Linear models for GWAS

I: Introduction and linear regression

Christoph Lippert

Microsoft Research, Los Angeles, USA

Microsoft Research

Current topics in computational biology
UCLA
October 15^{th}, 2012

Course structure

October 15^{th}

- Introduction
  - Terminology
  - Study design
  - Data preparation
  - Challenges and pitfalls
  - Course overview
- Linear regression
  - Parameter estimation
  - Statistical testing

October 17^{th}

- Linear mixed models
  - Population structure correction
  - Parameter estimation
  - Variance component modeling
  - Phenotype prediction
- Further challenges and outlook

Course structure

October 15^{th}

- Introduction
  - Terminology
  - Study design
  - Data preparation
  - Challenges and pitfalls
  - Course overview
- Linear regression
  - Parameter estimation
  - Statistical testing

October 17^{th}

- Linear mixed models
  - Population structure correction
  - Parameter estimation
  - Variance component modeling
  - Phenotype prediction
- Further challenges and outlook

Course structure

October 15^{th}

- Introduction
  - Terminology
  - Study design
  - Data preparation
  - Challenges and pitfalls
  - Course overview
- Linear regression
  - Parameter estimation
  - Statistical testing

October 17^{th}

- Linear mixed models
  - Population structure correction
  - Parameter estimation
  - Variance component modeling
  - Phenotype prediction
- Further challenges and outlook

Course structure

October 15^{th}

- Introduction
  - Terminology
  - Study design
  - Data preparation
  - Challenges and pitfalls
  - Course overview
- Linear regression
  - Parameter estimation
  - Statistical testing

October 17^{th}

- Linear mixed models
  - Population structure correction
  - Parameter estimation
  - Variance component modeling
  - Phenotype prediction
- Further challenges and outlook

Course structure

October 15^{th}

- Introduction
  - Terminology
  - Study design
  - Data preparation
  - Challenges and pitfalls
  - Course overview
- Linear regression
  - Parameter estimation
  - Statistical testing

October 17^{th}

- Linear mixed models
  - Population structure correction
  - Parameter estimation
  - Variance component modeling
  - Phenotype prediction
- Further challenges and outlook

Outline

Outline

Introduction
- Why QTL mapping
- Terminology & background
- Methodological challenges

Linear Regression

Hypothesis Testing

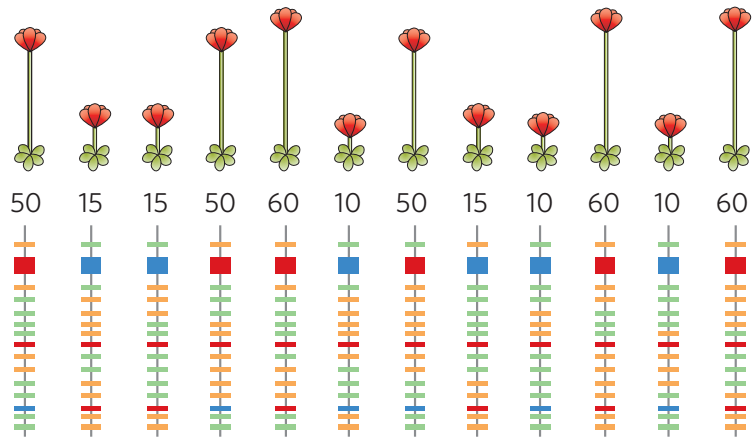
Multiple Hypothesis Testing

Model Checking

Genotype to phenotype mapping

Given:

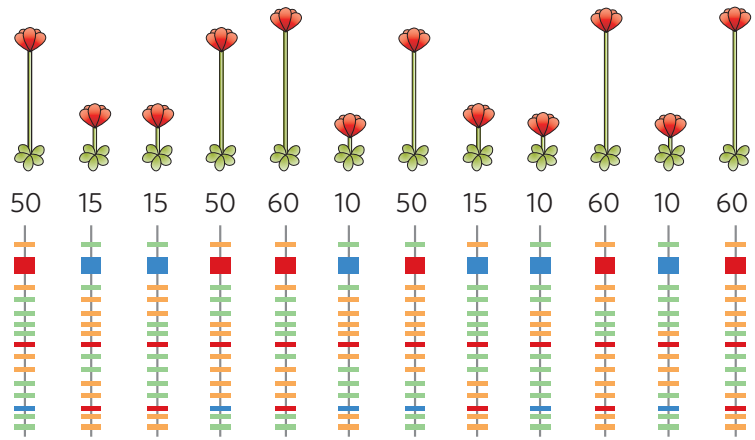
- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, . . .



Genotype to phenotype mapping

Given:

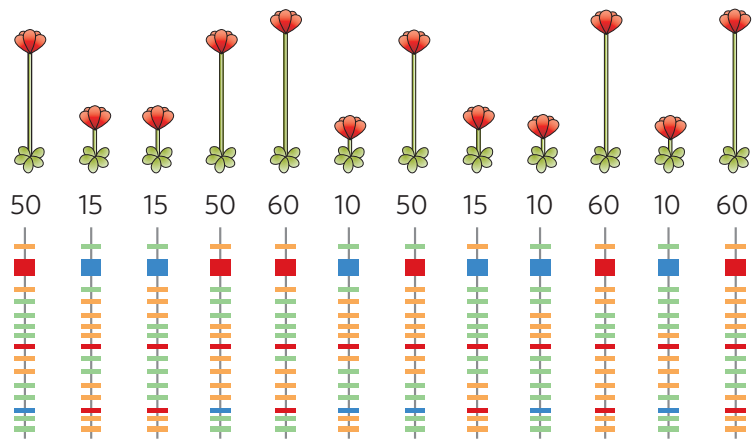
- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, . . .



Genotype to phenotype mapping

Given:

- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, ...



Genotype to phenotype mapping

Given:

- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, ...

Goal:

- Identify causal loci that explain phenotypic differences.

Genotype to phenotype mapping

Given:

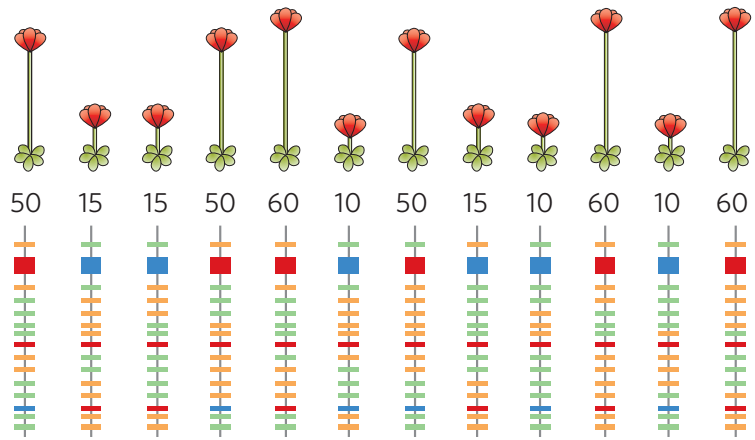
- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, ...

Goal:

- Identify causal loci that explain phenotypic differences.

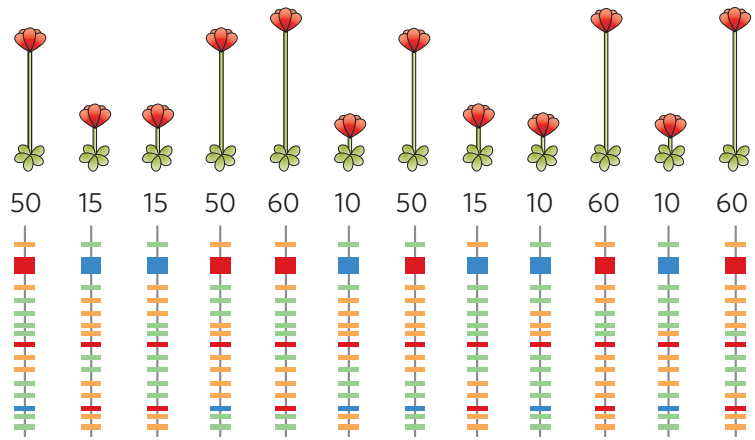
Use of GWAs in plant systems

- Basic biology
  - Understand the makeup of molecular pathways
  - Dissect the genetic component of natural variation.
  - Genotype-environment interactions
- Breeding
  - Mine for markers causal for phenotype to assist in breeding decisions.
  - Maximization of yield, pathogene resistance, etc.



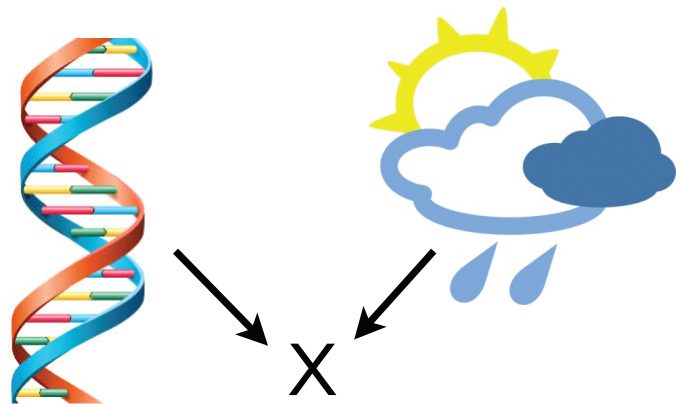
Use of GWAs in plant systems

- Basic biology
  - Understand the makeup of molecular pathways
  - Dissect the genetic component of natural variation.
  - Genotype-environment interactions
- Breeding
  - Mine for markers causal for phenotype to assist in breeding decisions.
  - Maximization of yield, pathogene resistance, etc.



Use of GWAs in plant systems

- Basic biology
  - Understand the makeup of molecular pathways
  - Dissect the genetic component of natural variation.
  - Genotype-environment interactions
- Breeding
  - Mine for markers causal for phenotype to assist in breeding decisions.
  - Maximization of yield, pathogene resistance, etc.



Personalized medicine & health

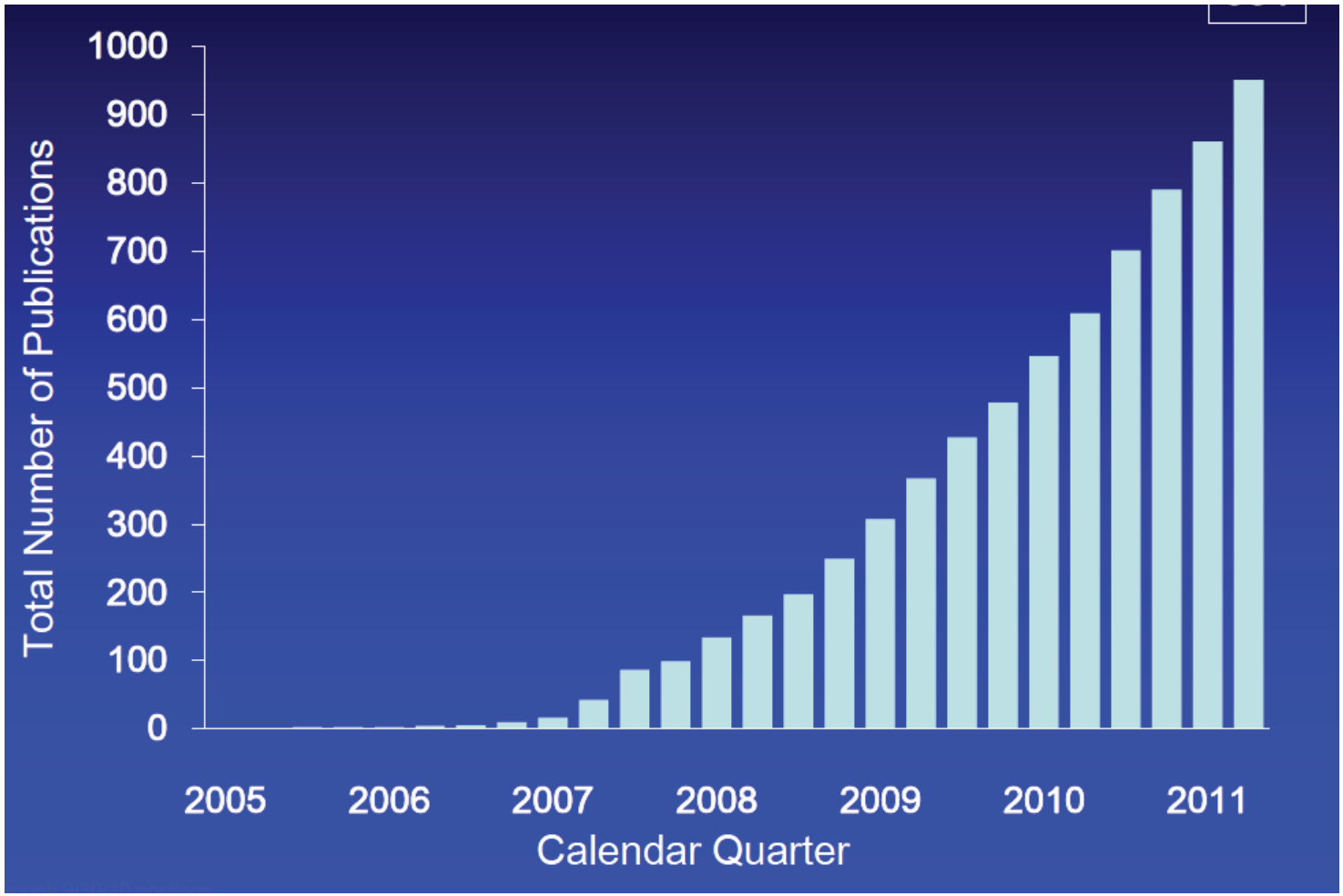
- Adapting treatment to the patients genetic make-up.
  - Targeting patients who can benefit.
  - Appropriate dosage of a drug by using genetic variants to understand drug metabolism (e.g., anti-depressants, beta blockers, opioid analgesics).
  - Disease subcategorization
- Risk prediction
  - Known causal variants help to identify individuals with higher risk to develop a particular disease.
  - Improved monitoring of high-risk groups.

Personalized medicine & health

- Adapting treatment to the patients genetic make-up.
  - Targeting patients who can benefit.
  - Appropriate dosage of a drug by using genetic variants to understand drug metabolism (e.g., anti-depressants, beta blockers, opioid analgesics).
  - Disease subcategorization
- Risk prediction
  - Known causal variants help to identify individuals with higher risk to develop a particular disease.
  - Improved monitoring of high-risk groups.

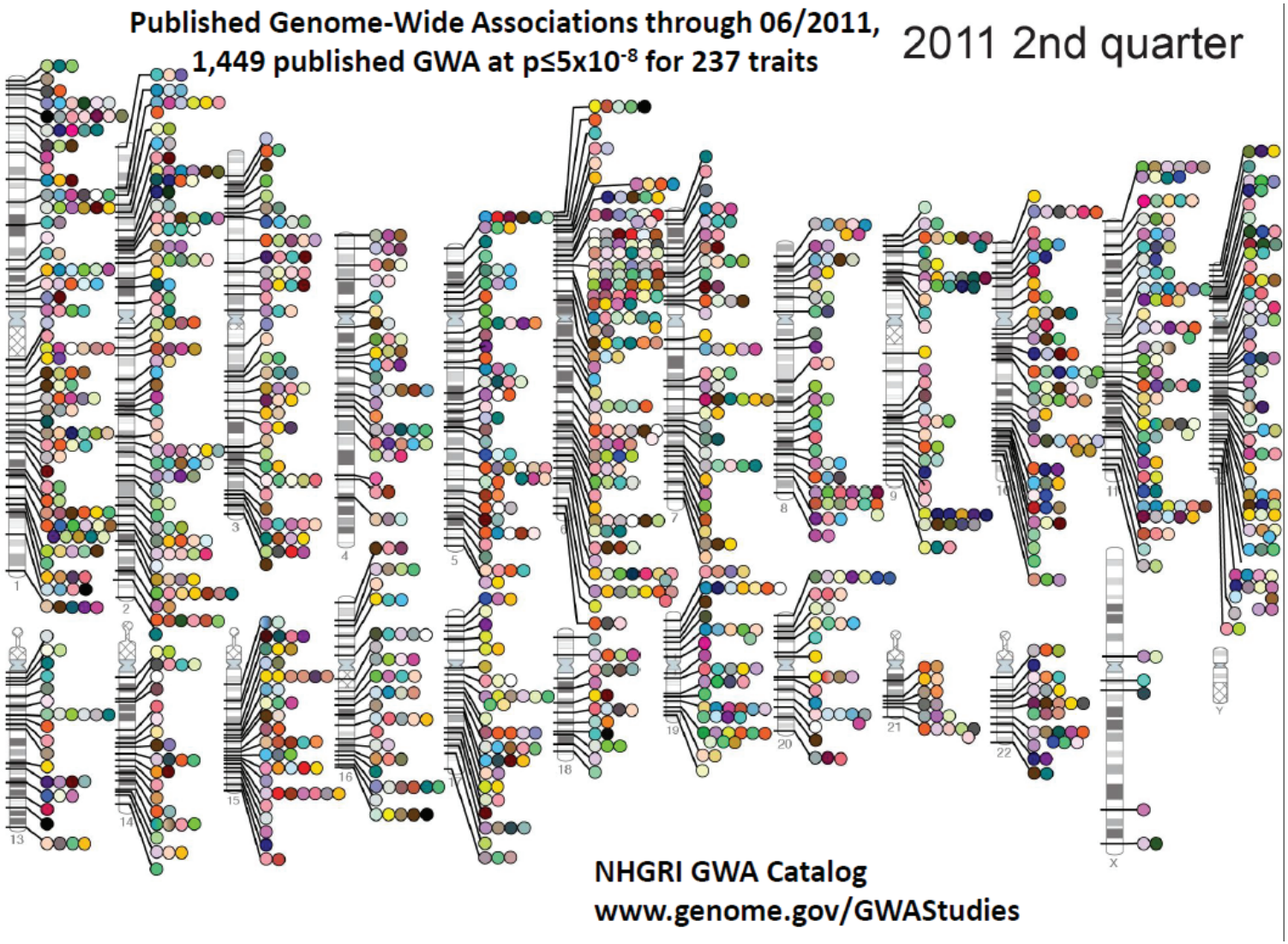
Personalized medicine & health

Publication boost



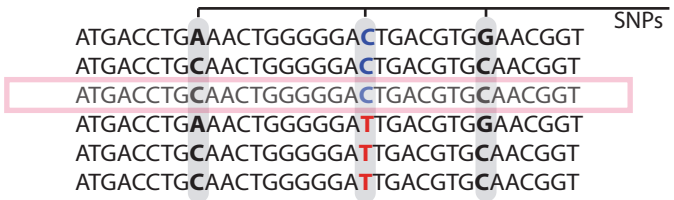
Personalized medicine & health

Publication boost



Some definitions

- Genotype denotes the genetic state of an individual.
  - Denoted by $\mathbf{x}^n$ for individual n.
- Phenotype denotes the state of a trait of an individual.
  - Denoted by $\mathbf{y}^n$ for individual n.
- A locus is a position or limited region in the genome.
  - Denoted by $\mathbf{x}_s$ for locus (or SNP) s.
- An allele is the genetic state of a locus.



