Estimating Heritability under Case-Control Sampling

Yu Xinyi

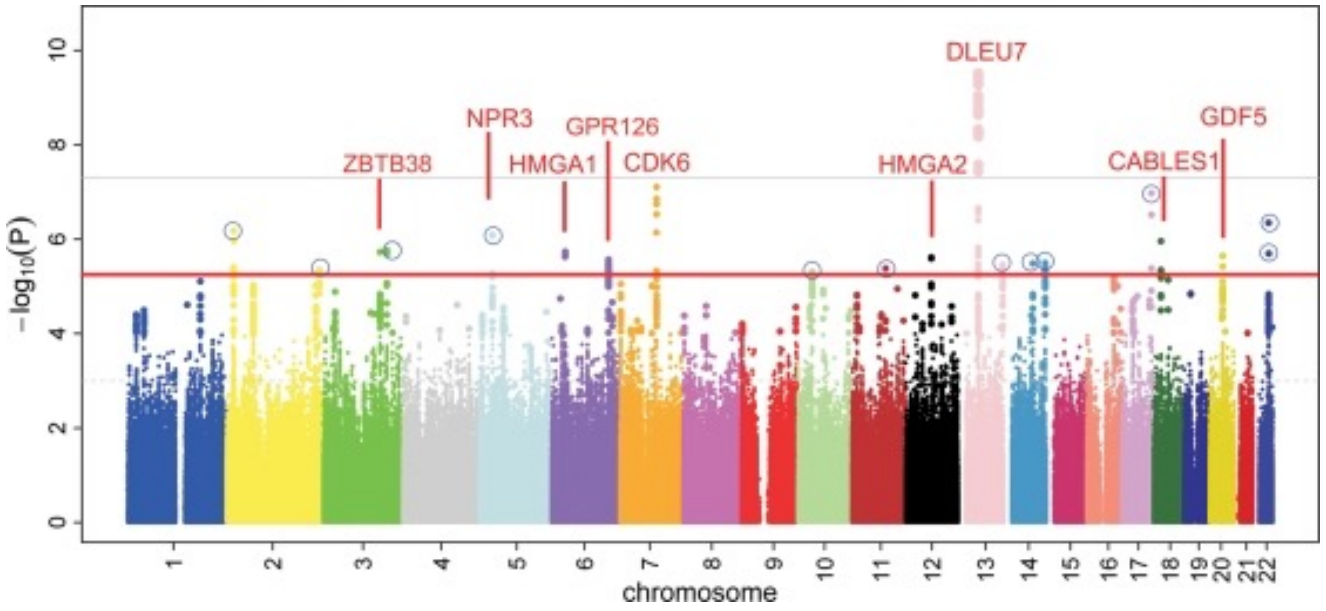
July 9, 2021

Outline

- Background
- Estimating heritability – quantitative traits
- Estimating heritability – disease traits
- Simulations and applications
- Derivation - PCGC regression
- Discussion

Background

- **Heritability** is the proportion of phenotype variance that is due to genetic effects.
- **Broad-sense heritability** (H^2) is the proportion of phenotype variance due to all genetic effects.
- **Narrow-sense heritability** (h^2) is the proportion of phenotype variance due to additive genetic effects.
- h^2_{GWAS} is the proportion of phenotype variance explained by genome-wide significant SNPs.



Manhattan plot of the height association study result

Background

- Genome-Wide Association Studies (GWASs) have uncovered thousands of genetic variants associated with hundreds of diseases.
- However, the variants that reach statistical significance typically explain only a small fraction of the heritability.
- **Missing heritability:** $h^2_{GWAS} \ll h^2$.

Trait / Disease	h^2 Pedigree Studies	h^2 GWAS Hits
Bipolar disorder	0.6-0.7	0.02
BMI	0.4-0.6	0.01-0.02
Crohn's disease	0.6-0.8	0.1
Height	0.8	0.1
HDL cholesterol	0.5	0.1
Type 2 diabetes	0.3-0.6	0.05-0.1
Schizophrenia	0.7-0.8	0.01

Visscher et al. Five year of GWAS discovery. 2012 AJHG

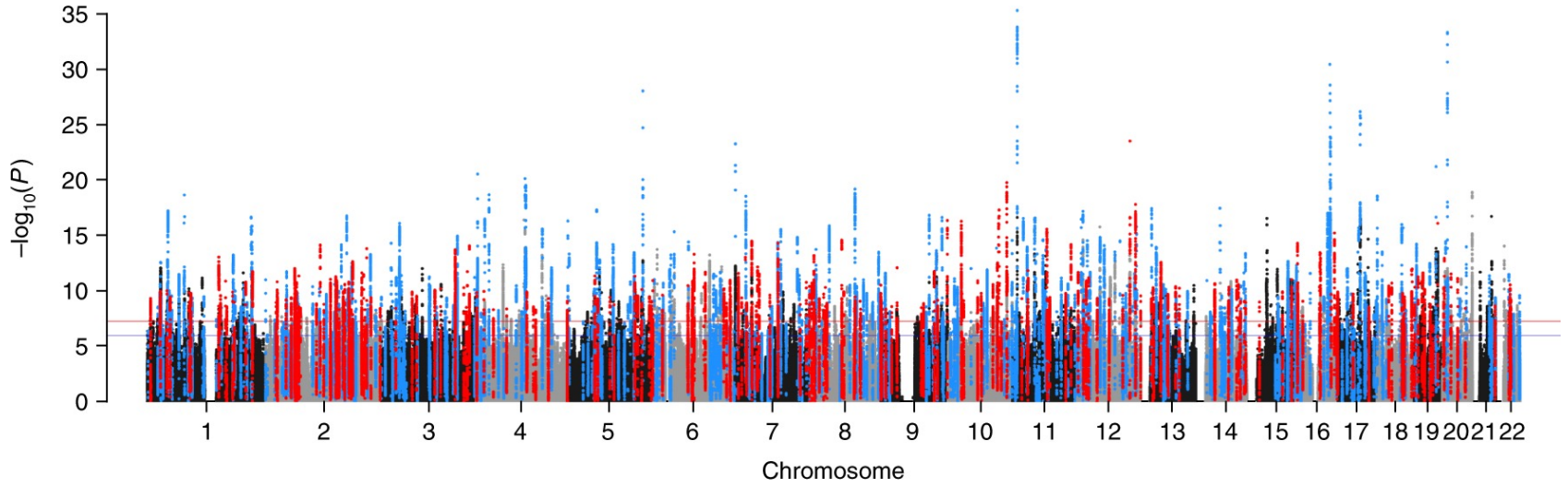
Background

- Where is the “missing heritability”?

  Common causal variants of exceedingly low effect sizes

  - Most complex traits are extremely polygenic.

$$h^2_{GWAS} \ll h^2$$



Manhattan plot of novel 535 loci discovered by a genome-wide association meta-analysis of 757,601 individuals across blood pressure traits. Image from Evangelous et al. 2018 Nature Genetics.

Background

$h^2_{GWAS} \ll h^2$

- Where is the "missing heritability"?

  Common causal variants of exceedingly low effect sizes

  - 10 common risk variants, each explaining 1/10 of h^2?

    100 common risk variants, each explaining 1/100 of h^2?

    1,000 common risk variants, each explaining 1/1,000 of h^2?

    10,000 common risk variants, each explaining 1/10,000 of h^2?

  - Infinitesimal model:

    All M common SNPs in the genome are risk variants, each with effect size $\sim N(0, h^2/M)$.

  - What will happen if we try to estimate effect size of each variant and sum the square of these estimated effect sizes up as estimation of heritability?

Background

- Let's look at a toy example.

$$y_i = \sum_{k=1}^{m} x_{ik}\beta_k + e_i, i = 1,2, \dots, n$$

- Suppose we have 1000 SNPs, equally explaining heritability:

  $\beta_k \sim N\left(0, \frac{h^2}{1000}\right), e_i \sim N(0, 1 - h^2), h^2 = 0.5.$

- x_k have mean 0 and variance 1 so that $y_i \sim N(0, 1)$.
- Do simple linear regression on each SNP to get estimated effect size $\hat{\beta}_k$ and calculate $\sum_{k=1}^{m} \hat{\beta}_k^2$.
- $\sum_{k=1}^{m} \hat{\beta}_k^2$ exceeds h^2: we add too much noise because the sample size is too small.
- We do not want to include that much noise into our estimation.
- But we only have a small number of samples.

