IGES 118
ASHG 3513W

Extending SNP-based heritability analysis: how many variants show strong effect in a GWAS

Fumihiko Takeuchi , Norihiro Kato
National Center for Global Health and Medicine, Japan

2018.10.15 (IGES), 17 (ASHG)
@San Diego

Poster available http://www.fumihiko.takeuchi.name
Code available https://github.com/fumi-github/Popcorn-t
Hopefully appears in *Nature Communications*. Takeuchi et al. “Interethnic analyses of blood pressure loci in populations of East Asian and European descent”

Summary of Part 1.

We modeled LD-score regression in two ancestries, with per-SNP heritability depending on functional annotations.

Previous studies on LD-score regression

	Per-SNP heritability is identically distributed for all SNPs	Per-SNP heritability depends on allele-frequency and functional annotations
A population of one ancestry	Bulik-Sullivan et al. (2015) *Nat Genet* 47:291	Finucane et al. (2015) *Nat Genet* 47:1228 Speed et al. (2017) *Nat Genet* 49:986 Gazal et al. (2017) *Nat Genet* 49:1421
Two populations of different ancestries; "trans-ancestry genetic correlation"	Brown et al. (2016) *AJHG* 99:76	This study

Genotype model

- N_1 individuals in study 1
- N_2 individuals in study 2
- M SNPs
- Genotype matrix
  - X_1 of study 1 ($N_1 \times M$ dim.)
  - X_2 of study 2 ($N_2 \times M$ dim.)
  - Coded by the standardized allele dose
- LD matrix
  - Correlation between SNPs
  - Σ_1 for population 1 ($M \times M$ dim.)
  - Σ_2 for population 2 ($M \times M$ dim.)
- Studies 1, 2 are derived from populations 1, 2, respectively

Genotype-phenotype model

- Phenotype vector
  - SNPs and non-genetic factors contribute additively to a quantitative phenotype
  - $\mathbf{Y}_1 = X_1\boldsymbol{\beta}_{1,.} + \boldsymbol{\varepsilon}_1$ in study 1 (N_1 dim.)
  - $\mathbf{Y}_2 = X_2\boldsymbol{\beta}_{2,.} + \boldsymbol{\varepsilon}_2$ in study 2 (N_2 dim.)
- Allele substitution effect of SNP j in populations 1 and 2 are $\beta_{1,j}$ and $\beta_{2,j}$
- Phenotype variance not owing to SNPs
  - $\boldsymbol{\varepsilon}_1 \sim \mathcal{N}\left(\mathbf{0},\ (1-h_1^2)\, I_{N_1}\right)$
  - $\boldsymbol{\varepsilon}_2 \sim \mathcal{N}\left(\mathbf{0},\ (1-h_2^2)\, I_{N_2}\right)$
  - h_1^2 and h_2^2 are the heritability attributable to the M SNPs in populations 1 and 2

Allele substitution effect of SNPs

- Allele substitution effect $\beta_{1,j}$ and $\beta_{2,j}$ of SNP j in populations 1 and 2

$$\begin{pmatrix} \beta_{1,j}/w_{1,j} \\ \beta_{2,j}/w_{2,j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \frac{1}{M} \begin{pmatrix} h_1^2 & h_X \\ h_X & h_2^2 \end{pmatrix} \right)$$

- $w_{1,j}$ and $w_{2,j}$ are pre-defined positive weights, coding dependence of per-SNP heritability on
  - Allele frequency [Speed et al.]
  - LD-related functional annotations: predicted allele age, levels of LD, recombination rate, nucleotide diversity, background selection statistic and CpG-content [Gazal et al.]
- h_X is genetic covariance
- Genetic correlation between the populations is $\rho = h_X/\sqrt{h_1^2 h_2^2}$