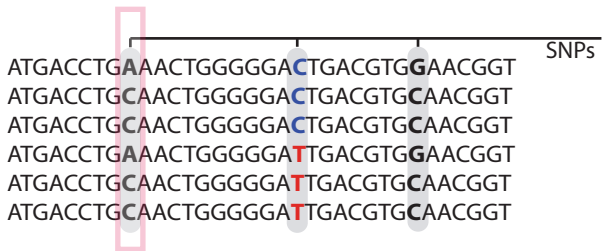
Some definitions

- Genotype denotes the genetic state of an individual.
  - Denoted by $\mathbf{x}^n$ for individual n.
- Phenotype denotes the state of a trait of an individual.
  - Denoted by $\mathbf{y}^n$ for individual n.
- A locus is a position or limited region in the genome.
  - Denoted by $\mathbf{x}_s$ for locus (or SNP) s.
- An allele is the genetic state of a locus.

image source: Wikipedia

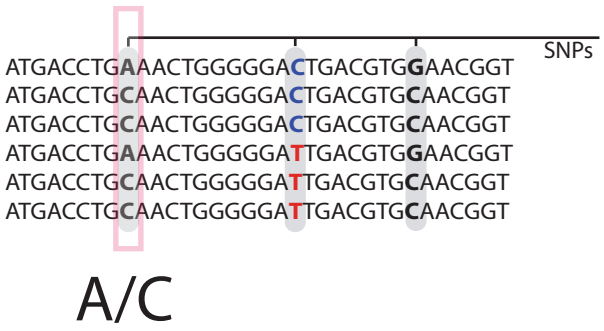
Some definitions

- Genotype denotes the genetic state of an individual.
  - Denoted by $\mathbf{x}^n$ for individual n.
- Phenotype denotes the state of a trait of an individual.
  - Denoted by $\mathbf{y}^n$ for individual n.
- A locus is a position or limited region in the genome.
  - Denoted by $\mathbf{x}_s$ for locus (or SNP) s.
- An allele is the genetic state of a locus.



Some definitions

- Genotype denotes the genetic state of an individual.
  - Denoted by $\mathbf{x}^n$ for individual n.
- Phenotype denotes the state of a trait of an individual.
  - Denoted by $\mathbf{y}^n$ for individual n.
- A locus is a position or limited region in the genome.
  - Denoted by $\mathbf{x}_s$ for locus (or SNP) s.
- An allele is the genetic state of a locus.



More definitions

- An organism/cell is haploid if it only has one chromosome set or identical chromosome sets.
  - e.g. *A. thaliana*, sperm cells or inbred lab strains
- An organism/cell is diploid if it has two separately inherited homologous chromosomes.
  - e.g. *human*
- An organism/cell is polyploid if it has more than two homologous chromosomes.
  - e.g. *sugar cane* is hexaploid.

image source: Wikipedia

More definitions

- An organism/cell is haploid if it only has one chromosome set or identical chromosome sets.
  - e.g. *A. thaliana*, sperm cells or inbred lab strains
- An organism/cell is diploid if it has two separately inherited homologous chromosomes.
  - e.g. *human*
- An organism/cell is polyploid if it has more than two homologous chromosomes.
  - e.g. *sugar cane* is hexaploid.

image source: Wikipedia

More definitions

- An organism/cell is haploid if it only has one chromosome set or identical chromosome sets.
  - e.g. *A. thaliana*, sperm cells or inbred lab strains
- An organism/cell is diploid if it has two separately inherited homologous chromosomes.
  - e.g. *human*
- An organism/cell is polyploid if it has more than two homologous chromosomes.
  - e.g. *sugar cane* is hexaploid.

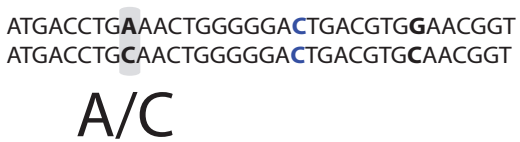
image source: Wikipedia

Even more definitions

- Haplotype denotes an individual's state of a single set of chromosomes (paternal or maternal).
- A locus is homozygous if the paternal and maternal haplotypes are identical.
- A locus is heterozygous if it differs between paternal and maternal haplotypes.

Even more definitions

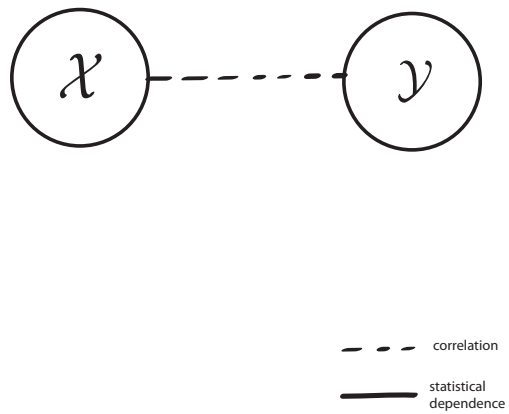
- Haplotype denotes an individual's state of a single set of chromosomes (paternal or maternal).
- A locus is homozygous if the paternal and maternal haplotypes are identical.
- A locus is heterozygous if it differs between paternal and maternal haplotypes.



Statistical association

Association is any relationship between two measured quantities that renders them statistically dependent.

- Direct association
- Indirect association
  - Can be beneficial
    - e.g.: Linkage
  - Can be harmful
    - e.g.: Population structure

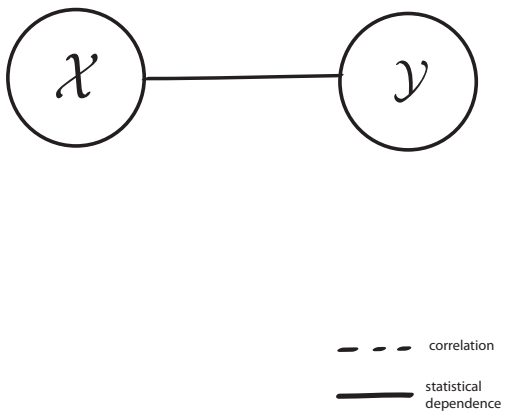


[Upton and Cook, 2002]

Statistical association

Association is any relationship between two measured quantities that renders them statistically dependent.

- Direct association
- Indirect association
  - Can be beneficial
    - e.g.: Linkage
  - Can be harmful
    - e.g.: Population structure

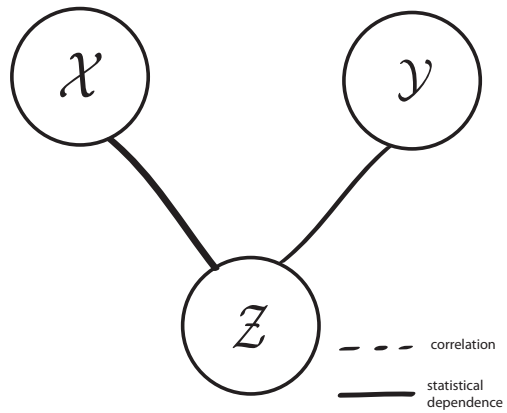


[Upton and Cook, 2002]

Statistical association

Association is any relationship between two measured quantities that renders them statistically dependent.

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure

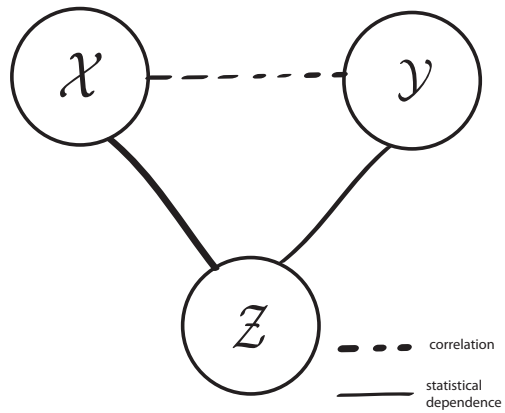


[Upton and Cook, 2002]

Statistical association

Association is any relationship between two measured quantities that renders them statistically dependent.

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure

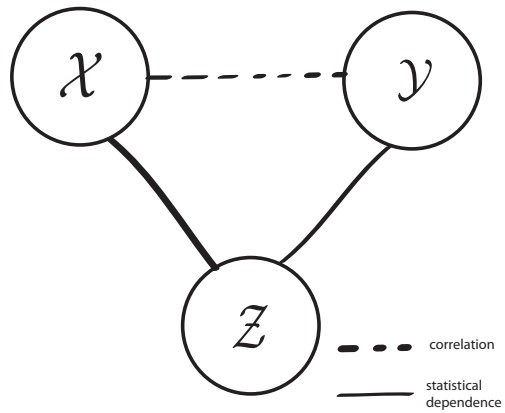


[Upton and Cook, 2002]

Statistical association

Association is any relationship between two measured quantities that renders them statistically dependent.

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure

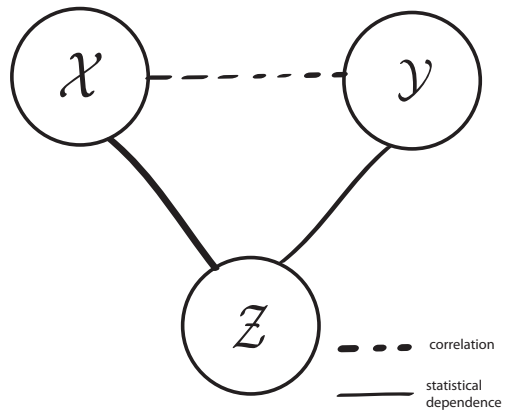


[Upton and Cook, 2002]

Statistical association

Association is any relationship between two measured quantities that renders them statistically dependent.

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure



[Upton and Cook, 2002]

Result

Example GWAS on *A. thaliana*

- Phenotype: Flowering time at 10 degrees
- Test every SNP in the genome for association with floweringtime
- Position vs. Log10(P-value) (Manhattan plot)

[Atwell et al., 2010]

Result

Example GWAS on *A. thaliana*

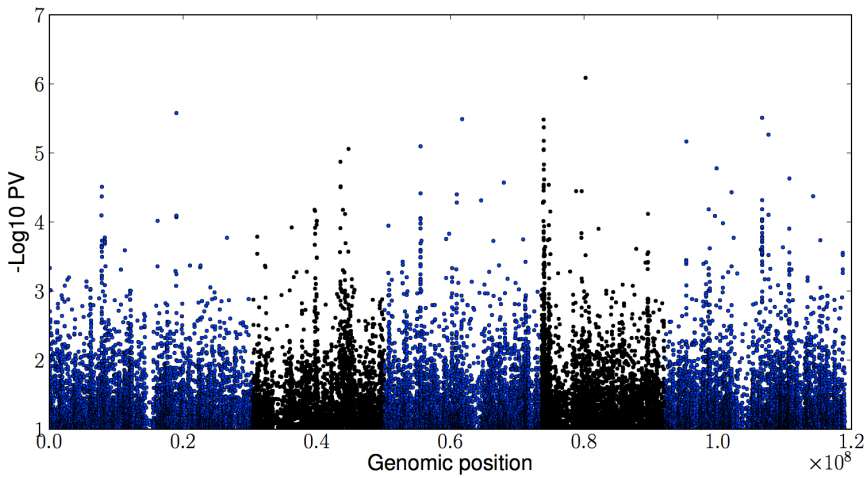
- Phenotype: Flowering time at 10 degrees
- Test every SNP in the genome for association with floweringtime
- Position vs. Log10(P-value) (Manhattan plot)

[Atwell et al., 2010]

Result

Example GWAS on *A. thaliana*

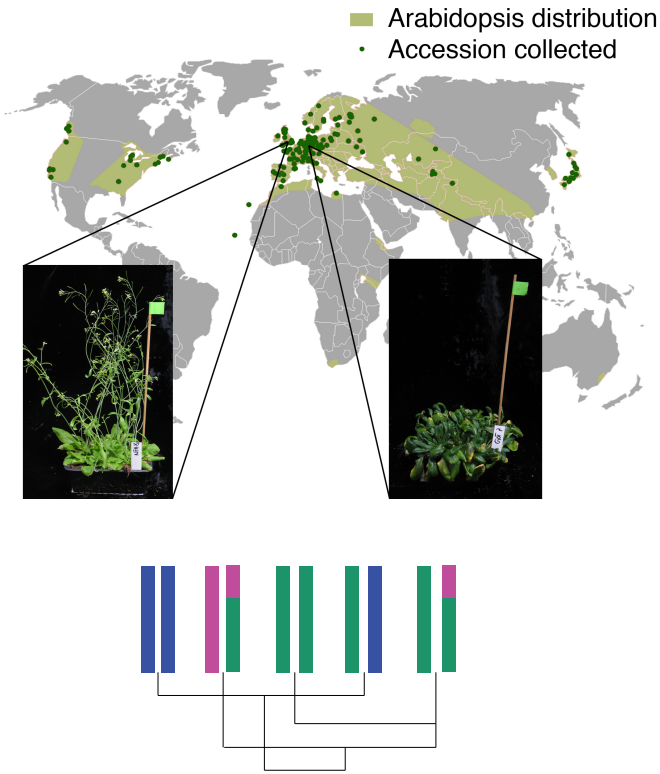
- Phenotype: Flowering time at 10 degrees
- Test every SNP in the genome for association with floweringtime
- Position vs. Log10(P-value) (Manhattan plot)



[Atwell et al., 2010]

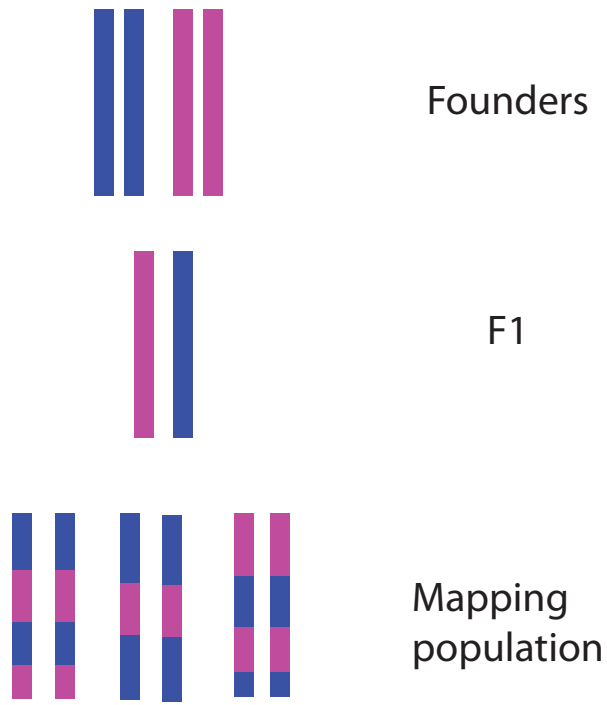
Genetic designs

- Natural population
  - Global sampling of plants, human or animals.
  - Samples may exhibit varying degrees of relatedness.
  - Typically diploid.
- Inbred F2 crosses
  - Mapping of the differences of founder strains
  - Plant- and animal systems
  - No relatedness
  - Typically haploid.
- Multi-parent crosses
  - Increased genetic diversity
  - No relatedness
  - Typically haploid.



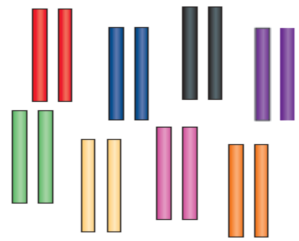
Genetic designs

- Natural population
  - Global sampling of plants, human or animals.
  - Samples may exhibit varying degrees of relatedness.
  - Typically diploid.
- Inbred F2 crosses
  - Mapping of the differences of founder strains
  - Plant- and animal systems
  - No relatedness
  - Typically haploid.
- Multi-parent crosses
  - Increased genetic diversity
  - No relatedness
  - Typically haploid.

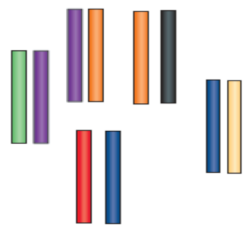


Genetic designs

- Natural population
  - Global sampling of plants, human or animals.
  - Samples may exhibit varying degrees of relatedness.
  - Typically diploid.
- Inbred F2 crosses
  - Mapping of the differences of founder strains
  - Plant- and animal systems
  - No relatedness
  - Typically haploid.
- Multi-parent crosses
  - Increased genetic diversity
  - No relatedness
  - Typically haploid.

