



**HAL**  
open science

## Deriving stratified effects from joint models investigating gene-environment interactions

Vincent Laville, Timothy Majarian, Paul de Vries, Amy Bentley, Mary Feitosa, Yun Sung, D. Rao, Alisa Manning, Hugues Aschard, Charge  
Gene-Lifestyle Interactions Working Group

► **To cite this version:**

Vincent Laville, Timothy Majarian, Paul de Vries, Amy Bentley, Mary Feitosa, et al.. Deriving stratified effects from joint models investigating gene-environment interactions. BMC Bioinformatics, 2020, 21 (1), pp.251. 10.1186/s12859-020-03569-4 . pasteur-03278592

**HAL Id: pasteur-03278592**

**<https://pasteur.hal.science/pasteur-03278592>**

Submitted on 5 Jul 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

SOFTWARE

Open Access

# Deriving stratified effects from joint models investigating gene-environment interactions



Vincent Laville<sup>1\*</sup> , Timothy Majarian<sup>2</sup>, Paul S. de Vries<sup>3</sup>, Amy R. Bentley<sup>4</sup>, Mary F. Feitosa<sup>5</sup>, Yun J. Sung<sup>5</sup>, D. C. Rao<sup>5</sup>, Alisa Manning<sup>2,6</sup>, Hugues Aschard<sup>1,7\*</sup> and on behalf of the CHARGE Gene-Lifestyle Interactions Working Group

\* Correspondence: [vincent.laville@pasteur.fr](mailto:vincent.laville@pasteur.fr); [hugues.aschard@pasteur.fr](mailto:hugues.aschard@pasteur.fr)

<sup>1</sup>Department of Computational Biology, USR 3756 CNRS, Institut Pasteur, Paris, France  
Full list of author information is available at the end of the article

## Abstract

**Background:** Models including an interaction term and performing a joint test of SNP and/or interaction effect are often used to discover Gene-Environment (GxE) interactions. When the environmental exposure is a binary variable, analyses from exposure-stratified models which consist of estimating genetic effect in unexposed and exposed individuals separately can be of interest. In large-scale consortia focusing on GxE interactions in which only the joint test has been performed, it may be challenging to get summary statistics from both exposure-stratified and marginal (i.e. not accounting for interaction) models.

**Results:** In this work, we developed a simple framework to estimate summary statistics in each stratum of a binary exposure and in the marginal model using summary statistics from the “joint” model. We performed simulation studies to assess our estimators’ accuracy and examined potential sources of bias, such as correlation between genotype and exposure and differing phenotypic variances within exposure strata. Results from these simulations highlight the high theoretical accuracy of our estimators and yield insights into the impact of potential sources of bias. We then applied our methods to real data and demonstrate our estimators’ retained accuracy after filtering SNPs by sample size to mitigate potential bias.

**Conclusions:** These analyses demonstrated the accuracy of our method in estimating both stratified and marginal summary statistics from a joint model of gene-environment interaction. In addition to facilitating the interpretation of GxE screenings, this work could be used to guide further functional analyses. We provide a user-friendly Python script to apply this strategy to real datasets. The Python script and documentation are available at <https://gitlab.pasteur.fr/statistical-genetics/j2s>.

**Keywords:** Gene-environment interaction, Binary exposure, Stratified analysis, Summary statistics



## Background

Gene-Environment (GxE) interactions are of great interest in deciphering biological mechanisms underlying complex human traits and diseases. Several theoretical approaches [1–3] and applications [4–7] have recently been published that identify such GxE interactions. A strategy to detect these interactions applies linear regression models including a GxE interaction term and testing for the hypothesis of null main genetic effect size and GxE interaction effect size, also referred to as the “joint” test [8, 9]. Although several interactions have been associated with different traits using this joint test, the main limitation is that of large sample sizes requirements to reach a suitable statistical power [10]. The Gene-Lifestyle Interaction Working Group is an international, large-scale, multi-ancestry initiative within the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium that aims to systematically evaluate genome-wide GxE interactions on cardiovascular disease related traits using genotypic data from up to 610,475 individuals [11]. This working group has already unraveled significant GxE interactions using the joint test [12–15]. Nevertheless, in the case of binary exposures, alternative approaches can be of interest, notably to identify differential genetic effects between unexposed and exposed individuals. This strategy requires summary statistics computed in each group of individuals separately, which may not always be available in large-scale consortia. Because of logistical challenges, it can be difficult to obtain these summary statistics in such consortia including tens of individual cohorts.

