Linear models III: Improved linear mixed models

Christoph Lippert[1] **Oliver Stegle**[2]

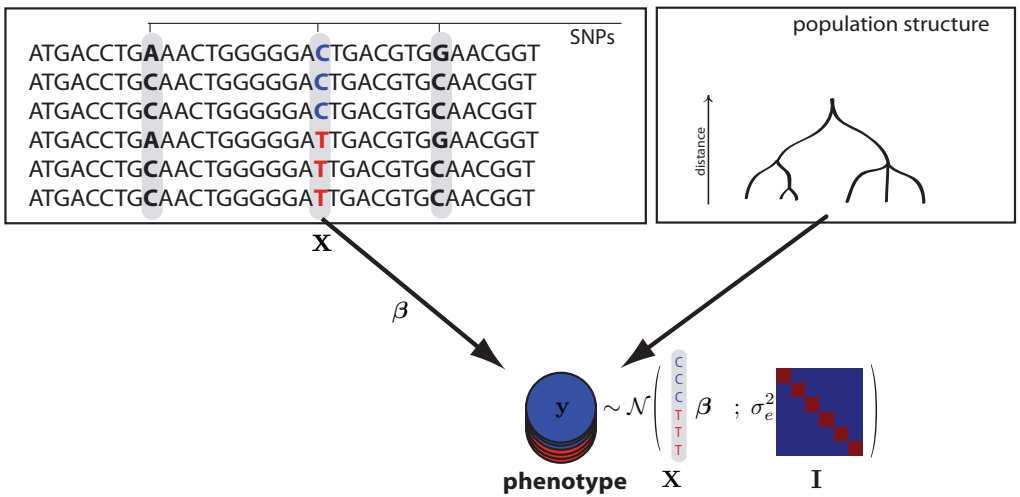
[1] Microsoft Research, Los Angeles, USA
[2] Max-Planck-Institutes Tübingen, Germany

Basel
09. September 2012

Microsoft Research

Correcting for population stratification using linear mixed models

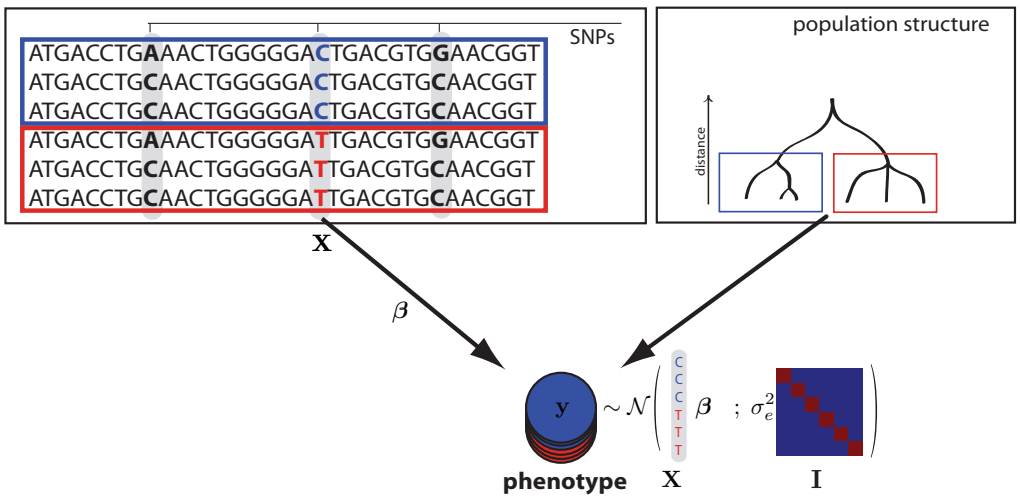
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs*



[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

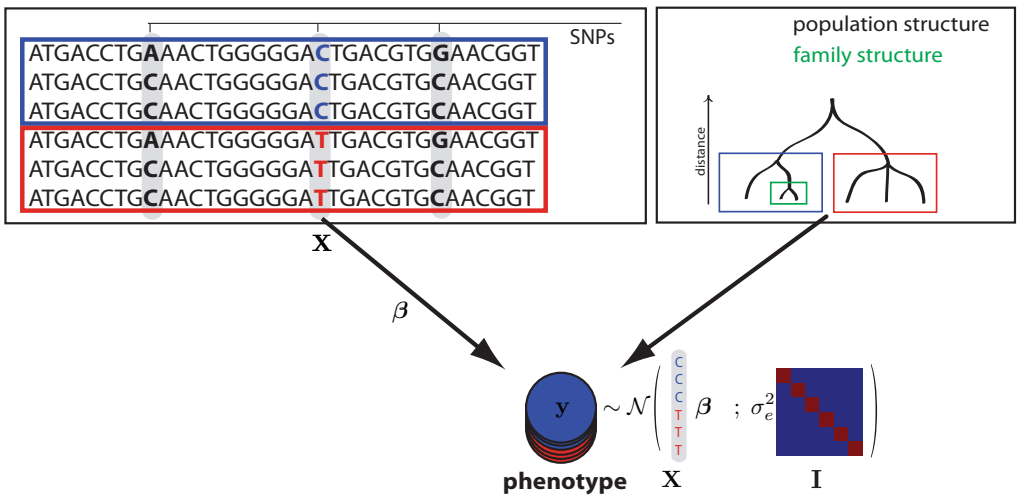
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - models the probability that two individuals share causal SNPs



[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

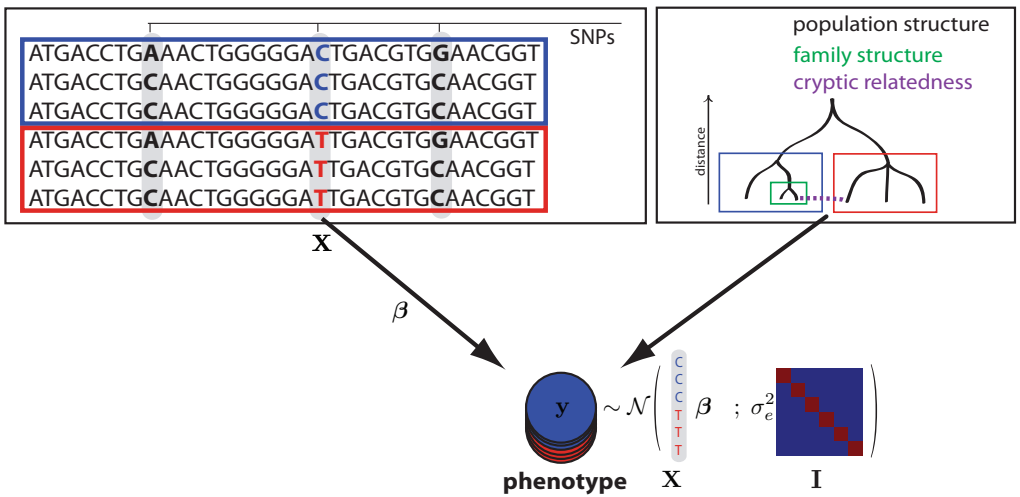
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs*



[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

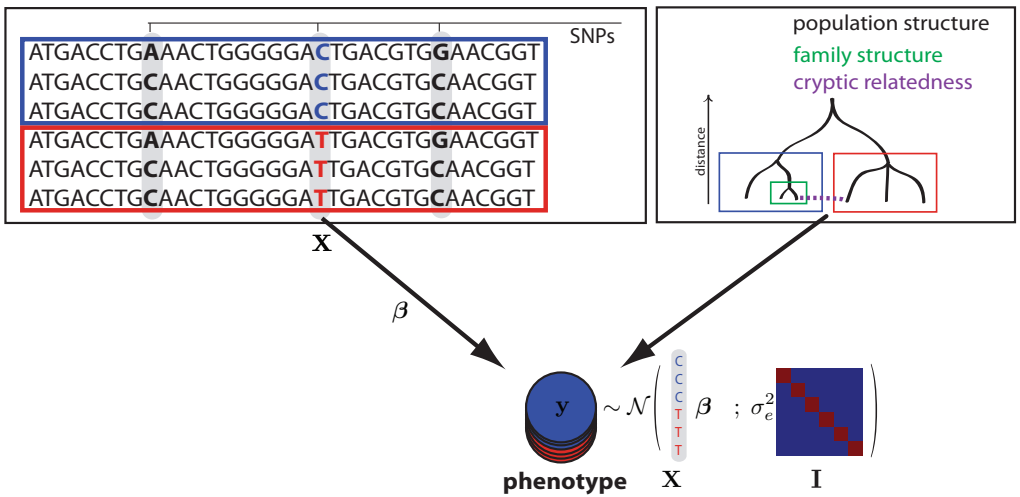
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - models the probability that two individuals share causal SNPs

