Published in final edited form as:

Genet Epidemiol. 2011 July ; 35(5): 303–309. doi:10.1002/gepi.20578.

On the Follow-Up of Genome-Wide Association Studies: An Overall Test for the Most Promising S.N.P.'s

Peter J. Lipman^{*1}, Michael H. Cho², Per Bakke^{3,4}, Amund Gulsvik^{3,4}, Xiangyang Kong⁵, David Lomas⁶, Wayne Anderson⁵, Edwin Silverman², and Christoph Lange¹

¹Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

²Channing Laboratory, Brigham and Women's Hospital, Boston, MA 02115, USA

³Department of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway

⁴Institute of Medicine, University of Bergen, Bergen, Norway

⁵GlaxoSmithKline Research and Development, Research Triangle Park, NC 27709, USA

⁶Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK

Abstract

Even in large-scale genome-wide association studies, only a fraction of the true associations are detected at the genome-wide significance level. When few or no associations reach the significance threshold, one strategy is to follow-up on the most promising candidates, i.e. the single nucleotide polymorphisms with the smallest association-test p-values, by genotyping them in additional studies. In this communication, we propose an overall test for genome-wide association studies that analyzes the SNP's with the most promising p-values simultaneously and thereby allows an early assessment of whether the follow-up of the selected SNP's is likely promising. We theoretically derive the properties of the proposed overall test under the null hypothesis and assess its power based on simulation studies. An application to a GWAS for chronic obstructive pulmonary disease suggests that there are true association signals among the top SNPs and that an additional follow-up study is promising.

Keywords

genome wide association studies; snps association tests; chronic obstructive pulmonary disease; statistical genetics; multiple testing

Introduction

When a statistical analysis involves multiple tests, such as a genome-wide association study (GWAS) that computes upwards of one million SNP-association tests, one must correct the nominal p-values to account for the multiple opportunities to achieve statistical significance. Even in large studies, the combination of weak association signals and the conservative

^{*}For correspondence, please contact Peter Lipman at the Department of Biostatistics, Harvard School of Public Health, 4th Floor, Building 2, 677 Huntington Ave, Boston, MA 02115; pjlipman@fas.harvard.edu; 443-799-0171.

correction for multiple comparisons (such as the widely-used Bonferroni correction) often leads to one's failing to reject the null hypothesis when the null hypothesis may in fact be false. In this communication, we develop an overall statistical test that is applied to the set of most promising association tests from the original study. The overall test will conclude whether there is evidence that the null hypothesis is in fact false among the most promising candidates, even though no individual test achieves a global significance level. The conclusions of the test are used to determine whether a follow-up study on the top hits is recommended. This single test is statistically straight-forward and easily implemented. Using simulations, power is assessed under realistic scenarios of GWAS's. The test can also be modified to help determine the number of top hits which should be further examined in the follow-up study.

Materials and Methods

Suppose we have a GWAS and no SNP achieves statistical significance after Bonferroni correction. Although we fail to reject the null hypothesis of no genetic association for each SNP, it is possible that we are making some type-2 errors, since Bonferroni correction is conservative, our GWAS may be underpowered, and/or the true associations may be relatively weak. Therefore, we may want to perform a follow-up study on an independent population, testing only the top hits from the original study. In this paper, we propose a statistical test, using the original study, which will conclude whether this follow-up study is likely promising.

Assume we are willing to perform a follow-up study on a predetermined percentage, $100 \times \theta_F$, of the top SNP's. Under the global null hypothesis that no associations exist, all the p-values from the original GWAS tests follow a Uniform(0,1) distribution and the top SNPs follow a Uniform(0, θ_F) distribution. Under the alternative hypothesis that at least association does exist, the p-values from candidate SNPs will be smaller than expected under the null; they will be clustered towards 0.