Z-statistics observed in GWAS

- *Z*-statistics for genotype-phenotype association of SNP *j* in studies 1 and 2

$$\begin{pmatrix}\mathbf{Z}_{1,\cdot}\\ \mathbf{Z}_{2,\cdot}\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{N_1}}X_1{}'\mathbf{Y}_1\\ \frac{1}{\sqrt{N_2}}X_2{}'\mathbf{Y}_2\end{pmatrix}$$

$$=\begin{pmatrix}\frac{1}{\sqrt{N_1}}X_1{}'X_1W_1 & O\\ O & \frac{1}{\sqrt{N_2}}X_2{}'X_2W_2\end{pmatrix}\begin{pmatrix}W_1{}^{-1}\boldsymbol{\beta}_{1,\cdot}\\ W_2{}^{-1}\boldsymbol{\beta}_{2,\cdot}\end{pmatrix}$$

(*1) SNP component

$$+\begin{pmatrix}\frac{1}{\sqrt{N_1}}X_1{}' & O\\ O & \frac{1}{\sqrt{N_2}}X_2{}'\end{pmatrix}\begin{pmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{pmatrix}$$

(*2) residual component

- (*1) follows

$$\mathcal{N}\left(\mathbf{0},\ \frac{1}{M}\begin{pmatrix}h_1^2N_1\left(\Sigma_1W_1{}^2\Sigma_1+\frac{M}{N_1}\Sigma_1\right) & h_{\times}\sqrt{N_1N_2}\Sigma_1W_1W_2\Sigma_2\\ h_{\times}\sqrt{N_1N_2}\Sigma_2W_2W_1\Sigma_1 & h_2^2N_2\left(\Sigma_2W_2{}^2\Sigma_2+\frac{M}{N_2}\Sigma_2\right)\end{pmatrix}\right)$$

- (*2) follows

$$\mathcal{N}\left(\mathbf{0},\ \begin{pmatrix}(1-h_1^2)\Sigma_1 & O\\ O & (1-h_2^2)\,\Sigma_2\end{pmatrix}\right)$$

- Thus,

$$\begin{pmatrix}\mathbf{Z}_{1,\cdot}\\ \mathbf{Z}_{2,\cdot}\end{pmatrix}\sim$$

$$\mathcal{N}\left(\mathbf{0},\ \begin{pmatrix}\frac{h_1^2}{M}N_1\Sigma_1W_1{}^2\Sigma_1+\Sigma_1 & \frac{h_{\times}}{M}\sqrt{N_1N_2}\Sigma_1W_1W_2\Sigma_2\\ \frac{h_{\times}}{M}\sqrt{N_1N_2}\Sigma_2W_2W_1\Sigma_1 & \frac{h_2^2}{M}N_2\Sigma_2W_2{}^2\Sigma_2+\Sigma_2\end{pmatrix}\right)$$

Estimating heritability parameters

- Variance of Z-statistics is parametrized by
  - Heritability h_1^2, h_2^2, $h_\times$
  - LD matrices Σ_1, Σ_2
  - Weighted LD score matrices $\Sigma_1 {W_1}^2 \Sigma_1$, $\Sigma_2 {W_2}^2 \Sigma_2$, $\Sigma_1 W_1 W_2 \Sigma_2$
- Weighted LD scores for SNP j in population 1, 2 or trans-population are
  - $\ell_{1,j} = \left(\Sigma_1 {W_1}^2 \Sigma_1\right)_{j,j}$
  - $\ell_{2,j} = \left(\Sigma_2 {W_2}^2 \Sigma_2\right)_{j,j}$
  - $\ell_{\times,j} = (\Sigma_1 W_1 W_2 \Sigma_2)_{j,j} = (\Sigma_2 W_2 W_1 \Sigma_1)_{j,j}$
- LD scores were computed from 1000 Genomes Project data using the Popcorn program [Brown et al.] https://github.com/brielin/popcorn
- Z-statistics are observed in GWAS
- Heritability parameters h_1^2, h_2^2 and $h_\times$ can be fitted by performing maximum-likelihood estimation

