HERITABILITY ESTIMATION USING A REGULARIZED REGRESSION APPROACH (HERRA)

Malka Gorfine, Tel Aviv University, Israel

Joint work with Li Hsu, FHCRC, Seattle, USA

The concept of heritability

Heritability summarizes how much of the variation in a trait (e.g. height, diabetes, Alzheimer disease) is due to variation in genetic factors.

Broad-sense heritability: includes also dominance and epistasis effects.

Narrow-sense heritability: additive genetic variation.

The mystery of missing heritability

Heritability estimated by **family data**

is much higher then

overall heritability estimated by **GWAS data** of non-related individuals using all significant SNPs' effects

genome-wide association study – examination of many (e.g. 500K) genetic variants in different individuals to study the association between the variants and the trait

Examples: height, Alzheimer disease, diabetes, among many others.

Example: colorectal cancer (CRC)

- CRC is one of the most common diagnosed cancers in development countries. In USA: 5.2% for men, 4.8% for women.
- Heritability estimate ranging 12-35% from twins and family data.
- 31 SNPs were identified as associated with CRC risk.
- The Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) including over 40,000 participants. 3 platforms: 300K, 550K and 730K. Of the 31 known CRC susceptibility SNPs, 18, 30, 26 of the SNPs or proxies (r2>0.8) were available on the 300K, 550K and 730K, respectively.
- **GECCO data: heritability estimate based on the 31 identified CRC SNPs (or their proxies) = 0.65% (SE=0.18%).**

Possible reasons for the missing heritability

- **Unidentified variants with small effect size.**
- Some mutations causing variation are not in perfect LD with any of the SNPs.
- Rare variants not captured by current genotyping platforms.
- Missing epistatic interaction in the model.
- Missing gene-environment interaction in the model.
- Inflated heritability estimates based on twins studies.

Hill et al. (2008); Manolio et al. (2009); Eichler et al. (2010); Dickson et al. (2010); Gibson (2011); Wray et al. (2011); Visscher et al. (2012); Zuk et al. (2012); Zaitlen et al. (2013);

GCTA – State of the art heritability estimator

Mixed effects models were applied in quantitative genetics by animal breeders decades ago.

Yang et al. (2010) introduced the mixed effects approach for heritability estimation using GWAS data of apparently unrelated individuals.

Their method is applied by the GCTA software.

The key advantage: estimating heritability of a trait without explicitly identifying the causal genetic loci.

