Cancer Informatics

Libertas Academica
FREEDOM TO RESEARCH

METHODOLOGY

An Efficient Gatekeeper Algorithm for Detecting GxE

Jimmy T. Efird

Biostatistics and Data Management Unit, Center for Health of Vulnerable Populations, University of North Carolina,
237-A McIver Building, Administrative Drive, Greensboro, NC 27402-6170, USA. Email: jimmy.efird@stanfordalumni.org

Abstract: The risk for many complex diseases is believed to be a result of the interactive effects of genetic and environmental factors. Developing efficient techniques to identify gene-environment interactions (GxE) is important for unraveling the etiologic basis of many modern day diseases including cancer. The problem of false positives and false negatives continues to pose significant roadblocks to detecting GxE and informing targeted public health screening and intervention. A heuristic gatekeeper method is presented to guide the selection of single nucleotide polymorphisms (SNPs) in the design phase of a GxE study.

Keywords: gene-environment interaction, multiplicity corrected confidence intervals, SNP microarrays

Cancer Informatics 2010:9 115–120

This article is available from http://www.la-press.com.

Introduction

Advances in bioinformatics and genomics have opened the door for personalized medicine, enabling epidemiologists to identify genetic variations that predispose some, but not others, to disease. Single nucleotide polymorphisms (SNPs), which on average occur in about every 1,000 base pairs throughout the human genome, also may be useful for determining variability in individual response to treatment and could potentially lead to the development of novel therapeutics custom tailored to patients' genetic profiles. However, studies often have failed to yield consistent findings or definitive results, in part because analyses have not accounted for gene-environment interactions (GxE).

While genetics play a significant role in many diseases, few common medical disorders are explained by a single SNP or genetic mutation. Rather, environmental factors are thought to modulate an individual's genetic predisposition for certain diseases.[1–3] In the case of GxE, risk for disease occurs only when genetic and environmental factors are present in combination, while individual factors alone convey little or no risk for disease. Correctly identifying GxE is particularly difficult in the context of high density SNP arrays, because the number of multiple comparisons can be in the thousands.

In this paper, an efficient gatekeeper algorithm is presented to identify GxE. The technique involves computing a multiplicity adjusted lower bound on an indirect estimate for GxE. The indirect estimate is then used to independently screen for GxE in a direct disease association study in order to correctly identify risk or propensity for disease.

Methodology

The method for indirectly estimating the odds ratio (OR) for GxE from a case-control study and the technique for computing multiplicity corrected confidence intervals for a relative effects estimate have been separately described in previous publications and are only briefly summarized below.[4,5] Their combination forms the basis for the procedure to screen for GxE.

Indirect OR estimate and confidence interval (CI) for GxE

The odds ratio (OR) for environmental exposure (E) associated with disease (D) in the population may be expressed as

$$\mathrm{OR(E|D)} = \frac{\mathrm{P(E|D)}/\mathrm{P(\bar{E}|D)}}{\mathrm{P(E|\bar{D})}/\mathrm{P(\bar{E}|\bar{D})}}. \qquad (1)$$

Assuming that D is relatively rare in both exposed and unexposed populations, and that genotype (G) is independent of environmental exposure (E) [i.e., $\mathrm{P(G|E)=P(G|\bar{E})=g}$], equation number (1) simplifies to

$$\frac{\left[\mathrm{P(D|GE)g} \Big/ \mathrm{P(D|\bar{G}\bar{E})}\right] + \left[\mathrm{P(D|\bar{G}E)(1-g)} \Big/ \mathrm{P(D|\bar{G}\bar{E})}\right]}{\left[\mathrm{P(D|G\bar{E})g} \Big/ \mathrm{P(D|\bar{G}\bar{E})}\right] + (1-g)}. \qquad (2)$$