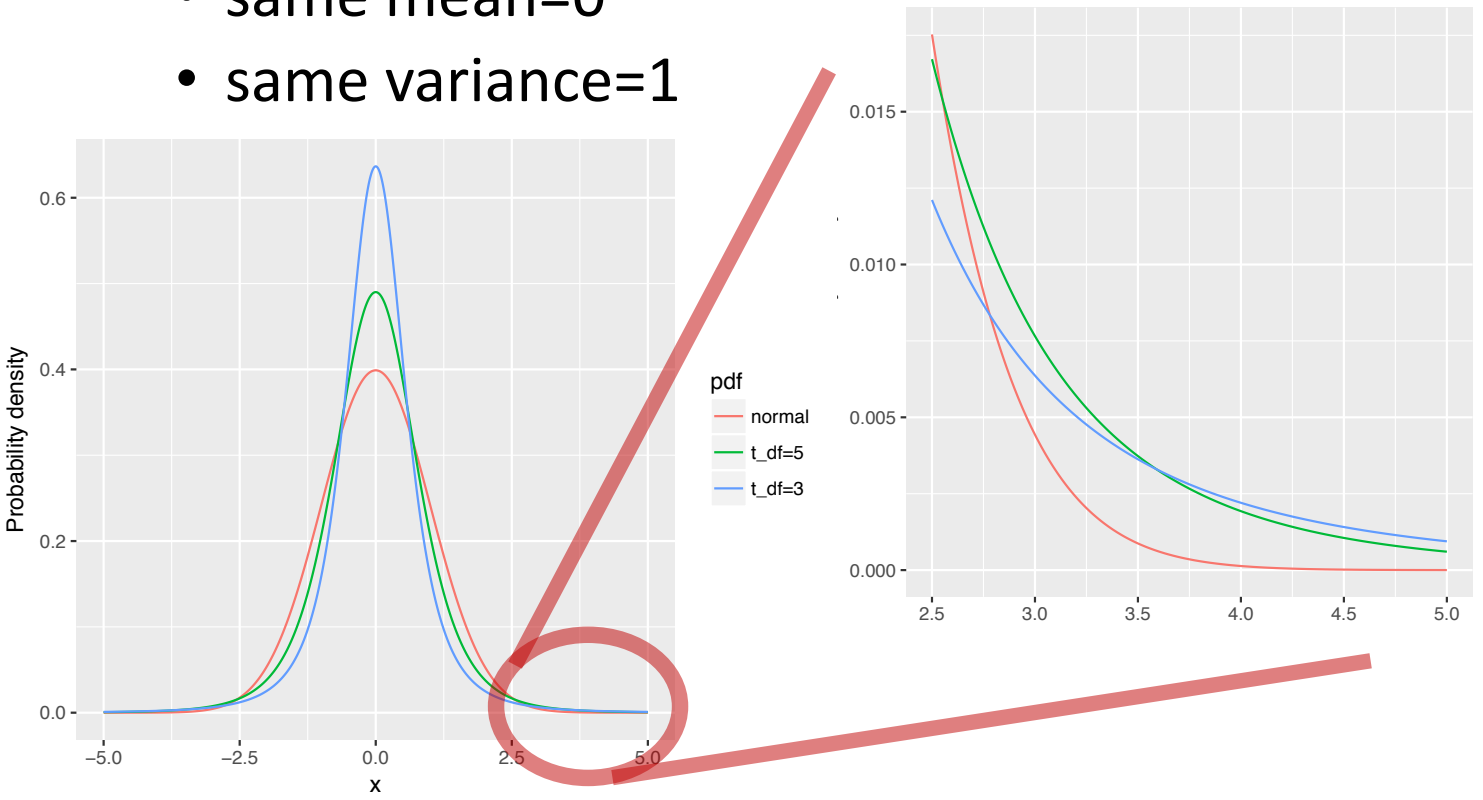
Summary of Part 2.

