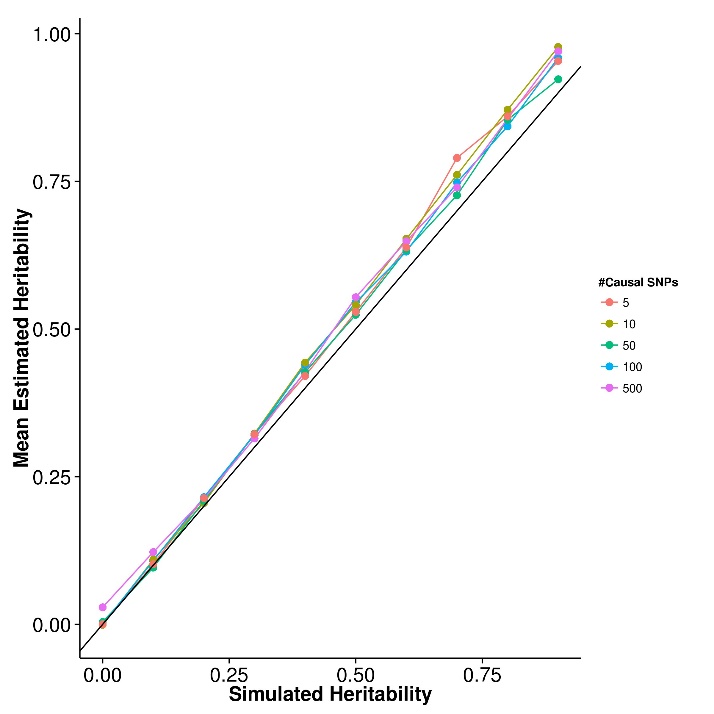
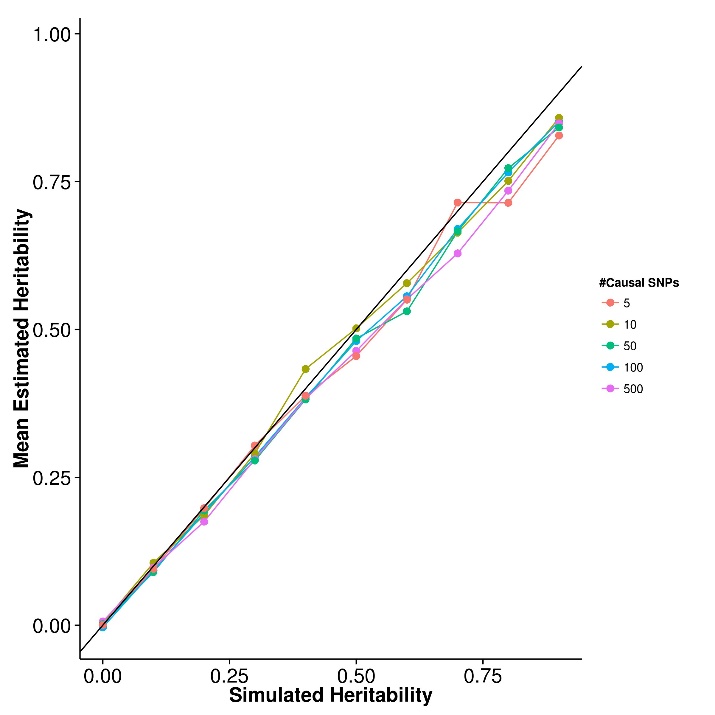
