Overview and introduction

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

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

Basel
09. September 2012

Time table

- 09:10–09:20 Welcome
- 09:20–10:00 Introduction and background
- 10:00–10:30 Linear models I
- 10:30–11:00 **Coffee break**
- 11:00–11:30 Linear models I contd.
- 11:30–11:45 Demonstrations I
- 11:45–12:15 Linear models II
- 12:15–13:30 **Lunch**
- 13:30–14:00 Linear models II contd.
- 14:00–14:30 Advanced mixed models
- 14:30–15:00 Demonstrations II
- 15:00–15:30 **Coffee break**
- 15:30–16:15 High-dimensional traits, gene expression
- 16:15–17:00 Discussion, questions, etc.

Time table

- 09:10–09:20 Welcome
- 09:20–10:00 Introduction and background
- 10:00–10:30 Linear models I
- 10:30–11:00 **Coffee break**
- 11:00–11:30 Linear models I contd.
- 11:30–11:45 Demonstrations I
- 11:45–12:15 Linear models II
- 12:15–13:30 **Lunch**
- 13:30–14:00 Linear models II contd.
- 14:00–14:30 Advanced mixed models
- 14:30–15:00 Demonstrations II
- 15:00–15:30 **Coffee break**
- 15:30–16:15 High-dimensional traits, gene expression
- 16:15–17:00 Discussion, questions, etc.

Time table

- 09:10–09:20 Welcome
- 09:20–10:00 Introduction and background
- 10:00–10:30 Linear models I
- 10:30–11:00 **Coffee break**
- 11:00–11:30 Linear models I contd.
- 11:30–11:45 Demonstrations I
- 11:45–12:15 Linear models II
- 12:15–13:30 **Lunch**
- 13:30–14:00 Linear models II contd.
- 14:00–14:30 Advanced mixed models
- 14:30–15:00 Demonstrations II
- 15:00–15:30 **Coffee break**
- 15:30–16:15 High-dimensional traits, gene expression
- 16:15–17:00 Discussion, questions, etc.

Time table

- 09:10–09:20 Welcome
- 09:20–10:00 Introduction and background
- 10:00–10:30 Linear models I
- 10:30–11:00 **Coffee break**
- 11:00–11:30 Linear models I contd.
- 11:30–11:45 Demonstrations I
- 11:45–12:15 Linear models II
- 12:15–13:30 **Lunch**
- 13:30–14:00 Linear models II contd.
- 14:00–14:30 Advanced mixed models
- 14:30–15:00 Demonstrations II
- 15:00–15:30 **Coffee break**
- 15:30–16:15 High-dimensional traits, gene expression
- 16:15–17:00 Discussion, questions, etc.

Outline

Outline

Why QTL mapping

Terminology & background

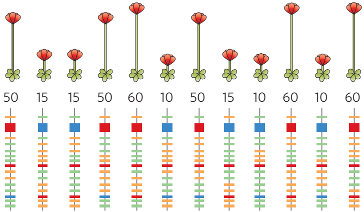
Methodological challenges

Tutorial outline & resources

Genotype to phenotype mapping

Given:

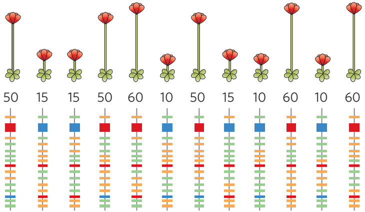
- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, . . .



Genotype to phenotype mapping

Given:

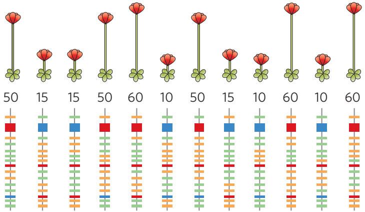
- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, ...



Genotype to phenotype mapping

Given:

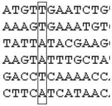
- Genotype for multiple individuals
  - Single nucleotide polymorphisms (SNPs), microsatelite markers
- Quantitative traits (phenotypes) for the same individuals
  - disease, height, gene-expression, ...



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, . . .



ATGTTGAATCTG
AAAGTGAAATGT
TATTATACGAAG
AAGTATTTGCTA
GACCTCAAAACC
CTTCATCATAAC

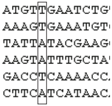
Goal:

- 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, . . .



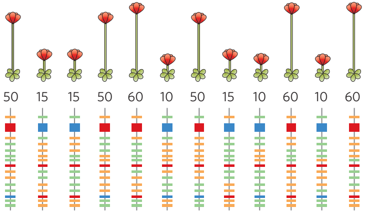
```
ATGTTGAATCTG
AAAGTGAAATGT
TATTATACGAAG
AAGTATTTGCTA
GACCTCAAAACC
CTTCATCATAAC
```

Goal:

- Identify causal loci that explain phenotypic differences.