Outline

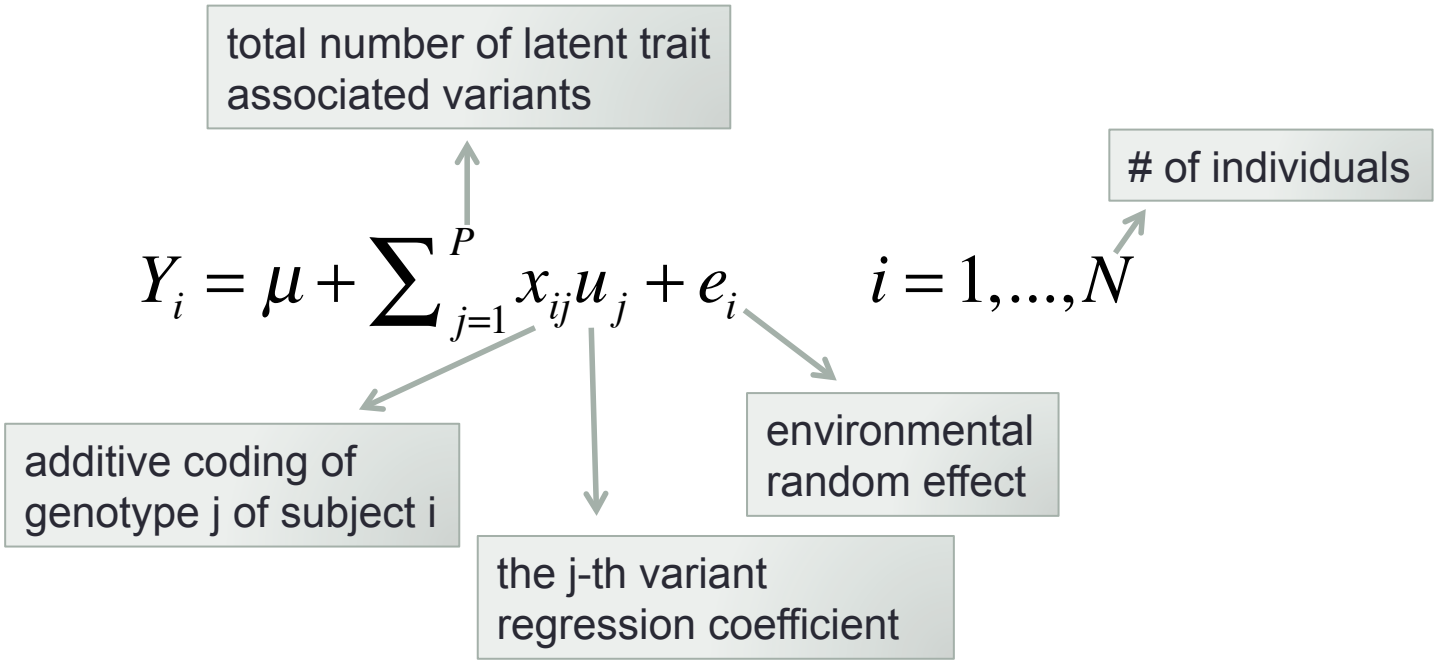
GCTA & LDAK : model, estimation procedure, properties

HERRA: model, estimation procedure, properties

GCTA & LDAK vs HERRA: simulation results

Real data analysis: CRC of GECCO data

GCTA model



$u_j \sim N\left(0,\sigma_u^2\right) \quad j = 1,\ldots,P \qquad\qquad e_i \sim N\left(0,\sigma_e^2\right) \quad i = 1,\ldots,N$

GCTA model – cont.

$$Y_i = \mu + \sum_{j=1}^{P} x_{ij} u_j + e_i \qquad i = 1,\ldots,N$$

standardized 0,1 or 2 for genotype qq, Qq or QQ, respectively

$$u_j \sim N\left(0,\sigma_u^2\right) \quad j = 1,\ldots,P$$

$$e_i \sim N\left(0,\sigma_e^2\right) \quad i = 1,\ldots,N$$

$$g_i = \sum_{j=1}^{P} x_{ij} u_j \quad \Rightarrow \quad E\left(g_i\right) = 0 \quad Var\left(g_i\right) = P\sigma_u^2 = \sigma_g^2$$

total additive genetic effects of subject *i*

total variance of additive genetic effects

GCTA model – cont.

$$Y_i = \mu + \sum_{j=1}^{P} x_{ij} u_j + e_i \qquad i = 1,\ldots,N$$

$$u_j \sim N\left(0,\sigma_u^2\right) \quad j = 1,\ldots,P$$

$$e_i \sim N\left(0,\sigma_e^2\right) \quad i = 1,\ldots,N$$

$$g_i = \sum_{j=1}^{P} x_{ij} u_j \quad Var\left(g_i\right) = P\sigma_u^2 = \sigma_g^2$$

NxN genetic relationship matrix

$$\left(Y_1,\ldots,Y_N\right)' \mid X \sim MVN\left(\mu, G\sigma_g^2 + I\sigma_e^2\right) \qquad G = \frac{1}{P} XX'$$

narrow-sense heritability

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

$$Var\left(Y_1\right) = \sigma_g^2 + \sigma_e^2$$

***G* and *Y* are required for MLE of the narrow-sense heritability parameter, but *G* is unknown.**

GCTA model – cont.

$$Y = (Y_1, \ldots, Y_N)' \,|\, X \sim MVN\left(\mu, G\sigma_g^2 + I\sigma_e^2\right) \qquad G = \frac{1}{P} XX'$$

