Robust estimation of R^2 for $Trait \sim PRS$ using GWAS summary statistics

Li Ge, lge7@wisc.edu
PhD Student in Biomedical Data Science
University of Wisconsin-Madison

2019-08-20, Rotation Advisor: Qiongshi Lu

Motivation

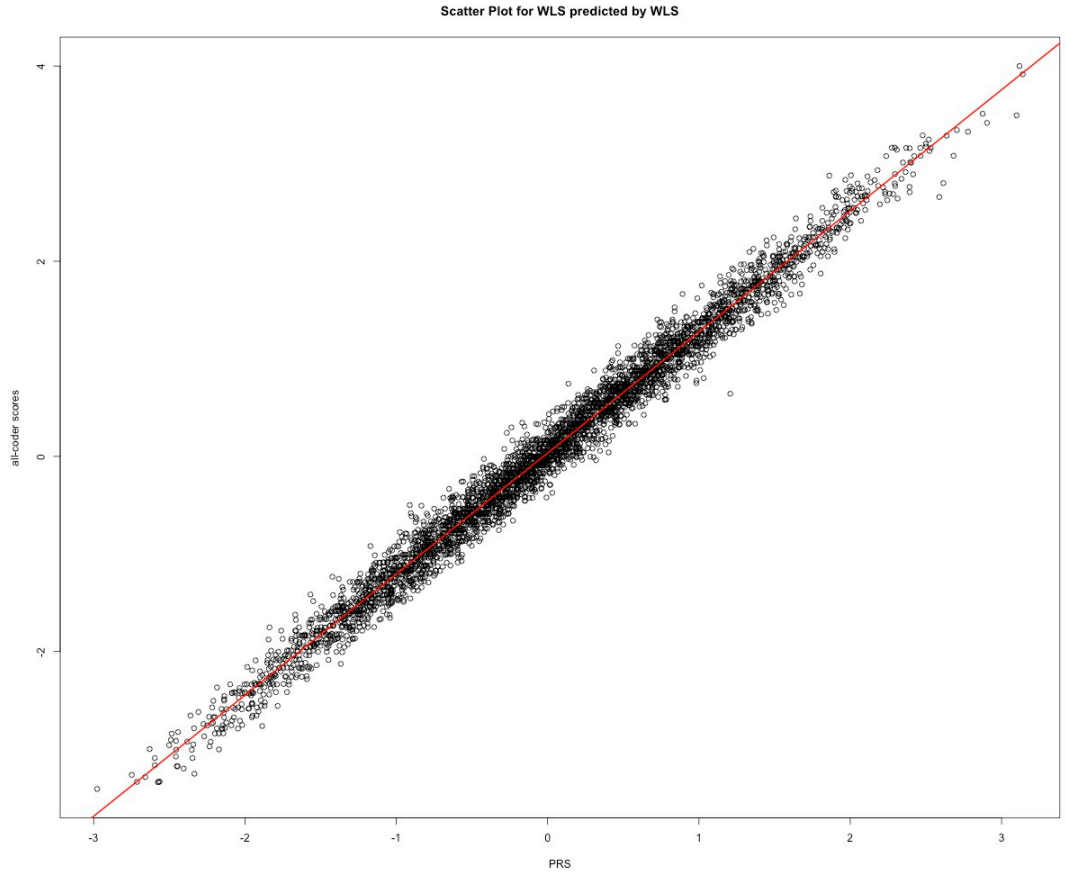
- How predictive is PRS? It is usually quantified by the R^2 of the regression of $Trait \sim PRS$.
- However, this process is often sabotaged by overlapping of training and testing samples (overfitting), resulting in inflated R^2 and effect sizes.
- We want to fix the sample-overlapping problem, i.e., to obtain a robust estimation of R^2.
- Specifically, we want to achieve it by only using GWAS summary statistics.

Polygenic Risk Scores (PRS)

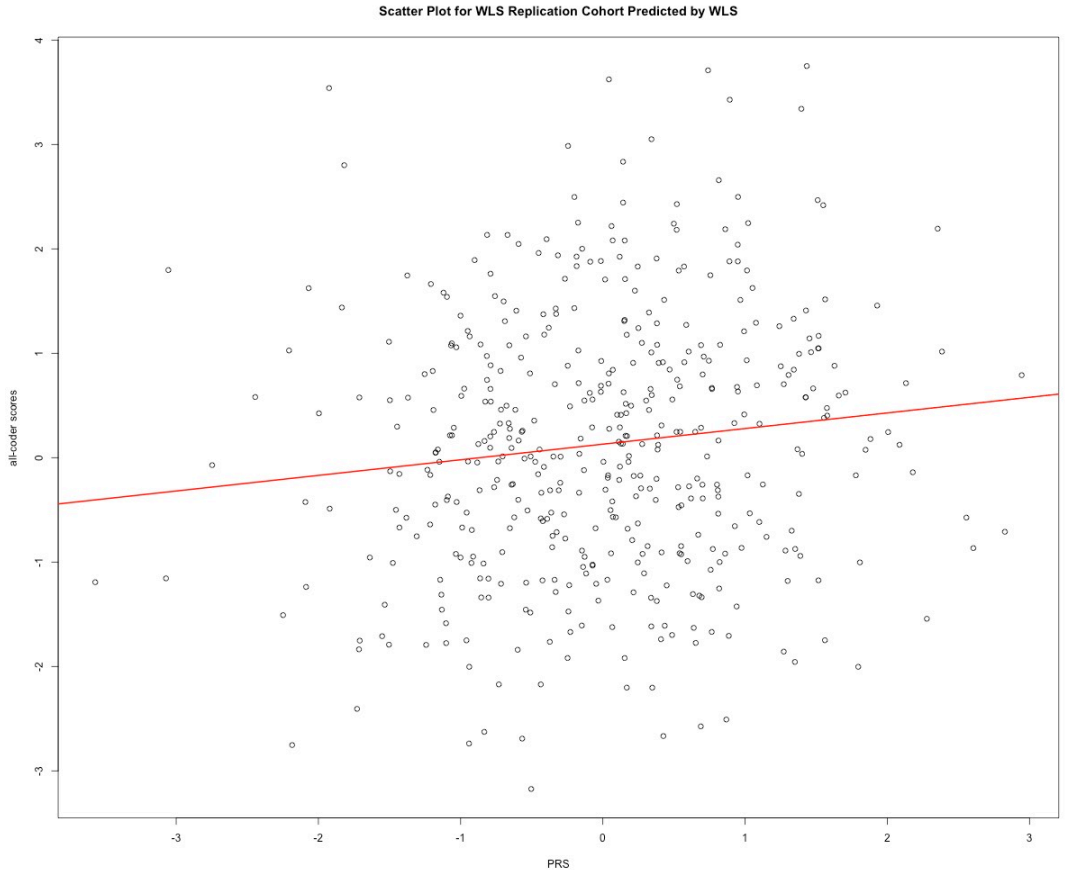
- A polygenic risk score (PRS) is a sum of trait-associated alleles across many genetic loci, typically weighted by effect sizes estimated from a genome-wide association study [1].
- Polygenic Risk Scores (PRS) have recently been used to summarize genetic effects among an ensemble of markers that do not individually achieve significance in a large-scale association study [2].
- There have also been interests in cross trait PRS analysis. For example, “polygenic risk scores for schizophrenia and bipolar disorder predict creativity” [3], etc. And the BADGERS [4].