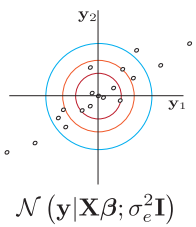


[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs*

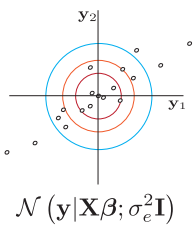


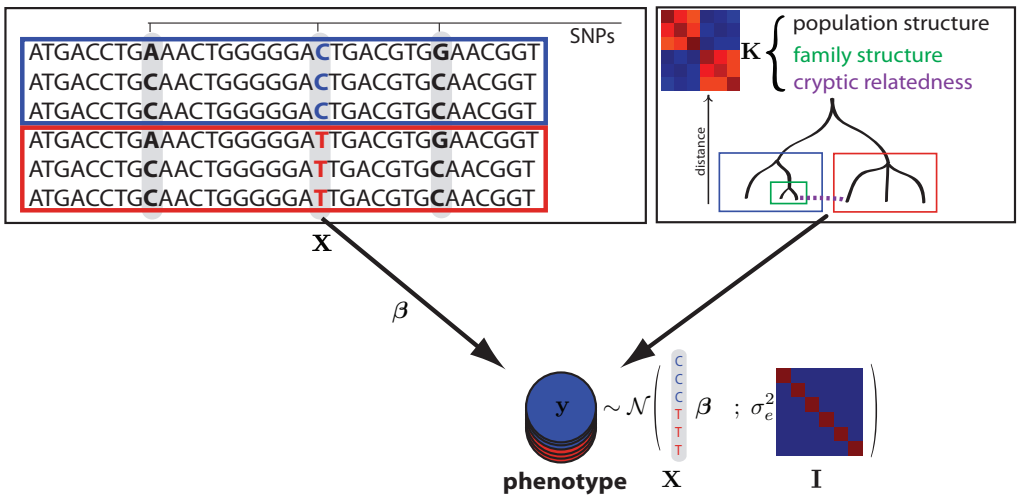


[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs.*

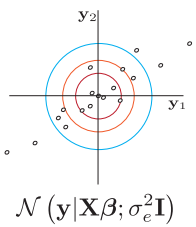


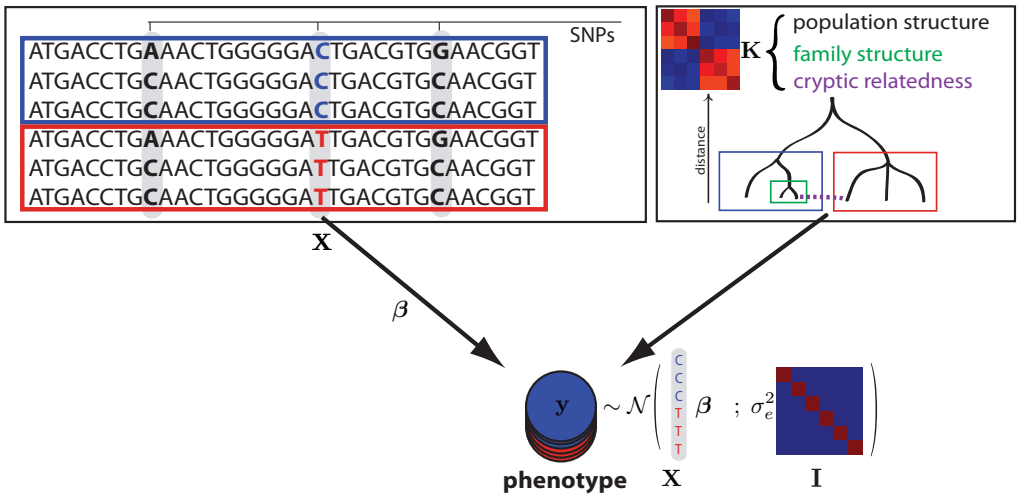


[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs.*

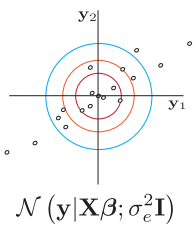


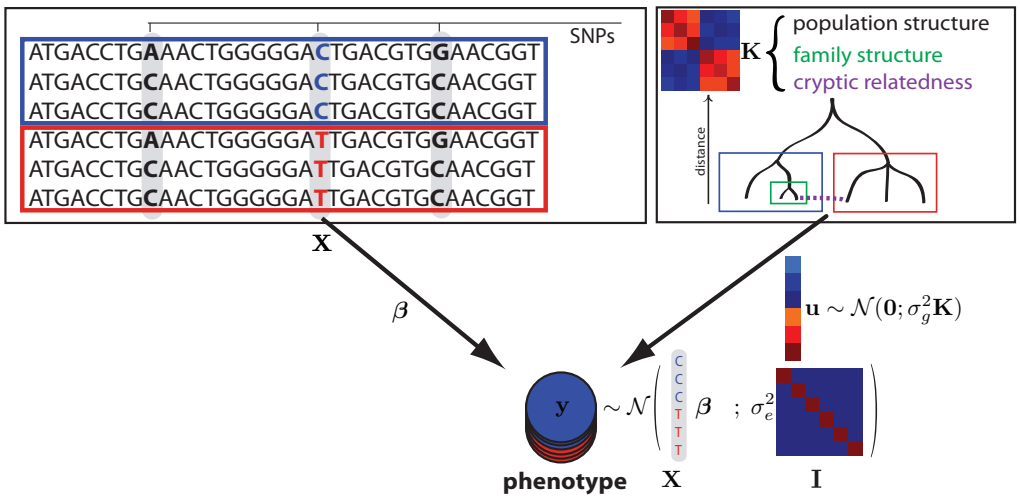


[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs.*

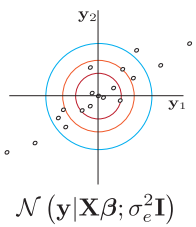


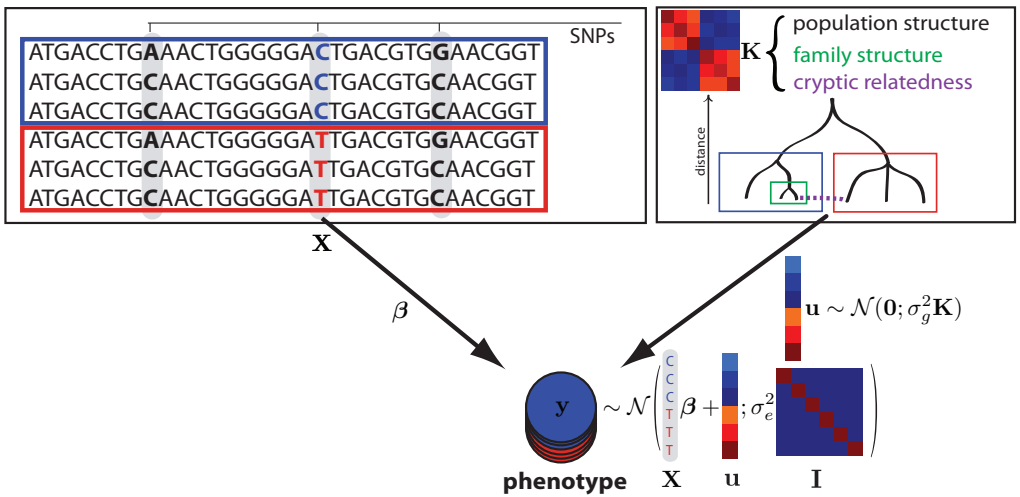


[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs.*

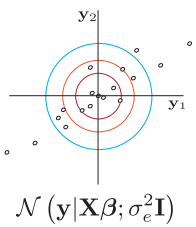


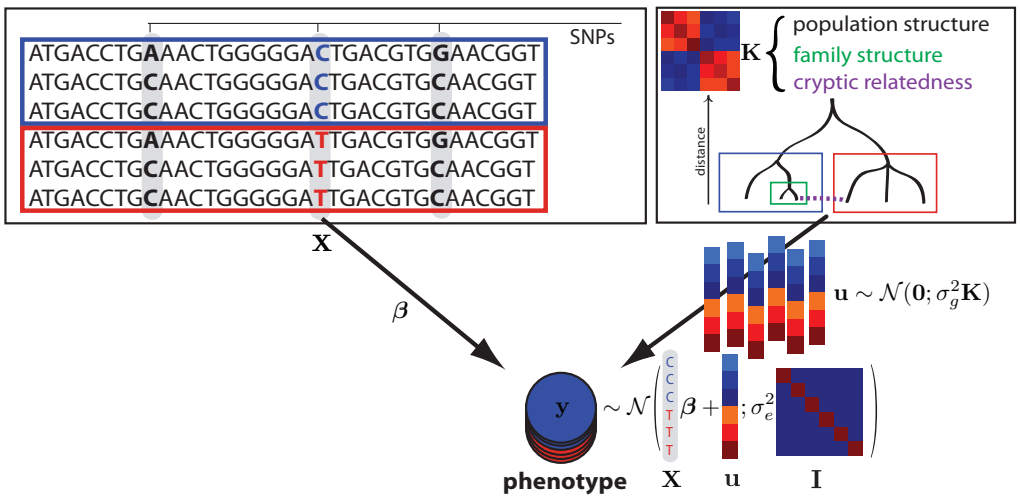


[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs.*

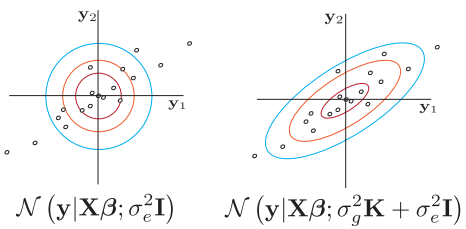


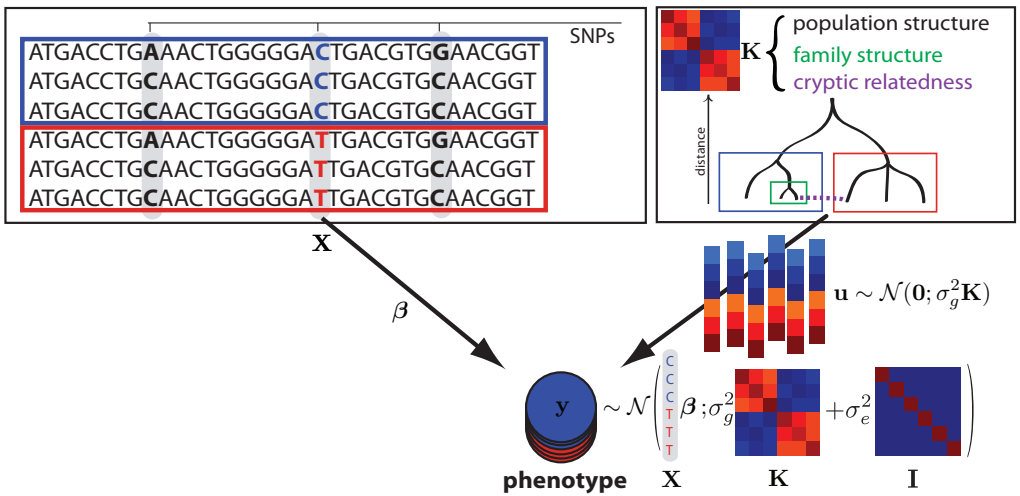


[Kang et al., 2008, 2010, Lippert et al., 2011]

Correcting for population stratification using linear mixed models

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness
- Genetic similarity matrix $\mathbf{K}$
  - Estimated from SNP data
  - *models the probability that two individuals share causal SNPs.*

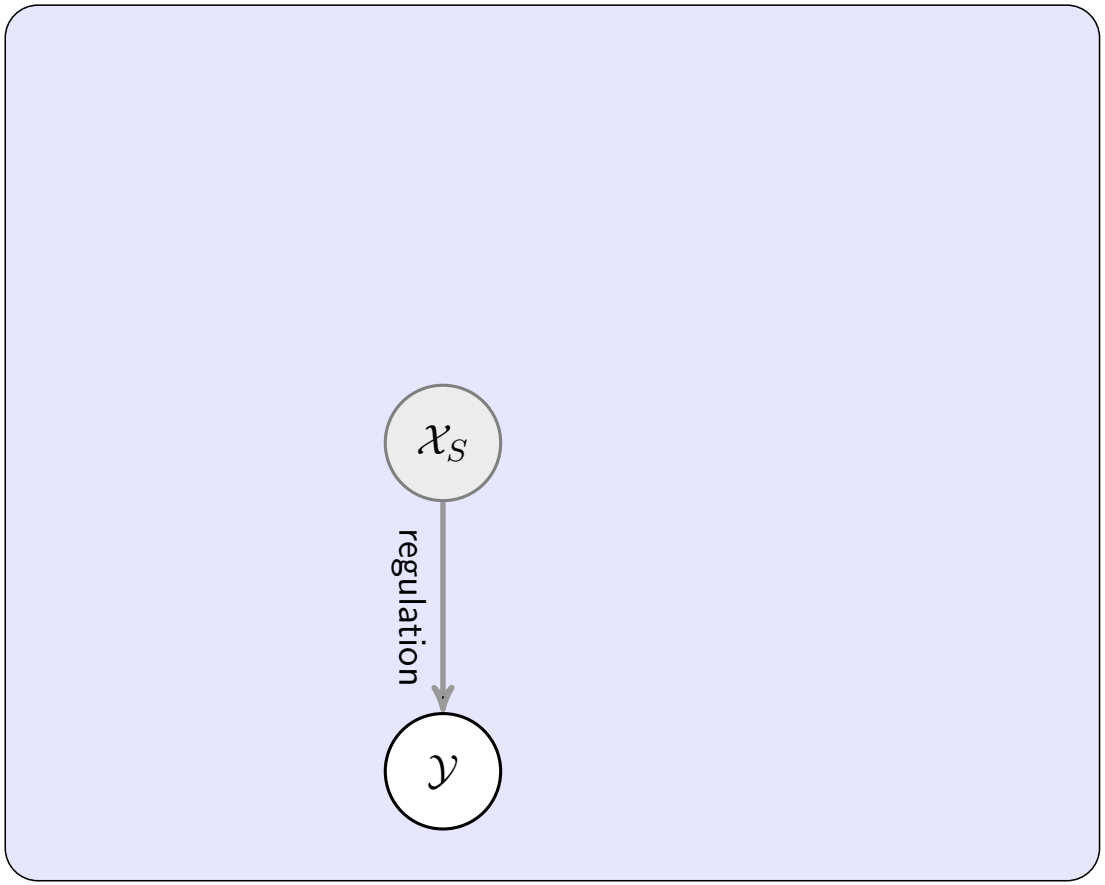




[Kang et al., 2008, 2010, Lippert et al., 2011]