Instead of the unknown *G*, GCTA uses a matrix that consists of:

1. **Genetic correlation matrix computed on the entire genotyped data (e.g. using 10^6*0.5 SNPs instead of only hundreds causal SNPs).**

2. Correcting the genetic correlation matrix for sampling bias, taking into account the inbreeding coefficient.

3. **Accounting for LD between causal variants and genotyped SNPs by using a simulation-based heuristic procedure with certain assumptions about the number of causal variants and their MAF.**

minor allele frequency

Modification and Extensions of GCTA

Golan and Rosset (2011):

"Replacing the correct genetic correlation matrix by a different matrix estimated from the data as if the latter matrix were the correct matrix is unfounded statistically."

"The very large number of SNPs used for estimating the genetic correlations — most of them likely not causative — masks out the correlations on the set of causal SNPs … this leads to inaccurate and suboptimal estimation of heritability."

Instead, they use the mixed model approach while

- treating the identity of causal SNPs as missing data
- finding the ML estimates based on intensive MCMC method.

Unfortunately, this approach is not tractable for problems with, for example, half a million genotyped SNPs.

Modification and Extensions of GCTA - LDAK

Speed et al. (2011):

Uneven LD between SNPs can generate large bias in the heritability estimator based on the mixed model approach.

Causal variants tend to be overestimated in regions of strong LD and underestimated in regions of low LD.

In practice, if some of the causal variants being tagged by multiple genotyped SNPs more than others, it distorts their contributions to the heritability estimator.

Instead, they use the mixed model approach, while replacing the observed correlation matrix by a weighted matrix consists of scaling SNP genotypes according to local patterns of LD.

The weights are identified by linear programming procedure.

HERRA – the proposed approach

Heritability
Estimation using
Regularized
Regression
Approach.
Not a mixed model approach.

total number of latent trait associated variants

$$Y_i = \mu + \sum_{j=1}^{P} x_{ij} u_j + e_i \qquad i = 1,\ldots,N$$

additive coding of variant *j* of subject *i*

the *j*-th variant **fixed** regression coefficient

environmental random effect

~~$u_j \sim N(0,\sigma_u^2) \quad j = 1,\ldots,P$~~

$e_i \sim N(0,\sigma_e^2) \quad i = 1,\ldots,N$

HERRA

$$Y_i = \mu + \sum_{j=1}^{P} x_{ij} u_j + e_i \qquad i = 1,\ldots,N$$

$$E\left(x_{ij}\right) = 0, \quad Var\left(x_{ij}\right) = 1, \quad e_i \sim N\left(0, \sigma_e^2\right) \quad i = 1,\ldots,N$$

Assuming the variants are independent for all *i* and *j* yields

$$\sigma_g^2 = \sum_{j=1}^{P} u_j^2$$

so

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{\sigma_g^2}{\sigma_Y^2} = 1 - \frac{\sigma_e^2}{\sigma_Y^2}$$

The main idea: estimate $\left(\sigma_Y^2, \sigma_e^2\right)$ **instead of** $\left(\sigma_g^2, \sigma_e^2\right)$.

These two options are not the same, since the identity of the causal SNPs is unknown and a working model is required

HERRA

$$Y_i = \mu + \sum_{j=1}^{P} x_{ij} u_j + e_i \qquad i = 1,\ldots,N$$

$$e_i \sim N\left(0, \sigma_e^2\right) \quad i = 1,\ldots,N$$

$$h^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2}$$

Working model:

$$Y_i = \mu + \sum_{j=1}^{M} z_{ij} u_j + e_i \qquad i = 1,\ldots,N$$

genotyped SNPs

By using high or ultrahigh dim. Approach, we can consistently estimate σ_e^2 ,

and estimate σ_Y^2 by $(N-1)^{-1} \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$

HERRA

$$Y_i = \mu + \sum_{j=1}^{M} z_{ij} u_j + e_i \qquad e_i \sim N\left(0, \sigma_e^2\right)$$

Estimator of σ_e^2 (in the spirit of Fan et al. 2012):

I. Apply a joint-type screening method (e.g. ITRRS, Fan & Lv, 2008) or a marginal-type sure independent screening (SIS, Fan & Lv, 2008), and reduce the ultra-high dim to a relatively large scale. This step is to filter out SNPs that are unlikely to be associated with the trait.

