Mixed linear model analyses of human complex traits using SNP data

Jian Yang

Queensland Brain Institute

The University of Queensland

Why do we need a mixed linear model?

Linear model

- $y = b_0 + x_1 b_1 + x_2 b_2 + \ldots + x_p b_p + e$

y = phenotype
x_i = independent variable

$y \sim N(b_0 + x_1 b_1 + x_2 b_2 + \ldots + x_p b_p, \sigma^2_e)$
b_0 = mean term
$b_1 \ldots b_p$ = effect sizes (regression coefficients)
e = residual, $e \sim N(0, \sigma^2_e)$

Linear model

- In matrix form

$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$

$\mathbf{y} = \{y_j\}_{n \times 1}$; $\mathbf{X} = \{X_{ij}\}_{n \times p}$; $\mathbf{b} = \{b_i\}_{p \times 1}$; $\mathbf{e} = \{e_j\}_{n \times 1}$

- Estimation

$\mathbf{b}\text{-hat} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$\text{var}(\mathbf{b}\text{-hat}) = \sigma^2_e(\mathbf{X}^T\mathbf{X})^{-1}$

Special cases

- Simple regression

$y = b_0 + x_1 b_1 + e$

b_1-hat = $\mathbf{b}$-hat = $\mathbf{X}_1^T \mathbf{y} / (\mathbf{X}_1^T \mathbf{X}_1)$

$E(b_1\text{-hat}) = b_1 = \text{cov}(x_1, y) / \text{var}(x_1)$

$\text{var}(b_1\text{-hat}) = \sigma^2_e / [n * \text{var}(x_1)]$

- Conditional analysis

$y \mid b_2 \ldots b_p = b_0 + x_1 b_1 + e$

Limitations

- n > p: sample size needs to be > than the number of parameters
- All the effect sizes are treated as fixed (we have no idea about the variation in effect sizes)
- What if n << p?

What is a mixed linear model (MLM)?

- $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$

Fixed effects: **b** (special case: **X** = **1** and **b** = b_0)

Random effects: **u** = $\{u_i\}$, **u** ~ N(**0**, $\sigma^2_u\mathbf{A}$)

A = correlation matrix between u_i and u_j

E(**y**) = **Xb**

var(**y**) = **V** = $\mathbf{ZAZ}^T\sigma^2_u + \mathbf{I}\sigma^2_e$

Parameter estimation

- Estimation of variance components (σ^2_u)
  logL = -1/2(log|$\mathbf{V}$| + log|$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$| + $\mathbf{y}^T\mathbf{P}$y
  $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$

- Prediction of random effects ($\mathbf{u}$)
  $\mathbf{u}$-hat = σ^2_u-hat $\mathbf{Z}^T\mathbf{P}\mathbf{y}$

- Estimation of fixed effects ($\mathbf{b}$)
  $\mathbf{b}$-hat = $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
  Linear model: $\mathbf{b}$-hat = $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

MLM analysis of human complex traits

- Animal and plant breeding
  - predicting breeding values
  - linkage mapping (QTL mapping)

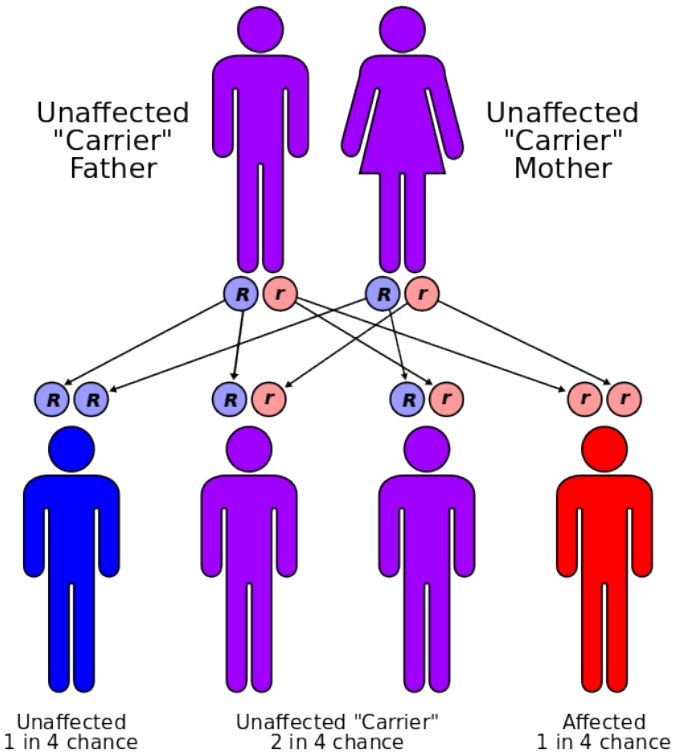
- Human genetics (before 2007)
  - pedigree based analysis of variance (heritability)
  - linkage mapping

- Human genetics (after 2007)
  - **estimating SNP-based heritability**
  - association analysis
  - genetic risk prediction