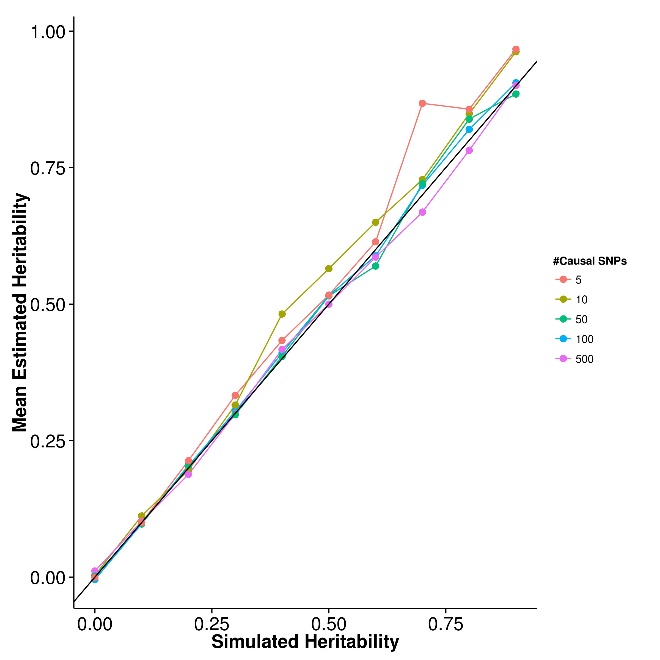
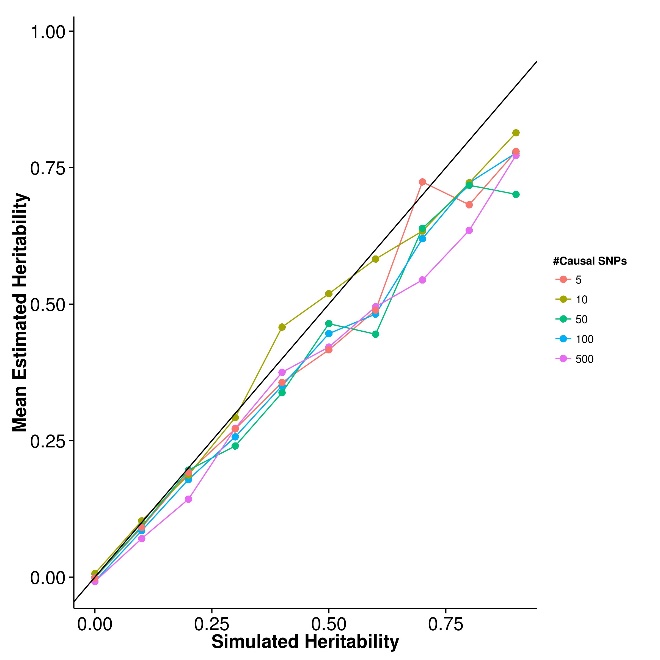
**Quantitative Trait**

