Estimation of genetic variation and SNP-heritability from GWAS data
Key concepts

• Dense SNP panels allow the estimation of the expected genetic covariance between distant relatives
• A model based upon estimated relationships from SNPs is equivalent to a model fitting all SNPs simultaneously
• The total genetic variance due to LD between common SNPs and (unknown) causal variants can be estimated
• Genetic variance captured by common SNPs can be partitioned across the genome
• Different methods to estimate relatedness from SNPs assume different genetic trait architectures
Estimation of SNP-heritability from GWAS data

Background

– 2008: GWAS was perceived by many to have failed as an experimental design

– Missing heritability: discrepancy between pedigree heritability and variance captured by associated SNPs
<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of loci</th>
<th>Percent of Heritability Measure Explained</th>
<th>Heritability Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-related macular degeneration</td>
<td>5</td>
<td>50%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>32</td>
<td>20%</td>
<td>Genetic risk (liability)</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>6</td>
<td>15%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>18</td>
<td>6%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>7</td>
<td>5.2%</td>
<td>Phenotypic variance</td>
</tr>
<tr>
<td>Height</td>
<td>40</td>
<td>5%</td>
<td>Phenotypic variance</td>
</tr>
<tr>
<td>Early onset myocardial infarction</td>
<td>9</td>
<td>2.8%</td>
<td>Phenotypic variance</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>4</td>
<td>1.5%</td>
<td>Phenotypic variance</td>
</tr>
</tbody>
</table>

Where is the Dark Matter?
Hypothesis testing vs. Estimation

GWAS = hypothesis testing
  – Stringent p-value threshold
  – Estimates of effects biased ("Winner’s Curse")

Can we estimate the total proportion of variation accounted for by all SNPs?
A model for a single causal variant

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>(1-p)^2</td>
<td>2p(1-p)</td>
<td>p^2</td>
</tr>
<tr>
<td>x</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>effect</td>
<td>0</td>
<td>b</td>
<td>2b</td>
</tr>
<tr>
<td>w = [x-E(x)]/σ_x</td>
<td>-2p/√{2p(1-p)}</td>
<td>(1-2p)/ √{2p(1-p)}</td>
<td>2(1-p)/ √{2p(1-p)}</td>
</tr>
</tbody>
</table>

\[ y_j = \mu' + x_{ij}b_i + e_j \quad x = 0, 1, 2 \{\text{standard association model}} \]

\[ y_j = \mu + w_{ij}u_j + e_j \quad u = b\sigma_x; \mu = \mu' + b\sigma_x \]
Multiple (M) causal variants

\[ y_j = \mu + \sum w_{ij} u_j + e_j \]

= \[ \mu + g_j + e_j \]

\[ y = \mu_1 + g + e \]

= \[ \mu_1 + Wu + e \]
Equivalence

Let \( u \) be a random variable, \( u \sim N(0, \sigma_u^2) \)

Then \( \sigma_g^2 = M\sigma_u^2 \)

\[
\text{var}(y) = WW' \sigma_u^2 + I\sigma_e^2 \\
= WW'(\sigma_g^2/M) + I\sigma_e^2 \\
= G\sigma_g^2 + I\sigma_e^2
\]

Model with individual genome-wide additive values using relationships \((G)\) at the causal variants is equivalent to a model fitting all causal variants

We can estimate genetic variance just as if we would do using pedigree relationships
If we estimate $\mathbf{G}$ from SNPs:

– lose information due to imperfect LD between SNPs and causal variants
– how much we lose depends on
  • density of SNPs
  • allele frequency spectrum of SNPs vs. causal variants
– estimate of variance $\rightarrow$ missing heritability

$\mathbf{G}$ from $M$ SNPs:

$$G_{jk} = \frac{(1/M) \sum (x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

$$= \frac{(1/M) \sum w_{ij}w_{ik}}{2p_i(1-p_i)}$$
Methods (Yang et al. 2010)

