Published in final edited form as:

Genet Epidemiol. 2013 September ; 37(6): 551–559. doi:10.1002/gepi.21741.

Empirical Hierarchical Bayes Approach to Gene-Environment Interactions: Development and Application to Genome-Wide Association Studies of Lung Cancer in TRICL

Melanie Sohns^{#1}, Elena Viktorova^{#1}, Christopher I. Amos², Paul Brennan³, Gord Fehrer⁴, Valerie Gaborieau³, Younghun Han², Joachim Heinrich⁵, Jenny Chang-Claude⁶, Rayjean J. Hung⁴, Martina Müller-Nurasyid^{7,8}, Angela Risch⁹, Duncan Thomas¹⁰, and Heike Bickeböllner¹

¹ Department of Genetic Epidemiology, University Medical Center, Georg-August University of Göttingen, Göttingen, Germany

² Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA

³ International Agency for Research on Cancer (IARC), Lyon, France

⁴ Prosserman Centre for Health Research, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

⁵ Institute of Epidemiology I, Helmholtz Zentrum München, Neuherberg, Germany

⁶ Unit of Genetic Epidemiology, Division of Cancer Epidemiology, Deutsches Krebsforschungszentrum, Heidelberg, Germany

⁷ Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology and Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany

⁸ Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

⁹ Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center, Heidelberg, Germany

¹⁰ Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

[#] These authors contributed equally to this work.

Abstract

The analysis of gene-environment (GxE) interactions remains one of the greatest challenges in the post-genome-wide-association-studies (GWAS) era. Recent methods constitute a compromise between the robust but underpowered case-control and powerful case-only methods. Inferences of the latter are biased when the assumption of gene-environment (G-E) independence fails. We propose a novel empirical hierarchical Bayes approach to GxE interaction (EHB-GE), which

benefits from greater power while accounting for population-based G-E dependence. Building on Lewinger et al.'s ([2007] *Genet Epidemiol* 31:871-882) hierarchical Bayes prioritization approach, the method utilizes posterior G-E association estimates in controls based on G-E information across the genome to adjust for it in resulting test statistics. These *posteriori* estimates are subtracted from the corresponding G-E association coefficients within cases.

We compared EHB-GE with rival methods using simulation. EHB-GE has similar or greater rank power to detect GxE interactions in the presence of large numbers of G-E associations with weak to strong effects or only a low number of such associations with large effect. When there are no or only a few weak G-E associations, Murcray et al.'s method ([2009] *Am J Epidemiol* 169:219-226) identifies markers with low GxE interaction effects better. We applied EHB-GE and competing methods to four lung cancer case-control GWAS from the TRICL/ILCCO consortium with smoking as environmental factor. Genes identified by the EHB-GE approach are reasonable candidates, suggesting usefulness of the method.

Keywords

population G-E association; GWAS; rank power; lung cancer

INTRODUCTION

Genome-wide association studies (GWAS) have identified thousands of genetic risk factors for complex diseases. However, identified variants only explain a part of the total heritability. Unexplained phenotypic variance may be partly due to undetected GxE interactions, which are characterized by a complex interplay of genetic and environmental contributors to the disease. The most effective way to detect GxE interactions in GWAS remains unresolved.

The standard case-control (CC) logistic regression tests for GxE interactions as departure from a multiplicative odds ratio model. Generally, it is underpowered to detect GxE interactions unless genetic and environmental factors are common and the interaction effect is strong [Mukherjee et al., 2012]. Unconditional logistic regression as commonly applied to case-control data does not account for preferential sampling through the cases, which contain most of the information on susceptibility to the disease, and hence underweights the more informative component of the sample. However, unconditional logistic regression is robust to gene-environment association.

The case-only test (CO) [Piegorsch et al., 1994] is a more powerful alternative under the assumptions of a rare disease and G-E independence in the source population. It does not maintain a pre-specified level of Type I error when G-E independence is violated [Albert et al., 2001; Chatterjee and Carroll, 2005; Satten and Epstein, 2004]. Therefore, it cannot be recommended for GWAS, in which G-E association cannot be ruled out but is expected to occur.

The intuitive two-step approach (TWO) first tests all SNPs in controls for G-E independence; based on the result, it employs the case-only or case-control method to test

for GxE interaction in the second step. For a study of moderate size, the power to detect G-E dependence is typically low and the approach may still remain significantly biased [Mukherjee and Chatterjee, 2008]. In addition, a procedure failing to account for the two-step uncertainty demonstrates an inflated level of Type I error even when G-E independence is given [Albert et al., 2001]. The correlation between the two steps leads to an inflated Type I error rate overall.

The two-step approach (MUR) by Murcray et al. [2009] combines the power of the case-only test with the protection from bias of the case-control estimator in a two-step procedure with the test statistics being independent from each other. The method screens for G-E association in the whole sample. Then, only SNPs exceeding a given significance threshold are tested for GxE interaction. Unfortunately, the power of the first step highly depends on the case-control ratio (*ccr*). A large number of controls compared to cases leads to decreasing power of step one and hence loss of power in the overall procedure [Murcray et al., 2011; Kraft et al., 2007].

The empirical Bayes-type shrinkage estimator (MUK-EB) [Mukherjee et al., 2008; Mukherjee and Chatterjee, 2008] combines the robust case-control with the efficient case-only estimator assuming G-E independence. This estimator is approximately robust to the presence of G-E association in the source population and performs comparably to the case-control estimator under large departures from independence. The method does not strictly maintain the correct Type I error rate under G-E dependence and moderate sample size. However, it maintains smaller mean squared error compared to the other estimators mentioned, irrespective of the G-E association.