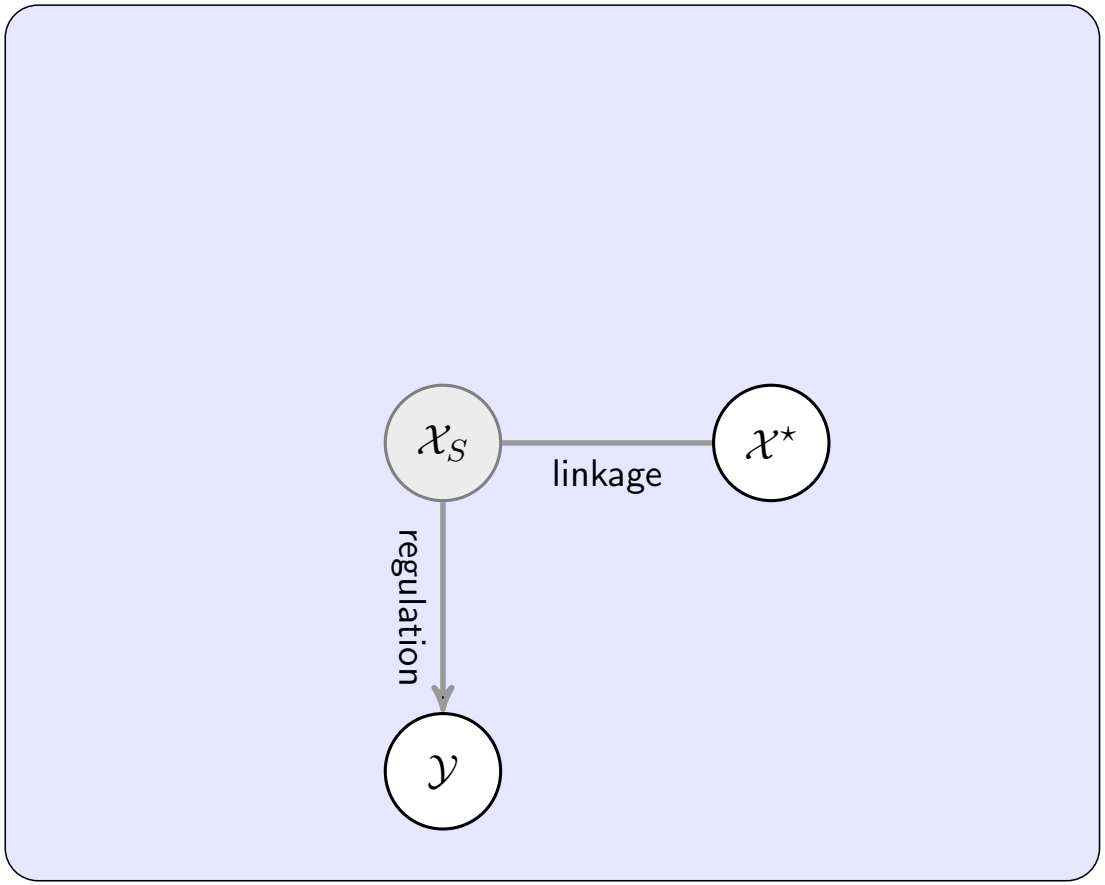
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



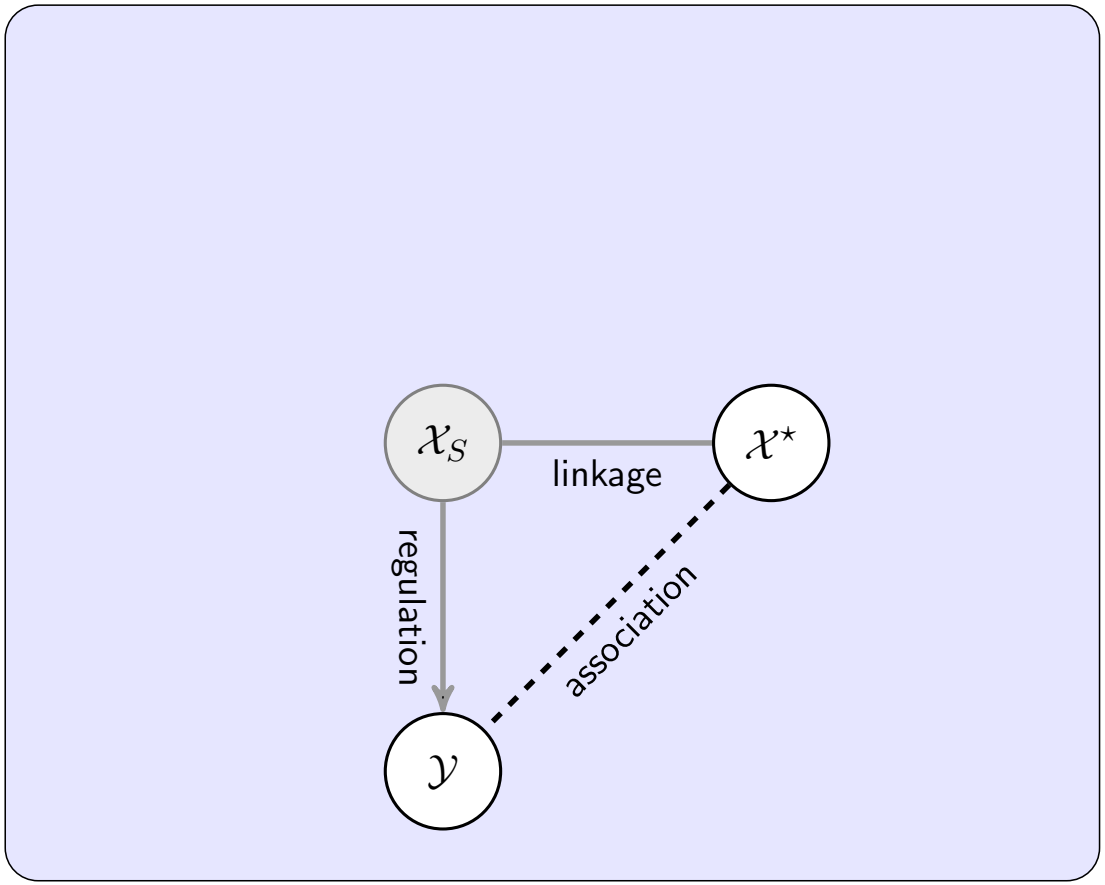
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



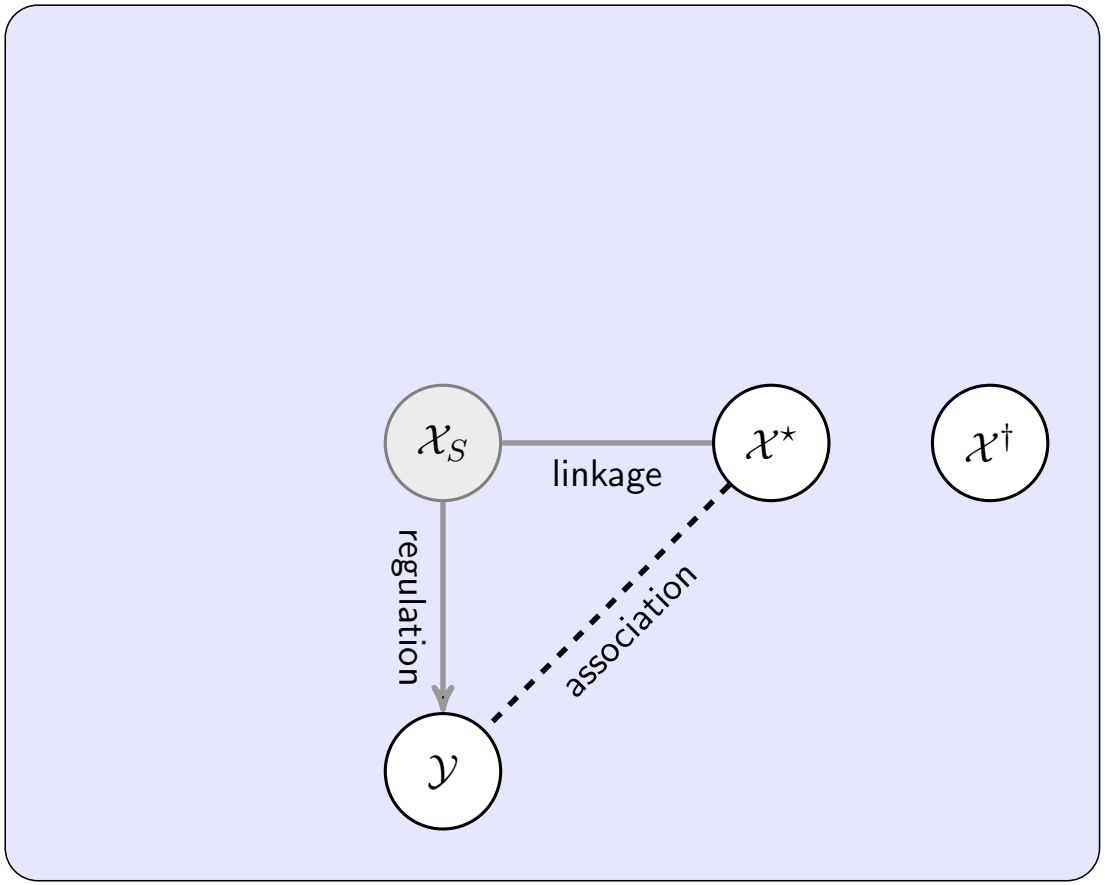
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



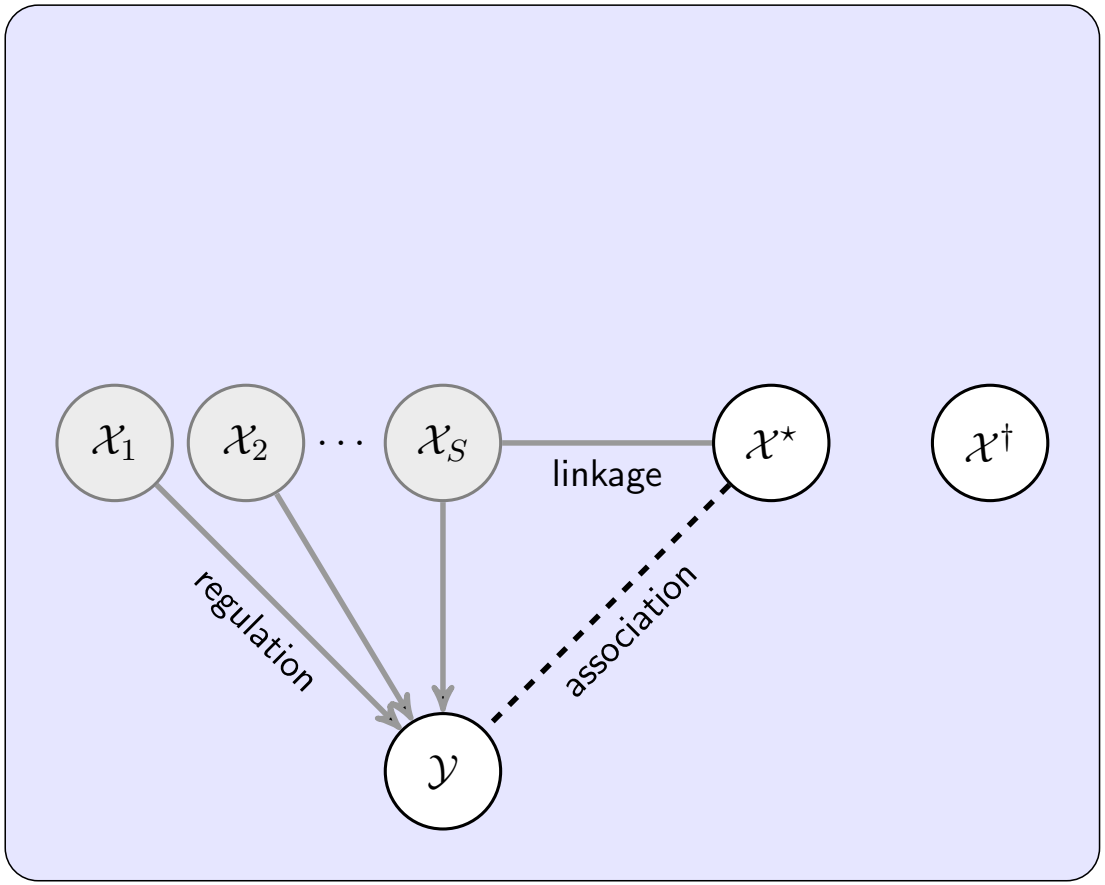
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