II. Use only the selected SNPs of Step I. Randomly split the sample into two equal subgroups. Apply a high-dim variable selection method (e.g. LASSO) to the 1st subset, and apply OLS to the 2nd subset using only those selected SNPs. Get an unbiased estimator of σ_e^2 .

III. Repeat Step II while switching the role of the 1st and 2nd datastes.

IV. The final estimator of σ_e^2 is defined as the mean of the above two estimators: $\hat{\sigma}_e^2$.

Which screening method should be used?

Sure Independence Screening (SIS, Fan and Lv, 2008):

- Independence screening is used as a fast but crude method of reducing the dimensionality to a more moderate size (usually below the sample size).
- Then, variable selection can be accomplished by some refined lower dimensional method (e.g. lasso).
- Independence screening recruits those features having the best **marginal** utility, which corresponds to the largest marginal correlation with the response in the context of least-squares regression.
- This fast feature selection method has a sure screening property - with probability tending to 1, the independence screening technique retains all of the important features in the model.

In our setting (with standardized predictors): $\omega = X^T Y$ (design matrix)

and keep those *c* variables with the largest $|\omega_j|$.

For unltrahigh-dimension data, we recommend on a joint-type screener such as the Iteratively Ridge Regression Screener (ITRRS, Fan and Lv, 2008), so LD between SNPs is considered.

HERRA

$$Y_i = \mu + \sum_{j=1}^{M} z_{ij} u_j + e_i \qquad e_i \sim N\left(0, \sigma_e^2\right)$$

$$\hat{h}^2 = 1 - \frac{\hat{\sigma}_e^2}{\hat{\sigma}_Y^2}$$

Theorem – the oracle property:

Under some regularity conditions and $E\left(e^4\right) < \infty$. If a procedure satisfies the sure screening property (with probability tending to 1, the independence screening technique retains all of the important features in the model), then,

$$\sqrt{N}\left(\hat{\sigma}_e^2 - \sigma^2\right) \xrightarrow{D} N\left(0, E\left(e^4\right) - \sigma^4\right)$$

HERRA

$$Y_i = \mu + \sum_{j=1}^{M} z_{ij} u_j + e_i \qquad e_i \sim N\left(0, \sigma_e^2\right)$$

$$\hat{h}^2 = 1 - \frac{\hat{\sigma}_e^2}{\hat{\sigma}_Y^2}$$

By the delta method we get

$$\sqrt{N}\left(\hat{h}^2 - h^2\right) \xrightarrow{D} N\left(0, 4h^2\left(1-h^2\right)^2\right)$$

But we cannot use it for inference in practice!

Alternatively, use weighted bootstrap.

HERRA – accounting for known risk factors

The known risk factors (e.g. smoking, dietary variables) are included in the model to account for the confounding effect and reduce the error variance.

The risk factors W are included in the model, and are not subject to variable selection in any step.

The heritability estimator:

$$1 - \frac{\hat{\sigma}_e^2 + \hat{\beta}' \operatorname{var}(W) \hat{\beta}}{\hat{\sigma}_Y^2}$$

regression coefficient estimators for W

known risk factors

What about categorical traits?

Assume a disease status outcome. The observed scale is 0/1.

Variances and heritability calculated on the observed scale are function of the prevalence of the trait in the population.

Liability model (Wright, 1934 ;Falconer, 1965):

- There is an underlying gradation of some attribute immediately related to the causation of the disease.
- If we could measure this attribute, it would give a graded scale of the degree of affectedness or of normality.
- All individuals above a certain value exhibited the disease and all below it did not.
- This hypothetical graded attribute will be referred as the individual liability to the disease.
- Liability is normally distributed.