The aim of this paper is to introduce a novel method to search for GxE interactions in a GWAS, which has the high power of the case-only design for testing GxE interactions, while accounting for population G-E association. We term it the empirical hierarchical Bayes method for GxE interactions (EHB-GE). We compared our method with previously discussed alternative estimators of GxE interaction in a simulation study as well as in real data applications.

METHODS

The Empirical Hierarchical Bayes Method for GxE Interactions (EHB-GE)

Based on the hierarchical mixture model described by Lewinger et al. [2007] we developed a novel empirical hierarchical Bayes method to detect GxE interactions in GWAS. Similarly to Mukherjee et al. [2008], we propose a compromise between the case-control and case-only test of GxE interaction, by subtracting a posterior estimate of the G-E association within controls from that within cases. For an individual, let the indicators D , G and E denote presence (e.g. $D = 1$) or absence (e.g. $D = 0$) of disease, minor allele (i.e. carriers) of the considered SNP, or environment, respectively. Estimates of G-E association within cases (likewise for controls with $\beta_{controls}$) can be obtained from the following logistic regression model

$$\text{logit}(P(E=1|G, D=1)) = \alpha_{cases} + \beta_{cases}G$$

Under the rare disease assumption and G-E independence in the source population, $\beta_{controls} = 0$. Then testing for $\beta_{cases} = 0$ provides a valid case-only test for GxE interaction. However, when such assumptions are not true, the $\beta_{controls}$ coefficients should be estimated and subtracted from the β_{cases} coefficients to result in an unbiased estimate of GxE interaction. We propose the following novel hierarchical test statistic for GxE interaction in GWAS (calculated for each SNP)

$$T_M^{EHB-GE} = \frac{\hat{\beta}_M^{cases} - \text{sgn}_M \times \text{posterior}(\hat{\beta}_M^{controls})}{\sqrt{(\hat{\sigma}_M^{cases})^2 + \text{posterior}((\hat{\sigma}_M^{controls})^2)}} = \frac{\hat{\beta}_M^{cases} - \text{sgn}_M \hat{\lambda}_M}{\sqrt{(\hat{\sigma}_M^{cases})^2 + \text{Var}[\lambda_M | \hat{\beta}_M^{controls}]}} ,$$

where $M = 1, \dots, N$, are the markers (with total number N), $\beta_M^{cases}, \beta_M^{controls}$ are the corresponding logistic regression coefficients for G-E association among cases or controls, respectively, with standard deviations $\sigma_M^{cases}, \sigma_M^{controls}$, $\hat{\lambda}_M$ are the posterior estimators of $\hat{\beta}_M^{controls}$, and $\text{sgn}_M = \text{sign}(\hat{\beta}_M^{controls})$. The corresponding test statistics to test for the association within cases (likewise for controls) are $T_M^{cases} = \hat{\beta}_M^{cases} / \hat{\sigma}_M^{cases}$. For G-E associated SNPs $\hat{\beta}_M^{cases} = \hat{\beta}_M^{controls} \neq 0$, and possibly $\hat{\sigma}_M^{cases} \neq \hat{\sigma}_M^{controls}$. The latter can particularly occur in the presence of an environmental main effect or different numbers of cases and controls in the sample. Unequal expected values of the test statistics result in $(T_M^{cases} - T_M^{controls})$ or $(T_M^{cases} - \text{sgn}_M \hat{\lambda}_M)$ not being distributed around 0, even if the null hypothesis is true. Thus, association may be misinterpreted as interaction. To overcome this, we use a hierarchical Bayes model for the estimated effects $\beta_M^{controls}$ and calculate $\hat{\beta}_M^{cases} - \text{sgn}_M \hat{\lambda}_M$. The model is given by

$$\begin{array}{ll} \text{Level 1} & |\hat{\beta}_m^{controls}| | \lambda_M \sim \hat{\sigma}_{M_i}^{controls} \chi_1(\lambda_M), \\ \text{Level 2} & \lambda_M | \theta, \sigma, p \sim p \sigma \chi_1(\theta) + (1-p) \delta(0), \end{array}$$

where λ_M are noncentrality parameters of the χ distribution with one degree of freedom ($\chi_1(\lambda_M)$) and p is the prior probability that marker M is associated with the disease, i.e. $\lambda_M > 0$. Given $\lambda_M > 0$, λ_M is assumed to have a χ_1 distribution with noncentrality parameter θ as the measure of association strength and a scaling parameter $\sigma > 0$. Given $\lambda_M = 0$, $\delta(0)$ denotes a point mass at zero. The probability density function, the *prior* probability, marginal distribution, and *posterior* expected values are derived in **Appendix A** using an approximation by Kass and Steffey [1989] for the variance. Based on the maximum

likelihood estimates (MLE) $\hat{\Theta} = (\hat{\theta}, \hat{\sigma}, \hat{p})$, the posterior expectation of noncentrality parameter $(\hat{\lambda}_M)$ and variance of λ_M are

$$\hat{\lambda}_M = E \left[\lambda_M | \hat{\beta}_m^{controls} | \hat{\theta}, \hat{\sigma}, \hat{p} \right],$$

$$Var \left[\lambda_M | \hat{\beta}_M^{controls} \right] = E \left[Var \left[\lambda_M | \hat{\beta}_M^{controls}, \Theta \right] \right] + Var \left[E \left[\lambda_M | \hat{\beta}_M^{controls}, \Theta \right] \right].$$