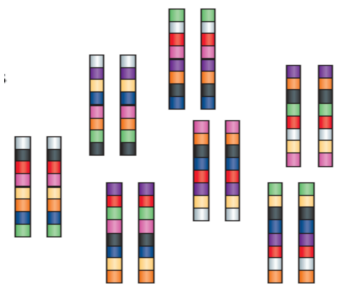


n founders



n (n-1)
F1

n (n-1)
out crosses



selfing

Genetic designs

Genotype encoding

A simple encoding scheme, ignoring dominance:

- A locus is heterozygous if it differs between paternal and maternal haplotypes.
  - heterozygous allele usually encoded as 1
- A locus is homozygous if it matches between paternal and maternal haplotypes.
  - homozygous *major* allele usually encoded as 0
  - homozygous *minor* allele usually encoded as 2

ATGACCTG**A**AACTGGGGGA**C**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGGA**C**TGACGTG**C**AACGGT

Genetic designs

Genotype encoding

A simple encoding scheme, ignoring dominance:

- A locus is heterozygous if it differs between paternal and maternal haplotypes.
  - heterozygous allele usually encoded as 1
- A locus is homozygous if it matches between paternal and maternal haplotypes.
  - homozygous *major* allele usually encoded as 0
  - homozygous *minor* allele usually encoded as 2

ATGACCTG**A**AACTGGGGGA**C**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGGA**C**TGACGTG**C**AACGGT

A/C

Genetic designs

Genotype encoding

A simple encoding scheme, ignoring dominance:

- A locus is heterozygous if it differs between paternal and maternal haplotypes.
  - heterozygous allele usually encoded as 1
- A locus is homozygous if it matches between paternal and maternal haplotypes.
  - homozygous *major* allele usually encoded as 0
  - homozygous *minor* allele usually encoded as 2

ATGACCTG**A**AACTGGGGGA**C**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGGA**C**TGACGTG**C**AACGGT

A/A

Linkage Disequilibrium

Physical linkage

- Recombination causes linkage between loci.
- Linkage is not uniform along the chromosome.
- Recombination hotspots on the chromosome lead to conserved haplotype blocks in strong linkage.
- Linkage can be used to chose *tag*-SNPs to cover all linked regions.
  - Tradeoff between resolution and genotyping cost.

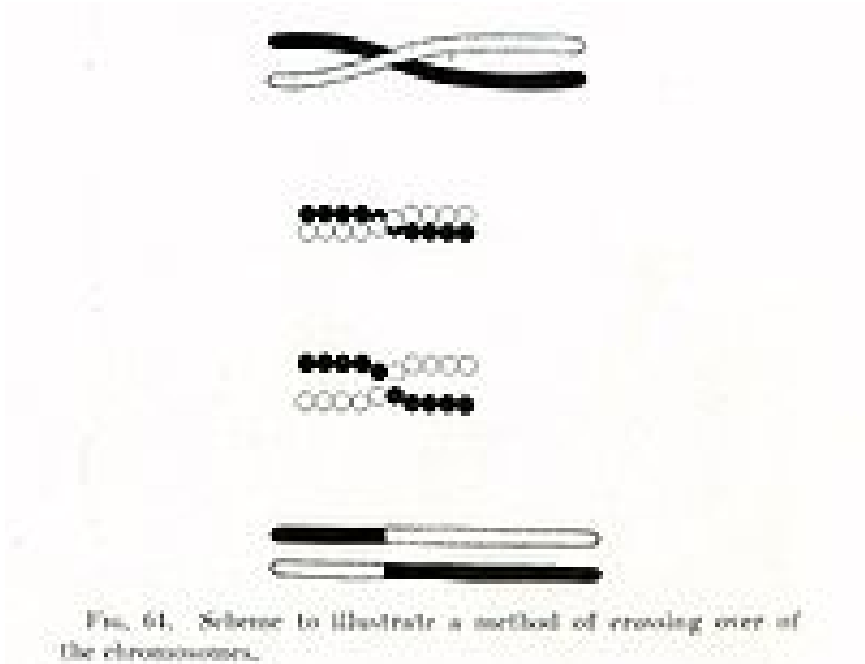


image source: Wikipedia

Linkage Disequilibrium

Physical linkage

- Recombination causes linkage between loci.
- Linkage is not uniform along the chromosome.
- Recombination hotspots on the chromosome lead to conserved haplotype blocks in strong linkage.
- Linkage can be used to chose *tag*-SNPs to cover all linked regions.
  - Tradeoff between resolution and genotyping cost.

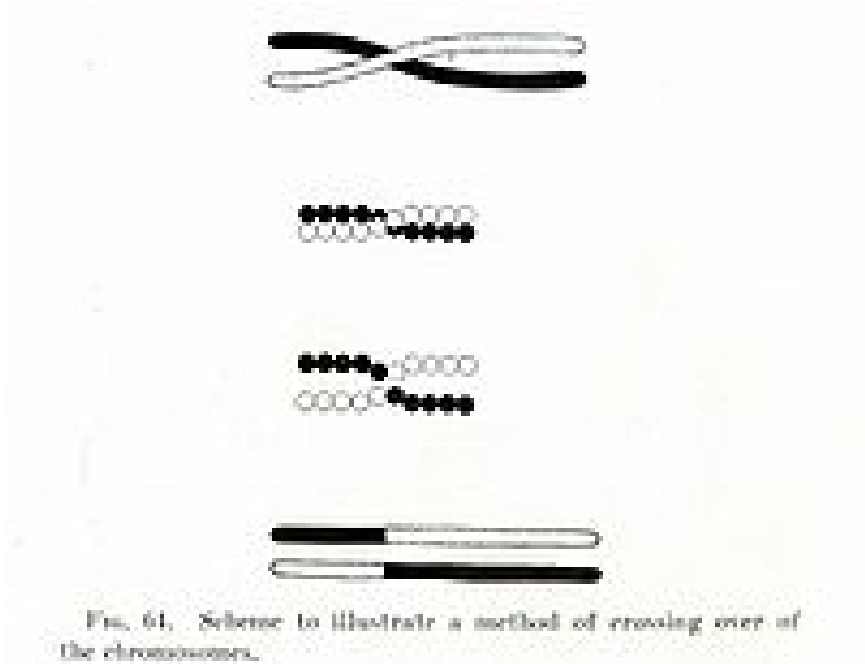


image source: Wikipedia

Linkage Disequilibrium

Physical linkage

- Recombination causes linkage between loci.
- Linkage is not uniform along the chromosome.
- Recombination hotspots on the chromosome lead to conserved haplotype blocks in strong linkage.
- Linkage can be used to chose *tag*-SNPs to cover all linked regions.
  - Tradeoff between resolution and genotyping cost.

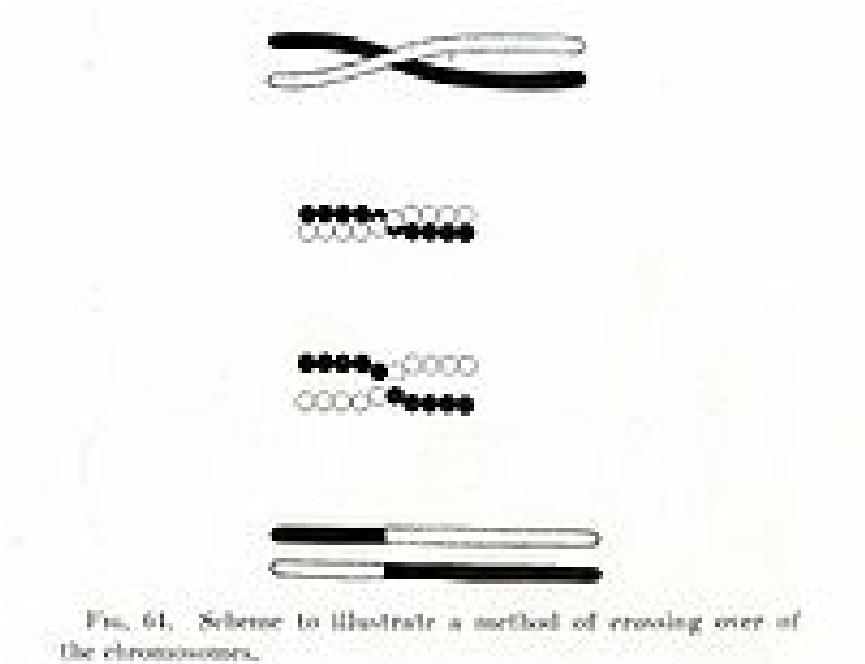
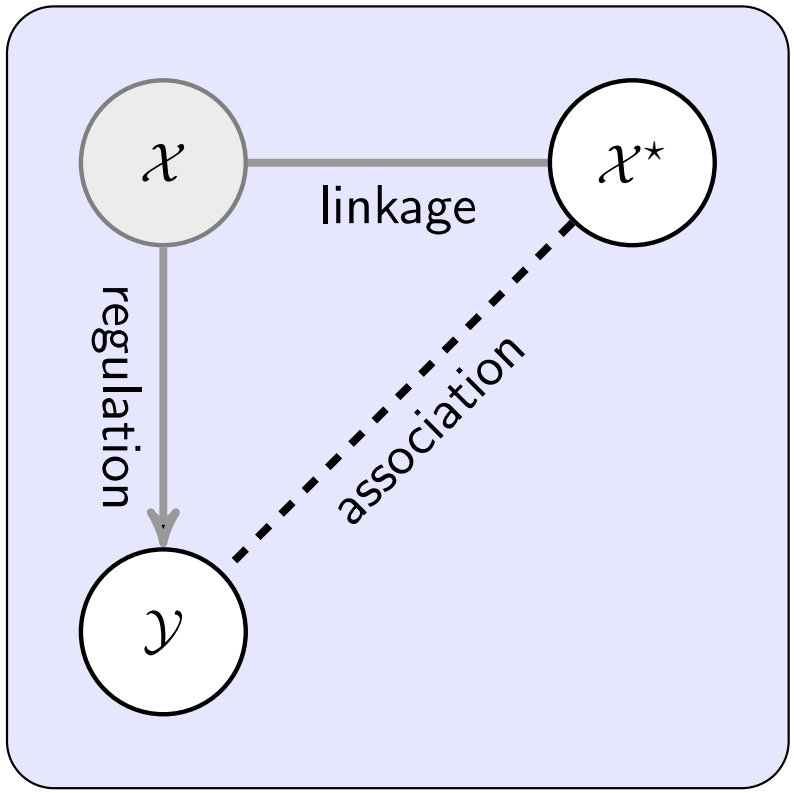


image source: Wikipedia

Linkage Disequilibrium

Physical linkage

- Recombination causes linkage between loci.
- Linkage is not uniform along the chromosome.
- Recombination hotspots on the chromosome lead to conserved haplotype blocks in strong linkage.
- Linkage can be used to chose *tag*-SNPs to cover all linked regions.
  - Tradeoff between resolution and genotyping cost.



Phenotypes

- Binary
  - case, control
  - e.g. disease status
- Continuous
  - Gaussian
    - height
  - Non-Gaussian
    - survival time, cell counts
    - gene-expression
- Multivariate
  - Images, videos
- Other

Phenotypes

- Binary
  - case, control
- Continuous
  - Gaussian
  - Non-Gaussian
- Multivariate
- Other

- e.g. disease status
- height
- survival time, cell counts
- gene-expression
- Images, videos

Phenotypes

- Binary
  - case, control
- Continuous
  - Gaussian
  - Non-Gaussian
- Multivariate
- Other

- e.g. disease status
- height
- survival time, cell counts
- gene-expression
- Images, videos

Phenotypes

- Binary
  - case, control
- Continuous
  - Gaussian
  - Non-Gaussian
- Multivariate
- Other

- e.g. disease status
- height
- survival time, cell counts
- gene-expression
- Images, videos

Phenotypes

- Binary
  - case, control
- Continuous
  - Gaussian
  - Non-Gaussian
- Multivariate
- Other

- e.g. disease status
- height
- survival time, cell counts
- gene-expression
- Images, videos

Phenotypes

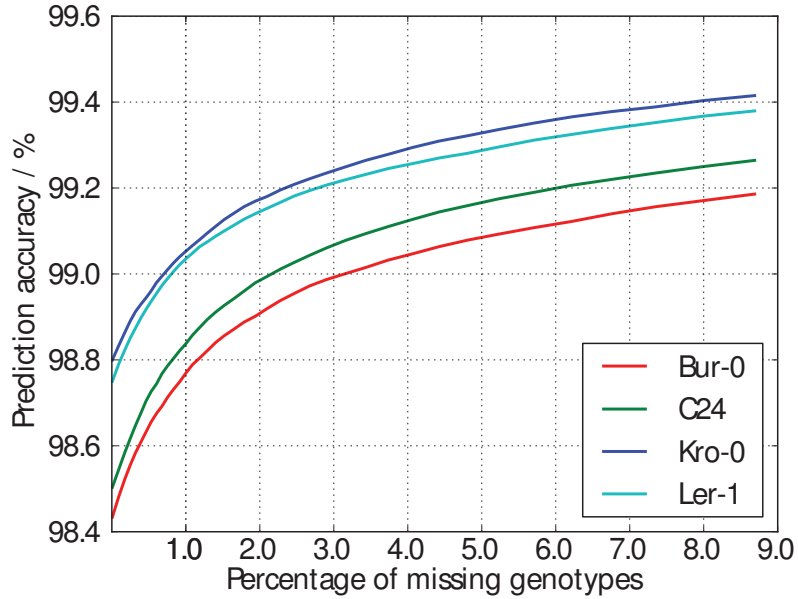
- Binary
  - case, control
- Continuous
  - Gaussian
  - Non-Gaussian
- Multivariate
- Other

- e.g. disease status
- height
- survival time, cell counts
- gene-expression
- Images, videos

Preprocessing

Genotype

- Imputation of missing values
  - Hidden Markov Models and related approaches
  - Beagle, IMPUTE
- In GWAS based on full sequencing data, some alleles may be rare or even private.
  - Model designs need to be adapted
  - Rare variances filtered out



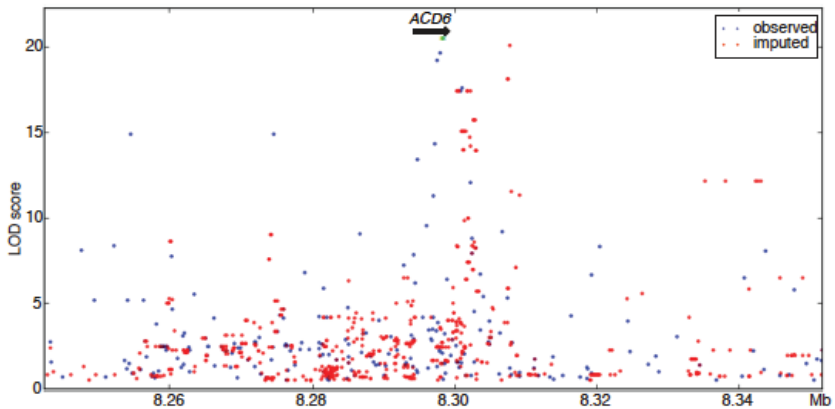
Genotype imputation accuracy from SNP-chip to 80Genomes reference panel [Cao et al., 2011].

[Browning and Browning, 2009]

Preprocessing

Genotype

- Imputation of missing values
  - Hidden Markov Models and related approaches
  - Beagle, IMPUTE
- In GWAS based on full sequencing data, some alleles may be rare or even private.
  - Model designs need to be adapted
  - Rare variances filtered out



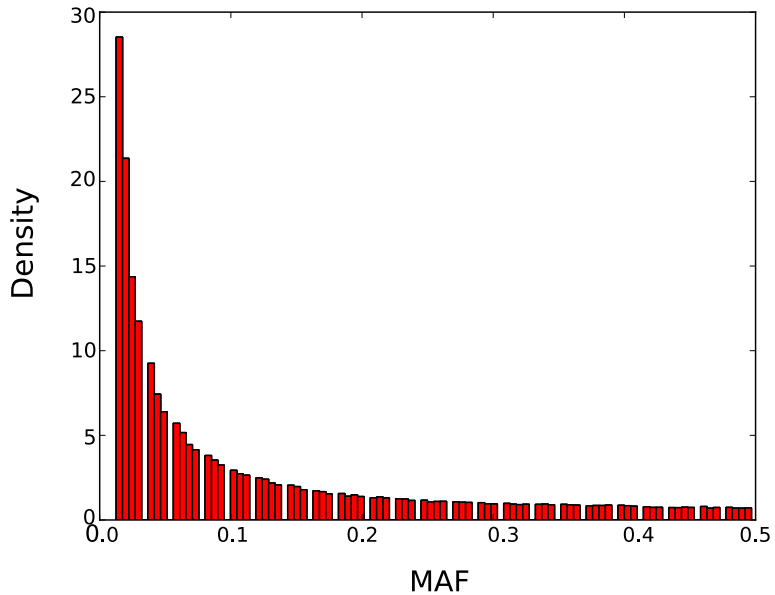
Genotype imputation accuracy from SNP-chip to 80Genomes reference panel [Cao et al., 2011].