Considering the simple case when $\mathrm{OR(G\bar{E}|D)} = \mathrm{OR(\bar{G}E|D)} = 1$, the OR for GxE given disease may be written as

$$\mathrm{OR(GE|D)} = [\mathrm{OR(E|D)} - 1 + \mathrm{g}]/\mathrm{g}, \qquad (3)$$

where $[\mathrm{OR(E|D)}| -1] \geq \mathrm{g}$ by unity constraints on the joint conditional probabilities. Treating (g) as fixed, the $(1-\alpha/2) \times 100\%$ CI for OR(GE|D) is approximately equal to

$$\exp\left\{ \log\{[\mathrm{OR(E|D)} - 1 + \mathrm{g}]/\mathrm{g}\} \pm z_{(1-\alpha/2)} \frac{\mathrm{OR(E|D)}\sqrt{\left(\mathrm{var}\left(\mathrm{OR(E|D)}\right)\right)}}{\{\mathrm{OR(E|D)} - 1 + \mathrm{g}\}} \right\}, \qquad (4)$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2) \times 100$ percentile of a standard normal distribution.

Multiplicity corrected CIs

Given a set of (i) SNPs, the P value corresponding to the statistical significance of $\mathrm{OR(GE|D)}_i$ is computed as

$$\mathrm{p}_i = 2\left[1 - \Phi_{(z_i)}\right], \qquad (5)$$

where

$$\Phi\left(z_i\right) = \int_{-\infty}^{z_i} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx, \qquad (6)$$

$$z_i = \left|\frac{\log[\mathrm{OR(GE|D)}_i]}{\mathrm{SE}[\log[\mathrm{OR(GE|D)}_i]]}\right| = \left|\frac{\log[\mathrm{OR(GE|D)}_i]}{-\{\log[\mathrm{LCI}[\mathrm{OR(GE|D)}_i] - \log[\mathrm{OR(GE|D)}_i]\}/Z_{(1-\alpha/2)}}\right|, \qquad (7)$$

and LCI is the (1–α/2) × 100% lower CI from equation 4. Ordering the P values (p_i's) from the lowest to highest values, i.e., $p_{(1)} \leq p_{(2)} \leq \ldots p_{(i)} \leq \ldots p_{(n)}$ (with arbitrary ordering in the case of ties), the multiplicity corrected P values denoted by "*" are computed as

$$p^*_j = p_{(j)}(n-j+1), \qquad (8)$$

where j ranges from 1 to n in a 1:1 identity mapping with the i values, and $p^*_{(j)}$ is bounded by unity. The multiplicity corrected (1–α/2) × 100% CI for $\mathrm{OR(GE|D)}_{(i)}$ is then computed as

$$\mathrm{CI}_{(1-\alpha/2)} = \exp\left\{\log(\mathrm{OR}(\mathrm{GE}\,|\,\mathrm{D}_i) \pm Z_{(1-\alpha/2)}\mathrm{SE}^*[\log(\mathrm{OR}(\mathrm{GE}\,|\,\mathrm{D})_i)]\right\}, \qquad (9)$$

where

$$\mathrm{SE}^*\left[\log\left(\mathrm{OR}(\mathrm{GE}\,|\,\mathrm{D})_i\right)\right] = \frac{\log\left(\mathrm{OR(GE|D)}_i\right)}{\Phi^{-1}\left[1-\frac{p^*_{(j)}}{2}\right]}. \qquad (10)$$

Screening for GxE in a direct association study

A SNP will be selected as a possible candidate in a direct disease association study for GxE if the (1–α/2) × 100% multiplicity corrected lower CI (MCLCI) estimate for OR(GE|D) is greater than an *a priori* specified threshold value, for example, OR = 3.0. Letting α_1 and α_2 denote the type I error for the indirect and direct tests, the statistical significance of the overall procedure will be protected at $\alpha \leq \alpha_1 + \alpha_2$, where the upper bound holds under independence of the indirect and direct tests. The significance level of the likelihood ratio test for the global null hypothesis $\boldsymbol{\beta} = 0$, where $\boldsymbol{\beta}$ denotes the vector of β coefficients in a direct multivariable logistic regression model, may be set to a nominal value (e.g., ≤0.001), such that the significance level for the overall procedure (indirect and direct combined) will be protected at an α-level only slightly greater than the type I error for the indirect test. Furthermore, the total number of SNPs allowed to enter the direct model may be fixed at a small number, for example ≤10, based on the rank order of the lower 95% CIs for SNPs passing the OR threshold value.