Simulation Study

In 1000 replications of 3000 individuals each (case-control ratio $ccr = 1:1, 1:2$ and $2:1$) we generated disease as affected/unaffected, environment as exposed/unexposed and SNP genotypes as carrier/non-carrier (all binary) by logistic regression as $\text{logit}(P(D = 1|E, G)) = p_d + \beta_{G \times E} G \times E$, $\beta_{G \times E} = \log(OR_{G \times E})$, where disease prevalence is $p_d = 0.01, 0.5, 0.1$. We simulated 10,000 SNPs in total. A *single*, so-called “*interacting*” SNP with $OR_{G \times E} = 1.2, 1.5, 2, 2.5$, and 3 was simulated. Additionally, up to 1000 so-called “G-E” SNPs ($N_{G-E} = 0, 1, 5, 10, 25, 50, 75, 100, 200, \dots, 1000$) with low, medium, and high regression coefficients for G-E association were simulated from a normal distribution $N(0, \log(1.5)/2)$ and $N(0.7, 0.1)$ for low and medium associations, respectively, and for high associations for the log of the regression coefficients from $N(0, \log(1.5)/2)$. We named the other SNPs with no true G-E interaction or association “dummy” SNPs. For the interacting SNP the minor allele frequency (MAF) was $p_g = 0.1, 0.3, 0.5$. For all other SNPs, the MAF was sampled from a Beta distribution $B(1, 3)$ truncated to $[0.01, 0.5]$ and the exposure frequency was $p_e = 0.1, 0.3, 0.5$. The frequency of carriers of the minor allele was derived under the assumption of Hardy-Weinberg equilibrium.

We compared EHB-GE and competing methods (CC, CO, TWO, MUK-EB and MUR) by the “**rank power**”, defined as percentage of simulated replicates where the true interacting SNP is within the top ranking positions, according to absolute value of the corresponding test statistics (top 1, 10, 25, 50 or 100). We plotted rank power of EHB-GE on the *x-axis* against the difference in rank power between EHB-GE and a competitor on the *y-axis*. Hence, positive values on the *y-axis* represent a gain in rank power by EHB-GE. We evaluated rank power of the methods because, especially in the context of GxE interaction, GWAS is often regarded as a screening process, selecting a subset of top SNPs for further investigation [Hirschhorn, 2005; Thomas et al., 2009; Ziegler et al., 2008]. Rank power addresses this issue, evaluating the ability of a particular method to detect the true positive effect for follow-up, even in the situations in which moderate to small effect size does not reach genome-wide significance level.

RESULTS

Evaluation of EHB-GE through Simulation

We first note that, as necessary, rank power for EHB-GE increased with increasing GxE effect size, increasing frequencies p_g, p_e and decreasing p_d . The rank power for different *ccrs* depended heavily on the number of G-E associations and their effect size. **Figure 1**

illustrates different *ccrs* and different G-E association numbers and effect sizes for $OR_{G \times E} = 2$, $p_g = 0.3$, $p_e = 0.1$, $p_d = 0.5$. Overall $ccr = 1:1$ was the most preferable and $ccr = 1:2$ the most unfavorable scenario. When only low G-E associations were present (left figure), a larger number of cases than controls was the best. However, for a large number of medium or high G-E associations, the rank power of this unbalanced scenario decreased. The balanced $ccr = 1:1$ is the best ratio in situations in which a large number of G-E associations in the sample are suspected independently of the G-E association effect size. Although the difference in rank power between the *ccrs* was larger for larger $OR_{G \times E}$, these tendencies hold when varying the interaction effect size. True GxE interaction is easier to detect for a more frequent environmental factor, but this is counterbalanced by a greater probability of finding a G-E association, which can obscure the true GxE interaction. For $ccr = 1:1$ and low $p_e = 0.1$, the scenario with low G-E association effect attained the highest power, and for common exposure, $p_e = 0.3$ EHB-GE method even reached the highest power for high association effects (data not shown).

Comparison of EHB-GE with other GxE interaction methods

We compared the rank power of EHB-GE with that of five other methods (CC, CO, TWO, MUK-EB and MUR) (see **Table 1**). In all scenarios without G-E associations (“none”), the method of Murcay et al. [2009] yielded 100% rank power. The power of EHB-GE for these scenarios was nearly the same as for the case-only test. On average MUK-EB and TWO lost around 5 – 10% rank power compared to EHB-GE. If G-E associations were present, EHB-GE always outperformed CC. The performance of MUR depended heavily on the number of population G-E associations and their strength. When they increased, the rank power of MUR decreased rapidly. MUR outperformed others, with nearly 100% rank power when only a small number, e.g. 100, of G-E associations with low effect appeared in the population. The power of EHB-GE and CO was very similar in the absence or presence of few G-E associations with low effects. With increasing number or strength of the G-E associations, EHB-GE retained the power level that lowered for the CO test. Results for $ccr = 1:2$ and $2:1$ are similar (data not shown).