[Browning and Browning, 2009]

Preprocessing

Genotype

- Imputation of missing values
  - Hidden Markov Models and related approaches
  - Beagle, IMPUTE
- In GWAS based on full sequencing data, some alleles may be rare or even private.
  - Model designs need to be adapted
  - Rare variances filtered out



Minor allele frequency from 160 *A. thaliana* lines; 2.3 million genome-wide SNPs from NGS sequencing

[Browning and Browning, 2009]

Preprocessing

Phenotype

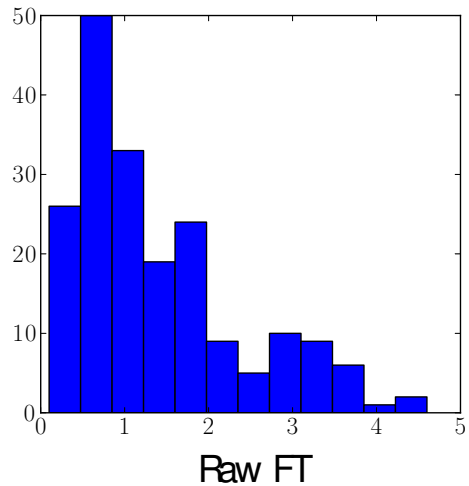
- Most parametric models are based on Gaussianity assumptions
- Phenotype residuals are often non-Gaussian
- Phenotype transformation on suitable scale
  - Use of prior knowledge
    - Growth rates, generation doubling time, etc.
  - Variance stabilization
  - Box-Cox transformation

[Spitzer, 1982]

Preprocessing

Phenotype

- Most parametric models are based on Gaussianity assumptions
- Phenotype residuals are often non-Gaussian
- Phenotype transformation on suitable scale
  - Use of prior knowledge
    - Growth rates, generation doubling time, etc.
  - Variance stabilization
  - Box-Cox transformation



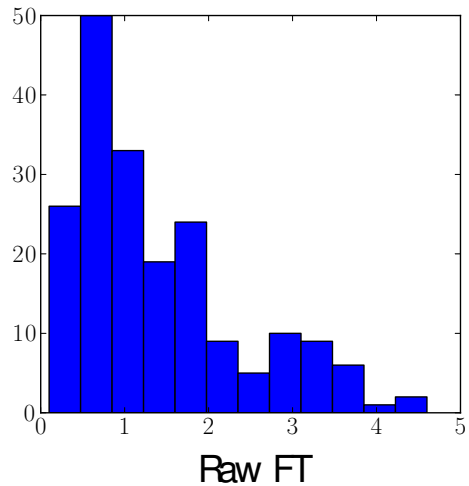
Raw and Box-Cox transformed flowering phenotypes at 10C [Atwell et al., 2010].

[Spitzer, 1982]

Preprocessing

Phenotype

- Most parametric models are based on Gaussianity assumptions
- Phenotype residuals are often non-Gaussian
- Phenotype transformation on suitable scale
  - Use of prior knowledge
    - Growth rates, generation doubling time, etc.
  - Variance stabilization
  - Box-Cox transformation



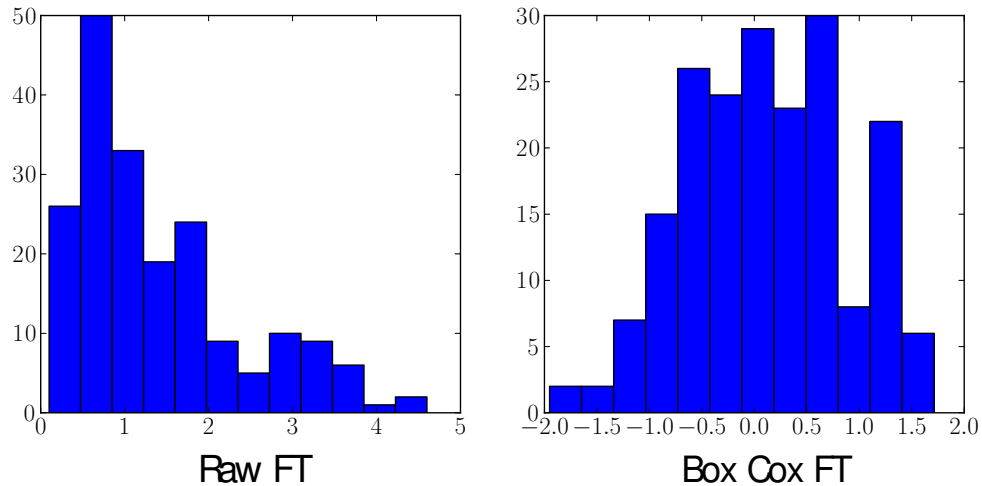
Raw and Box-Cox transformed flowering phenotypes at 10C [Atwell et al., 2010].

[Spitzer, 1982]

Preprocessing

Phenotype

- Most parametric models are based on Gaussianity assumptions
- Phenotype residuals are often non-Gaussian
- Phenotype transformation on suitable scale
  - Use of prior knowledge
    - Growth rates, generation doubling time, etc.
  - Variance stabilization
  - Box-Cox transformation



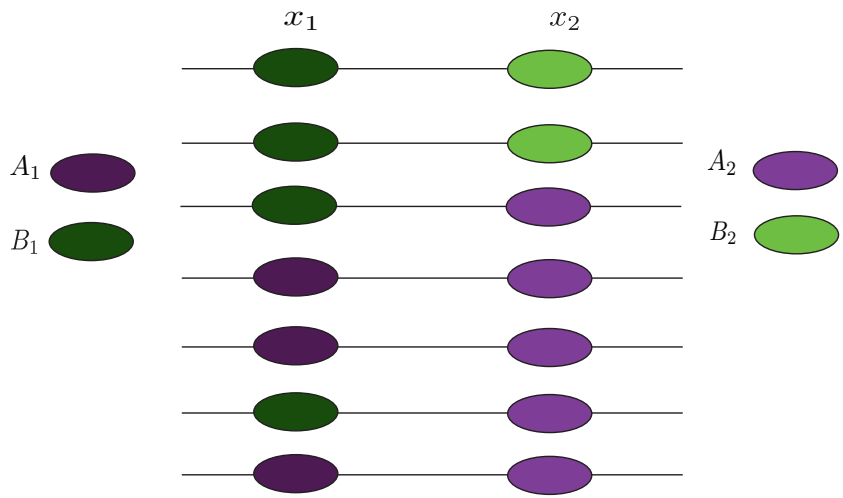
Raw and Box-Cox transformed flowering phenotypes at 10C [Atwell et al., 2010].

[Spitzer, 1982]

Linkage Disequilibrium

Gametic Phase Disequilibrium

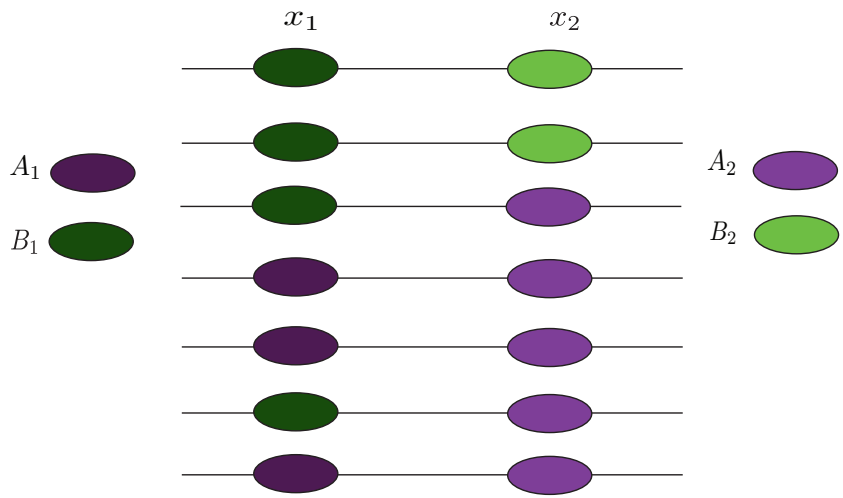
- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AB} - f_A f_B$
  - $r^2 = \frac{D^2}{f_A f_a f_B f_b}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.



Linkage Disequilibrium

Gametic Phase Disequilibrium

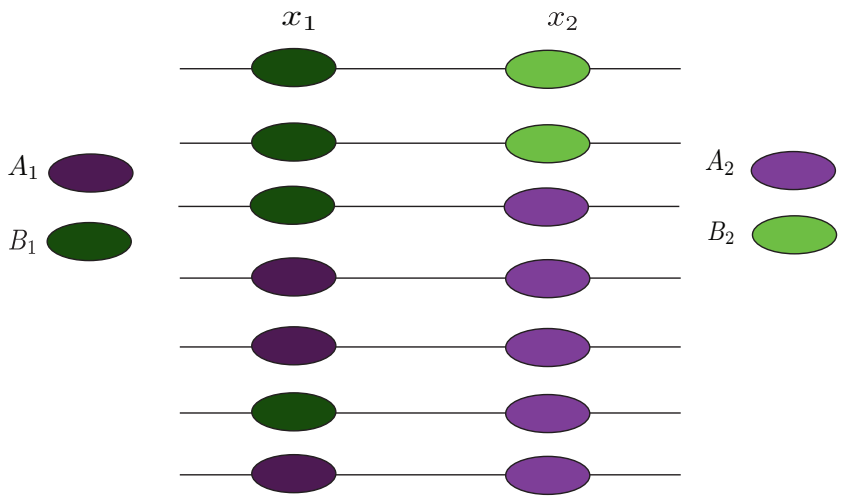
- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AB} - f_A f_B$
  - $r^2 = \frac{D}{f_A f_a f_B f_b}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.



Linkage Disequilibrium

Gametic Phase Disequilibrium

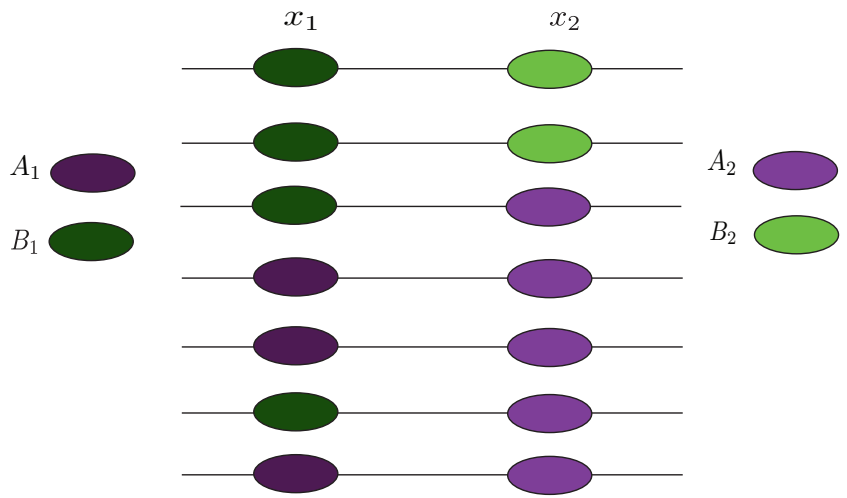
- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_A f_A$.
  - $r^2 = \frac{D^2}{f_A f_a f_B f_b}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.



Linkage Disequilibrium

Gametic Phase Disequilibrium

- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_{.A} f_{A.}$
  - $r^2 = \frac{D^2}{f_{AA} f_{AB} f_{BA} f_{BB}}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.

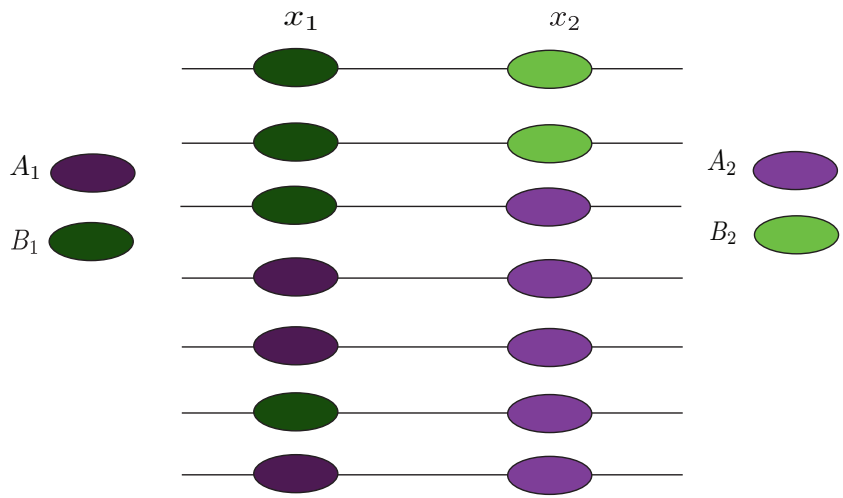


	$x_2 = A_2$	$x_2 = B_2$	
$x_1 = A_1$	f_{AA}	f_{AB}	$f_{A.}$
$x_1 = B_1$	f_{BA}	f_{BB}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

Linkage Disequilibrium

Gametic Phase Disequilibrium

- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_{.A}f_{A.}$.
  - $r^2 = \frac{D^2}{f_{AA}f_{AB}f_{BA}f_{BB}}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.

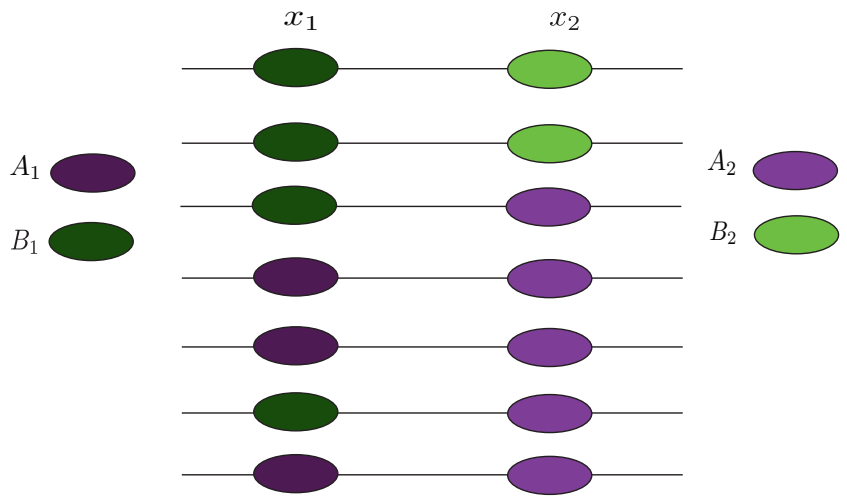


	$x_2 = A_2$	$x_2 = B_2$	
$x_1 = A_1$	f_{AA}	f_{AB}	$f_{A.}$
$x_1 = B_1$	f_{BA}	f_{BB}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

Linkage Disequilibrium

Gametic Phase Disequilibrium

- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_{.A}f_{A.}$.
  - $r^2 = \frac{D^2}{f_{AA}f_{AB}f_{BA}f_{BB}}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.

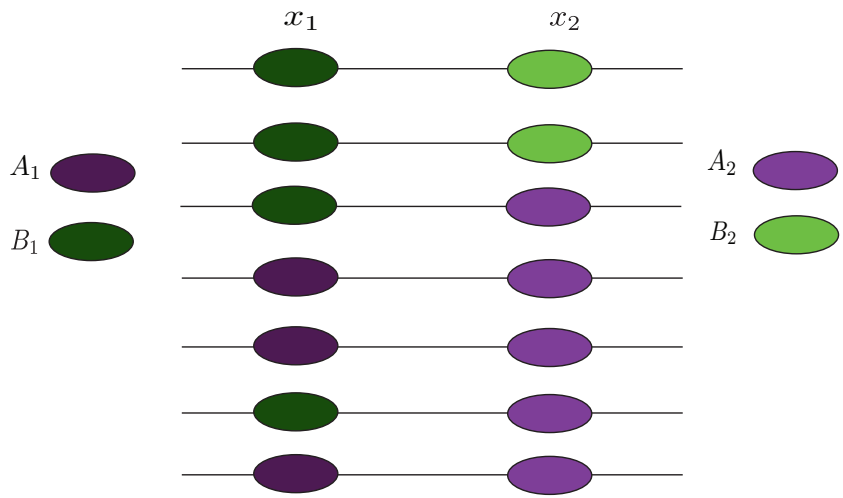


	$x_2 = A_2$	$x_2 = B_2$	
$x_1 = A_1$	f_{AA}	f_{AB}	$f_{A.}$
$x_1 = B_1$	f_{BA}	f_{BB}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

Linkage Disequilibrium

Gametic Phase Disequilibrium

- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_{.A}f_{A.}$.
  - $r^2 = \frac{D^2}{f_{AA}f_{AB}f_{BA}f_{BB}}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.