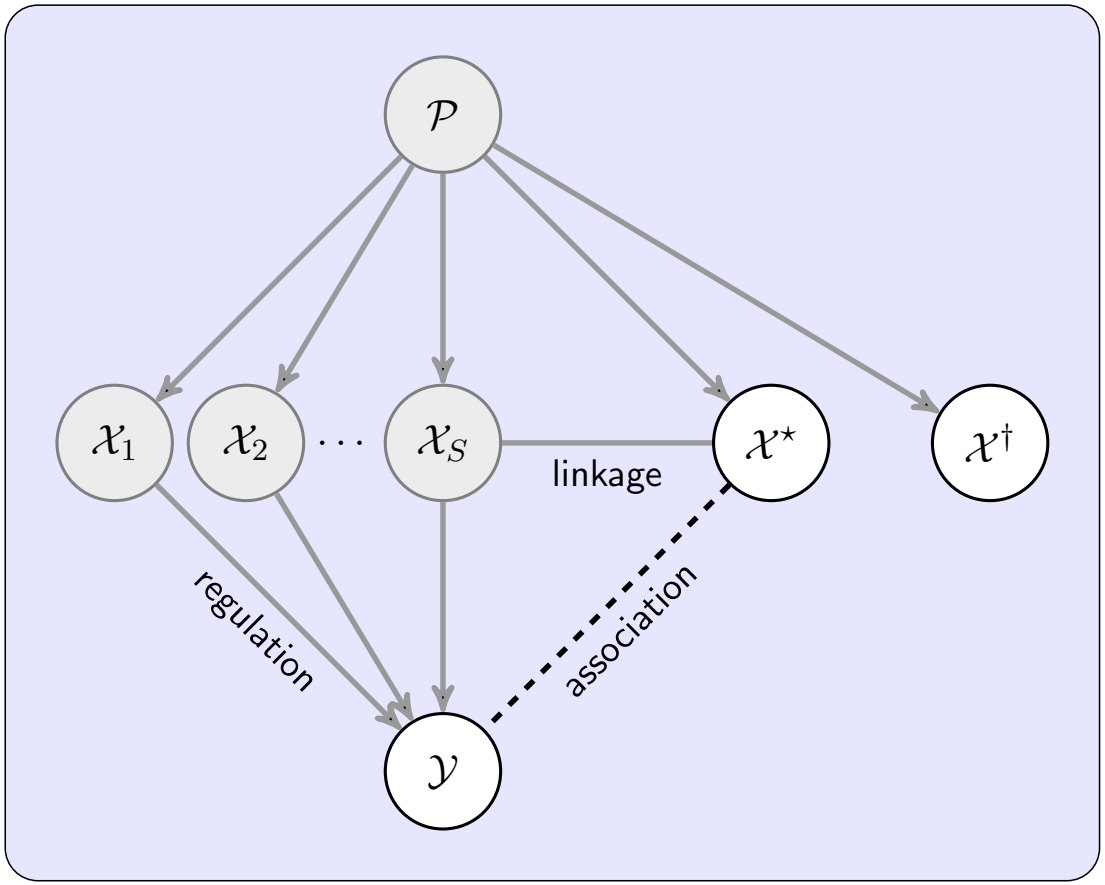
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



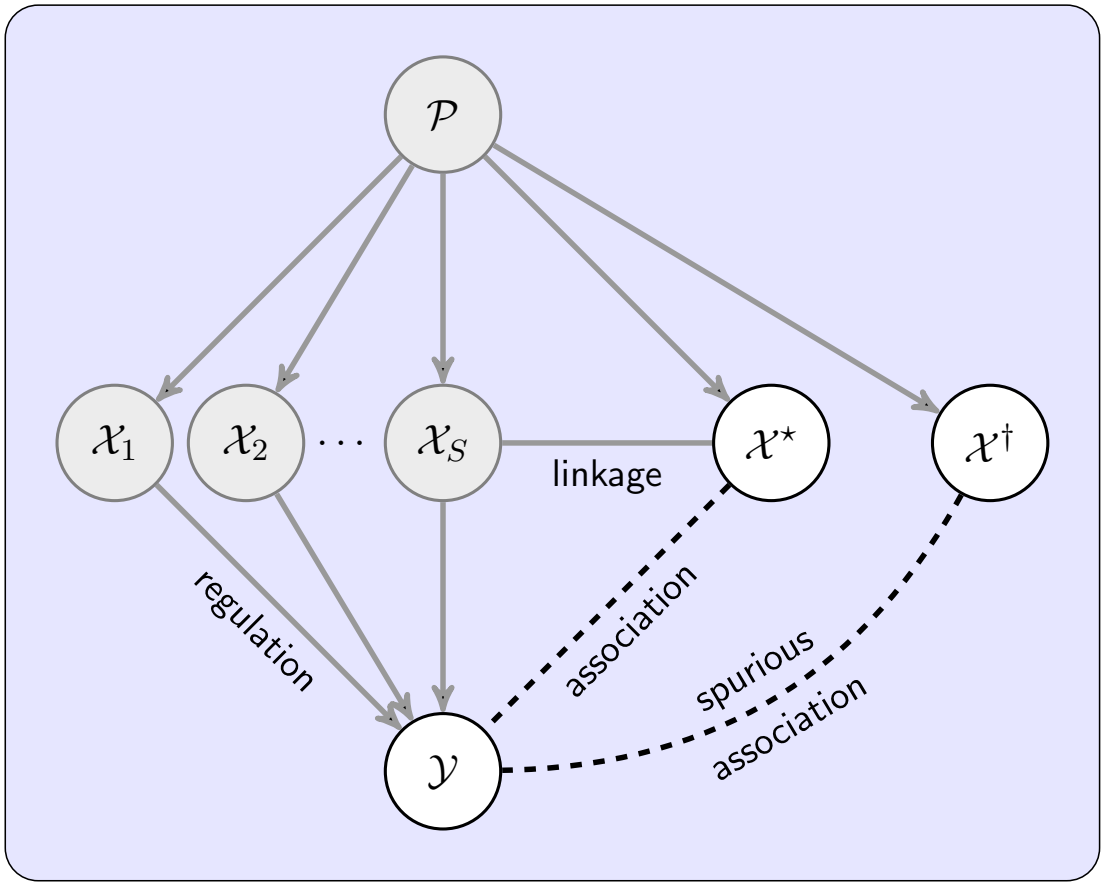
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



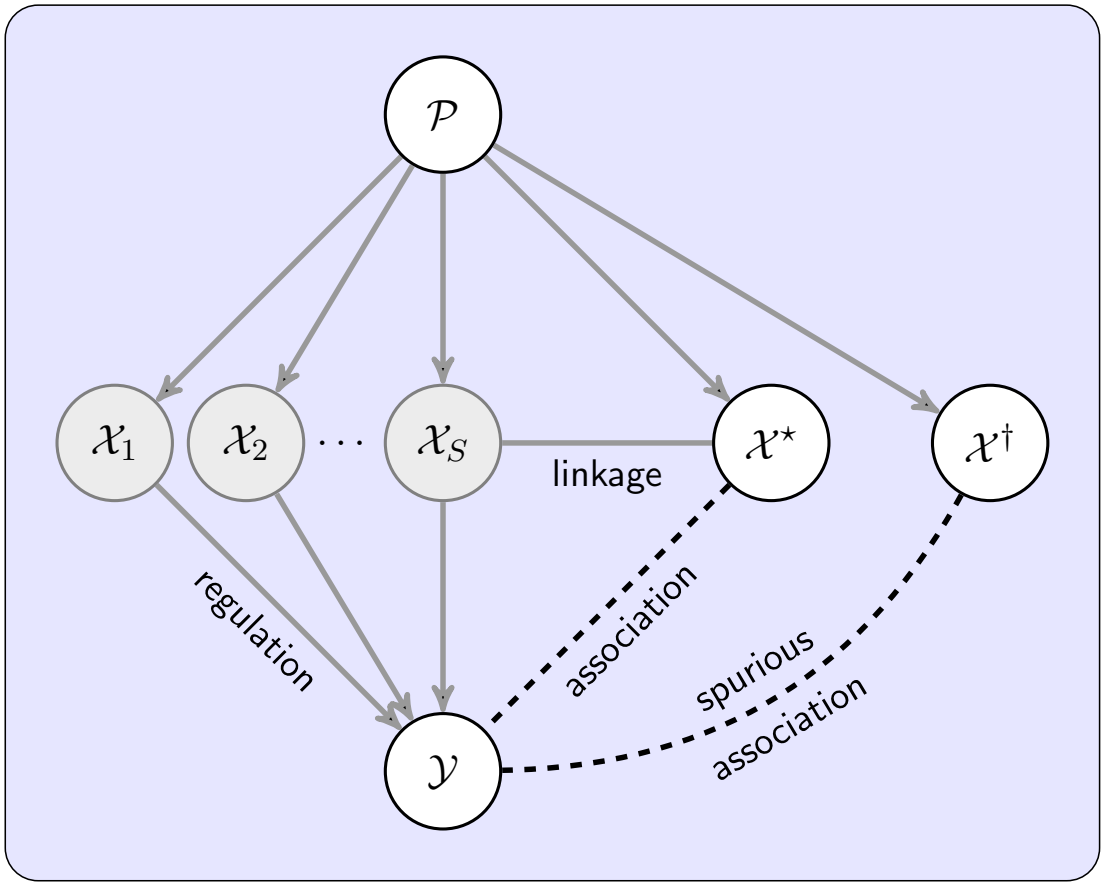
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