Background

- How about setting a strict threshold to include those variants we are more confident (with larger signals)?
- Threshold: $0.05/m$ (Bonferroni correction). SNPs whose p-value below this threshold will be included.
- Now sum the estimated effect sizes of these selected SNPs.
- When the sample size (n=500, n=1000) is small relative to the number of variants (m=1000), the aggregate effect size of significant SNPs are far from total heritability ($h^2 = 0.5$).
- But as sample size increases, this value will grow to approach the total heritability.

Background

- In a real GWAS, there are hundreds of thousands SNPs (e.g $m \approx 300{,}000$) with dozens of thousands samples (e.g $n \approx 34{,}000$).
- This sample size is already not small, but still too small to detect all associated SNPs.
- We need to find some other way to estimate heritability.
- Start from formally defining heritability estimation problem in mathematical language.

Many sequence variants affecting diversity of adult human height

Nature Genetics **40**, 609–615 (2008) | Cite this article

Abstract

Adult human height is one of the classical complex human traits[1]. We searched for sequence variants that affect height by scanning the genomes of 25,174 Icelanders, 2,876 Dutch, 1,770 European Americans and 1,148 African Americans. We then combined these results with previously published results from the Diabetes Genetics Initiative on 3,024 Scandinavians[2] and tested a selected subset of SNPs in 5,517 Danes. We identified 27 regions of the genome with one or more sequence variants showing significant association with height. The estimated effects per allele of these variants ranged between 0.3 and 0.6 cm and, taken together, they explain around 3.7% of the population variation in height. The genes neighboring the identified loci cluster in biological processes related to skeletal development and mitosis. Association to three previously reported loci are replicated in our analyses[3,4,5], and the strongest association was with SNPs in the *ZBTB38* gene.

All these samples were genotyped with SNP chips containing a superset of the HapMap panel on the 317K Illumina chip. After controlling for quality, we had 304,226 SNPs available for analysis. All

Estimating Heritability – Quantitative Traits

- Our interest is in estimating the heritability of a trait attributed to common variants.
- General model for quantitative traits:

Normalized genotype: $g_i = (g_{i1}, g_{i2}, \cdots, g_{im})$

Normalized phenotype ← $p_i = \psi(g_i, e_i)$ → Environmental effect

- Additive model for quantitative traits:

$$p_i = g_i + e_i$$

No G×G or G×E interactions

$$g_i = \sum_k u_k g_{ik}$$

u_k: Effect size of the kth variant

- Narrow-sense heritability:

Additive contribution of genes

$$h^2 = \sum_k u_k^2$$

Estimating Heritability – Quantitative Traits

- Estimating heritability from **genome-wide significant variants**:

$$h^2_{GWAS} = \sum_{k \in S} \hat{u}^2_k$$

- Problem: many variants associated with the trait are excluded because they have not reached statistical significance in the sample due to low effect sizes or low MAF ⇒ severely underestimates h^2.
- Estimating the aggregate impact of all variants:
  - We are interested in $\sum u_k^2$ rather than the individual effect sizes u_k ("nuisance parameters").
  - Treat u_k as random variables, equally contribute to the total heritability: $var(u_k) \equiv \sigma_u^2 = \frac{h^2}{m}$.
  - Under additive model, we have the relationship:

$$\begin{cases} corr(p_i, p_j) = E(p_i p_j) = h^2 G_{ij} \\ G_{ij} = corr(g_i, g_j) = \frac{1}{m} \sum_{k=1}^{m} g_{ik} g_{jk} \end{cases}$$

PCGC: Phenotype Correlation - Genotype Correlation

$$\Leftrightarrow \qquad \boxed{corr(p_i, p_j) = h^2 \; corr(g_i, g_j)}$$

Estimating Heritability – Quantitative Traits

Additive model:

$$p_i = \sum_{k=1}^{m} g_{ik} u_k + e_i$$

Under independence assumption and suppose p_i and g_{ik} have been normalized to have mean 0 and variance 1, then:

$$\begin{aligned}
corr(p_i, p_j) = \mathbb{E}[p_i p_j] &= \mathbb{E}\left[(\sum_{k=1}^{m} g_{ik} u_k + e_i)(\sum_{k=1}^{m} g_{jk} u_k + e_j) \right] \\
&= \mathbb{E}\left[\sum_{k=1}^{m} g_{ik} g_{jk} u_k^2 \right] + \mathbb{E}\left[\sum_{k \neq k'} g_{ik} g_{jk'} u_k u_{k'} \right] + 2\mathbb{E}\left[e_i \sum_{k=1}^{m} g_{ik} u_k \right] + \mathbb{E}\left[e_i e_j \right] \\
&= \mathbb{E}\left[\sum_{k=1}^{m} g_{ik} g_{jk} u_k^2 \right] \\
&= \sum_{k=1}^{m} g_{ik} g_{jk} \mathbb{E}[u_k^2] \\
&= \frac{1}{m} \sum_{k=1}^{m} g_{ik} g_{jk} \cdot m\mathbb{E}[u_k^2] \\
&= h^2 G_{ij} = h^2 corr(g_i, g_j)
\end{aligned}$$

Estimating Heritability – Quantitative Traits

- Improving heritability estimates with REML:
  - PCGC regression estimator is a moments-based estimator, which looks only at pairs of individuals at a time.
  - In contrast, maximum likelihood estimator can extract more information by looking at the entire cohort simultaneously.
  - Maximum-likelihood estimation relies on an explicit probabilistic model:

$$p_i = g_i + e_i = \sum_{k=1}^{m} g_{ik} u_k + e_i$$

$$u_k \sim N(0, h^2/m), \qquad e_i \sim N(0, 1 - h^2)$$

Joint distribution: $\boldsymbol{p} \sim MVN(0, Gh^2 + I(1 - h^2))$ G is the genetic relationship matrix

  - This model is a special case of random effects model.
  - The problem of estimating heritability can be viewed as estimating variance components.

Estimating Heritability – Quantitative Traits

- This random effects model approach enables us to estimate the aggregate contribution of common SNPs with only thousands of samples.
- A big part of the gap between h^2_{GWAS} and h^2 has been filled, thanks to the powerful statistical tool.

Common SNPs explain a large proportion of the heritability for human height

Abstract

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

Yang et al. 2010 Nature Genetics

Estimating Heritability – Disease Traits

- Estimating heritability for quantitative traits may be considered largely solved.
- The primary focus of medical genetics is disease traits – which are binary (0/1) rather than quantitative.
- Binary variables are not continuous while people are more willing to work with continuous variables.
- Consider introducing an underlying quantitative trait to connect with the observed disease trait.
- Deal with the disease trait through this quantitative trait so that statistical methods developed for quantitative traits can be applied.

Estimating Heritability – Disease Traits

- **Liability**: an underlying unobserved continuous trait.
- e.g Type 2 diabetes: liability l is fasting blood glucose (空腹血糖).
- **Liability threshold model**:

$$y_i = \mathbb{I}\{l_i > t\}$$

$$l_i = g_i + e_i$$

$$g_i \sim N\left(0, h_l^2\right), \quad e_i \sim N(0, 1 - h_l^2)$$

- The value of t determines the prevalence of the disease. e.g $t = 1.5 \Leftrightarrow$ prevalence $\approx 7\%$.
- **Two types of heritability**:

	On liability scale	**On observed scale**
Model	$l = g_{liab} + e$	$y = g_{obs} + \epsilon$
Heritability	$h_l^2 = \frac{\sigma_{g_{liab}}^2}{\sigma_l^2}$	$h_o^2 = \frac{\sigma_{g_{obs}}^2}{\sigma_y^2}$

Estimating Heritability – Disease Traits

- We are more interested in estimating h_l^2:
  - h_l^2 are not affected by disease prevalence, so that they can be compared across diseases or across populations.
- Estimating h_l^2 using maximum-likelihood approach? Challenging.