We modeled allele substitution effects of causal variants by the *t*-distribution (instead of normal).

This enabled power calculation of GWAS.

Tail-heaviness of probability distributions

- Probability distributions with
  - same mean=0
  - same variance=1



heavier tail

- *t*-distribution (df=3)
- *t*-distribution (df=5)
- normal distribution *equals* *t*-distribution (df=∞)

df, degrees of freedom

Previous models for allele substitution effect of causal variants

Realistically model the heavy tail

	Normal distribution	Mixture of null and normal	Mixture of null and 2 normals	*t*-distribution	Mixture of null and *t*-distribution
Genomic selection methods for breeding	Unrealistic cf. Meuwissen et al. (2013) *Annu Rev Anim Biosci* 1:221	BayesC	BayesR	BayesA	BayesB
LD-score regression	Bulik-Sullivan et al. (2015) *Nat Genet* 47:291 Most studies	Zhang et al. (2018) *Nat Genet* 50:1318		This study	

Which purpose requires modeling tail heaviness?

- Knowing variance is enough
  - Tail heaviness is irrelevant
  - Estimate heritability == variance in conventional LD-score regression
- Tail heaviness becomes useful
  - Quantify SNPs with strong effects
    - Many for lipid, few for BMI and height
  - Calculate power of GWAS
    - How many loci would attain genome-wide significance under given sample size?
  - Maybe better LD-score regression fitting

GWAS *Z*-statistics is not normally distributed

- Squared *Z*-statistics
  - Should be χ^2 dist. if *Z*-statistics were normal
  - Actually, the tail is heavier
  - Not due to LD

Trait	Pareto α
SBP (EAS)	2.7
HDL (EUR)	2.1
BMI (EUR)	3.5
Height (EUR)	2.4

Tail heaviness was measured using Hill estimator

- Normality assumption for (*1) was incorrect

χ^2 distribution

Parameters	$k \in \mathbb{N}_{>0}$ (known as "degrees of freedom")
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$

t-distribution

Parameters	$\nu > 0$ degrees of freedom
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$

$x^{-(\nu+1)}$

Pareto distribution

Parameters	$x_m > 0$ scale $\alpha > 0$ shape (
Support	$x \in [x_m, \infty)$
PDF	$\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$
CDF	$1 - \left(\frac{x_m}{x}\right)^\alpha$

$x^{-(\alpha+1)}$

Z-statistics observed in GWAS; under *t*-distribution model

- *Z*-statistics for genotype-phenotype association of SNP *j* in studies 1 and 2

$$\begin{pmatrix}\mathbf{Z}_{1,\cdot}\\ \mathbf{Z}_{2,\cdot}\end{pmatrix}$$=(*1) SNP component

+(*2) residual component

- (*1) follows

$$t\left(\mathbf{0},\ \frac{1}{M}\begin{pmatrix} h_1^2 N_1\left(\Sigma_1 {W_1}^2 \Sigma_1 + \frac{M}{N_1}\Sigma_1\right) & h_\times\sqrt{N_1 N_2}\Sigma_1 W_1 W_2 \Sigma_2 \\ h_\times\sqrt{N_1 N_2}\Sigma_2 W_2 W_1 \Sigma_1 & h_2^2 N_2\left(\Sigma_2 {W_2}^2 \Sigma_2 + \frac{M}{N_2}\Sigma_2\right)\end{pmatrix},\ \nu\right)$$

  - *t*-distribution instead of normal
  - ν degrees of freedom

- (*2) follows

$$\mathcal{N}\left(\mathbf{0},\ \begin{pmatrix} c_1(1-h_1^2)\Sigma_1 & O \\ O & c_2(1-h_2^2)\,\Sigma_2\end{pmatrix}\right)$$

  - Additional parameters c_1, c_2 to account for population stratification

- Maximum-likelihood estimation as in [Brown et al.]
  - For efficient computation, set zero the off-diagonal elements of variance matrices

