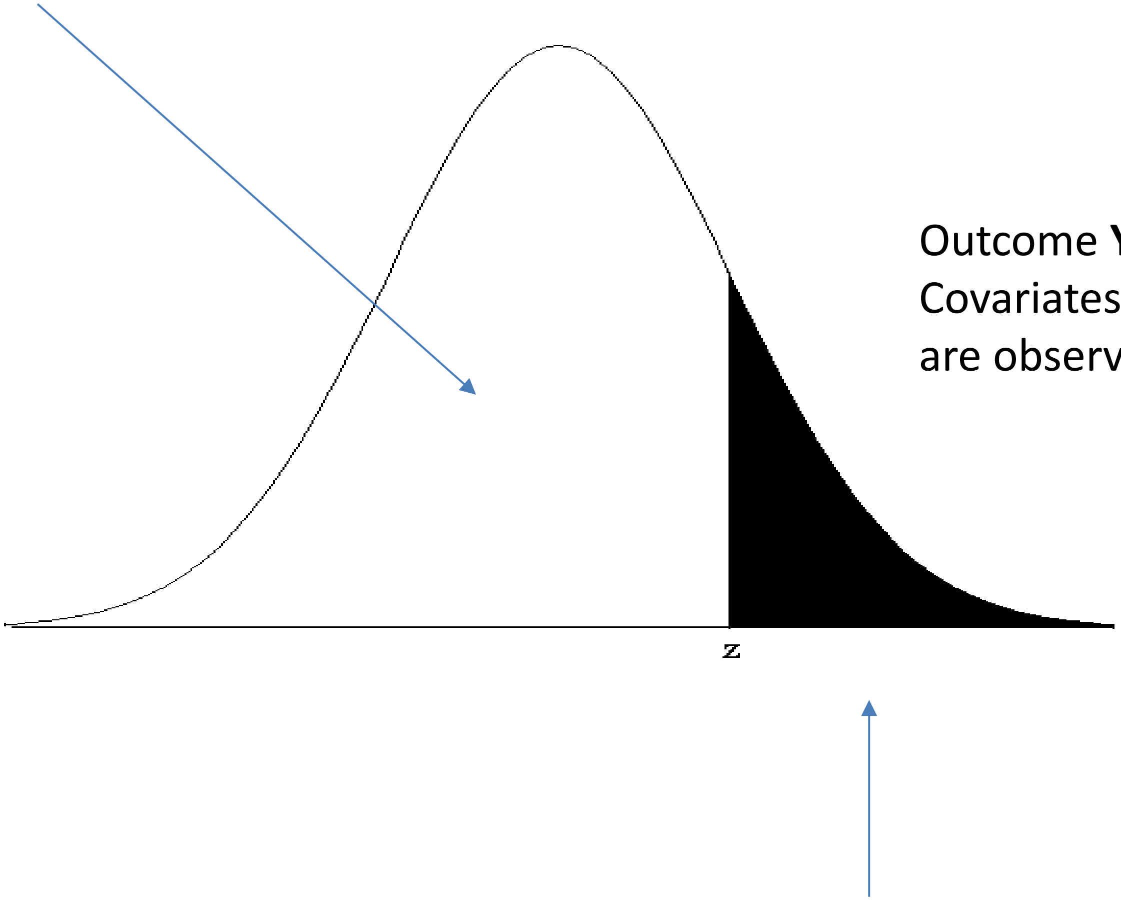
Mapping genes from a single tail sample of the phenotype distribution by generating pseudo observations

Mikko J. Sillanpää
Research Unit of Mathematical Sciences
University of Oulu

I am talking about following work:

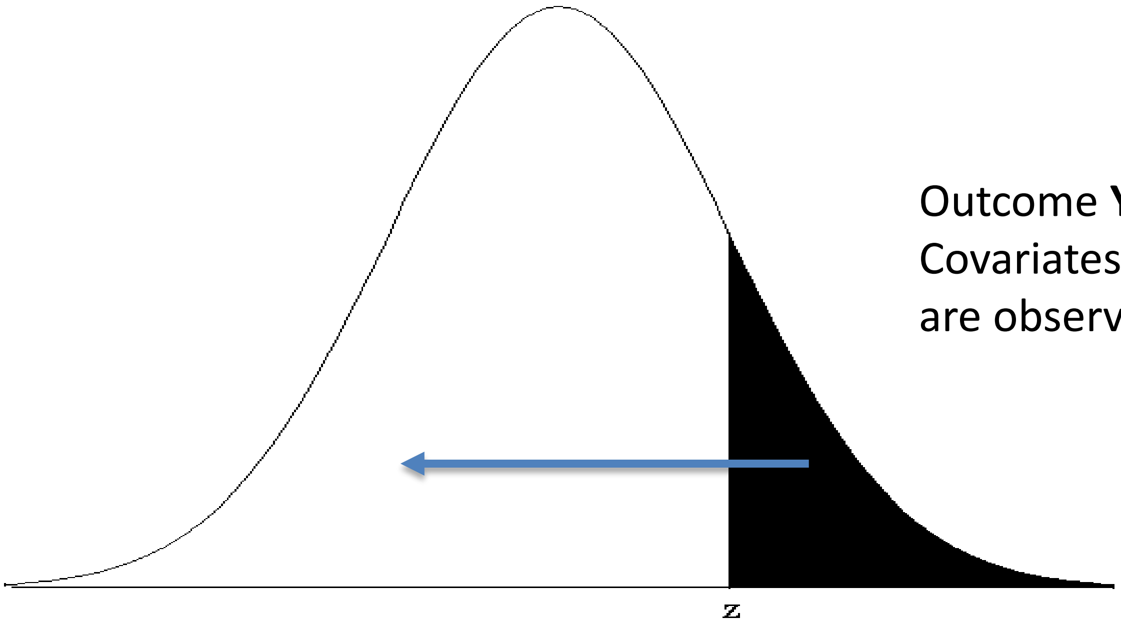
- Sillanpää MJ, Hoti F (2007) Mapping quantitative trait loci from a single-tail sample of the phenotype distribution including survival data. *Genetics* **177**: 2361-2377.

This part is systematically missing (MNAR)



Outcome **Y** (*continuous*)
Covariates **X** (*discrete*)
are observed

cases only

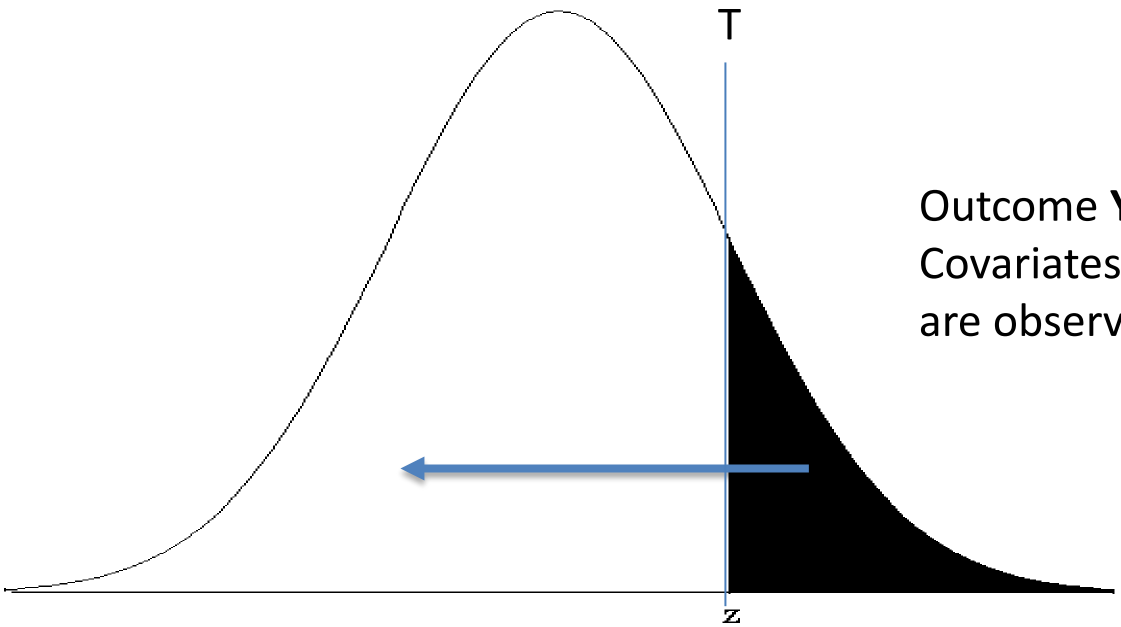


Outcome **Y** (*continuous*)
Covariates **X** (*discrete*)
are observed

1. We create maximally dissimilar antisib (, **X**) for each individual

2. Conditionally on **X**, we create **Y** using data-augmentation

$y_i^{h} \le T \le y_i^{o}$ for all pairs i.