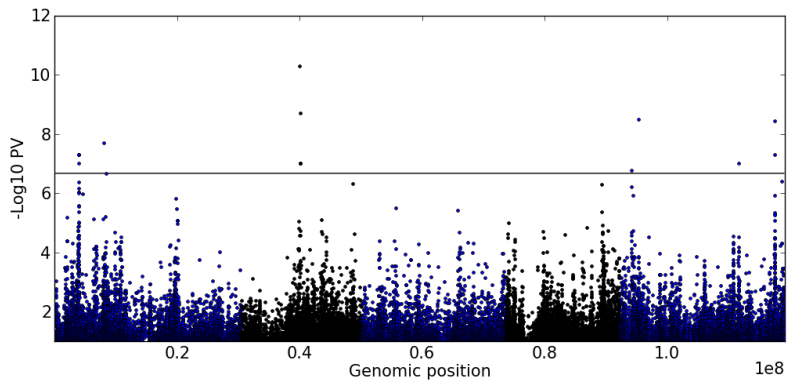


	$x_2 = A_2$	$x_2 = B_2$	
$x_1 = A_1$	f_{AA}	f_{AB}	$f_{A.}$
$x_1 = B_1$	f_{BA}	f_{BB}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

Challenges we are going to address

Multiple hypothesis testing

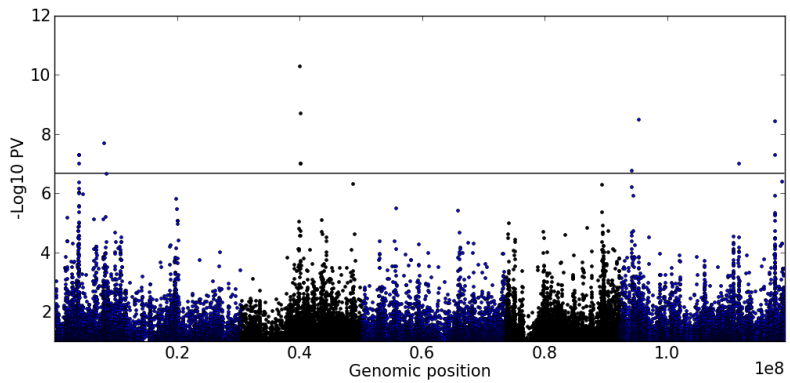
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges we are going to address

Multiple hypothesis testing

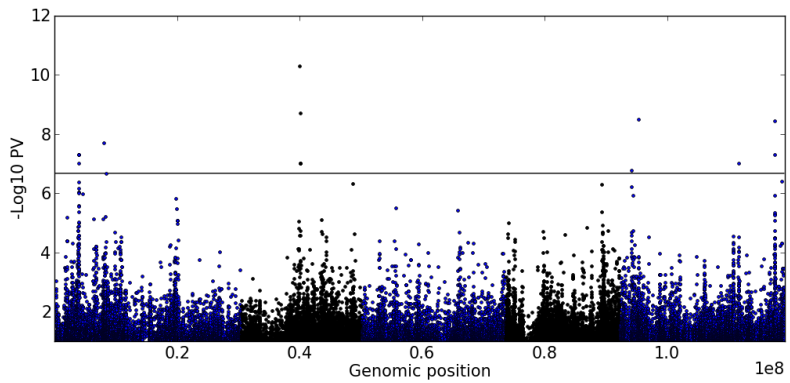
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges we are going to address

Multiple hypothesis testing

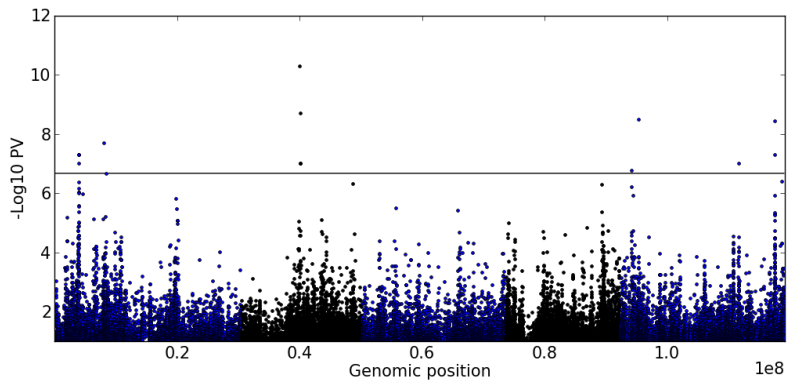
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges we are going to address

Multiple hypothesis testing

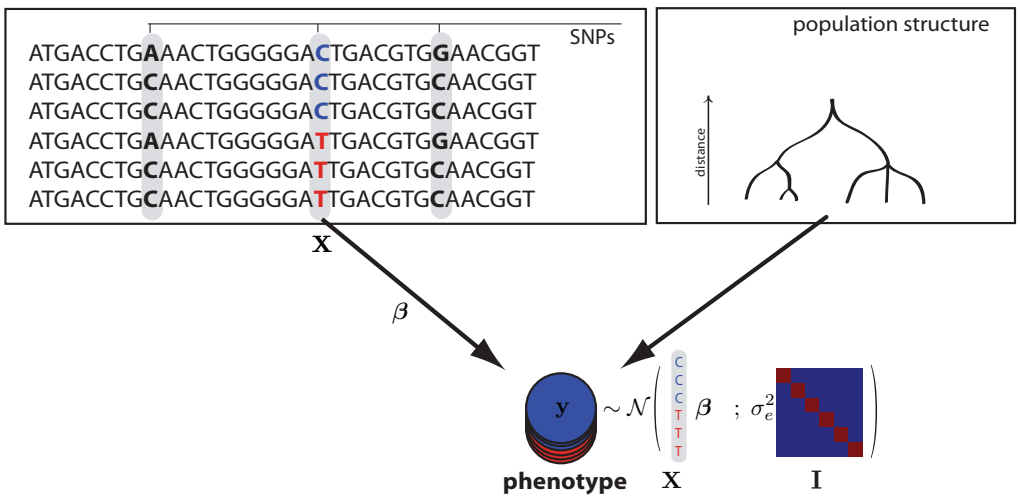
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges we are going to address

Population structure

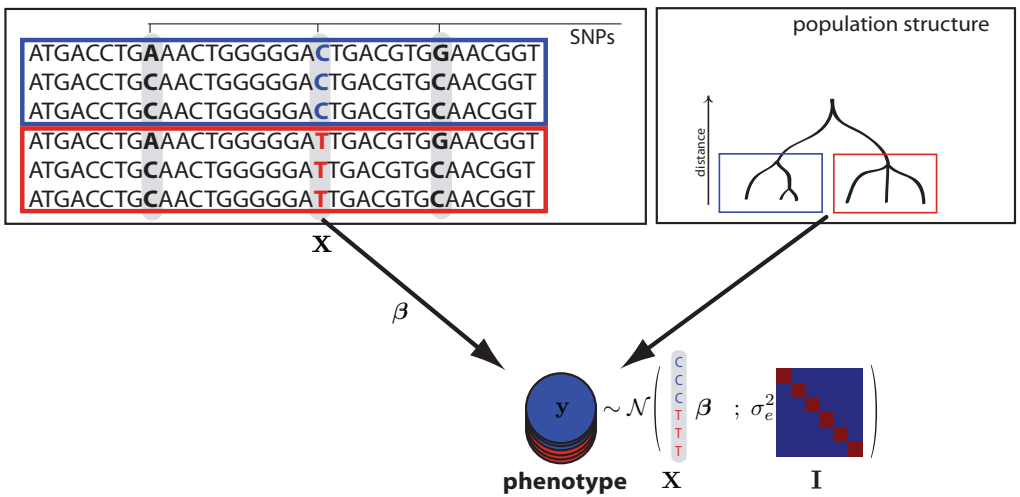
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness



Challenges we are going to address

Population structure

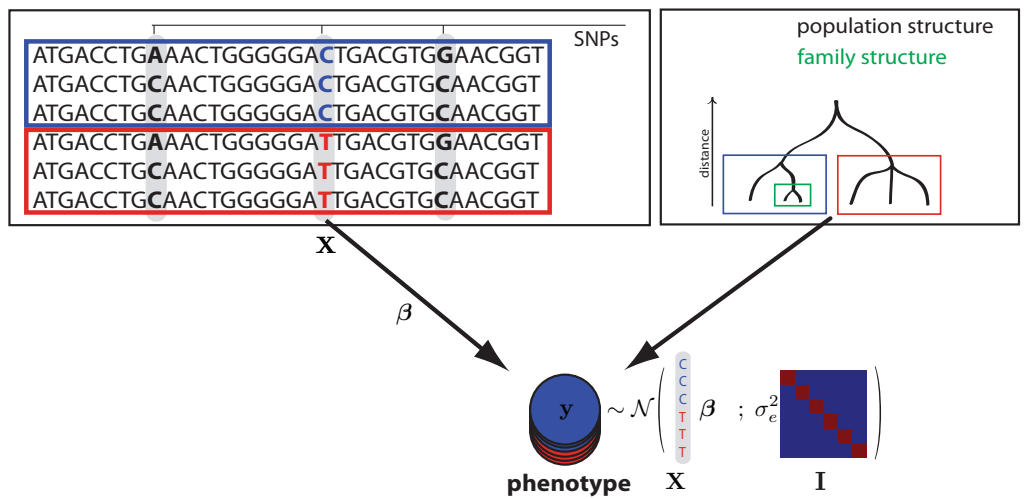
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness



Challenges we are going to address

Population structure

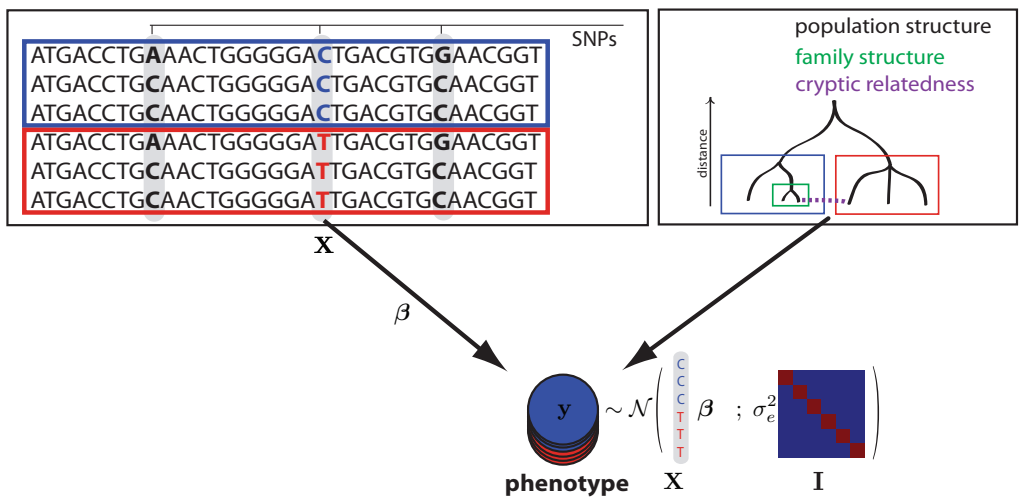
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness



Challenges we are going to address

Population structure

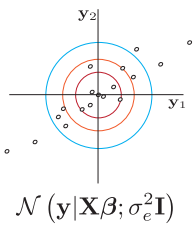
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness

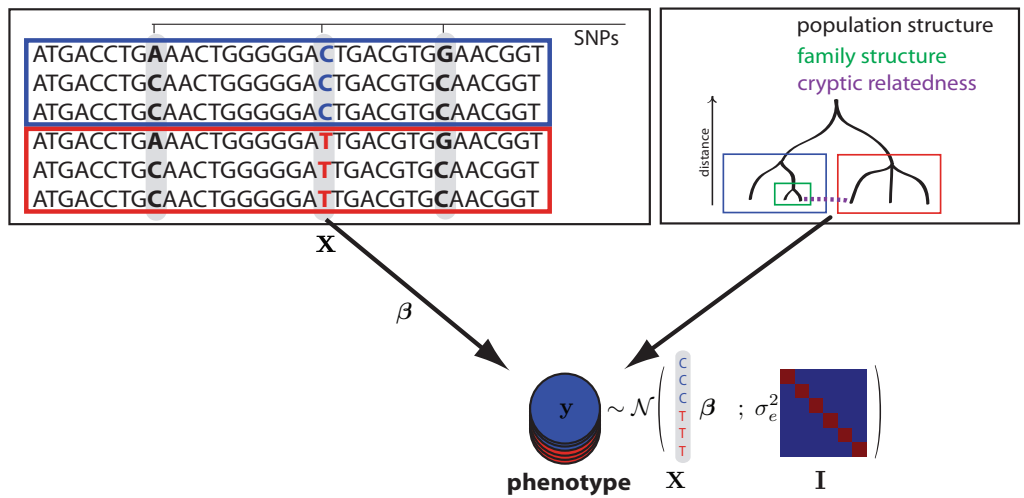


Challenges we are going to address

Population structure

- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness





Challenges we are going to address

Population structure

- GWA on inflammatory bowel disease (WTCCC)
- 3.4k cases, 11.9k controls
- Methods
  - Linear regression
  - Likelihood ratio test

Challenges we are going to address

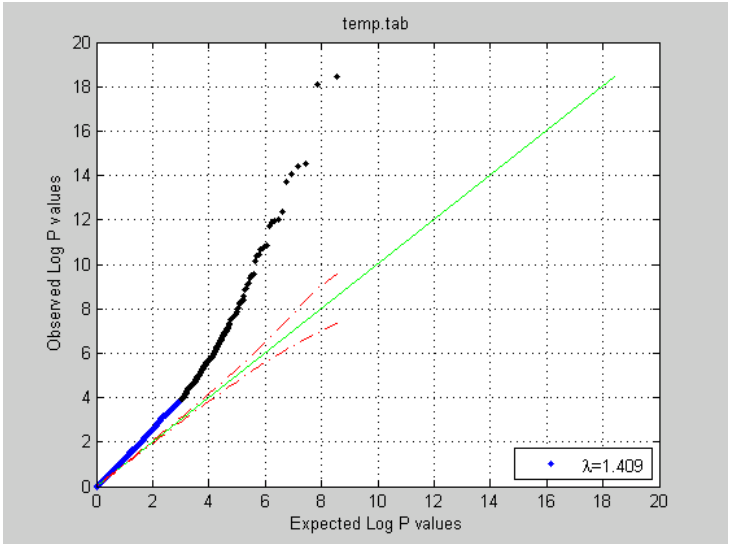
Population structure

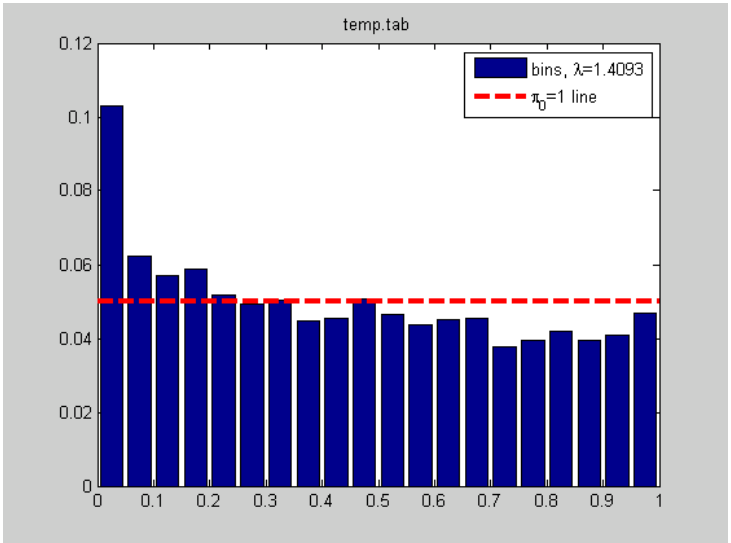
- GWA on inflammatory bowel disease (WTCCC)
- 3.4k cases, 11.9k controls
- Methods
  - Linear regression
  - Likelihood ratio test

Challenges we are going to address

Population structure

- GWA on inflammatory bowel disease (WTCCC)
- 3.4k cases, 11.9k controls
- Methods
  - Linear regression
  - Likelihood ratio test





Challenges we are going to address

Statistical power and resolution

- Small number of samples, large number of hypotheses
- Rare variants
- Small effect sizes
- Complex phenotypes have multiple regulators
- Increase power by
  - Conditioning on covariates and known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges we are going to address

Statistical power and resolution

- Small number of samples, large number of hypotheses
- Rare variants
- Small effect sizes
- Complex phenotypes have multiple regulators
- Increase power by
  - Conditioning on covariates and known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges we are going to address

Statistical power and resolution

- Small number of samples, large number of hypotheses
- Rare variants
- Small effect sizes
- Complex phenotypes have multiple regulators
- Increase power by
  - Conditioning on covariates and known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges we are going to address

Statistical power and resolution

- Small number of samples, large number of hypotheses
- Rare variants
- Small effect sizes
- Complex phenotypes have multiple regulators
- Increase power by
  - Conditioning on covariates and known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges we are going to address

Statistical power and resolution

- Small number of samples, large number of hypotheses
- Rare variants
- Small effect sizes
- Complex phenotypes have multiple regulators
- Increase power by
  - Conditioning on covariates and known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges we are going to address

Statistical power and resolution

- Small number of samples, large number of hypotheses
- Rare variants
- Small effect sizes
- Complex phenotypes have multiple regulators
- Increase power by
  - Conditioning on covariates and known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges we are going to address

Statistical power and resolution

- Small number of samples, large number of hypotheses
- Rare variants
- Small effect sizes
- Complex phenotypes have multiple regulators
- Increase power by
  - Conditioning on covariates and known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Outline

Outline

Introduction
- Why QTL mapping
- Terminology & background
- Methodological challenges

Linear Regression

Hypothesis Testing

Multiple Hypothesis Testing

Model Checking

Regression

Noise model and likelihood

- Given a dataset $\mathcal{D} = \{\mathbf{x}^n, y^n\}_{n=1}^N$, where $\mathbf{x}^n = \{x_1^n, \ldots, x_S^n\}$ is S dimensional, fit parameters $\boldsymbol{\theta}$ of a regressor f with added Gaussian noise:

$$y^n = f(\mathbf{x}^n; \boldsymbol{\theta}) + \epsilon^n \quad \text{where} \quad p(\epsilon \,|\, \sigma^2) = \mathcal{N}\left(\epsilon \,\middle|\, 0, \sigma^2\right).$$

- Equivalent likelihood formulation:

$$p(\mathbf{y} \,|\, \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \,\middle|\, f(\mathbf{x}^n), \sigma^2\right)$$

Regression

Choosing a regressor

- Choose f to be linear:

$$p(\mathbf{y} \,|\, \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta} + c, \sigma^2\right)$$

- Consider bias free case, $c = 0$, otherwise include an additional column of ones in each $\mathbf{x}^n$.

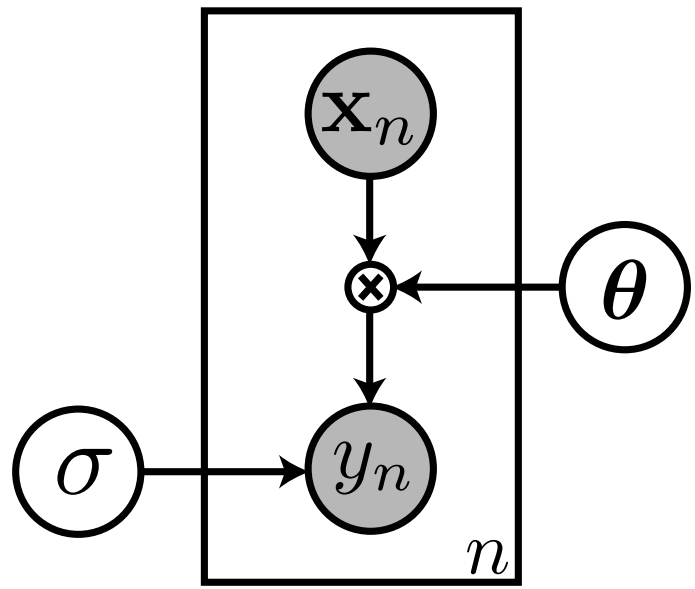
Regression

Choosing a regressor

- Choose f to be linear:

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \mid \mathbf{x}^n \cdot \boldsymbol{\theta} + c, \sigma^2\right)$$

- Consider bias free case, $c = 0$, otherwise include an additional column of ones in each $\mathbf{x}^n$.