Subsequently for overall comparison of simulated scenarios, we plotted the rank power of EHB-GE versus the difference in rank power of EHB-GE and each of the methods (**Figure 2** illustrates these plots for rank power regarding the top 25 ranks and $ccr = 1:1$). Our EHB-GE approach demonstrated remarkably greater rank power in almost all simulation scenarios compared to the CC test (**Figure 2a**). The superiority of EHB-GE compared to CO could be confirmed (**Figure 2b**). A clear triangular structure was observed, in which the diagonal consists of the high association scenarios. Irrespective of the choice of other parameters, CO has almost no rank power in these situations. The points along the horizontal line represent the similarity of CO and EHB-GE given a low number and low strength of G-E associations. The vertical line represents situations in which EHB-GE reached nearly 100% rank power and CO between 0 and 100%. **Figure 2c** demonstrates the comparison between TWO and EHB-GE. TWO sometimes outperforms EHB-GE in the case of moderate number of G-E associations with medium to high effect size. EHB-GE tends to outperform TWO in the case of low number of such effects. EHB-GE also reached higher power in the presence of G-E associations with low effect size irrespective of their number. The EHB-GE method was

remarkably better for almost all the parameter combinations than MUK-EB (**Figure 2d**), except for some cases of medium G-E association effect. Finally, the rank power of MUR is higher compared to EHB-GE when there are only a small number of G-E associations with low effect size. However, for a larger number of higher G-E association effects, EHB-GE reaches up to 50% more rank power. When there are G-E associations with medium effect size, irrespective of their number, EHB-GE always outperformed MUR for $ccr = 1:1$ (**Figure 2e**). The same behavior of these competing tests in terms of the rank power compared to EHB-GE for the top 50 or 100 SNPs and in the presence of an environmental main effect is observed (data not shown).

Application to ILCCO/TRICL GWAS

We applied EHB-GE and the other competing approaches to four lung cancer case-control GWASs from the International Lung Cancer Consortium (ILCCO) / Interdisciplinary Research in Cancer of the Lung (TRICL) consortium. Four participating case-control GWAS were analyzed including the German lung cancer GWAS (GLC) [Holle et al., 2005; Sauter et al., 2008; Wichmann et al., 2005], the Texas lung cancer GWAS (MD Anderson Cancer Center, MDACC) [Amos et al., 2008; Hung et al., 2008; Wang et al., 2008], the Toronto lung cancer GWAS (Samuel Lunenfeld Research Institute, SLRI) [Hung et al., 2008a; Hung et al., 2008b] and the Central Europe lung cancer GWAS (Central Europe IARC, CE-IARC) [Amos et al., 2008]. Some characteristics of all datasets are given in **Table 2**.

Since tobacco smoking is clearly the main risk factor in lung cancer development, we investigated GxE interactions with smoking in all four GWAS. We defined smoking status of a particular individual by two different approaches, i.e., *ever* = 1 and *never* = 0 smokers or *heavy* = 1 and *moderate* = 0 smokers. *Never* smokers were defined as those individuals who consumed no more than 100 cigarettes during their life-time. All others were called *ever* smokers. The MDACC study collected only *ever* smokers. We defined *moderate* smokers as those consuming fewer than or equal to 20 pack-years and *heavy* smokers as those smoking more than 20 pack-years, where a pack year is calculated as cigarettes per day $\times$ years smoked divided by the number of cigarettes (20) per pack [Amos et al., 2008]. Two models with two different smoking variables were applied to each of the GWAS. In the following results, we will use the abbreviations NE for the *never* versus *ever* smokers model and MH for the *moderate* versus *heavy* model. To guarantee high quality of the data, standard quality control procedures were performed, mainly using PLINK and EIGENSOFT [Price et al., 2006].

For each of the four datasets, the NE and MH models were fitted and each of the competing methods applied. The classical case-control test for GxE interaction did not identify any SNP markers with a genome-wide significant interaction effect, i.e. with $p\text{-value} < 10^{-8}$ in any of the four GWAS. The EHB-GE and case-only test identified one SNP, rs13244987, using the NE model in GLC with genome-wide significant p-value for interaction and one neighboring SNP, rs13438768, with a nearly genome-wide significant p-value slightly larger than 10^{-8} . Both SNPs belong to the miscRNA gene LOC645249. Since this analysis is focused on the methodology and does not contain all available lung cancer GWAS, we shall

refrain from in-depth discussions regarding the biology of the findings. To verify that findings by CO approach were not false positive signals, we tested for G-E dependence in controls. None of the markers revealed evidence of strong G-E association for either smoking model (data not shown).

Since in the CC test none of the SNPs demonstrated genome-wide significant interaction, selecting the top SNPs for follow-up is reasonable. In this respect and for each GWAS, we compared the top 100 SNPs yielded by our proposed EHB-GE method to those of the competing methods.

The top 100 SNPs obtained by the different GxE methods within the study and smoking model displayed similar trends. Results of EHB-GE were in close agreement with those of the case-only test. Particularly in SLRI and GLC, the top 100 SNPs ranked by EHB-GE and CO were identical, in CE-IARC and MDACC there were at most three discordant SNPs. Results of both tests moderately coincided with those of the intuitive two-step test and Mukherjee's empirical Bayes-type shrinkage estimator.

As an example, **Table 3** displays the results for the top 100 SNPs from the GLC Study. Approximately 75-90 SNPs in the top 100 agreed for EHB-GE and MUK-EB in the NE model, with nearly 60-70 SNPs in the MH model. For TWO, we observed 48-72 common SNPs using NE and 41-64 common ones with MH. With MH we had around 20-40 SNPs in common for EHB-GE and CC. Notably, in SLRI we saw greater discordance of the SNPs for both smoking models compared to the other three GWAS datasets. The intuitive two step approach demonstrated a greater agreement with case-control test of the top 100 SNPs, 40-60 SNPs for MH and around 60 for NE. There were no common SNPs in the top 100 identified by MUR and CC in any of the analyses. For the *never* versus *ever* model, the overlap of MUR with other methods was limited to only few SNPs. However, for *moderate* versus *heavy* we observed up to 17 common SNPs chosen by EHB-GE and MUR. Please note that, among the top 100 pairs, in GLC at most 25 pairs, in SLRI, CE-IARC and MDACC at most 6 pairs of SNPs were correlated with $r^2 > 0.8$.