Outcome **Y** (*continuous*)
Covariates **X** (*discrete*)
are observed

1. We create maximally dissimilar antisib (, **X**) for each individual

2. Conditionally on **X**, we create **Y** using data-augmentation

Human data:

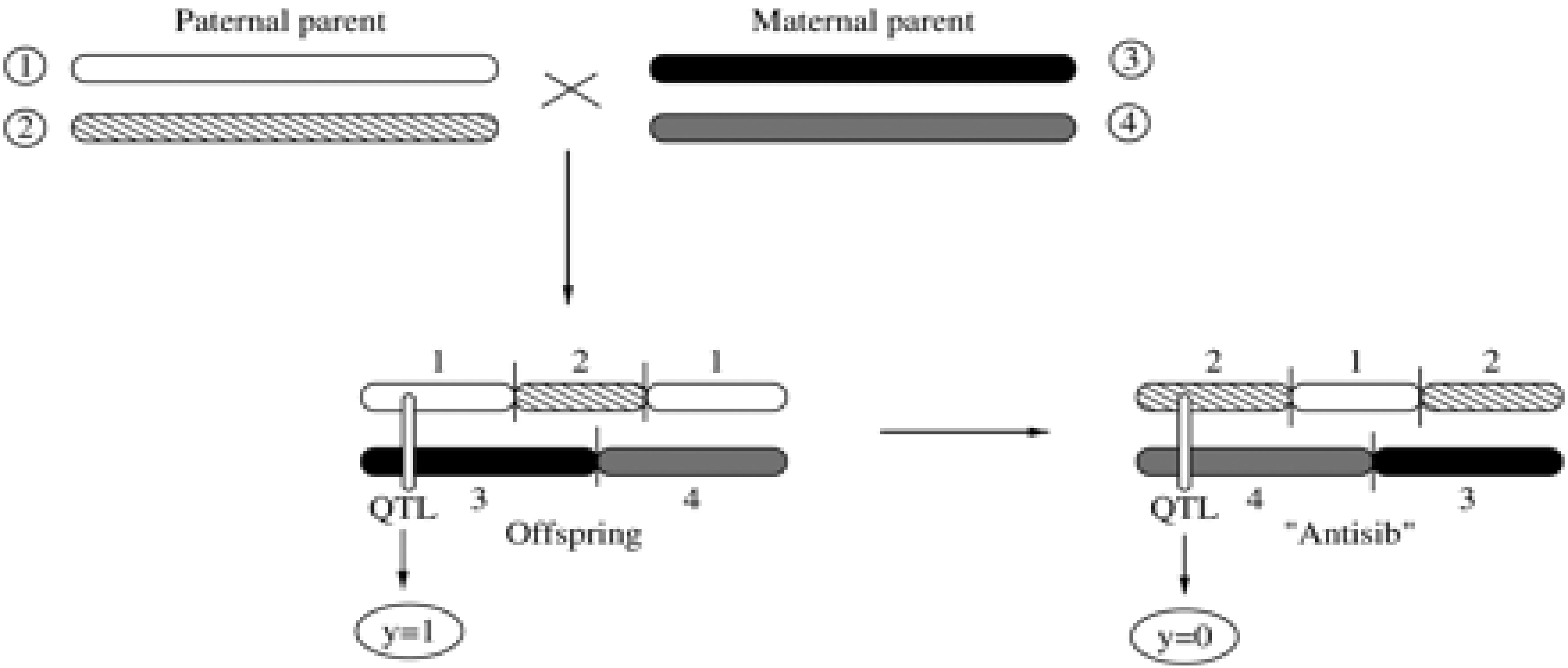
Collect cancer patients with high disease severity score and their parents

Info from parents -> create pseudo observations

Plant data:

A single family can have 300-500 offspring

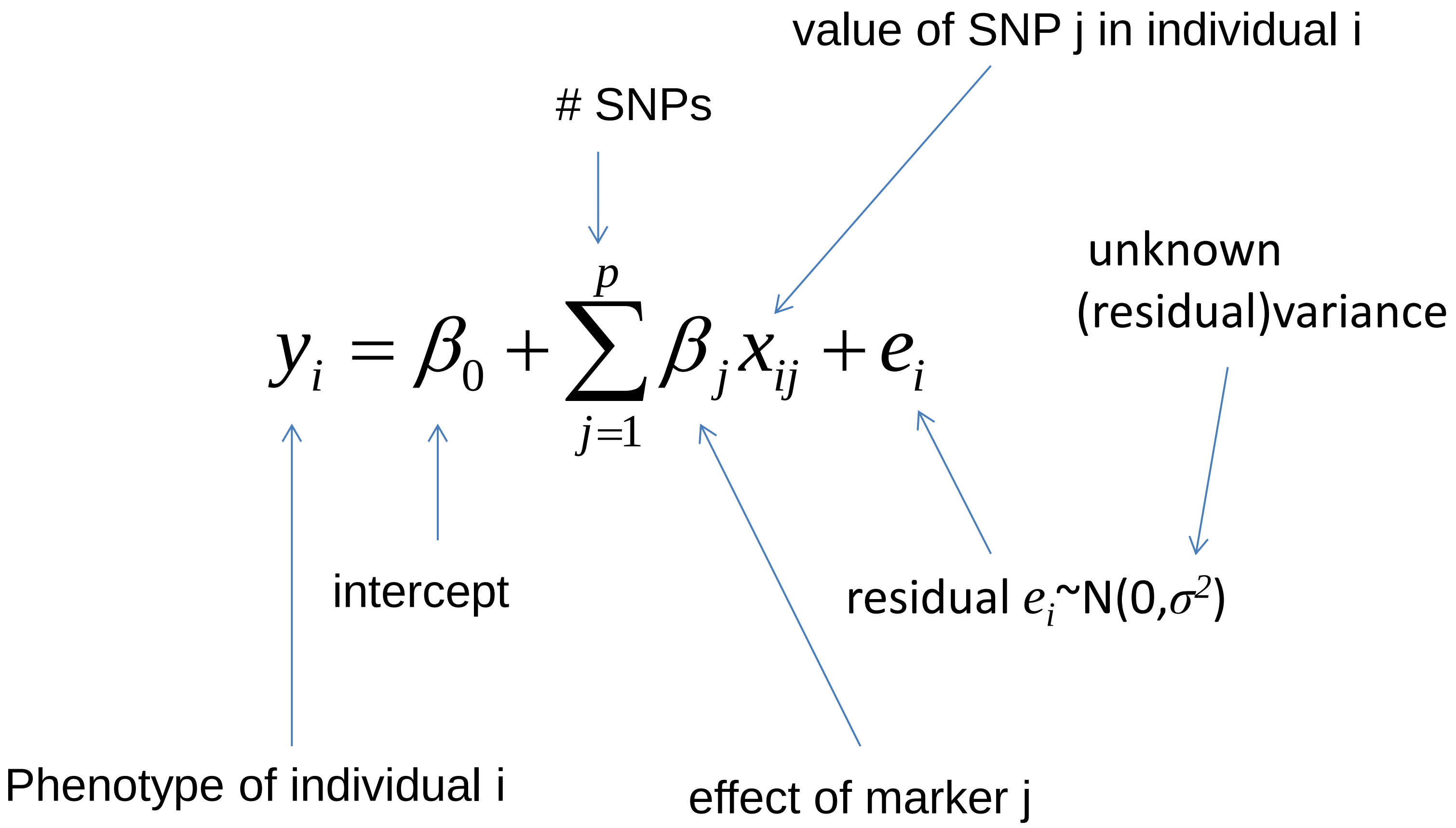
-> easy to obtain parental data


Paternal parent
Maternal parent
①
②
③
④
1
2
1
QTL
3
4
Offspring
y=1
2
1
2
QTL
4
3
"Antisib"
y=0

If parental mating type AA x AB:

Observed sib		Anti-sib
AA	->	AB
AB	->	AA

Regression model



value of SNP j in individual i

SNPs

unknown (residual)variance

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i$$

intercept

residual e_i~N(0,σ^2)

Phenotype of individual i

effect of marker j

To do variable selection, we set shrinkage-inducing priors for regression coefficients

Prior shrinks un-important coefficients to zero

Automatic relevance determination prior – No tuning parameters!

MCMC estimation of model parameters

- Given parents, anti-sib covariates **X** are created only once before running MCMC.

- (1) Conditionally on **X** and model parameters, **Y** is created using data-augmentation at every MCMC round.