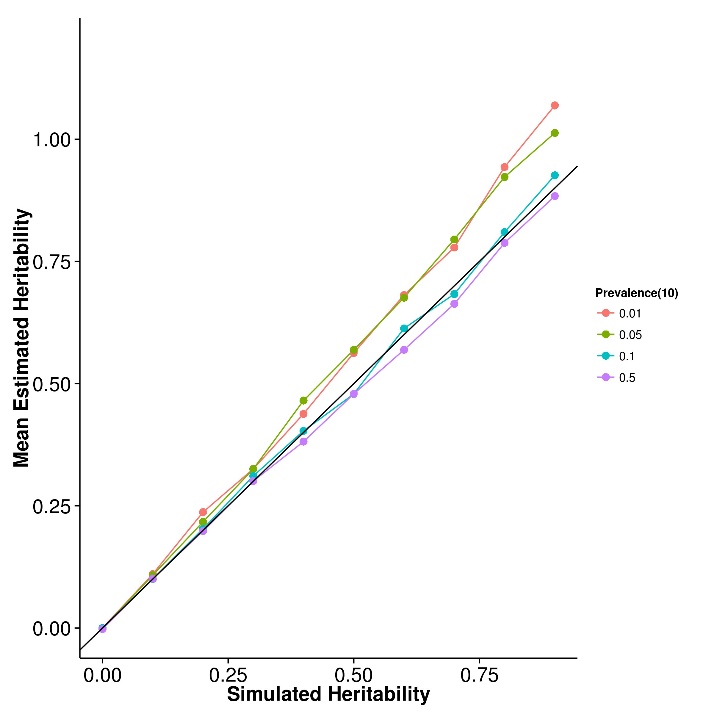


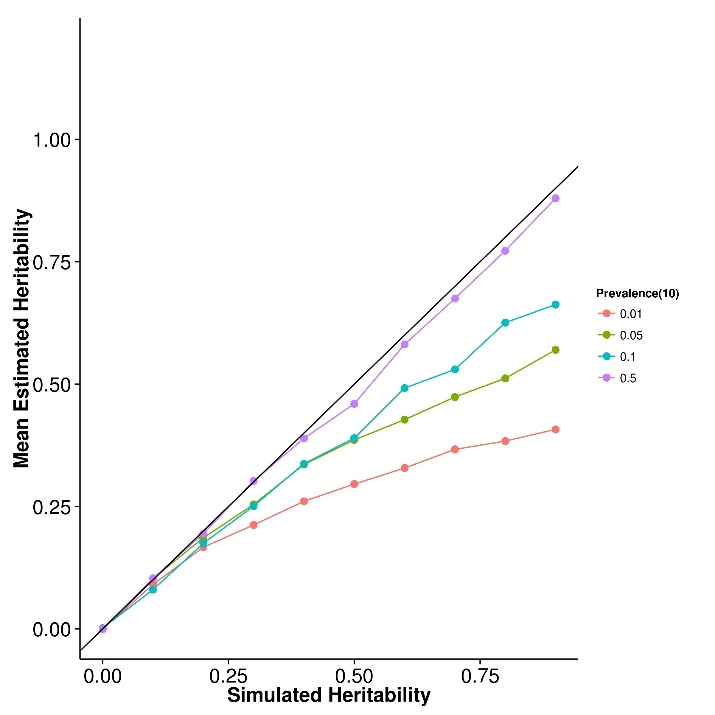
