Phylogenetic methods for quantitative trait mapping with complex data sets

Katie Thompson[1]
Joint work with: Laura Kubatko[2]

[1]Dr. Bing Zhang Department of Statistics, University of Kentucky
[2]Department of Statistics, The Ohio State University

October 24, 2024

OUTLINE

- Introduction
- Motivation
- Proposed Methods
- Simulation Study
- Conclusions
- Future Directions

INTRODUCTION

Motivation: Search for Single Nucleotide Polymorphisms (SNPs) and/or external covariates associated with quantitative traits

Goals:

- Detection of Associated SNPs
- Localization of Associated SNPs
- Detection of Associated Covariates

Aims of Proposed Work:

- Combine ideas from stochastic processes and phylogenetics to simulate genetic and trait data
- Identify SNPs and/or external covariates associated with quantitative traits using a proposed likelihood score statistic

EXAMPLE: Outbred Mice Study

(Zhang et al. 2012)

Mice Data:

- Organisms: 288 outbred male mice
- Genetic Data: Genome-wide Association Study SNP data
- Quantitative Trait Data: High-Density Lipoprotein (HDL) level for each mouse

EXAMPLE: Deer Mice Study

(Linnen et al. 2013)

Deer Mice (*Peromyscus maniculatus*) Study:

- *Organisms:* 91 wild-caught mice from the edge of the Nebraska Sand Hills
- *Genetic Data:* SNP data
- *Quantitative Traits:* nine quantitative color phenotypes
- *Covariates:* weight, body length, tail length, ear length, foot length, sex, and pregnancy status

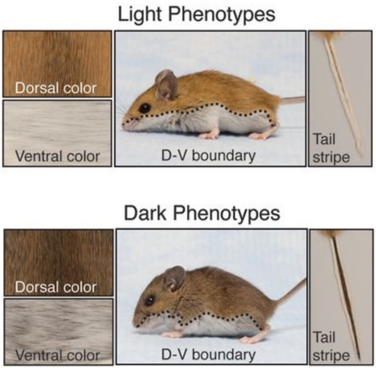


Figure 2A, Linnen et al. 2013

EXAMPLE: Deer Mice Study (Linnen et al. 2013)

- Researchers are interested in identifying regions of the genome contributing to mouse coat color.
- Previous work has shown that much of the variation in coat color appears to be controlled by a single gene, *Agouti*.

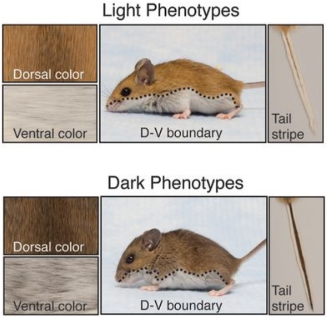
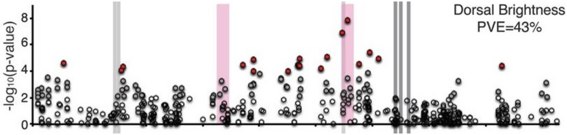


Figure 2A, Linnen et al. 2013



Excerpt from Figure 2C, Linnen et al. 2013

INTRODUCTION: The Data

Phenotypic Data:

- Quantitative trait data
- One observation per individual in the study

INTRODUCTION: The Data

Phenotypic Data:

- Quantitative trait data
- One observation per individual in the study

SNP Data:

- Can be represented as a collection of binary random variables

INTRODUCTION: The Data

Phenotypic Data:

- Quantitative trait data
- One observation per individual in the study

SNP Data:

- Can be represented as a collection of binary random variables

Data Example: (3 diploid individuals)

Person	DNA Sequences	SNP Data
1	...AACTGGTCCAACGTC...	...000...
1	...AACTGGTCCACCGTC...	...010...
2	...AACTTGTCCAACATC...	...101...
2	...AACTGGTCCACCATC...	...011...
3	...AACTTGTCCAACGTC...	...100...
3	...AACTGGTCCAACATC...	...001...

INTRODUCTION: The Data

Phenotypic Data:

- Quantitative trait data
- One observation per individual in the study

SNP Data:

- Can be represented as a collection of binary random variables

Data Example: (3 diploid individuals)

Person	DNA Sequences	SNP Data	Trait Data
1	...AACTGGTCCAACGTC...	...000...	175.8
1	...AACTGGTCCACCGTC...	...010...	175.8
2	...AACTTGTCCAACATC...	...101...	115.6
2	...AACTGGTCCACCATC...	...011...	115.6
3	...AACTTGTCCAACGTC...	...100...	157.3
3	...AACTGGTCCAACATC...	...001...	157.3

INTRODUCTION: The Data

Quantitative Trait Data:

- One observation per individual in the study

SNP Data:

- Can be represented as a collection of binary random variables

Covariate Data:

- One observation per individual in the study

Data Example: 3 diploid individuals

Person	SNP Data	Covariate Data	Trait Data
1	...000...	32.2	175.8
1	...010...	32.2	175.8
2	...101...	28.2	115.6
2	...011...	28.2	115.6
3	...100...	30.2	157.3
3	...001...	30.2	157.3

INTRODUCTION: Previous Methods

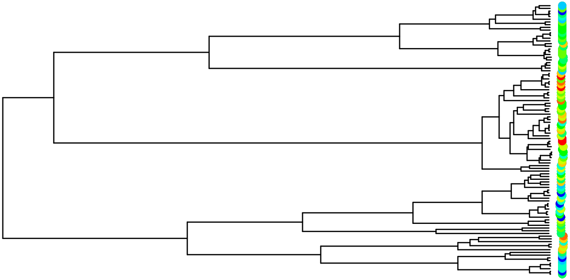
Regression-based Methods:

- Tend to detect large genetic signals
- Assume observations have means that are related directly to their genotype and covariate value.
- Assume observations are independent

Phylogenetic Methods:

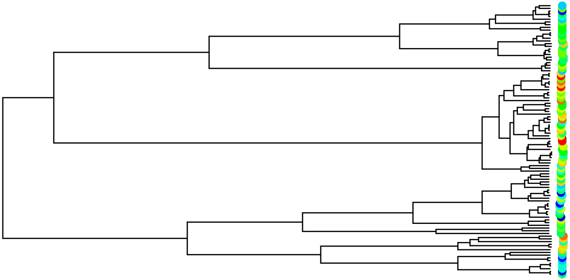
- Use the relationships within each SNP to gain information about the correlation structure among individuals.
- Use this correlation structure to help improve data analysis.
- Assume observations are normally distributed
- Assume observations have means that related directly to their covariate value and their evolutionary history.