Example of Overfitting

100% overlapping

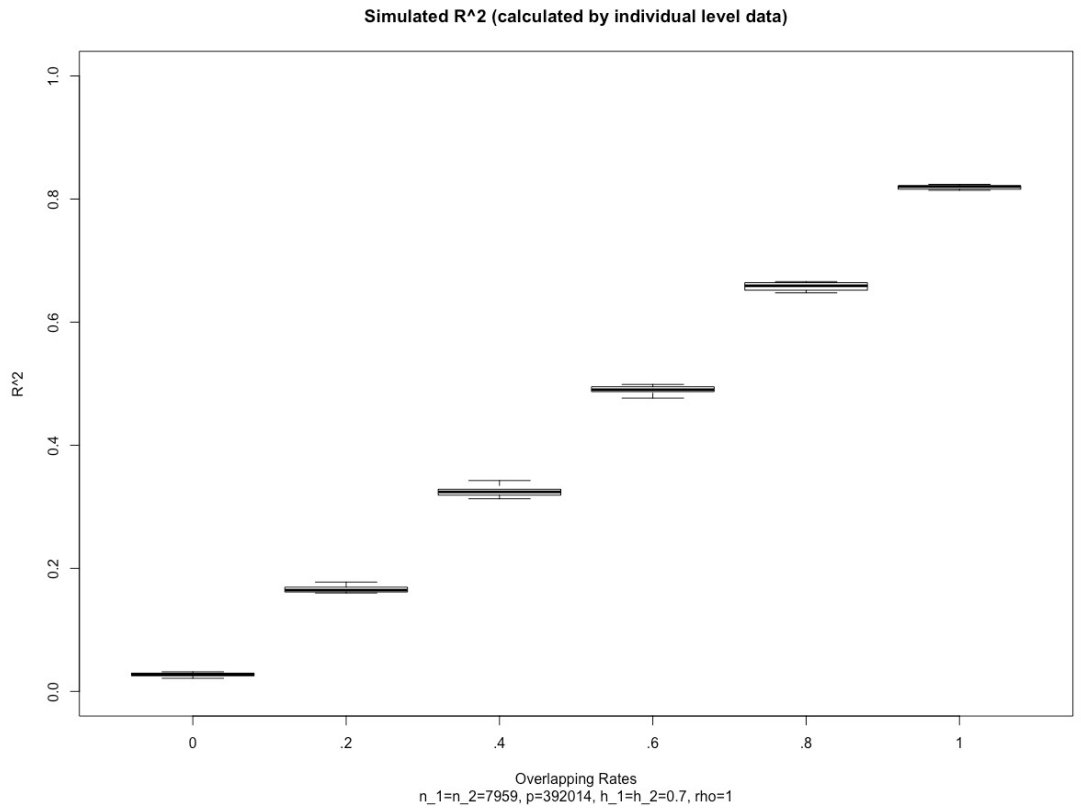


zero overlapping

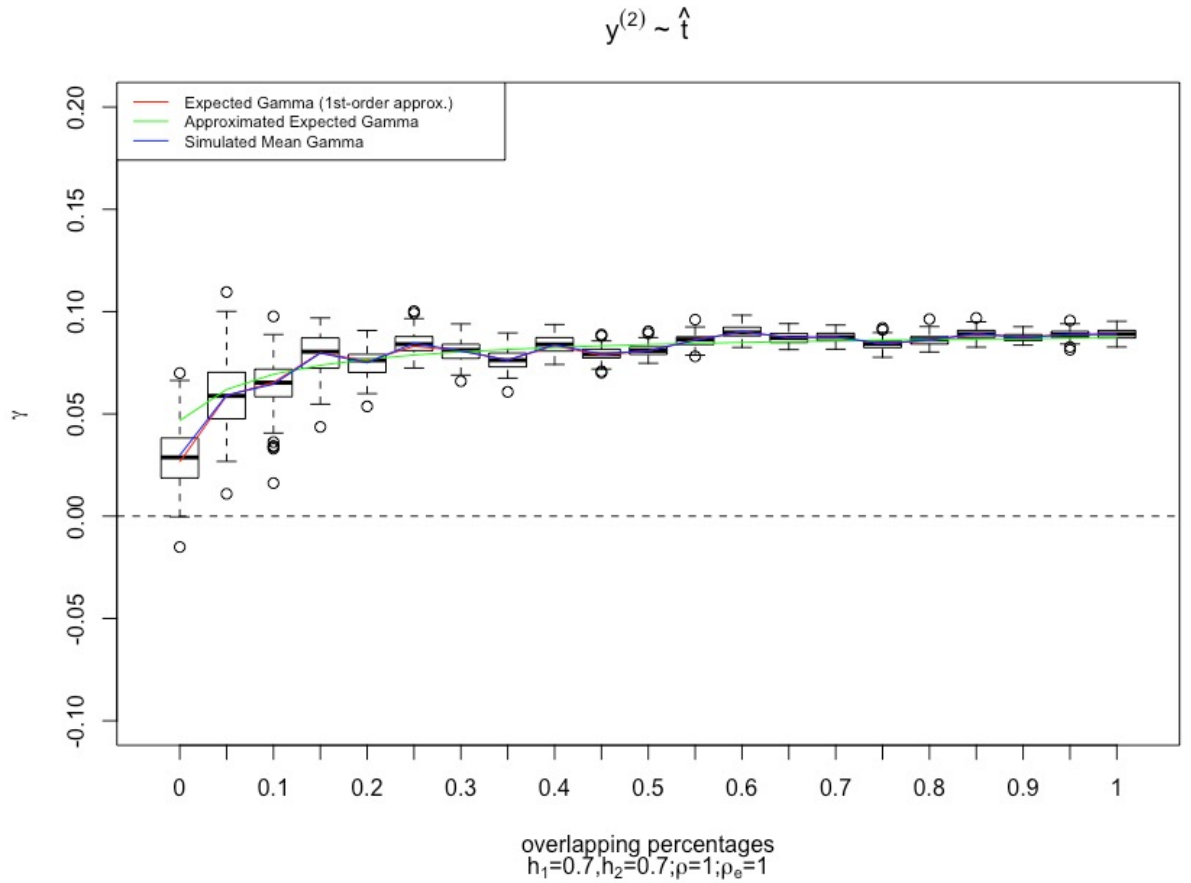


Example of Overfitting

inflated R^2



inflated effect sizes



Methods Overview

- Goal – derive the expected R^2 assuming no sample overlapping
- Model Setup
- Assumptions
- Derivation
- Results

Model Setup

$$y^{(1)} = X^{(1)}w^{(1)} + \epsilon, \qquad \epsilon \sim N(0, (1 - h_1^2)I)$$
$$y^{(2)} = X^{(2)}w^{(2)} + \delta, \qquad \delta \sim N(0, (1 - h_2^2)I)$$

- $y^{(j)} \in R^{n_j \times 1}$- quantitative trait j, has n_j samples.
- $X^{(j)} \in R^{n_j \times p}$- genotypic data (design matrix of trait j), each contains n_j samples, p SNPs, has been standardized.
- $w^{(j)} \in R^{p \times 1}$ - effect sizes of trait j, corresponding to p SNPs.
- $\epsilon \in R^{n_1 \times 1}, \delta \in R^{n_2 \times 1}$ - non-genetic (environmental) factors, random vectors.
- h_j^2 - heritability of trait j, stands for the degree of variation in a phenotypic trait in a population that is due to genetic variation between individuals in that population.
- This is a polygenic model and the effect sizes w have infinitesimal prior.

Model Setup

- Genome-wide Association Studies (GWAS) are generally conducted by performing marginal linear regression, i.e., regress the trait on each SNP.
  - It is computationally feasible.
  - It is theoretically unstable to estimate full polygenic model (n << p).
  - It can also tag indirect association because of linkage disequilibrium (LD), which is actually helpful.
- Summary Statistics
  - $\hat{w} = \frac{1}{n} X^T y$, $se(\hat{w})$, etc.
  - They are largely available and sharable.

Model Setup

- Overlapping setting

$$\begin{pmatrix} y^{(1,s)} \\ y^{(1,*)} \end{pmatrix} = \begin{pmatrix} X^{(1,s)} \\ X^{(1,*)} \end{pmatrix} w^{(1)} + \begin{pmatrix} \epsilon^{(s)} \\ \epsilon^{(1,*)} \end{pmatrix}$$

$$\begin{pmatrix} y^{(2,s)} \\ y^{(2,*)} \end{pmatrix} = \begin{pmatrix} X^{(2,s)} \\ X^{(2,*)} \end{pmatrix} w^{(2)} + \begin{pmatrix} \delta^{(s)} \\ \delta^{(2,*)} \end{pmatrix}$$

  - $X^{(1,s)}$ and $X^{(2,s)}$ are the genotype of overlapping samples. They not strictly the same, since they might be standardized separately. But if the sample size n_1 and n_2 are relatively large enough, we may regard them as the same $X^{(s)}$ in the calculation.