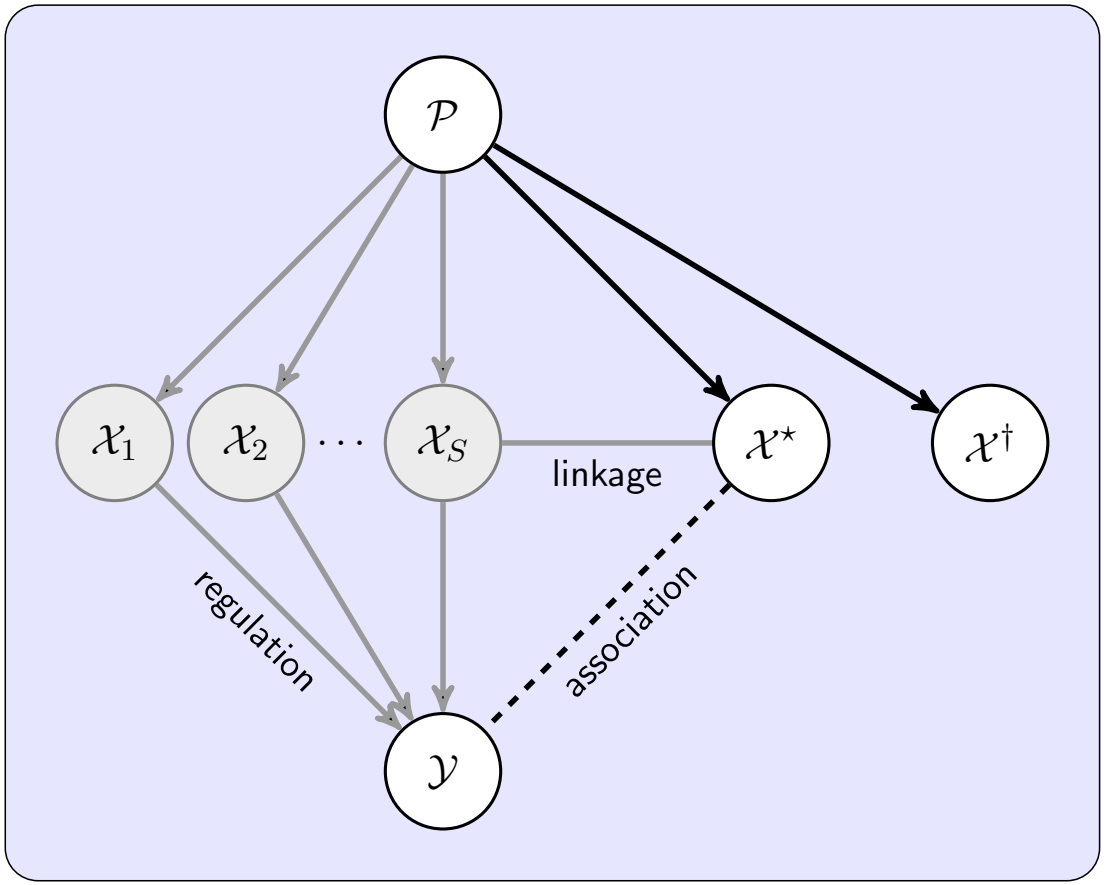
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



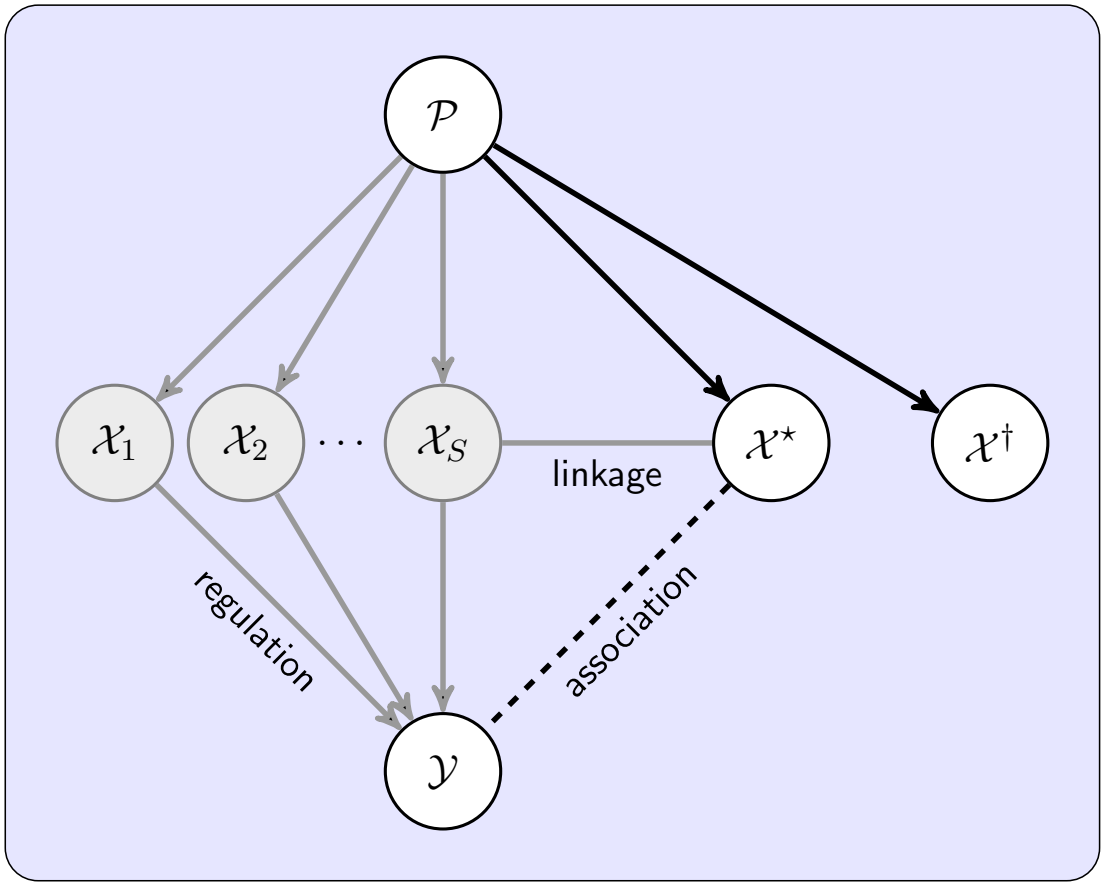
A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



A graphical model view of GWAS

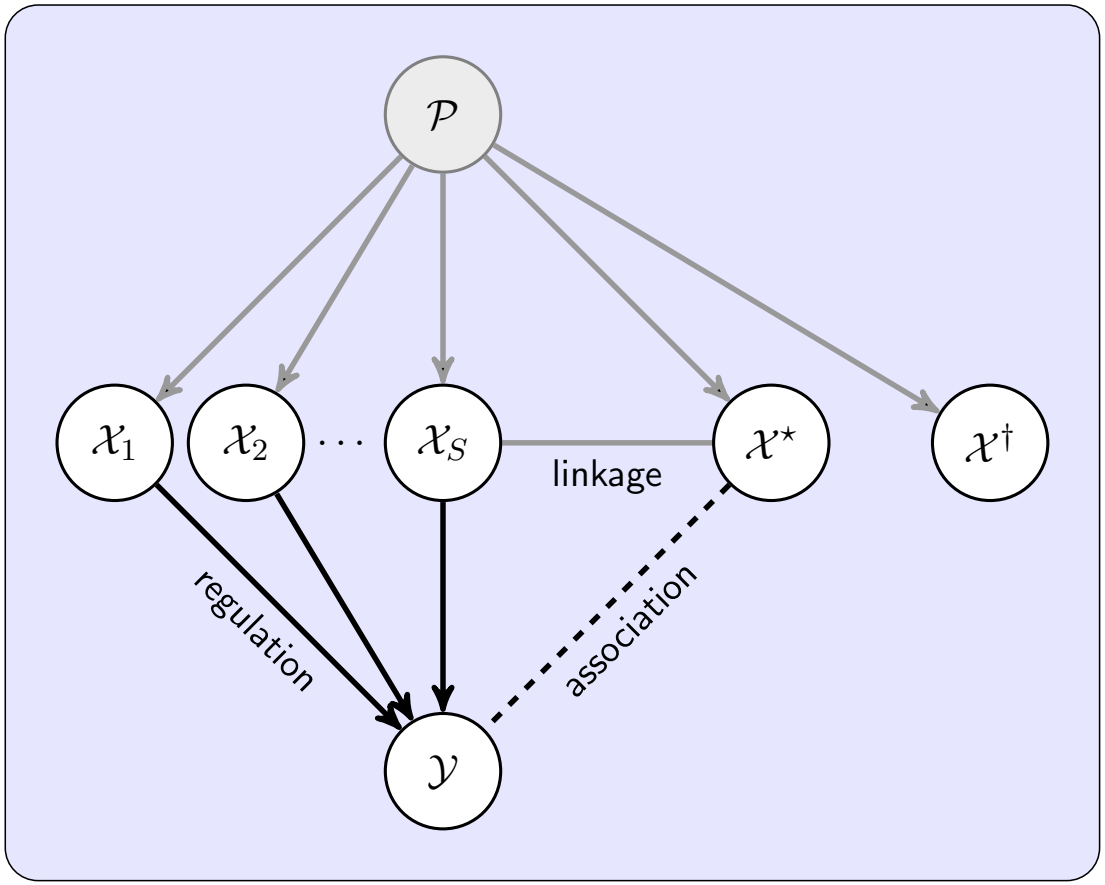
- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.

A graphical model view of GWAS

- Goal: genotype-phenotype relatioships
- Associations via linkage disequislibrium
- SNPs confounded by population structure
- Spurious associations if not taken into account
- Alternatively condition on causal SNPs



A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.

Linear mixed models

Bayesian linear regression view

- *A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.*
- For linear similarities, a LMM is equivalent to a *linear regression*.
- All SNPs are used as regression covariates.
- Uncertainty about identity of regulative SNPs expressed by using a Bayesian prior distribution on the regression weights.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}\right).$$

[Listgarten et al., 2012]

Linear mixed models

Bayesian linear regression view

- *A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.*
- For linear similarities, a LMM is equivalent to a *linear regression*.
- All SNPs are used as regression covariates.
- Uncertainty about identity of regulative SNPs expressed by using a Bayesian prior distribution on the regression weights.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \sigma_e^2\mathbf{I}\right).$$

[Listgarten et al., 2012]

Linear mixed models

Bayesian linear regression view

- *A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.*
- For linear similarities, a LMM is equivalent to a *linear regression*.
- All SNPs are used as regression covariates.
- Uncertainty about identity of regulative SNPs expressed by using a Bayesian prior distribution on the regression weights.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2 \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathsf{T}} + \sigma_e^2 \mathbf{I}\right).$$

$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2 \delta \mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2 \mathbf{I}\right) \mathrm{d}\boldsymbol{\theta}.$$

[Listgarten et al., 2012]

Linear mixed models

Bayesian linear regression view

- *A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.*
- For linear similarities, a LMM is equivalent to a *linear regression*.
- All SNPs are used as regression covariates.
- Uncertainty about identity of regulative SNPs expressed by using a Bayesian prior distribution on the regression weights.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \sigma_e^2\mathbf{I}\right).$$

$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2\delta\mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2\mathbf{I}\right)\mathrm{d}\boldsymbol{\theta}.$$

[Listgarten et al., 2012]

Linear mixed models

Bayesian linear regression view

- *A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.*
- For linear similarities, a LMM is equivalent to a *linear regression*.
- All SNPs are used as regression covariates.
- Uncertainty about identity of regulative SNPs expressed by using a Bayesian prior distribution on the regression weights.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2 \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathsf{T}} + \sigma_e^2 \mathbf{I}\right).$$

$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2 \delta \mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2 \mathbf{I}\right) \mathrm{d}\boldsymbol{\theta}.$$

Can we do better than using **all** SNPs for correction?

