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 ~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 $\sim\log(n)$, 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

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 ~80% caused by genetics

   → Architecture discussions: „missing heritability“ → 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 $g \sim N(0, \sigma_g^2)$

enviromental 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

$h^2 := \sigma_{g,additive}^2 / \sigma_p^2$

Estimate heritability based on genotype data: GREML

Estimate narrow-sense heritability $h\downarrow SNPs\uparrow 2$ based on n samples and m SNPs via

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

where $e\sim N(0,\sigma\downarrow e\uparrow 2\ I\downarrow n)$, $u\sim N(0,\sigma\downarrow u\uparrow 2\ I\downarrow n)$ and $W\downarrow ij=g\downarrow ij-2p\downarrow j/\sqrt{2p\downarrow j}(1-p\downarrow j)$

Yang et al., Nat Genet 2010

Estimate heritability based on genotype data: GREML

Estimate narrow-sense heritability $h\downarrow SNPs\uparrow 2$ based on n samples and m SNPs via

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

where $e\sim N(0,\sigma\downarrow g\uparrow 2\ I\downarrow n\ )$, $u\sim N(0,\sigma\downarrow u\uparrow 2\ I\downarrow n\ )$ and $W\downarrow ij=g\downarrow ij-2p\downarrow j\ /\sqrt{2p\downarrow j\ (1-p\downarrow j\ )}$

Yang et al., Nat Genet 2010

assumptions:

- samples unrelated
- $Wu\approx W\downarrow causal\ u\downarrow causal$
- effect size $\sim 1/\sqrt{p(1-p)}$ (selection)

Estimate heritability based on genotype data: GREML

Estimate $\sigma\downarrow g\uparrow 2$ and $\sigma\downarrow e\uparrow 2$ via REML based on

$$\rightarrow \text{Var}(y)=\sigma\downarrow g\uparrow 2\ G+\sigma\downarrow e\uparrow 2\ I\downarrow n\ ,$$

where $G=WW\uparrow T\ /m$ denotes $n\ x\ n$ Genetic Relationship Matrix (GRM), $\sigma\downarrow g\uparrow 2 =m\sigma\downarrow u\uparrow 2$

Estimate heritability based on genotype data: GREML

Estimate $\sigma\downarrow g\uparrow 2$ and $\sigma\downarrow e\uparrow 2$ via REML based on

$$\rightarrow \text{Var}(y)=\sigma\downarrow g\uparrow 2\ G+\sigma\downarrow e\uparrow 2\ I\downarrow n\ ,$$

where $G=WW\uparrow T\ /m$ denotes $n\ x\ n$ Genetic Relationship Matrix (GRM), $\sigma\downarrow g\uparrow 2 =m\sigma\downarrow u\uparrow 2$

recall $G=WW\uparrow T\ /m\approx W\downarrow causal\ W\downarrow causal\uparrow T\ /m\downarrow causal$

Estimate heritability based on genotype data: GREML

Estimate $\sigma\downarrow g\uparrow 2$ and $\sigma\downarrow e\uparrow 2$ via REML based on

$$\rightarrow \mathrm{Var}(y)=\sigma\downarrow g\uparrow 2\ G+\sigma\downarrow e\uparrow 2\ I\downarrow n\ ,$$

where $G=WW\uparrow T\ /m$ denotes $n\ x\ n$ Genetic Relationship Matrix (GRM), $\sigma\downarrow g\uparrow 2 =m\sigma\downarrow u\uparrow 2$

recall $G=WW\uparrow T\ /m\approx W\downarrow causal\ W\downarrow causal\uparrow T\ /m\downarrow causal$

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):

  $h_{array}^2 < h_{imputed}^2 < h_{WGS}^2 \approx h^2$ (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

- 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

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

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 $WW↑/m \approx W↓causal\, W↓causal↑/m↓causal$ assumption

  might be biased → 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 $h\downarrow array\uparrow 2 \sim 45\%$

Genetic correlation

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

GREML-LDMS:

- imputing captures ~97% of common genetic variation
- Human height: $h\downarrow imputed\uparrow 2 \sim 55\%$, $h\downarrow common\uparrow 2 \sim 47\%$, $h\downarrow rare\uparrow 2 \sim 8\%$
- BMI: $h\downarrow imputed\uparrow 2 \sim 27\%$, $h\downarrow common\uparrow 2 \sim 25\%$, $h\downarrow rare\uparrow 2 \sim 2\%$
- missing heritability → hidden heritability
- Evolutionary theory: height-associated variants have been under selection

References

- Concepts, estimation and interpretation of SNP-based heritability. Yang J et al. Nat Genet. 2017
- Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Yang J et al. Nat Genet. 2015
- Estimating missing heritability for disease from genome-wide association studies. Lee SH et al. AJHG 2011.
- GCTA: a tool for genome-wide complex trait analysis. Yang J et al. AJHG 2011
- 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)