$$LL = \log \mathrm{P}(\boldsymbol{y}; h_l^2) = \log \int \varphi(\boldsymbol{l}; h_l^2)\, d\boldsymbol{l}$$

$$\varphi = MVN(0, Gh_l^2 + (1 - h_l^2)I_n)$$

- Viewing disease trait as quantitative trait to first estimate h_o^2 then transfer to h_l^2? Probably.
  - Relationship between h_l^2 and h_o^2 if samples are randomly selected:

$$h_l^2 = \frac{K(1-K)}{\varphi(t)^2} h_o^2$$

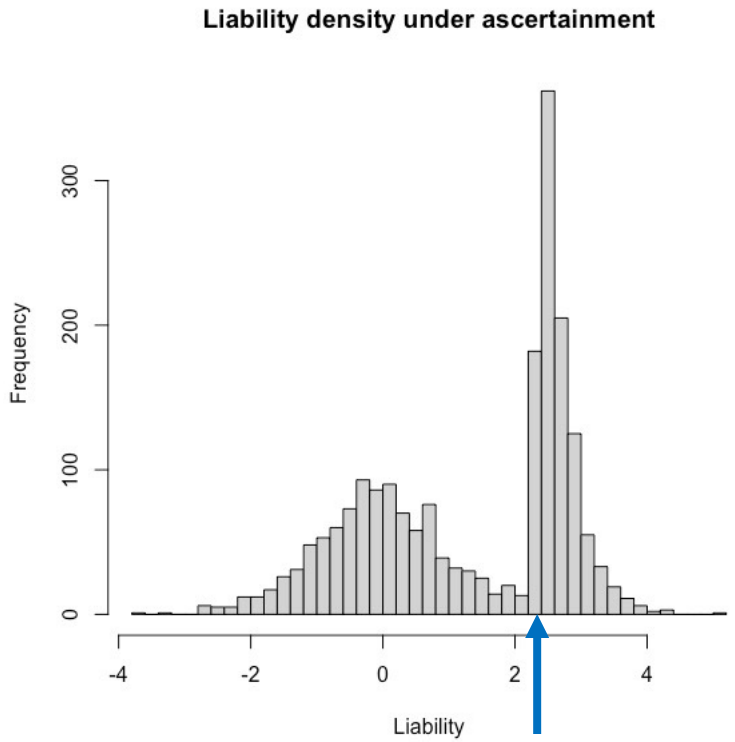
Estimating Heritability – Disease Traits

- In real disease studies, cases are usually oversampled – **ascertainment bias/case-control sampling**.

```
### liability plot
# paramters
h2l = 0.5
GE = rmvnorm(1000000, c(0, 0), diag(c(h2l, 1-h2l)))
g = GE[,1]
e = GE[,2]
l = g + e
# parameters
K = 0.01
t = qnorm(1 - K)
P = 0.5
n = 2000 # total sample size in the study
n_case = n*P # number of cases in the study
n_control = n*(1 - P) # number of controls in the study
# case-control sampling
idx_case_popu = which(l > t)
idx_control_popu = which(l < t)
idx_case_study = sample(idx_case_popu, n_case)
idx_control_study = sample(idx_control_popu, n_control)
# liability distribution in the study
l_case_study = l[idx_case_study]
l_control_study = l[idx_control_study]
l_study = c(l_case_study, l_control_study)
hist(main='Liability density under ascertainment',
     x=l_study, breaks=50, xlab='Liability')
```

Cases are oversampled

→ Nonnormal liability



Threshold: t=2.33

Estimating Heritability – Disease Traits

- This is one notable consequence cased by ascertainment bias: marginal distributions of genetic effect (g_{liab}), environmental effect (e), as well as liability (l) are no longer normal.
- Lee et al. (2011) adjust the transformation formula to account for the induced nonnormality:

$$\hat{h}_l^2 = \frac{K^2(1-K)^2}{P(1-P)\varphi(t)^2}\hat{h}_o^2$$

- They still follow the idea of transforming the $\hat{h}_o^2$ estimated by REML to get $\hat{h}_l^2$.
- It turns out that this approach (REML) suffers severe underestimation of h_l^2 under ascertainment bias.
- Reason: there is another subtle consequence caused by ascertainment bias is ignored.

Estimating Heritability – Disease Traits

- **“Induced” G×E interactions**:
  - Although there is no G×E interactions in the population, there is an obvious interaction between g_{liab} and e under case-control sampling.
  - Not only is normality assumption violated, but the independence assumption ($g_{liab} \perp e$) does not hold any more.
  - There is a positive correlation between genetic and environmental effects.

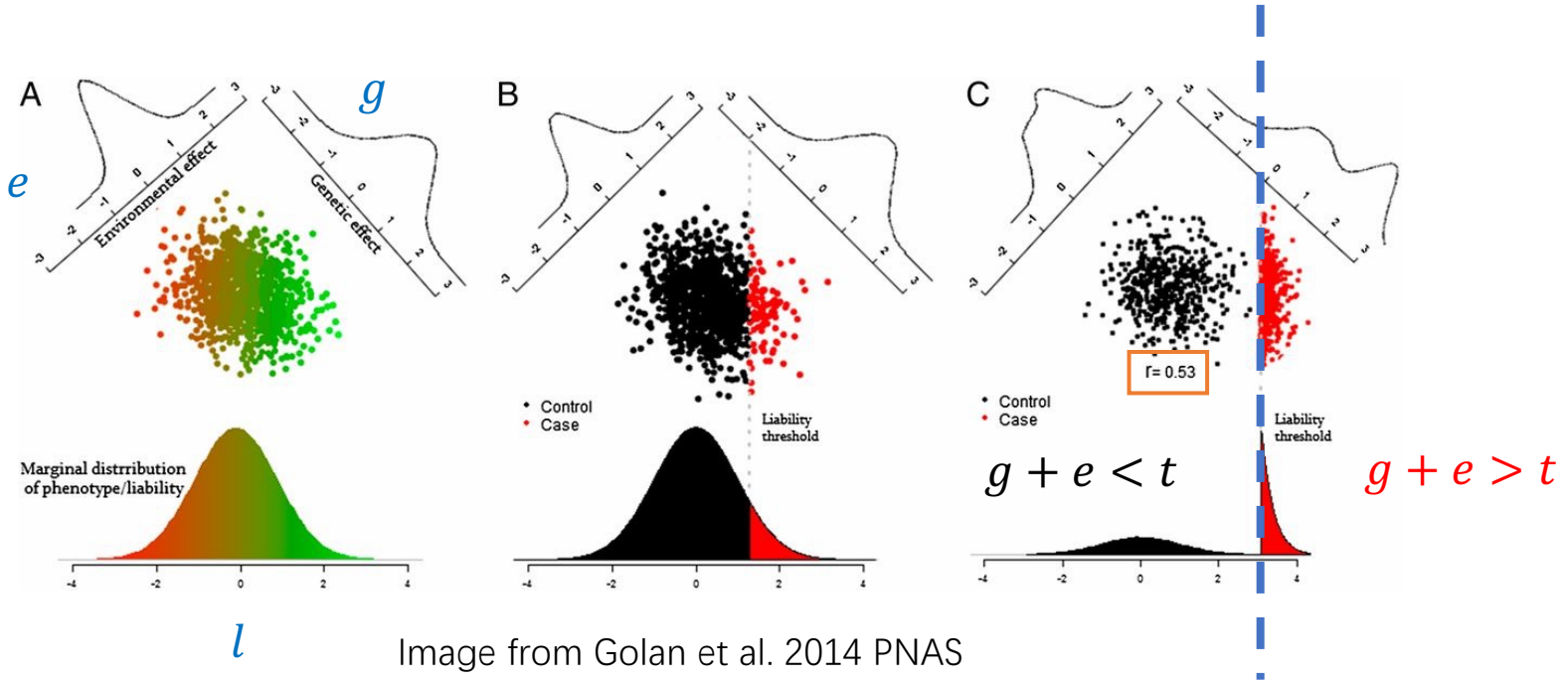
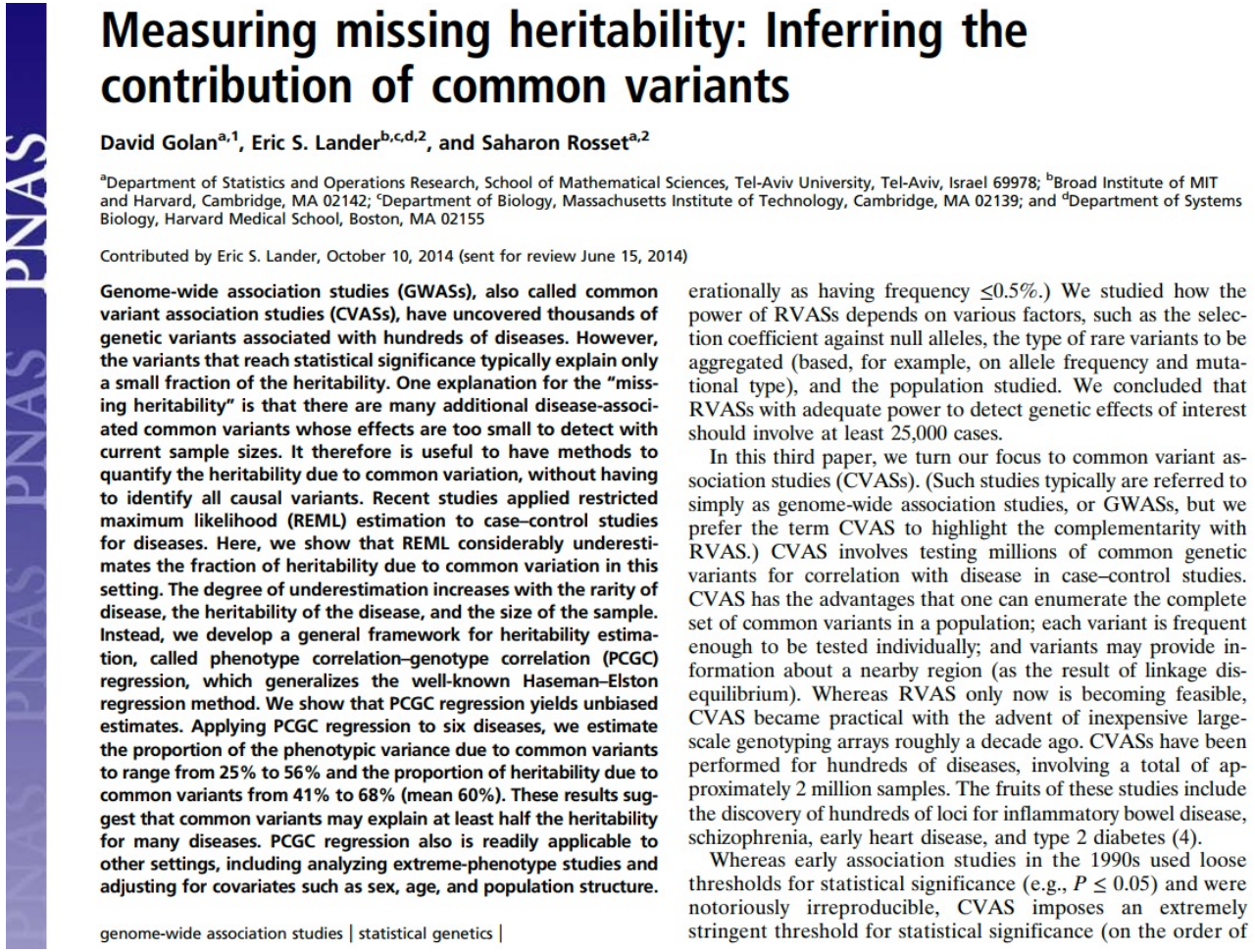