Background

Mendelian traits

Cystic fibrosis



Complex traits

Human height

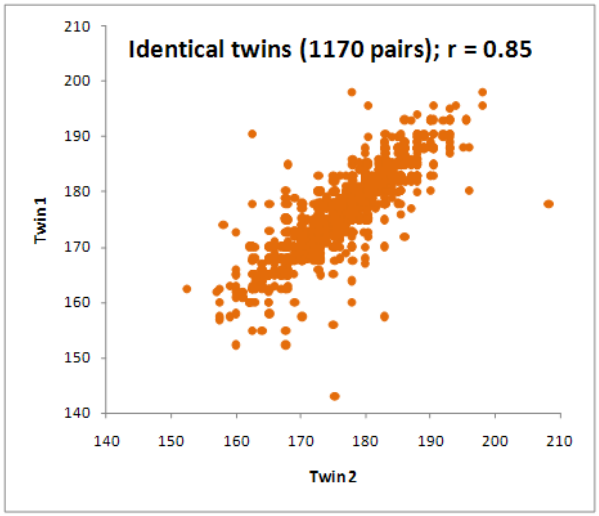
Schizophrenia

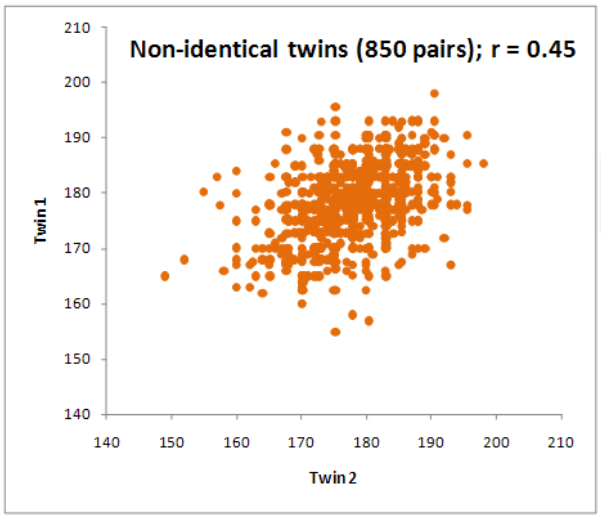
Obesity

Major questions

- Are these traits heritable?
- If so, what is the heritability?
- How many genes involved and where are they located?

Resemblance between twins for human height





Heritability = ~80%

Resemblance between relatives for body mass index (BMI)

Relatedness	Correlation
Full-sibs	0.36
Father-son	0.28

Heritability = 40%~60%

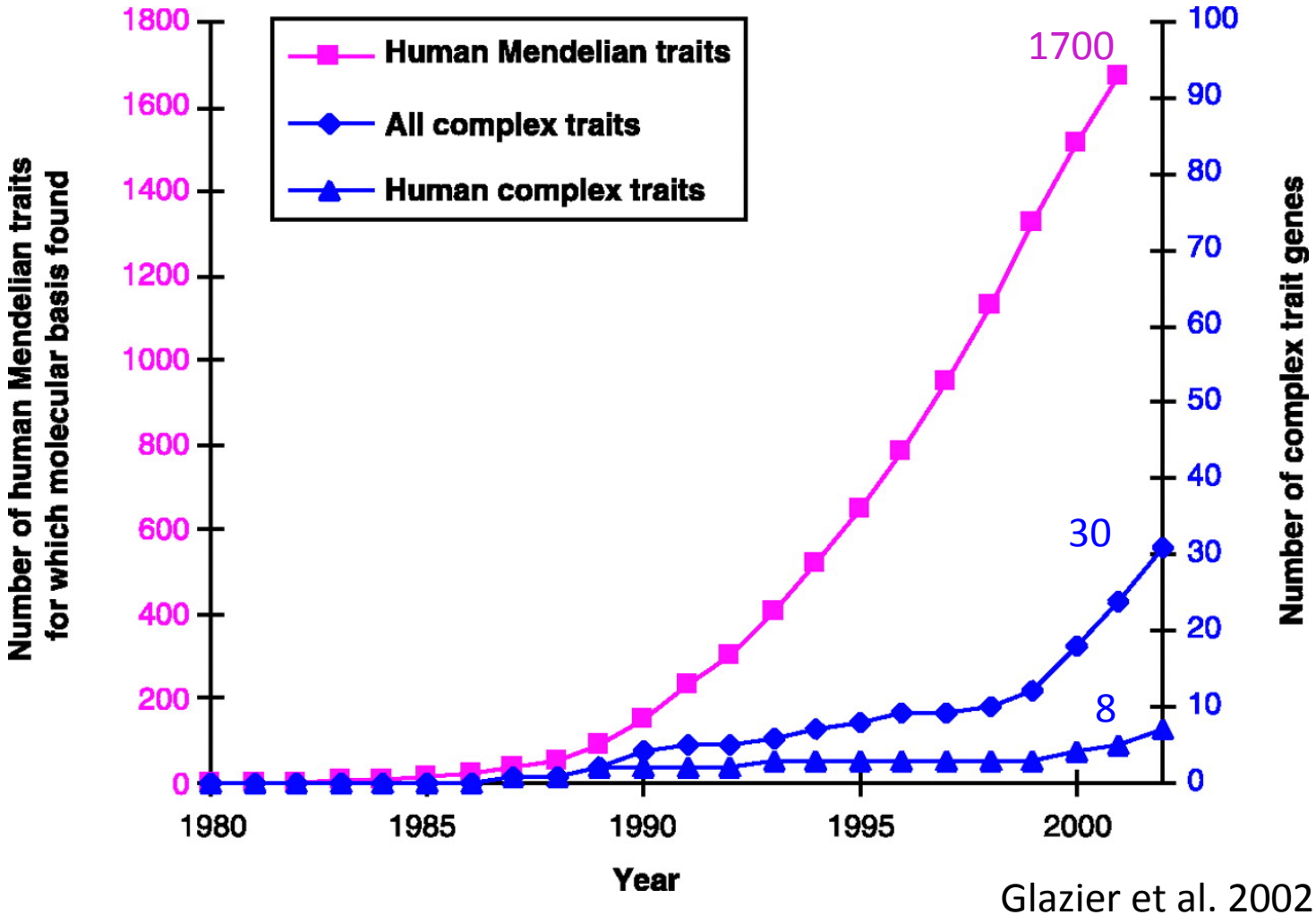
Risk of schizophrenia (%)