INTRODUCTION: Motivating Example



INTRODUCTION: Motivating Example

$t = 0.2211$
with 21.65 df
p-value $= 0.8271$



INTRODUCTION: Previous Methods

Regression-based Methods:

- Tend to detect large genetic signals
- Assume observations have means that are related directly to their genotype and covariate value.
- Assume observations are independent

Phylogenetic Methods:

- Use the relationships within each SNP to gain information about the correlation structure among individuals.
- Use this correlation structure to help improve data analysis.
- Assume observations are normally distributed
- Assume observations have means that related directly to their covariate value and their evolutionary history.

INTRODUCTION: The Data

How the Data are Used:

- Use **SNP data** to learn about evolutionary relationships
- Use **trait data** and **covariate data** to find connections between the trait and the SNP and/or the covariate

Note:

- Relationships among SNPs exist due to evolution of genetic data.
- Relationships among trait values are imposed by the relationships among SNPs and environmental covariates.

INTRODUCTION: The Phylogenetic Framework

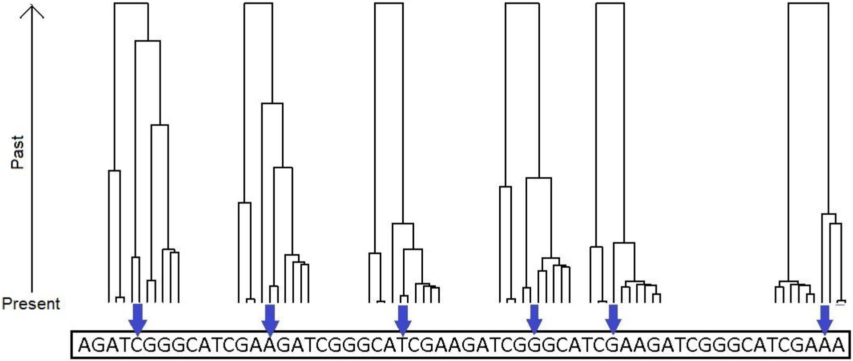
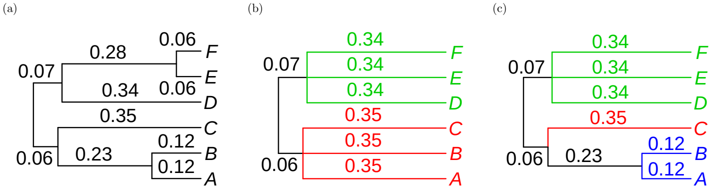


Figure: SNPs along a Chromosome

PHYLOGENETIC METHOD: Step 1

At each SNP,

- Estimate or use the underlying phylogenetic tree, Θ.
- Partition the estimated tree into k clusters using the $(k-1)$ earliest edges.

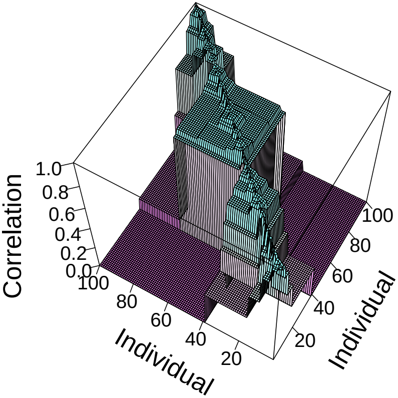


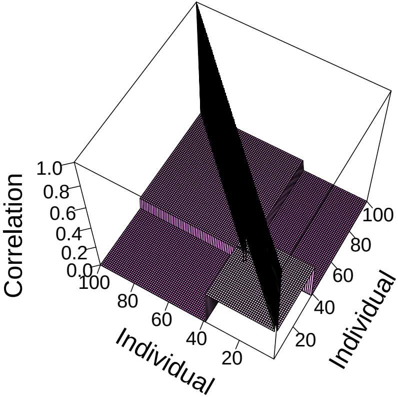
$$\left(\begin{array}{ccccccc} & A & B & C & D & E & F \\ A & 0.41 & 0.29 & 0.06 & 0.00 & 0.00 & 0.00 \\ B & 0.29 & 0.41 & 0.06 & 0.00 & 0.00 & 0.00 \\ C & 0.06 & 0.06 & 0.41 & 0.00 & 0.00 & 0.00 \\ D & 0.00 & 0.00 & 0.00 & 0.41 & 0.07 & 0.07 \\ E & 0.00 & 0.00 & 0.00 & 0.07 & 0.41 & 0.35 \\ F & 0.00 & 0.00 & 0.00 & 0.07 & 0.35 & 0.41 \end{array}\right)$$

$$\left(\begin{array}{ccccccc} & A & B & C & D & E & F \\ A & 0.41 & 0.06 & 0.06 & 0.00 & 0.00 & 0.00 \\ B & 0.06 & 0.41 & 0.06 & 0.00 & 0.00 & 0.00 \\ C & 0.06 & 0.06 & 0.41 & 0.00 & 0.00 & 0.00 \\ D & 0.00 & 0.00 & 0.00 & 0.41 & 0.07 & 0.07 \\ E & 0.00 & 0.00 & 0.00 & 0.07 & 0.41 & 0.07 \\ F & 0.00 & 0.00 & 0.00 & 0.07 & 0.07 & 0.41 \end{array}\right)$$

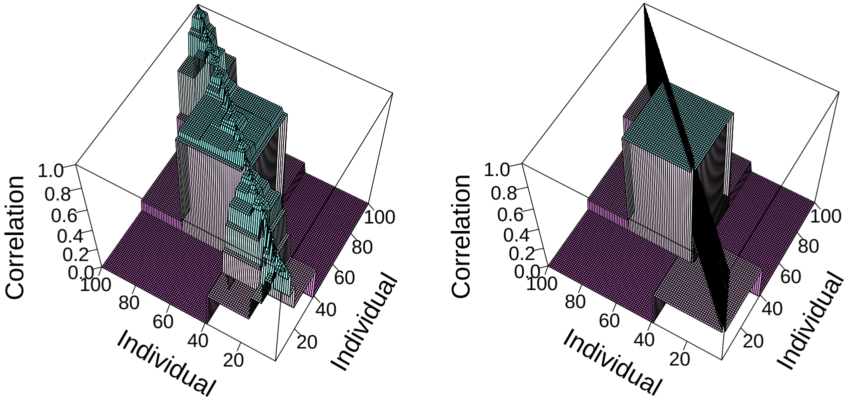
$$\left(\begin{array}{ccccccc} & A & B & C & D & E & F \\ A & 0.41 & 0.29 & 0.06 & 0.00 & 0.00 & 0.00 \\ B & 0.29 & 0.41 & 0.06 & 0.00 & 0.00 & 0.00 \\ C & 0.06 & 0.06 & 0.41 & 0.00 & 0.00 & 0.00 \\ D & 0.00 & 0.00 & 0.00 & 0.41 & 0.07 & 0.07 \\ E & 0.00 & 0.00 & 0.00 & 0.07 & 0.41 & 0.07 \\ F & 0.00 & 0.00 & 0.00 & 0.07 & 0.07 & 0.41 \end{array}\right)$$