**(a)** SHREK **(b)** GCTA

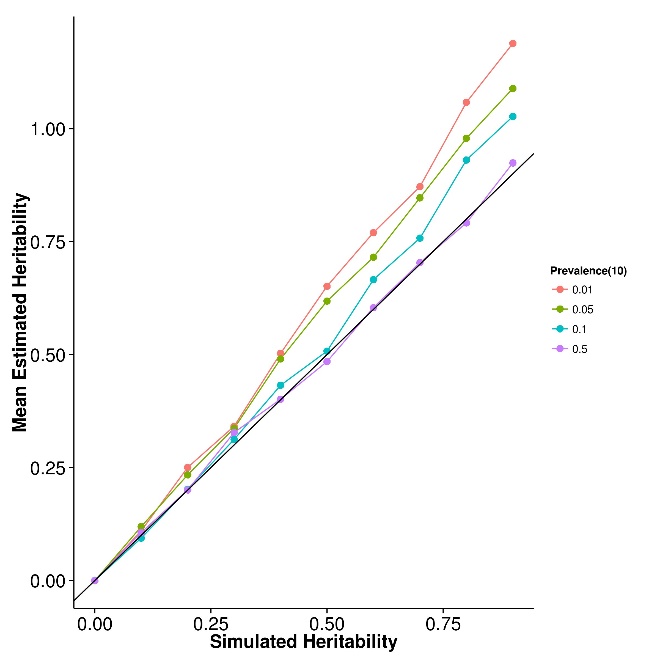
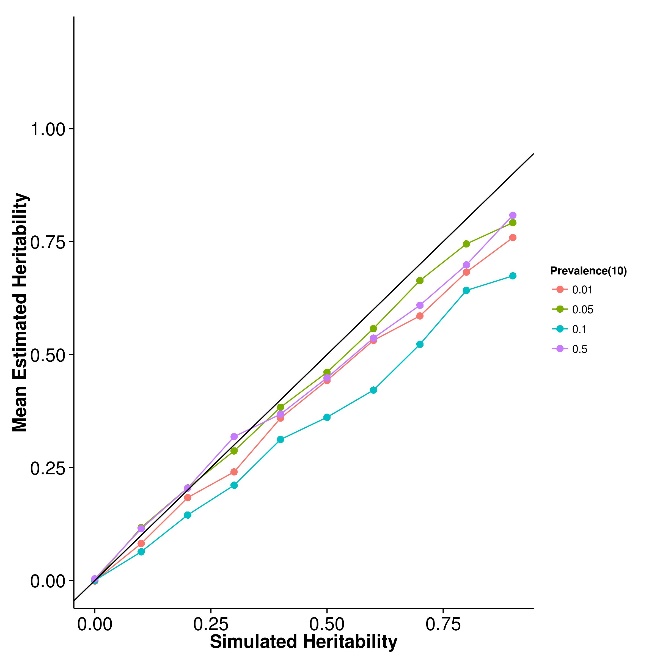