	Population	1st degree relative	MZ twin
Schizophrenia	1	10	50

Heritability = ~80%

Complex traits such as height, BMI and SCZ are highly heritable.

Identifying genes underlying complex traits



Glazier et al. 2002 Science

8 genes for human complex traits before 2002

Genome-wide Association Study (GWAS)

Linear model (simple regression)

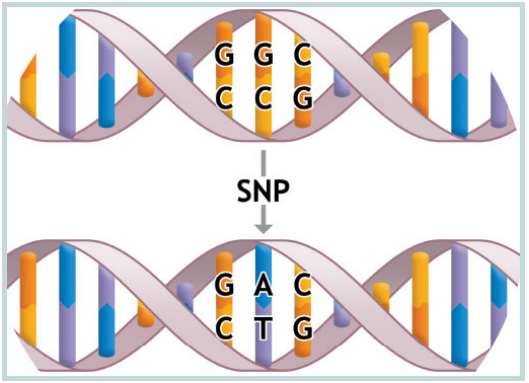
$y = b_0 + x_1 b_1 + e$

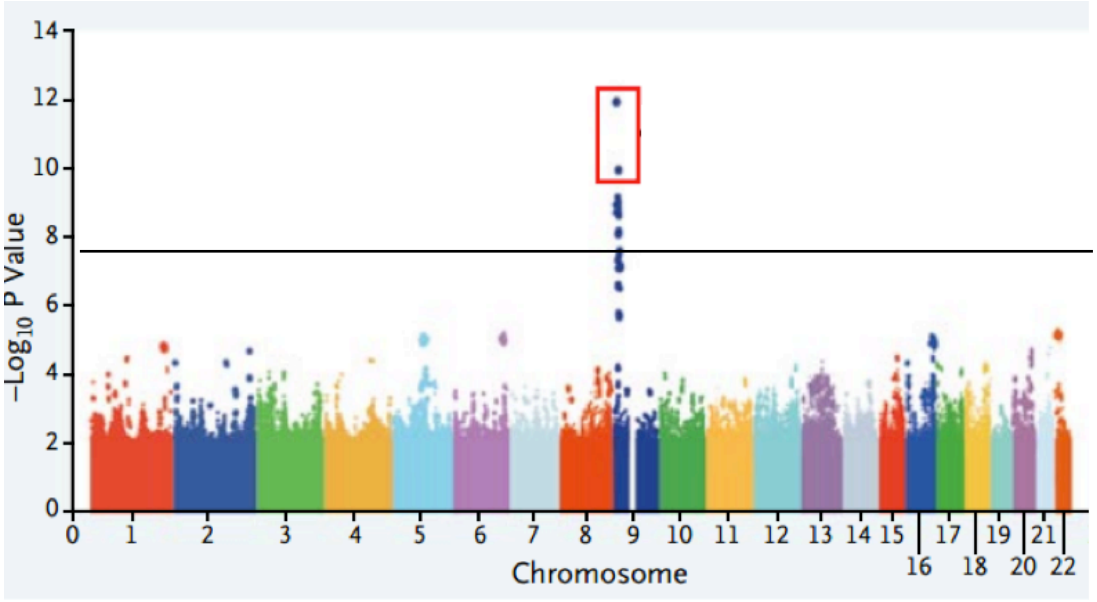
y = trait value

x_1 = SNP genotype (0, 1 or 2)

b_1-hat = $\mathbf{X}_1^T\mathbf{y} / (\mathbf{X}_1^T\mathbf{X}_1) = \text{cov}(x_1,y) / \text{var}(x_1)$