PHYLOGENETIC METHOD: Step 1

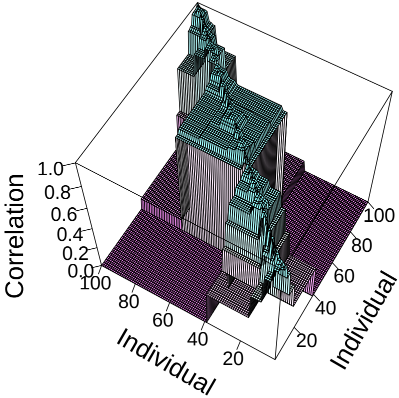


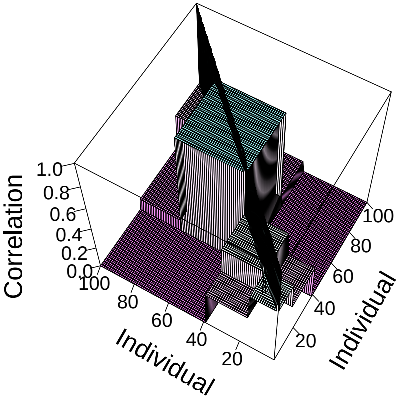


PHYLOGENETIC METHOD: Step 1



PHYLOGENETIC METHOD: Step 1

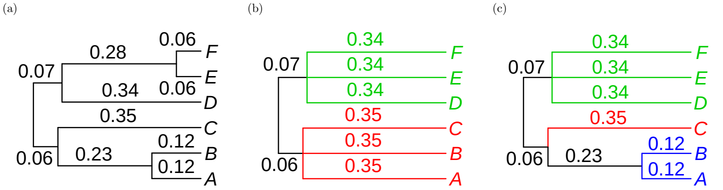




PHYLOGENETIC METHOD: Step 1

At each SNP,

- Estimate or use the underlying phylogenetic tree, Θ.
- Partition the estimated tree into k clusters using the $(k-1)$ earliest edges.



$$\left(\begin{array}{ccccccc} & A & B & C & D & E & F \\ A & 0.41 & 0.29 & 0.06 & 0.00 & 0.00 & 0.00 \\ B & 0.29 & 0.41 & 0.06 & 0.00 & 0.00 & 0.00 \\ C & 0.06 & 0.06 & 0.41 & 0.00 & 0.00 & 0.00 \\ D & 0.00 & 0.00 & 0.00 & 0.41 & 0.07 & 0.07 \\ E & 0.00 & 0.00 & 0.00 & 0.07 & 0.41 & 0.35 \\ F & 0.00 & 0.00 & 0.00 & 0.07 & 0.35 & 0.41 \end{array}\right)$$

$$\left(\begin{array}{ccccccc} & A & B & C & D & E & F \\ A & 0.41 & 0.06 & 0.06 & 0.00 & 0.00 & 0.00 \\ B & 0.06 & 0.41 & 0.06 & 0.00 & 0.00 & 0.00 \\ C & 0.06 & 0.06 & 0.41 & 0.00 & 0.00 & 0.00 \\ D & 0.00 & 0.00 & 0.00 & 0.41 & 0.07 & 0.07 \\ E & 0.00 & 0.00 & 0.00 & 0.07 & 0.41 & 0.07 \\ F & 0.00 & 0.00 & 0.00 & 0.07 & 0.07 & 0.41 \end{array}\right)$$

$$\left(\begin{array}{ccccccc} & A & B & C & D & E & F \\ A & 0.41 & 0.29 & 0.06 & 0.00 & 0.00 & 0.00 \\ B & 0.29 & 0.41 & 0.06 & 0.00 & 0.00 & 0.00 \\ C & 0.06 & 0.06 & 0.41 & 0.00 & 0.00 & 0.00 \\ D & 0.00 & 0.00 & 0.00 & 0.41 & 0.07 & 0.07 \\ E & 0.00 & 0.00 & 0.00 & 0.07 & 0.41 & 0.07 \\ F & 0.00 & 0.00 & 0.00 & 0.07 & 0.07 & 0.41 \end{array}\right)$$

PHYLOGENETIC METHOD: Step 2

For *n* diploid individuals, assume the following model for trait data, $\boldsymbol{Y}_{\boldsymbol{n}\times\mathbf{1}}$:

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ \boldsymbol{Y_g} &\sim N\left(ZD\boldsymbol{\mu}, ZVZ^T\sigma^2\right)\end{aligned}$$

Phylogenetic Tree Parameters:

PHYLOGENETIC METHOD: Step 2

For n diploid individuals, assume the following model for trait data, $\boldsymbol{Y_{n\times 1}}$:

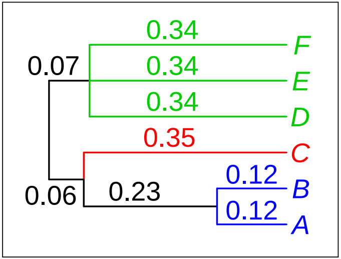
$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ \boldsymbol{Y_g} &\sim N\left(ZD\boldsymbol{\mu}, ZVZ^T\sigma^2\right)\end{aligned}$$

Phylogenetic Tree Parameters:

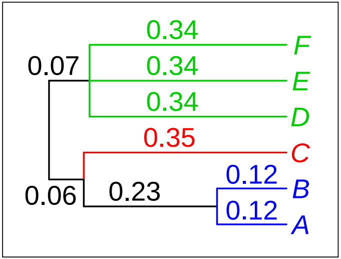
$$\begin{aligned}D(\Theta) &= 2n \times k \text{ matrix with elements:} \\ & \quad D_{ij} = \begin{cases} 1, & \text{if tip } i \text{ is in cluster } j \\ 0, & \text{otherwise} \end{cases} \\ \boldsymbol{\mu}(\Theta) &= (\mu_1, \mu_2, \ldots, \mu_k)^T = \text{ vector of within-cluster trait means} \\ V(\Theta) &= \text{ variance-covariance structure determined by the estimated phylogeny,} \\ & \quad \text{with elements:} V_{ij}(\Theta) = \text{the length of shared time in the evolutionary} \\ & \qquad \text{history of tips } i \text{ and } j\end{aligned}$$