To benefit from these consortia in which only summary statistics in the joint testing framework may be available, we developed a simple tool to infer summary statistics in the groups of unexposed and exposed individuals separately, as well as summary statistics from the regression model without the GxE interaction term. First, we showed that these summary statistics can be efficiently derived from the joint model assuming independence between genotypes and exposure. We then performed a series of simulations to assess the accuracy of these estimations and to examine the impact of different potential sources of bias. Finally, we applied our pipeline to real data from the Gene-Lifestyle Interactions Working group within the CHARGE Consortium.

## Theoretical derivations

Consider a trait  $Y$ , a dichotomous exposure  $E$  and a SNP  $G$  coded as the number of minor alleles. A framework to test Gene-Environment interactions is based on the generalized linear model:

$$g(\mathbb{E}[Y|G]) = \alpha + \beta G + \gamma E + \delta GE$$

where  $g$  denotes either the identity function if  $Y$  is a quantitative trait or the logit function if  $Y$  is a binary phenotype.

The marginal model (i.e excluding the interaction term) in unexposed individuals ( $E = 0$ ), exposed individuals ( $E = 1$ ) and all individuals are defined as:

$$\begin{aligned} g(\mathbb{E}[Y|G, E = 0]) &= \alpha_{unexp} + \beta_{unexp} G \\ g(\mathbb{E}[Y|G, E = 1]) &= \alpha_{exp} + \beta_{exp} G \\ g(\mathbb{E}[Y|G]) &= \alpha_{marg} + \beta_{marg} G + \gamma E \end{aligned}$$

Assuming independence between the genotypes and the exposure (i.e  $\mathbb{E}[G|E = 0] = \mathbb{E}[G|E = 1] = G$ ), the joint model can be used to retrieve the marginal genetic effects  $\beta_{unexp}$  and  $\beta_{exp}$  in unexposed ( $e = 0$ ) and exposed ( $e = 1$ ) individuals respectively:

$$g(\mathbb{E}[Y|G, E = e]) = \alpha + \beta G + \gamma e + \delta Ge$$

$$= \alpha + (\beta + \delta e)G + \gamma e$$

Then setting  $e$  to either 0 or 1, marginal effect sizes in unexposed individuals  $\widehat{\beta}_{unexp}$  and in exposed individuals  $\widehat{\beta}_{exp}$  can be derived from the genetic and interaction effect sizes ( $\hat{\beta}$  and  $\hat{\delta}$  respectively) estimated in the joint model:

$$\widehat{\beta}_{unexp} = \hat{\beta}, \sigma_{\widehat{\beta}_{unexp}} = \sigma_{\hat{\beta}}$$

$$\widehat{\beta}_{exp} = \hat{\beta} + \hat{\delta}, \sigma_{\widehat{\beta}_{exp}} = \sqrt{\sigma_{\hat{\beta}}^2 + \sigma_{\hat{\delta}}^2 + 2 \text{cov}(\sigma_{\hat{\beta}}, \sigma_{\hat{\delta}})}$$

where  $\sigma_{\hat{\beta}}$  and  $\sigma_{\hat{\delta}}$  denote respectively the standard errors of the genetic effect and interaction effect in the joint model.

