

Evaluation of Power and Risk Prediction Utility of Future Genome-wide Association Studies

Ju-Hyun Park

Biostatistics Branch, Division of Cancer Epidemiology and Genetics,
National Cancer Institute, USA

Joint work with Nilanjan Chatterjee

7/30/2009

IMS New Researchers Conference

Background

- Genome-wide association study (GWAS) examines hundreds of thousand of markers across the whole genome
 - Multiple testing requires a stringent criteria, usually 10^{-7} , and hence a sufficiently large sample size is needed to obtain power to detect
- GWAS has become a common tool for identifying genetic variants that predispose to various complex quantitative and qualitative traits
- The success of the first generation study and the ever decreasing cost of genotyping has generated tremendous enthusiasm for continuing the effort
- Consortium of existing GWAS to increase sample size and hence power

Challenges for next GWAS

- Efficient study designs of next GWAS - sample size and power calculation
- Only a small fraction of genetic variance has been explained
 - Many more to be found
- The loci which remain undetected are likely to have smaller effects
 - Larger studies are needed
- Standard power calculations are not appropriate
 - Probability of detecting a susceptibility loci with fixed effect size
 - How many such loci?

The Basic Idea

- Studies have reported susceptibility SNPs in a range of effect sizes with low power
 - There must be more SNPs to be found in those range

- For a given effect size,

$$\# \text{ of loci discovered} = \# \text{ of underlying susceptibility loci} \times \text{power}$$

- For example, if a study detected 2 loci, each with 25% power, one could expect there are all together $2/0.25 = 8$ loci with similar effect size

Definition of Effect Size for SNPs

- A single measure that captures both regression effect (β) and allele frequency (f)

$$\text{effect size } (es) = \beta^2 \times 2f(1 - f)$$

- es is the contribution of the SNP to the genetic variance of a trait under an additive model with Hardy-Weinberg Equilibrium
- es is also proportional to the non-centrality parameter for the one d.f. chi-square test for trend for association
 - For linear regression, the proportionality constant is $n/\text{var}(y)$

Estimation of Total Number of Underlying SNPs

- From the established relationship,

$$\# \text{ of underlying susceptibility loci } (N) = \frac{\# \text{ of loci discovered}}{\text{power}}$$

- For a given effect size, its power is a point-wise estimate, which can be unstable
 - Bin-approach? How many bins?
 - Borrow information by smoothing?

Notation

- Data available
 - Let es_1, es_2, \dots, es_M be M distinct effect sizes observed and pow_1, \dots, pow_M be corresponding powers of detecting the effect sizes
- Parameters to be estimated
 - Let N_i be the total number of underlying SNPs at es_i (including both seen and unseen)
 - $N = \sum_{i=1}^M N_i$ be the total number of underlying SNPs within the observed range of effect sizes

Smoothing

- We consider three approaches
 - Nonparametric
 - Parametric
 - Semiparametric

Nonparametric Approach

- Loess (or Lowess) (Locally weighted polynomial regression, Cleveland, 1979; Cleveland and Devlin (1988))

$$1/pow_i = \mu(es_i) + \epsilon_i$$

- Linear function used to model $\mu(\cdot)$

Parametric Approach

- Weighted log-likelihood

$$\log(L) = \sum_i^M 1/pow_i \log(f_{\theta}(es_i))$$

- Considered Exponential and Weibull distributions
- Equate the expected and observed number of total discovery to estimate the total number of underlying susceptibility SNPs (in the observed range of effect sizes)

$$\sum_{i=1}^M N \times \frac{f_{\theta}(es_i)}{\sum_{j=1}^M f_{\theta}(es_j)} \times pow_i = \# \text{ of findings}$$

Semiparametric Approach

- Infinite mixture models (Bayesian perspective)

$$f^*(es_i^*) = \sum_{h=1}^{\infty} p_h f_{\theta_h}(es_i^*), \quad i = 1, \dots, N^*$$