Genotyping for the four studies was performed on two different genotyping platforms (see **Table 2**), leading to availability of some SNPs in one study and not in the other. Therefore, we selected the top 100 genes and not SNPs as previously described, to compare ranking by each GxE method across the GWAS studies. Comparison of the 100 top ranked genes *per GxE method between the GWAS studies* suggested low concordance. For the CC test and our new hierarchical modeling approach, the results are presented in **Table 4**. We observed only up to 5 common genes among 100 across studies for any of the methods. The number of common genes occurring for at least two studies within the top 100 ranked genes varied per method from 14 to 26. In total, we noted 73 such genes across the different methods. No two approaches yielded the same sets of genes, indicating a large variability between studies applying these methods.

DISCUSSION

Recently, much effort has been devoted to developing novel methods to search for GxE interactions capable of reaching high power consistently and maintaining a pre-specified level of Type I error [Mukherjee et al., 2008; Murcray et al., 2009; Murcray et al., 2011, Thomas et al., 2012]. To date, no single method has been found that reliably outperforms others for all sets of parameters [Thomas et al., 2012]. Usually, GxE interactions methods have low power to identify significant interactions, possibly due to the moderate to low true GxE effect sizes, limited sample size, or enormous number of tests performed in a single study. Additionally, the presence of population G-E association and failure of the methods to account for it properly are two of the major reasons for inflated Type I error rate. Although the case-only test has greater power compared to the case-control method, it produces a large number of false positive results when G-E association is present. Thomas et al. [2012] proposed to use the case-only test as a screening tool in an initial discovery sample followed by replication. Unfortunately, this is not always a viable solution, due to the excessive number of false positive findings or lack of additional data. Following up the top ranked SNPs is most often a useful alternative strategy. We evaluated the rank power of different GxE interaction methods. Our simulation study revealed that, in many situations, our novel empirical hierarchical Bayes method for GxE interaction has greater rank power compared to other methods. In all scenarios, the EHB-GE outperformed the traditional CC test. Rank power of EHB-GE declined only slightly with increasing number of G-E associations compared to other methods. The performance of EHB-GE was similar to that of CO if there were no or few weak G-E associations. However, with increasing number or strength of G-E associations, the case-only test lost power. In the situation of no or only a few weak G-E association effects, the method proposed by Murcray et al. [2009] demonstrated the best power. However, the power of this method decreased with increasing number or strength of the G-E associations. MUK-EB and the TWO methods showed slightly lower power than EHB-GE for most scenarios.

We analyzed four lung cancer GWASs and found that the results of EHB-GE closely agree with those of the CO method. Since we expected the CO test to produce false positive findings due to G-E association, this was not anticipated. Hence, we tested for population G-E association in controls. Although with smoking we expected the presence of population G-E association, no such signals were found in controls. Therefore, it is understandable that the results for EHB-GE and CO were similar. To explain the similarity between EHB-GE and CO methods, the hyperparameter estimates of EHB-GE are depicted in **Table 5**. For GLC and SLRI, the estimates of μ representing the strength of a population G-E association were less than 10^{-5} . Hence, the posterior estimates λ_M for the single markers M subtracted of the β_M^{cases} were close to zero; consequently the top 100 SNPs were the same for EHB-GE and CO. For the MDACC and CE-IARC GWAS, the μ estimates were larger, e.g., for the NE model, 0.017 from the CE-IARC and 0.066 from the MH study. Differences across the studies may depend on the actual sample sizes in each of the GWAS, although we cannot prove that. Nevertheless, for CE-IARC and MDACC, we see up to 3 SNPs discordant for CO and EHB-GE in the top 100. Therefore, we can conclude that, when G-E associations are absent, our EHB-GE method performs similarly to the CO test, which is known to have

superior power to identify GxE interactions in the absence of distorting G-E associations. Not all the results are displayed. In particular, the inclusion of main effects into the model yielded qualitatively the same results.

The empirical hierarchical Bayes for GxE interactions method demonstrated high power while accounting for population G-E associations. It should be the first choice for follow-up SNP selection, particularly when G-E associations are suspected. Murcray's method is recommended as a reasonable complement to EHB-GE when the G-E effects are expected to be small.

Acknowledgments

This research was supported by NIH grant TRICL 5U19CA148127-03 and DFG GRK 1034.

Authors of the manuscript have declared no conflict of interest for this work.

Appendix A

The density functions for the hierarchical model and the hyperparameters $\Theta = (\theta, \sigma, p)$ are

$$f\left(\hat{\beta}_m^{controls} \mid \lambda_M\right) = \left(\varphi\left(\frac{\|\hat{\beta}_M^{controls} - \lambda_M\|}{\hat{\sigma}_M^{controls}}\right) + \varphi\left(\frac{\|\hat{\beta}_M^{controls} + \lambda_M\|}{\hat{\sigma}_M^{controls}}\right) \right) / \hat{\sigma}_M^{controls},$$

$$g(\lambda_M \mid \theta, \sigma, p) = p \left(\varphi\left(\frac{\lambda_M - \theta}{\sigma}\right) + \varphi\left(\frac{\lambda_M + \theta}{\sigma}\right) \right) / \sigma + (1 - p) \delta(0)$$

where $\phi(\cdot)$ is the standard normal density. To obtain the estimates for the hyperparameters of that model, we need the marginal likelihood function. The marginal distribution is given by