The general idea of the test is to partition the p-values into a predetermined number of bins ($J+1$); one bin collects all the (large) p-values that are contributed by the SNP's which will be of no interest in the follow-up study, and the other J bins partition the p-values falling between 0 and θ_F . We then count how many p-values fall into each bin. If the bins contain more p-values than expected under the null - Uniform(0,1) - distribution, there is evidence to reject the null hypothesis and conclude that a follow-up study is recommended. Partitioning the p-values into bins performs the test non-parametrically.

More formally, let θ_F be the fixed percentage of SNP's one is willing to resequence, let y_i be the p-values (not significant after Bonferroni correction) from the SNP-association tests ($i = 1, \dots, n$) and let $J+1$ represent the number of bins used to partition your p-values. Define the $J+1$ test statistics as follows:

$$X_j = \sum_{i=1}^n 1 \left\{ \frac{(j-1)\theta_F}{J} < y_i \leq \frac{j\theta_F}{J} \right\} \text{ for } j=1, \dots, J \quad (1)$$

$$X_{J+1} = n - \sum_{j=1}^J X_j$$

where $1\{\cdot\}$ is an indicator function.

Under the null hypothesis, the test statistics, which are defined above as the number of p-values that fall into the bins, jointly follow a multinomial distribution, with bin probabilities θ_F/J (and the last bin has probability $1-\theta_F$). Under the alternative hypothesis, the test statistics also jointly follow a multinomial distribution, with elevated (unknown) bin probabilities for the lower J bins.

More formally, under H_0 ,

$$\{X_j\} \sim \text{multinomial}(n, p_1, \dots, p_J, 1 - \sum_{j=1}^J p_j)$$

$$p_1 = \dots = p_J = \frac{\theta_F}{J}$$

And under H_A ,

$$\{X_j\} \sim \text{multinomial}(n, p_1, \dots, p_J, 1 - \sum_{j=1}^J p_j)$$

$$p_1, \dots, p_J \geq \frac{\theta_F}{J}, \exists j \text{ s.t. } p_j > \frac{\theta_F}{J}$$

With the distribution of the test statistics known under the null and alternative hypotheses, we formulate a likelihood ratio test for the p_i . Using the multinomial distribution, the likelihood of the data is:

$$L = \binom{n}{x_1 \dots x_J} p_1^{x_1} \dots p_J^{x_J} (1 - \sum_{j=1}^J p_j)^{(n - \sum_{j=1}^J x_j)} \quad (2)$$

and the Likelihood Ratio Test Statistic is $\lambda = \frac{\sup_{H_0} L}{\sup_{H_0 \cup H_A} L}$.

Under H_0 , using Maximum Likelihood Estimation, $\hat{p}_{j0} = \frac{\theta_F}{J}$ since the bin probabilities are fixed. Under H_A , a closed form solution for the maximum likelihood estimates of the bin probabilities, subject to the elevated probability constraint, is derived using Lagrange

multipliers. Please see the Supplementary Text for the closed form solution of $p_{jA}^{\hat{}}$, the maximum likelihood estimates under the alternative. Thus,

$$\lambda = \left(\frac{\theta_F}{J\hat{p}_{1A}}\right)^{x_1} \cdots \left(\frac{\theta_F}{J\hat{p}_{JA}}\right)^{x_J} \left(\frac{1-\theta_F}{\hat{p}_{J+1A}}\right)^{(n-\sum_{j=1}^J x_j)} \quad (3)$$

and small values of the likelihood ratio test statistic imply evidence to reject the null hypothesis.

COPD Application: Data Description

After verifying the properties of the test statistic, the method was applied to a Chronic Obstructive Pulmonary Disease (COPD) GWAS dataset. Details of the included study populations and methods for the GWAS have been previously described [Bell et al. 1972] [Cho et al. 2010] [?] [Hersh et al. 2005] [Vestbo et al. 2008] [Zhu et al. 2007]. Cases and controls were obtained from the Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE; SCO104960, NCT00292552) study [Cho et al. 2010]; the National Emphysema Treatment Trial (NETT) (cases) [?] and Normative Aging Study (NAS) (controls) [Bell et al. 1972] [Hersh et al. 2005]; and the GenKOLs cohort from Bergen, Norway [Zhu et al. 2007] [Pillai et al. 2009]. All subjects were self-identified whites; cases were GOLD Stage II or greater [Rabe et al. 2007] in severity, without alpha-1 antitrypsin deficiency; controls were current or former smokers with normal lung function.

