^{p} \beta_j(t_1)x_{ij} + e_i(t_1)$$
$$y_i(t_2) = \sum_{j=1}^{p} \beta_j(t_2)x_{ij} + e_i(t_2)$$
$$\vdots$$
$$y_i(t_k) = \sum_{j=1}^{p} \beta_j(t_k)x_{ij} + e_i(t_k)$$

- **We consider the genetic effects $\beta(t_1), \ldots, \beta(t_k)$ jointly as a trend function over time.**





How to model residual covariance?

$$
\begin{aligned}
y_i(t_1) &= \beta(t_1)x_i + e_i(t_1) \\
y_i(t_2) &= \beta(t_2)x_i + e_i(t_2) \\
&\vdots \\
y_i(t_k) &= \beta(t_k)x_i + e_i(t_k)
\end{aligned}
$$

If the distribution of traits is normal, the residual terms **e**$_i$=[$e(t_1)$, …, $e(t_k)$] can be specified as $\mathbf{e}_i \sim N(0, \sigma^2\Sigma)$. The covariance matrix **Σ** describes the temporal correlation among non-QTL (*i.e.,* environmental) factors.

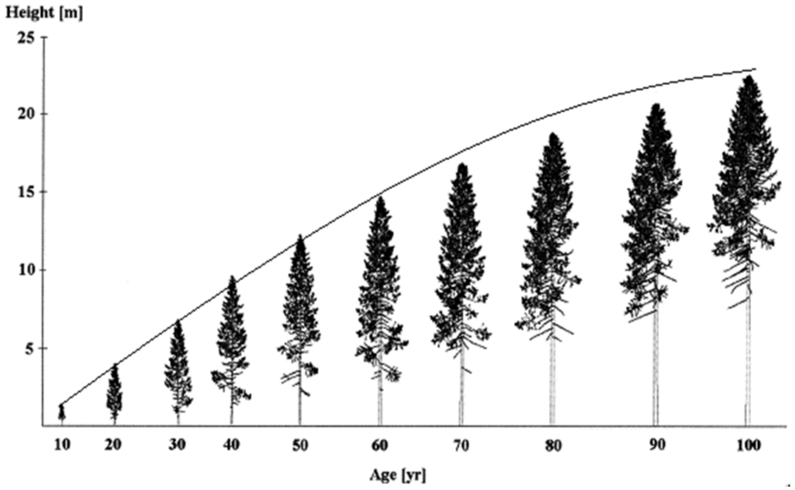
We consider two possible covariance structures (i) diagonal, and (ii) AR(1)

$$
\Sigma_{\text{Diag}} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_k^2 \end{bmatrix}
\quad
\Sigma_{\text{AR(1)}} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{k-1} \\ \rho & 1 & \rho & \cdots & \rho^{k-2} \\ \rho^2 & \rho & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \rho \\ \rho^{k-1} & \rho^{k-2} & \cdots & \rho & 1 \end{bmatrix}, \quad 0 < \rho < 1
$$

Parametric methods

- Used when the curve of dynamic traits is simple
- Model $\beta(t)$ as a known parametric function
- Example: logistic growth curve

Growth trajectory of Scots pine



Likelihood function

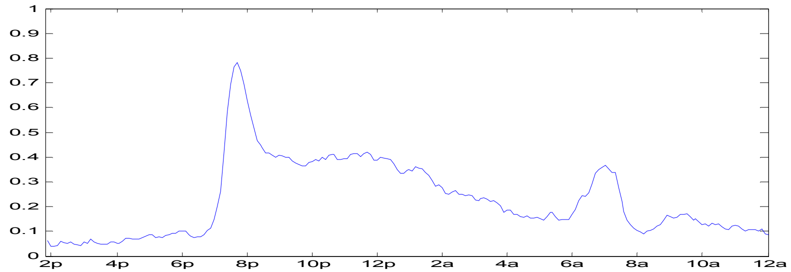
$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} N(\mathbf{y}_i \mid x\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\beta} = \left\{ \frac{a}{1 + b\exp(ct_r)} \right\}_{r=1}^{k}$$