Similarly, summary statistics in the marginal model (excluding the interaction term) can be derived from the joint model:

$$g(\mathbb{E}[Y|G]) = g(\mathbb{E}[Y|E = 0]) \times P(E = 0) + g(\mathbb{E}[Y|E = 1]) \times P(E = 1)$$

$$= [\alpha + \beta G] \times (1 - \mu_E) + [\alpha + \beta G + \gamma + \delta G] \times \mu_E$$

$$= (\alpha + \gamma \mu_E) + (\beta + \delta \mu_E)G$$

Hence, the marginal genetic effect  $\widehat{\beta}_{marg}$  and its standard error  $\sigma_{\widehat{\beta}_{marg}}$  are equal to:

$$\widehat{\beta}_{marg} = \hat{\beta} + \hat{\delta} \mu_E$$

$$\sigma_{\widehat{\beta}_{marg}} = \sqrt{\sigma_{\hat{\beta}}^2 + \mu_E^2 \sigma_{\hat{\delta}}^2 + 2 \mu_E \text{cov}(\sigma_{\hat{\beta}}, \sigma_{\hat{\delta}})}$$

### Implementation

We developed a Python script to derive summary statistics in the marginal model and in each group of individuals separately. As input, the script takes one file with the summary statistics from the joint model, that are genetic and interaction effect sizes, their standard errors, the correlation between the two effect sizes and the sample size per SNP corresponding to the number of genotypes available for this SNP (which may differ from the sample size of the study because of missing data). This file corresponds to the output of the METAL software to meta-analyze GxE screenings using the joint test [9]. In addition to this file, the script also takes two arguments that are the total sample size  $N$  of the study and the number of exposed individuals  $N_e$  included in the study. These two arguments are used to infer the sample sizes  $N_v \times (N - N_e) / N$  and  $N_v \times N_e / N$  in the group of unexposed and exposed individuals respectively for each SNP, where  $N_v$  is the sample size for the SNP. We also implemented a filtering procedure to exclude variants with a low sample size compared to the distribution of the sample sizes: a SNP with a sample size below the 9th decile of the sample size distribution divided by 1.5 is excluded from the analysis. As output, the script generates a single file containing the genetic effect size and its standard error in the group of unexposed individuals, in the group of exposed individuals and in the total sample. The script and a

detailed documentation using an example are available at <https://gitlab.pasteur.fr/statistical-genetics/j2s>.