Genotyping was performed on the Illumina Quad 610 or HumanHap 550 arrays (Illumina, Inc; San Diego, CA) with a uniform set of quality control criteria applied to both the raw genotype intensity and SNP call data, as detailed previously [Cho et al. 2010]. Additional subjects were removed as ancestry outliers based on principal components. After quality control, a total of 499,506 markers on 2940 cases and 1380 controls remained. Genome-wide association tests were performed under an additive model, adjusting for ancestry based principal components [Price et al. 2006], pack-years of cigarette smoking, and age, using PLINK version 1.05 [Purcell et al. 2007]. The genomic inflation factor for this analysis was 1.02.

Results

Simulations under the Null Hypothesis

The exact distribution of the likelihood ratio test statistic λ does not follow the standard chi-squared distribution due to the elevated probability constraints under the alternative. Therefore, we simulate the likelihood ratio test statistic under the null hypothesis in order to determine the rejection region for a given α -level. For the following calculations, we assume we have p -values from 1,000,000 tests (i.e. $n=1,000,000$), which follow a Uniform(0,1) distribution. We assume that we are willing to perform a follow-up study on the top 10,000 (1%, i.e. $\theta_F=.01$) tests and create 11 bins ($J=10$). The exact cut-off values are determined using the histogram presented in Figure 1 and are reported in Table 1.

Simulations for Power

We then simulated data under the alternative hypothesis and used the cut-off values determined above for power calculations at $\alpha = .05$ level. We simulated between 60 and 390 disease susceptibility loci (DSL's), each with odds ratios of 1.2. One-third of the DSL's has minor allele frequency .1, .2, and .3. The DSL's are simulated under an additive model. We ran the simulations under 3 disease prevalence rates: 5%, 10% and 14%. For each DSL, we simulated a 2 by 3 genotype table with 2000 cases and 2000 controls, then computed an Armitage trend test to calculate the p-values. We focus our efforts on simulating many common DSL's with small effect sizes, for this scenario is most difficult with current methodology. For the non-DSL's, p-values are simulated from a Uniform(0,1) distribution. Only the p-values that are not Bonferroni-significant are used in the test. Results (Table 2, Figure 2) are similar for all 3 disease prevalence rates and are reported below. Although power is minimal when only few SNPs are DSL's, it is important to note that on average, in all simulations, only ~2.8% of the DSL are Bonferroni significant, so the vast majority remain undetected by the standard methods.

Power simulations were also performed while varying the bin number parameter, J , to determine reasonable values for J in real applications. Using the same values for all other parameters as in the previous simulations and disease prevalence set to 5%, we found power is at nearly equivalent levels for $10 \leq J \leq 100$, although power started to drop off for $J > 100$.

These results are reported below in Table 3 and Figure 3. Under the null hypothesis, $\frac{n*\theta_F}{J}$ is the expected number of p-values in each of the first J bins. From the above results, we have

reasonable power when $100 \leq \frac{n*\theta_F}{J} \leq 1000$ and we suggest, in practice, to choose J such that this inequality holds.

COPD Application Results

For the COPD dataset, there were originally two statistically significant p-values on chromosome 4 (rs7671167 and rs1903003, $p=8.59e-8$, $7.18e-8$ (.042, .036 after Bonferroni correction)). We removed these p-values from the dataset and used our method to test if there was a evidence in the original tests that a follow-up study on the top hits would be beneficial. In order to apply this method to the COPD GWAS dataset, we simulated the null distribution of the likelihood ratio test statistic, using 499,504 p-values from a Uniform(0,1) distribution, to determine the cut-off points for the test statistic. This null distribution (10,000 simulations) is presented in Figure 4 below. The histogram of the p-values from the GWAS is presented in Figure 5 below. We used the top 1% of the SNPs (i.e. $\theta_F = .01$, willing to resequence roughly 5000 SNPs), and created 11 bins ($J=10$) to calculate the test statistic. These parameters were also used in the null distribution simulation. From the null distribution simulations, the cutoff value for the likelihood ratio statistic λ at the $0.05-\alpha$ level is 0.003. The observed value in this dataset for λ is 1.99×10^{-6} , $p < 0.0001$. We therefore reject the null hypothesis and recommend performing a follow-up study on the top 1% of the SNP's in the original GWAS, as there are likely marginally significant SNP's that are, in fact, truly associated with the disease.