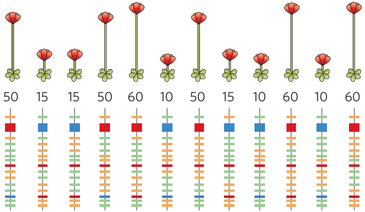
Use of GWAs in plant systems

- 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.



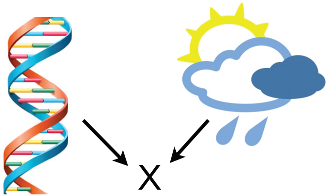
Use of GWAs in plant systems

- 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

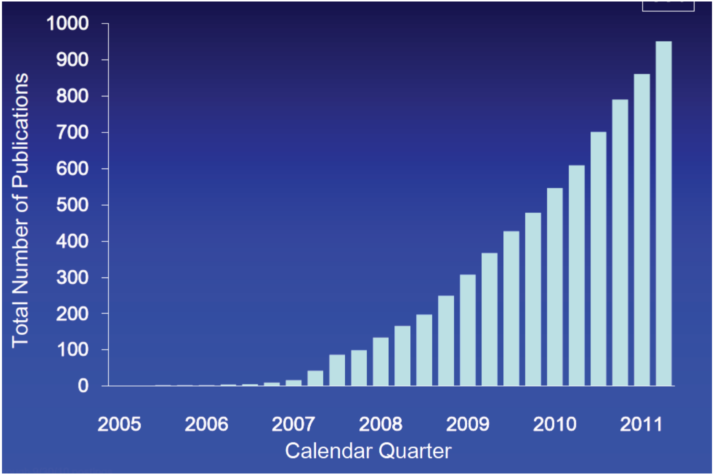
- Adapting treatment to the patients genetic make-up.
  - Targeting patients who can benefit.
  - Appropriate dosage of a drug by using genetic variants to understand drug metabolism (e.g., anti-depressants, beta blockers, opioid analgesics).
  - Disease subcategorization
- Risk prediction
  - Known 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.

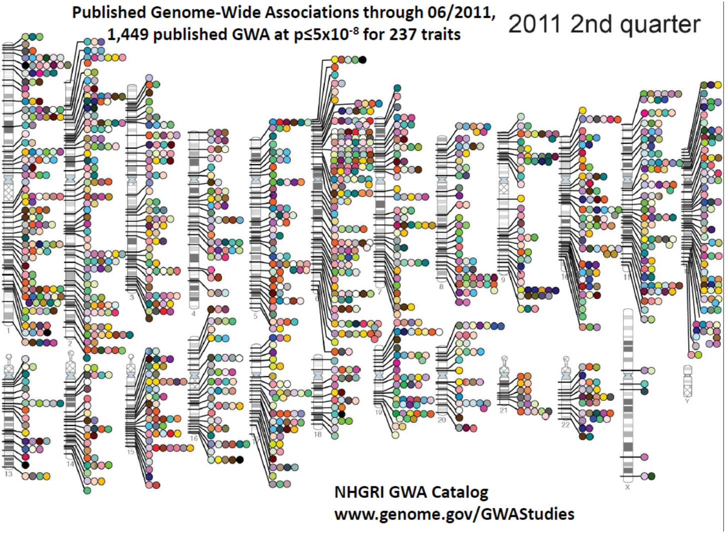
Personalized medicine & health

Publication boost



Personalized medicine & health

Publication boost



Published Genome-Wide Associations through 06/2011, 1,449 published GWA at p≤5x10^{-8} for 237 traits

2011 2nd quarter

NHGRI GWA Catalog
www.genome.gov/GWAStudies

Outline

Why QTL mapping

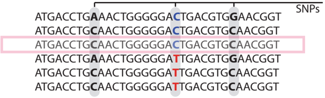
Terminology & background

Methodological challenges

Tutorial outline & resources

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.


SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT

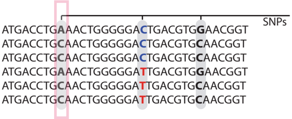
Some definitions

- 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

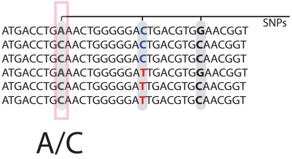
- Genotype denotes the genetic state of an individual.
  - Denoted by $\mathbf{x}^n$ for individual n.
- Phenotype denotes the state of a trait of an individual.
  - Denoted by $\mathbf{y}^n$ for individual n.
- A locus is a position or limited region in the genome.
  - Denoted by $\mathbf{x}_s$ for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



ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT

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.


SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
A/C

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

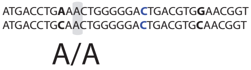
More definitions

- 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.


ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
A/A

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.

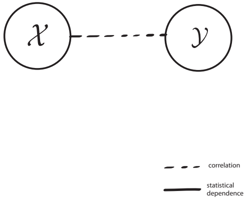
ATGACCTG**A**AACTGGGGGA**C**TGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGGA**C**TGACGTG**C**AACGGT

A/C

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

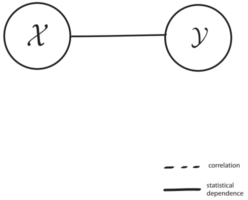


[Upton and Cook, 2002]

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

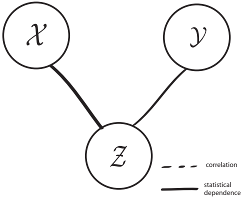


[Upton and Cook, 2002]

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

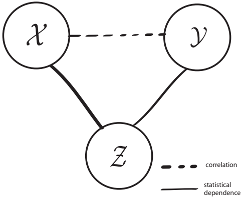


[Upton and Cook, 2002]

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

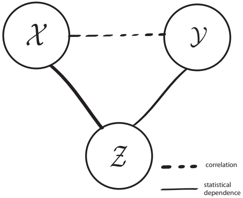


[Upton and Cook, 2002]

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

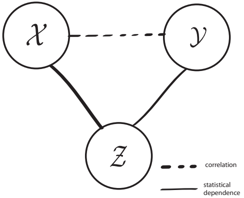


[Upton and Cook, 2002]

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*

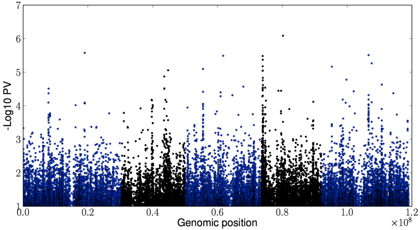
- Phenotype: Flowering time at 10 degrees
- Test every SNP in the genome for association with floweringtime
- Position vs. Log10(P-value) (Manhattan plot)

[Atwell et al., 2010]

Result

Example GWAS on *A. thaliana*

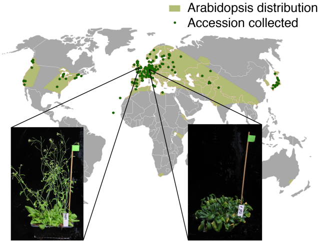
- Phenotype: Flowering time at 10 degrees
- Test every SNP in the genome for association with floweringtime
- Position vs. Log10(P-value) (Manhattan plot)

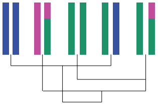


[Atwell et al., 2010]

Genetic designs

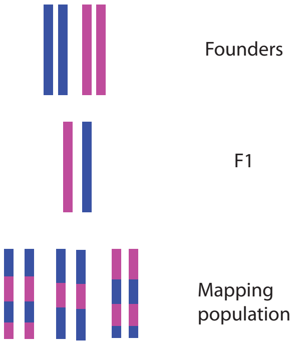
- Natural population
  - Global sampling of plants, human or animals.
  - Samples may exhibit varying degrees of relatedness.
  - Typically diploid.
- Inbred F2 crosses
  - Mapping of the differences of founder strains
  - Plant- and animal systems
  - No relatedness
  - Typically haploid.
- Multi-parent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





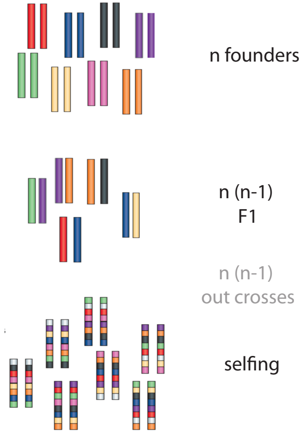
Genetic designs

- 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.



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**AACTGGGGACTGACGTG**G**AACGGT
ATGACCTG**C**AACTGGGGACTGACGTG**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.

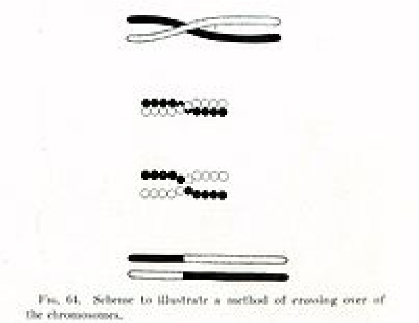


FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

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.