GCTA with all-or-none traits

$$D_i = \alpha + \sum_{j=1}^{P} x_{ij} v_j + \varepsilon_i \qquad i = 1,\ldots,N$$

binary outcome

$$v_j \sim N\left(0,\sigma_v^2\right) \quad j = 1,\ldots,P \qquad\qquad \varepsilon_i \sim N\left(0,\sigma_\varepsilon^2\right) \quad i = 1,\ldots,N$$

observed scale heritability

$$h_o^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2} = \frac{P\sigma_v^2}{P\sigma_v^2 + \sigma_\varepsilon^2}$$

Apply the Robertson transformation (Dempster and Lerner, 1950) and get

liability scale heritability

$$h_l^2 = h_o^2 K\left(1-K\right)/z^2$$

normal density at the threshold

trait prevalence in the population

GCTA with all-or-none trait – cont.

$$D_i = \alpha + \sum_{j=1}^{P} x_{ij} v_j + \varepsilon_i \qquad i = 1,\ldots,N$$

$$h_o^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2} = \frac{P\sigma_v^2}{P\sigma_v^2 + \sigma_\varepsilon^2}$$

GCTA uses mixed model effects **with the aforementioned calibration steps** for estimating $\left(\sigma_g^2, \sigma_\varepsilon^2\right)$, and then the Robertson transformation is recommended.

$$h_l^2 = h_o^2 K\left(1 - K\right) / z^2$$

HERRA with all-or-none traits

$$D_i = \alpha + \sum_{j=1}^{P} x_{ij} v_j + \varepsilon_i \qquad i = 1,\ldots,N$$

~~$v_j \sim N\left(0,\sigma_v^2\right) \quad j = 1,\ldots,P$~~

$$\varepsilon_i \sim N\left(0,\sigma_\varepsilon^2\right) \quad i = 1,\ldots,N$$

$$h_o^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2} = \frac{\sum_{j=1}^{P} v_j^2}{\sum_{j=1}^{P} v_j^2 + \sigma_\varepsilon^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_D^2}$$

HERRA estimates σ_D^2 by $\bar{D}\left(1-\bar{D}\right)$ and σ_ε^2 by $\hat{\sigma}_\varepsilon^2$ and then uses

$$h_l^2 = h_o^2 K\left(1-K\right)/z^2$$

thus, it provides a consistent heritability estimator.

HERRA with all-or-none traits

$$E(Y_i)=0, \quad \mathrm{var}(Y_i)=1, \quad Y_i \,|\, x_i \sim N(\gamma^T x_i, \sigma_e^2) \quad i=1,\ldots N$$

Working model: $D_i = I(Y_i > c)$, $D_i = \alpha + \sum_{j=1}^{P} x_{ij} v_j + \varepsilon_i$ $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

$$\tilde{\sigma}_\varepsilon^2 = \frac{1}{N-p}\sum_{i=1}^{N}\left(D_i - \hat{\alpha} - x^T\hat{v}\right)^2$$

OLS

$$\hat{\alpha} \to \Pr(D=1) = K, \quad \hat{v} \to E(xD)$$

$$\tilde{\sigma}_\varepsilon^2 \to K(1-K) - E(Dx^T)E(xD)$$

By some algebra we get

$$E(Dx^T)E(xD) = (1-\sigma_e^2)\{\phi(c)\}^2 = h_l^2 z^2$$

$$\tilde{h}_o^2 = 1 - \frac{\tilde{\sigma}_\varepsilon^2}{\hat{\sigma}_D^2} \to 1 - \frac{K(1-K) - h_l^2 z^2}{K(1-K)} = \frac{h_l^2 z^2}{K(1-K)} = h_o^2$$