PHYLOGENETIC METHOD: Parameter Example ($k = 3$)

$$\boldsymbol{\mu}(\Theta) = (\mu_1, \mu_2, \mu_3)^T$$



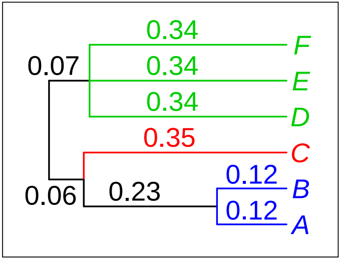
PHYLOGENETIC METHOD: Parameter Example ($k = 3$)



$$\boldsymbol{\mu}(\Theta) = (\mu_1, \mu_2, \mu_3)^T$$

$$D(\Theta) = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{bmatrix} j=1 & j=2 & j=3 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

PHYLOGENETIC METHOD: Parameter Example ($k = 3$)



$$\boldsymbol{\mu}(\Theta) = (\mu_1, \mu_2, \mu_3)^T$$

$$D(\Theta) = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{bmatrix} j=1 & j=2 & j=3 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$V(\Theta) = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{bmatrix} A & B & C & D & E & F \\ 0.41 & 0.29 & 0.06 & 0 & 0 & 0 \\ 0.29 & 0.41 & 0.06 & 0 & 0 & 0 \\ 0.06 & 0.06 & 0.41 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.41 & 0.07 & 0.07 \\ 0 & 0 & 0 & 0.07 & 0.07 & 0.41 \\ 0 & 0 & 0 & 0.07 & 0.41 & 0.07 \end{bmatrix}$$

PHYLOGENETIC METHOD: Step 2

For n diploid individuals, assume the following model for trait data, $\boldsymbol{Y}_{\boldsymbol{n}\times\mathbf{1}}$:

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ \boldsymbol{Y_g} &\sim N\left(ZD\boldsymbol{\mu}, ZVZ^T\sigma^2\right)\end{aligned}$$

Phylogenetic Tree Parameters:

$$\begin{aligned}D(\Theta) &= 2n \times k \text{ matrix with elements:} \\ & \quad D_{ij} = \begin{cases} 1, & \text{if tip } i \text{ is in cluster } j \\ 0, & \text{otherwise} \end{cases} \\ \boldsymbol{\mu}(\Theta) &= (\mu_1, \mu_2, \ldots, \mu_k)^T = \text{ vector of within-cluster trait means} \\ V(\Theta) &= \text{ variance-covariance structure determined by the estimated phylogeny,} \\ & \quad \text{with elements:} V_{ij}(\Theta) = \text{the length of shared time in the evolutionary} \\ & \quad \quad \text{history of tips } i \text{ and } j\end{aligned}$$

PHYLOGENETIC METHOD: Step 2

For n diploid individuals, assume the following model for trait data ($\boldsymbol{Y}_{\boldsymbol{n}\times\mathbf{1}}$):

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ \boldsymbol{Y_g} &\sim N\left(ZD\boldsymbol{\mu}, ZVZ^T\sigma^2\right) \end{aligned}$$

Genetic Component Notation/Parameters:

$$\begin{aligned} Z_{n\times 2n} &= n \times 2n \text{ matrix that maps each tip to diploid individual} \\ &\quad\ \text{assuming each chromosome contributes equally to } \boldsymbol{Y_g} \\ \sigma^2 &= \text{variance due to genetic component of trait} \end{aligned}$$

PHYLOGENETIC METHOD: Step 2

For n diploid individuals, assume the following model for trait data ($\boldsymbol{Y_{n \times 1}}$):

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\
\boldsymbol{Y_g} &\sim N\left(ZD\boldsymbol{\mu}, ZVZ^T\sigma^2\right) \\
\boldsymbol{Y_e} &\sim N\left(X\boldsymbol{\beta}, I\nu^2\right)
\end{aligned}$$

PHYLOGENETIC METHOD: Step 2

For n diploid individuals, assume the following model for trait data ($\boldsymbol{Y}_{\boldsymbol{n}\times\mathbf{1}}$):

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e}\\ \boldsymbol{Y_g} &\sim N\left(ZD\boldsymbol{\mu}, ZVZ^T\sigma^2\right)\\ \boldsymbol{Y_e} &\sim N\left(X\boldsymbol{\beta}, I\nu^2\right)\end{aligned}$$

Environmental Parameters:

$$\boldsymbol{X}^T = \begin{bmatrix}1 & 1 & \dots & 1\\ X_1 & X_2 & \dots & X_n\end{bmatrix}$$

where $X_i =$ value of the covariate for the i^{th} observation

$$\boldsymbol{\beta} = (\beta_0, \beta_1)^T = \text{ vector of regression coefficients}$$

$$\nu^2 = \text{ variance due to environmental component of trait}$$

PHYLOGENETIC METHOD: Step 3

Estimation in a Bayesian Framework

The Likelihood:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ \text{where } \boldsymbol{Y} | \boldsymbol{Y_g}, \boldsymbol{\mu}, \boldsymbol{\beta}, \nu^2, \sigma^2 &\sim N(\boldsymbol{X\beta} + \boldsymbol{Y_g}, \nu^2 \boldsymbol{I}) \text{ and} \\ \boldsymbol{Y_g} | \boldsymbol{\mu}, \boldsymbol{\beta}, \nu^2, \sigma^2 &\sim N(\boldsymbol{ZD\mu}, \sigma^2 \boldsymbol{ZVZ^T}) \end{aligned}$$

Prior Distributions:

- $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, u^2 \boldsymbol{I})$
- $\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_0, w^2 \boldsymbol{I})$
- $\sigma^2 \sim$ Inverse Gamma $(Shape = a, Scale = b)$
- $\nu^2 \sim$ Inverse Gamma $(Shape = c, Scale = d)$
- and we assume all parameters are independent!

Note: The **conditional posterior distributions** have closed forms!

PHYLOGENETIC METHOD

Advantages of the phylogenetic method:

- Allow for and uses covariance among the observations
- Clustering uses the broad-scale evolutionary relationships to remain computationally feasible
- Bayesian framework produces posterior means for estimates

Ways to Assess Performance of Phylogenetic Method:

- Simulation Study
- Real Data Analysis

METHODS: Data Simulation

Data needed for a simulation study include:

1. SNP data
2. Covariate data
3. Quantitative trait data that has
    a. a genetic component (related to a single SNP)
    b. an environmental component (related to an external covariate)

DATA SIMULATION METHOD

1. **Simulate SNP data.**

DATA SIMULATION METHOD

1. **Simulate SNP data.**
2. **Simulate covariate values** (X) for each diploid individual uniformly from a specified range.