Discussion

We have developed a method, applicable for large multi-test analyses, to test if there is evidence that type-2 errors occur for the non-significant p-values. These type-2 errors are prevalent due to the conservative nature of Bonferroni-correction and the possibility of underpowered studies and weak effect sizes. The method uses the fact that the null distribution of the p-values follows a Uniform(0,1) distribution, and non-parametrically tests if this holds in the data-set, against the alternative that the p-values are clustered towards 0. The method allows us to assess if a follow-up study would likely be beneficial. It can also be used to determine how many candidates to push forward into the follow-up by testing at different values of θ_F : either iteratively searching over different θ_F to maximize the test statistic or searching to minimize θ_F (while keeping statistical significance) for financial considerations. This method is widely applicable, as the null distribution is easy to simulate and the method can be implemented after any association test, since it only uses the p-values. We have demonstrated that the method achieves adequate power when the number of underlying DSL's is moderately large (greater than 250). We therefore expect this method to be most applicable when the disease is caused by many common SNP's that have small effect sizes. The SNP's in these situations, likely true for factors such as height, are often the most difficult to discover with current methods. We applied the method to a COPD GWAS, and concluded that a follow-up study on the top 1% of the SNPs would be beneficial; the data suggests many of non-significant but "suggestive" SNP's may, in fact, be associated with the disease.

Acknowledgments

The members of the Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE) Steering Committee are: H. Coxson (University of British Columbia, Vancouver, BC, Canada); L. Edwards (GlaxoSmithKline, Research Triangle Park, NC, USA); K. Knobil (Co-Chair; GlaxoSmithKline, Greenford, UK); D. Lomas (University of Cambridge, Cambridge, UK); W. MacNee (University of Edinburgh, Edinburgh, UK); E. Silverman (Brigham and Womens Hospital, Boston, MA, USA); R. Tal-Singer (GlaxoSmithKline, King of Prussia, PA, USA); J. Vestbo (Co-Chair; Hvidovre Hospital, Hvidovre, Denmark); and J. Yates (GlaxoSmithKline, Research Triangle Park).

The members of the ECLIPSE Scientific Committee are: A. Agusti (Son Dureta Hospital and Címera, Palma, Spain); P. Calverley (University Hospital Aintree, Liverpool, UK); B. Celli (Caritas St. Elizabeths Medical Center, Boston, MA, USA); C. Crim (GlaxoSmithKline, Research Triangle Park); G. Hagan (GlaxoSmithKline, Greenford); W. MacNee (Chair); S. Rennard (University of Nebraska, Omaha, NE, USA); R. Tal-Singer, E. Wouters (University of Maastricht, Maastricht, the Netherlands); and J. Yates.

We acknowledge the co-investigators in the NETT Genetics Ancillary Study including Joshua Benditt, Gerard Criner, Malcolm DeCamp, Philip Diaz, Mark Ginsburg, Larry Kaiser, Marcia Katz, Mark Krasna, Neil MacIntyre, Barry Make, Rob McKenna, Fernando Martinez, Zab Mosenifar, John Reilly, Andrew Ries, Paul Scanlon, Frank Scirba, and James Utz.