Equivalent graphical model

Linear Regression

Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\mathbf{y} \,|\, \boldsymbol{\theta}\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}, \sigma^2\right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression

Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\mathbf{y} \,|\, \boldsymbol{\theta}\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}, \sigma^2\right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression

Maximum likelihood

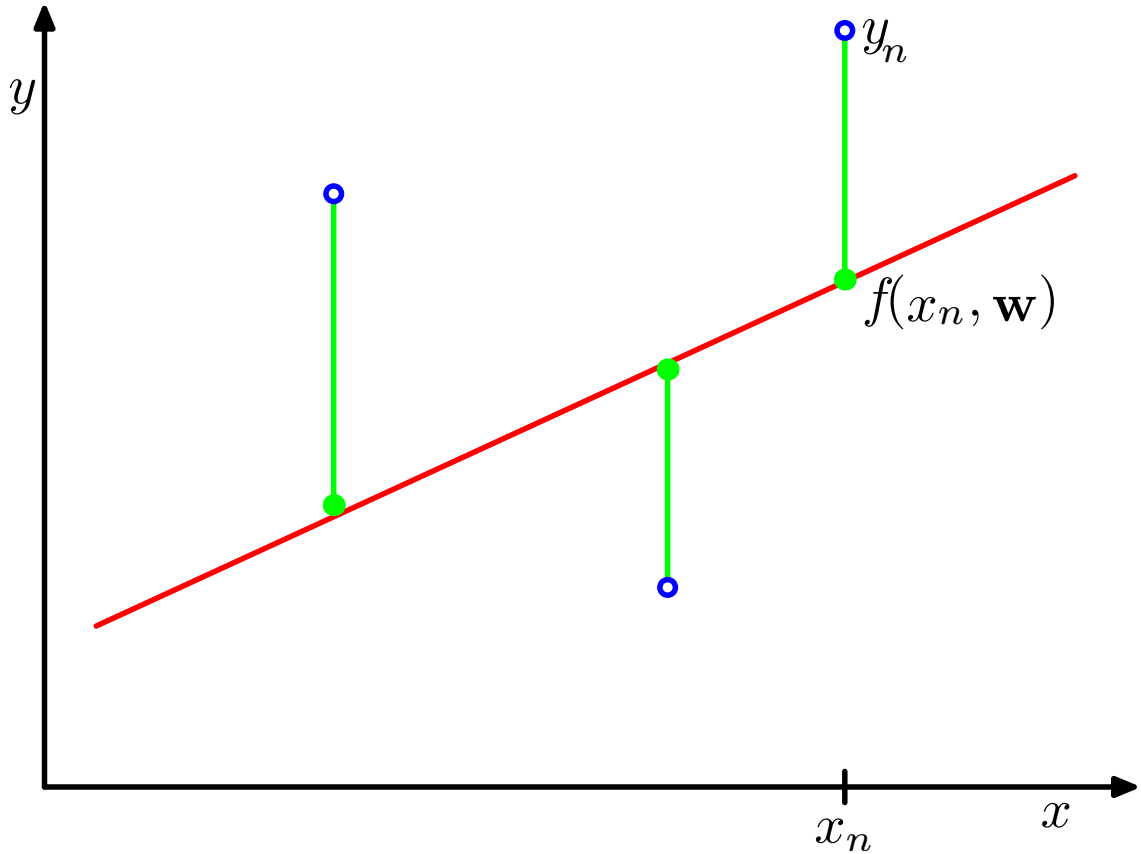
- Taking the logarithm, we obtain

$$\ln p(\mathbf{y} \,|\, \boldsymbol{\theta}\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}, \sigma^2\right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression and Least Squares



(C.M. Bishop, Pattern Recognition and Machine Learning)

$$E(\boldsymbol{\theta}) = \frac{1}{2}\sum_{n=1}^{N}(y^n - \mathbf{x}^n \cdot \boldsymbol{\theta})^2$$

Linear Regression and Least Squares

- Derivative w.r.t. a single weight entry θ_i

$$\frac{d}{\mathrm{d}\theta_i} \ln p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\theta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta})^2 \right]$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta}) x_i$$

- Set gradient w.r.t. θ to zero

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta}) \mathbf{x}^{n\mathrm{T}} = 0$$

$$\Longrightarrow \boldsymbol{\theta}_{\mathsf{ML}} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}}_{\text{Pseudo inverse}} \mathbf{y}$$

- Here, the matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{bmatrix} x_1^1 & \dots & x_S^1 \\ \dots & \dots & \dots \\ x_1^N & \dots & x_S^N \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t. a single weight entry θ_i

$$\frac{d}{\mathrm{d}\theta_i} \ln p(\mathbf{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\theta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta})^2 \right]$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta}) x_i$$

- Set gradient w.r.t. $\boldsymbol{\theta}$ to zero

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta}) \mathbf{x}^{n\mathrm{T}} = 0$$

$$\implies \boldsymbol{\theta}_{\mathsf{ML}} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}}_{\text{Pseudo inverse}} \mathbf{y}$$

- Here, the matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{bmatrix} x_1^1 & \dots & x_S^1 \\ \dots & \dots & \dots \\ x_1^N & \dots & x_S^N \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t. a single weight entry θ_i

$$\frac{d}{\mathrm{d}\theta_i} \ln p(\mathbf{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\theta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta})^2 \right]$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta}) x_i$$

- Set gradient w.r.t. $\boldsymbol{\theta}$ to zero

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y^n - \mathbf{x}^n \cdot \boldsymbol{\theta}) \mathbf{x}^{n\mathrm{T}} = 0$$

$$\Longrightarrow \boldsymbol{\theta}_{\mathsf{ML}} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}}_{\text{Pseudo inverse}} \mathbf{y}$$

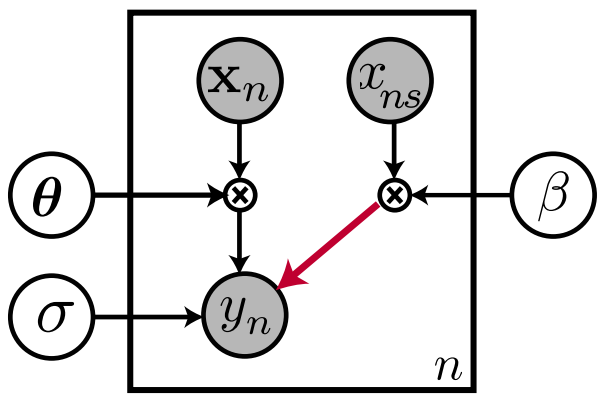
- Here, the matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{bmatrix} x_1^1 & \dots & x_S^1 \\ \dots & \dots & \dots \\ x_1^N & \dots & x_S^N \end{bmatrix}$

Testing in Linear Regression

Likelihood Ratio Test

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \mid \mathbf{x}^n \cdot \boldsymbol{\theta} + x_s^n \beta, \sigma^2\right)$$

- x_s^n: SNP to be tested
- $\mathbf{x}^n$: regression covariates (including bias term)
  - Race
  - Known background SNPs
  - Environment



Equivalent graphical model

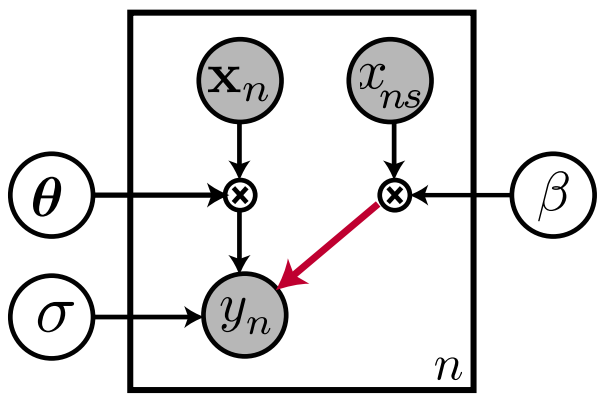
x^n: regression covariates

Testing in Linear Regression

Likelihood Ratio Test

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \mid \mathbf{x}^n \cdot \boldsymbol{\theta} + x_s^n \beta, \sigma^2\right)$$

- x_s^n: SNP to be tested
- $\mathbf{x}^n$: regression covariates (including bias term)
  - Race
  - Known background SNPs
  - Environment



Equivalent graphical model

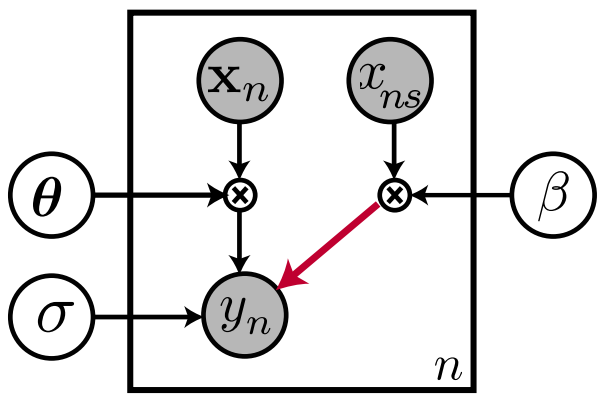
x^n: regression covariates

Testing in Linear Regression

Likelihood Ratio Test

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \mid \mathbf{x}^n \cdot \boldsymbol{\theta} + x_s^n \beta, \sigma^2\right)$$

- x_s^n: SNP to be tested
- $\mathbf{x}^n$: regression covariates (including bias term)
  - Race
  - Known background SNPs
  - Environment



Equivalent graphical model

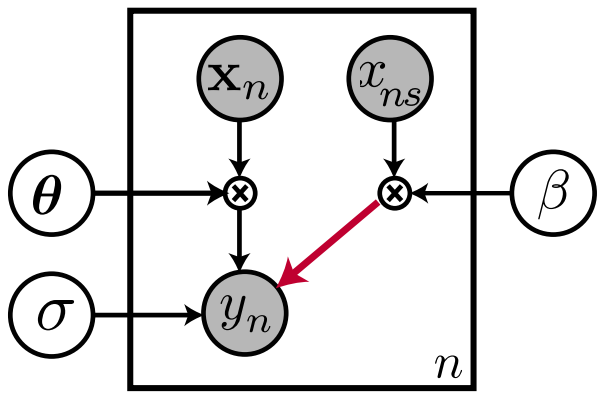
x^n: regression covariates

Testing in Linear Regression

Likelihood Ratio Test

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta} + x_s^n \beta, \sigma^2\right)$$

- Test $\mathcal{H}_0 : \beta = 0$
- The ratio of the ML estimator and the ML_0 estimator restricted to $\mathcal{H}_0$ is a common test statistic.



Equivalent graphical model

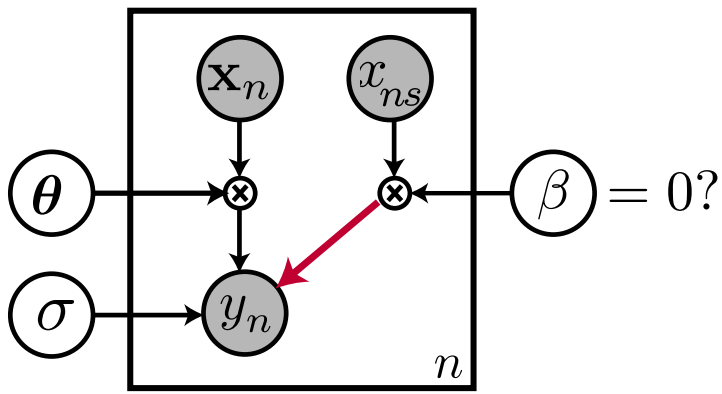
x^n: regression covariates

Testing in Linear Regression

Likelihood Ratio Test

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta} + x_s^n \beta, \sigma^2\right)$$

- Test $\mathcal{H}_0 : \beta = 0$
- The ratio of the ML estimator and the ML_0 estimator restricted to $\mathcal{H}_0$ is a common test statistic.



Equivalent graphical model

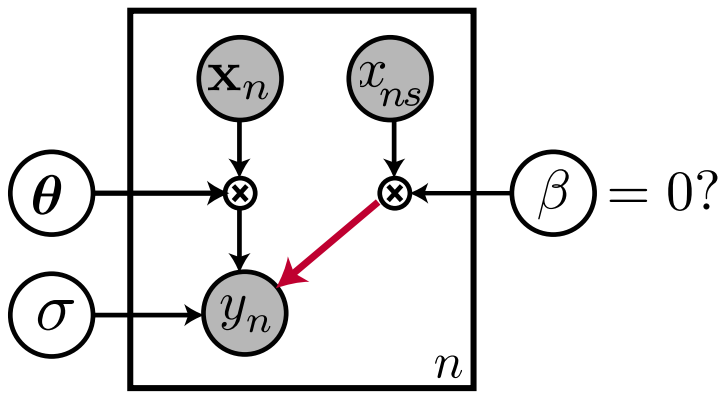
x^n: regression covariates

Testing in Linear Regression

Likelihood Ratio Test

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta} + x_s^n \beta, \sigma^2\right)$$

- Test $\mathcal{H}_0 : \beta = 0$
- The ratio of the ML estimator and the ML_0 estimator restricted to $\mathcal{H}_0$ is a common test statistic.



Equivalent graphical model

x^n: regression covariates

$$\frac{\prod_{n=1}^{N} \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\mathrm{ML}} + x_s^n \beta_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}}\right)}{\prod_{n=1}^{N} \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\mathrm{ML}_0} + x_s^n 0, \sigma^2_{\mathrm{ML}_0}\right)}$$

Outline

Introduction
- Why QTL mapping
- Terminology & background
- Methodological challenges

Linear Regression

Hypothesis Testing

Multiple Hypothesis Testing

Model Checking

Hypothesis Testing

Example:

- Given a sample $\mathcal{D} = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

Hypothesis Testing

Example:

- Given a sample $\mathcal{D} = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

Hypothesis Testing

Example:

- Given a sample $\mathcal{D} = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

Hypothesis Testing

Example:

- Given a sample $\mathcal{D} = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Hypothesis Testing

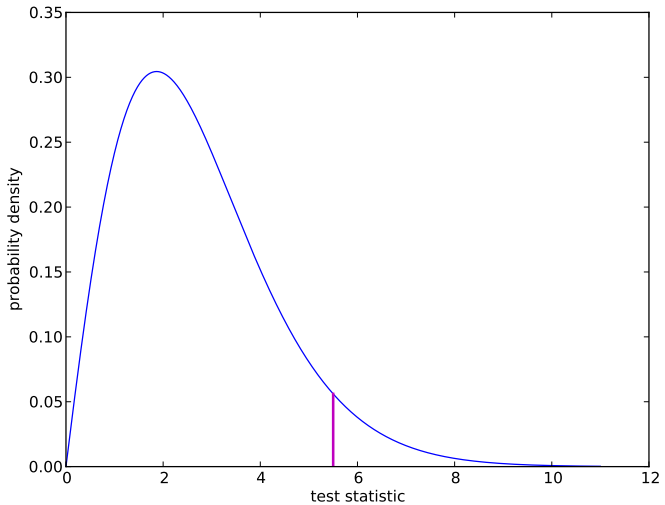
Example:

- Given a sample $\mathcal{D} = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- To show that $\beta_s \neq 0$ we can perform a statistical test that tries to reject $\mathcal{H}_0$.
- **type 1 error:** $\mathcal{H}_0$ is rejected but does hold.
- **type 2 error:** $\mathcal{H}_0$ is accepted but does not hold.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

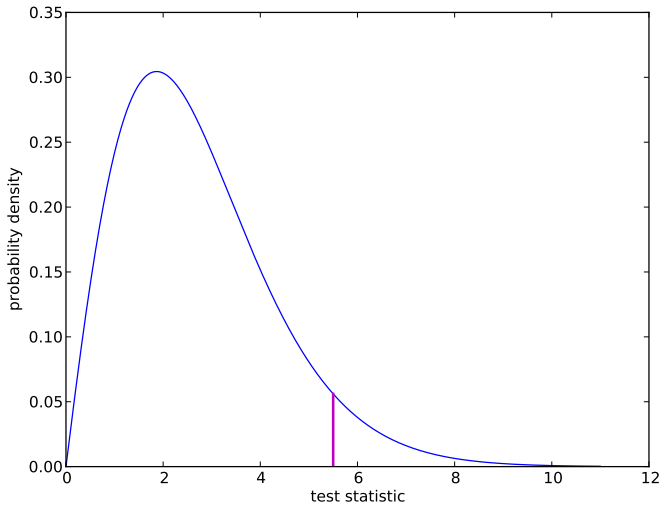
Hypothesis Testing

- Given a sample $\mathcal{D} = \{x^1, \ldots, x^N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** defines the values of the test statistic that lead to a rejection of the test.



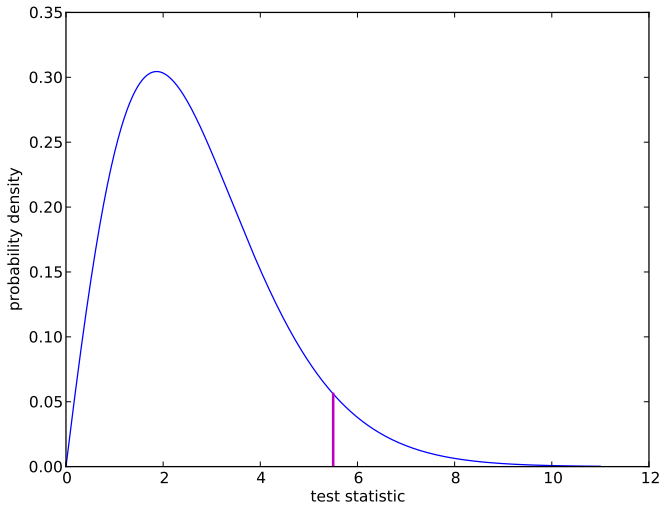
Hypothesis Testing

- Given a sample $\mathcal{D} = \{x^1, \ldots, x^N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** defines the values of the test statistic that lead to a rejection of the test.