$SE^2(b_1\text{-hat}) = \sigma^2_e / [n\ \text{var}(x_1)]$



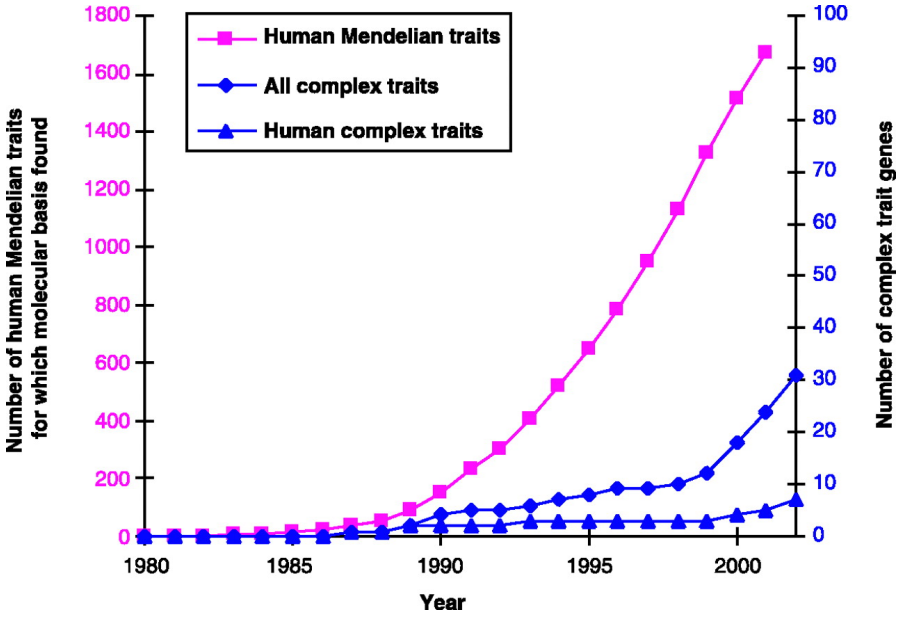


Genome-wide threshold P = 5×10^{-8}

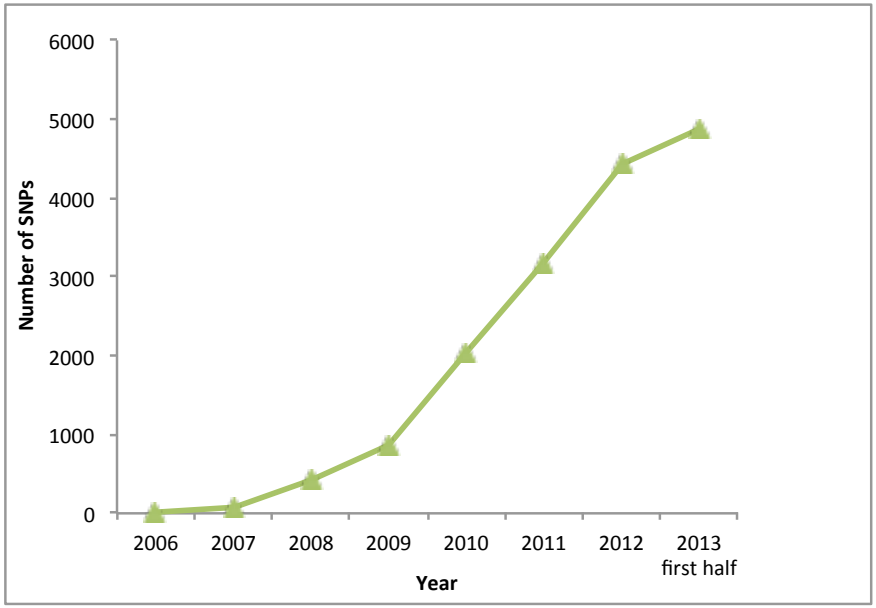
Manolio 2010 NEJM

An explosion of gene discoveries

Prior to GWAS



GWAS



Glazier et al. 2002 Science

~5000 genetic variants associated with ~650 traits / diseases

The missing heritability problem

Height:

- 180 loci
- ~180K samples
- < 10% of variance explained
- heritability = ~80%

Lango Allen et al. 2010 Nature

Schizophrenia:

- 22 loci
- ~21K cases / ~38K controls
- < 3% of variance explained
- heritability = ~80%

Ripke et al. 2013 Nat Genet

BMI:

- 32 loci
- ~250K samples
- ~1% of variance explained
- Heritability = 40% ~ 60%

Speliotes et al. 2010 Nat Genet

Fitting all SNPs in a MLM

- $\mathbf{y} = \mathbf{Wu} + \mathbf{e}$
  $\mathbf{W} = \{w_{ij}\}_{n \times m}$, w_{ij} = standardised SNP genotype
  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_u)$
  $\text{var}(\mathbf{y}) = \mathbf{ZZ}^T\sigma^2_u + \mathbf{I}\sigma^2_e$
  variance explained $= m\sigma^2_u / (m\sigma^2_u + \sigma^2_e)$

- Let $\mathbf{g} = \mathbf{Zu}$
  $\mathbf{y} = \mathbf{g} + \mathbf{e}$
  $\mathbf{g} \sim \mathbf{N(0, A}\sigma^2_g)$, $\mathbf{A}$ = genetic relationship matrix
  $\text{var}(\mathbf{y}) = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$
  variance explained $= \sigma^2_g / (\sigma^2_g + \sigma^2_e)$