## Results

### Simulation study

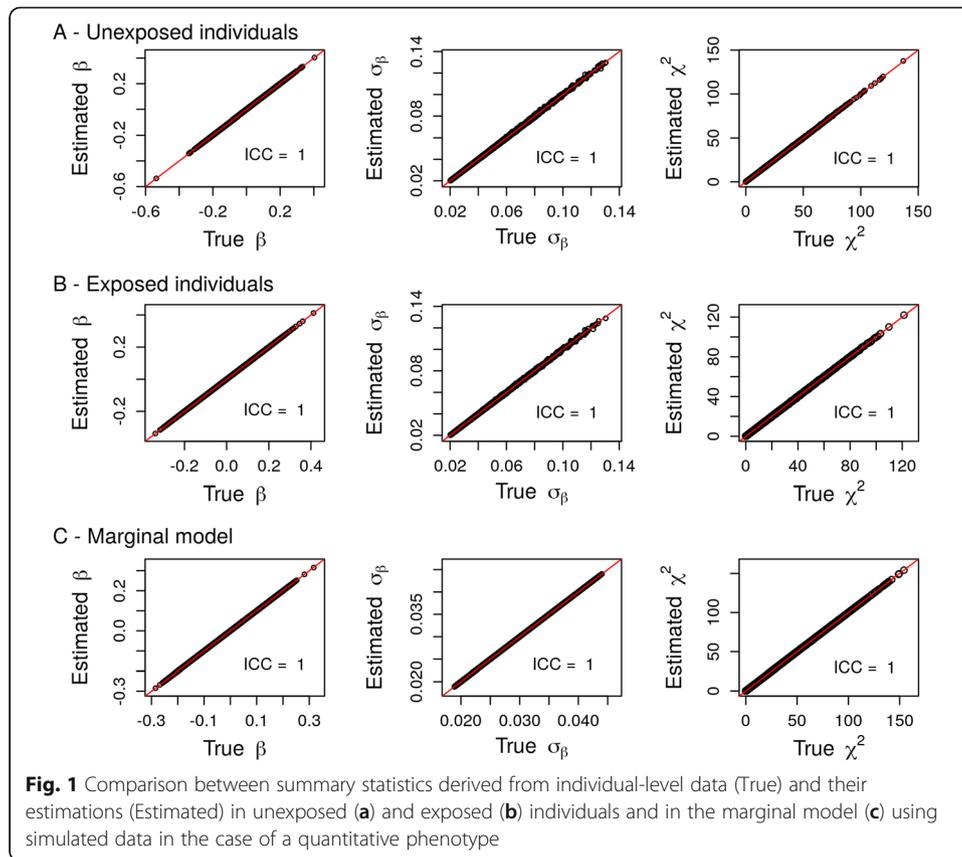
First, we performed a simulation study to assess the accuracy of the estimations obtained from the theoretical results described above. In each of the 1000 replicates, we simulated 10,000 genotypes of a SNP with a random MAF between 1 and 50% and a binary exposure with a random probability of being exposed ranging from 0.1 to 0.5. Then, we simulated a continuous phenotype  $Y = \beta_G G + \beta_E E + \beta_{GE} G \times E + \varepsilon$  as a linear combination of the SNP  $G$ , the exposure  $E$  and the  $G \times E$  interaction term with randomly chosen effect sizes  $\beta_G$ ,  $\beta_E$  and  $\beta_{GE}$  and a random noise  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ . The effect sizes  $\beta_G$ ,  $\beta_E$  and  $\beta_{GE}$  were drawn from a uniform distribution on  $[0.05; 0.2]$  with a randomly and equiprobably chosen sign. Note that in this design, genotypes  $G$  and exposure  $E$  were drawn independently. Then, on the one hand, we computed the summary statistics from the joint model including the GxE interaction term using individual level data. On the other hand, we applied linear regressions without the GxE interaction term in each group of individuals (unexposed and exposed) separately and in the pooled sample to compute the summary statistics of the genetic effect in each group of individuals and in the marginal model. Using the estimators derived from the joint model, we also inferred these summary statistics in each group and in the marginal model using our pipeline. Comparisons of the empirical and inferred summary statistics showed high accuracy of the estimators, with intraclass correlation coefficient (ICC) between “real” and “estimated” equal to 1 in all scenarios (Fig. 1).

We also performed this simulation study for a binary trait. For each of the 10,000 replicates, we generated random effect sizes  $\beta_G$ ,  $\beta_E$  and  $\beta_{GE}$  as described above and then simulated a binary outcome from a Bernoulli distribution, with the probability of being a case as a binary trait from using a logistic model as  $P(Y = 1) = \frac{1}{1 + e^{\alpha + \beta_G G + \beta_E E + \beta_{GE} G \times E + \varepsilon}}$ , where  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ . We then conducted the same analyses as for quantitative traits by performing logistic regressions instead of linear regressions to compare the summary obtained using individual-level data to those estimated by our pipeline. As for quantitative traits, the estimator was highly accurate (Figure S1).

### Potential bias sources

We performed several complementary simulation studies to assess the contribution of several bias sources. Each time, we generated genotypes for 50,000 individuals and repeated the analysis 10,000 times.