- Dirichlet Process Mixture (DPM) of Exponentials
 - Dirichlet Process (Ferguson, 1973;1974), denoted by $DP(\alpha, G_0)$,

$$p_h = V_h \prod_{l < h} (1 - V_l),$$

where $V_h \stackrel{iid}{\sim} \text{Beta}(1, \alpha)$ and $\theta_h \stackrel{iid}{\sim} G_0$, $h = 1, \dots, \infty$

Finding Proper Power and Effect Sizes from Existing Studies

- Be aware of winner's curse problem
 - Ideally estimate effect sizes from replication study that is independent of discovery phase
- Multi-stage design
- Selective sampling
 - Studies for breast, prostate and colon cancers sampled subjects based on family history and young age-at-onset
- Multiple independent studies

Power=Probability of detecting from at least one

Examples

- Adult Height
- Crohn's Disease
- Cancers (Breast, Prostate, and Colon)

Adult Height

- Highly heritable (up to 80-90% of total variation has been attributed to genes)
- Three independent large GWAS (Gudbjartsson et al., 2008; Weedon et al., 2008; Lettre et al., 2008) have reported a total of 40 loci by studying a total of 65,000 subjects Initial scan using 15K-30K subjects following replication of a small number of top SNPs
- 33 loci which reached genome-wide significance in the first stage
- Estimate of effects sizes from the second stage