Image from Golan et al. 2014 PNAS

Estimating Heritability – Disease Traits

- Is there some other way to estimate h_l^2 while avoiding these consequences caused by ascertainment bias ?
- One approach: PCGC regression.



PNAS PNAS PNAS

Measuring missing heritability: Inferring the contribution of common variants

David Golan[a,1], Eric S. Lander[b,c,d,2], and Saharon Rosset[a,2]

[a]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel 69978; [b]Broad Institute of MIT and Harvard, Cambridge, MA 02142; [c]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [d]Department of Systems Biology, Harvard Medical School, Boston, MA 02155

Contributed by Eric S. Lander, October 10, 2014 (sent for review June 15, 2014)

Genome-wide association studies (GWASs), also called common variant association studies (CVASs), have uncovered thousands of genetic variants associated with hundreds of diseases. However, the variants that reach statistical significance typically explain only a small fraction of the heritability. One explanation for the "missing heritability" is that there are many additional disease-associated common variants whose effects are too small to detect with current sample sizes. It therefore is useful to have methods to quantify the heritability due to common variation, without having to identify all causal variants. Recent studies applied restricted maximum likelihood (REML) estimation to case–control studies for diseases. Here, we show that REML considerably underestimates the fraction of heritability due to common variation in this setting. The degree of underestimation increases with the rarity of disease, the heritability of the disease, and the size of the sample. Instead, we develop a general framework for heritability estimation, called phenotype correlation–genotype correlation (PCGC) regression, which generalizes the well-known Haseman–Elston regression method. We show that PCGC regression yields unbiased estimates. Applying PCGC regression to six diseases, we estimate the proportion of the phenotypic variance due to common variants to range from 25% to 56% and the proportion of heritability due to common variants from 41% to 68% (mean 60%). These results suggest that common variants may explain at least half the heritability for many diseases. PCGC regression also is readily applicable to other settings, including analyzing extreme-phenotype studies and adjusting for covariates such as sex, age, and population structure.

genome-wide association studies | statistical genetics |

erationally as having frequency ≤0.5%.) We studied how the power of RVASs depends on various factors, such as the selection coefficient against null alleles, the type of rare variants to be aggregated (based, for example, on allele frequency and mutational type), and the population studied. We concluded that RVASs with adequate power to detect genetic effects of interest should involve at least 25,000 cases.

In this third paper, we turn our focus to common variant association studies (CVASs). (Such studies typically are referred to simply as genome-wide association studies, or GWASs, but we prefer the term CVAS to highlight the complementarity with RVAS.) CVAS involves testing millions of common genetic variants for correlation with disease in case–control studies. CVAS has the advantages that one can enumerate the complete set of common variants in a population; each variant is frequent enough to be tested individually; and variants may provide information about a nearby region (as the result of linkage disequilibrium). Whereas RVAS only now is becoming feasible, CVAS became practical with the advent of inexpensive large-scale genotyping arrays roughly a decade ago. CVASs have been performed for hundreds of diseases, involving a total of approximately 2 million samples. The fruits of these studies include the discovery of hundreds of loci for inflammatory bowel disease, schizophrenia, early heart disease, and type 2 diabetes (4).

Whereas early association studies in the 1990s used loose thresholds for statistical significance (e.g., $P \leq 0.05$) and were notoriously irreproducible, CVAS imposes an extremely stringent threshold for statistical significance (on the order of

Estimating Heritability – Disease Traits

- Idea of PCGC regression: the heritability of a trait controls the strength of the relationship between genotype correlation and phenotype correlation.
- Mathematical expression:

$$\mathbb{E}(p_i p_j) = f(h^2, G_{ij})$$

- A special case: $f(h^2, G_{ij}) = h^2 G_{ij}$ for additive quantitative traits without ascertainment bias.
- In general, f could be a complicated function. So, we consider approximating f by a Taylor series at $G_{ij} = 0$.
- Under the assumption that *g* and *e* are normally distributed,

$$f(h^2, G_{ij}) = ch^2 G_{ij} + o(G_{ij}), \qquad c = \frac{P(1-P)\varphi(t)^2}{K^2(1-K)^2} \qquad \Longrightarrow \qquad \boxed{\mathbb{E}(p_i p_j) \approx ch^2 G_{ij}}$$

- Regressing phenotypic correlation ($p_i p_j$) onto genetic correlation (G_{ij}) and adjusting for the constant c yield an unbiased estimator of h_l^2.

```r
# pcgc function
pcgc_reg = function(y, X, K, P){

  t = qnorm(1 - K)
  n = nrow(X)
  # phenotype correlations
  pheno_corr = outer(y, y) # phenotype correlation matrix
  # genotype correlations
  geno_corr = X%*%t(X)/p # genetic correlation matrix
  # multiply genotype correlations with the constant c
  c = P*(1 - P)*dnorm(t)^2/(K^2*(1 - K)^2)
  geno_corr_c = c*geno_corr
  # vectorization
  pheno_corr_vec = pheno_corr[upper.tri(pheno_corr)]
  geno_corr_c_vec = geno_corr_c[upper.tri(geno_corr_c)]

  # regress phenotype correlations on genotype correlations
  data = data.frame(y=pheno_corr_vec, X=geno_corr_c_vec)
  fit = lm(y~X, data=data)
  h2l = summary(fit)$coefficients[2,1] # liability scale heritability

  return(h2l)

}

# parameters
K = 0.1
P = 0.3
t = qnorm(1 - K)
h2l = 0.5
p = 10000
# sample size
n = 4000
n_case = n*P
n_control = n*(1 - P)
mafs = runif(p, 0.05, 0.5)
```