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of Causal SNPs | SHREK | LDSC | LDSC-In | GCTA |
| 5 | 0.0235 | 0.0576 | 0.0828 | 0.0365 |
| 10 | 0.0231 | 0.0343 | 0.0555 | 0.0189 |
| 50 | 0.0196 | 0.0157 | 0.0494 | 0.0114 |
| 100 | 0.0210 | 0.0129 | 0.0363 | 0.00961 |
| 500 | 0.0205 | 0.0115 | 0.0308 | 0.00887 |

**Case Control (10 causal SNPs)**



**(a)** SHREK **(b)** GCTA

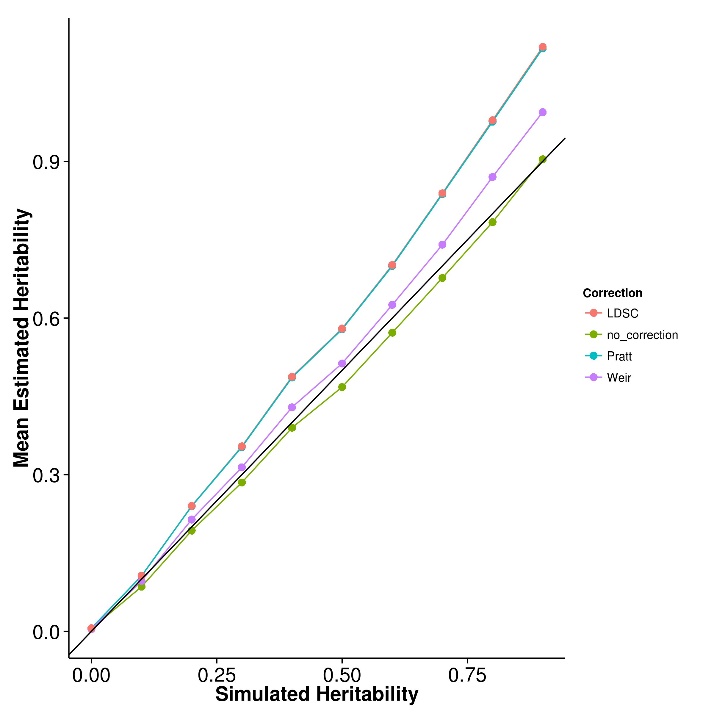


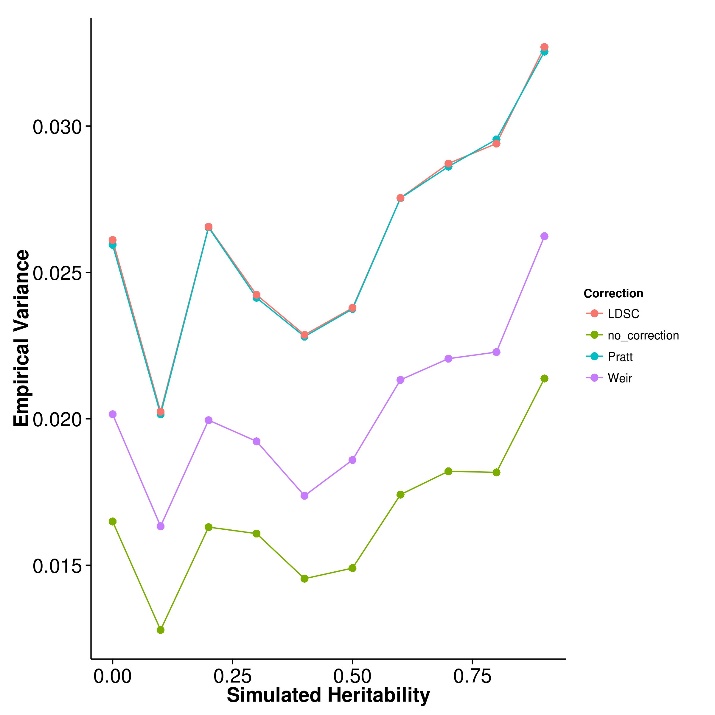
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Population  Prevalence | Number of  Causal SNPs | SHREK | LDSC | LDSC-In | GCTA |
| 0.01 | 10 | **0.0145** | 0.0361 | 0.0164 | 0.0675 |
| 0.01 | 50 | 0.0135 | 0.0254 | **0.00791** | 0.0702 |
| 0.01 | 100 | 0.0128 | 0.0227 | **0.0102** | 0.0698 |
| 0.01 | 500 | **0.0126** | 0.0214 | 0.0150 | 0.0710 |
| 0.05 | 10 | 0.0110 | 0.0201 | **0.00983** | 0.0302 |
| 0.05 | 50 | **0.00453** | 0.00974 | 0.0115 | 0.0299 |
| 0.05 | 100 | **0.00569** | 0.0113 | 0.00981 | 0.0304 |
| 0.05 | 500 | **0.00540** | 0.00999 | 0.0171 | 0.0305 |
| 0.1 | 10 | **0.00512** | 0.0109 | 0.0301 | 0.0165 |
| 0.1 | 50 | **0.00381** | 0.00824 | 0.0105 | 0.0152 |
| 0.1 | 100 | **0.00418** | 0.00802 | 0.0163 | 0.0148 |
| 0.1 | 500 | **0.00400** | 0.00740 | 0.0141 | 0.0155 |
| 0.5 | 10 | 0.00560 | 0.00749 | 0.0219 | **0.00410** |
| 0.5 | 50 | 0.00362 | 0.00528 | 0.0232 | **0.00244** |
| 0.5 | 100 | 0.00356 | 0.00460 | 0.0208 | **0.00225** |
| 0.5 | 500 | 0.00338 | 0.00365 | 0.0159 | **0.00200** |