Robertson transformation

Age-at-onset outcome

Consider a parametric accelerated failure time model

$$Y_i^o = \mu + \sum_{j=1}^{P} x_{ij} u_j + e_i$$

$$Y_i^o = \log T_i \qquad e_i \sim N\left(0, \sigma_e^2\right)$$

failure time

$$Y_i = \min\left(Y_i^o, C_i\right)$$

observed time

log-scale censoring time

$$\delta_i = I\left(Y_i^o \le C_i\right)$$

The observed data:

$$\left\{Y_i, X_i^{'}, \delta_i\right\} \qquad i = 1, \ldots, N$$

Age-at-onset outcome

Our goal is estimating

$$h^2 = 1 - \frac{\sigma_e^2}{\sigma_{Y^o}^2}$$

Let

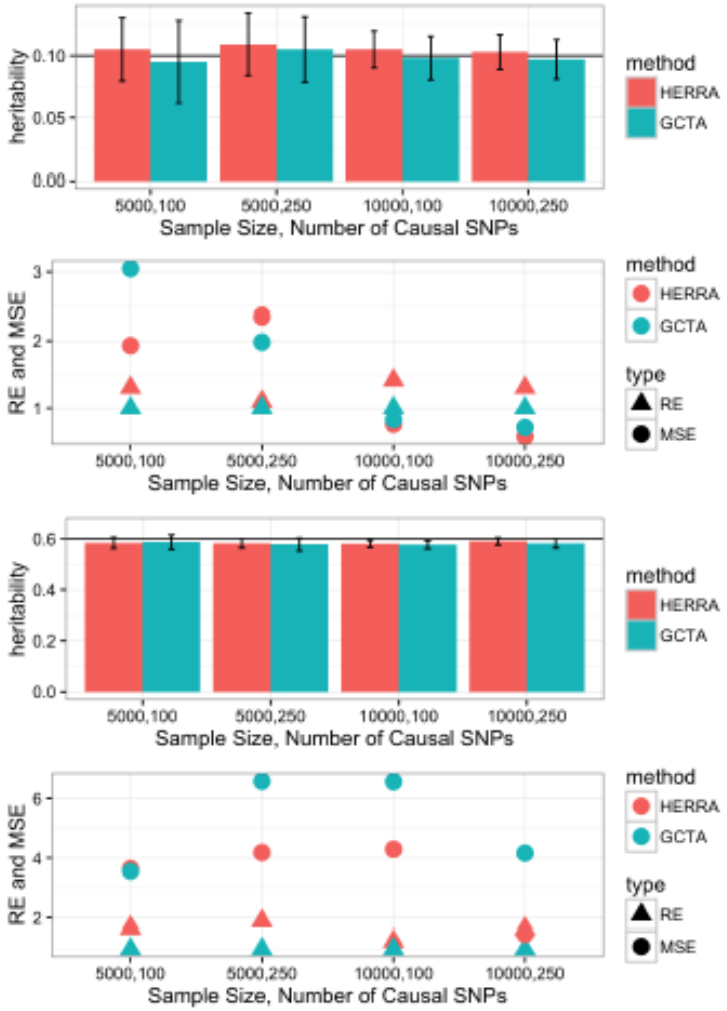
$$W_i = \delta_i / \hat{S}_C(Y_i)$$

KM estimator of the censoring survival distribution

Then, the above variances are estimated by the IPCW approach.