- Estimate realised relationship matrix from SNPs
- Estimate additive genetic variance

\[ y = Xb + e = Wu + e, \text{ var}(y) = G\sigma_g^2 + I\sigma_e^2 \]

\[ G_{jk} = (1/M) \sum \{ x_{ij} - 2p_i \}(x_{ik} - 2p_i) / \{2p_i(1-p_i)\} \]

\[ = (1/M) \sum w_{ij}w_{ik} \]

- Base population = current population
- Weighting scheme 1
Statistical analysis

$$\text{var}(y) = V = G\sigma_g^2 + I\sigma_e^2$$

\(y\) standardised \(\sim\) \(N(0,1)\)

No fixed effects other than mean

\(G\) estimated from SNPs

Residual maximum likelihood (REML)
Results

$h^2 \sim 0.5$ (SE 0.1)

[Yang et al. 2010, Nature Genetics]
Checking for population structure

### Table 1

**Estimates of the Variance Explained by the SNPs on Even Chromosomes from 10 Simulation Replicates**

<table>
<thead>
<tr>
<th>Replicate</th>
<th>$h^2$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.045</td>
<td>0.055</td>
</tr>
<tr>
<td>2</td>
<td>0.025</td>
<td>0.057</td>
</tr>
<tr>
<td>3</td>
<td>0.0</td>
<td>0.058</td>
</tr>
<tr>
<td>4</td>
<td>0.0</td>
<td>0.057</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>0.059</td>
</tr>
<tr>
<td>6</td>
<td>0.0</td>
<td>0.056</td>
</tr>
<tr>
<td>7</td>
<td>0.057</td>
<td>0.056</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
<td>0.062</td>
</tr>
<tr>
<td>9</td>
<td>0.0</td>
<td>0.057</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
<td>0.054</td>
</tr>
</tbody>
</table>

*Note:* A total of 1,000 causal variants were simulated on the odd chromosomes, with a total heritability of 0.8. Genetic variance was estimated from a relationship matrix constructed from all SNPs on the even chromosomes. The same genotypes were used as in Yang et al. (2010). If there is population structure then estimated relatedness on the even chromosomes is correlated with relatedness on the odd chromosomes (where the causal variants are simulated) and therefore genetic variance will be associated with the even chromosomes.
Genetic variance associated with all SNPs can be estimated from GWAS data

- use SNPs to estimate $G$
- use phenotypes on “unrelated” individuals and $G$ to estimate genetic variance

Empirical results: most additive genetic variation for height is captured by common SNPs

- little ‘missing’ heritability
- GWAS works fine

Conclusions Yang et al. 2010
Partitioning of genetic variation

\[ y = \text{mean} + g_1 + g_2 + g_3 + g_4 + g_5 + e \]
\[ \text{var}(g_i) = (\mathbf{W}_i\mathbf{W}_i'/M_i)\sigma_i^2 \text{ for SNPs in group } i \]

Examples of groupings:
- chromosome
- genome annotation
- MAF
- LD
Application (2): partitioning variation

If we can estimate the variance captured by SNPs genome-wide, we should be able to partition it and attribute variance to regions of the genome.

“Population based linkage analysis”
**Example on quantitative traits**

<table>
<thead>
<tr>
<th>Trait</th>
<th>n</th>
<th>$h^2_G$ (s.e.)</th>
<th>$P$</th>
<th>$h^2_G$ (s.e.)</th>
<th>$P$</th>
<th>Heritability</th>
<th>GWAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>11,576</td>
<td>0.448 (0.029)</td>
<td>4.5 x 10^{-69}</td>
<td>0.419 (0.030)</td>
<td>7.9 x 10^{-48}</td>
<td>80–90%^{32}</td>
<td>~10%^{23}</td>
</tr>
<tr>
<td>BMI</td>
<td>11,558</td>
<td>0.165 (0.029)</td>
<td>3.0 x 10^{-10}</td>
<td>0.159 (0.029)</td>
<td>5.3 x 10^{-9}</td>
<td>42–80%^{25,26}</td>
<td>~1.5%^{14}</td>
</tr>
<tr>
<td>vWF</td>
<td>6,641</td>
<td>0.252 (0.051)</td>
<td>1.6 x 10^{-7}</td>
<td>0.254 (0.051)</td>
<td>2.0 x 10^{-7}</td>
<td>66–75%^{33,34}</td>
<td>~13%^{15}</td>
</tr>
<tr>
<td>QTi</td>
<td>6,567</td>
<td>0.209 (0.050)</td>
<td>3.1 x 10^{-6}</td>
<td>0.168 (0.052)</td>
<td>5.0 x 10^{-4}</td>
<td>37–60%^{35,36}</td>
<td>~7%^{16}</td>
</tr>
</tbody>
</table>