$$m\left(\hat{\beta}_M^{controls} \mid \theta, \sigma, p\right) = \int f\left(\hat{\beta}_m^{controls} \mid \lambda_M\right) g(\lambda_M \mid \theta, \sigma, p) d\lambda_M$$

$$= p \frac{\varphi(D_{+M}) + \varphi(D_{-M})}{\sqrt{(\hat{\sigma}_M^{controls})^2 + \sigma^2}} + (1 - p) 2\varphi\left(\frac{\|\hat{\beta}_M^{controls}\|}{\hat{\sigma}_M^{controls}}\right) / \hat{\sigma}_M^{controls},$$

where $D_{+M} = \frac{\|\hat{\beta}_M^{controls} + \theta\|}{\sqrt{(\hat{\sigma}_M^{controls})^2 + \sigma^2}}$, $D_{-M} = \frac{\|\hat{\beta}_M^{controls} - \theta\|}{\sqrt{(\hat{\sigma}_M^{controls})^2 + \sigma^2}}$. The likelihood of the model is given

by $L = \prod_M m\left(\hat{\beta}_M^{controls} \mid \theta, \sigma, p\right)$ and is maximized with respect to Θ . Based on the MLEs $\hat{\Theta} = (\hat{\theta}, \hat{\sigma}, \hat{p})$, the posterior expectation of λ_M is straight-forward and the posterior variance can be derived using an approximation by Kass and Steffey [1989]

$$Var\left[\lambda_M \mid \hat{\beta}_M^{controls}\right] = E\left[Var\left[\lambda_M \mid \hat{\beta}_M^{controls}, \Theta\right]\right] + Var\left[E\left[\lambda_M \mid \hat{\beta}_M^{controls}, \Theta\right]\right]$$

$$\approx Var\left[\lambda_M \mid \hat{\beta}_M^{controls}, \hat{\Theta}\right] + \Sigma_{j,h} \tilde{\tau}_h \tilde{\delta}_{Mj} \tilde{\delta}_{Mh}$$

$$= E\left[\lambda_M^2 \mid \hat{\beta}_M^{controls}, \hat{\Theta}\right] - \left(E\left[\lambda_M \mid \hat{\beta}_M^{controls}, \hat{\Theta}\right]\right)^2 + \Sigma_{j,h} \tilde{\tau}_h \tilde{\delta}_{Mj} \tilde{\delta}_{Mh}$$

where $\tilde{\tau}_{jh}$ is the (j,h) -component of the inverse negative Hessian of the marginal log-

likelihood evaluated at the MMLE, $\tilde{\Sigma} = \left(-D^2 \log(L)(\hat{\Theta}) \right)^{-1}$, $\tilde{\delta}_{Mk}$ is given by the Jacobian of the posterior expectation with

$$\tilde{\delta}_{Mk} = \left(\frac{\partial}{\partial \Theta_k} \right) E \left[\lambda_M | \hat{\beta}_M^{controls}, \Theta \right] |_{\Theta = \hat{\Theta}} \quad \text{at} \quad \hat{\Theta}. \quad E \left[\lambda_M | \hat{\beta}_M^{controls}, \hat{\Theta} \right] \text{ is known and}$$

$$\begin{aligned} & E \left[\lambda_M^2 | \hat{\beta}_M^{controls}, \hat{\Theta} \right] \\ &= E \left[\lambda_M^2 | \lambda_M > 0, \hat{\beta}_M^{controls}, \hat{\Theta} \right] \\ &\quad \times P \left[\lambda_M > 0 | \hat{\beta}_M^{controls}, \hat{\Theta} \right] \\ &+ E \left[\lambda_M^2 | \lambda_M = 0, \hat{\beta}_M^{controls}, \hat{\Theta} \right] \\ &\quad \times P \left[\lambda_M = 0 | \hat{\beta}_M^{controls}, \hat{\Theta} \right] \end{aligned}$$

where the first summand in the above can be calculated and

$P \left[\lambda_M^2 | \lambda_M = 0, \hat{\beta}_M^{controls}, \hat{\Theta} \right] = 0$. The Jacobian and Hessian used in the correction term are given by

$$\tilde{\nabla}_{Mi} = \begin{pmatrix} \tilde{\delta}_{M\theta} \\ \tilde{\delta}_{M\sigma} \\ \tilde{\delta}_{Mp} \end{pmatrix} \quad \text{and} \quad \tilde{\Sigma} = \begin{pmatrix} \widetilde{\tau_{\theta\theta}} & \widetilde{\tau_{\theta\sigma}} & \widetilde{\tau_{\theta p}} \\ \widetilde{\tau_{\sigma\theta}} & \widetilde{\tau_{\sigma\sigma}} & \widetilde{\tau_{\sigma p}} \\ \widetilde{\tau_{p\theta}} & \widetilde{\tau_{p\sigma}} & \widetilde{\tau_{pp}} \end{pmatrix}$$

Finally,

$$Var \left[\lambda_M | \hat{\beta}_M^{controls} \right] = E \left[\lambda_M^2 | \hat{\beta}_M^{controls}, \hat{\Theta} \right] - \left(E \left[\lambda_M | \hat{\beta}_M^{controls}, \hat{\Theta} \right] \right)^2 + \left(\tilde{\nabla}_M \right)^T \tilde{\Sigma} \tilde{\nabla}_M.$$

Detailed formulas and program code can be obtained via email from the authors on request.