In practice, a significant gain in power may be realized by using a meta-analysis estimate for OR(E|D) and 95% CI. Because a meta-analysis combines several studies, the resulting confidence interval will be more precise than a single case-control estimate of the effect.

Example (hypothetical)

A population-based meta-analysis of several case-control studies estimates that children living on a farm have a 1.5-fold OR(E|D) (95% CI = 1.1676 – 1.9270) for childhood brain cancer compared with controls. The aim of a future association study is to determine whether GxE are occurring between a panel of 100 innate immunity SNPs and exposure to farm life. The study investigator is interested in finding interactions with an OR(GE|D) ≥ 3.0. The population allele frequencies (g) for the 100 SNPs and computed 95% MCLCI for the indirect estimates of OR(GE|D) are shown in Table 1. Upon examining Table 1, the investigator observes that 7 SNPs (highlighted in gray in the 3 rightmost columns) have a MCLCI ≥ 3.0 for the indirect estimate of OR(GE|D) and these will be included in a new association study to directly test for GxE. Assuming that the type I error rate for the direct test will be controlled at $\alpha_2 = 0.001$, the statistical significance of the overall procedure will be protected at $\alpha \leq 0.051$.

Table 1. Multiplicity corrected 95% lower confidence intervals (LCI) for OR(GE|D) given the population allele frequency (g) for 100 innate immunity SNPs and OR(E|D) = 1.5 (95% LCI = 1.1676).

g*	OR (GE|D)	Multiplicity corrected 95% LCI	g*	OR (GE|D)	Multiplicity corrected 95% LCI	g*	OR (GE|D)	Multiplicity corrected 95% LCI	g*	OR (GE|D)	Multiplicity corrected 95% LCI
74.0	1.68	1.24	49.4	2.01	1.21	39.4	2.27	1.25	26.8	2.87	1.48
73.0	1.68	1.22	49.0	2.02	1.21	39.0	2.28	1.26	25.8	2.94	1.51
72.0	1.69	1.21	48.6	2.03	1.21	38.6	2.30	1.26	24.8	3.02	1.54
71.0	1.70	1.20	48.2	2.04	1.21	38.2	2.31	1.27	23.8	3.10	1.58
70.0	1.71	1.19	47.8	2.05	1.21	37.8	2.32	1.27	22.8	3.19	1.62
69.0	1.72	1.19	47.4	2.05	1.21	37.4	2.34	1.27	21.8	3.29	1.66
68.0	1.74	1.18	47.0	2.06	1.21	37.0	2.35	1.28	20.8	3.40	1.71
67.0	1.75	1.18	46.6	2.07	1.21	36.6	2.37	1.28	19.8	3.53	1.76
66.0	1.76	1.18	46.2	2.08	1.21	36.2	2.38	1.29	18.8	3.66	1.82
65.0	1.77	1.18	45.8	2.09	1.21	35.8	2.40	1.29	17.8	3.81	1.89
64.0	1.78	1.17	45.4	2.10	1.21	35.4	2.41	1.30	16.8	3.98	1.96
63.0	1.79	1.17	45.0	2.11	1.22	35.0	2.43	1.30	15.8	4.16	2.04
62.0	1.81	1.17	44.6	2.12	1.22	34.6	2.45	1.31	14.8	4.38	2.14
61.0	1.82	1.17	44.2	2.13	1.22	34.2	2.46	1.32	13.8	4.62	2.25
60.0	1.83	1.17	43.8	2.14	1.22	33.8	2.48	1.32	12.8	4.91	2.38
59.0	1.85	1.18	43.4	2.15	1.22	33.4	2.50	1.33	11.8	5.24	2.53
58.0	1.86	1.18	43.0	2.16	1.23	33.0	2.52	1.33	10.8	5.63	2.70
57.0	1.88	1.18	42.6	2.17	1.23	32.6	2.53	1.34	9.8	6.10	2.92
56.0	1.89	1.18	42.2	2.18	1.23	32.2	2.55	1.35	8.8	6.68	3.18
55.0	1.91	1.18	41.8	2.20	1.23	31.8	2.57	1.35	7.8	7.41	3.51
54.0	1.93	1.19	41.4	2.21	1.24	31.4	2.59	1.36	6.8	8.35	3.94
53.0	1.94	1.19	41.0	2.22	1.24	31.0	2.61	1.37	5.8	9.62	4.52
52.0	1.96	1.19	40.6	2.23	1.24	29.8	2.68	1.40	4.8	11.42	5.34
51.0	1.98	1.20	40.2	2.24	1.25	28.8	2.74	1.42	3.8	14.16	6.60
49.8	2.00	1.20	39.8	2.26	1.25	27.8	2.80	1.45	2.8	18.86	8.80