Estimate parameters *a*, *b* and *c* by maximum likelihood

Non-parametric methods

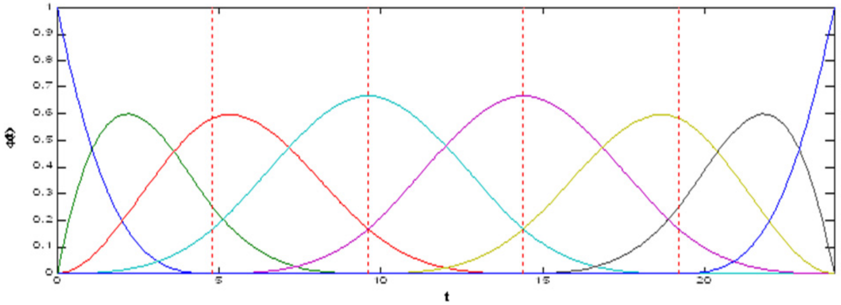
- When the curve of dynamic traits is complicated, we cannot use any known function to describe it



Active state probability of mouse (Xiong *et al*. 2011, Genetics)

- Basis expansions: represent $\beta(t)$ as a linear combination of some basis functions, $\beta(t)=\alpha_1\Phi_1(t)+\alpha_2\Phi_2(t)+\cdots+\alpha_{\mathrm{m}}\Phi_m(t)$

- We choose B-spline basis functions.



Cubic B-spline bases with 4 interior knots

P-spline: penalized B-spline

•In B-splines, choosing an appropriate number of knots is crucial:

-too few or two many knots: underfitting or overfitting

•P-spline (Eilers and Marx 1996, Stat. Sci.):

-pre-specify a relatively large number of knots

-add a difference penalty to the "likelihood" function of $\beta(\mathrm{t})$ in order to avoid overfitting

$$\lambda(\alpha_2-\alpha_1)^2+\lambda(\alpha_3-\alpha_2)^2+\cdots+\lambda(\alpha_m-\alpha_{m-1})^2$$

-In Bayesian statistics, the difference penalty is corresponding to a random walk prior (Lang and Brezger 2004, J. Comput. Graph. Stat.)

Our Bayesian hierarchical model

$p(\boldsymbol{\theta} \mid \mathbf{Y})$ $p(\mathbf{Y} \mid \boldsymbol{\theta})$ $p(\boldsymbol{\theta})$

- Posterior $\propto$ Likelihood × Prior

↓

$$\prod_{i=1}^{n} \text{MVN}(\mathbf{y}_i \mid \boldsymbol{\beta}_0 + \sum_{i=1}^{p} x_{ij} \boldsymbol{\beta}_j, \boldsymbol{\Sigma}),$$

↓

$$\boldsymbol{\beta}_j = \boldsymbol{\Psi} \boldsymbol{\alpha}_j$$

- Prior:

- random walk prior for $\boldsymbol{\alpha}_j$: $p(\boldsymbol{\alpha}_j \mid \tau_j^2) p(\tau_j^2) = \text{MVN}(\boldsymbol{\alpha}_j \mid \mathbf{0}, \tau_j^2 \mathbf{K}^{-1}) \mathbf{IG}(\tau_j^2 \mid 0.0001, 0.0001),$

  where matrix **K** contains the information of the difference penalty

- non-informative priors for $\boldsymbol{\Sigma}_{\text{Diag}}$ or $\boldsymbol{\Sigma}_{\text{AR(1)}}$:

  $p(\boldsymbol{\Sigma}_{\text{Diag}}) \propto \prod_{l=1}^{k} \frac{1}{\sigma_l^2}$, or $p(\boldsymbol{\Sigma}_{\text{AR(1)}}) \propto \frac{1}{\sigma^2} 1_{(0<\rho<1)}$

200 time points = 200 traits

MCMC estimation of parameters of multitrait methods is SLOW!!

Thus, we need faster estimation approaches

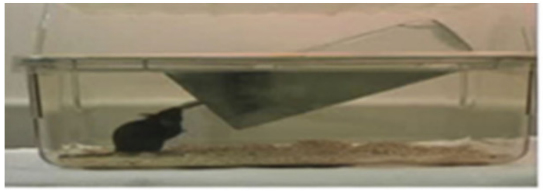
Computation and posterior inference

- Variational Bayes (VB): a deterministic approximation algorithm for posterior inference (Beal 2003, PhD thesis)

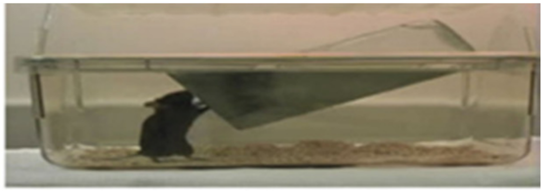
- is used to estimate the mode of the posterior distribution and for variable selection.