- $\text{var}(\mathbf{y}) = (1/m)\mathbf{ZZ}^T(m\sigma^2_u) + \mathbf{I}\sigma^2_e$
  $A = \mathbf{ZZ}^T / m$

Reconciling family studies and GWAS

Family studies: comparing phenotypic similarity to family relatedness
– Our method: comparing phenotypic similarity to genetic similarity (estimated from SNPs) in unrelated individuals

GWAS: testing a SNP at a time in unrelated samples
– Our method: Estimating the contribution from all SNPs together

ANALYSIS

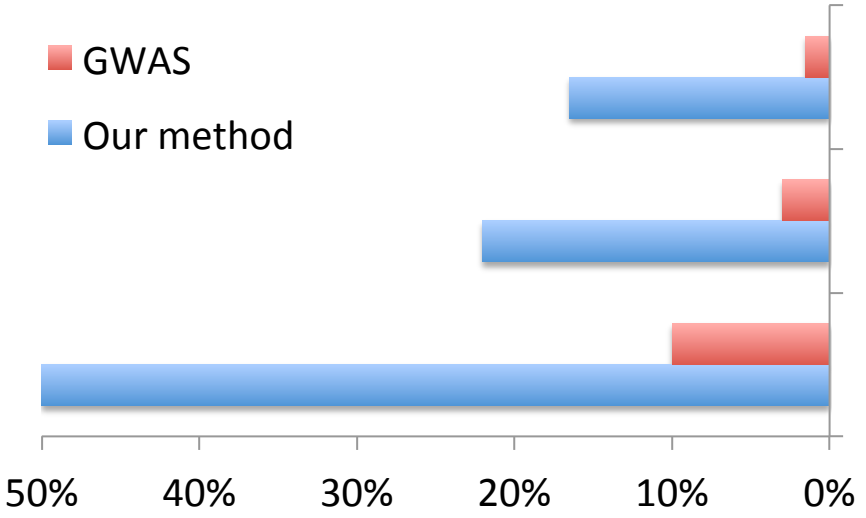
nature genetics

~50% of variation explained by all SNPs for height vs. ~10% from GWAS

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

GWAS vs All-SNP estimation



Obesity (BMI)	Yang et al. 2011 Nat Genet
Schizophrenia	Lee et al. 2012 Nat Genet
Height	Yang et al. 2010 Nat Genet

Many genetic variants each with a small effect contributing to the trait variation

Genome partitioning

- Single component MLM

  $\mathbf{y} = \mathbf{g} + \mathbf{e}$ (or $\mathbf{y} = \mathbf{W}\mathbf{u} + \mathbf{e}$)

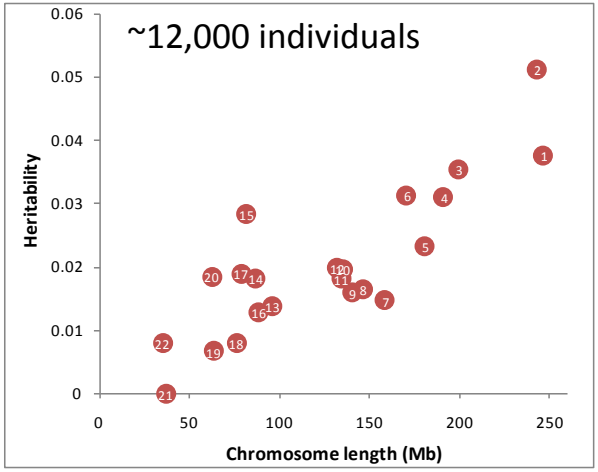
- Multi-component MLM

  $\mathbf{y} = \mathbf{g}_1 + \mathbf{g}_2 + \ldots + \mathbf{g}_{22} + \mathbf{e}$

  $\text{var}(\mathbf{y}) = \mathbf{A}_1\sigma^2_{g1} + \mathbf{A}_2\sigma^2_{g2} + \ldots + \mathbf{A}_{22}\sigma^2_{g22} + \mathbf{I}\sigma^2_{e}$

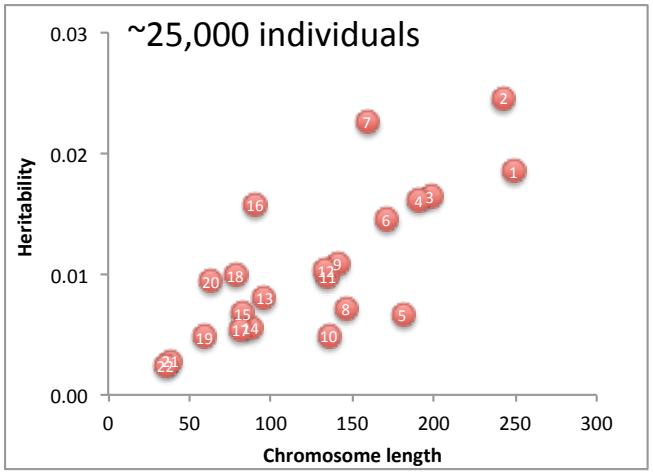
Partitioning the genetic variance into individual chromosomes