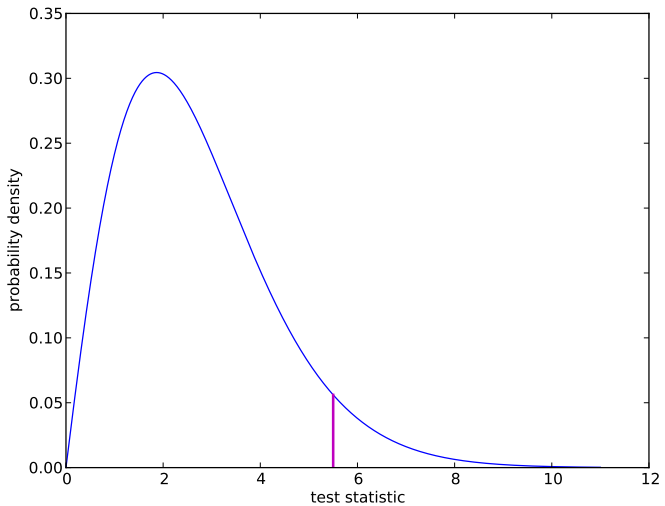
Hypothesis Testing

- Given a sample $\mathcal{D} = \{x^1, \ldots, x^N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** defines the values of the test statistic that lead to a rejection of the test.



Hypothesis Testing

- Given a sample $\mathcal{D} = \{x^1, \ldots, x^N\}$.
- Test whether $\mathcal{H}_0 : \beta_s = 0$ (null hypothesis) or $\mathcal{H}_1 : \beta_s \neq 0$ (alternative hypothesis) is true.
- The **significance level** α defines the threshold and the sensitivity of the test. This equals the probability of a type-1 error.
- Usually decision is based on a **test statistic**.
- The **critical region** defines the values of the test statistic that lead to a rejection of the test.



P-value

definition

- Probability of observing a test statistic at least as extreme (e.g. likelihood ratio statistic), given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- Possible to generate artificial null-distribution by **permutations**

P-value

definition

- Probability of observing a test statistic at least as extreme (e.g. likelihood ratio statistic), given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- Possible to generate artificial null-distribution by **permutations**

P-value

definition

- Probability of observing a test statistic at least as extreme (e.g. likelihood ratio statistic), given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- Possible to generate artificial null-distribution by **permutations**

P-value

definition

- Probability of observing a test statistic at least as extreme (e.g. likelihood ratio statistic), given that $\mathcal{H}_0$ is true.
- Significance level α becomes threshold on P-value.
- Need to know the null distribution of test statistics. (usually unknown)
- Possible to generate artificial null-distribution by **permutations**

P-value

Permutation procedure

Repeat M times:

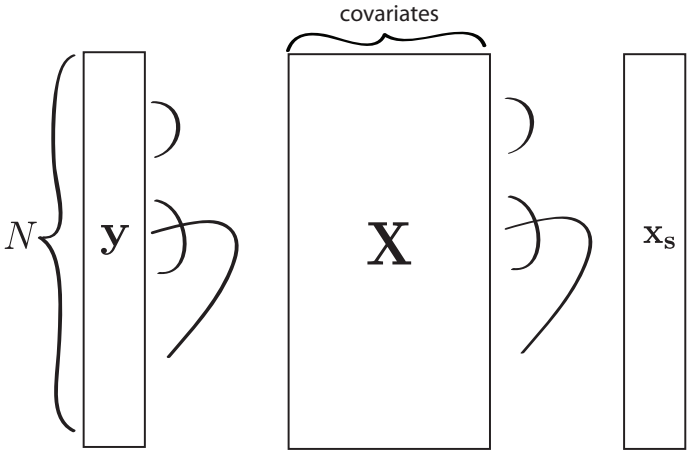
- Permute phenotype $\mathbf{y}$ and covariates $\mathbf{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution

P-value

Permutation procedure

Repeat M times:

- Permute phenotype $\mathbf{y}$ and covariates $\mathbf{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution

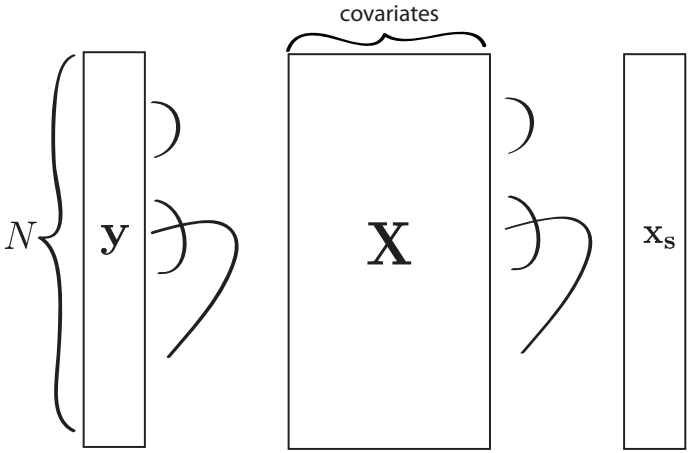


P-value

Permutation procedure

Repeat M times:

- Permute phenotype $\mathbf{y}$ and covariates $\mathbf{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution

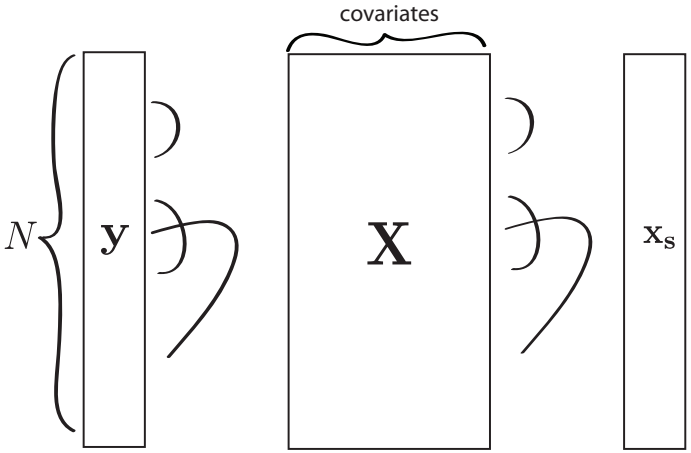


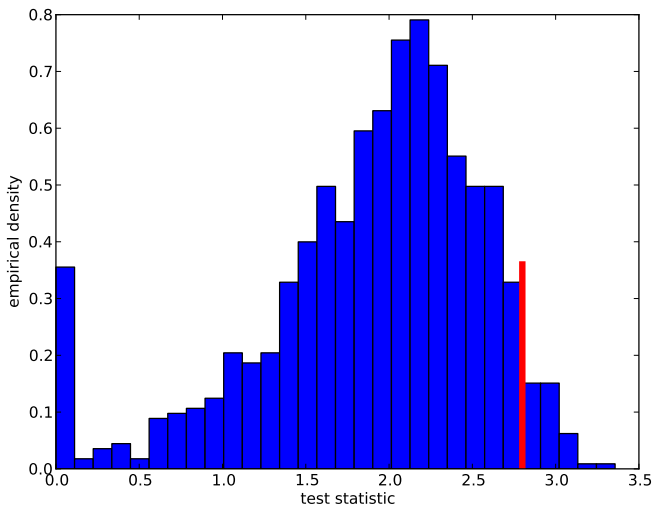
P-value

Permutation procedure

Repeat M times:

- Permute phenotype $\mathbf{y}$ and covariates $\mathbf{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution





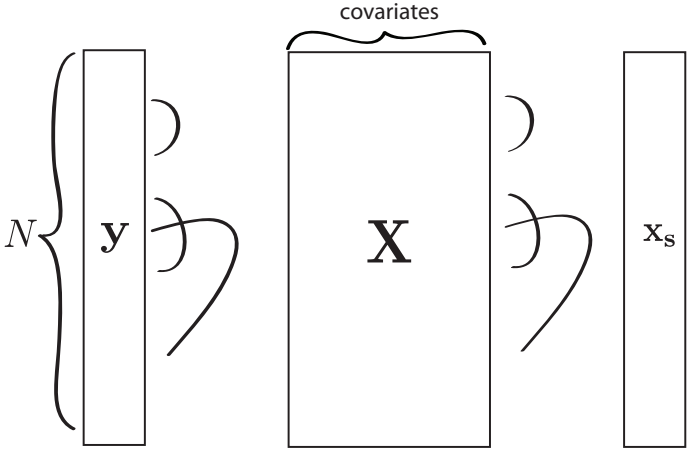
P-value

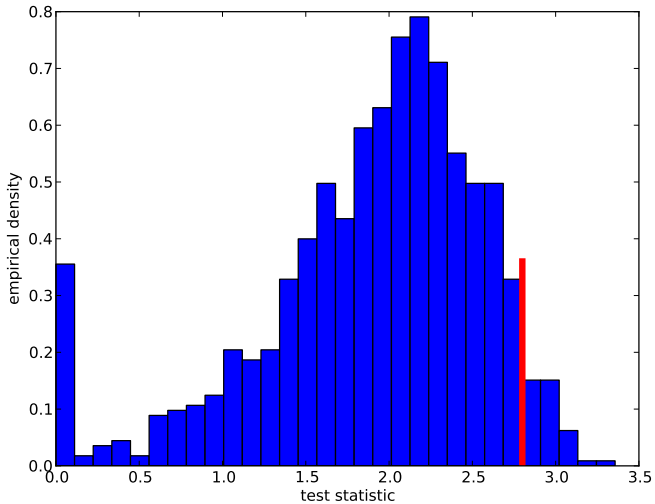
Permutation procedure

Repeat M times:

- Permute phenotype $\mathbf{y}$ and covariates $\mathbf{x}$ jointly over individuals.
- Compute permuted test statistic
- Add test statistic to emprirical null distribution

The P-value is the quantile of real test statistic in artificial null distribution.

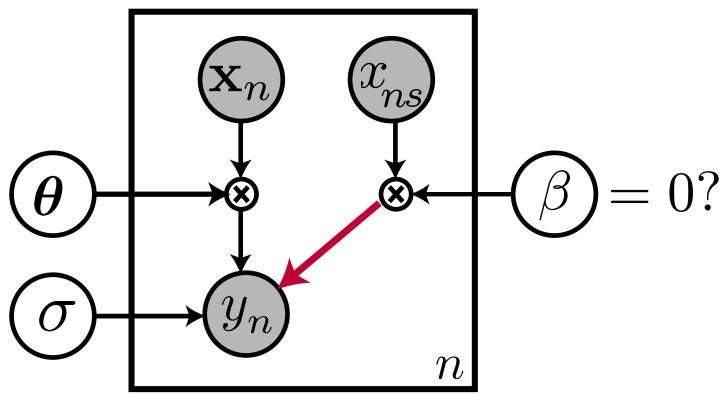




Testing in Linear Regression

Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:



Equivalent graphical model

x^n: regression covariates

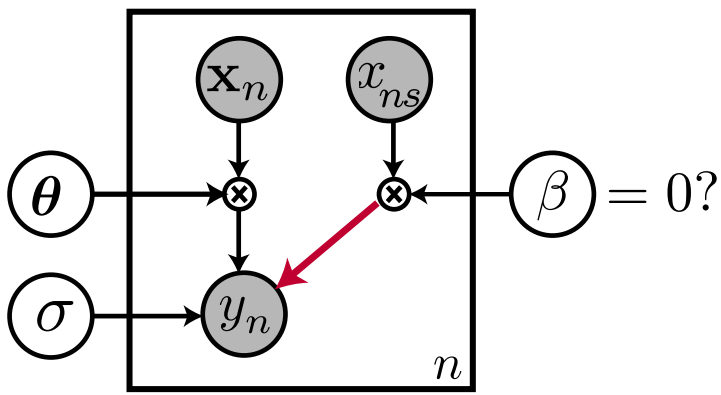
- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree of freedom.
- P-value = 1-CDF(2LR).

Testing in Linear Regression

Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:

$$\text{LR} = \sum_{n=1}^{N} \log \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\text{ML}} + x_s^n \beta_{\text{ML}}, \sigma^2_{\text{ML}}\right) - \sum_{n=1}^{N} \log \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\text{ML}_0}, \sigma^2_{\text{ML}_0}\right)$$



Equivalent graphical model

x^n: regression covariates

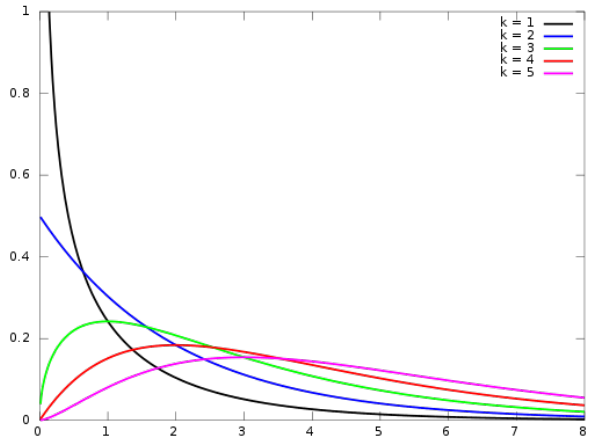
- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree of freedom.
- P-value = 1-CDF(2LR).

Testing in Linear Regression

Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:

$$\text{LR} = \sum_{n=1}^{N} \log \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\text{ML}} + x_s^n \beta_{\text{ML}}, \sigma^2_{\text{ML}}\right) - \sum_{n=1}^{N} \log \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\text{ML}_0}, \sigma^2_{\text{ML}_0}\right)$$



(source: Wikipedia)

- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree of freedom.
- P-value = 1-CDF(2LR).

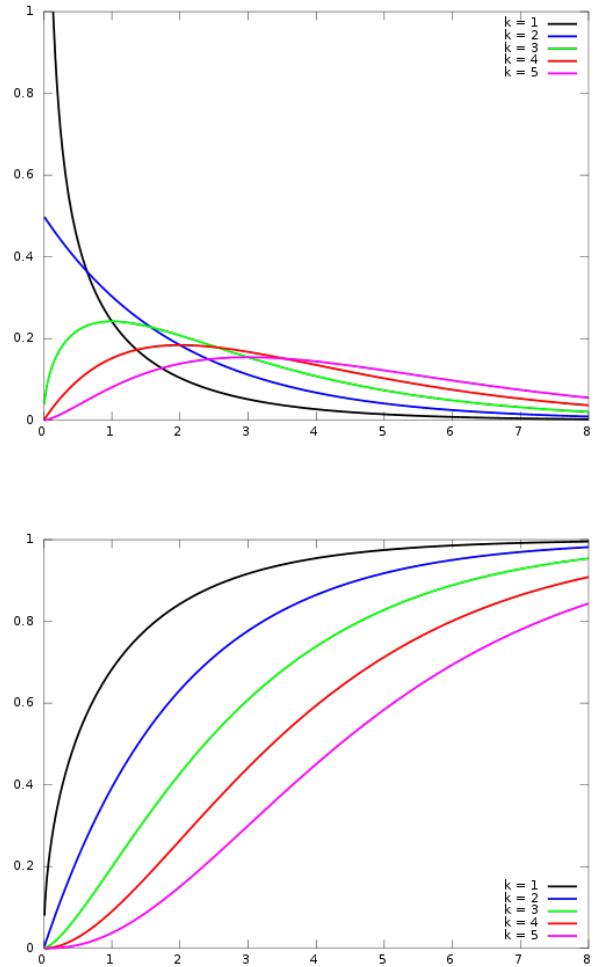
Testing in Linear Regression

Likelihood Ratio Test revisited

- Can equivalently compute log-likelihood ratio:

$$\text{LR} = \sum_{n=1}^{N} \log \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\text{ML}} + x_s^n \beta_{\text{ML}}, \sigma^2_{\text{ML}}\right) - \sum_{n=1}^{N} \log \mathcal{N}\left(y^n \,\middle|\, \mathbf{x}^n \cdot \boldsymbol{\theta}_{\text{ML}_0}, \sigma^2_{\text{ML}_0}\right)$$

- Wilks' theorem: 2LR follows a Chi-square distribution with 1 degree of freedom.
- P-value = 1-CDF(2LR).



(source: Wikipedia)

Outline

Introduction
- Why QTL mapping
- Terminology & background
- Methodological challenges

Linear Regression

Hypothesis Testing

Multiple Hypothesis Testing

Model Checking

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- At $\alpha = 0.01$ we would expect 10000 type-1 errors!
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^S \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- At $\alpha = 0.01$ we would expect 10000 type-1 errors!
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^S \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- At $\alpha = 0.01$ we would expect 10000 type-1 errors!
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^S \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- At $\alpha = 0.01$ we would expect 10000 type-1 errors!
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^S \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- At $\alpha = 0.01$ we would expect 10000 type-1 errors!
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^S \to 1$.
- Individual P-values < 0.01 are not significant anymore.

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Motivation

- Significance level α equals probability of type-1 error.
- In GWAS we perform $S = 10^6$ tests
- At $\alpha = 0.01$ we would expect 10000 type-1 errors!
- Probability of at least 1 type-1 error is $1 - (1 - \alpha)^S \to 1$.
- Individual P-values < 0.01 are not significant anymore.

Need to correct for multiple hypothesis testing!