DATA SIMULATION METHOD

1. **Simulate SNP data.**
2. **Simulate covariate values** (X) for each diploid individual uniformly from a specified range.
3. **Simulate quantitative trait data:** For some $\rho \in [0, 1]$, let

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ &= \rho \boldsymbol{T_g} + (1 - \rho) \boldsymbol{T_e}, \end{aligned}$$

DATA SIMULATION METHOD

1. **Simulate SNP data.**
2. **Simulate covariate values** (X) for each diploid individual uniformly from a specified range.
3. **Simulate quantitative trait data:** For some $\rho \in [0, 1]$, let

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ &= \rho \boldsymbol{T_g} + (1 - \rho)\boldsymbol{T_e}, \text{ where} \end{aligned}$$

   a. $\boldsymbol{T_g}$**: Genetic Component**
      - Simulate data along the “disease” tree using a two-target Ornstein-Uhlenbeck process.

DATA SIMULATION METHOD

1. **Simulate SNP data.**
2. **Simulate covariate values** (X) for each diploid individual uniformly from a specified range.
3. **Simulate quantitative trait data:** For some $\rho \in [0, 1]$, let

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ &= \rho \boldsymbol{T_g} + (1-\rho)\boldsymbol{T_e}, \text{ where} \end{aligned}$$

   a. $\boldsymbol{T_g}$**: Genetic Component**
      - Simulate data along the "disease" tree using a two-target Ornstein-Uhlenbeck process.
      - Randomly pair the tips to create individuals.

DATA SIMULATION METHOD

1. **Simulate SNP data.**
2. **Simulate covariate values** (X) for each diploid individual uniformly from a specified range.
3. **Simulate quantitative trait data:** For some $\rho \in [0, 1]$, let

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ &= \rho \boldsymbol{T_g} + (1 - \rho) \boldsymbol{T_e}, \text{ where} \end{aligned}$$

   a. $\boldsymbol{T_g}$**: Genetic Component**
      - Simulate data along the “disease” tree using a two-target Ornstein-Uhlenbeck process.
      - Randomly pair the tips to create individuals.
      - Average the trait across SNP copies to find $\boldsymbol{T_g}$.

DATA SIMULATION METHOD

1. **Simulate SNP data.**
2. **Simulate covariate values** (X) for each diploid individual uniformly from a specified range.
3. **Simulate quantitative trait data:** For some $\rho \in [0, 1]$, let

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{Y_g} + \boldsymbol{Y_e} \\ &= \rho \boldsymbol{T_g} + (1 - \rho) \boldsymbol{T_e}, \text{ where} \end{aligned}$$

   a. $\boldsymbol{T_g}$: **Genetic Component**
      - Simulate data along the “disease” tree using a two-target Ornstein-Uhlenbeck process.
      - Randomly pair the tips to create individuals.
      - Average the trait across SNP copies to find $\boldsymbol{T_g}$.

   b. $\boldsymbol{T_e}$: **Environmental Component**
      - $\boldsymbol{T_e} \sim N(X\boldsymbol{\eta}, \tau^2 I)$, where $\boldsymbol{\eta}$ and τ^2 are fixed.

SIMULATION STUDY: Data Simulation

Data Simulation Process:

1. Simulate chromosomes in an ARG framework (SNP data).
2. Simulate covariate data.
3. Simulate quantitative trait data.
    - Simulate the genetic and environmental components of the trait.
    - Take a weighted average of these components to find the quantitative trait value.

Simulated Data:

- A matrix of SNP values at tips of phylogenies
- A vector of covariate values for each diploid individual.
- A vector of quantitative trait values, $\boldsymbol{Y}$, for each diploid individual.

DATA SIMULATION METHOD

In the interim steps, the data looks like this:

i	Person	SNP Data	Covariate	T_g	T_e	Y $(\rho = 0.5)$
1	1	$1 \ldots 000 \ldots 0$	25.2	180.35	173.44	176.90
2	2	$1 \ldots 010 \ldots 0$	32.3	184.65	191.95	188.30
3	1	$1 \ldots 101 \ldots 1$	25.2	180.35	173.44	176.90
4	3	$1 \ldots 011 \ldots 0$	29.5	182.45	179.19	180.82
5	2	$0 \ldots 100 \ldots 1$	32.3	184.65	191.95	188.30
6	3	$0 \ldots 001 \ldots 1$	29.5	182.45	179.19	180.82

DATA SIMULATION METHOD

The *observed* data is:

i	Person	SNP Data	Covariate	$\boldsymbol{Y}$ $(\rho = 0.5)$
1	1	$1\ldots000\ldots0$	25.2	176.90
2	2	$1\ldots010\ldots0$	32.3	188.30
3	1	$1\ldots101\ldots1$	25.2	176.90
4	3	$1\ldots011\ldots0$	29.5	180.82
5	2	$0\ldots100\ldots1$	32.3	188.30
6	3	$0\ldots001\ldots1$	29.5	180.82

SIMULATION PARAMETERS

Parameters for Genetic Component of Trait:

- $Y_i(0) = 90$
- $\delta_1 = 80, \delta_2 = 100$
- $\alpha = 10$
- $\sigma_Y = 20$

Parameters for Environmental Component of Trait:

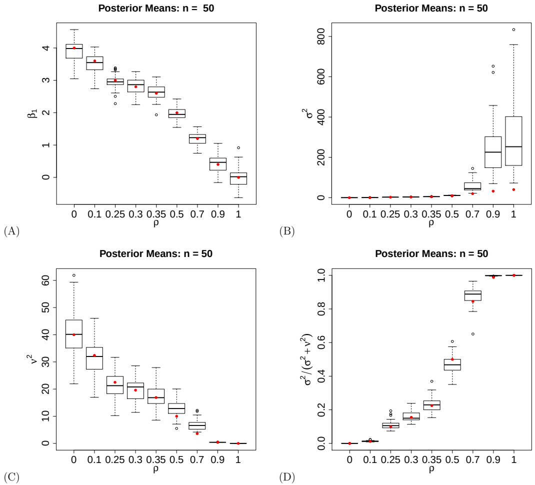
- $r = 1$ covariate
- X_i are independent draws from a Uniform$(25, 35)$ distribution
- $\boldsymbol{\eta} = (\eta_0, \eta_1)^T = (10, 2.5)^T$
- $\tau = 15$
- ρ: varied

SIMULATION STUDY: DATA ANALYSIS PROCESS

For each simulated data set:

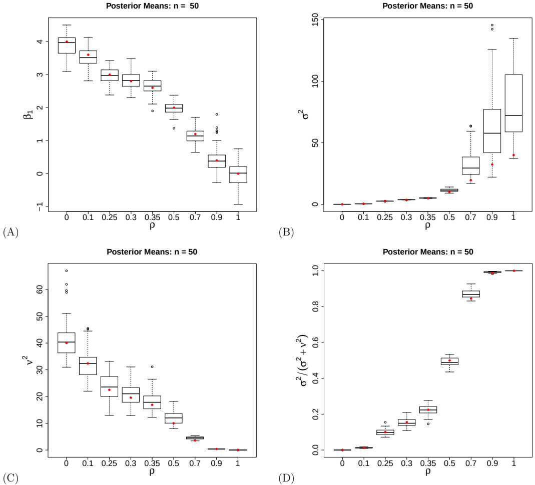
- Using the phylogenetic tree, trait, and covariate data, estimate parameters using the posterior means from the Gibbs sampler.
- *Note:* True trees are used in this simulation study and the number of clusters is set to $k = 5$.