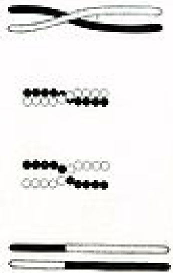


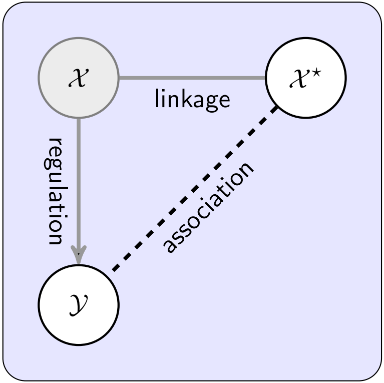
Fig. 61. Scheme to illustrate a method of crossing over of the chromosomes.

image source: Wikipedia

Linkage Disequilibrium

Physical linkage

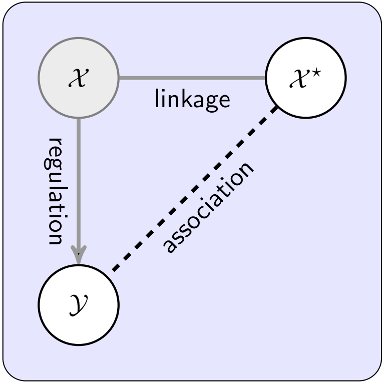
- Recombination causes linkage between loci.
- Linkage is not uniform along the chromosome.
- Recombination hotspots on the chromosome lead to conserved haplotype blocks in strong linkage.
- Linkage can be used to chose *tag*-SNPs to cover all linked regions.
    - Tradeoff between resolution and genotyping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



Linkage Disequilibrium

Physical linkage

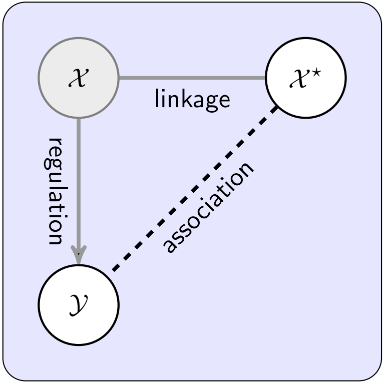
- Recombination causes linkage between loci.
- Linkage is not uniform along the chromosome.
- Recombination hotspots on the chromosome lead to conserved haplotype blocks in strong linkage.
- Linkage can be used to chose *tag*-SNPs to cover all linked regions.
    - Tradeoff between resolution and genotyping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



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
- 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

Phenotypes

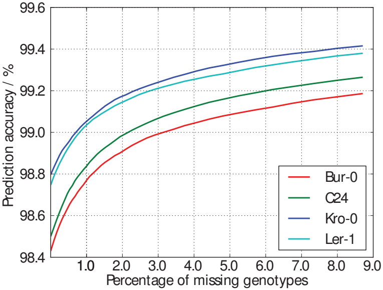
- Binary
  - case, control
- Continuous
  - Gaussian
  - Non-Gaussian
- Multivariate
- Other

- 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



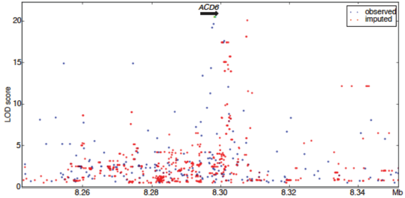
Genotype imputation accuracy from SNP-chip to 80Genomes reference panel [Cao et al., 2011].

[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



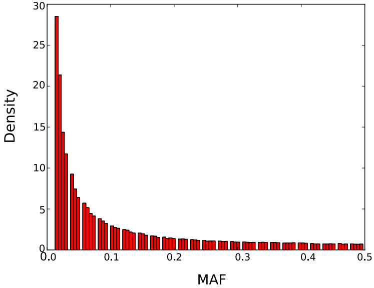
Genotype imputation accuracy from SNP-chip to 80Genomes reference panel [Cao et al., 2011].

[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

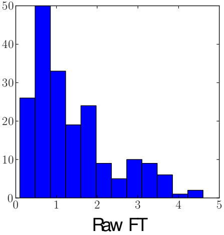
Phenotype

- 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



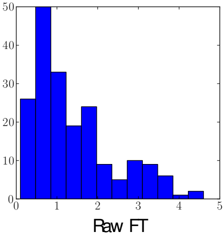
Raw and Box-Cox transformed flowering phenotypes at 10C [Atwell et al., 2010].