- Correlated non-genetic factors (for overlapping samples)

  - $\begin{pmatrix} \epsilon \\ \delta \end{pmatrix} \sim N\left(0, \begin{pmatrix} (1-h_1^2)I_{n_1} & \rho_e J_s \\ \rho_e J_s^T & (1-h_2^2)I_{n_2} \end{pmatrix}\right)$, $J_s = \begin{pmatrix} I_s & 0 \\ 0 & 0 \end{pmatrix}_{n_1 \times n_2}$
  - $\rho_e = r_e\sqrt{(1-h_1^2)(1-h_2^2)}$ is the non-genetic covariance, r_e is the non-genetic correlation.

Model Setup

- Polygenic Risk Score (PRS)

$$\hat{t} = X^{(2)}\hat{w}^{(1)} = \frac{1}{n_1} X^{(2)} X^{(1)^T} y^{(1)}$$

- Consider the simple linear regression $y^{(2)} \sim \hat{t}$

$$y^{(2)} = \alpha + \gamma\hat{t} + \xi$$

  - How to estimate the effect size γ using summary statistics? (BADGERS)
  - How to estimate the R^2 using summary statistics?
  - What if there is sample overlapping?

Assumptions

- Genotypic data has normal prior

$$\begin{pmatrix} X^{(s)} \\ X^{(1,*)} \\ X^{(2,*)} \end{pmatrix} \sim N_{n_1+n_2-s,p}(0, I \otimes \Sigma)$$

  - All individuals are independent.
  - SNPs has correlation (LD) matrix $\Sigma_{p\times p}$.

- Effect sizes have infinitesimal prior

$$\begin{pmatrix} w_i^{(1)} \\ w_i^{(2)} \end{pmatrix} \sim N\left(0, \frac{1}{p}\begin{pmatrix} h_1^2 & \rho \\ \rho & h_2^2 \end{pmatrix}\right), i = 1, \ldots, p$$

  - $\rho = rh_1h_2$ is the genetic covariance, r is the genetic correlation.

Derivation

- $E(R^2) = E\big(E(R^2|X,w)\big) = E\left(E\left(\frac{y^{(2)^T}\hat{t}(\hat{t}^T\hat{t})^{-1}\hat{t}^T y^{(2)}}{y^{(2)^T}y^{(2)}}|X,w\right)\right)$

- $E(\hat{\gamma}) = E\big(E(\hat{\gamma}|X,w)\big) = E\left((\hat{t}^T\hat{t})^{-1}\hat{t}^T y^{(2)}|X,w\right)$

Results

- Expected R^2 (no sample overlapping)

$$E(R^2)\Big|_{s=0} \approx \frac{(1-h_1^2)(1-h_2^2)L_2 + (1-h_2^2)h_1^2(n_1+p)\frac{L_2}{p} + (1-h_1^2)h_2^2(n_2+p)\frac{L_2}{p} + r^2h_1^2h_2^2n_1n_2\frac{L_2^2}{p^2}}{n_2\left(L_2 + h_1^2n_1\frac{L_3}{p}\right)} \to r^2h_2^2\frac{L_2^2}{pL_3} + \frac{1-h_2^2}{n_2}\frac{L_2}{L_3}, \mathrm{n}_1 \to \infty$$

  - $L_2 = tr(\Sigma^2) = \mathrm{sum}(\text{LD scores})$
  - $L_3 = tr(\Sigma^3) \geq \frac{L_2^2}{p}$, hard to estimate.

- Expected effect size γ estimation (sample overlapping considered)

$$E(\hat{\gamma}) \approx n_1 \frac{s\rho_e p\sqrt{(1-h_1^2)(1-h_2^2)} + \frac{rh_1h_2}{p}\left((n_1n_2+s)L_2 + sp^2\right)}{(1-h_1^2)\left((n_1n_2+s)L_2 + sp^2\right) + \frac{h_1^2}{p}\left(\left((2n_1+4)s + n_1n_2(n_1+1)\right)L_3 + p\left((2n_1+3)s + n_1n_2\right)L_2 + sp^3\right)}$$

$$= \frac{rh_1h_2n_1L_2}{pL_2 + h_1^2(n_1+1)L_3}\Big|_{s=0} \to r\frac{h_2}{h_1}\frac{L_2}{L_3}, n_1 \to \infty$$

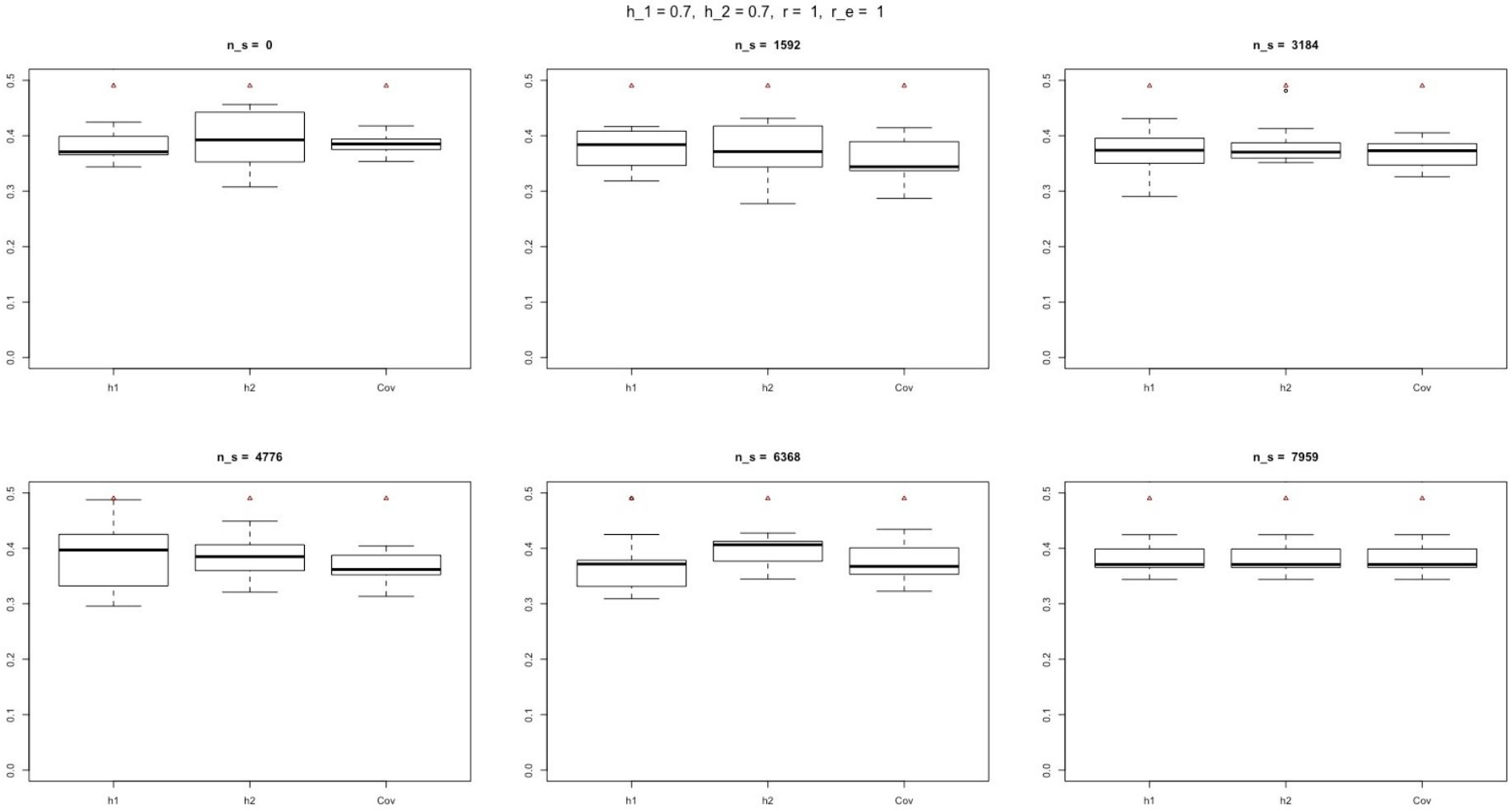
Proposed method for robust R^2 estimation

- Use LD score regression to estimate heritability and genetic correlation since it is robust to the sample overlapping problem.
- Use the formula we just derived, plug in the estimated heritability and genetic correlation to get a (hopefully) good estimation of R^2.