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

- Probability of at least one type-1 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \frac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

- Probability of at least one type-1 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \frac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

- Probability of at least one type-1 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \frac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

- Probability of at least one type-1 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \frac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Multiple Hypothesis Testing

Family-Wise Error Rate (FWER)

- Probability of at least one type-1 error.
- Correct by bounding the FWER.
- Bonferroni correction: $P_B = P \cdot S$
- Equivalently $P < \frac{\alpha}{S}$ significant.
- Bounds the FWER $1 - (1 - \alpha/S)^S$ by α

	$\mathcal{H}_0$ holds	$\mathcal{H}_0$ doesn't hold
$\mathcal{H}_0$ accepted	true negatives	false negatives type-2 error
$\mathcal{H}_0$ rejected	false positives type-1 error	true positives

Outline

Introduction
- Why QTL mapping
- Terminology & background
- Methodological challenges

Linear Regression

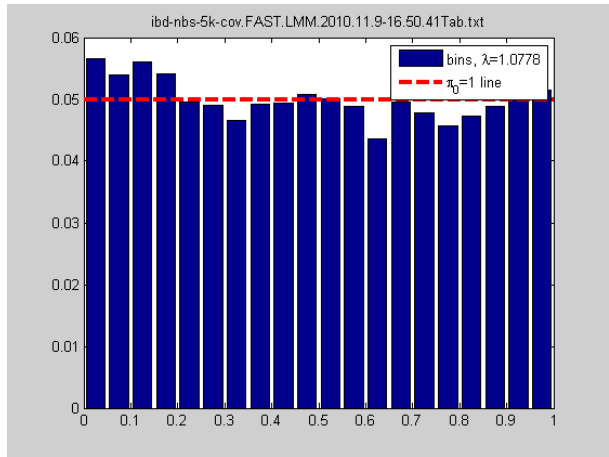
Hypothesis Testing

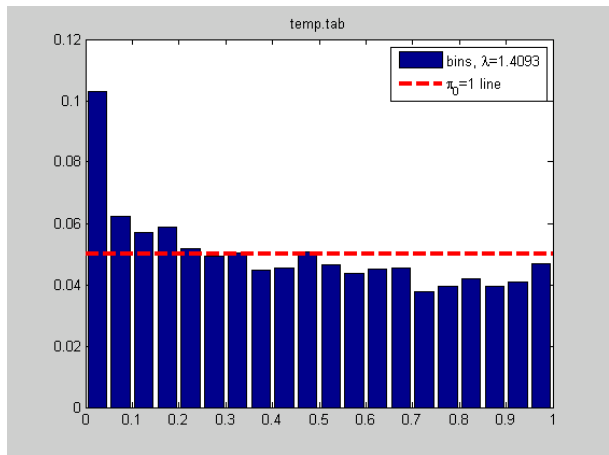
Multiple Hypothesis Testing

Model Checking

Model Checking

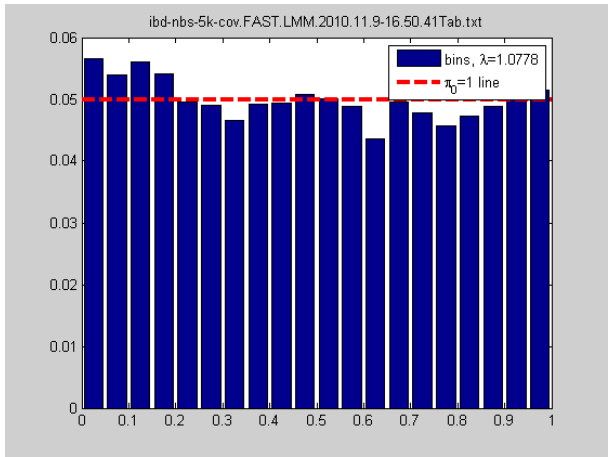
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

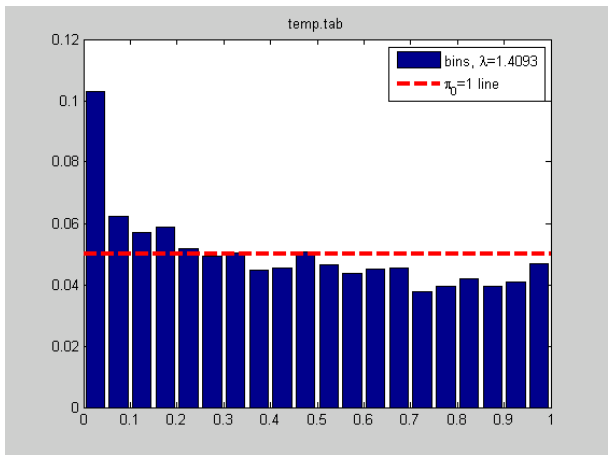




Model Checking

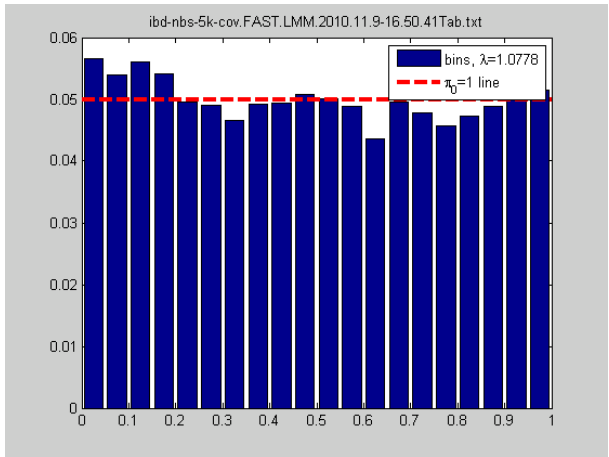
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

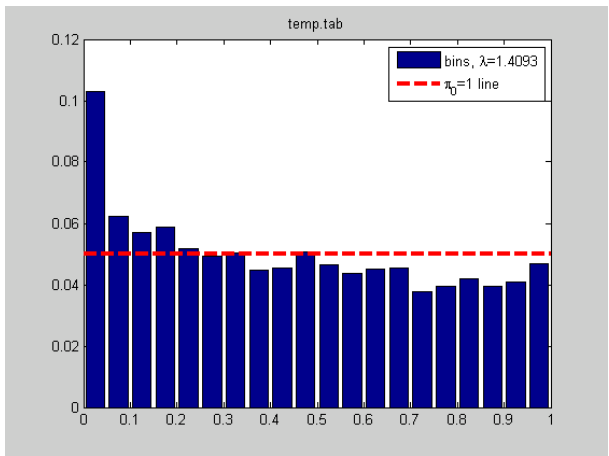




Model Checking

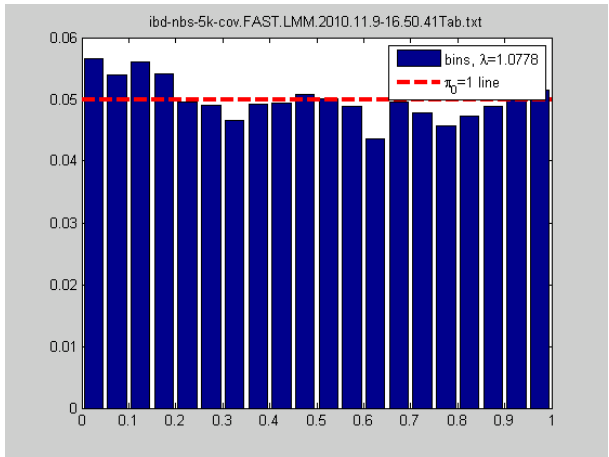
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

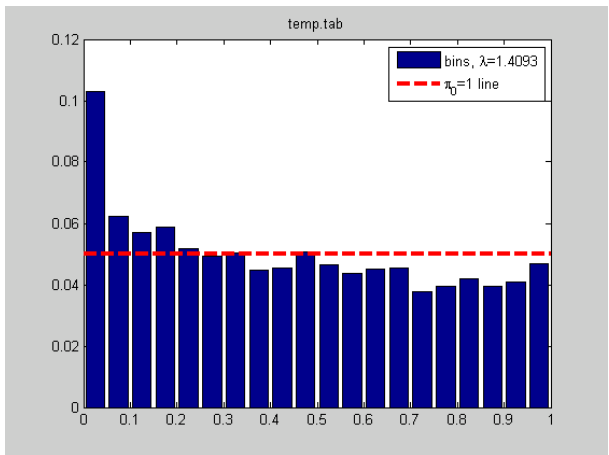




Model Checking

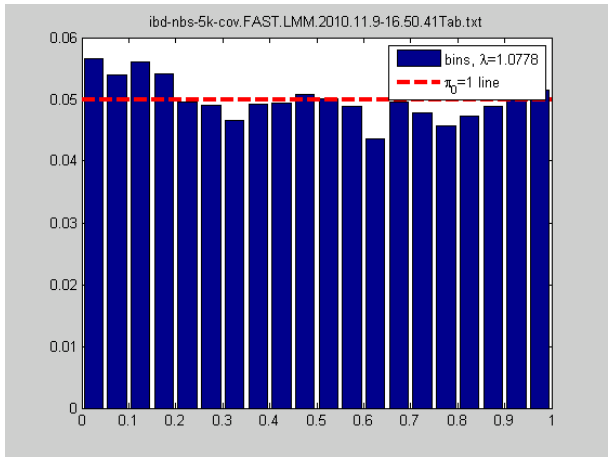
- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution

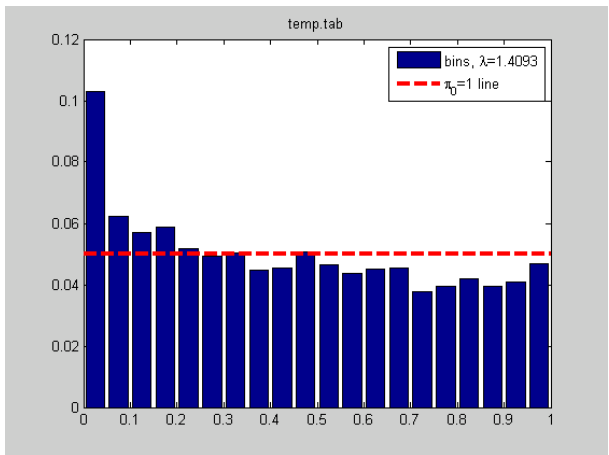




Model Checking

- Do my estimated P-values match the true null distribution?
  - By definition uniformly distributed under null distribution.
- Do the empirical results match my assumptions on the null model?
- In GWAS we perform a large number of tests. (usually in the order of 10^6)
- Use the strong prior knowledge that in GWAS almost all of the test SNPs have no effect on the phenotype.
- Empirical test statistics should follow the null distribution





Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

- Sort test statistics
- Plot test statisitcs against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.

Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

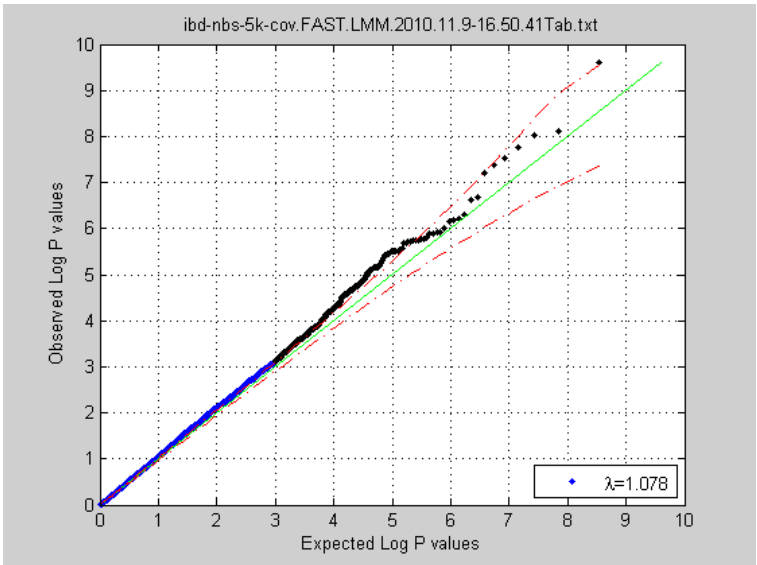
- Sort test statistics
- Plot test statisitcs against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.

Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

- Sort test statistics
- Plot test statisitcs against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.

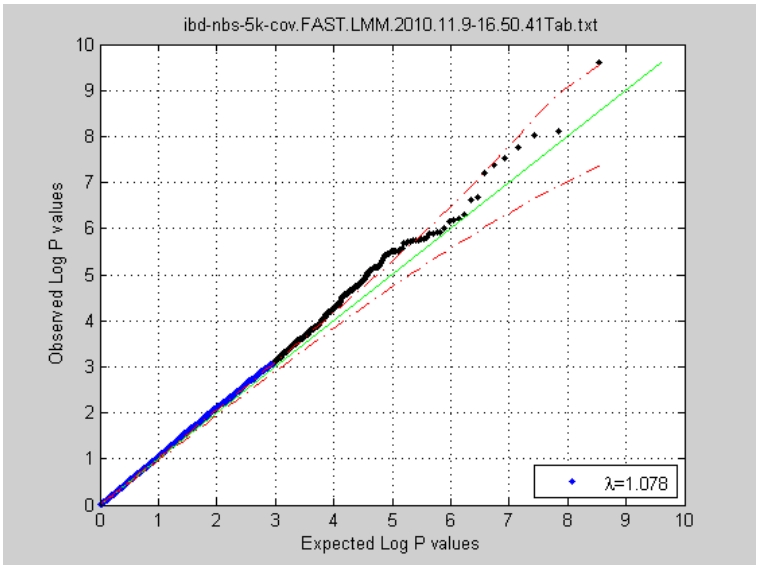


Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

- Sort test statistics
- Plot test statisitcs against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.

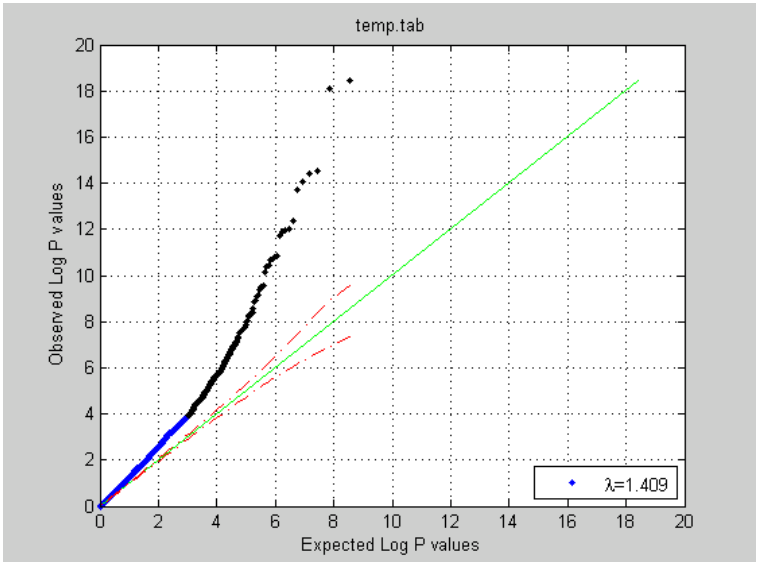


Model Checking

QQ-plot

Compare quantiles of the empirical test statistic distribution to assumed null distribution.

- Sort test statistics
- Plot test statisitcs against (y-axis) quantiles of the theoretical null-distribution (x-axis)
  - for example: 2LR vs. χ^2_1
- If the plot is close to the diagonal, the distributions match up
- Deviation from the diagonal indicates inflation or deflation of test statistics.



Summary

- Introduction
  - Genetics terminology
  - Study design
  - Data preparation
- Challenges and pitfalls
  - Power
  - Multiple hypothesis testing
  - Population structure
- Linear regression for association studies.
- Hypothesis testing
- Multiple hypothesis testing correction.
- Model checking.

Summary

- Introduction
  - Genetics terminology
  - Study design
  - Data preparation
- Challenges and pitfalls
  - Power
  - Multiple hypothesis testing
  - Population structure
- Linear regression for association studies.
- Hypothesis testing
- Multiple hypothesis testing correction.
- Model checking.

Summary

- Introduction
  - Genetics terminology
  - Study design
  - Data preparation
- Challenges and pitfalls
  - Power
  - Multiple hypothesis testing
  - Population structure
- Linear regression for association studies.
- Hypothesis testing
- Multiple hypothesis testing correction.
- Model checking.

Summary

- Introduction
  - Genetics terminology
  - Study design
  - Data preparation
- Challenges and pitfalls
  - Power
  - Multiple hypothesis testing
  - Population structure
- Linear regression for association studies.
- Hypothesis testing
- Multiple hypothesis testing correction.
- Model checking.

Summary

- Introduction
  - Genetics terminology
  - Study design
  - Data preparation
- Challenges and pitfalls
  - Power
  - Multiple hypothesis testing
  - Population structure
- Linear regression for association studies.
- Hypothesis testing
- Multiple hypothesis testing correction.
- Model checking.

Summary

- Introduction
  - Genetics terminology
  - Study design
  - Data preparation
- Challenges and pitfalls
  - Power
  - Multiple hypothesis testing
  - Population structure
- Linear regression for association studies.
- Hypothesis testing
- Multiple hypothesis testing correction.
- Model checking.

Summary

- Introduction
  - Genetics terminology
  - Study design
  - Data preparation
- Challenges and pitfalls
  - Power
  - Multiple hypothesis testing
  - Population structure
- Linear regression for association studies.
- Hypothesis testing
- Multiple hypothesis testing correction.
- Model checking.

Acknowledgements

- **Joint course material**
  O. Stegle
- **Why QTL mapping**
  D. Weigel, K. Borgwardt

References I

S. Atwell, Y. Huang, B. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. Tarone, T. Hu, et al. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, 2010.

B. Browning and S. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009.

J. Cao, K. Schneeberger, S. Ossowski, T. Günther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, X. Wang, F. Ott, J. Müller, C. Alonso-Blanco, K. Borgwardt, K. Schmid, and D. Weigel. Whole-genome sequencing of multiple *arabidopsis thaliana* populations. *Nature Genetics*, 43(10):956–963, 10 2011. doi: 10.1038/ng.911.

J. Spitzer. A primer on box-cox estimation. *The Review of Economics and Statistics*, 64(2): 307–313, 1982.

G. Upton and I. Cook. Oxford dictionary of statistics, 2002.