[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



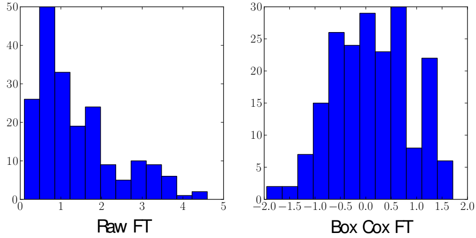
Raw and Box-Cox transformed flowering phenotypes at 10C [Atwell et al., 2010].

[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



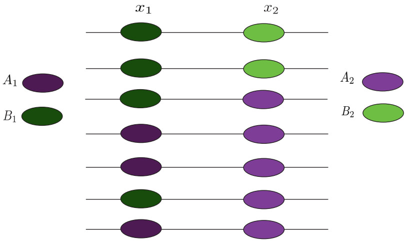
Raw and Box-Cox transformed flowering phenotypes at 10C [Atwell et al., 2010].

[Spitzer, 1982]

Linkage Disequilibrium

Gametic Phase Disequilibrium

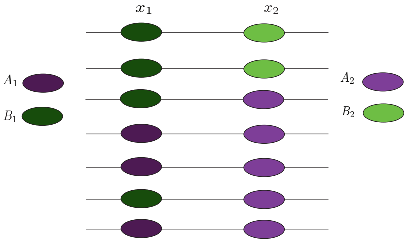
- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_{.A} f_{A.}$
  - $r^2 = \frac{D^2}{f_{AA} f_{AB} f_{BA} f_{BB}}$
-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



Linkage Disequilibrium

Gametic Phase Disequilibrium

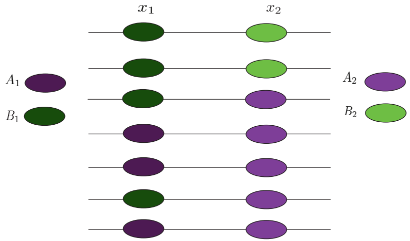
- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_{A\cdot} f_{\cdot A}$
  - $r^2 = \frac{D^2}{f_{AA} f_{AB} f_{BA} f_{BB}}$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



Linkage Disequilibrium

Gametic Phase Disequilibrium

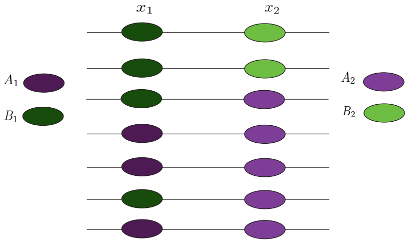
- Association between two loci.
- Deviation from random co-inheritance between loci.
- LD can be caused by recombination, population structure, epistasis
- Measures of LD between two loci x_1 and x_2 are D and r^2.
  - $D = f_{AA} - f_{\cdot A} f_{A \cdot}$
  - $r^2 = \frac{D^2}{f_{AA} f_{AB} f_{BA}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



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.
    - $D = f_{AA} - f_{.A} f_{A.}$.
    - $r^2 = \dfrac{D^2}{f_{AA} f_{AB} f_{BA} f_{BB}}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.

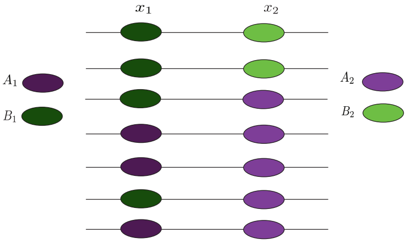


	$x_2 = A_2$	$x_2 = B_2$	
$x_1 = A_1$	f_{AA}	f_{AB}	$f_{A.}$
$x_1 = B_1$	f_{BA}	f_{BB}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

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.

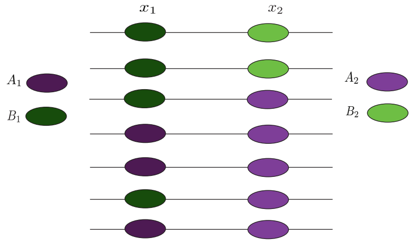


	$x_2 = A_2$	$x_2 = B_2$	
$x_1 = A_1$	f_{AA}	f_{AB}	$f_{A.}$
$x_1 = B_1$	f_{BA}	f_{BB}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

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.
  - $D = f_{AA} - f_{.A} f_{A.}$.
  - $r^2 = \dfrac{D^2}{f_{AA} f_{AB} f_{BA} f_{BB}}$
- $D \neq 0$ and $r^2 \neq 0$ are indicators of LD.