RESULTS: Known Phylogenies, Informative Priors



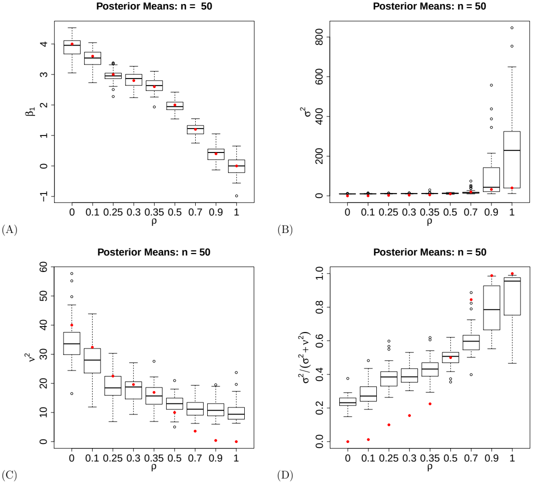
(Figure 3; Thompson et al. 2016)

RESULTS: Estimated Phylogenies, Informative Priors



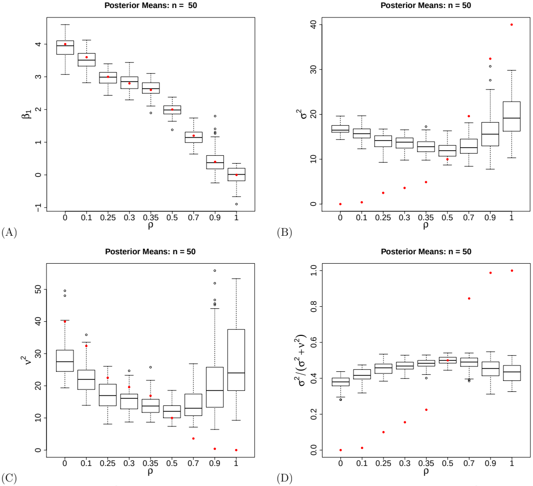
(Figure 4; Thompson et al. 2016)

RESULTS: Known Phylogenies, Vague Priors



(Figure 5; Thompson et al. 2016)

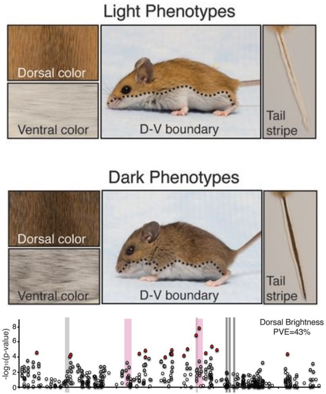
RESULTS: Estimated Phylogenies, Vague Priors



(Figure 6; Thompson et al. 2016)

RESULTS: Real Data Analysis

- *Organisms:* 91 wild-caught mice
- *Genetic Data:* SNP data
- *Quantitative Traits:* nine quantitative color phenotypes
- *Covariates:* Include weight, body length, tail length
- Goal: To identify regions of the genome contributing to mouse coat color after accounting for population structure covariates.
- Previous work showed that much of coat color variation appears to be controlled by a single gene, *Agouti*.



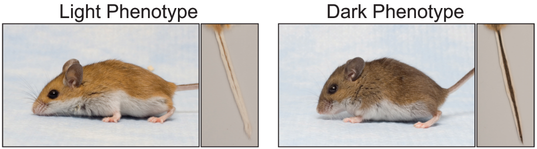
(Linnen et al. 2013)

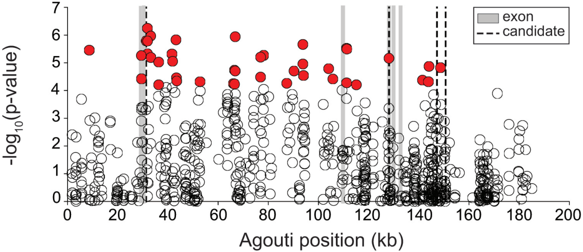
REAL DATA ANALYSIS ALGORITHM

For the real data set:

- Computationally phase the data using Beagle.
- At each SNP, estimate the phylogenetic tree using Blossoc and branch lengths using approximate MLEs.
- Using the phylogenetic tree, trait, and covariate data, estimate the parameters using the posterior means from the Gibbs sampler.
- *Note:* Estimated trees are used in this simulation study and the number of clusters is set to $k = 5$.

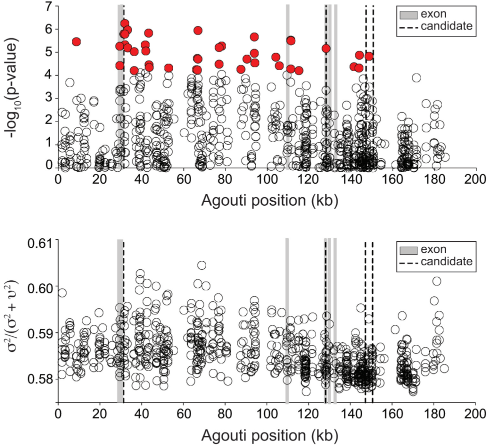
RESULTS: Real Data Analysis





(Figure 8; Thompson et al. 2016)

RESULTS: Real Data Analysis



(Figure 8; Thompson et al. 2016)

CONCLUSIONS

- Posterior means provide good estimates of environmental parameters, even when two SNPs are considered.
- Using an evolutionary framework to approach problem is more realistic than non-tree based approximations.
- Use of the broad-scale evolutionary relationships among SNPs makes the technique computationally feasible.
- This model allows for analysis on a per-individual basis while preserving per-chromosomal estimation of evolutionary history at each SNP.

FUTURE DIRECTIONS

Related problems of interest:

- the analysis of related genetic and environmental components
- the study of multivariate traits (multiple traits affected by one SNP)
- developing a way to control for other associated SNPs present in the genetic data
- the analysis of data with population structure

Thank You!

Questions?