[Yang et al. 2011, Nature Genetics]
Partitioning on chromosomes

longer chromosomes explain more variation

[Yang et al. 2011, Nature Genetics]
Results are consistent with reported GWAS

**Variance explained by GIANT height SNPs on each chromosome**

- **BMI (11,586 unrelated)**
- **Variance explained by chromosome (adjusted for the FTO SNP)**
- **Variance explained by chromosome (no adjustment)**

\( R^2 = 0.511 \)

- **Height (11,586 unrelated)**

[Yang et al. 2011, Nature Genetics]
Inference robust with respect to genetic architecture

Variance explained by each chromosome vs. Chromosome length (Mb)

- vWF (ARIC)
  - Slope = 6.9 × 10^{-5}
  - P = 0.524
  - R^2 = 0.021

- vWF (6,662 unrelated)
  - ABO
  - Variances explained by chromosome (adjusted for the ABO SNP)
  - Variances explained by chromosome (no adjustment)

[Yang et al. 2011, Nature Genetics]
Partitioning on genome annotation

Genic regions explain variation disproportionately

**Height (combined)**
- 17,277 protein coding genes
- $h_{ge}^2 = 0.328$ (s.e. = 0.024)
- $h_{gi}^2 = 0.126$ (s.e. = 0.022)
- Coverage of genic regions = 49.4%
- $P$(observed vs. expected) = $2.1 \times 10^{-10}$

**BMI (combined)**
- 17,277 protein coding genes
- $h_{ge}^2 = 0.117$ (s.e. = 0.023)
- $h_{gi}^2 = 0.047$ (s.e. = 0.022)
- Coverage of genic regions = 49.4%
- $P$(observed vs. expected) = 0.022

---

[Yang et al. 2011, Nature Genetics]
Application (3): Using imputed sequence data

How much information is gained by using SNP array data imputed to a fully sequenced reference?

How much is lost relative to whole genome sequencing?

Partition variation according to MAF and LD

[Yang et al. 2015 Nature Genetics]
Accounting for LD and MAF spectrum allows unbiased estimation of genetic variance

[Yang et al. 2015 Nature Genetics]
Very little difference in “taggability” between SNP chips

Genetic variation captured after imputation:
96% due to common variants
73% due to rare variants

[Yang et al. 2015 Nature Genetics]
n = 45k data on height and BMI

Totals
~60% for height
~30% for BMI

[V Yang et al. 2015 Nature Genetics]
Partitioning variance of height

- Total variance
- Heritability (based on Twin or family studies)
- SNP heritability from imputation to sequenced reference
- SNP-heritability (variance explained by all genotyped SNPs on the Chip)
- Variance explained by genome wide significant SNPs

h² overestimation?
untagged rare variants?
better tagging of ungenotyped variants
sample size / power
Estimated relatedness and trait architecture

\[
\hat{G}_{jk} = \left( \frac{1}{\sum_{i=1}^{m} [2p_i(1 - p_i)]^{1+s}} \right) \sum_{i=1}^{m} (z_{ij}z_{ik}) [2p_i(1 - p_i)]^s
\]