Case study: mouse behavioral data

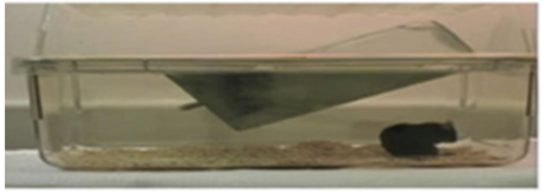
- Xiong *et al*. (2011, Genetics), Goulding *et al*. (2008, PNAS)
- Genotype: 89 backcross individuals, 233 SNPs distributed over 19 chromosomes.
- Phenotype: active state probability. 222 repeated measurements under a 12h:12h light:dark cycle
- We applied a logit transformation ($\log(\frac{y}{1-y})$) to make the phenotypic data more normally disributed



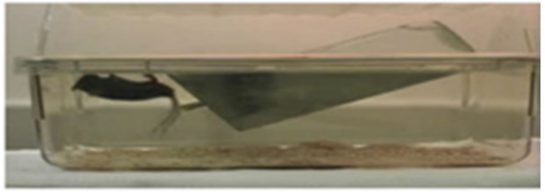
Drink



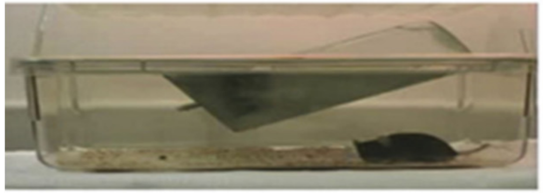
Eat



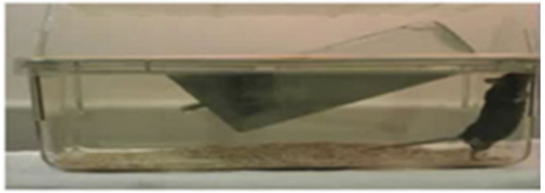
Groom



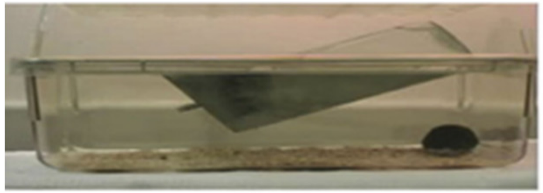
Hang



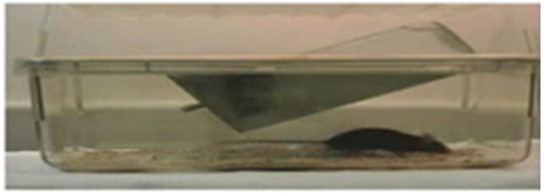
Micromovement



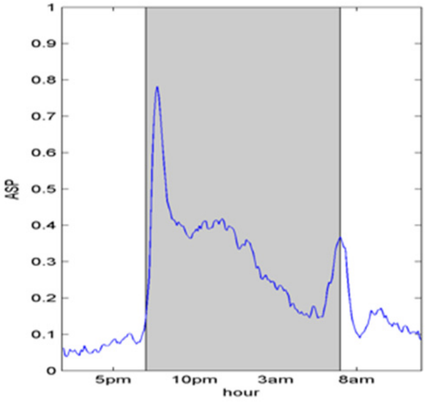
Rear

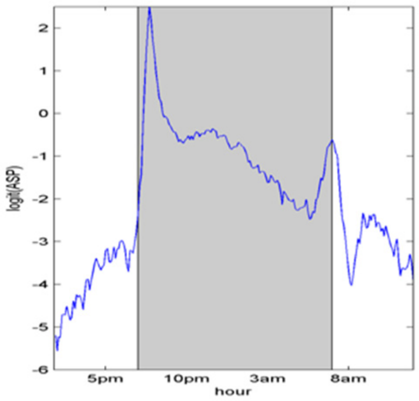


Rest



Walk



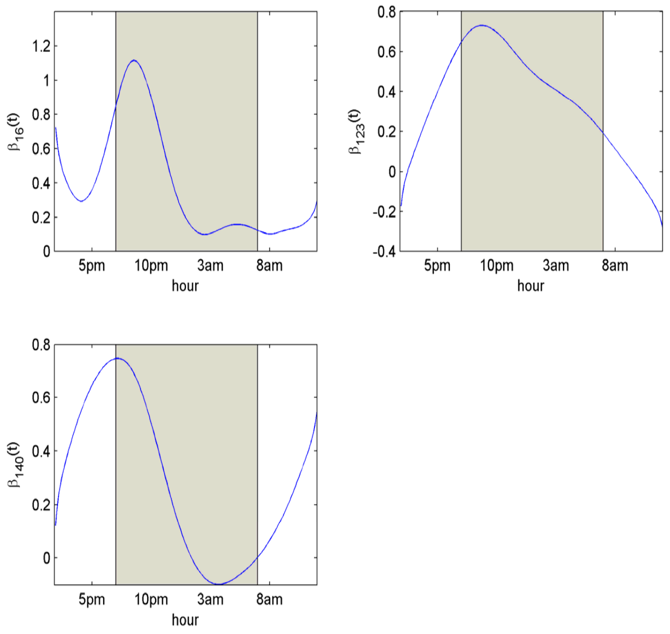


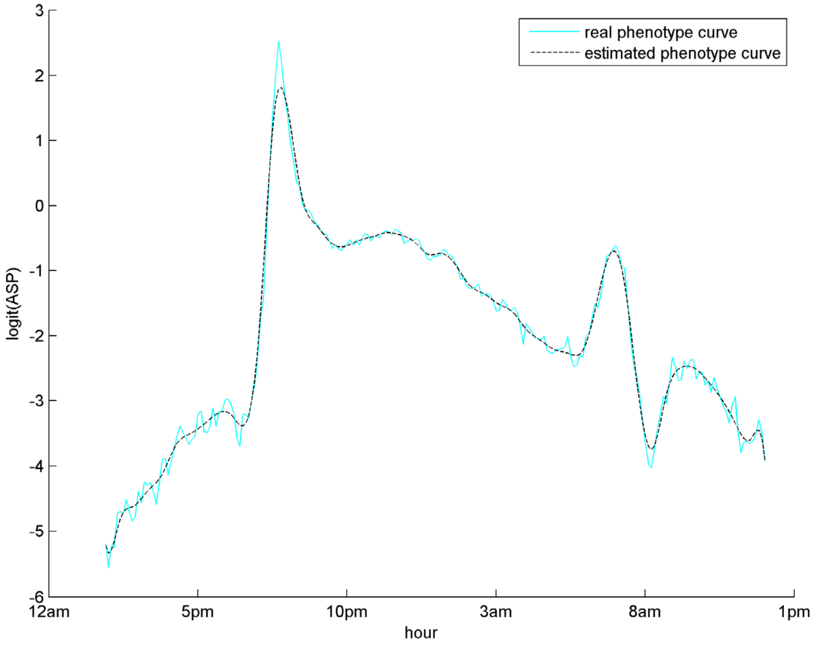
Case study: mouse behavioral data

- Previously analyzed by a single loci approach of Xiong *et al*. (2011)
- Our Bayesian model search algorithm (assuming AR(1) residual covariance) detects 3 important markers.
- 2 markers with largest effects are located on chromosome 1 and 9, respectively. This agrees with the findings in Xiong *et al*. (2011).