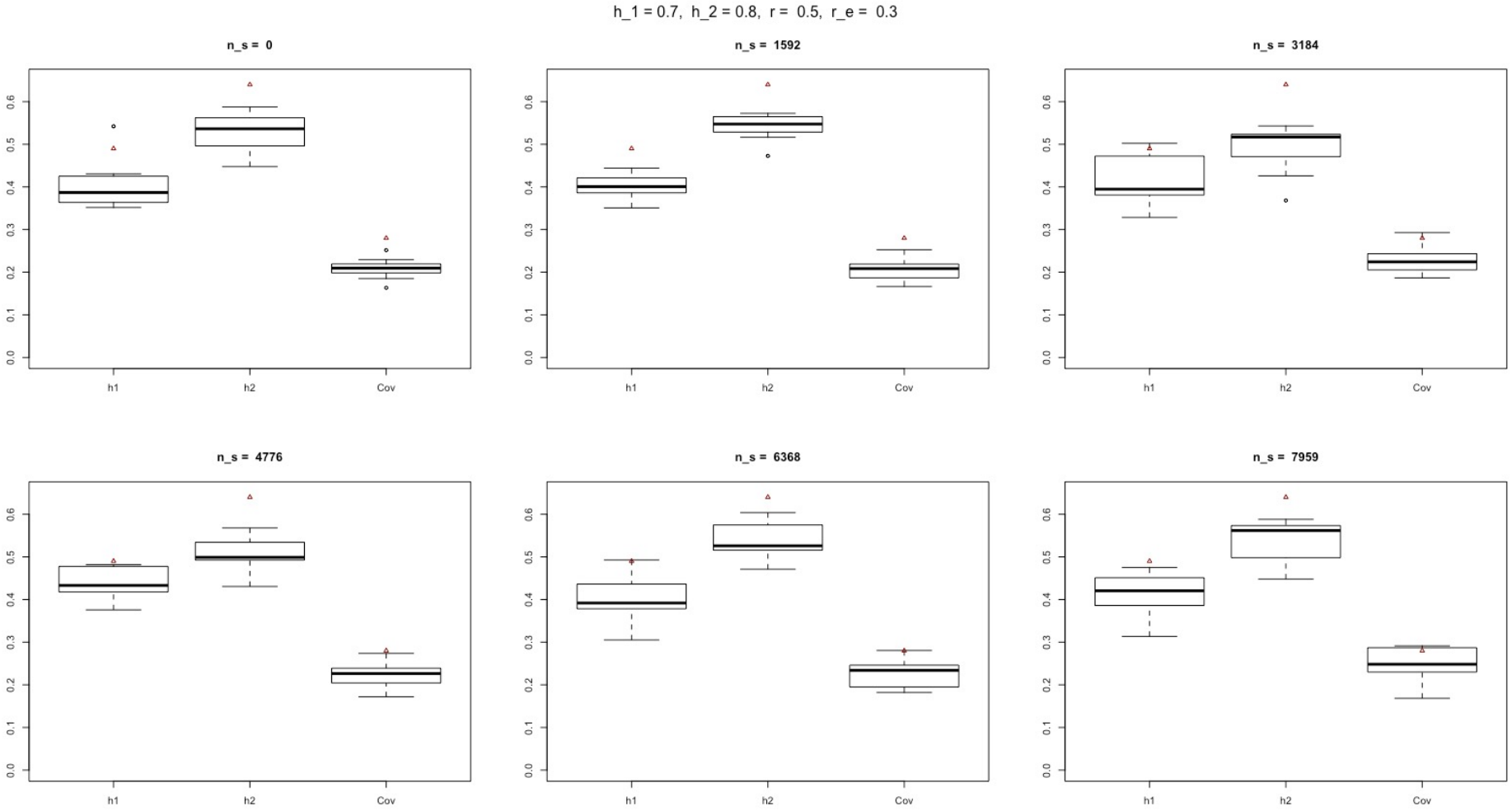
Simulation Pipeline

- Data source: WTCCC genotype data ($n = 15918, p = 336345$)
- QC (MAF cutoff = 0.05)
- Create sample-overlapping training and testing set (equal size)
  - Overlapping rate $(0\%, 20\%, 40\%, 60\%, 80\%, 100\%)$
- Use R and GCTA to simulate phenotype.
  - For every set of parameters (h_1, h_2, r, r_e), repeat 10 times, on 1 training dataset and 6 testing dataset.
- Use PLINK to perform GWAS.
- Use PRSice to calculate PRS.
  - Regress the testing phenotype on PRS to get empirical R^2
- Use LDSC to estimate heritability and genetic covariance [5, 6].
  - Plug the estimation into our formula to get robust inferred R^2

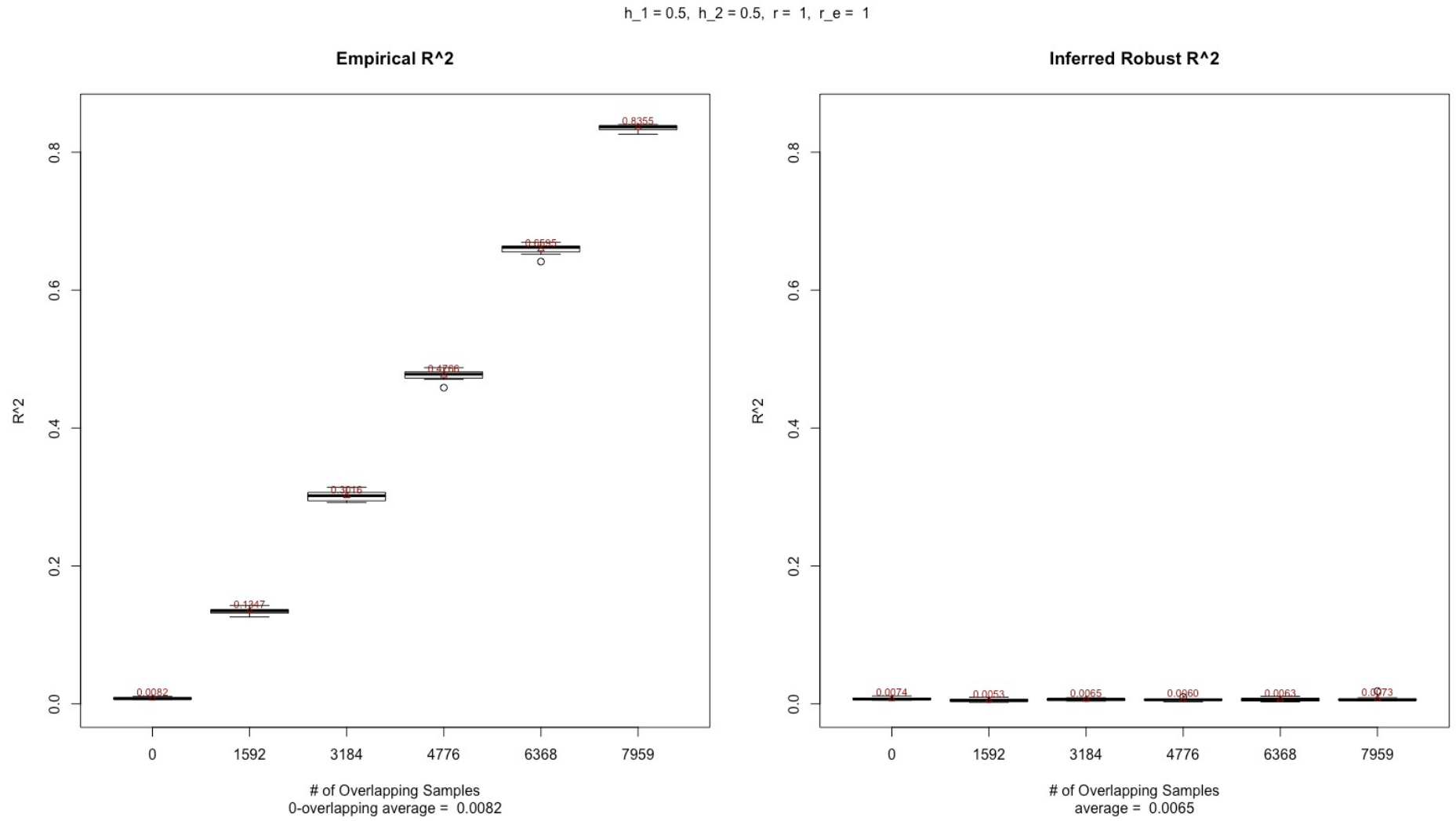
Simulation Results (LDSC)