If \( G \) describes the genetic covariance between individuals \( \text{var}(g) = G\sigma_g^2 \), then what is the equivalent linear model in terms of SNP effects?
Equivalent models

\[ \mathbf{y} = 1 \mu + \sum_{j}^{M} \mathbf{X}_j \beta_j + e \]

\[ \beta_j \sim N(0, \{2p_j(1 - p_j)\}^S \sigma^2_\beta) \]

\[ h_j^2 = 2p_j(1 - p_j)E(\beta^2) = \left\{2p_j(1 - p_j)\right\}^{1+S} \sigma^2_\beta \]
\[ S = -1 \]

\[ \beta_j \sim N(0, \{2p_j(1 - p_j)\}^{-1} \sigma^2_\beta) \]

\[ h^2_j = \{2p_j(1 - p_j)\}^0 \sigma^2_\beta = \sigma^2_\beta \]

- Weighting scheme 1
- All SNPs contribute equally to heritability
- Rare variants have bigger effects
- “Purifying selection model”
\[ S = 0 \]

\[ \beta_j \sim N \left( 0, \left\{ 2p_j (1 - p_j) \right\}^0 \sigma^2_\beta \right) \sim N(0, \sigma^2_\beta) \]

\[ h_j^2 = \left\{ 2p_j (1 - p_j) \right\}^1 \sigma^2_\beta \]

- Weighting scheme 2
- Common SNPs contribute more to heritability
- Rare and common variants have same effects
- "Neutral model"
Weighting scheme and genetic architecture

• Weighting schemes 1 and 2 can be justified in two ways:
  – IBD vs IBS
  – *A priori* assumption about the relationship between allele frequency of effect size (natural selection)

• Can we estimate genetic architecture from the data?
Bayesian mixture model (BayesS)

\[ y = 1\mu + \sum_j X_j \beta_j + e \]
\[ \beta_j \sim N \left( 0, \left( 2p_j q_j \right)^S \sigma^2_{\beta} \right) \pi + \phi(1 - \pi) \]

- S measures the relationship between effect size and MAF \( p_j \)
  - \( S = 0 \): independence
  - \( S < 0 \): negatively related (rare variant tends to have large effect)
  - \( S > 0 \): positively related (common variant tends to have large effect)
  - \textit{GCTA default}: \( S = -1 \)
- \( \pi \) is the polygenicity (proportion of SNPs with non-zero effects)
- \( h^2_{SNP} = \text{Var}(g)/\sigma^2_e \) where \( g = \sum_j X_j \beta_j \)
- Simultaneously estimate SNP effects and genetic architecture parameters using MCMC
- Account for LD between SNPs

Zeng .... Yang 2017 (BioRxiv)
Direction of $S$ distinguishes stabilising selection from directional and disruptive selection

$$\beta_j \sim N\left(0, 1 + \left(2p_j q_j\right)^S \sigma^2_\beta\right)$$
• Both height and BMI have been under selection
• Selection has been stronger for height than BMI-associated SNPs
• Height is more heritable than BMI.
• BMI is more polygenic than height.

Zeng .... Yang 2017 (BioRxiv)
29 traits in UKB

Zeng .... Yang 2017 (BioRxiv)
24 traits with significant $\hat{S}$
Summarize over 29 traits

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>-0.350</td>
<td>-0.367</td>
</tr>
<tr>
<td>$h^2_{\text{SNP}}$</td>
<td>0.223</td>
<td>0.222</td>
</tr>
<tr>
<td>$\pi$</td>
<td>5.8%</td>
<td>5.1%</td>
</tr>
</tbody>
</table>
Multiple methods to estimate additive genetic variance

Individual-level data
- GREML
- Haseman-Elston regression
\[(y_j y_k) = \text{mean} + bG_{jk} + e\]

Summary data
- LDscore regression

Consideration:
- data availability
- model assumptions
- computation
Key concepts

- Dense SNP panels allow the estimation of the expected genetic covariance between distant relatives
- A model based upon estimated relationships from SNPs is equivalent to a model fitting all SNPs simultaneously
- The total genetic variance due to LD between common SNPs and (unknown) causal variants can be estimated
- Genetic variance captured by common SNPs can be partitioned across the genome
- Different methods to estimate relatedness from SNPs assume different genetic trait architectures