ON THE EFFICIENCY OF GENOME-WIDE SCANS: A MULTIPLE HYPOTHESIS TESTING PERSPECTIVE

Lei Sun[1]

Una din caracteristicile studiilor genetice pentru întregul genom este intensitatea slabă a efectelor genetice precum şi raritatea lor. Aceasta conduce la serioase dificultăţi legate de testarea ipotezelor statistice pentru efectele diferitelor gene considerate în cadrul studiului. În această lucrare găsim frontierele regiunilor care definesc descoperiri reale atât pentru intensitatea semnalului genetic, cât şi a funcţiei sale de densitate. Demonstrăm că aplicarea obişnuită a analizei genetice pentru întregul genom nu produce analize cu putere statistică mare, iar semnalele autentice nu pot fi detectate cu precizie. În încheiere, discutăm câteva strategii moderne pentru a îmbunătăţi puterea statistică cu ajutorul metodei de prioritizare a genomului.

A main characteristic of the current high-throughput genome-wide studies is that the signals to be detected are weak in strength and low in density. This leads to statistical challenges in the context of high-dimensional hypothesis testing. We first show the boundaries for signal strength and density that allow for efficient true discoveries. We then demonstrate why the agnostic approach to the genome is not powerful, in that most of the underlying signals cannot be detected with good precision. Lastly, we discuss some emerging methods developed to improve power via prioritization of the genome.

Keywords: False Discovery Rate, Non-Discovery Rate, Genome-wide Association Studies, Power, Stratification.

MSC2000: 53C 05.

1. Introduction

Due to the recent advance in high-throughput genotyping technology, most current association studies rely on genome-wide scans, in particular the genome-wide association study (GWAS) design. The high-throughput scans simultaneously investigate many genetic variants (i.e. SNPs) to identify the ones that are truly associated with complex human diseases or traits.

Briefly, for a phenotype of interest Y (e.g. a quantitative trait such as height, or a binary disease outcome such as being diabetic or not), genotype data of 100K or more SNPs, scattered more or less randomly across the human genome, are collected for a large number (typically 2K-5K) of subjects. The association between the

[1]Associate Professor, Division of Biostatistics, Dalla Lana School of Public Health, and Department of Statistics, University of Toronto, Toronto, Canada, e-mail: sun@utstat.toronto.edu

	Declared non-significant	Declared significant	Total counts
Truth: H_0	U	V	m_0
Truth: H_1	T	S	m_1
Total	$m-R$	R	m

TABLE 1. *Summary of events for multiple hypothesis testing*

phenotype and any given SNP is typically investigated via linear or logistic regression of Y on X, where X denotes the number of copies of the mutation allele of the SNP ($X = 0$, 1 or 2), with or without adjusting for environmental factors. The statistical evidence for significant association must be adjusted for the fact that 100K or more hypotheses (one for each SNP) have been performed in a single genome-wide scan. Although this approach has lead to many successful findings, there is no shortage of skepticism or debate about its efficiency, particularly in light of the hundreds of millions of dollars spent. For example, the New England Journal of Medicine recently published a series of commentaries on the utilities of GWAS with conflicting views [5, 6, 7].

Here we address the efficiency issue of the GWAS design from the multiple hypothesis testing point of view, in the context of false discovery rate (FDR) control [1]. We choose FDR instead of the traditional family-wise error rate (FWER) control, because a conclusion of low power with FDR control could be easily extended to the situation when the more stringent FWER was used to control the type 1 error rate. In Section 2 we show the level of FDR and power as a function of the signal strength and density, where power is measured by (1 - the non-discovery rate) (NDR) [2]. Applications to several published GWAS data in Section 3 demonstrate that most of the underlying signals cannot be detected with good precision. In Section 4, we discuss a couple of recent methods proposed to improve power of GWAS by prioritization of the genome, and make concluding remarks.

2. **Boundaries for Efficient Signal Discovery**

When a large number of hypotheses tests are carried out simultaneously, the resulting true/false positives/negatives can be summarized in Table 1. Using the notations in Table 1, FDR $= \mathrm{E}[V/R]$ [1] and NDR $= \mathrm{E}[T]/m_1$ [2].