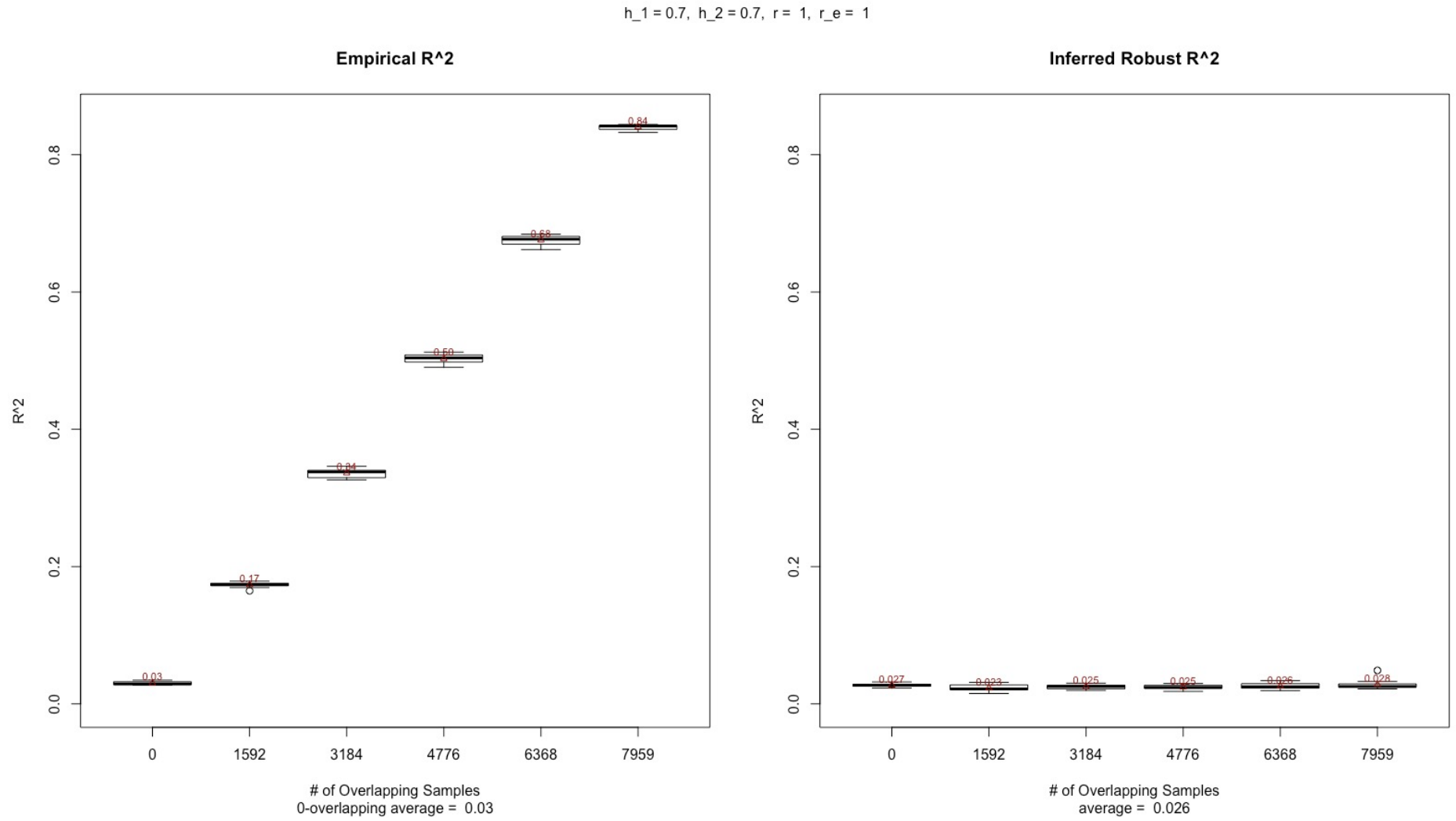
Simulation Results (LDSC)