References:

- Thompson, K.L., C.R. Linnen, and L. Kubatko. 2016. Tree-based quantitative trait mapping the the presence of external covariates. *Statistical Applications in Genetics and Molecular Biology*, 15: 473-490.
- Linnen, C. R. *et al.* 2013. Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. *Science*, 339(6125):1312–6.
- Zhang W. *et al.* 2012. Genome-wide association mapping of quantitative traits in outbred mice. *G3(Bethesda)* 2(2):167–174.

Supplemental Results

Modeling External Covariates

DATA ANALYSIS ALGORITHM

LSS-C/LSS-I: At each SNP site, do the following.

- Estimate the marginal tree topology and branch lengths.
- Calculate LSS-C/LSS-I using the estimated phylogeny, covariate data, and trait data.

Previous Methods: At each SNP site, calculate the Likelihood Ratio Test Statistic and the p-value from `SNPassoc`.

Data Analysis for Each Method:

- Detection of SNP/Covariate Analysis: Use permutation testing to check if any SNP along the chromosome or the covariate is detected.
  - Permute the trait values across the tips of the estimated phylogeny.
  - Recalculate each statistic using the permuted data.
- Localization Analysis: Record distance (in base pairs) between the most maximally-scored SNP and the associated SNP.

RESULTS: Detection and Localization

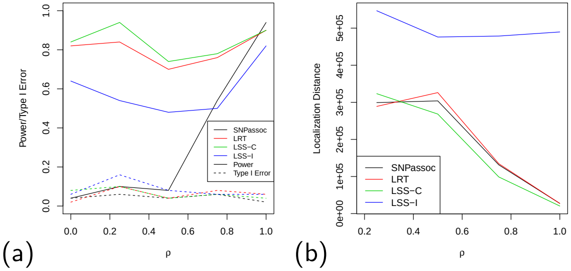


Figure: Power and localization in covariate analysis

RESULTS: Example Replication

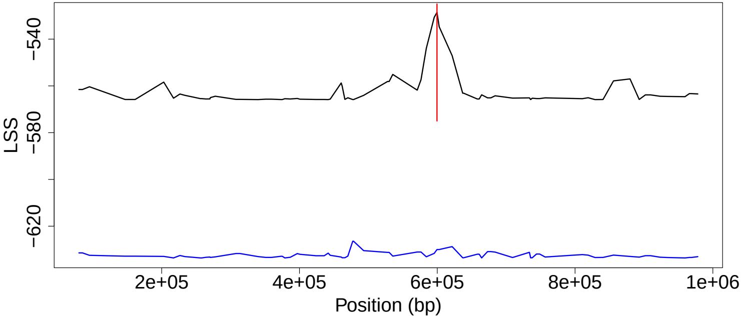


Figure: Example of behavior of LSS-C across a chromosome

Legend:

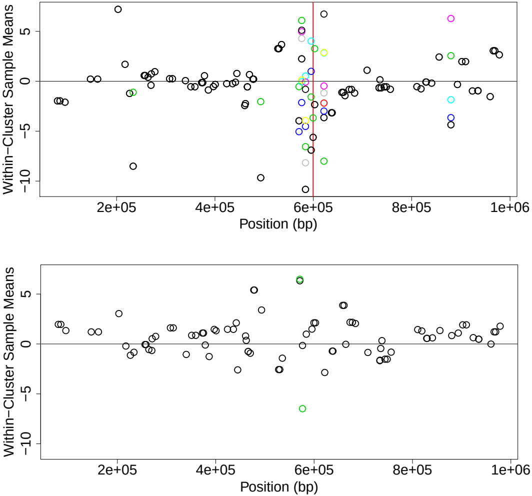
– Truly-associated SNP (located at red line) and related environmental covariate present

– No associated SNP nor related environmental covariate present

RESULTS: Example Replication

Figure: Behavior of within-cluster mean estimates along a chromosome (using the chromosomal model)

- (a) Truly-associated SNP and related covariate present
- (b) Neither a truly-associated SNP nor a related covariate present



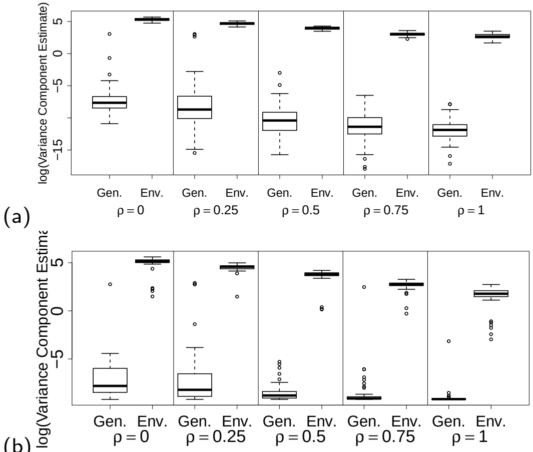
(a)

(b)

RESULTS: Estimates at the Maximally-Scored SNP

Figure: Estimates of genetic and environmental variances at the maximally-scored SNP

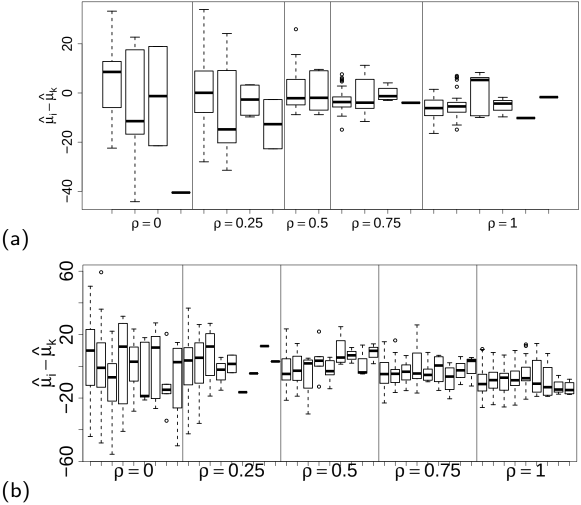
1. LSS-C
2. LSS-I



RESULTS: Estimates at the Maximally-Scored SNP

Figure: Estimated differences in cluster means at the maximally-scored SNP

- (a) LSS-C
- (b) LSS-I

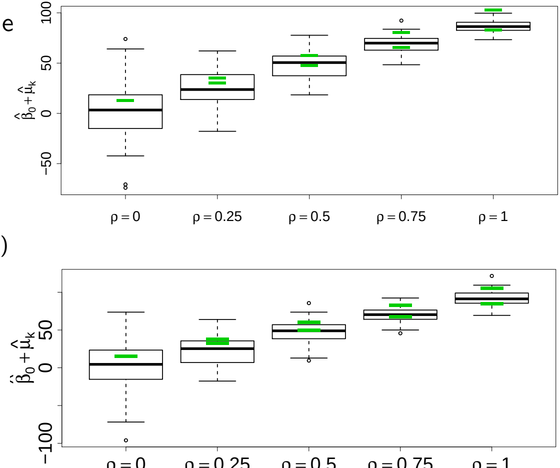


RESULTS: Estimates at the Maximally-Scored SNP

Figure: Estimates of $\beta_0 + \mu_k$ at the maximally-scored SNP

(a) LSS-C

(b) LSS-I



(a)