Height



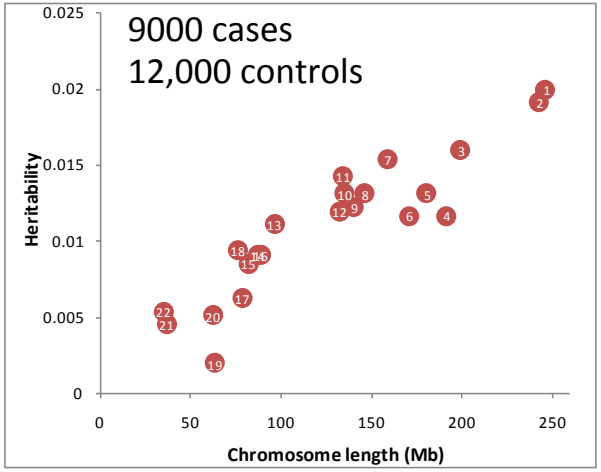
Yang et al. 2011 Nat Genet

BMI



Yang et al. unpublished

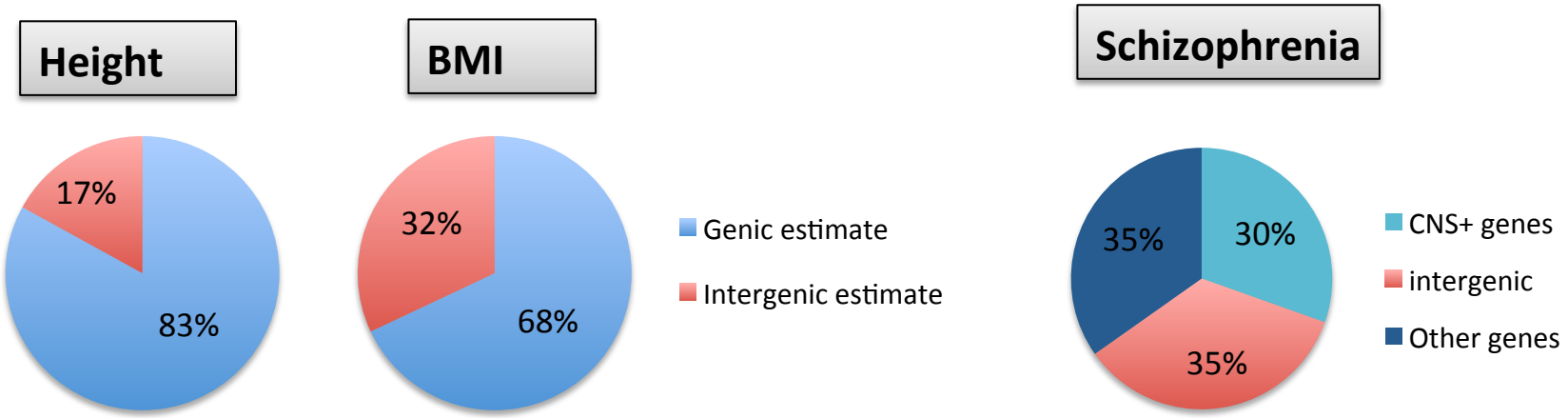
Schizophrenia



Lee et al. 2012 Nat Genet

Genetic variants distributed across the whole genome

Partitioning the genetic variance based on functional annotation



Yang et al. 2011 Nat Genet

Lee et al. 2012 Nat Genet

Genetic signals are enriched in or close to functional genes

More …

- Bivariate analysis – estimating the genetic correlation between two traits or two diseases using SNP data (Deary et al. 2012 Nature; Lee et al. 2013 Nat Genet)

- Fitting a mixture distribution rather than a single normal distribution to the random effects – e.g. Zhou et al. 2013 PLoS Genet

Linear model (simple regression based association test)

$y = b_0 + x_1 b_1 + e$
y = trait value; x_1 = SNP genotype (0, 1 or 2)