Simulation Results – continuous trait

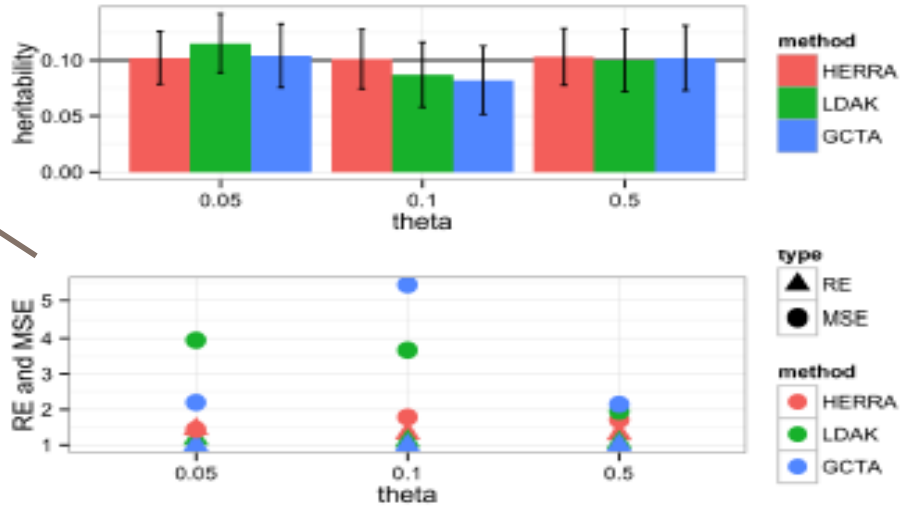
We phased chromosome 22 of a GWAS which included 6006 subjects. Random pairs of haplotypes generated the simulated data.