To calculate the level of FDR and NDR, we use a simple model for which analytical results can be derived and key issues can be better understood. We assume that all tests are independent of each other and the test statistics are normally distributed. For the m_0 true null hypotheses, we assume $Z|H_0 \sim N(0,1)$, and for the m_1 true signals, $Z|H_1 \sim N(\mu,1)$ with $\mu > 0$. Note that μ measures the strength of a signal which obviously depends on the sample size, and in genetic association studies also on the effect size (e.g. relative risk) and the frequency of the

		N=2K	N=5K	N=10K
$r = 1.2$	$pA = 1\%$	0.7	1.0	1.5
	$pA = 5\%$	1.4	2.3	3.2
	$pA = 50\%$	2.8	4.4	6.2
	$pA = 99\%$	0.5	0.7	1.1
$r = 1.5$	$pA = 1\%$	1.7	2.6	3.7
	$pA = 5\%$	3.5	5.5	7.8
	$pA = 50\%$	5.5	8.8	12.4
	$pA = 99\%$	0.8	1.3	1.8

TABLE 2. *Values of μ, the signal strength, for a SNP with a range of relative risk r and mutation allele frequency pA, and detected in a sample of size N for association with a disease with population prevalence $K = 5\%$.*

mutation allele of a truly associated SNP. The proportion of the alternatives among all hypotheses performed, $\pi_1 = m_1/m$, is a measure of signal density.

We consider μ ranging from 1 to 6, and π_1 ranging from .001% to 10%. The lower boundaries of $\mu = 1$ and $\pi_1 = 0.001\%$ are chosen to reflect the characteristics of current GWAS. For example, for a SNP that is truly associated with a disease with population prevalence of $K = 5\%$, the corresponding association statistic based on the Armitage trend test [10] has $\mu = 1.4$, if this SNP has a relative risk of $r = 1.2$ and mutation allele frequency of $pA = 5\%$, and the sample used to detect the association is $N = 2K$ (1K cases and 1K controls). The signal strength is reduced to $\mu = 0.7$ if $pA = 1\%$. The low signal density is determined by the GWAS design in which > 100K SNPs are selected to evenly cover the genome regardless of the phenotype of interest. Therefore only a small proportion of these SNPs are expected to be associated with any given phenotype as evident in the applications below. Table 2 shows the values of μ for different combinations of r, pA and N. The upper boundaries of $\mu = 6$ and $\pi_1 = 10\%$ are unrealistic for high-throughput genome-wide scans, but they were chosen to provide insights to the necessary signal strength and density needed for efficient discoveries.

Using the above model the considering a fixed rejection region approach in which we reject all hypotheses with test statistics $Z \leq t$, we have

$$\begin{aligned} \text{NDR} &= 1 - \Phi(\mu - t), \\ \text{FDR} &= \frac{(1-\pi_1)\,\Phi(-t)}{(1-\pi_1)\,\Phi(-t) + \pi_1\,\Phi(\mu - t)}, \end{aligned}$$

where Φ is the cdf of standard normal distribution. Of particular importance is the fact that NDR and FDR depend on μ and π, the signal strength and density, but they do not directly depend on m, the total number of hypotheses. Figure 1 show NDR vs. FDR for different combinations of π_1 and μ, left for $\pi_1 = 0.01\%$ and

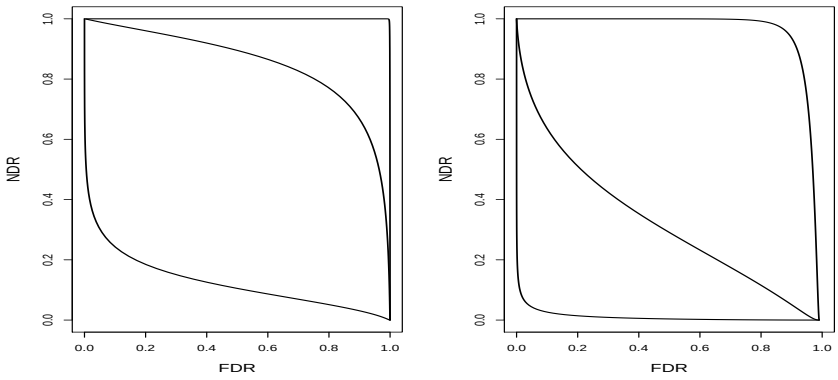

FIGURE 1. *NDR vs. FDR for different combinations of signal density (π_1) and strength (μ). Left: $\pi_1 = 0.01\%$; Right: $\pi_1 = 1\%$. Each curve is for a signal strength level, $\mu = 1$ (top right), $\mu = 3$ (middle) and $\mu = 5$ (bottom left).*

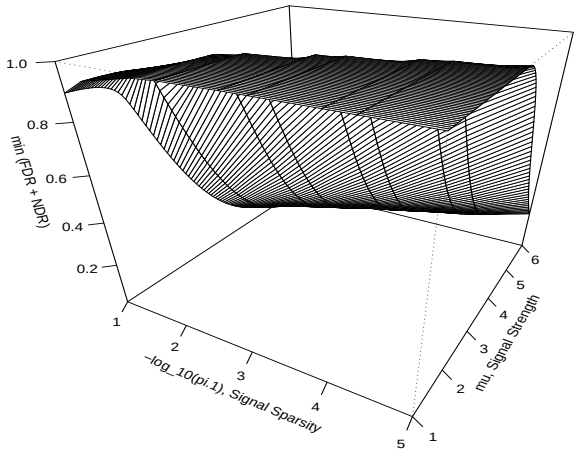

FIGURE 2. *min(NDR+ FDR) vs. $-log(\pi_1)$, signal sparsity, and μ, signal strength.*

right for $\pi_1 = 1\%$ (results qualitatively similar for other π_1 values). The trade of between type 1 error rate as measured by FDR and type 2 error rate by NDR is as expected and was discussed in detail in [2]. Clearly both π_1 and μ are crucial factors in determining the error rates.