Calculate phenotype correlations & genotype correlations

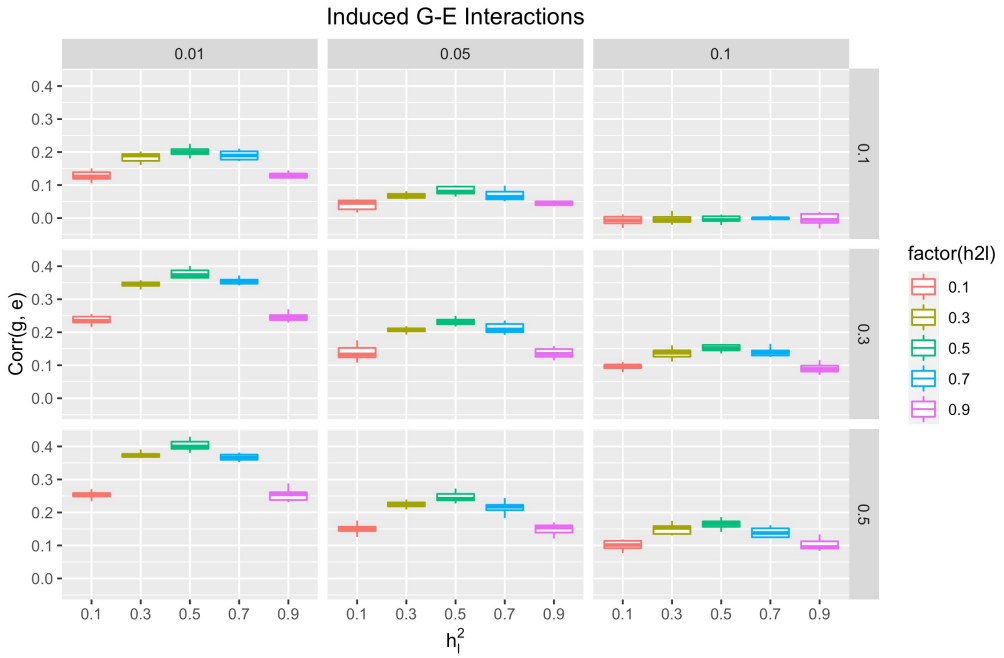
Simple linear regression

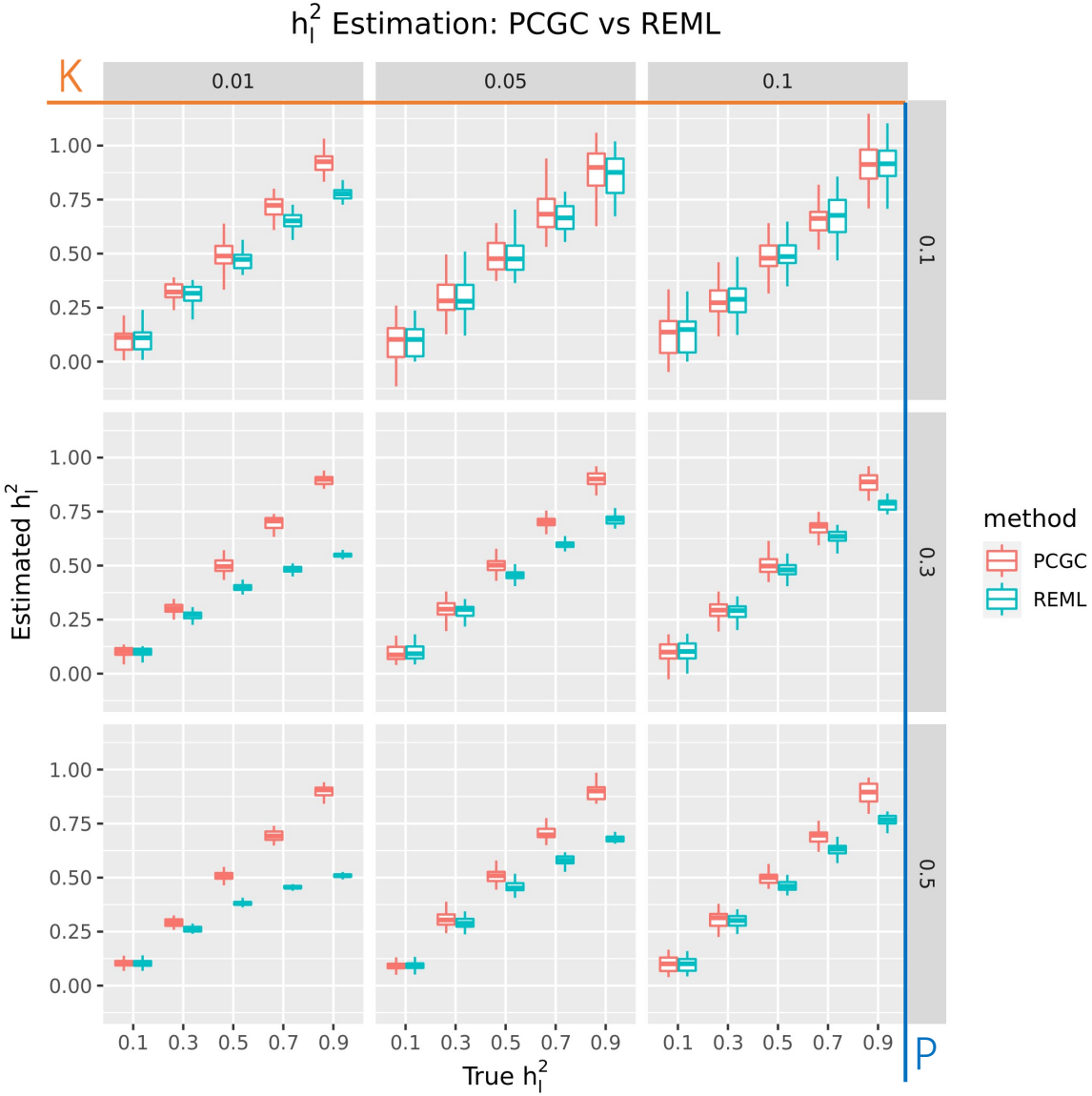
```r
## generate data
beta = rnorm(p, 0, sqrt(h2l/p)) # SNP effect sizes
n1 = n0 = 0
X1 = matrix(NA, n_case, p)
X0 = matrix(NA, n_control, p)
l1 = l0 = c()
while(n1 < n_case | n0 < n_control){
  # generate a genotype
  g = rbinom(p, 2, mafs)
  # standarized the genotype
  x = (g - 2*mafs)/sqrt(2*mafs*(1 - mafs))
  # liability for this individual
  l = sum(x*beta) + rnorm(1, 0, sqrt(1-h2l))
  # phenotype
  if(l > t & n1 < n_case){
    n1 = n1 + 1
    X1[n1,] = x
    l1[n1] = l
  }else{
    if(n0 < n_control){
      n0 = n0 + 1
      X0[n0,] = x
      l0[n0] = l
    }
  }
}
X = rbind(X1, X0)
l = c(l1, l0)
y = c(rep(1, n_case), rep(0, n_control))
y_std = (y - P)/sqrt(P*(1 - P))

## estimate heritability using pcgc regression
fit_pcgc = pcgc_reg(y=y_std, X=X, K=K, P=P)
```