- If diagonal residual covariance is assumed, the method tends to find many false positive signals

Marker ID	Chromosome	Location (cM)	P-value (based on Wald test)
16 (rs1347625)	1	81.40	1×10^{-11}
123 (rs6207781)	9	20.74	6×10^{-8}
140 (rs3654717)	10	55.93	2×10^{-6}

Case study: mouse behavioral data





Estimated effect curves

Re-estimated phenotype mean trajectory

Summary

- Benefits of functional mapping:

  (1)increase the power to detect QTLs by borrowing strength from nearby time points

  (2) control the false positives by incorporating the residual covariances

  (3) better interpretation of the results

- Compared to other function mapping approaches, our method is

  + fast
  + easy to use, suitable for many different types of dynamic traits
  - uncertainty measure is inaccurate, due to the approximation nature of VB

Acknowledgements

- Finnish Doctoral Program of Population Genetics
- Academy of Finland

More time consuming to have additional traits than markers especially with AR(1)

MCMC infeasible for 200 traits

Roughly

Mouse data, 200 traits, AR(1),

VB - half an hour

Mouse data, 200 traits, Diagonal

VB - several minutes

Simulated data, 100 traits, AR (1),

VB – 15-20 minutes

Simulated data, 100 traits, Diagonal

VB- several minutes