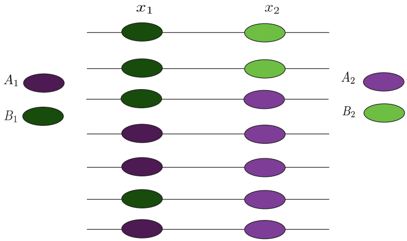


	$x_2 = A_2$	$x_2 = B_2$	
$x_1 = A_1$	f_{AA}	f_{AB}	$f_{A.}$
$x_1 = B_1$	f_{BA}	f_{BB}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

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.
  - $D = f_{AA} - f_{.A} f_{A.}$.
  - $r^2 = \dfrac{D^2}{f_{AA} f_{AB} f_{BA} f_{BB}}$
- $D \neq 0$ 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}	$f_{B.}$
	$f_{.A}$	$f_{.B}$	

Outline

Why QTL mapping

Terminology & background

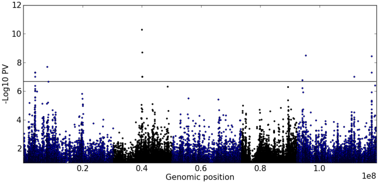
Methodological challenges

Tutorial outline & resources

Challenges

Multiple hypothesis testing

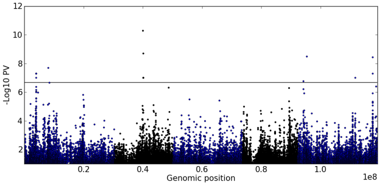
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges

Multiple hypothesis testing

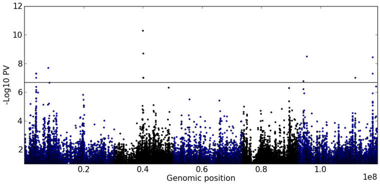
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges

Multiple hypothesis testing

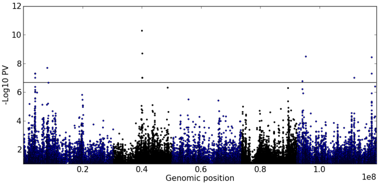
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges

Multiple hypothesis testing

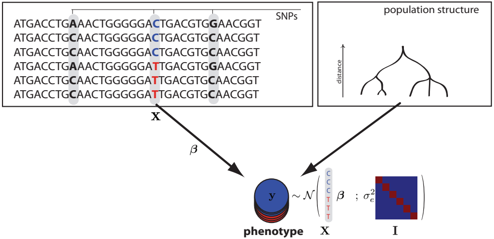
- In GWAS, the number of statistical tests commonly is on the order of 10^6.
- At significane level of 0.01 we would expect 10,000 false positives
- Thus, individual P-values < 0.01 are not significant anymore.
- Correction for multiple hypothesis testing is critical!



Challenges

Population structure

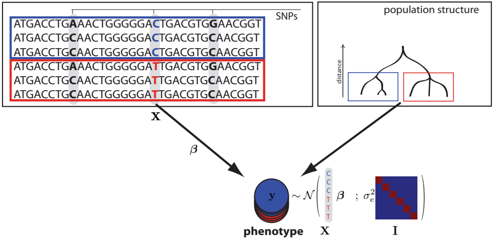
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness


SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
X
population structure
distance
β
y
phenotype
$\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}\ ;\ \sigma_e^2 \mathbf{I})$
C C C T T T
X
I

Challenges

Population structure

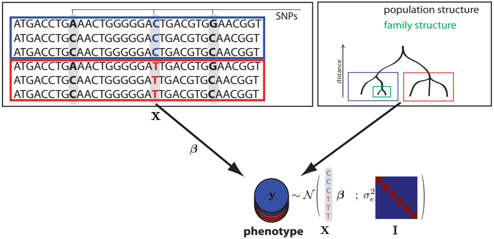
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness


SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
X
β
population structure
distance
y
~ N(C C C T T T β ; σ²_e)
phenotype
X
I

Challenges

Population structure

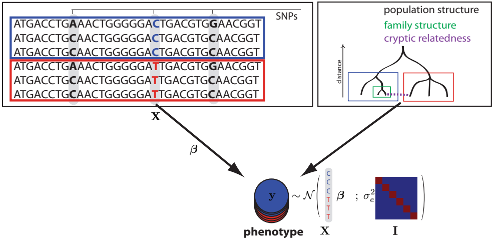
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness


SNPs
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
X
population structure
family structure
distance
β
y
phenotype
~ N
C
C
C
T
T
T
β
; σ_e^2
X
I

Challenges

Population structure

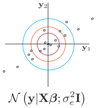
- Confounding structure leads to false positives.
  - Population structure
  - Family structure
  - Cryptic relatedness