Case-control sampling

Simulations

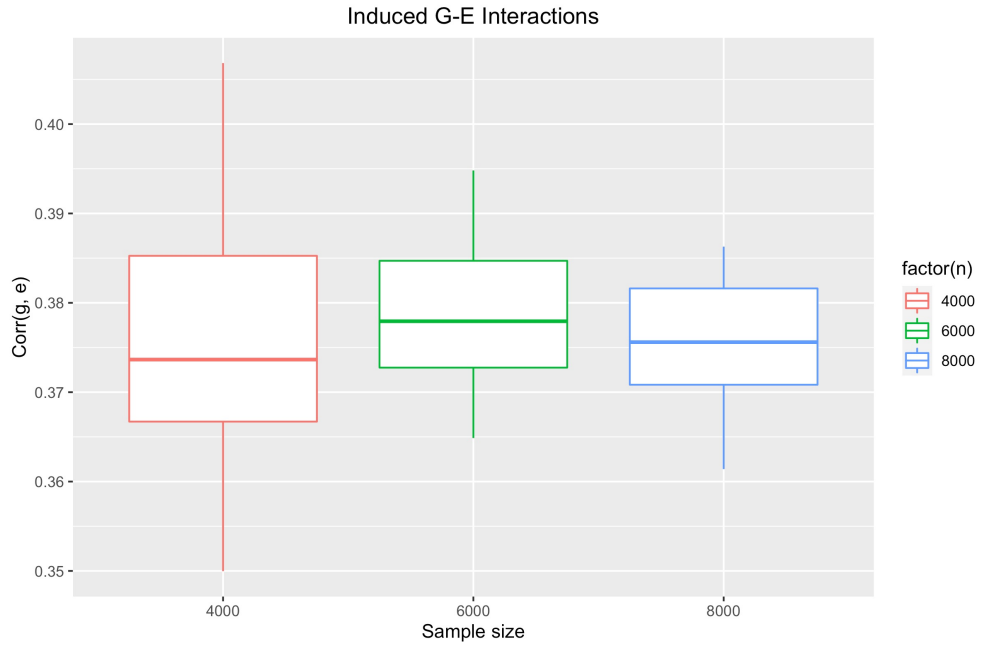
- PCGC vs REML under different scenarios.
- $n = 4000,\ p = 10000.$
- $K = \{0.01, 0.05, 0.1\},\ P = \{0.1, 0.3, 0.5\},\ h_l^2 = \{0.1, 0.3, 0.5, 0.7, 0.9\}.$
- PCGC estimates are unbiased under all settings.
- For a certain value of h_l^2, the bias of REML estimate increases with the increasing ascertainment bias.
- In each grid, the bias of REML estimate increases with the true h_l^2.

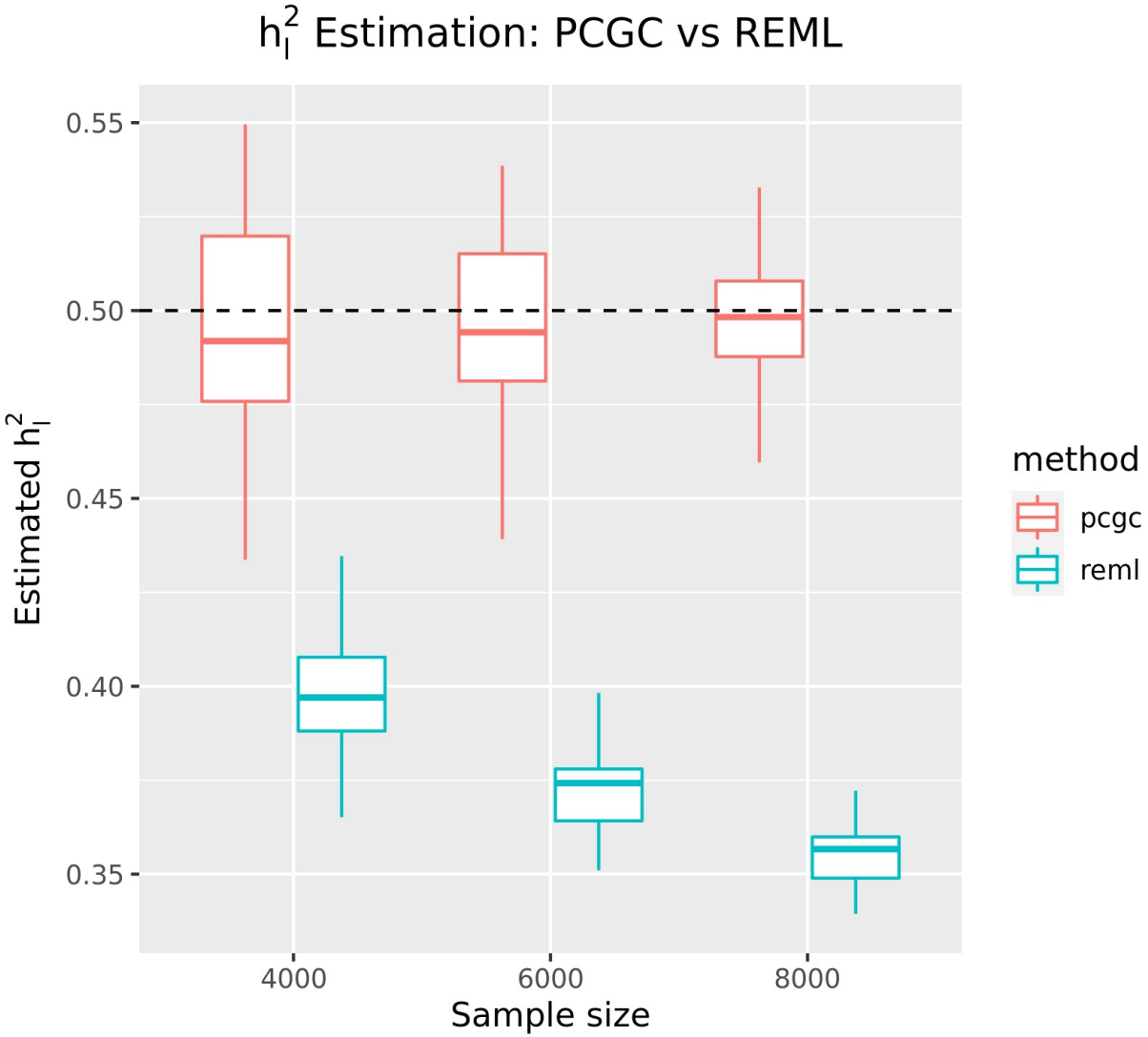




Simulations

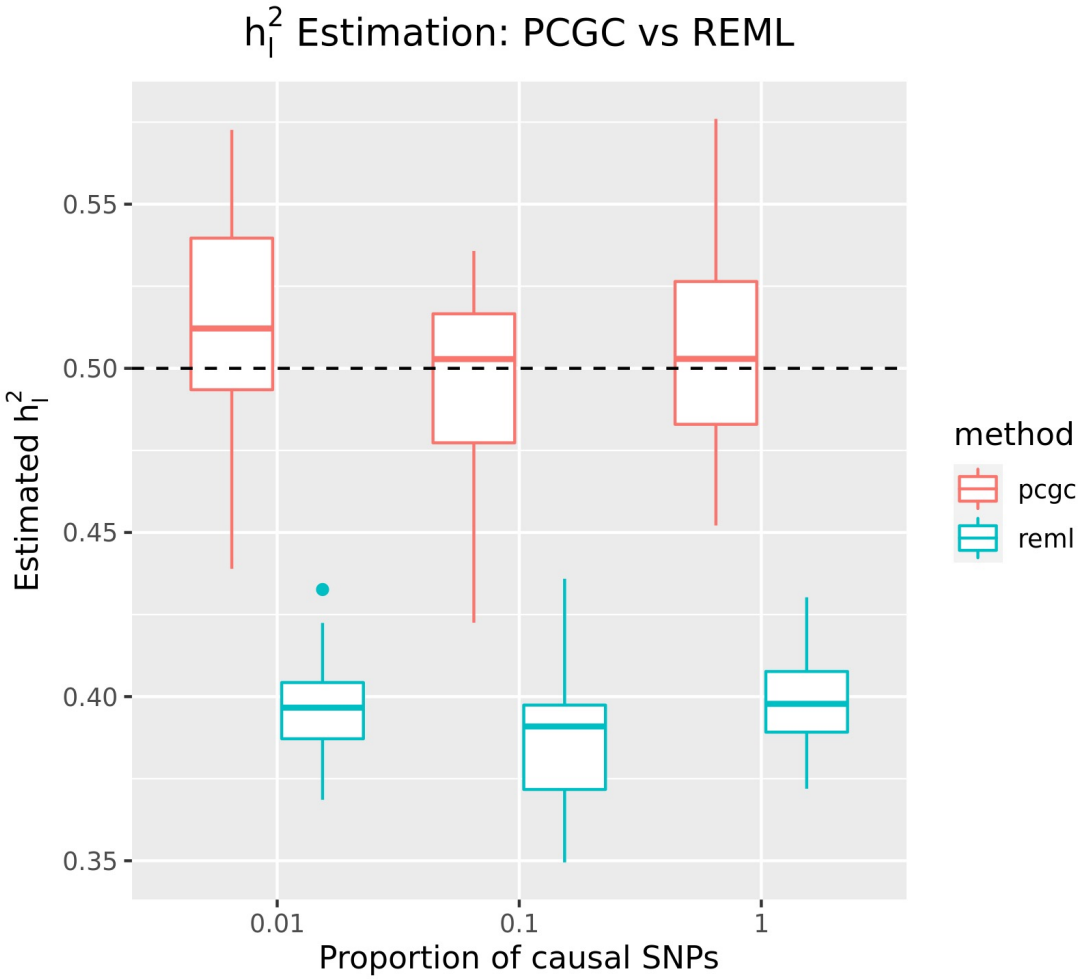
- PCGC vs REML with different sample sizes.
- $K = 0.01,\ P = 0.3,\ h_l^2 = 0.5,\ p = 10000.$
- $n = \{4000, 6000, 8000\}.$
- PCGC estimates are unbiased with different sample sizes.
- The bias of REML increases with increasing sample size.