- (2) Conditionally on (**X**, **Y**), sample new values for model parameters.

- (3) Go back to 1 until convergence

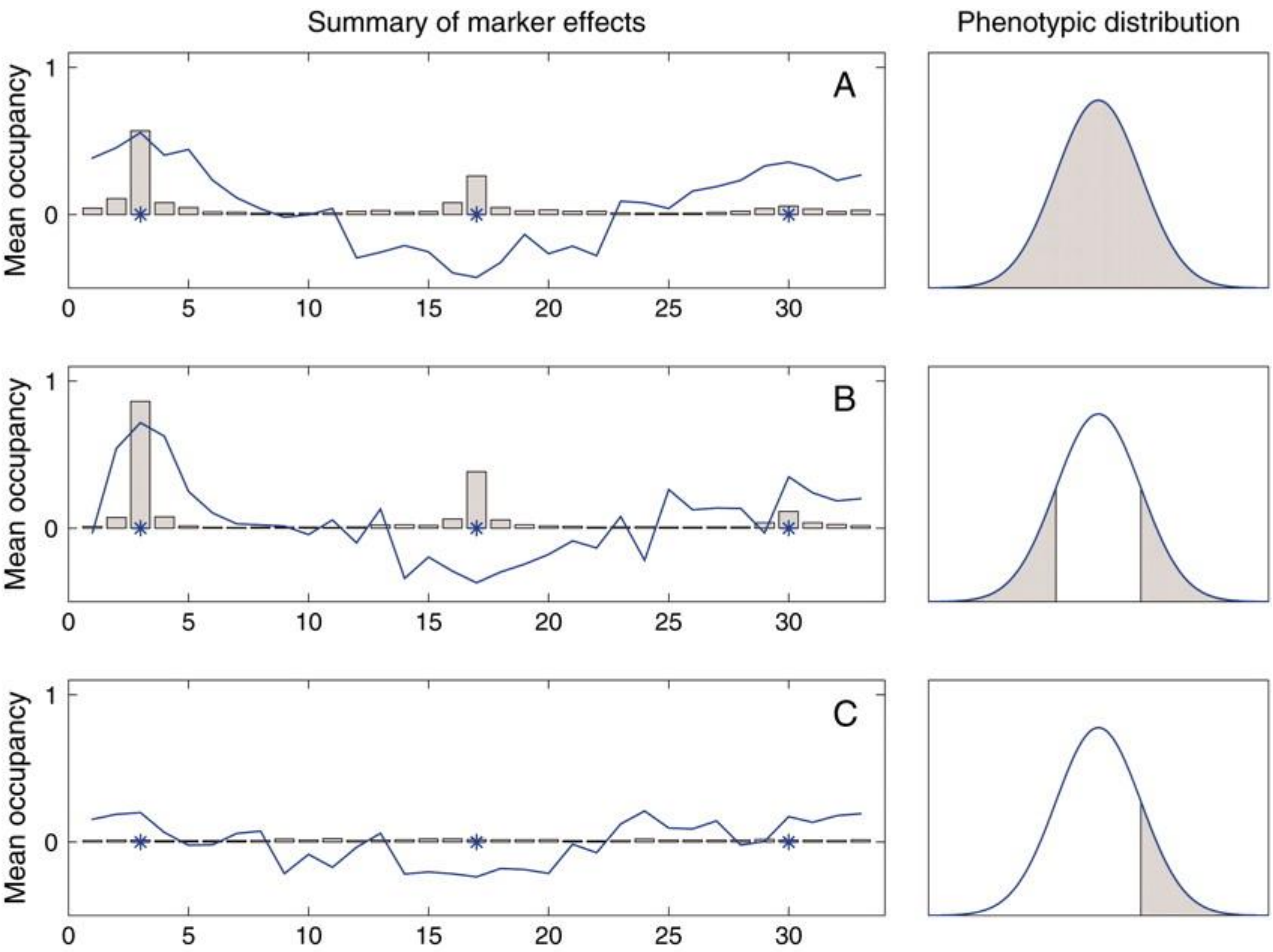
Bayesian variable selection

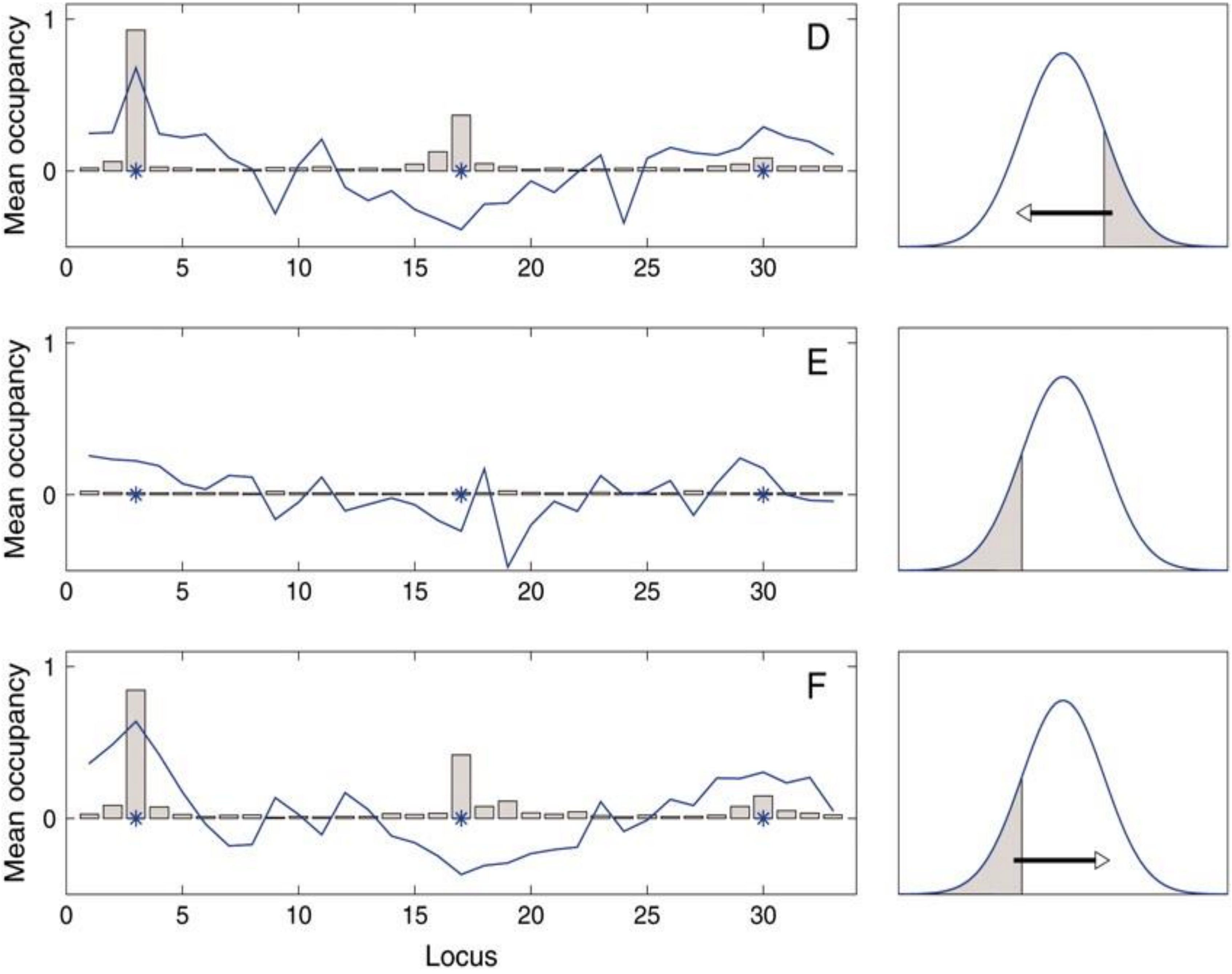
100 replicated data sets

ORIGINAL DATA: 250 backcross individuals
with parental mating type AB x AA

SAMPLING: 40 RANDOM INDIVIDUALS

1. Whole distribution
2. Both tails
3. Right tail (analysis assuming normality)
4. Right tail (analysis using pseudo-observations)
5. Left tail (analysis assuming normality)
6. Left tail (analysis using pseudo-observations)


Summary of marker effects
Phenotypic distribution
Mean occupancy
A
B
C
1
0
0
5
10
15
20
25
30


D
E
F
Mean occupancy
1
0
0
5
10
15
20
25
30
Locus

Conclusion

Missing Not At Random (MNAR):

”Probability of missing data is systematically related to the hypothetical values that are missing”

- Parental info -> create pseudo observations
- You got positions of signals right
- Effect sizes are over-estimated