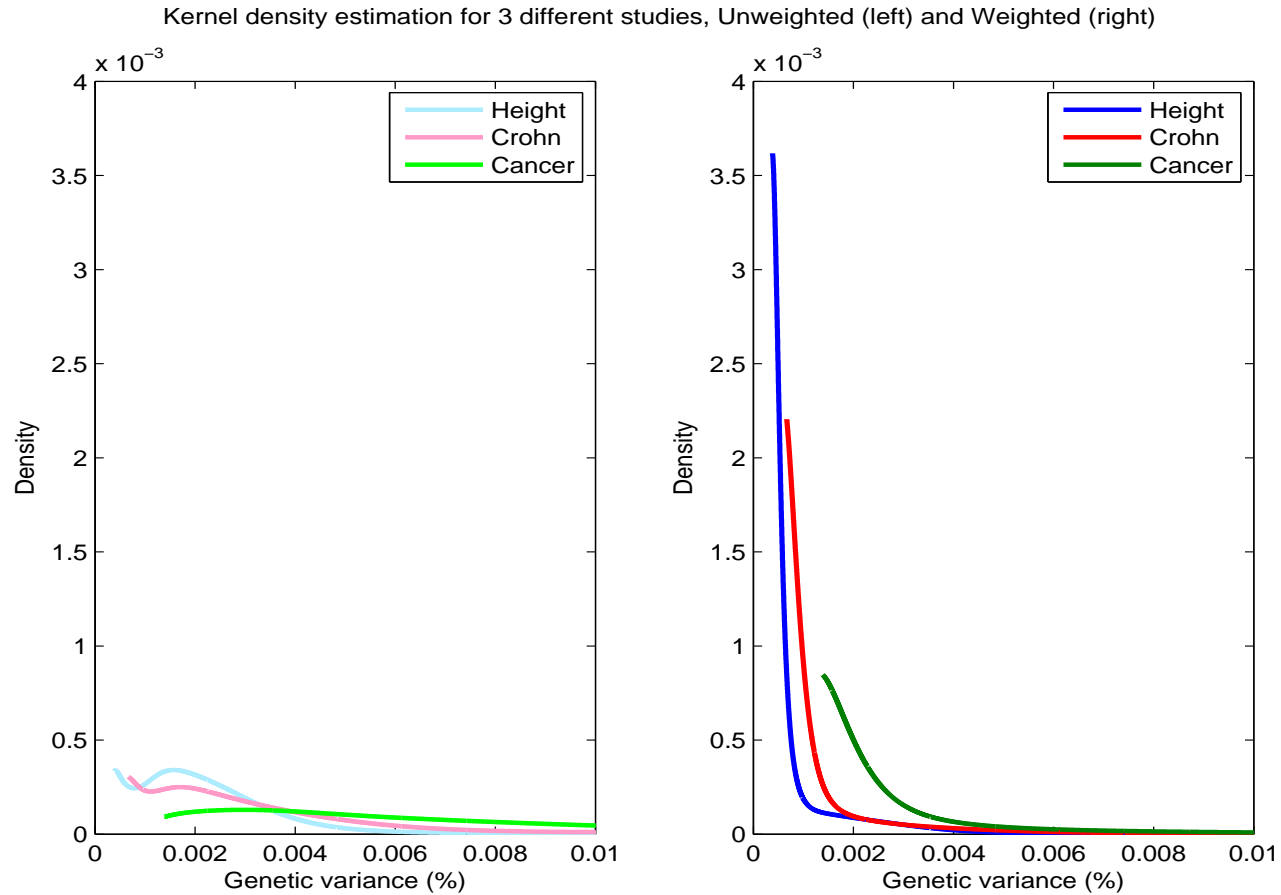
Crohn's Disease

- A common inflammatory bowel disease with very strong familial component (Sibling relative risk between 25-30)
- A recent two-stage GWAS (Berrett et al., 2008) reported 33 susceptibility loci
 - Initial scan involving 3,230 cases and 4,829 controls
 - Follow-up of 79 SNPs with p-value $< 5 \times 10^{-5}$ on 2,325 cases and 1,809 population controls
- 32 loci that reached genome-wide significance after combined analysis of first- and second-stage case-control studies
- Independent estimate of effect size from case-parent trios

Breast, Prostate and Colon Cancers

- Three common cancers with modest familial aggregation (sibling relative-risk=2) and no major environmental risk factor like smoking
- Various US and UK GWAS (Easton et al., 2007 for Breast; Thomas et al., 2008 for Prostate; COGENT Study, 2008 for Colon) have all together reported 10 susceptibility SNPs for each cancer
- Due to sparse data for each specific cancer, we pool the susceptibility SNPs from all three to obtain an "average" estimate of the distribution of effect sizes for these three cancers
- 22 SNPs (7 from Prostate, 5 from Breast, and 10 from Colon)
- Power calculations accounted for combination of multi-stage and selective sampling designs

Density of Effect Sizes for Susceptibility SNPs



Estimates of Total # Susceptibility SNPs within Observed Effect Sizes

	Loess	Tr. Exp	Tr. WB	MIX
Height	187	95	205	154
Crohn	161	78	183	135
Cancers	194	93	171	132

Tr. Exp: Truncated Exponential, Tr. WB: Truncated Weibull,
MIX: Mixture of truncated exponentials

Power Calculation Based on Distribution of Effect Sizes

- X = Total number of SNPs that could be discovered from a study
- We can write $X = X_1 + \dots + X_M$

$$X_l \sim \text{Binomial}(N_l, \text{pow}_l(\text{sample size}))$$

- $\text{pow}_l(\text{sample size})$ can be obtained from standard power calculation methods for a fixed effect size
- We can compute $E(X)$, $Pr(X \geq k)$ as a measure of power that will account for the likely distribution of effect sizes
- We can determine power for "new" discovery in a next GWAS by subtracting known loci from N_l , $l = 1, \dots, M$

Sample Size Calculation

- Required for a Single Stage GWAS to detect at least 5 novel loci with 80% power

	First Generation	Next Generation
Height	8,000	24,000
Crohn	3,000	10,000
Cancer	14,000	22,000