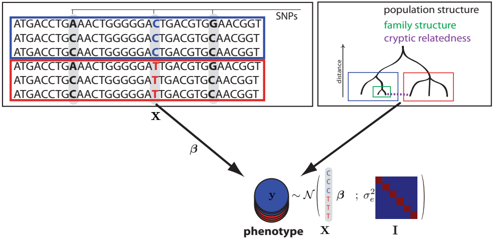

SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
population structure
family structure
cryptic relatedness
distance
X
β
y
phenotype
$\sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta} ; \sigma_e^2 \mathbf{I} \right)$
C C C T T T
X
I

Challenges

Population structure

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




SNPs
ATGACCTGAAACTGGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGGATTGACGTGCAACGGT
X
population structure
family structure
cryptic relatedness
distance
β
y
phenotype
$\sim \mathcal{N}$
C C C T T T
β
$; \sigma_e^2$
X
I

Challenges

Population structure

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

Challenges

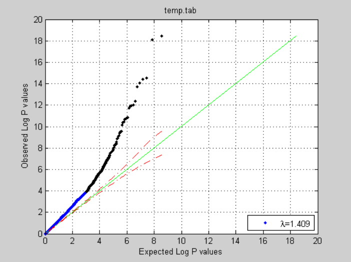
Population structure

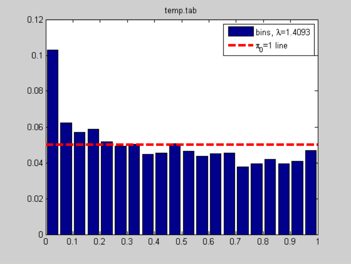
- GWA on inflammatory bowel disease (WTCCC)
- 3.4k cases, 11.9k controls
- Methods
  - Linear regression
  - Likelihood ratio test

Challenges

Population structure

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

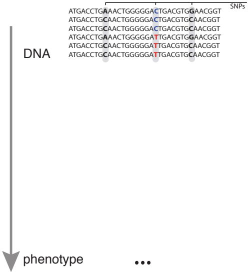




Challenges

Background variation and confounding

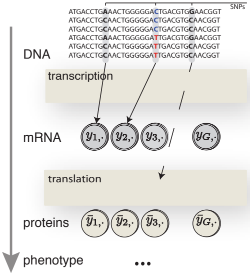
- Genotype is not the sole cause of phenotype variability
- Environment (known and unknown)
- Covariates


SNPs
DNA
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
phenotype
...

Challenges

Background variation and confounding

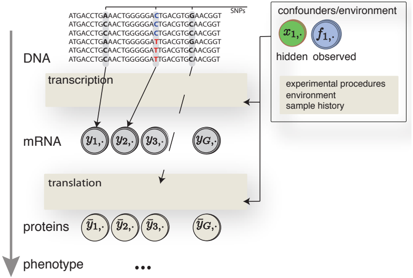
- Genotype is not the sole cause of phenotype variability
- Environment (known and unknown)
- Covariates


SNPs
ATGACCTGAAACTGGGGACTGACGTGGAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGCAACTGGGGACTGACGTGCAACGGT
ATGACCTGAAACTGGGGATTGACGTGGAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
ATGACCTGCAACTGGGGATTGACGTGCAACGGT
DNA
transcription
mRNA
$y_{1,\cdot}$
$y_{2,\cdot}$
$y_{3,\cdot}$
$y_{G,\cdot}$
translation
proteins
$\bar{y}_{1,\cdot}$
$\bar{y}_{2,\cdot}$
$\bar{y}_{3,\cdot}$
$\bar{y}_{G,\cdot}$
phenotype
...

Challenges

Background variation and confounding

- Genotype is not the sole cause of phenotype variability
- Environment (known and unknown)
- Covariates



Challenges

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 known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges

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 known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges

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 known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges

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 known effects
    - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges

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 known effects
    - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges

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 known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges

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 known effects
  - Testing compound hypotheses (e.g. test all (rare) variants in a window)

Challenges

Non-independent traits

- Measured expression levels for thousands of genes
- Growth related phenotypes
  - e.g. weight, BMI
- Increase power by exploiting phenotype correlations
- Use correlations to estimate hidden common causes

Challenges

Non-independent traits

- Measured expression levels for thousands of genes
- Growth related phenotypes
  - e.g. weight, BMI
- Increase power by exploiting phenotype correlations
- Use correlations to estimate hidden common causes

Challenges

Non-independent traits

- Measured expression levels for thousands of genes
- Growth related phenotypes
  - e.g. weight, BMI
- Increase power by exploiting phenotype correlations
- Use correlations to estimate hidden common causes

Challenges

Non-independent traits

- Measured expression levels for thousands of genes
- Growth related phenotypes
  - e.g. weight, BMI
- Increase power by exploiting phenotype correlations
- Use correlations to estimate hidden common causes