Figure 2 summarizes the results more concisely. The 3D plot shows the smallest combined error rate, min(FDR+NDR), achievable for each combination of signal strength (μ) and sparsity (density on the $-log_{10}(\pi_1)$ scale). It is painfully clear that

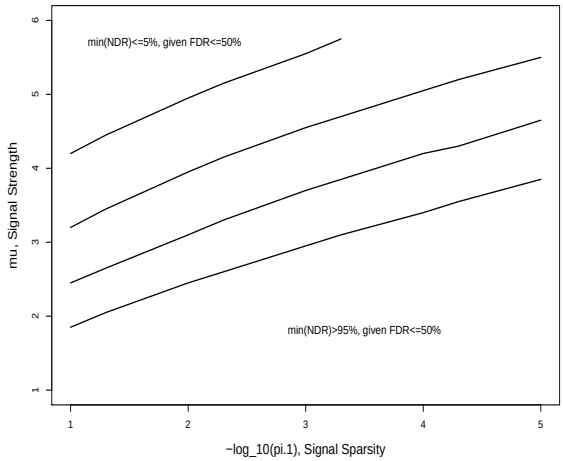

FIGURE 3. *Boundaries for signal sparsity and strength defined by min(NDR) $\leq c$, where $c = 5\%$ (top line), 30%, 70% or 95% (bottom line), given FDR controlled at $\leq 50\%$.*

if the signal density is less than 1% ($-log_{10}(\pi_1) > 2$ in sparsity), then it is difficult to contain the combined error rate, unless the signal strength is extremely strong.

In practice, one may not want to treat the two types of error equally. Stratified plots similar to those in Figure 2 could be drawn, stratified by a desirable level of FDR. The splicing of the surface at min(NDR) $= c$ provides the boundaries for signal detection at a pre-specified error rate which is shown in Figure 3 for FDR $\leq 50\%$. (Pattens for other FDR levels are characteristically similar.) Specifically, area above each line defines the range for π_1 and μ that could control min(NDR) at a c level, where $c = 5\%, 30\%, 70\%$ or 95%, conditional on FDR being controlled at 50% or less. Results show that if signal is sparse ($\pi_1 < 1\%$) and weak ($\mu < 1.5$), then one cannot identify a mere 5% of the signals (NDR $> 95\%$) even allowing half of the positives to be false (FDR $= 50\%$).

3. **GWAS Applications**

We applied the method to several published GWAS data, including studies of Parkinson disease by [8] and seven common diseases by [12]. Table 3 shows the phenotype of interest, the sample size, the number of SNPs tested, the estimated proportion of the signals, and the minimal achievable combined error rate in each of the studies. In all cases, the signal density is extremely low which in turn results in high error rate. Figure 4 provides a representative NDR vs. FDR curve using the WTCCC Coronary Artery Disease data. The figure clearly demonstrates the inefficiency of the GWAS design, because more than 95% of the underlying signals will be missed (NDR $> 95\%$) even we allow half of the discoveries to be false (FDR $= 50\%$).

Study (N)	Phenotype	m	$\hat{\pi}_1$	min(FDR+NDR)
Maraganore				
($\approx$ 500+500)	Parkinson Disease	197,222	2.4%	0.974
WTCCC				
($\approx$ 3K+2K)	Bipolar Disorder	360,971	5.7%	0.939
	Coronary Artery Disease	360,971	3.8%	0.959
	Crohn's Disease	360,971	5.4%	0.940
	Hypertension	360,971	3.0%	0.960
	Rheumatoid Arthritis	360,971	2.4%	0.973
	Type 1 Diabetes	360,971	3.0%	0.951
	Type 2 Diabetes	360,971	5.4%	0.944

TABLE 3. *GWAS application results. N is the sample size (controls+cases), m is the total number of SNPs, i.e. the number of hypotheses performed in each study, $\hat{\pi}_1$ is the estimated signal density, and min (FDR+NDR) is the minimal achievable combined error rate.*

