Complex diseases and heritability
Motivation/introduction

1. GWAS: single variant tests for common genetic variants, genetic effects tagged by linkage disequilibrium (LD)
Motivation/introduction

1. GWAS: single variant tests for common genetic variants, genetic effects tagged by linkage disequilibrium (LD)

2. Complex disease = complex architecture
   • # genome-wide significance hits $\sim \log(n)$, small odds ratios
   • example human height: first hits explained only 5% of phenotypic variance, family studies estimated $\sim 80\%$ caused by genetics
Motivation/introduction

1. GWAS: single variant tests for common genetic variants, genetic effects tagged by linkage disequilibrium (LD)

2. Complex disease = complex architecture
   • # genome-wide significance hits ~\log(n), small odds ratios
   • example human height: first hits explained only 5% of phenotypic variance, family studies estimated ~80% caused by genetics
   → Architecture discussions: „missing heritability“ → common and rare variants contribute Visscher et al., AJHG 2012
1. GWAS: single variant tests for common genetic variants, genetic effects tagged by linkage disequilibrium (LD)

2. Complex disease = complex architecture
   • # genome-wide significance hits $\sim \log(n)$, small odds ratios
   • example human height: first hits explained only 5% of phenotypic variance, family studies estimated $\sim 80\%$ caused by genetics

   $\Rightarrow$ Architecture discussions: „missing heritability“ $\Rightarrow$ common and rare variants contribute Visscher et al., AJHG 2012

3. Idea: use all available SNPs to infer genetic architecture based on heritability estimates
Heritability $h^2$

Quantitative trait $y \sim N(0, \sigma_p^2)$, where Yang et al., Nat Genet 2010

$y = g + e$

Genetic effect $g \sim N(0, \sigma_g^2)$
Enviromental effect $e \sim N(0, \sigma_e^2)$
Heritability $h^2$

Quantitative trait $y \sim N(0, \sigma_{p}^2)$, where Yang et al., Nat Genet 2010

$$y = g + e$$

Genetic effect $g \sim N(0, \sigma_{g}^2)$, environmental effect $e \sim N(0, \sigma_{e}^2)$

Heritability

$$h_{full}^2 := \sigma_{g}^2 / \sigma_{p}^2 = \sigma_{g}^2 / \sigma_{g}^2 + \sigma_{e}^2$$

Narrow-sense heritability

$$h^2 := \sigma_{g, \text{additive}}^2 / \sigma_{p}^2$$
Estimate heritability based on genotype data: GREML

Estimate narrow-sense heritability $h_{SNPs}^2$ based on $n$ samples and $m$ SNPs via

$$y = g + e = Wu + e$$

where $e \sim N(0, \sigma_e^2 I_n)$, $u \sim N(0, \sigma_u^2 I_n)$ and $W_{ij} = g_{ij} - 2p_{ij} / \sqrt{2p_{ij}(1-p_{ij})}$

Yang et al., Nat Genet 2010
Estimate heritability based on genotype data: GREML

Estimate narrow-sense heritability $h^2_{SNPs \uparrow 2}$ based on $n$ samples and $m$ SNPs via

\[ y = g + e = Wu + e \]

where $e \sim N(0, \sigma_g^2 I \downarrow n)$, $u \sim N(0, \sigma_u^2 I \downarrow n)$ and $W_{ij} = g_{ij} - 2p_{ij} / \sqrt{2p_{ij}(1-p_{ij})}$

Yang et al., Nat Genet 2010

assumptions:

- samples unrelated
- $Wu \approx W_{causal} u_{causal}$
- effect size $\sim 1/\sqrt{p(1-p)}$ (selection)
Estimate heritability based on genotype data: GREML

Estimate $\sigma_{g^2}$ and $\sigma_{e^2}$ via REML based on

$$\text{Var} (y) = \sigma_{g^2} G + \sigma_{e^2} I_n,$$

where $G = WW^T / m$ denotes $n \times n$ Genetic Relationship Matrix (GRM), $\sigma_{g^2} = m \sigma_{u^2}$
Estimate heritability based on genotype data: GREML

Estimate $\sigma^2_g$ and $\sigma^2_e$ via REML based on

$$\rightarrow \text{Var}(y) = \sigma^2_g G + \sigma^2_e I \downarrow n,$$

where $G = W W^\top / m$ denotes $n \times n$ Genetic Relationship Matrix (GRM), $\sigma^2_g = m \sigma^2_u$

recall $G = W W^\top / m \approx W \downarrow \text{causal} W \downarrow \text{causal}^\top / m \downarrow \text{causal}$
Estimate heritability based on genotype data: GREML