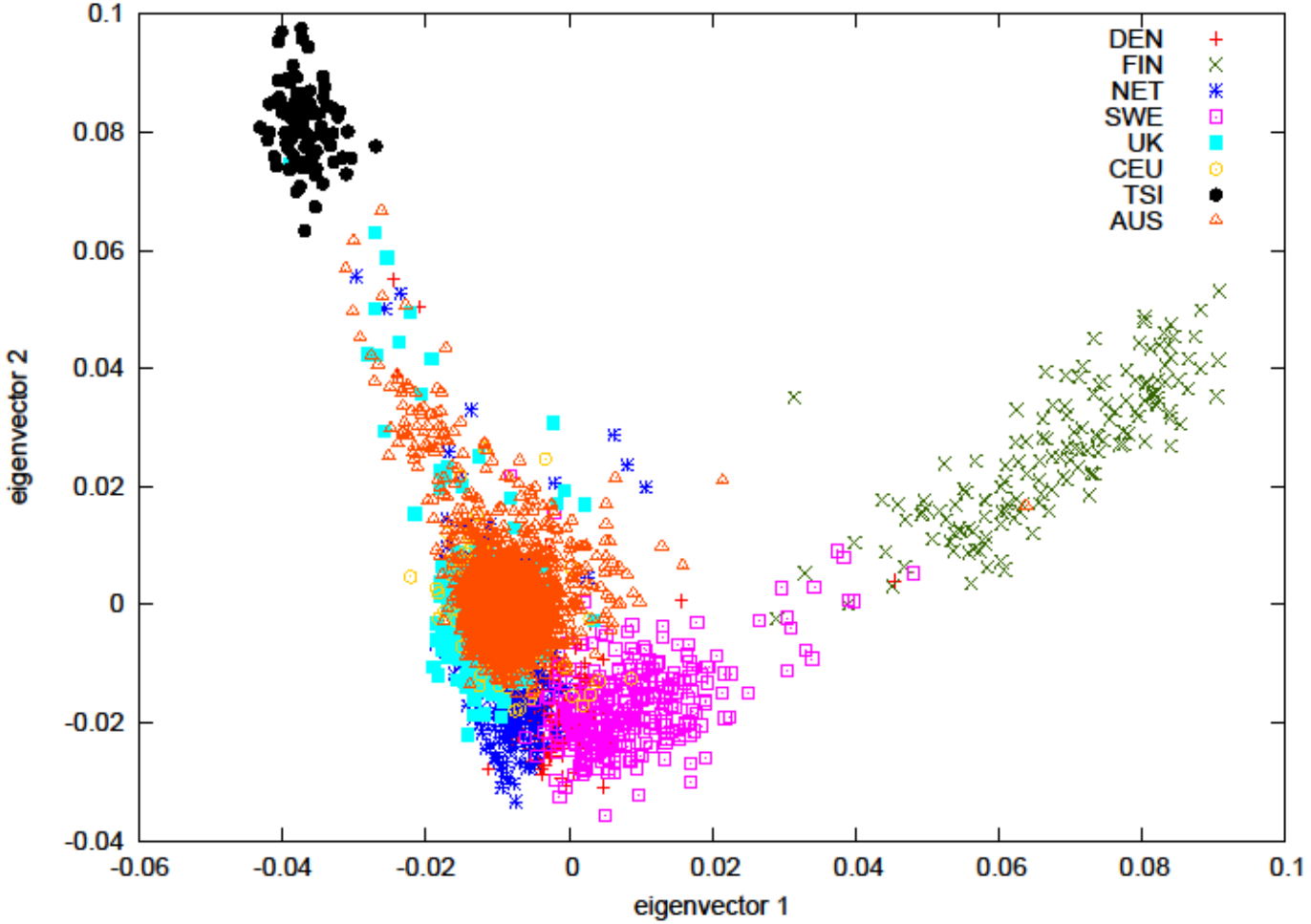
b_1-hat = $\mathbf{X}_1^T\mathbf{y} / (\mathbf{X}_1^T\mathbf{X}_1) = \text{cov}(x_1,y) / \text{var}(x_1)$
SE(b_1-hat) = $\sigma^2_e / [n \text{ var}(x_1)]$

Assumption: e is independent and identically distributed

Issues:

1) Relatedness: there are relatives in the sample – inflated false positive rate
2) Population stratification: individuals of different ancestries – spurious association; e.g. trait = eating with chopsticks, data = a random sample of US population.

Population stratification estimated from SNP data



Solution: MLM based association analysis

- $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$ or $\mathbf{y} = \mathbf{Xb} + \mathbf{g} + \mathbf{e}$

  $\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$

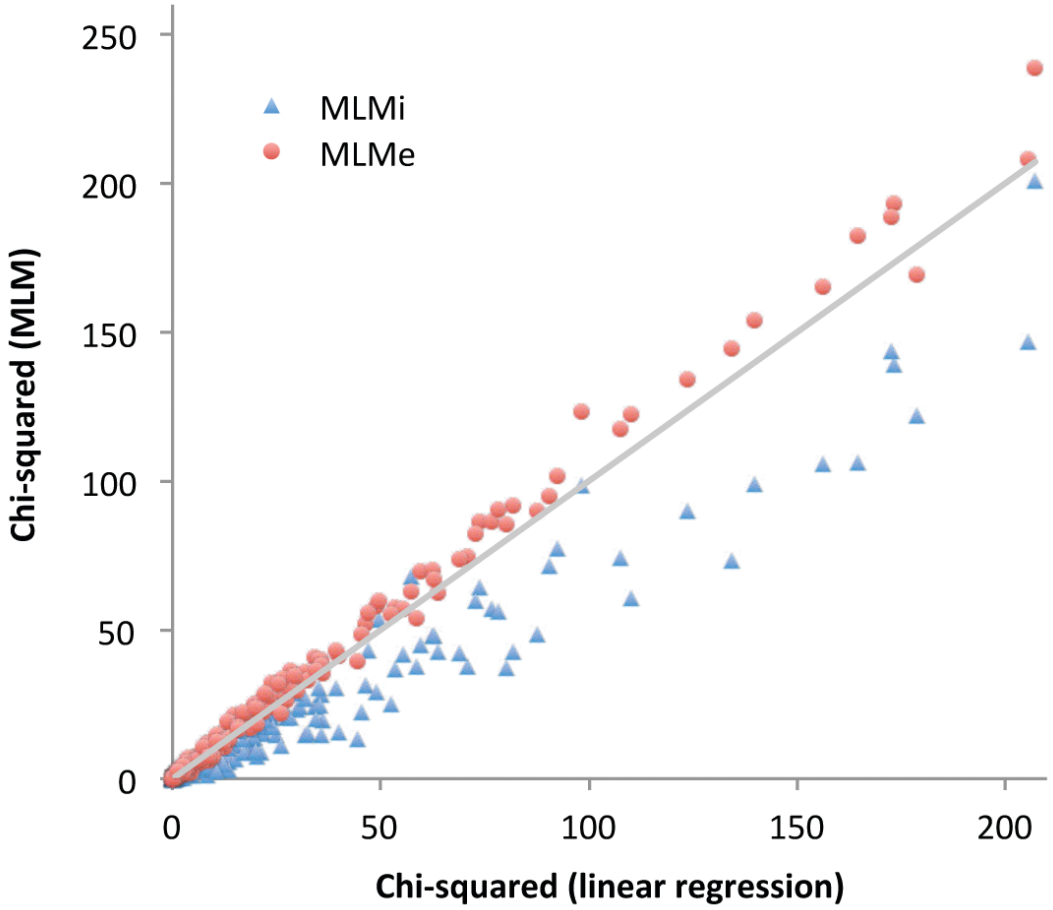
- Testing for fixed effects given sample structure