[Listgarten et al., 2012]

Linear mixed models

Bayesian linear regression view

- *A LMM accounts for model misspecification when testing a univariate model when in reality the trait is multi-factorial.*
- For linear similarities, a LMM is equivalent to a *linear regression*.
- All SNPs are used as regression covariates.
- Uncertainty about identity of regulative SNPs expressed by using a Bayesian prior distribution on the regression weights.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \sigma_e^2\mathbf{I}\right).$$

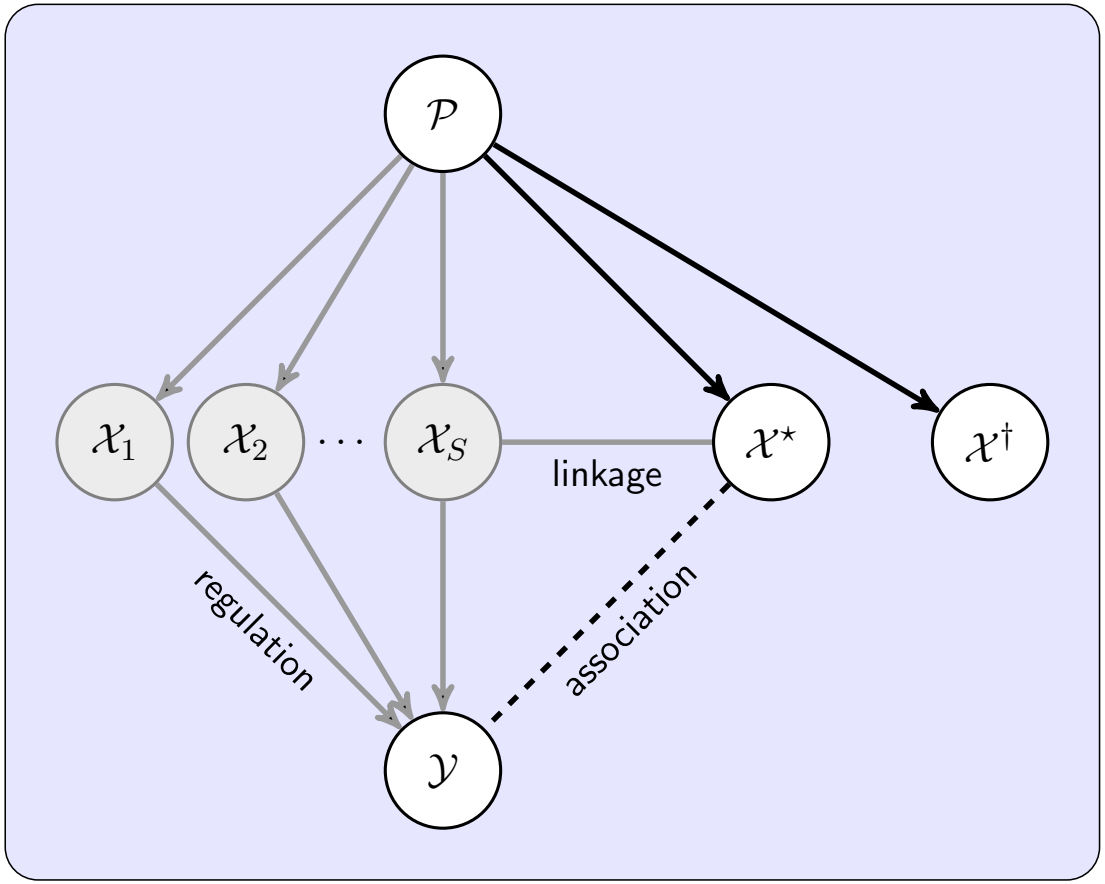
$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2\delta\mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2\mathbf{I}\right)\mathrm{d}\boldsymbol{\theta}.$$

Can we do better than using **all** SNPs for correction?
$\rightarrow$ Select SNPs by their association to the phenotype.

[Listgarten et al., 2012]

FaST-LMM **select**

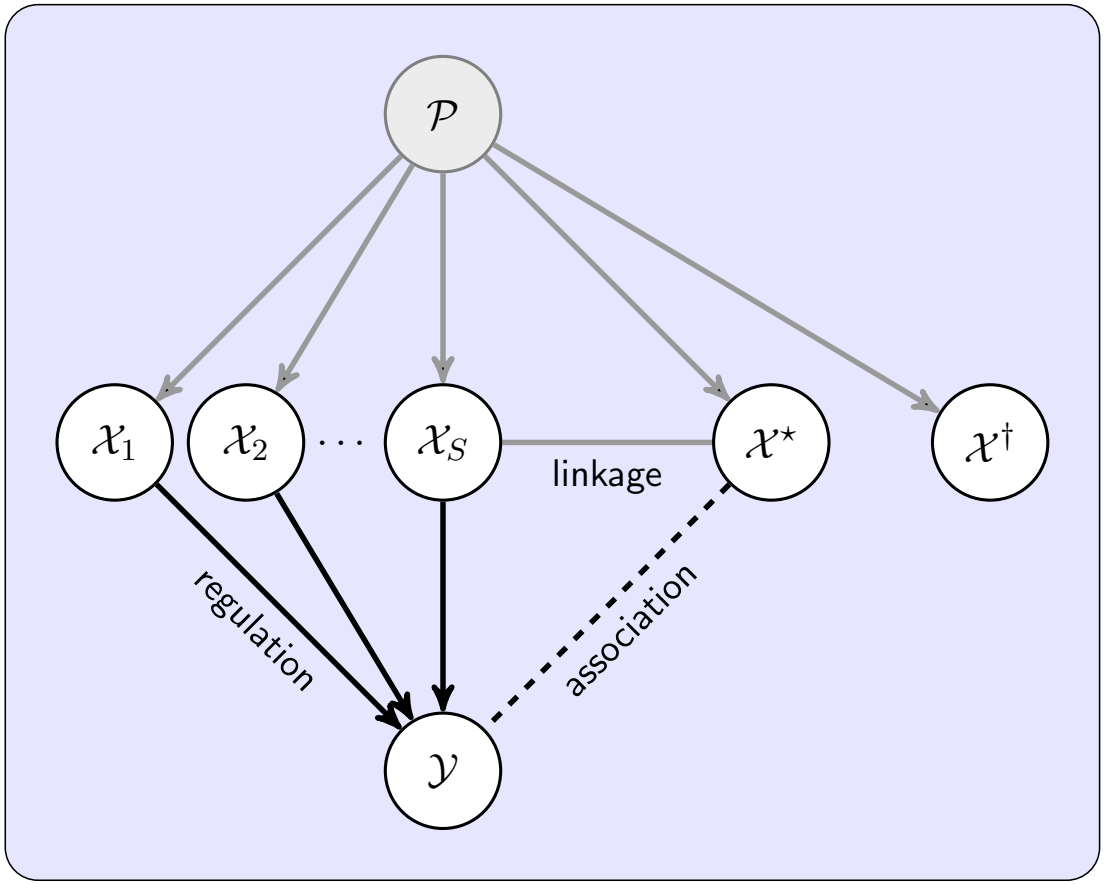
- Compute similarity matrix based on highly associated SNPs
- Equivalent to linear regression conditioned on these SNPs
- Conditioning on SNPs in LD blocks association.
  $\rightarrow$ False negative results!
- Remove SNPs in LD from computation of similarity matrix



[Listgarten et al., 2012]

FaST-LMM **select**

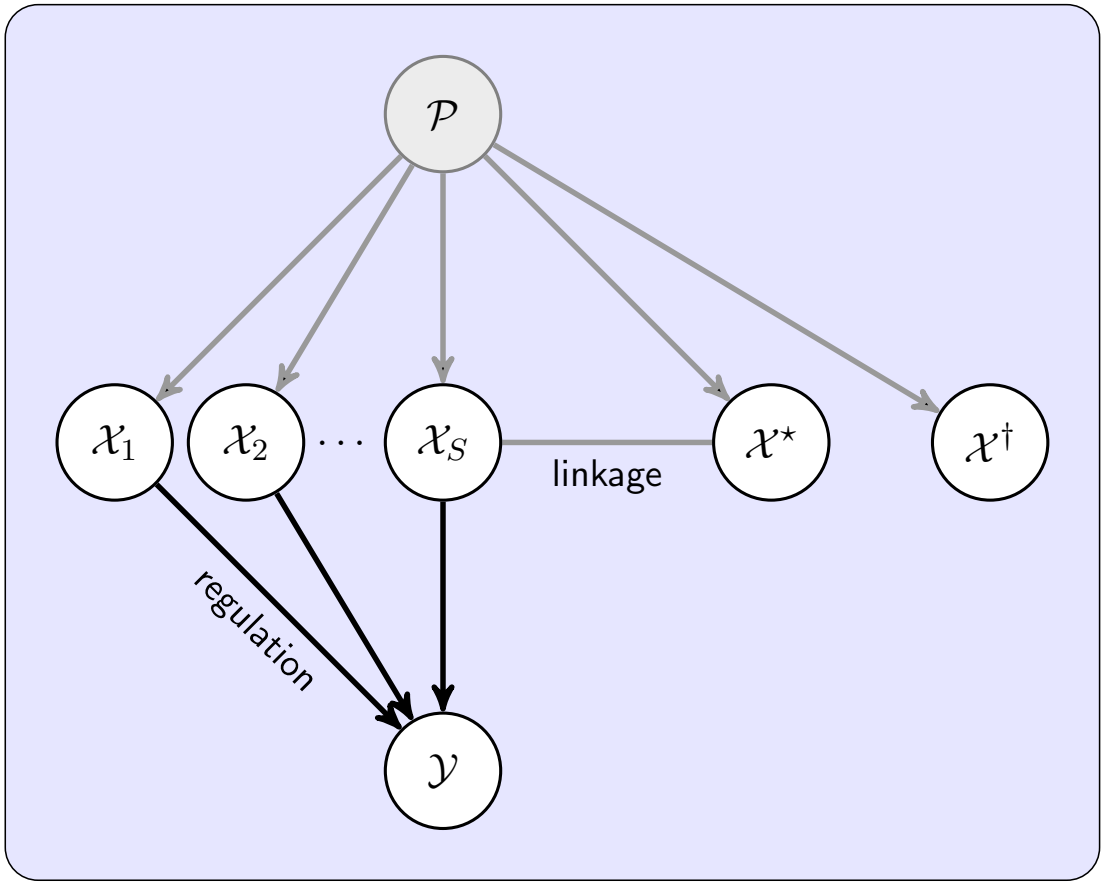
- Compute similarity matrix based on highly associated SNPs
- Equivalent to linear regression conditioned on these SNPs
- Conditioning on SNPs in LD blocks association.
  $\rightarrow$ False negative results!
- Remove SNPs in LD from computation of similarity matrix



[Listgarten et al., 2012]