- For a case/control study, # of cases = Half of total sample size

Estimating Possible Total Genetic Variance

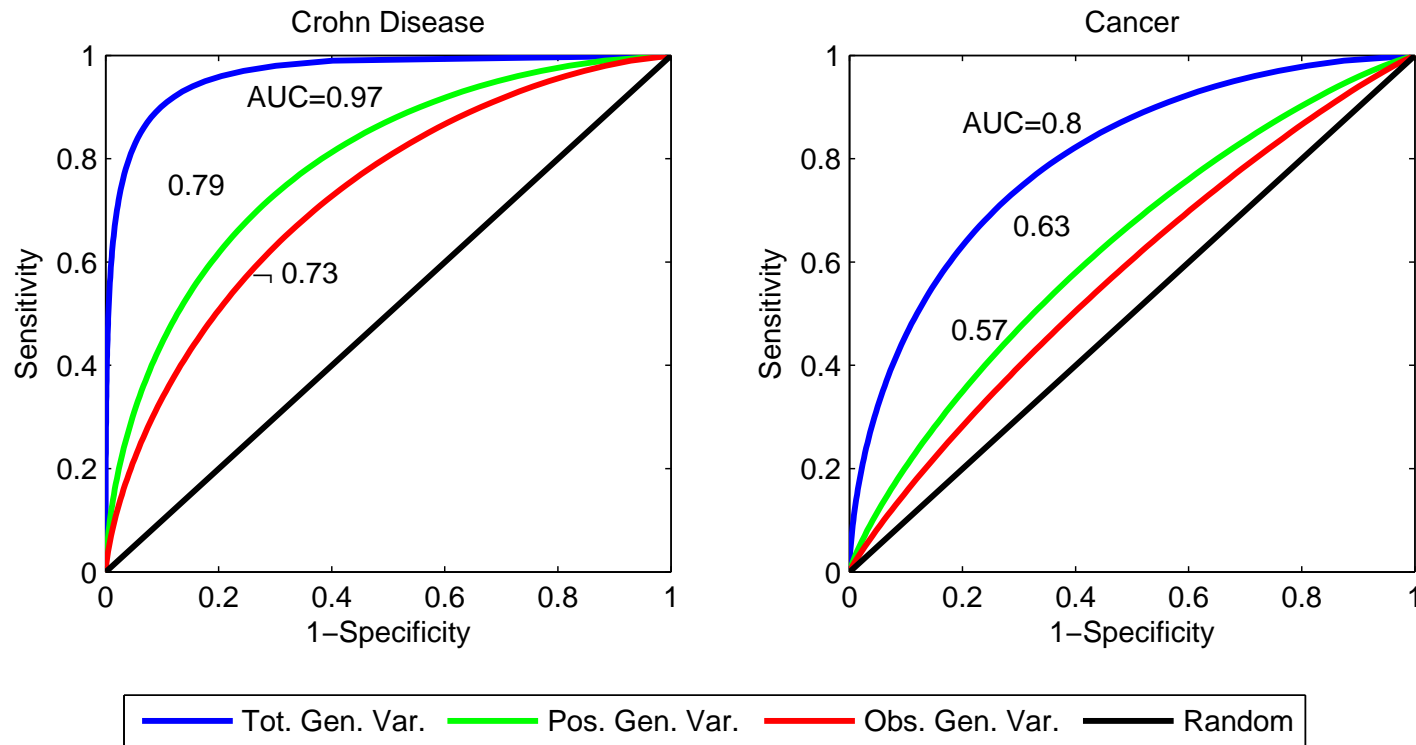
- Estimate of total genetic variance (OTG) that can be explained by all the seen and unseen SNPs within the range of observed effect sizes

$$OTG = \sum_{i=1}^M \hat{N}_i e s_i$$

$$POTG = \frac{PTG}{\text{Total Genetic Variance}}$$

- We estimated PPTG=15% for adult height, 20% for Crohn's disease and 17% for cancer

ROC Curve for Risk Discrimination (Pharoah et al., 2002)



Summary

- Used accessible empirical information from the existing GWAS such as sample sizes, ORs, and allele frequencies, not data itself.
- Considered various estimating methods for the total number of underlying SNPs at observed effect sizes
- Allowed power and sample size calculations to account for the distribution of susceptibility SNPs with similar effects.
- Provided researchers with a tool for evaluating the potential public health benefits of their studies in a designing stage