  $\mathbf{b}\text{-hat} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

  $\text{var}(\mathbf{b}\text{-hat}) = \sigma^2_e(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$

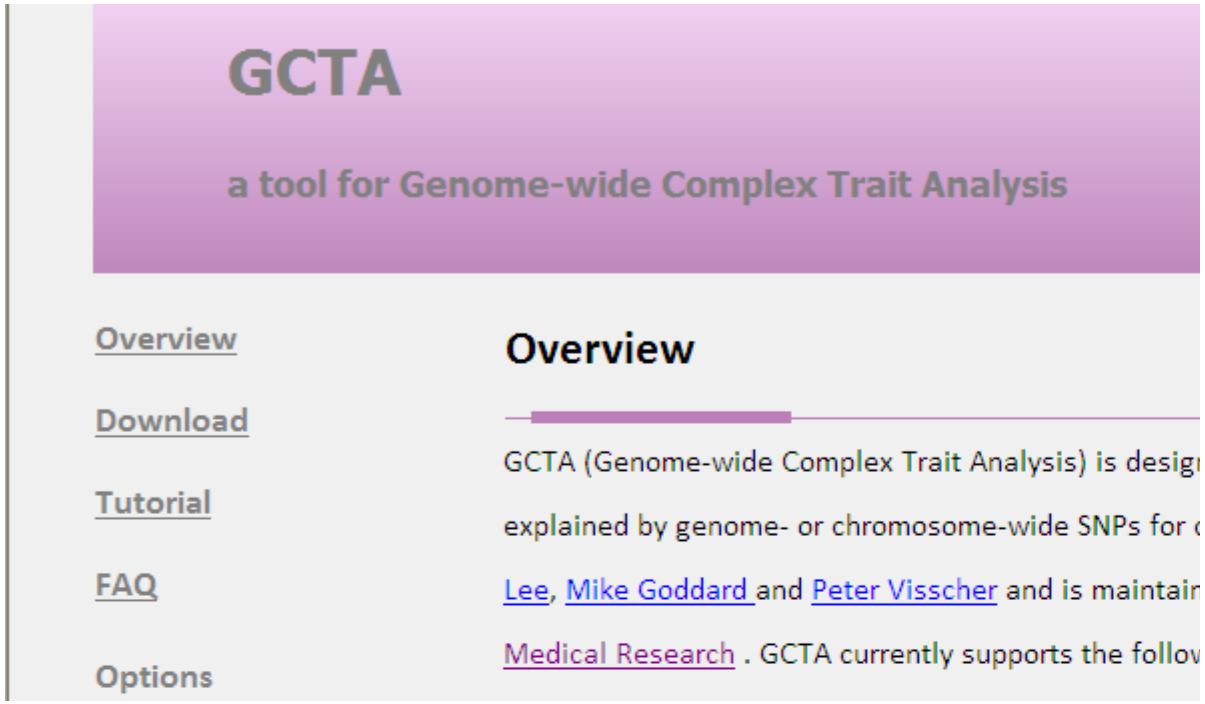
- Issue: a SNP is fitted twice.

Excluding the SNP from calculating the genetic relationship matrix



Software tool

http://gump.qimr.edu.au/gcta/



REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

Acknowledgements

Complex Traits Genomics Group (UQ)

- Peter Visscher
- Naomi Wray
- Hong Lee

University of Melbourne

- Mike Goddard

QIMR cohort

- Nick Martin
- Grant Montgomery

GENEVA Consortium

- Teri Manolio
- Bruce Weir

dbGaP

Australian Government
National Health and Medical Research Council
NHMRC

Australian Government
Australian Research Council

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

The Australian Neurogenetics Conference

at the Queensland Brain Institute (QBI), The University of Queensland, on September 11th and 12th, 2014

http://web.qbi.uq.edu.au/anc2014/