REFERENCES

- Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol.* 2001; 154:687–693. [PubMed: 11590080]
- Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 2008; 40:616–622. [PubMed: 18385676]
- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika.* 2005; 92:399–418.
- Hirschhorn JN. Genetic approaches to studying common diseases and complex traits. *Pediatr Res.* 2005; 57:74–77.
- Holle R, Happich M, Löwel H, Wichmann HE. KORA – a research platform for population based health research. *Gesundheitswesen Bundesverband Der Ärzte Des Öffentlichen Gesundheitsdienstes Germany.* 2005; 67:19–25.
- Hung RJ, Christiani DC, Risch A, Popanda O, Haugen A, Zienolddiny S, Benhamou S, Bouchardy C, Lan Q, Spitz MR, Wichmann HE, LeMarchand L, Vineis P, Matullo G, Kiyohara C, Zhang ZF,

Pezeshki B, Harris C, Mechanic L, Seow A, Ng DPK, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, Rudnai P, Fabianova E, Mates D, Foretova L, Janout V, Bencko V, Caporaso N, Chen C, Duell EJ, Goodman G, Field JK, Houlston RS, Hong YC, Landi MT, Lazarus P, Muscat J, McLaughlin J, Schwartz AG, Shen H, Stucker I, Tajima K, Matsuo K, Thun M, Yang P, Wiencke J, Andrew AS, Monnier S, Boffetta P, Brennan P. International Lung Cancer Consortium pooled analysis of sequence variants in DNA repair and cell cycle pathways. *Cancer Epidemiol Biomarkers Prev.* 2008a; 17:3081–3089. [PubMed: 18990748]

Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokian HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-de-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcátová I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Bolland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature.* 2008b; 452:633–637. [PubMed: 18385738]

Kass RE, Steffey D. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *JASA.* 1989; 84:717–726.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered.* 2007; 63:111–119. [PubMed: 17283440]

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol.* 2007; 31:871–882. [PubMed: 17654612]

Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol.* 2008; 32:615–626. [PubMed: 18473390]

Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics.* 2008; 64:685–694. [PubMed: 18162111]

Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am J Epidemiol.* 2012; 175:177–190. [PubMed: 22199027]

Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol.* 2009; 169:219–226. [PubMed: 19022827]

Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. Sample size requirements to detect gene environment interactions in genome wide association studies. *Genet Epidemiol.* 2011; 35:201–210. [PubMed: 21308767]

Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Stat in Med.* 1994; 13:153–162. [PubMed: 8122051]

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]

Satten GA, Epstein MP. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol.* 2004; 27:192–201. [PubMed: 15372619]

Sauter W, Rosenberger A, Beckmann L, Kropp S, Mittelstrass K, Timofeeva M, Wölke G, Steinwachs A, Scheiner D, Meese E, Sybrecht G, Kronenberg F, Dienemann H, LUCY-Consortium. Chang-Claude J, Illig T, Wichmann HE, Bickeböller H, Risch A. Matrix metalloproteinase1 (MMP1) is associated with early-onset lung cancer. *Cancer Epidemiol Biomarkers Prev.* 2008; 17:1127–1135.

Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO. Methodological Issues in Multistage Genome-Wide Association Studies. *Statist Sci.* 2009; 24:414–429.

Thomas DC, Lewinger JP, Murcray CE, Gauderman WJ. Invited commentary: Ge-whiz! ratcheting gene-environment studies up to the whole genome and the whole exposome. *Am J Epidemiol.* 2012; 175:203–207. discussion 208–209. [PubMed: 22199029]

- Wang Y, Broderick P, Matakidou A, Eisen T, Houlston RS. Role of 5p15.33 (TERT-CLPTM1L), 6p21.33 and 15q25.1 (CHRNA5-CHRNA3) variation and lung cancer risk in never-smokers. *Carcinogenesis*. 2010; 31:234–238. [PubMed: 19955392]
- Wichmann HE, Gieger C, Illig T. KORA-gen – resource for population genetics, controls and a broad spectrum of disease phenotypes. *Das Gesundheitswesen*. 2005; 67:26–30.
- Ziegler A, König IR, Thompson JR. Biostatistical aspects of genomewide association studies. *Biom J*. 2008; 50:8–28. [PubMed: 18217698]

Figure 1.

Rank power to detect a GxE interaction in the top 25 ranking SNPs by EHB-GE with various ratios of cases and controls ($ccr = 1:1, 1:2$ and $2:1 \Rightarrow 1,500:1,500, 1,000:2,000$ and $2,000:1,000$ cases and controls) and varying number of G-E associations N_{G-E} with different effect sizes $OR_{G-E}(\text{low})$, $OR_{G-E}(\text{medium})$ and $OR_{G-E}(\text{high})$, $OR_{GxE} = 2$, $p_g = 30\%$, $p_e = 10\%$, $p_d = 5\%$.

Figure 2.