**Real Data**

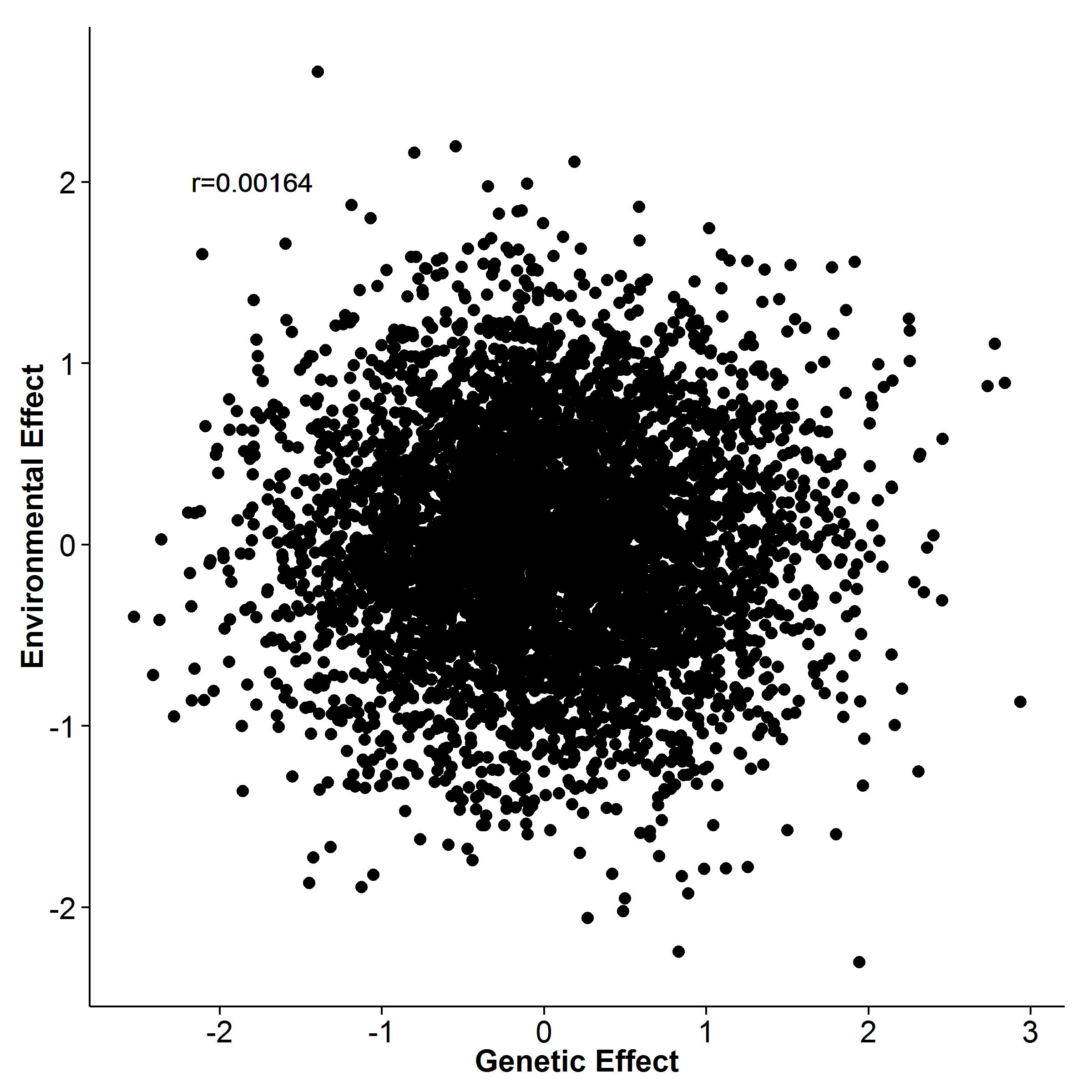
|  |  |  |  |
| --- | --- | --- | --- |
|  | Major Depression Disorder | Bipolar | Schizophrenia |
| SHREK | 0.252 (0.0273) | 0.308 (0.0167) | 0.185 (0.00450) |
| LDSC | 0.232 (0.0217) | 0.265 (0.0152) | 0.198 (0.0057) |
| LDSC-In | 0.154 (0.033) | 0.181 (0.0203) | 0.135 (0.0072) |