FaST-LMM **select**

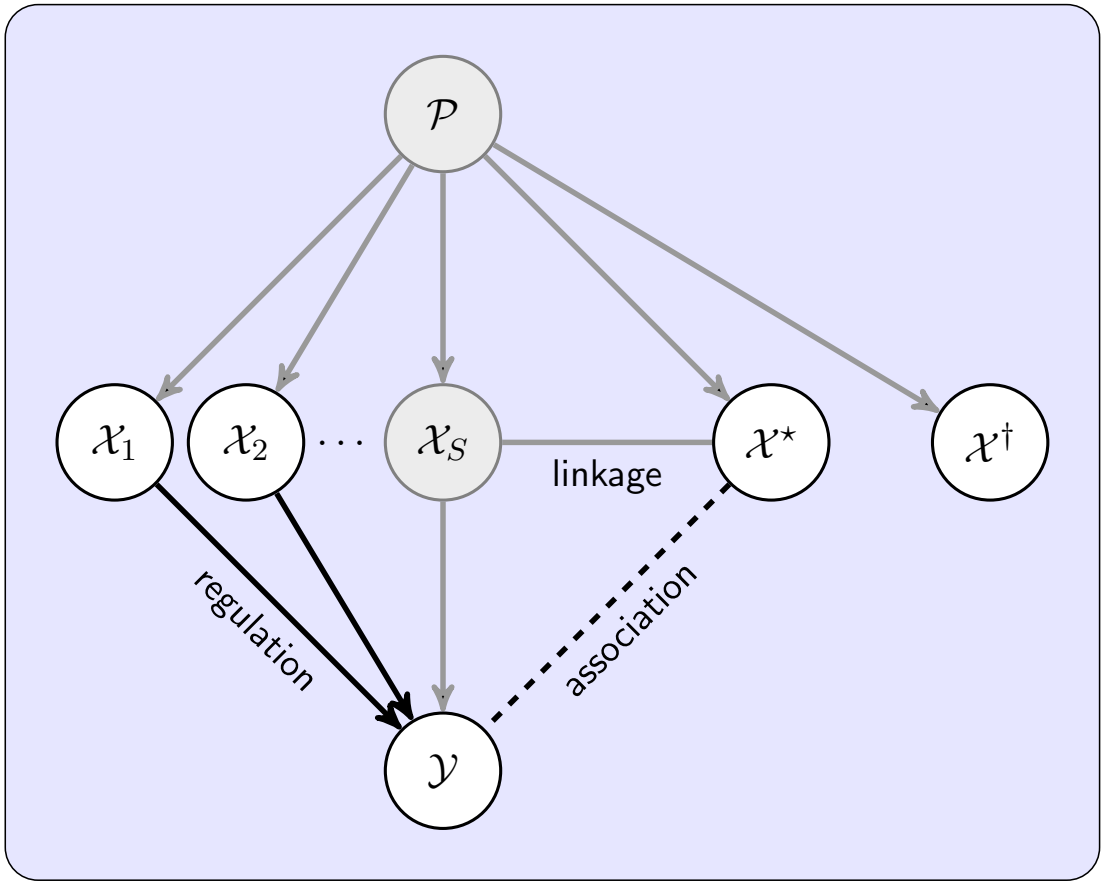
- Compute similarity matrix based on highly associated SNPs
- Equivalent to linear regression conditioned on these SNPs
- Conditioning on SNPs in LD blocks association.
  $\rightarrow$ False negative results!
- Remove SNPs in LD from computation of similarity matrix



[Listgarten et al., 2012]

FaST-LMM **select**

- Compute similarity matrix based on highly associated SNPs
- Equivalent to linear regression conditioned on these SNPs
- Conditioning on SNPs in LD blocks association.
  $\rightarrow$ False negative results!
- Remove SNPs in LD from computation of similarity matrix



[Listgarten et al., 2012]

Proximal contamination

- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathsf{T}} + \sigma_e^2\mathbf{I}\right).$$

$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2\delta\mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2\mathbf{I}\right) \mathrm{d}\boldsymbol{\theta}.$$

[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathsf{T}} + \sigma_e^2\mathbf{I}\right).$$

$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2\delta\mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2\mathbf{I}\right) \mathrm{d}\boldsymbol{\theta}.$$

[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathsf{T}} + \sigma_e^2\mathbf{I}\right).$$

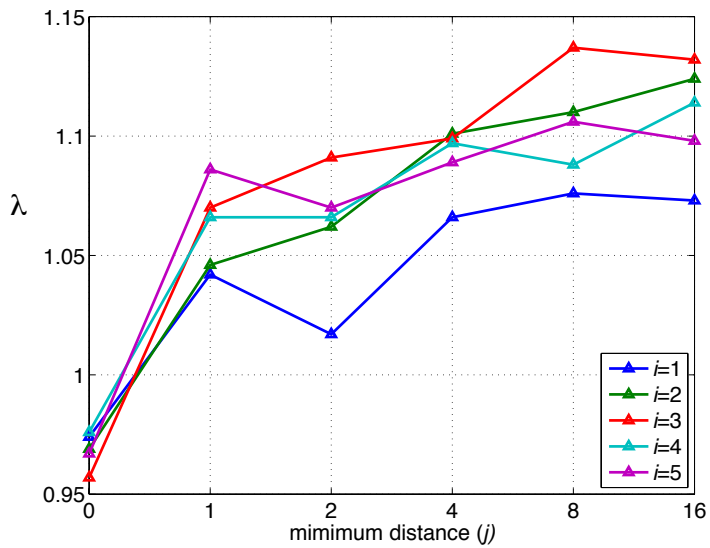
$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2\delta\mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2\mathbf{I}\right) \mathrm{d}\boldsymbol{\theta}.$$

[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.

- Proximal contamination on IBD phenotype [WTCCC, 2007]
- Test SNPs are equally distant sampled along the chromosome
- Compute 6 similarity matrices containing same number of SNPs at growing distance to test SNPs.

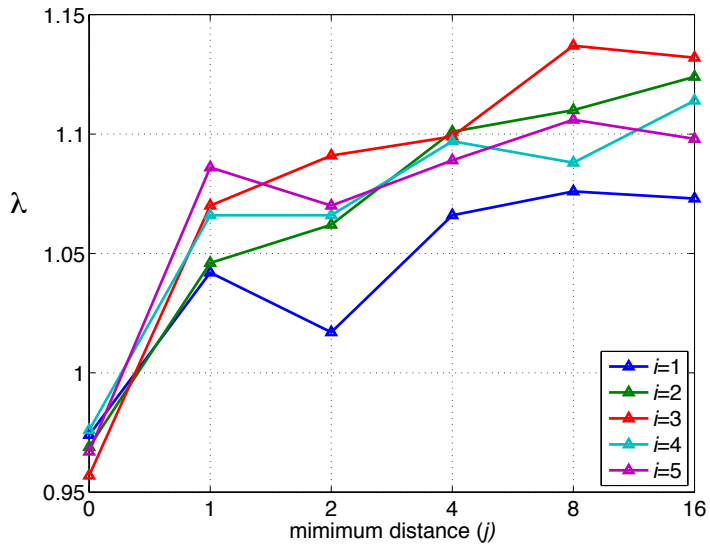


[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.

- Proximal contamination on IBD phenotype [WTCCC, 2007]
- Test SNPs are equally distant sampled along the chromosome
- Compute 6 similarity matrices containing same number of SNPs at growing distance to test SNPs.

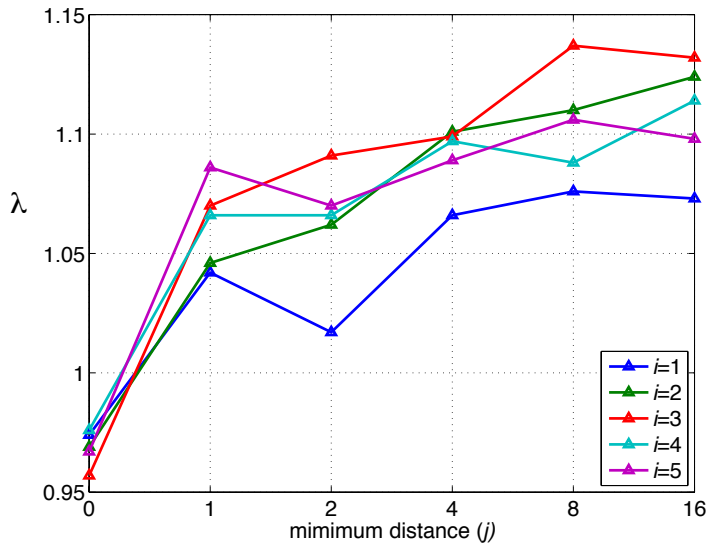


[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.

- Proximal contamination on IBD phenotype [WTCCC, 2007]
- Test SNPs are equally distant sampled along the chromosome
- Compute 6 similarity matrices containing same number of SNPs at growing distance to test SNPs.

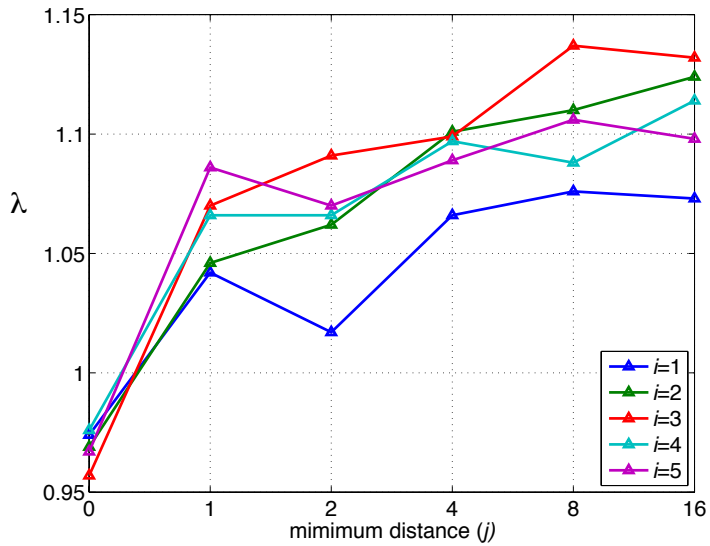


[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.

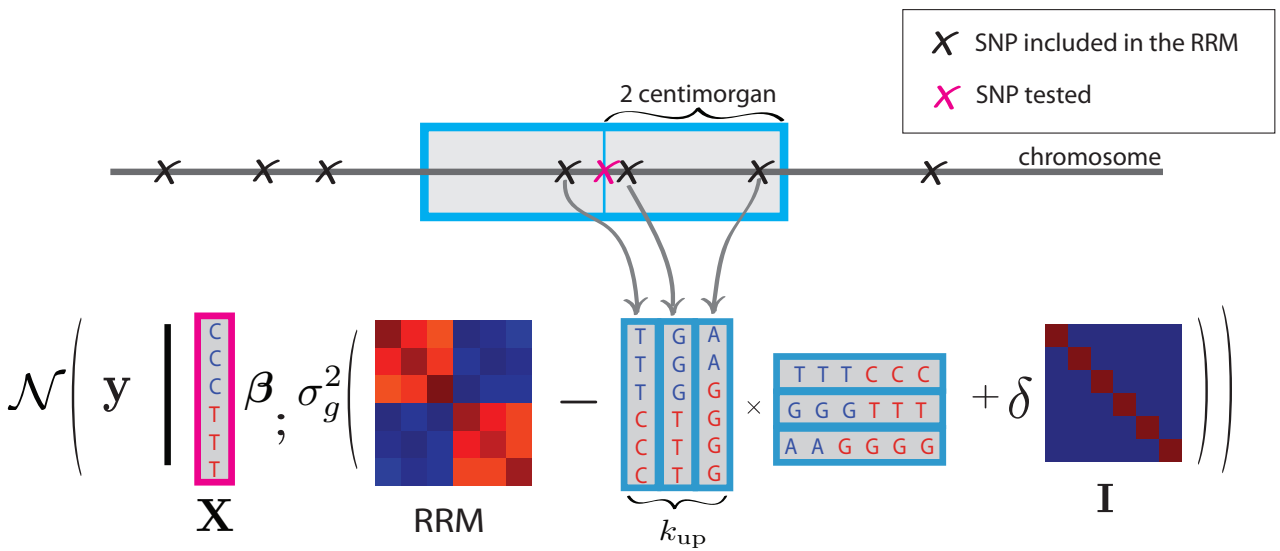
- Proximal contamination on IBD phenotype [WTCCC, 2007]
- Test SNPs are equally distant sampled along the chromosome
- Compute 6 similarity matrices containing same number of SNPs at growing distance to test SNPs.