Rank power comparison to detect a GxE interaction in the top 25 ranking SNPs between EHB-GE and competing method. On the *x-axis* EHB-GE rank power, on the *y-axis* difference between the rank power of EHB-GE and competing method: a) Case-Control (CC), b) Case-Only (CO), c) Intuitive two-step approach (TWO), d) Mukherjee's empirical Bayes-type shrinkage estimator (MUK-EB), e) Murcray's two step approach (MUR)

Table 1Rank power for G×E interacting SNP in top 25 ranking, $ORGxE = 2$, $p_d = 1\%$, $ccr = 1:1$.

p_e	OR_{G-E}	N_{G-E}	p_g	ehb-ge	CC	CO	TWO	MUK-EB	MUR
0.1	none	0	0.1	0.458	0.119	0.459	0.402	0.302	1.000
			0.3	0.867	0.405	0.867	0.813	0.729	1.000
			0.1	0.443	0.119	0.446	0.401	0.302	0.992
	low	100	0.3	0.856	0.405	0.856	0.809	0.727	0.997
			0.1	0.414	0.119	0.389	0.384	0.291	0.128
			0.3	0.830	0.408	0.812	0.792	0.719	0.403
	high	500	0.1	0.471	0.114	0.001	0.447	0.355	0.109
			0.3	0.898	0.445	0.012	0.861	0.767	0.442
			0.1	0.414	0.116	0.000	0.428	0.345	0.079
	none	0	0.3	0.867	0.446	0.000	0.836	0.763	0.376
			0.1	0.839	0.412	0.837	0.794	0.721	1.000
			0.3	0.997	0.855	0.997	0.977	0.961	1.000
0.3	low	100	0.1	0.823	0.411	0.814	0.792	0.720	0.538
			0.3	0.996	0.858	0.996	0.976	0.961	0.889
			0.1	0.794	0.411	0.607	0.765	0.71	0.316
	high	500	0.3	0.991	0.858	0.976	0.975	0.96	0.843
			0.1	0.833	0.393	0.000	0.783	0.705	0.324
			0.3	0.993	0.864	0.001	0.977	0.961	0.852
	none	0	0.1	0.815	0.393	0.000	0.781	0.700	0.274
			0.3	0.993	0.866	0.000	0.976	0.961	0.838

p_e = exposure frequency, OR_{G-E} , N_{G-E} = strength, number of G-E associations p_g = MAF interacting SNP, EHB-GE = empirical hierarchical Bayes approach, CC = case-control, CO = case-only, TWO = intuitive two-step approach, MUK-EB = Mukherjee's empirical Bayes-type shrinkage estimator, MUR = Murcray's two step approach.

Table 2

Overview of four TRICL/ILCCO lung cancer GWAS

	GLC	CE-IARC	MDACC	SLRI
Principal Investigator	Heinrich, Risch, Bickeboller	Brennan	Amos	Hung
Location	Germany	Czech Republic, Hungary, Poland, Romania, Russia, Slovakia	Houston, Texas, USA	greater Toronto area, Canada
Design	Population-based	Hospital-based	Hospital-based	Hospital and population-based
Number of cases/controls	514/488	1989/2625	1154/1137	332/505
Genome-wide SNP chip	HumanHap550K	HumanHap300K	HumanHap300K	HumanHap300K
Control recruitment	KORA cohort	Non-tobacco related diseases	Cancer free patients; only ever smokers	Family, medicine, clinic, tax files

Table 3

Overlap of the top 100 SNPs for the different GxE interaction methods in GLC GWAS

“never” vs. “ever” smokers	“moderate” vs. “heavy” smokers						
		CC	CO	TWO	MUK-EB	MUR	EHB-GE
	CC		9	43	15	0	9
	CO	30		64	60	16	100
	TWO	59	66		56	11	64
	MUK-EB	31	77	63		4	60
	MUR	0	2	1	0		16
	EHB-GE	30	100	66	77	2	

CC = case-control; CO = case-only; TWO = intuitive two-step approach; MUK-EB = Mukherjee's empirical Bayes-type shrinkage estimator; MUR = Murcray's two step approach; EHB-GE = empirical hierarchical Bayes approach;

Table 4

Overlap of the top 100 genes between smoking models and studies for EHB-GE and CC

			EHB-GE						
			GLC		CE-IARC		MDACC	SLRI	
			MH	NE	MH	NE	MH	MH	NE
CC	GLC	MH		2	1	0	2	0	0
		NE	1		0	1	1	3	2
	CE-IARC	MH	0	2		4	2	0	2
		NE	2	2	4		0	0	1
	MDACC	MH	2	1	4	2		0	4
	SLRI	MH	3	0	3	2	1		5
		NE	0	1	3	3	1	3	

NE = *never* vs. *ever*, MH = *moderate* vs. *heavy*; CC = case-control, EHB-GE = empirical hierarchical Bayes; Studies: GLC, CE-IARC, MDACC, SLRI.

Table 5

EHB-GE hyperparameter estimates and variance estimates for the different models

	GLC		CE-IARC		MDACC	SLRI	
	NE	MH	NE	MH	MH	NE	MH
θ	3×10^{-6}	10^{-6}	0.017	0.066	0.079	10^{-6}	10^{-6}
p	0.004	0.005	0.152	0.039	0.003	0.004	0.006
σ	2.18×10^{-5}	10^{-10}	10^{-10}	10^{-10}	1.5×10^{-5}	7.59×10^{-6}	10^{-10}
<i>posterior</i> ($\sigma_{controls}$)	3.09×10^{-5}	1.68×10^{-4}	0.0119	0.0129	4.6×10^{-3}	2.95×10^{-5}	3.46×10^{-4}
Mean [*] (var EHB-GE/ var MUK-EB)	0.949	0.752	0.965	0.845	0.822	0.873	0.816

θ, σ, p level 2 parameters of the hierarchical model; θ strength of a population-based G-E correlation, σ scaling parameter, p proportion of SNPs with G-E correlation; *posterior* ($\sigma_{controls}$) estimated average posterior variance of G-E correlations within controls;

* ratio calculated for each SNP (comparing Variance of EHB-GE with that of the MUK-EB method); NE = *never vs. ever*, MH = *moderate vs. heavy*; Studies: GLC, CE-IARC; MDACC; SLRI