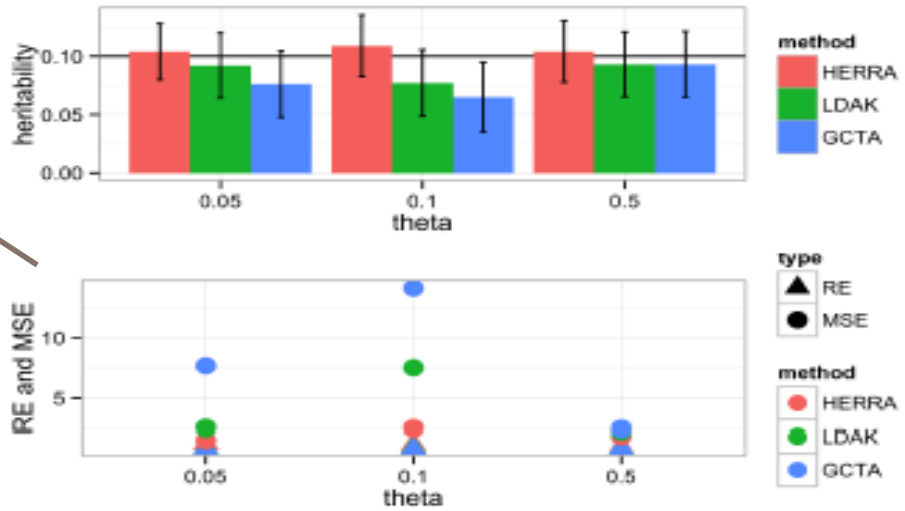
Simulation Results – continuous trait

One chromosome, 5000 observations, 60 causal SNPs, theta = MAF of causal SNPs

causal SNPs are included in the analysis

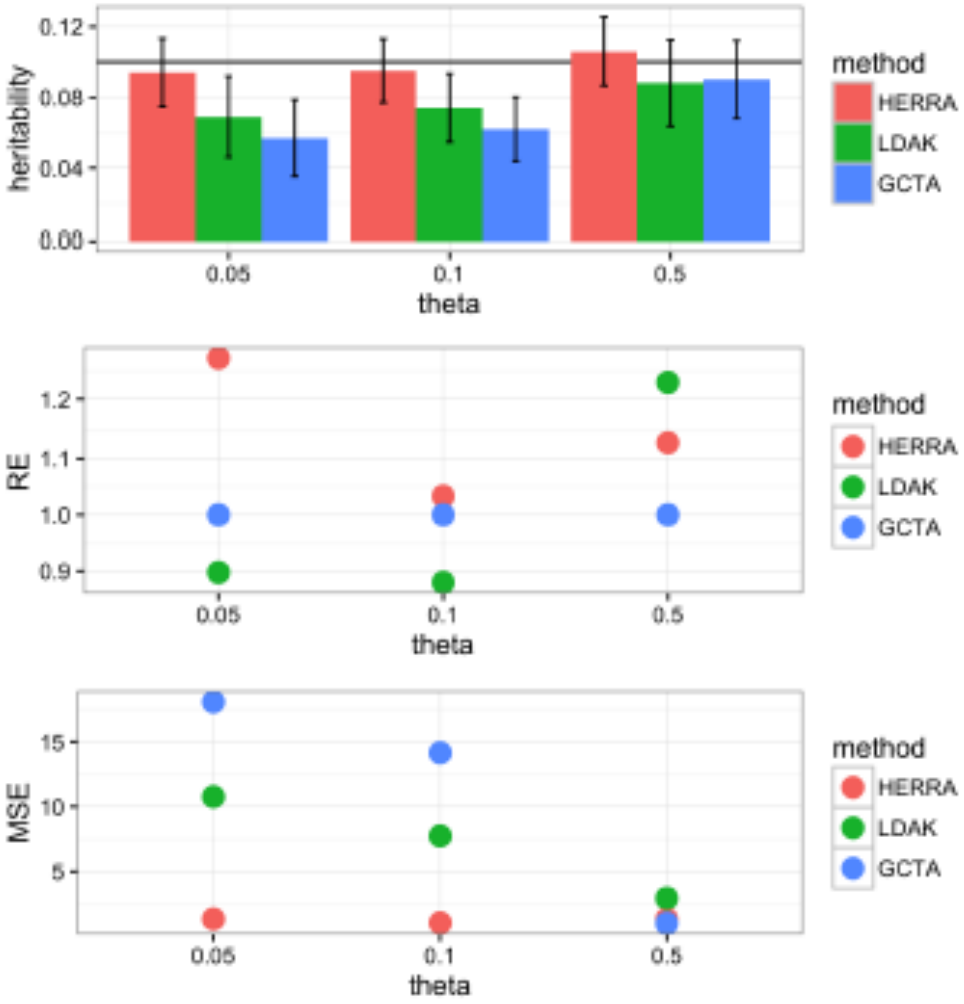


causal SNPs are **NOT** included in the analysis



Simulation Results – continuous trait

5 chromosome, 5000 observations, 250 causal SNPs, theta = MAF of causal SNPs, M=35760, N=10000



GECCO data

A consortium comprised of a coordinating center at the FHCRC and investigators from 16 cohort and case-control studies in North America, Australia and Europe.

For illustration we focus on the largest subset of sample that are genotyped using illumina 300K: 4312 cases, 4356 controls, 248,977 SNPs with MAF > 0.01.

Results are adjusted for age, gender and study center.

We first conducted **ITRRS** (6 iterations) for each chromosome, combined the selected SNPs, split the data and conducted two 10-fold lasso regressions.

SE – by 100 weighted bootstrap samples.

GECCO data – results (liability scale)

HERRA: heritability = 0.110, SE = 0.00519

GCTA: heritability = 0.068, SE = 0.017

LDAK: heritability = 0.072, SE = 0.021

Our estimate is larger than the GCTA or LDAK with smaller SE, and is closer to heritability estimates from twins and family data, which ranges from0.12 to 0.35.

Heritability estimate based on the 31 identified CRC SNPs (or their proxies) = 0.0065 (SE=0.0018).

GECCO data – results (liability scale)

Sensitivity Analysis

	5 Iterations		6 Iterations	
Shrinkage	Observed	Liability	Observed	Liability
0.0060	0.201	0.091	0.207	0.094
0.0080	0.256	0.116	0.236	0.107
0.0100	0.265	0.120	0.244	0.110
0.0102	0.221	0.100	0.218	0.099
0.0104	0.253	0.114	0.257	0.116

Open Questions

- Variance estimation – weighted bootstrap
- How many ITRRS?
- Setting tuning parameters of the ridge regression
- How to apply with 8M SNPs?
- Should this model be adopted for risk prediction?
- Adding GxG or GxE interactions

END