Estimate $\sigma^2_g$ and $\sigma^2_e$ via REML based on

$$\rightarrow \text{Var}(y) = \sigma^2_g G + \sigma^2_e I \downarrow n,$$

where $G = WW^\top / m$ denotes $n \times n$ Genetic Relationship Matrix (GRM), $\sigma^2_g = m\sigma^2_u$

recal $G = WW^\top / m \approx W\downarrow \text{causal} W\downarrow \text{causal}^\top / m\downarrow \text{causal}$

Implemented in GCTA software (cnsgenomics.com/software/gcta)
Some remarks
Some remarks

- GRM $G$ is NOT the sample covariance matrix of genotypes
Some remarks

• GRM $G$ is NOT the sample covariance matrix of genotypes

• Model accounts for LD, problems of the GRM Kumar et al., PNAS 2016; Yang et al., Nat Genet 2017
Some remarks

• GRM $G$ is NOT the sample covariance matrix of genotypes

• Model accounts for LD, problems of the GRM Kumar et al., PNAS 2016; Yang et al., Nat Genet 2017

• Expected heritability depends on SNP set (tagging argument):
  
  \[ h_{\text{array}}^2 < h_{\text{imputed}}^2 < h_{\text{WGS}}^2 \approx h^2 \]  
  
  (large sample size required)  

  Yang et al., Nat Genet 2015
Some remarks

- GRM $G$ is NOT the sample covariance matrix of genotypes

- Model accounts for LD, problems of the GRM Kumar et al., PNAS 2016; Yang et al., Nat Genet 2017

- Expected heritability depends on SNP set (tagging argument):
  \[ h_{array}^2 < h_{imputed}^2 < h_{WGS}^2 \approx h^2 \]  (large sample size required)
  Yang et al., Nat Genet 2015

- Population stratification and imputing differences cause problems
Extensions

- $\gamma$ affection status:
  \[\gamma = 1 \text{ iff } g + e > t \text{ and } \gamma = 0 \text{ otherwise, where } t \text{ is } (1-K) \text{ quantile of the standard normal distribution} \]

  difference: observed scale/ liability scale
Extensions

• $y$ affection status:

\[ y = 1 \text{ iff } g + e > t \text{ and } y = 0 \text{ otherwise}, \text{ where } t \text{ is } (1 - K) \text{ quantile of the standard normal distribution} \]

Lee et al., AJHG 2011

difference: observed scale/ liability scale

• Genetic covariance/genetic correlation between two diseases

Cross-Disorder Group PGC et al, Nat Genet 2013; Yang et al., AJHG 2011
Extensions

- $y$ affection status:
  
  $y=1$ iff $g+e > t$ and $y=0$ otherwise, where $t$ is $(1-K)$ quantile of the standard normal distribution Lee et al., AJHG 2011
  
  difference: observed scale/ liability scale

- Genetic covariance/genetic correlation between two diseases
  Cross-Disorder Group PGC et al, Nat Genet 2013; Yang et al., AJHG 2011

- Implicit assumption: causal variants uniformly distributed
  
  recall $WW^\uparrow /m \approx W\downarrow causal W\downarrow causal^\uparrow /m\downarrow causal$ assumption
  
  might be biased $\rightarrow$ GREML-LDMS with 28 (7x4) categories to estimate heritabilities in MAF and LD bins separately Yang et al., Nat Genet 2015
Some results

First GCTA analysis:
  • human height $h\downarrow \text{array} \uparrow_2 \sim 45\%$

Genetic correlation
  • Genetic correlation SCZ/BIP $\sim 68\%$

GREML-LDMS:
  • imputing captures $\sim 97\%$ of common genetic variation
  • Human height: $h\downarrow \text{imputed} \uparrow_2 \sim 55\%, h\downarrow \text{common} \uparrow_2 \sim 47\%, h\downarrow \text{rare} \uparrow_2 \sim 8\%$
  • BMI: $h\downarrow \text{imputed} \uparrow_2 \sim 27\%, h\downarrow \text{common} \uparrow_2 \sim 25\%, h\downarrow \text{rare} \uparrow_2 \sim 2\%$
  • missing heritability $\rightarrow$ hidden heritability
  • Evolutionary theory: height-associated variants have been under selection
References


- Five years of GWAS discovery. Visscher PM et al. AJHG 2012

- Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Cross-Disorder Group PGC et al. Nat Genet 2013

- Common SNPs explain a large proportion of the heritability. Yang et al. 2010. Nat Genet 2010
Additional approaches/results

**Mixed model heritability estimation:** Yang/Visscher (Queensland, Australia), Price/Loh (Boston), Heckerman/Listgarten (Amazon/Berkeley), Speed/Balding (London/Melbourne)

**Mixed models also for:**
- Prediction (connection to PRS)
- Association testing

**Regression approach (more robust):** PCGC regression (Golan et al., PNAS 2014)

**Heritability estimation based on summary statistics:** LD Score regression (Bulik-Sullivan et al., Nat Genet 2015)