References

- Bell B, Rose C, Damon H. The Normative Aging Study: an interdisciplinary and longitudinal study of health and aging. *Aging Hum Dev.* 1972; 3:5–17.
- Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, DeMeo DL, Hunninghake GM, Litonjua AA, Sparrow D, Lange C, Won S, Murphy JR, Beaty TH, Regan EA, Make BJ, Hokanson JE, Crapo JD, Kong X, Anderson WH, Tal-Singer R, Lomas DA, Bakke P, Gulsvik A, Pillai SG, Silverman EK. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet.* 2010; 42(3):200–202. [PubMed: 20173748]

- Fishman A, Martinez F, Naunheim K, Piantadosi S, Wise R, Ries A, Weinmann G, Wood DE. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med*. 2003; 348:2059–2073. [PubMed: 12759479]
- Hersh CP, Demeo DL, Lange C, Litonjua AA, Reilly JJ, Kwiatkowski D, Laird N, Sylvia JS, Sparrow D, Speizer FE, Weiss ST, Silverman EK. Attempted replication of reported chronic obstructive pulmonary disease candidate gene associations. *Am J Respir Cell Mol Biol*. 2005; 33:71–78. [PubMed: 15817713]
- Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Feng S, Hersh CP, Bakke P, Gulsvik A, Ruppert A, Ldrup Carlsen KC, Roses A, Anderson W, Rennard SI, Lomas DA, Silverman EK, Goldstein DB, ICGN Investigators. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*. 2009; 5(3):e1000421. [PubMed: 19300482]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38:904–909. [PubMed: 16862161]
- Purcell S, Neale B, Todd-Browne K, Thomas L, Ferreira MAR, Bender D, Maller J, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81:559–575. [PubMed: 17701901]
- Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, Fukuchi Y, Jenkins C, Rodriguez-Roisin R, van Weel C, Zielinski J. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD Executive Summary. *Am. J. Respir. Crit. Care Med*. 2007; 176:532555.
- Vestbo J, Anderson W, Coxson HO, Crim C, Dawber F, Edwards L, Hagan G, Knobil K, Lomas DA, MacNee W, Silverman EK, Tai-Singer R, ECLIPSE Investigators. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J*. 2008; 31:869–873. [PubMed: 18216052]
- Zhu G, Warren L, Aponte J, Gulsvik A, Bakke P, ICGN Investigators. Anderson WH, Lomas DA, Silverman EK, Pillai SG. The SERPINE2 gene is associated with chronic obstructive pulmonary disease in two large populations. *Am J Respir Crit Care Med*. 2007; 176:167–173. [PubMed: 17446335]

Appendix: Test Statistic Derivation

Under H_0 ,

$$\{X_j\} \sim \text{multinomial}(n, p_1, \dots, p_J, 1 - \sum_{j=1}^J p_j)$$

$$p_1 = \dots = p_J = \frac{\theta_F}{J}$$

Under H_A ,

$$\{X_j\} \sim \text{multinomial}(n, p_1, \dots, p_J, 1 - \sum_{j=1}^J p_j)$$

$$p_1, \dots, p_J \geq \frac{\theta_F}{J}, \exists j \text{ s.t. } p_j > \frac{\theta_F}{J}$$

With the exact distribution of the test statistics known under the null and alternative hypotheses, we formulate the likelihood ratio test for the p_i . Using the multinomial distribution, the likelihood of the data is:

$$L = \binom{n}{x_1 \dots x_J} p_1^{x_1} \dots p_J^{x_J} (1 - \sum_{j=1}^J p_j)^{(n - \sum_{j=1}^J x_j)}$$

The Likelihood Ratio Test Statistic is $\lambda = \frac{\sup_{H_0} L}{\sup_{H_0 \cup H_A} L}$.