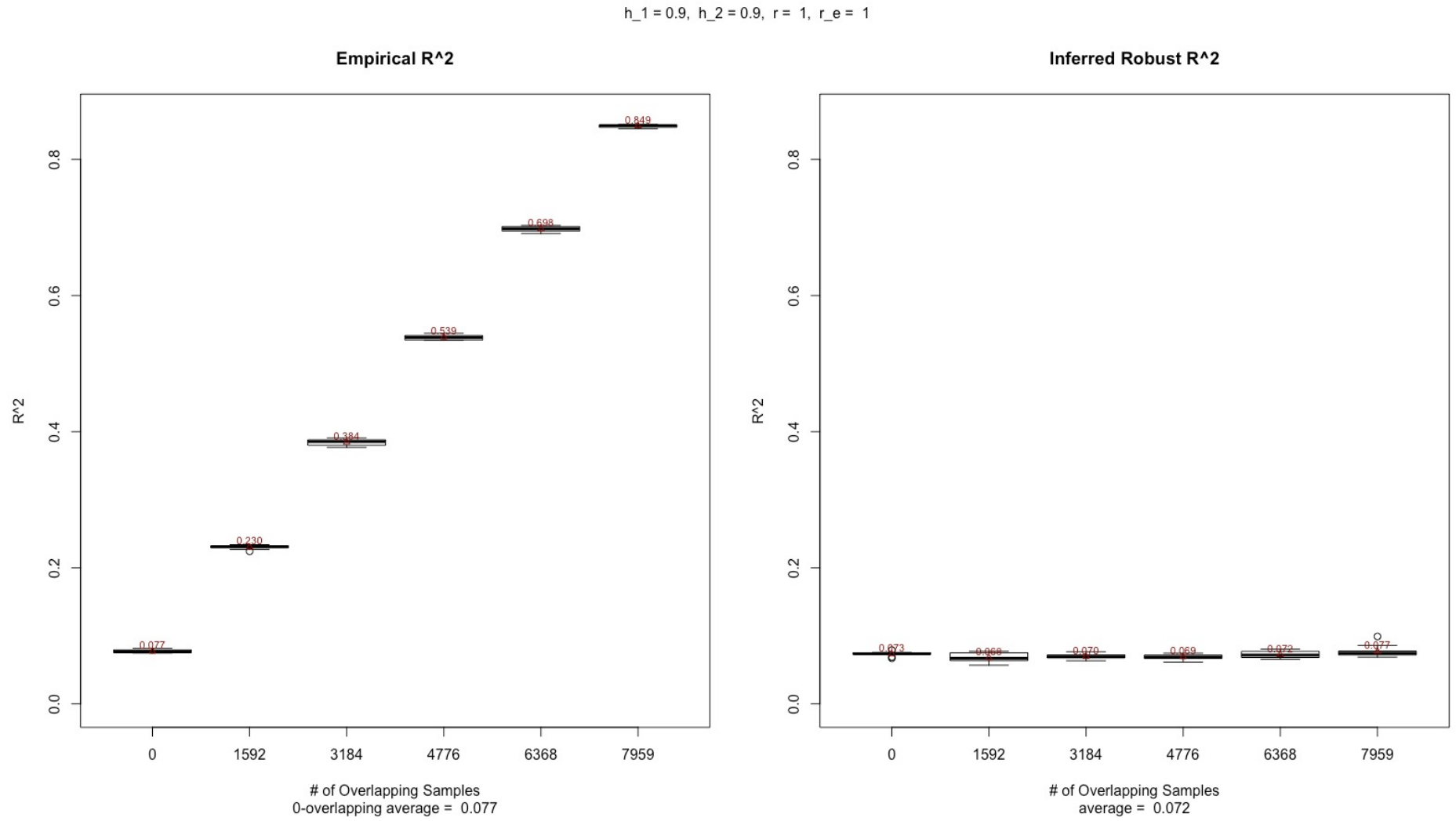
Simulation Results



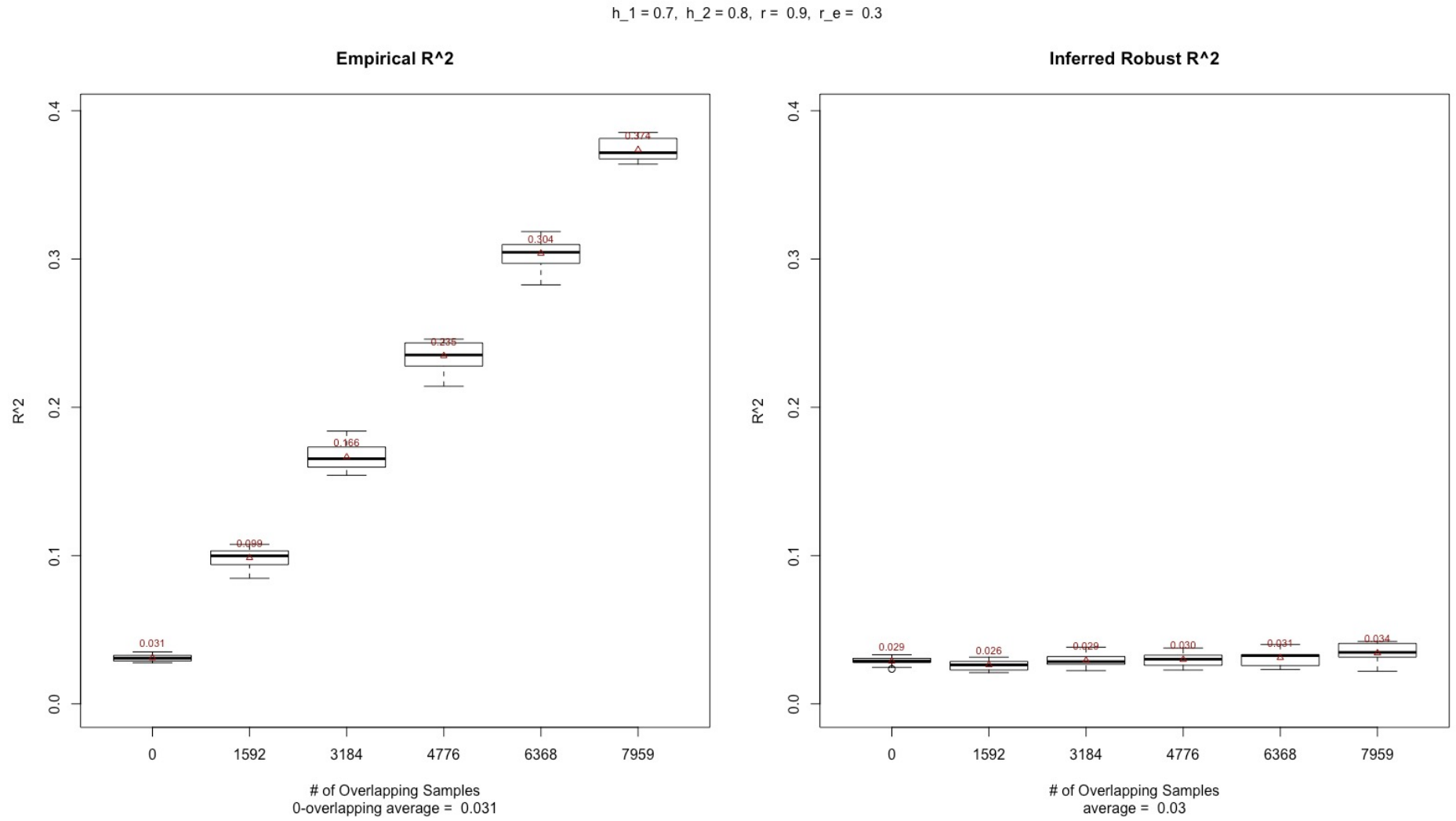
Simulation Results



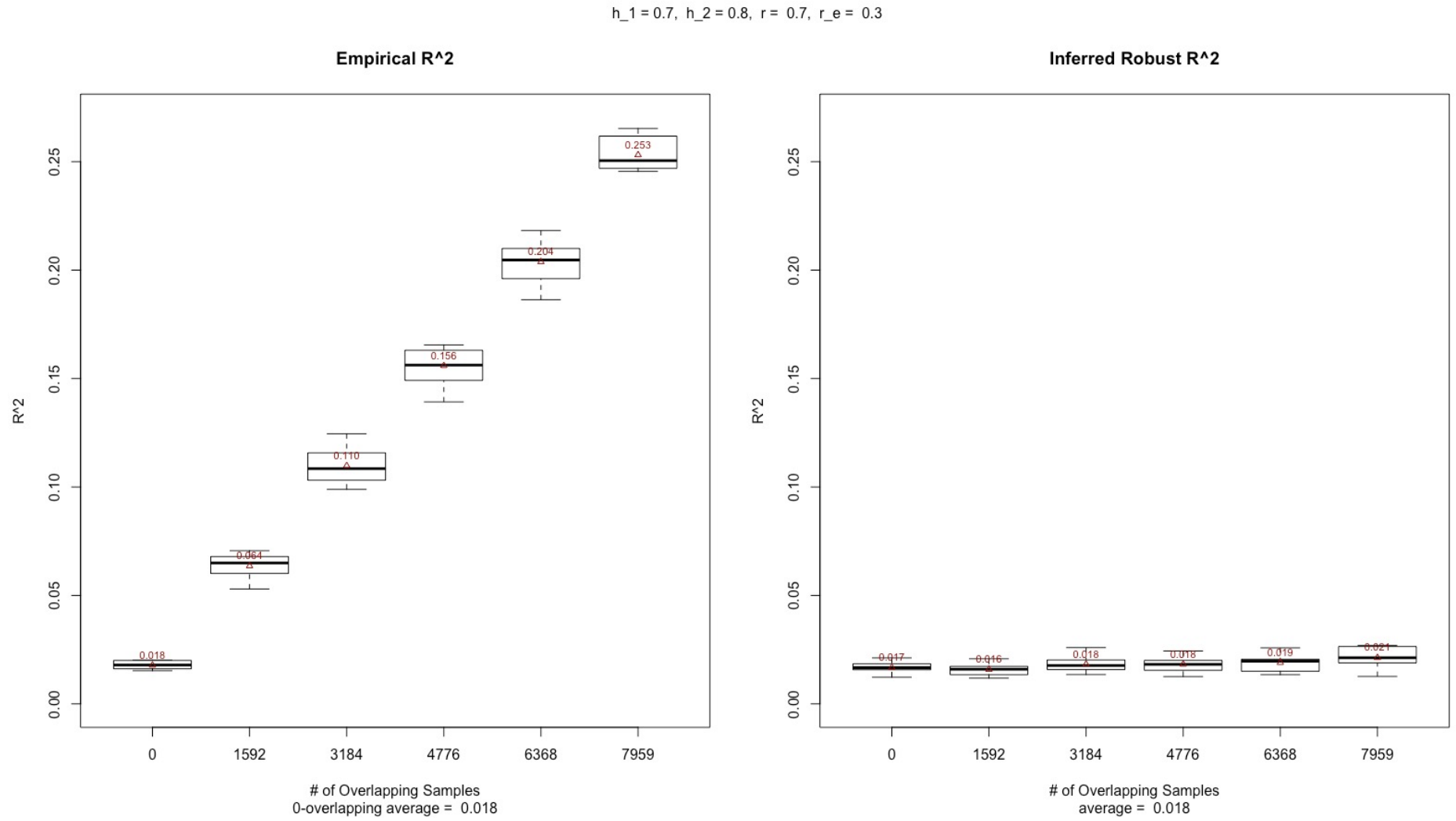
Simulation Results



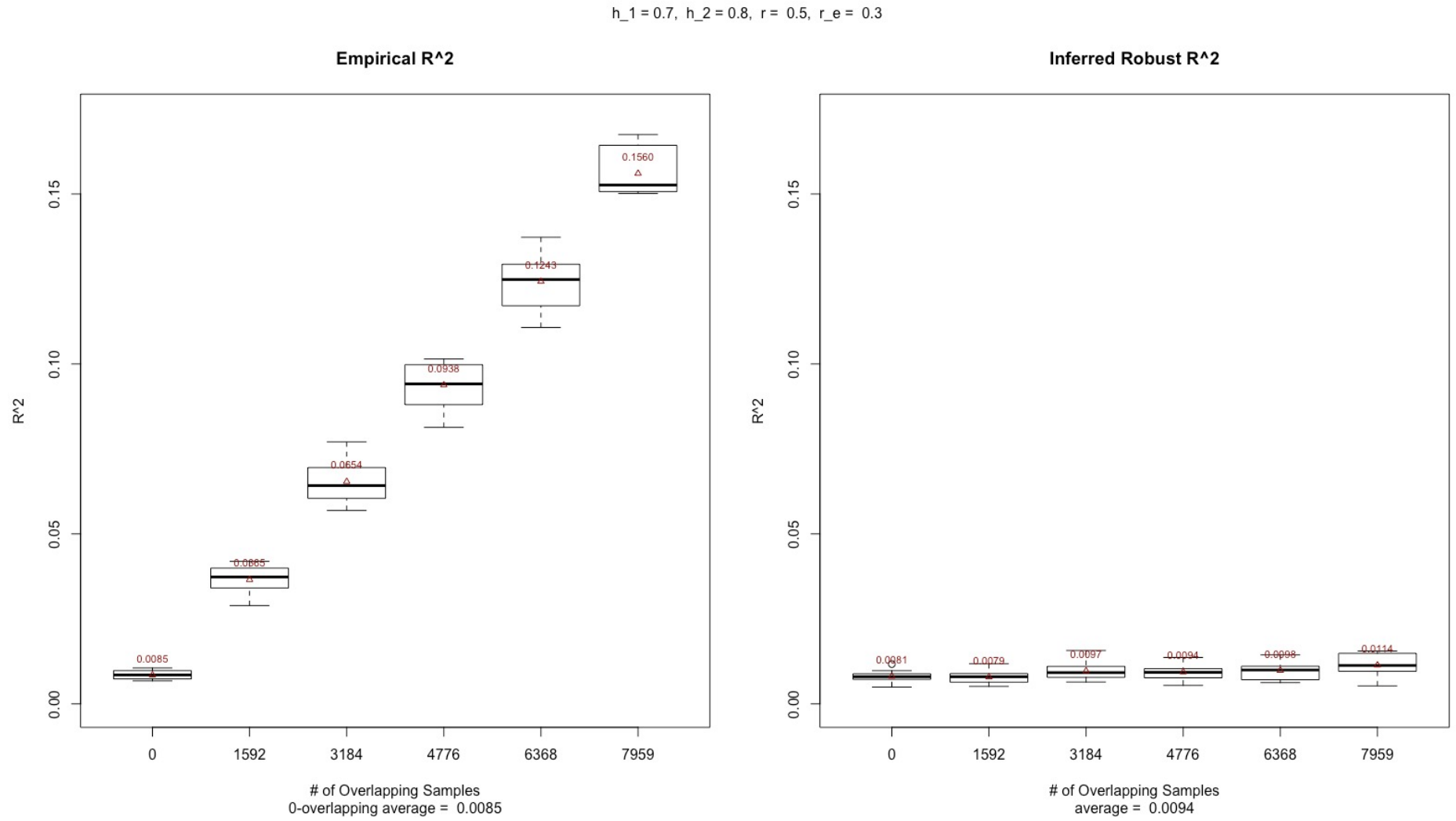
Simulation Results



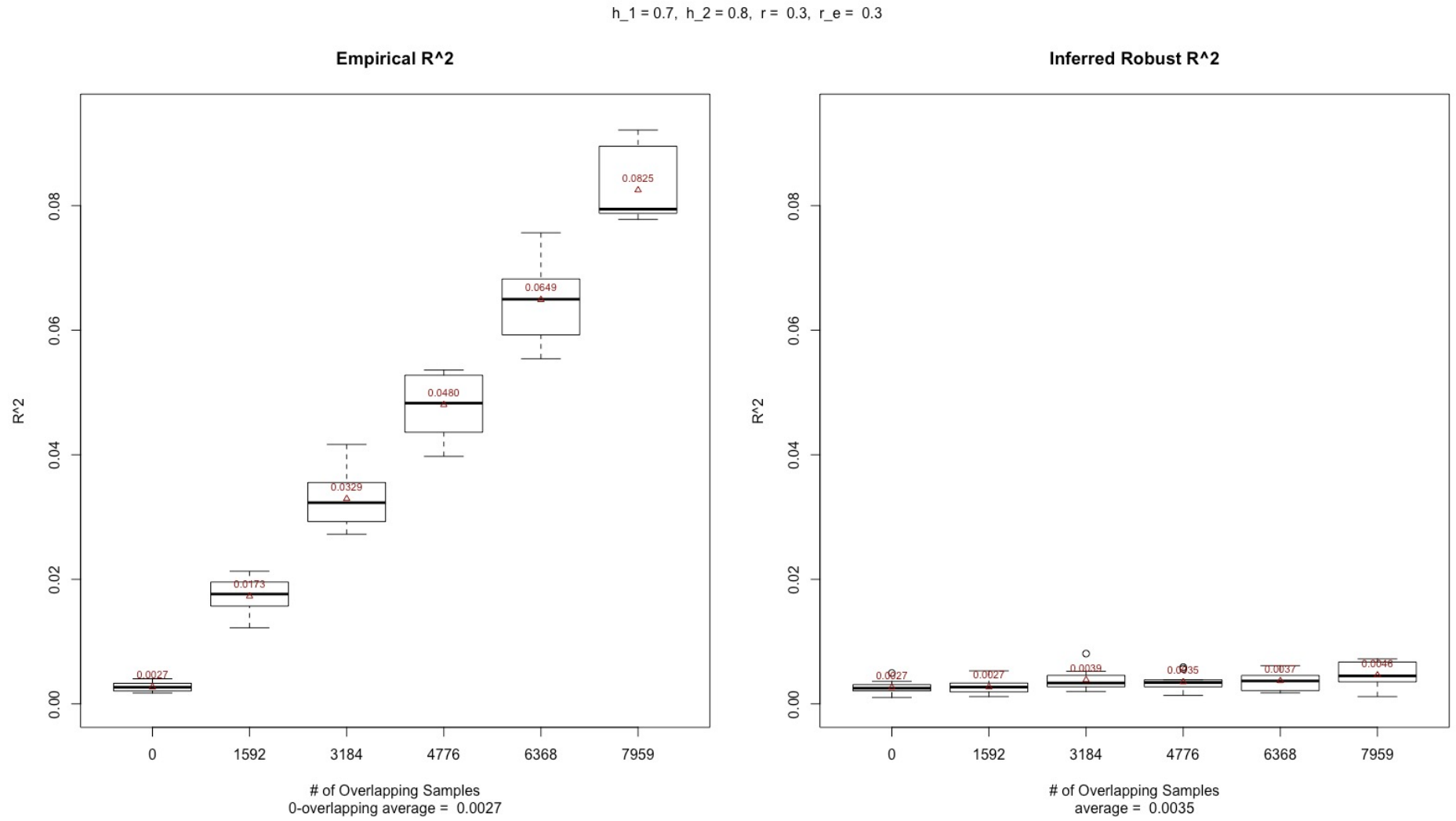
Simulation Results



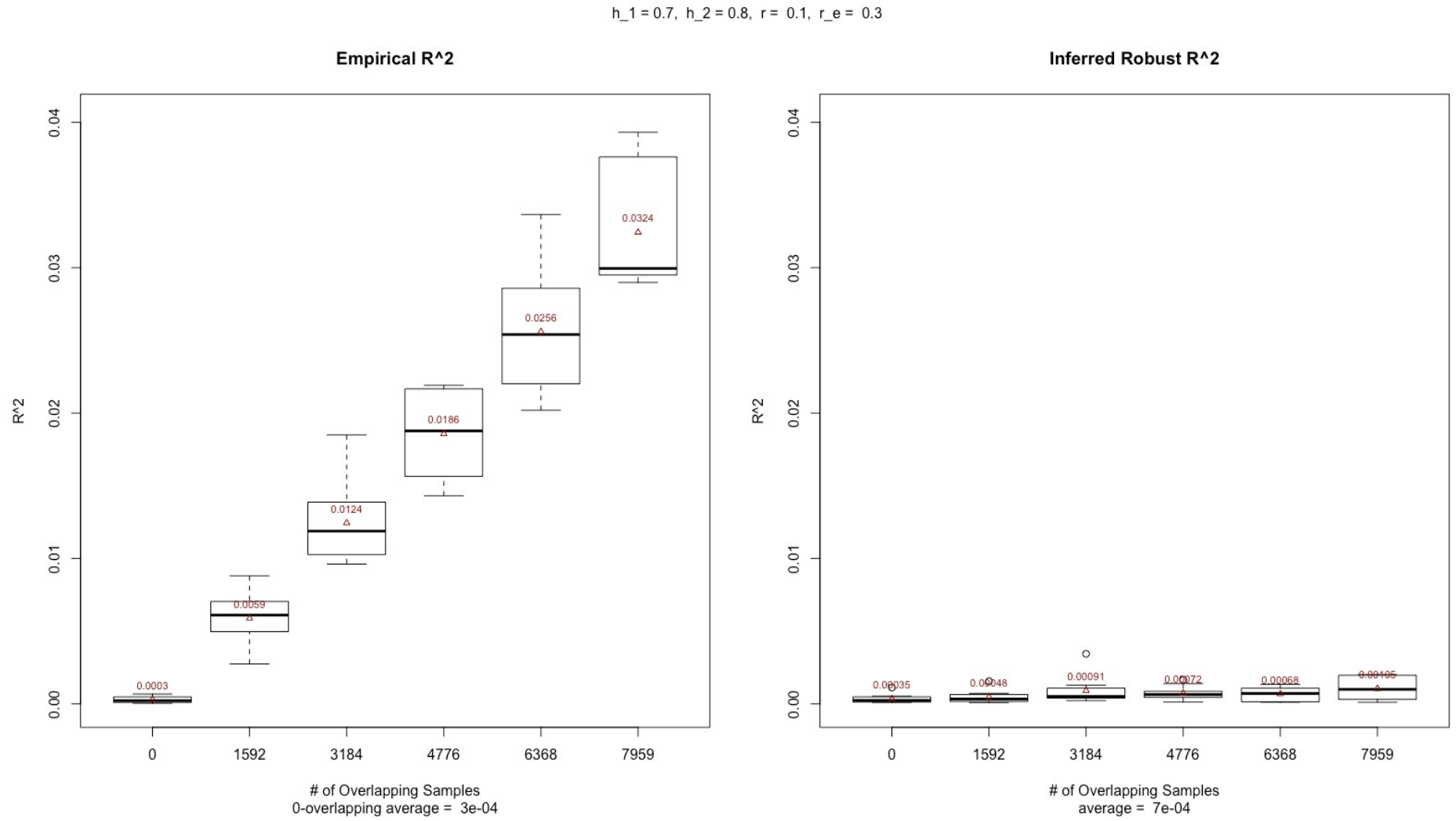
Simulation Results



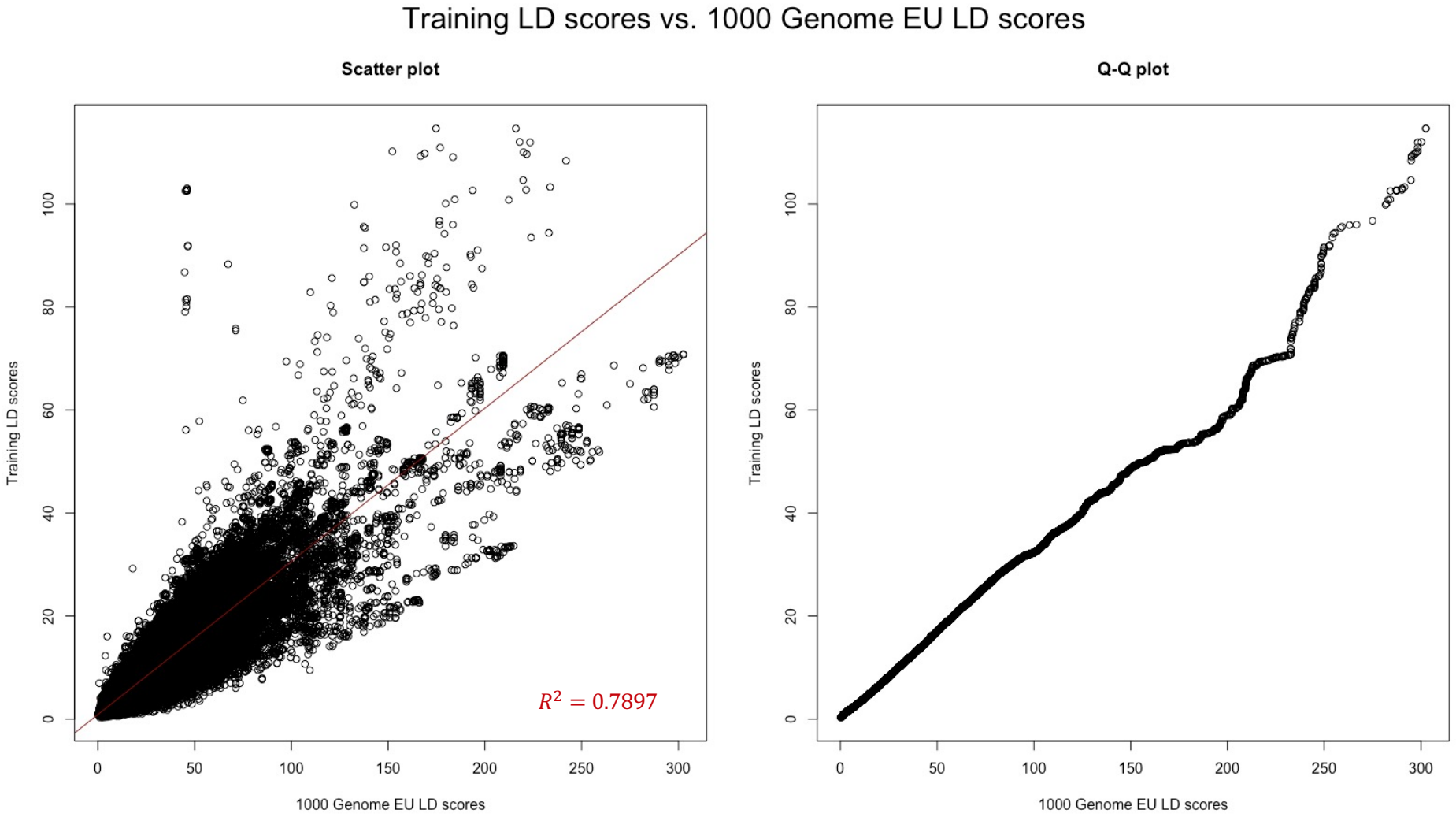
Simulation Results



Simulation Results



Comparison of LD scores



Future Work

- Try our method on real data.
- How do we quantify our estimator's variability?
- Change number of SNPs in PRS calculation.