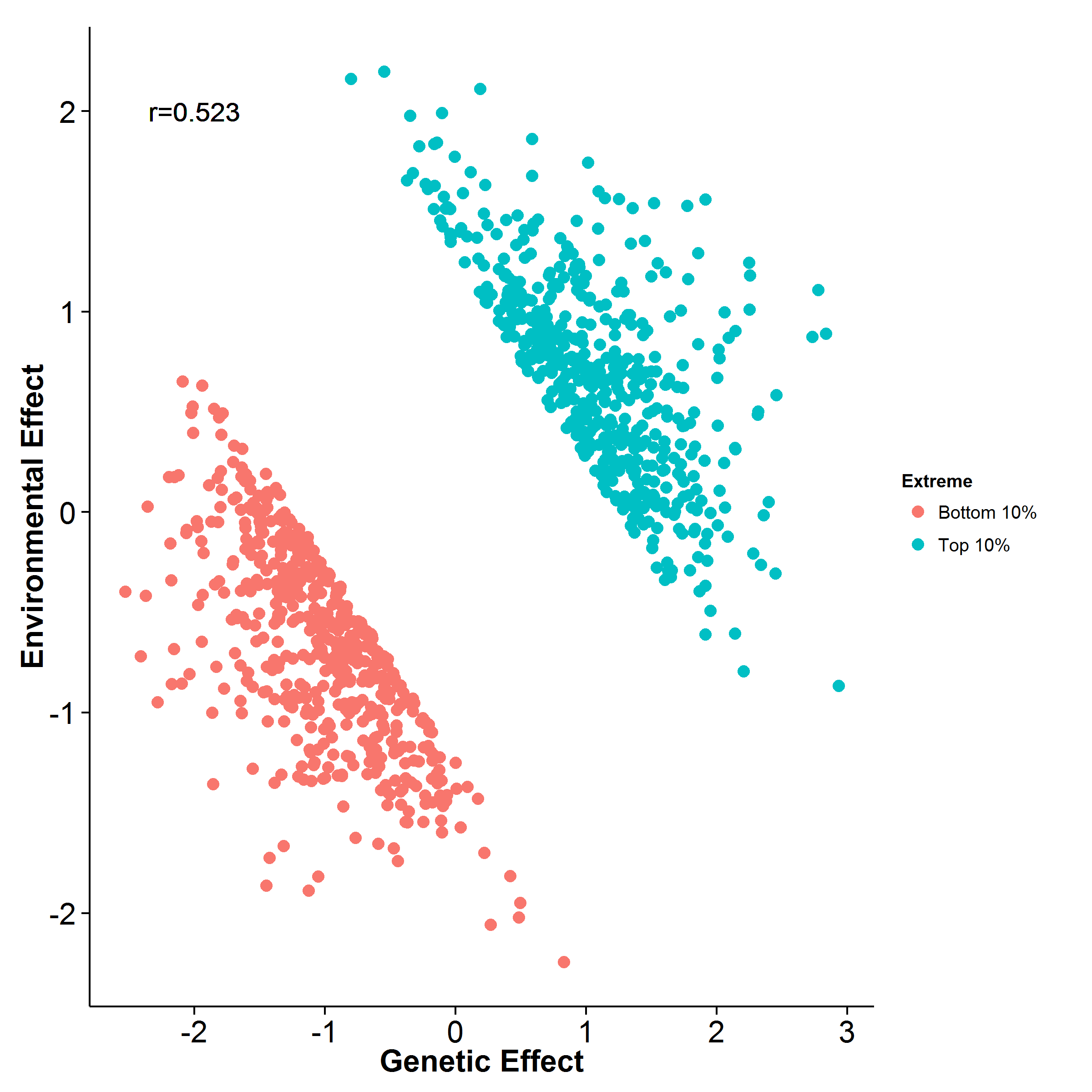
**LD Correction (With 50,000 SNPs)**



**(a)** Mean Estimation **(b)** Empirical Variance

**Extreme Sample Selection**



**(a)** No Extreme sampling **(b)** Extreme Sampling

Here we simulated the genetic and environmental effect independently. When extreme sampling was performed, an artificial correlation was observed. This might be the main reason why the estimates from GCTA are downward biased.

# Problem Observed

1. Inflated estimates in case control / extreme phenotype selection
   * Bias increase as prevalence decrease
2. LD Correction
   * With more SNP, correction introduce bias
3. Still produce funny result, e.g. all negative, for some real data (e.g. PGC Autism data)

# Current Progress

1. Re-implementing the programme with different C++ library
   * Expect to increase speed by 3 fold and improve memory footprint
   * Previous implementation require 2 days to finish 1 chromosome using PGC SCZ data
2. Try to “remove” perfect LD from the LD matrix
   * Previous experience suggest that a unreasonable result is usually a product of failed decomposition
   * By removing the perfect LD from the LD matrix, we can improve the condition number of the matrix by 100 fold (however, the condition number is still large)
   * For any pair of SNPs in perfect LD, e.g. SNP A and SNP B, we will use the mean of summary statistic of the SNP pairs as the input to our algorithm
3. Try to implement the more “sophisticated” method of SE estimation
   * This involve more complicated algorithm
4. After re-implementing SHREK, we will re-implement the simulation script for better simulation