1. Fit ${h_1}^2$, ν_1, c_1 in study 1
2. Fit ${h_2}^2$, ν_2, c_2 in study 2
3. Fit h_x, ν in whole

Estimated heritability parameters

Trait	Heritability (SE)		Genetic correlation (SE)	df of *t*-distribution
	EAS	EUR		
SBP	0.11 (0.01)	0.09 (0.01)	0.90 (0.04)	4.0
DBP	0.09 (0.01)	0.09 (0.01)	0.85 (0.05)	3.9
HDL	0.09 (0.01)	0.15 (0.02)	0.99 (0.00)	3.1
LDL	0.07 (0.01)	0.12 (0.01)	0.76 (0.09)	3.1
TC	0.09 (0.01)	0.15 (0.02)	0.77 (0.10)	2.9
TG	0.07 (0.01)	0.13 (0.01)	0.99 (0.00)	3.4
T2D	0.10 (0.01)	0.08 (0.00)	0.99 (0.00)	4.1
BMI	0.16 (0.01)	0.11 (0.00)	0.94 (0.02)	5.1
HEIGHT	0.22 (0.02)	0.31 (0.01)	0.94 (0.03)	3.7

EAS, East Asian; EUR, European-descent
df, degrees of freedom

- Heritability ranged 0.07—0.31
- Genetic correlation ranged 0.77—0.99
- Estimated degrees of freedom were ~3 for lipids, ~5 for BMI, and ~4 for other traits, indeed indicating heavy tail for lipids

Estimating the power of GWAS

- Using the estimated heritability parameters,
- We obtain the probability distribution of the standardized effect-size of SNPs

$$\begin{pmatrix} \frac{1}{\sqrt{N_1}}\mathbf{Z}_{1,\cdot} \\ \frac{1}{\sqrt{N_2}}\mathbf{Z}_{2,\cdot} \end{pmatrix} \sim t\left(\mathbf{0},\ \frac{1}{M}\begin{pmatrix} h_1^2\left(\Sigma_1 {W_1}^2 \Sigma_1 + \frac{M}{N_1}\Sigma_1\right) & h_\times \Sigma_1 W_1 W_2 \Sigma_2 \\ h_\times \Sigma_2 W_2 W_1 \Sigma_1 & h_2^2\left(\Sigma_2 {W_2}^2 \Sigma_2 + \frac{M}{N_2}\Sigma_2\right) \end{pmatrix},\ \nu\right)$$

$$+\mathcal{N}\left(\mathbf{0},\ \begin{pmatrix} (1-h_1^2)\frac{1}{N_1}\Sigma_1 & O \\ O & (1-h_2^2)\frac{1}{N_2}\Sigma_2 \end{pmatrix}\right)$$

$$\xrightarrow[N_1,N_2\to\infty]{} t\left(\mathbf{0},\ \frac{1}{M}\begin{pmatrix} h_1^2\Sigma_1 {W_1}^2 \Sigma_1 & h_\times \Sigma_1 W_1 W_2 \Sigma_2 \\ h_\times \Sigma_2 W_2 W_1 \Sigma_1 & h_2^2 \Sigma_2 {W_2}^2 \Sigma_2 \end{pmatrix},\ \nu\right)$$

- Perform numerical sampling from above distribution
- Estimate the power of a GWAS of a given sample size (ie, how many SNPs/loci could attain genome-wide significance)

- Power of SBP (systolic blood pressure) GWAS

			Number of loci detectable in a single EAS GWAS, N_{EAS}=		
			100K	200K	500K
			25	73	276
Number of loci detectable in a single EUR GWAS, N_{EUR}=	100K	18	8	12	16
	200K	53	14	26	42
	500K	209	21	50	118