- How to better estimate LD related quantities by using publicly available data?
- What if we have individual-level testing set, can we improve?
- What if we assume heterogeneous SNP effect sizes?
- Would results still hold if we assume a mild assumption on genotype?

References

[1] Euesden, Jack, Cathryn M. Lewis, and Paul F. O'Reilly. "PRSice: polygenic risk score software." *Bioinformatics* 31.9 (2014): 1466-1468.

[2] Dudbridge, Frank. "Power and predictive accuracy of polygenic risk scores." *PLoS genetics* 9.3 (2013): e1003348.

[3] Power, Robert A., et al. "Polygenic risk scores for schizophrenia and bipolar disorder predict creativity." *Nature neuroscience* 18.7 (2015): 953.

[4] Yan, Donghui, et al. "Biobank-wide association scan identifies risk factors for late-onset Alzheimer's disease and endophenotypes." *bioRxiv* (2018): 468306.

[5] Bulik-Sullivan, Brendan K., et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." *Nature genetics* 47.3 (2015): 291.

[6] Bulik-Sullivan, Brendan, et al. "An atlas of genetic correlations across human diseases and traits." *Nature genetics* 47.11 (2015): 1236.

Thank you!
Q&A

Li Ge, lge7@wisc.edu
Ph.D. Student in Biomedical Data Science
University of Wisconsin-Madison
Rotation Advisor: Qiongshi Lu