[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

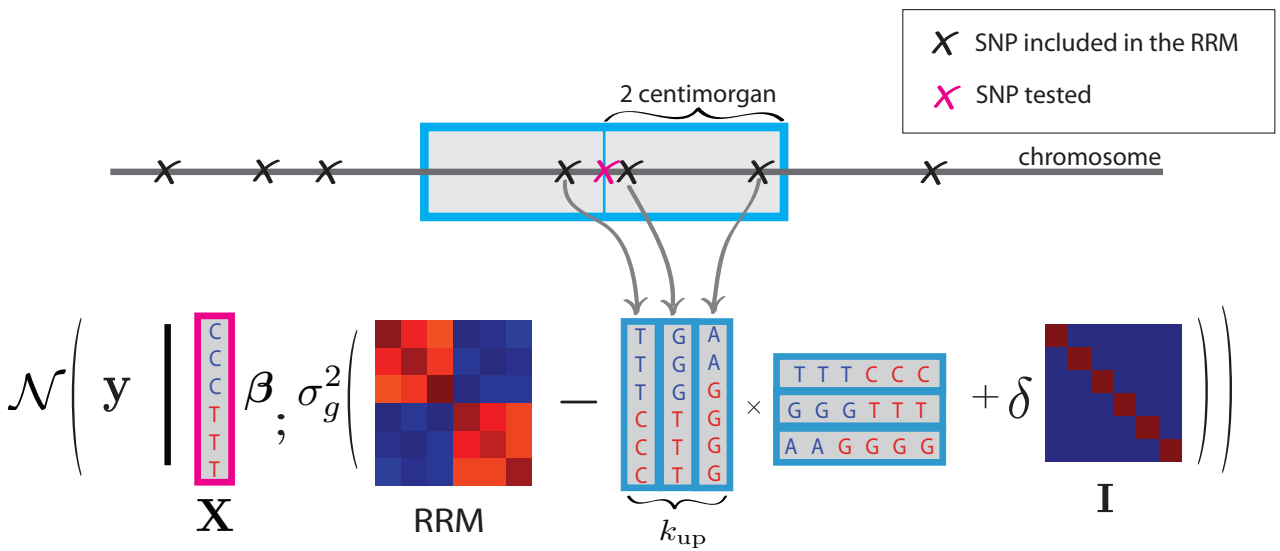
- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.



[Lippert et al., 2011, Listgarten et al., 2012]

Proximal contamination

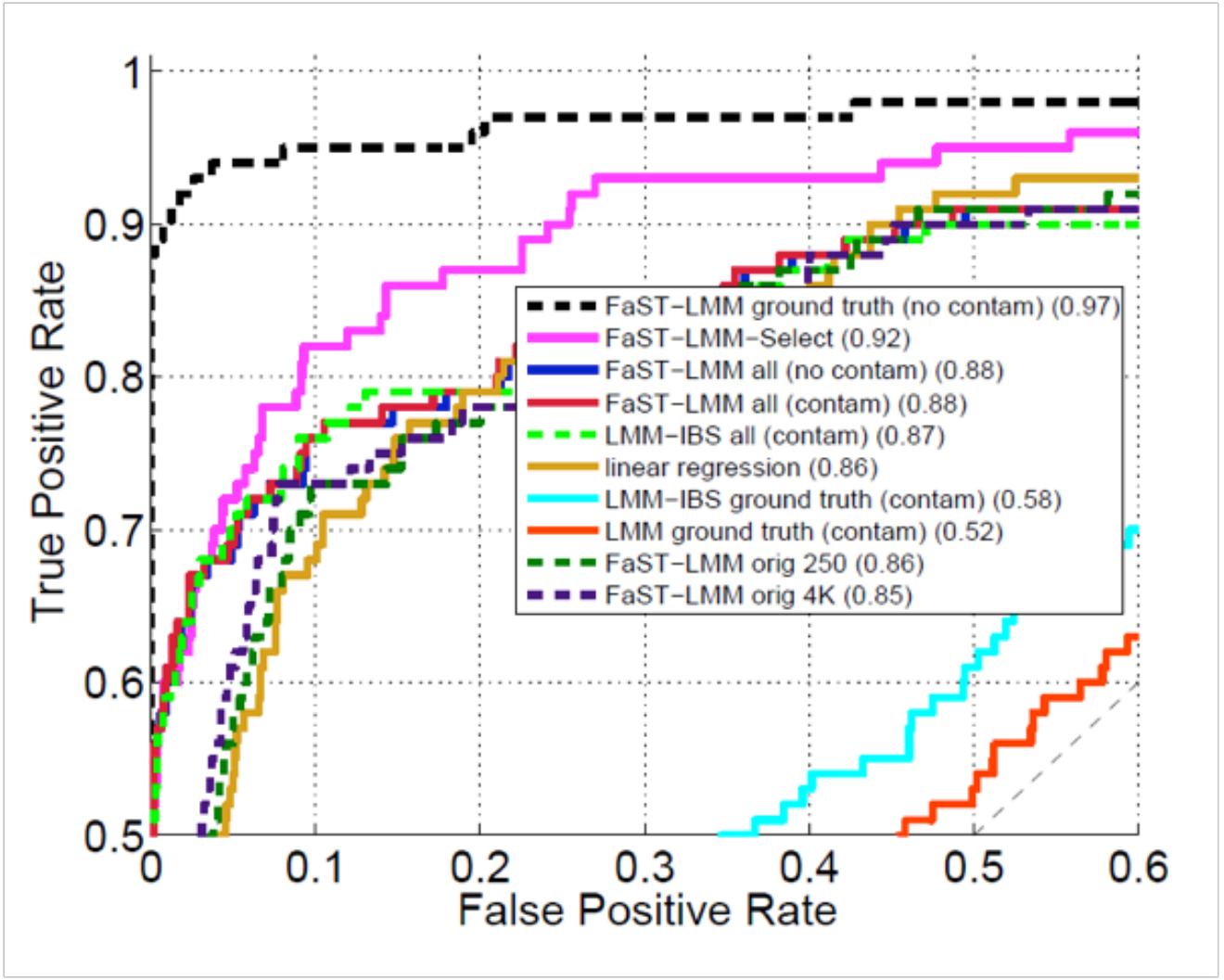
- SNPs that are in close proximity on the chromosome are highly correlated.
- Having a SNP in the similarity matrix that is close to the SNP tested is equivalent to conditioning on this SNP in the null-model.
- Correct by removing a sliding window around test-SNP from the similarity matrix.
- Correction is computed efficiently by subtracting a low-rank term.



[Lippert et al., 2011, Listgarten et al., 2012]

Power study

- 3000 individuals
- Two populations (6:4)
- Balding-Nichols model
- 100000 non-causal SNPs
- 100 causal SNPs ($\sigma_g^2 = \sigma_e^2 = 0.1$)
- Test-set of 5100 SNPs



[Listgarten et al., 2012]

Inflammatory bowel disease [WTCCC, Nature 2007]

Algorithm parameters				**Algorithm performance**					
Name	SNP selection method	#SNPs in matrix	Avoid prox contam	λ_{GC}	False Positives	True Positives	Runtime (min) without speedup	Runtime (min) with speedup	Memory use (GB)
FaST-LMM-Select	Select	310	yes	1.08	0	100	1.3×10^3	45	<1
FaST-LMM all	All	All	yes	1.09	2	108	4.0×10^6	4567	86
FaST-LMM orig 310	Equi-spaced	310	yes	1.26	15	128	1.1×10^3	6	<1
FaST-LMM orig 4K	Equi-spaced	4000	yes	1.17	8	114	2.1×10^5	30	2
Traditional	All	All	no	0.97	2	64	42	NA	45

SNPs considered True Positive if:

- Reported in WTCCC paper [WTCCC, Nature 2007]
- Reported in meta analysis [Franke et al., Nat Gen 2010]
- In major histocompatibility complex (MHC) region

[Burton et al., 2007]

Conclusions

- Relationship between linear mixed models and multi-variate modelling using linear regression
- Computing genetic similarities on associated SNPs improves correction *FaST-LMM* **select**
- Inclusion of SNPs in LD to the SNP tested causes deflated tests (proximal contamination)
- Efficient exclusion of SNPs in LD from the similarity matrix feasible using low-rank updates

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2 \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right).$$

$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2 \delta \mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2 \mathbf{I}\right) \mathrm{d}\boldsymbol{\theta}.$$

Conclusions

- Relationship between linear mixed models and multi-variate modelling using linear regression
- Computing genetic similarities on associated SNPs improves correction *FaST-LMM* **select**
- Inclusion of SNPs in LD to the SNP tested causes deflated tests (proximal contamination)
- Efficient exclusion of SNPs in LD from the similarity matrix feasible using low-rank updates

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \sigma_e^2\mathbf{I}\right).$$

$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2\delta\mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2\mathbf{I}\right)\mathrm{d}\boldsymbol{\theta}.$$

Conclusions

- Relationship between linear mixed models and multi-variate modelling using linear regression
- Computing genetic similarities on associated SNPs improves correction *FaST-LMM* **select**
- Inclusion of SNPs in LD to the SNP tested causes deflated tests (proximal contamination)
- Efficient exclusion of SNPs in LD from the similarity matrix feasible using low-rank updates

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta; \sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \sigma_e^2\mathbf{I}\right).$$

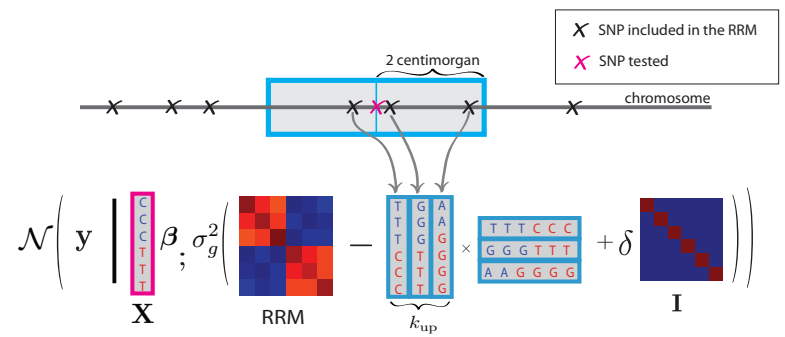
$$\propto \int \mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta + \tilde{\mathbf{X}}\boldsymbol{\theta}; \sigma_e^2\delta\mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}; \sigma_g^2\mathbf{I}\right)\mathrm{d}\boldsymbol{\theta}.$$

Conclusions

- Relationship between linear mixed models and multi-variate modelling using linear regression
- Computing genetic similarities on associated SNPs improves correction *FaST-LMM* **select**
- Inclusion of SNPs in LD to the SNP tested causes deflated tests (proximal contamination)
- Efficient exclusion of SNPs in LD from the similarity matrix feasible using low-rank updates

$$\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta;\sigma_g^2\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}+\sigma_e^2\mathbf{I}\right).$$

$$\propto\int\mathcal{N}\left(\mathbf{y}|\mathbf{x}\beta+\tilde{\mathbf{X}}\boldsymbol{\theta};\sigma_e^2\delta\mathbf{I}\right)\cdot\mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0};\sigma_g^2\mathbf{I}\right)\mathrm{d}\boldsymbol{\theta}.$$



References I

P. Burton, D. Clayton, L. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. Kwiatkowski, M. McCarthy, W. Ouwehand, N. Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 107, 2008.

H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, 42:348–354, Apr 2010.

C. Lippert, J. Listgarten, Y. Liu, C. Kadie, R. Davidson, and D. Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):838;835, 10 2011. doi: 10.1038/nmeth.1681.

J. Listgarten, C. Lippert, C. Kadie, R. Davidson, E. Eskin, and D. Heckerman. Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–526, 2012.