Simulations

- PCGC vs REML with different levels of polygenicity.
- $K = 0.01,\ P = 0.3,\ h_l^2 = 0.5,\ n = 4000,\ p = 10000.$
- Proportion of causal SNPs: $\{1\%, 10\%, 100\%\}$.
- PCGC estimates are unbiased under different degrees of polygenicity.



Estimating Heritability of Disease

- Apply PCGC regression and REML to case-control studies of disease.
- WTCCC studies of seven diseases: Bipolar disorder (BD), Crohn's disease (CD), Cardiovascular diseases (CAD), Hypertension (HT), Type 1 diabetes (T1D), Type 2 diabetes (T2D), Rheumatoid arthritis (RA).

Phenotype	Prevalence	REML	PCGC	Total heritability
Bipolar disorder	0.005	0.34	0.44	0.71
Crohn's disease	0.001	0.22	0.25	0.5-0.6
CAD	0.05	0.45	0.46	0.5-0.6
Hypertension	0.06	0.45	0.51	0.3-0.6
T1D (including MHC)	0.005	0.43	0.59	0.72-0.88
T2D	0.06	0.44	0.47	0.72
RA	0.005	0.29	0.33	0.6

Derivation of PCGC Regression

- Preliminary:
  - Liability threshold model: $l_i = g_i + e_i$
  - Assumptions: $g_i \perp e_i,\ g_i \sim N(0, \sigma_g^2),\ e_i \sim N(0, 1-\sigma_g^2)$
  - Other notations:

    $y_i = \mathbb{I}\{l_i > t\}$: phenotype $s_i = \mathbb{I}\{\text{individual } i \text{ is selected in the study}\}$

    $K = P(y_i = 1)$: population prevalence $t = \Phi^{-1}(1-K)$: threshold

    $P = P(y_i = 1 | s_i = 1)$: in-sample prevalence

    $P_{case} = P(s_i = 1 | y_i = 1)$: probability that a case is selected for the study

    $P_{control} = P(s_i = 1 | y_i = 0)$: probability that a control is selected for the study
  - Relationship among $K, P, P_{case}, P_{control}$:

$$\frac{KP_{case}}{(1-K)P_{control}} = \frac{P}{1-P}$$

$$\frac{P(y_i = 1)P(s_i = 1 | y_i = 1)}{P(y_i = 0)P(s_i = 1 | y_i = 0)} = \frac{P(y_i = 1 | s_i = 1)}{P(y_i = 0 | s_i = 1)}$$

  - Assume $P_{case} = 1$ (full ascertainment), then $P_{control} = \frac{K(1-P)}{P(1-K)}$.

Derivation of PCGC Regression

- Purpose: find the relationship between phenotype correlation and genotype correlation.
- Consider a pair of individuals (i, j) in the study:
  - Product of standardized phenotypes:

$$Z_{ij} = \frac{(y_i - P)(y_j - P)}{P(1-P)}$$

  - Genotype correlation: ρ
  - Indicator $\mathcal{S} = \mathbb{I}\{s_i = s_j = 1\}$
- Start from calculating $\mathbb{E}[Z_{ij} | \mathcal{S} = 1; \rho]$ (phenotype correlation in the study)

$$\mathbb{E}\left[Z_{ij} \middle| S = 1; \rho\right] = \frac{1-P}{P} \mathrm{P}\left(y_i = y_j = 1 \middle| S = 1; \rho\right) - \mathrm{P}\left(y_i \neq y_j \middle| S = 1; \rho\right) + \frac{P}{1-P} \mathrm{P}\left(y_i = y_j = 0 \middle| S = 1; \rho\right)$$

How to deal with conditional probability?

Derivation of PCGC Regression

$$\mathbb{P}(y_i = y_j = 1; \rho, \sigma_g^2) = \int_t^\infty \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2$$

- **Apply Bayes' rule** to eliminate the impact of ascertainment so that we can utilize the independence and normality assumptions on liability:

Unconditional probability

Due to case-control sampling

$$\mathrm{P}(y_i = y_j = 1 | \mathcal{S} = 1; \rho) = \frac{\mathrm{P}(\mathcal{S} = 1 | y_i = y_j = 1; \rho)\mathrm{P}(y_i = y_j = 1; \rho)}{\mathrm{P}(\mathcal{S} = 1; \rho)}$$

- Under the full ascertainment assumption, $\mathrm{P}(\mathcal{S} = 1 | y_i = y_j = 1; \rho) = 1$.
- For $y_i = y_j = 0$ and $y_i \neq y_j$, use $P_{control} = \frac{K(1-P)}{P(1-K)}$ to get:

$$\mathrm{P}(\mathcal{S} = 1 | y_i = y_j = 0; \rho) = \left(\frac{K(1-P)}{P(1-K)}\right)^2, \qquad \mathrm{P}(\mathcal{S} = 1 | y_i \neq y_j; \rho) = \frac{K(1-P)}{P(1-K)}$$

$$\Rightarrow P(y_i = y_j = 0 | \mathcal{S} = 1; \rho) = \left(\frac{K(1-P)}{P(1-K)}\right)^2 \frac{\mathrm{P}(y_i = y_j = 1; \rho)}{\mathrm{P}(\mathcal{S} = 1; \rho)}, \qquad P(y_i \neq y_j \,|\mathcal{S} = 1; \rho) = \frac{K(1-P)}{P(1-K)} \frac{\mathrm{P}(y_i \neq y_j; \rho)}{\mathrm{P}(\mathcal{S} = 1; \rho)}$$

Derivation of PCGC Regression

$$\mathbb{P}(y_i = y_j = 1; \rho, \sigma_g^2) = \int_t^\infty \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2$$

Integration of bivariate normal

- Using these results, we get:

$$\mathbb{E}[Z_{ij}|\mathcal{S} = 1; \rho] = \frac{\frac{1-P}{P}\mathbb{P}(y_i = y_j = 1; \rho) - \frac{K(1-P)}{P(1-K)}\mathbb{P}(y_i \neq y_j; \rho) + \frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^2 \mathbb{P}(y_i = y_j = 0; \rho)}{\mathbb{P}(\mathcal{S} = 1; \rho)} \equiv \frac{A(\rho)}{B(\rho)}$$

- Although we have expressed phenotype correlation as a function of genotype correlation ρ, this function takes a complex form, difficult for estimating $\sigma_g^2 = h_l^2$ (target parameter).
- Consider approximating the complicated function by using a Taylor series around $\rho = 0$:

$$\mathbb{E}[Z_{ij}|\mathcal{S} = 1; \rho] = \frac{A(0)}{B(0)} + \frac{A'(0)B(0) - A(0)B'(0)}{B(0)^2}\rho + \mathcal{O}(\rho^2) \Rightarrow \mathbb{E}[Z_{ij}|\mathcal{S} = 1; \rho] = \frac{A'(0)}{B(0)}\rho + \mathcal{O}(\rho^2)$$

- The remaining task is to find $A'(0)$.