Notes: The lightly highlighted area in the lower right hand corner of the table denotes the 7 SNPs that have multiplicity corrected 95% lower confidence intervals (MCLCI) exceeding the *a priori* specified threshold value of 3.0. The value 3.18 in the darkly highlighted area in the upper right hand corner of the above region corresponds to the MCLCI of the minimum SNP passing the threshold for entrance into the direct model. This value is used to conservatively estimate power, as higher values in the same column beneath 3.18 will yield smaller sample size estimates.
*Expressed as a percentage.

Power and sample size computation

Power and sample size for the direct study may be computed using standard maximum likelihood methods for a logistic regression model and setting the α-level equal to α_2.[6] When the joint distribution of covariates is unknown, the sample size for a multivariable model may be estimated by multiplying the univariate result times a variance inflation factor $1/(1-\rho^2_{1.2,3\ldots p})$, where $\rho^2_{1.2,3\ldots p}$ denotes the squared multiple correlation coefficient and p equals the number of model covariates.[7] In the example above, approximately $n_1 = 260$ cases and $n_2 = 260$ controls are needed in a direct study to have at least 80% power to detect an OR(GE|D) ≥ 3.18 (corresponding to the LCI of the minimum SNP passing the threshold for entrance into the direct model; upper right hand value in highlighted region of Table 1) at the $\alpha_2 = 0.001$ level of statistical significance (2-sided test), given $\rho^2_{1.2,3\ldots p} = 0.2$, P(E) = 0.10 and P(GE) = (0.88)(0.10) = 0.088.[8] Accordingly, the overall test procedure is protected at $\alpha \le 0.051$ (i.e., $\alpha_1 + \alpha_2 = 0.05 + 0.001 = 0.051$).

For models involving non-null main effects [i.e., $P(G=1, E=0) \neq 1$ and/or $P(G=0, E=1) \neq 1$], power may be computed by excluding these effects from the sample space when estimating the power for OR(GE|D).

Discussion

To the best of our knowledge, this method is the first to screen for GxE using an indirect estimate for OR(GE|D). Statistical power is significantly increased in this approach by eliminating SNPs prior to conducting a direct association study of GxE. Furthermore, study cost is greatly reduced since fewer SNPs need to be genotyped.

The approach has several advantages. For example, the derived indirect estimate requires only knowledge of the OR for environmental exposure OR(E|D) and the population allele frequency (g). Also, the model can detect interactions even when the OR for an environmental effect is null, i.e., OR(E|D) = 1.

An indirect estimate for OR(GE|D) can be computed regardless of whether a biologic rationale exists for the underlying effect. Since the inclusion of biologically irrelevant or non-functional SNPs will inflate type I error, pathway analysis and other molecular techniques are recommended to determine the relevance of SNPs prior to analysis.[9]

The face validity of the method is based on established probabilistic principles and theory.[10–12] Nonetheless, further validation of the technique will require testing its ability to detect biologically and clinically meaningful results that hold under replication in future independent studies. Furthermore, since the multiplicity corrected CIs used in the indirect screening phase of the method were derived heuristically and represent approximate estimates of the true interval widths, re-sampling methods are recommended in situations requiring exact coverage.

A practical limitation of the method is that genotype must be independent of environmental exposure. This assumption may be violated, for example, when an underlying gene affects behavior such that an individual is predisposed to seek (or avoid) the environmental exposure (e.g., a gene that causes craving for alcohol). Additionally, the method does not account for complex gene-environment interactions that may underlie multifactorial diseases.