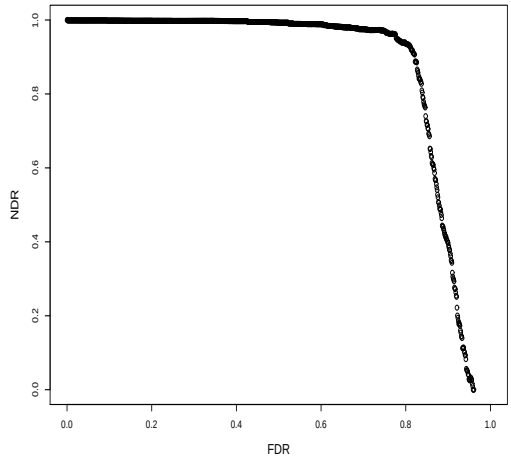

FIGURE 4. *NDR vs. FDR for the application to the WTCCC Coronary Artery Disease GWAS data.*

4. **Discussion and Conclusion**

Our analytical results were derived under the assumptions of independence and equal signal strength. The independence assumption is not a particular concern here because GWAS SNPs are typically selected to have small correlation among them so that the information provided by these SNPs do not overlap. We also considered the situation when the underlying signals have different signal strength, e.g. signal strength, μ, uniformly distributed between 1 and 6, or more signals with low signal

strength. Results are characteristically similar to those presented above. Observations from real data application also support what we concluded from the theoretical model. The boundaries shown in Figure 3 have potential connection with the work of [3] in which classifiable regions are given under a different parameterization of the signal density and strength. This is the subject of on-going research.

Both analytic and application results show that current GWAS are in fact underpowered even with a sample size of 2K or more. The root of the problem is that the statistical significance for each single SNP must be adjusted by a factor of 100K or more at the genome-wide level. Instead of this agnostic approach to the genome which leads to the same multiple hypothesis testing penalty for all SNPs, alternative methods have been proposed to improve power by giving different priorities to different parts of the genome. For example, [11] and [13] considered a stratified approach, and [4] and [9] a weighted method, both in the context of FDR control. The essence of these two methods is to prioritize the genome and maintain power to interrogate candidate regions within the GWAS design. These candidate regions are defined by available prior information such as linkage or gene-expression results, or biological knowledge, and they have higher prior odds to be associated with the phenotype of interests.

The multiple hypothesis testing issue will be more severe when we move forward from GWAS to the whole-genome sequencing design. In the latter, data of more than ten million SNPs are collected, providing a better coverage of the genome. Results in this report however show that the increased penalty associated with multiple hypothesis testing could potentially outstrip the benefits provided by the sequencing data. Therefore, it is critical to consider alternative methods to improve power of high-throughput genome-wide scans.

Acknowledgment

I would like to thank Professor Dan Nicolae and Professor Jiashun Jin for insightful discussions. This research is supported by research grants from Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institute of Health Research (CIHR).

REFERENCES

[1] *Y. Benjamini and Y. Hochberg.* Controlling the false discovery rate: A practical and powerful approch to multiple testing. J. Roy. Statist. Soc. Ser. B, **57**:289–300, 1995.

[2] *R. V. Craiu and L. Sun.* Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. Statistica Sinica, **18**:861–879, 2008.

[3] *D. Donoho and J. Jin.* Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist., **32**:962–994, 2004.

[4] *C. R. Genovese, K. Roeder and L. Wasserman.* False discovery control with p-value weighting. Biometrika, **93**, 2006.

[5] *D. B. Goldstein.* Common genetic variation and human traits. N Engl J Med, **360**:1696–1698, 2009.

[6] *J. N. Hirschhorn*. Genomewide association studies–illuminating biologic pathways. N Engl J Med, **360**:1699–1701, 2009.

[7] *P. Kraft and D. J. Hunter*. Genetic risk prediction–are we there yet? N Engl J Med, **360**:1701–1703, 2009.

[8] *D. M. Maraganore, M. de Andrade, T. G. Lesnick, K. J. Strain, M. J. Farrer, W. A. Rocca, P. V. K. Pant, K. A. Frazer, D. R. Cox and D. G. Ballinger*. High-resolution whole-genome association study of parkinson disease. Am J Hum Genet, **77**:685–693, 2005.

[9] *K. Roeder, S.-A. Bacanu, L. Wasserman and B. Devlin*. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet, **78**:243–252, 2006.

[10] *S. Slager and D. Schaid*. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Human Heredity, **52**:149–153, 2001.

[11] *L. Sun, R. V. Craiu, A. D. Paterson and S. B. Bull*. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet Epidemiol, **30**:519–530, 2006.

[12] *WTCCC*. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, **447**:661–678, 2007.

[13] *Y. J. Yoo, S. B. Bull, A. D. Paterson, D. Waggott and L. Sun*. Were genome-wide linkage studies a waste of time? exploiting candidate regions within genome-wide association studies. Genetic Epidemiology, **34**:107–118, 2010.