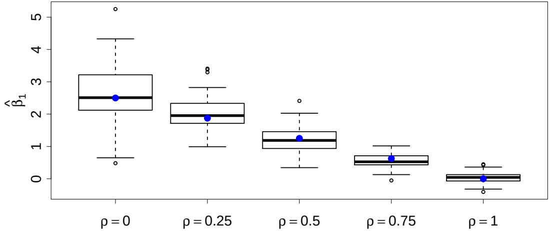
(b)

RESULTS: Estimates at the Maximally-Scored SNP

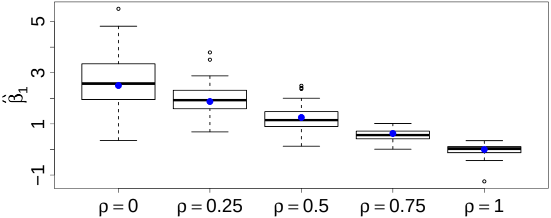
Figure: Estimates of β_1 at the maximally-scored SNP

(a) LSS-C

(b) LSS-I



(a)



(b)

RESULTS: Example Replication

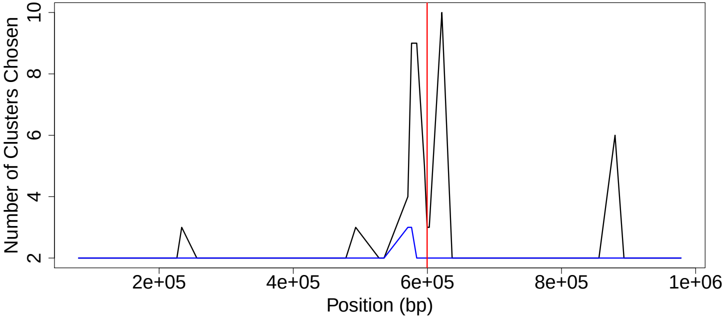


Figure: Example of number of clusters chosen by LSS across a chromosome

Legend:

– Truly-associated SNP (located at red line) and related environmental covariate present

– No associated SNP nor related environmental covariate present

RESULTS: Example Replication

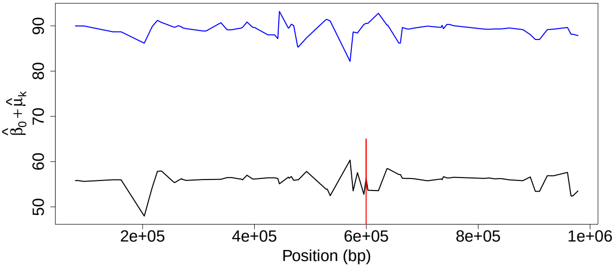


Figure: Example of behavior of baseline estimate across a chromosome.

Legend:

– Truly-associated SNP (located at red line) and related environmental covariate present

– No associated SNP nor related environmental covariate present

RESULTS: Example Replication

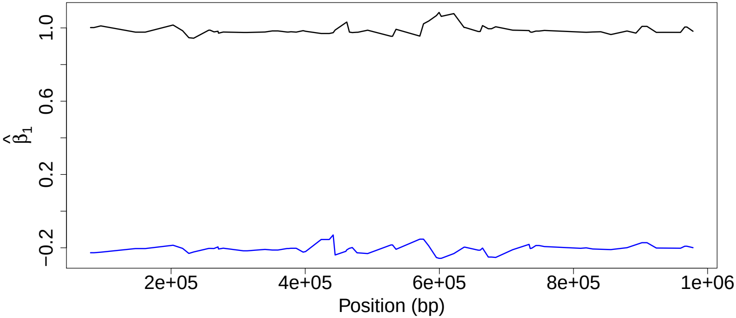


Figure: Example of behavior of estimation of β_1 across a chromosome

Legend:

– Truly-associated SNP (located at red line) and related environmental covariate present

– No associated SNP nor related environmental covariate present

RESULTS: Adjusting for External Covariates

ρ	$\tau = 5$		$\tau = 15$	
	SNPassoc		LSS	
0.00	0.00	0.04	0.02	0.06
0.25	0.06	0.06	0.04	0.06
0.50	0.02	0.04	0.10	0.04
0.75	0.06	0.06	0.04	0.00
1.00	0.06	0.02	0.06	0.04

Table: Type I error of LSS and SNPassoc when adjusting for environmental covariates

RESULTS: Adjusting for External Covariates

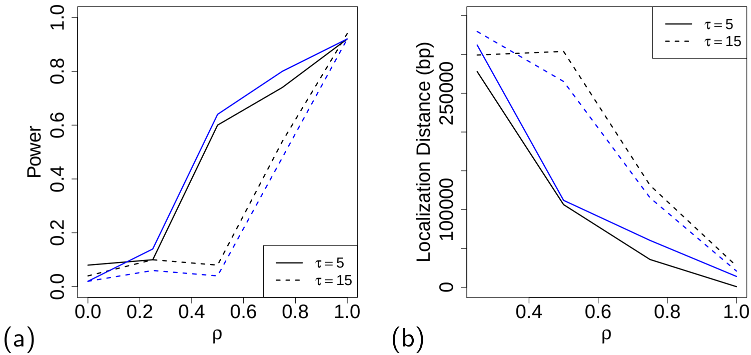


Figure: Power and localization when adjusting for covariates

Statistics: `SNPassoc`, LSS

RESULTS: Adjusting for External Covariates

ρ	$\tau = 5$		$\tau = 15$	
	SNPassoc	LSS	SNPassoc	LSS
0.00	0.08	0.02	0.04	0.02
0.25	0.10	0.14	0.10	0.06
0.50	0.60	0.64	0.08	0.04
0.75	0.74	0.80	0.54	0.48
1.00	0.92	0.92	0.94	0.92

Table: Power of Detection of LSS and SNPassoc when adjusting for environmental covariates

RESULTS: Adjusting for External Covariates

ρ	$\tau = 5$		$\tau = 15$	
	SNPassoc	LSS	SNPassoc	LSS
0.25	277926	312144	299226	329554
0.50	106454	111968	303868	265114
0.75	35738	60362	131350	115288
1.00	734	13926	27360	20830

Table: Average localization distance (bp) of LSS and SNPassoc when adjusting for environmental covariates