An implicit assumption of the method is that estimates for OR(E|D) and (g) remain unchanged in the population under consideration in the direct association study. However, this may not hold true when samples are collected based on strict population stratification or the target population has changed over time. Accordingly, a prudent comparison of known epidemiologic characteristics for the indirect and direct populations is advised prior to the implementation of this method. Additionally, the user must use caution in the interpretation of results when the decimal precision of estimates are limited.

When the allelic frequency of SNPs is very low, the multiplicity adjusted P values will approach zero. To remedy this limitation, a GMP-based implementation of the Schonhage-Strassen algorithm may be used to perform arbitrary-precision arithmetic.[13,14] This algorithm uses fast Fourier transforms in rings with 2^{2n+1} elements to enable multiplicative computation of factors near absolute zero.

The ultimate success of detecting GxE will depend on the accurate and precise measurement of environmental exposures on par with recent advances in genotyping technology. For many diseases, this will entail determining life-course environmental exposures from birth onward.[1] Parsimonious questionnaire design and the use of targeted biomarkers will play a key role in assessing environmental exposures in the context of GxE.

In summary, the failure to account for multiplicity in large scale GxE studies may lead to the misinterpretation of results. Furthermore, disease association studies for GxE are expensive and time consuming, and careful control of these factors is important to consider in study design.[15] The method presented in this paper offers an easy to implement and efficient means to identify GxE that will provide more efficacious use of research and clinical resources.

Acknowledgements

This manuscript was made possible by a grant from NCMHD/NIH (P20MD002289) entitled “Teamwork in Research and Intervention to Alleviate Disparities Project (TRIAD).” Elizabeth Tornquist (UNC-CH) and Debra C. Wallace (UNCG) offered valuable comments during the writing of this manuscript and their knowledge and insight are greatly appreciated.

Disclosures

This manuscript has been read and approved by the author. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The author and peer reviewers of this paper report no conflicts of interest. The author confirms that they have permission to reproduce any copyrighted material.

Abbreviations

CI, Confidence interval; GxE, gene-environment interaction; LCI, lower confidence interval; MCLCI, multiplicity corrected lower confidence interval; OR, odds ratio; SNP, single nucleotide polymorphism; SE, standard error.

References

1. Wild P. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2005;14:1847–50.
2. von Mutius E. Gene-environment interactions in asthma. *J Allergy Clin Immunol*. 2009;123:3–11.
3. Vercelli D. Gene-environment interactions: The road less traveled by in asthma genetics. *J Allergy Clin Immunol*. 2009;123:26–7.
4. Efird J. A method for indirectly estimating gene-environment effect modification and power given only genotype frequency and odds ratio of environmental exposure. *Eur J Epidemiol*. 2005;20:389–93.
5. Efird J, Searles Nielsen S. A method to compute multiplicity corrected confidence intervals for odds ratios and other relative effect estimates. *Int J Environ Res Public Health*. 2008;5:394–8.
6. Lyles R, Lin H, Williamson J. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Stat Med*. 2007;26:1632–48.
7. Whittemore A. Sample size for logistic regression with small response probability. *J Am Stat Assoc*. 1981;76:27–32.
8. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav Res Methods*. 2009;41:1149–60.
9. Jorgensen T, Ruczinski I, Kessing B, Smith M, Shugart Y, Alberg A. Hypothesis-driven candidate gene association studies: practical design and analytical considerations. *Am J Epidemiol*. 2009;170:986–93.
10. Roy S, Bose R. Simultaneous confidence interval estimation. *Ann of Math Stat*. 1953;24:513–36.
11. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75:800–2.
12. Marcus R, Peritz E, Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63:655–60.
13. Schonhage A, Strassen V. Schnelle multiplication grosser zahlen. *Computing*. 1971;7:281–92.
14. Gaudry P, Kruppa A, Zimmermann P. A GMP-based implementation of Schonhage-Strassen's large integer multiplication algorithm. Proceedings of the 2007 International Symposium on Symbolic and Algebraic Comutation, pp. 167–74.
15. Li C, Li M, Long J, Cai Q, Zheng W. Evaluating cost efficiency of SNP chips in genome-wide association studies. *Genet Epidemiol*. 2008;32:387–95.