EAS, East Asian; EUR, European-descent

Data sources

Trait	Ancestry	No. of SNPs	Sample size	Study name	Reference
SBP, DBP	EAS	6,233,864	130,777	This study	
SBP, DBP	EAS	2,485,253	27,868	iGEN-BP	Nat Genet 47:1282
SBP, DBP	EUR	2,149,719	35,344	iGEN-BP	Nat Genet 47:1282
SBP, DBP	EUR	2,398,700	69,909	ICBP	Nature 478:103
LDL, HDL, TG, TC	EAS	2,227,836	34,374	AGEN	Hum Mol Genet 26:1770
LDL, HDL, TG, TC	EUR	2,447,441	187,365	GLGC	Nat Genet 45:1274
T2D	EAS	479,088	25,066	BBJ	Nat Comm 7:10531
T2D	EUR	8,075,531	158,186	DIAGRAM	Diabetes 66:2888
BMI	EAS	5,961,600	158,284	BBJ	Nat Genet 49:1458
BMI	EUR	2,554,637	322,154	GIANT	Nature 518:197
Height	EAS	2,730,895	36,227	AGEN	Hum Mol Genet 24:1791
Height	EUR	2,550,858	253,280	GIANT	Nat Genet 11:1173

Thank you for making the data available!

What is LD-score regression? (1/2)

- Linkage disequilibrium (LD)
  - Correlation of alleles between a pair of SNPs
- LD score
  - Number of SNPs in LD (including itself)
  - 3.7 for SNP3, 1.2 for SNP5
- The effect-size of a SNP observed in GWAS = the sum of the allele substitution effects taken over nearby SNPs in LD
  - SNP3_A allele is observed as “risk”, but actually indirectly manifests the effects of SNP2_A and SNP4_G
- The idea of LD-score regression
  - Many SNPs in the genome have weak causal effect
  - Higher LD-score SNPs tend to show stronger association in GWAS
    - SNP3 > SNP5

Observed

		SNP1	SNP2	SNP3	SNP4	SNP5
Patient 1	Paternal chr.	A	A	A	G	C
	Maternal chr.	T	C	T	T	G
Patient 2	Paternal chr.	A	A	A	G	C
	Maternal chr.	T	C	T	T	G
Patient 3	Paternal chr.	T	C	T	T	C
	Maternal chr.	A	A	A	G	C
Control 1	Paternal chr.	A	A	A	T	G
	Maternal chr.	T	C	T	T	C
Control 2	Paternal chr.	T	C	T	T	G
	Maternal chr.	T	C	T	T	C
Control 3	Paternal chr.	T	C	T	T	C
	Maternal chr.	T	C	T	T	C

Disease risk allele
Disease protective allele
Non-causal allele
Minor allele
Major allele

What is LD-score regression? (2/2)

- A Higher LD-score SNP tends to show stronger association (larger Z-score) in GWAS
- From the gradient of this correlation, we can compute heritability and genetic correlation!
  - Heritability is the proportion of trait variance that can be explained by SNPs
  - Genetic correlation is the correlation of allele substitution effects of a SNP in two ancestries
- Compute the gradient using linear regression. Regression formulae for each SNP are
  - (Z-score in EAS GWAS)^2
    $=$ (Heritability in EAS, h_{EAS}^2) / (# SNPs) $\times$ (# samples) $\times$ (LD-score in EAS) $+ 1$
  - (Z-score in EUR GWAS)^2
    $=$ (Heritability in EUR, h_{EUR}^2) / (# SNPs) $\times$ (# samples) $\times$ (LD-score in EUR) $+ 1$
  - (Z-score in EAS GWAS) $\times$ (Z-score in EUR GWAS)
    $=$ (Covariance of heritabilities, h_X) / (# SNPs) $\times$ (# samples) $\times$ (Concordance of LD-score in two ancestries)
- (Genetic correlation EAS vs EUR) $= h_X/\mathrm{sqrt}(h_{EAS}^2\ h_{EUR}^2)$