$A(0) = 0, B(0) = \frac{K^2}{P^2}$

Derivation of PCGC Regression

- The three unconditional probabilities in $A(\rho)$:

$$\mathbb{P}(y_i = y_j = 1; \rho, \sigma_g^2) = \int_t^\infty \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 \qquad \mathbb{P}(y_i \neq y_j; \rho, \sigma_g^2) = 2\int_{-\infty}^t \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 \qquad \mathbb{P}(y_i = y_j = 0; \rho, \sigma_g^2) = \int_{-\infty}^t \int_{-\infty}^t f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2$$

- ρ appears in the covariance matrix of bivariate normal density f_{ρ,σ_g^2} :

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \sigma_g^2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (1 - \sigma_g^2) = \begin{pmatrix} 1 & \rho\sigma_g^2 \\ \rho\sigma_g^2 & 1 \end{pmatrix}$$

- Note that both ρ and σ_g^2 only appear as the product $\rho\sigma_g^2$.
- Hence, differentiating $A(\rho)$ w.r.t. ρ results in an expression of the form $f\left(\rho\sigma_g^2\right)\sigma_g^2$ (f is some function).
- Setting $\rho = 0$ yields $f(0)\sigma_g^2$, so the first-order approximation is of the form $f(0)\sigma_g^2\rho$.
- Straightforward calculation shows that:

$$\frac{d}{d\rho}\int_t^\infty \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 = \sigma_g^2 \varphi(t)^2 \qquad \frac{d}{d\rho}\int_{-\infty}^t \int_t^\infty f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 = -\sigma_g^2 \varphi(t)^2 \qquad \frac{d}{d\rho}\int_{-\infty}^t \int_{-\infty}^t f_{\rho,\sigma_g^2}(l_1, l_2) dl_1 dl_2 = \sigma_g^2 \varphi(t)^2$$

Derivation of PCGC Regression

- Using these results to get $A'(0)$:

$$A'(0) = \frac{P(1-P)}{K^2(1-K)^2}\sigma_g^2\varphi(t)^2$$

- Finally, we obtain the approximated relationship between $\mathbb{E}\left[Z_{ij}\middle|\mathcal{S}=1;\rho\right]$ and ρ:

$$\mathbb{E}\left[Z_{ij}\middle|\mathcal{S}=1;\rho\right] \approx \frac{P(1-P)\varphi(t)^2}{K^2(1-K)^2}\sigma_g^2\rho \equiv c\sigma_g^2\rho$$

- Hence, when the error term of this approximation is small, the slope obtained by regression Z_{ij} on G_{ij} (realization of genotype correlation ρ) is an unbiased estimator of $c\sigma_g^2$. Dividing by the constant c yields an unbiased estimator of $\sigma_g^2 = h_l^2$.

Discussion

- Different ideas on accounting for ascertainment bias:
  - Lee et al. (2011): treat 0/1 phenotype as quantitative phenotype and apply REML to get $\hat{h}_o^2$, then transform to $\hat{h}_l^2$ via some formula correcting for the ascertainment bias (**ex post correction**).
  - The underlying reason for the underestimation of h_l^2 using REML: under case-control sampling, the observed scale additive model $\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{e},\ \boldsymbol{y} \sim N(0, A\sigma_u^2 + \sigma_e^2 I)$ is further invalid due to the "induced" G×E interactions; while REML still uses this model to estimate h_o^2, yielding biased estimate $\hat{h}_o^2$ thus biased $\hat{h}_l^2$.
  - In contrast, **PCGC regression directly estimates $\boldsymbol{h_l^2}$**, and accounting for the ascertainment bias issue is intrinsically included in the derivation of phenotype correlation – genotype correlation. In this way, PCGC regression also avoids the impact of "induced" G×E interactions.

Discussion

- Maximum-likelihood method for estimating disease trait heritability under case-control sampling:

$$\text{Probit model: } P(y_i = 1|X_i, g_i, \boldsymbol{\beta}) = \Phi(X_i^T\boldsymbol{\beta} + g_i),\ \boldsymbol{g} \sim \text{MVN}(0, \theta \boldsymbol{Z}\boldsymbol{Z}^T)$$

$$\text{Likelihood: } L(\boldsymbol{\beta}, \theta) = P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \theta, \boldsymbol{\beta}) = \int P(\boldsymbol{g}|\boldsymbol{Z}, \theta) \prod_i P(y_i|X_i, g_i, \boldsymbol{\beta}) d\boldsymbol{g}$$

Approximation approach 2

Approximation approach 1

$$\text{Ascertained likelihood: } L^*(\boldsymbol{\beta}, \theta) = P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{s} = 1, \theta, \boldsymbol{\beta}) = \frac{P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \theta, \boldsymbol{\beta})}{P(\boldsymbol{s} = 1|\boldsymbol{X}, \boldsymbol{Z}, \theta, \boldsymbol{\beta})} \prod_i P(s_i = 1|y_i)$$

Approximation approaches without ascertainment bias (more involved under ascertainment bias):

$$1.\ \text{Pairwise likelihood: } P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \theta, \boldsymbol{\beta}) \approx \prod_{i \neq j} P(y_i, y_j|X_i, X_j, Z_i, Z_j, \theta, \boldsymbol{\beta})$$

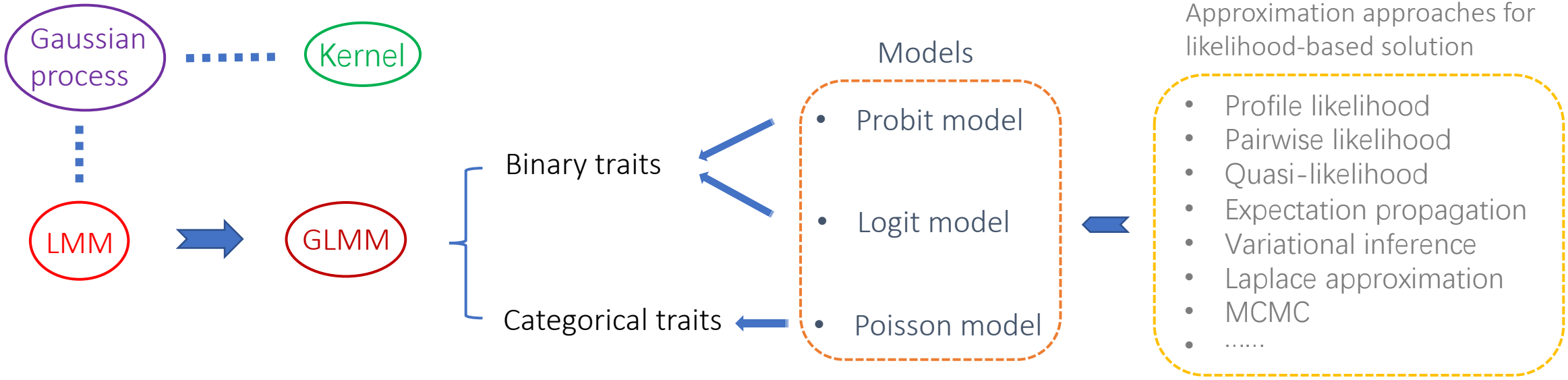
$$2.\ \text{Expectation propagation: } P(y_i|X_i, g_i, \boldsymbol{\beta}) \approx t_i(g_i) \triangleq r_i N(g_i; \tilde{\alpha}_i, \tilde{\gamma}_i)$$

Maximum Likelihood for Gaussian Process Classification and Generalized Linear Mixed Models under Case-Control Sampling

Omer Weissbrod OWEISSBROD@HSPH.HARVARD.EDU
Epidemiology Department
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

Discussion

- Summary
  - **Missing heritability**: deal with problems caused by low signal-to-noise ratio using powerful statistical tools.
  - **Estimating heritability of disease traits**: link disease traits with latent quantitative traits using liability threshold model.
  - **Ascertainment bias**: use conditional probability to account for case-control sampling, and transfer to unconditional probability to detach the effect caused by ascertainment bias.
- Connecting dots



Thank you!