Under H_0 , using Maximum Likelihood Estimation, $\hat{p}_{j0} = \frac{\theta_F}{J}$;

Under $H_0 \cup H_A$, we must maximize L subject to the constraints

$$p_1, \dots, p_J \geq \frac{\theta_F}{J}, p_{J+1} \leq 1 - \theta_F$$

Define the following:

$$I = \left\{ i: \frac{x_i}{n} \leq \frac{\theta_F}{J} \right\} \text{ for } i=1, \dots, J$$

$$J+1 \in I \text{ iff } \frac{x_{J+1}}{n} \geq 1 - \theta_F$$

$$\gamma = \frac{\theta_F}{J} \left(\sum_{i=1}^J 1_{\{i \in I\}} \right) + (1 - \theta_F)^{1_{\{J+1 \in I\}}}$$

Then,

$$\text{for } i=1, \dots, J, \hat{p}_{i_A} = \left(\frac{\theta_F}{J} \right)^{1_{\{i \in I\}}} \left(\frac{x_i}{\sum_{j \notin I} x_j} (1 - \gamma) \right)^{1 - 1_{\{i \in I\}}}$$

$$\hat{p}_{J+1_A} = (1 - \theta_F)^{1_{\{J+1 \in I\}}} \left(\frac{x_{J+1}}{\sum_{j \notin I} x_j} (1 - \gamma) \right)^{1 - 1_{\{J+1 \in I\}}}$$

Note that I holds the i 's whose typical estimate (the bin's sample average) does not follow the boundary constraint or exactly equals the boundary. The maximum likelihood estimates (mle's) of these i will be set to the boundary value. The parameter holds the total probability

from the estimates set to the boundary constraint. If the sample average within a bin does satisfy the constraint (in terms of strict inequalities), it's mle is the sample average of the non-boundary-mle bins, weighted by the total probability that has not yet been assigned (i.e. $1 - \gamma$).

Finally,

$$\lambda = \left(\frac{\theta_F}{J\hat{p}_{1A}} \right)^{x_1} \cdots \left(\frac{\theta_F}{J\hat{p}_{JA}} \right)^{x_J} \left(\frac{1 - \theta_F}{\hat{p}_{jA}} \right)^{(n - \sum_{j=1}^J x_j)} \quad (7)$$

and we reject H_0 when λ is small.

Histogram of lrt.null

Figure 1.
Likelihood ratio test statistic simulated null distribution

Power Results

Figure 2.
Power calculations by prevalence rates

Power Results by J

Figure 3.
Power calculations by different number of bins J

Histogram of lrt.null

Figure 4.
Likelihood ratio test statistic null distribution, COPD dataset

histogram of Fisher p-values, COPD dataset

Figure 5.
Histogram of p-values from COPD dataset

Table 1

Cut-off levels using simulated null distribution

alpha	cutoff
0.10	0.00897
0.05	0.00309
0.01	0.00039

Table 2

Power calculations by prevalence rates

SNPs	Prev=0.05	Prev=0.10	Prev=0.14
60	0.104	0.103	0.103
90	0.152	0.153	0.153
120	0.231	0.237	0.227
150	0.322	0.334	0.328
180	0.442	0.451	0.449
210	0.573	0.582	0.575
240	0.692	0.703	0.699
270	0.806	0.810	0.801
300	0.884	0.890	0.884
330	0.935	0.937	0.939
360	0.968	0.969	0.968
390	0.986	0.986	0.987

Table 3

Power calculations by different number of bins J

SNPs	J=1	J=10	J=20	J=50	J=100	J=150	J=200	J=300
60	0.101	0.107	0.109	0.104	0.098	0.100	0.095	0.060
90	0.137	0.162	0.158	0.171	0.156	0.161	0.148	0.064
120	0.183	0.239	0.248	0.249	0.246	0.247	0.229	0.077
150	0.238	0.345	0.364	0.384	0.378	0.373	0.333	0.078
180	0.307	0.456	0.496	0.533	0.533	0.529	0.439	0.071
210	0.376	0.582	0.633	0.681	0.694	0.684	0.551	0.059
240	0.434	0.712	0.763	0.812	0.814	0.787	0.605	0.044
270	0.515	0.811	0.868	0.902	0.899	0.856	0.606	0.030
300	0.600	0.886	0.919	0.951	0.952	0.871	0.569	0.021
330	0.657	0.940	0.966	0.977	0.967	0.851	0.503	0.015