Challenges

Non-independent traits

- Measured expression levels for thousands of genes
- Growth related phenotypes
  - e.g. weight, BMI
- Increase power by exploiting phenotype correlations
- Use correlations to estimate hidden common causes

Outline

Why QTL mapping

Terminology & background

Methodological challenges

Tutorial outline & resources

Topics covered

- Linear models 1
  - Significance testing, multiple hypothesis correction, correction for population structure
- Linear models 2
  - Composite variance analysis, multi-trait models, phenotype prediction, LASSO
- Advanced topics
  - Improved linear mixed models
  - Association mapping of high-dimensional traits
- Practical demonstations and demos to take away
- Opportunities for open discussion, questions and scientific exchange

Topics covered

- Linear models 1
  - Significance testing, multiple hypothesis correction, correction for population structure
- Linear models 2
  - Composite variance analysis, multi-trait models, phenotype prediction, LASSO
- Advanced topics
  - Improved linear mixed models
  - Association mapping of high-dimensional traits
- Practical demonstations and demos to take away
- Opportunities for open discussion, questions and scientific exchange

Topics covered

- Linear models 1
  - Significance testing, multiple hypothesis correction, correction for population structure
- Linear models 2
  - Composite variance analysis, multi-trait models, phenotype prediction, LASSO
- Advanced topics
  - Improved linear mixed models
  - Association mapping of high-dimensional traits

- Practical demonstations and demos to take away
- Opportunities for open discussion, questions and scientific exchange

Topics covered

- Linear models 1
  - Significance testing, multiple hypothesis correction, correction for population structure
- Linear models 2
  - Composite variance analysis, multi-trait models, phenotype prediction, LASSO
- Advanced topics
  - Improved linear mixed models
  - Association mapping of high-dimensional traits

- Practical demonstations and demos to take away
- Opportunities for open discussion, questions and scientific exchange

Topics covered

- Linear models 1
  - Significance testing, multiple hypothesis correction, correction for population structure
- Linear models 2
  - Composite variance analysis, multi-trait models, phenotype prediction, LASSO
- Advanced topics
  - Improved linear mixed models
  - Association mapping of high-dimensional traits

- Practical demonstations and demos to take away
- Opportunities for open discussion, questions and scientific exchange

Software used in examples

FaST-LMM

- Large-scale genome-wide association studies
- Command line tool
- Correction for population structure on up to 100k samples
- `http://mscompbio.codeplex.com`

[Lippert et al., 2011]

Software used in examples

FaST-LMM

- Large-scale genome-wide association studies
- Command line tool
- Correction for population structure on up to 100k samples
- `http://mscompbio.codeplex.com`

[Lippert et al., 2011]

Software used in examples

FaST-LMM

- Large-scale genome-wide association studies
- Command line tool
- Correction for population structure on up to 100k samples
- `http://mscompbio.codeplex.com`

[Lippert et al., 2011]

Software used in examples

FaST-LMM

- Large-scale genome-wide association studies
- Command line tool
- Correction for population structure on up to 100k samples
- `http://mscompbio.codeplex.com`

[Lippert et al., 2011]

Software used in examples
Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- http://github.com/PMBio/limix

Software used in examples

Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- `http://github.com/PMBio/limix`

Software used in examples

Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- `http://github.com/PMBio/limix`

Software used in examples
Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- `http://github.com/PMBio/limix`

Software used in examples

Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- `http://github.com/PMBio/limix`

Software used in examples

Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- `http://github.com/PMBio/limix`

Software used in examples

Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- `http://github.com/PMBio/limix`

Software used in examples

Limix

- Efficient open source C++ toolbox for advanced GWAS analyses
- Modular python interface (R coming soon)
- Variance component estimation
- Complex covariance modeling
- Multi-trait models
- Latent variable models
- `http://github.com/PMBio/limix`

Resources

Datasets and resources used in this tutorial

- Wellcome Trust Case Control Consortium [Burton et al., 2007]
  - Data access and details: `http://www.wtccc.org.uk/`.
- *A. thliana* GWAS on 107 phenotypes [Atwell et al., 2010]
  - Data publicly available `https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb/genomic-polymorphism-data-in-arabidopsis-thaliana`
- eQTL datasets from yeast [Smith and Kruglyak, 2008]
  - Data is also included in the examples of PEER [Stegle et al., 2012]
    - Data download: `http://www.nature.com/nprot/journal/v7/n3/extref/nprot.2011.457-S1.zip`

Questions?

Acknowledgements

- **Why QTL mapping**
  Detlef Weigel, Karsten 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.

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.

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.

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.

E. Smith and L. Kruglyak. Gene–environment interaction in yeast gene expression. *PLoS biology*, 6(4):e83, 2008.

References II

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

O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

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