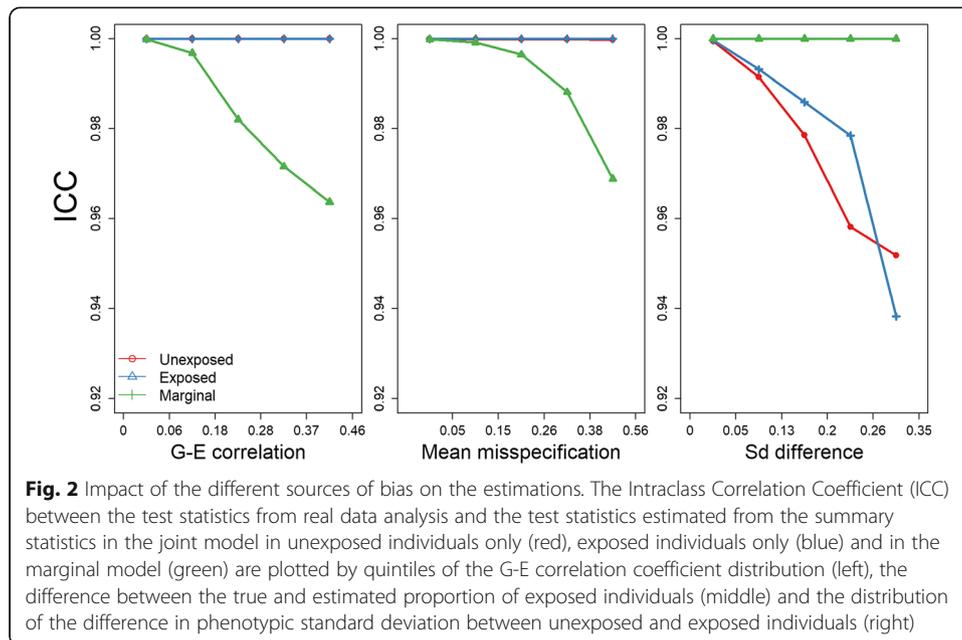
First, the estimators' derivation relies on the assumption that genotypes and environment are statistically independent. We performed a simulation study in which correlation existed between genotypes and the environment. We then compared our summary statistics estimated from the joint model to summary statistics derived using individual-level data (Figs. 2, S2). Relaxing the G-E independence assumption did not impact the estimator's accuracy when deriving stratified summary. However, estimations in the marginal model were slightly impacted by the correlation between  $G$  and  $E$ .



Indeed, inferred effect sizes are a little biased and effect sizes standard errors are over-estimated. Although estimation errors increase with the correlation, the impact on the test statistics remains very limited.

Second, bias can occur because of a misspecification of the proportion of exposed individuals. This is very likely to happen for SNPs with a low sample size (because of missing genotypes) compared to the maximum sample size. To evaluate the impact of such a misspecification, for each SNP, we selected a subset of individuals to include only a randomly selected proportion of individuals while intentionally misspecifying the proportion of exposed individuals. For each subset, we removed a randomly selected number of exposed individuals. We then compared the summary statistics obtained using the individual-level data and those estimated using the pipeline. As expected from the theoretical derivations detailed above, misspecifying the proportion of exposed individuals, quantified as  $|\mu_E - m_E| / \mu_E$  where  $\mu_E$  is the mean of the exposure in the whole sample and  $m_E$  is the mean of the exposure in the subsample, only impacted estimations in the marginal model including all individuals. Notably, the larger the difference between the true (in the subset of selected individuals) and the estimated (computed in the whole sample) proportions of exposed individuals, the larger are the discrepancies between the summary statistics (Figs. 2, S3).

Third, bias in our estimations can also occur due to differences in phenotypic variance between unexposed and exposed individuals. To explore this, we simulated a phenotype with exposure-dependent variance by adding statistical noise to the



phenotypes of exposed individuals and performed the same simulation study as described above. A different phenotypic variance in the two groups of individuals did not bias estimation of the summary statistics in the marginal model but it clearly biased the estimation of summary statistics in the exposed and unexposed individuals (Figs. 2, S4). Although this exposure-dependent phenotypic variance did not impact the estimation of the effect sizes, it biased the estimation of the effect size standard error. Standard errors tend to be overestimated in the group in which the phenotypic variance is the largest, leading to deflated test statistics and conversely. Importantly, the larger differences in phenotypic variance yielded larger induced biases.

Finally, we also generated data (50,000 individuals and 10,000 iterations) under a null model with neither a genetic effect nor an interaction effect to assess the control of the type I error rate and quantify the discrepancies in significance results that can arise because of these different sources of bias (Figures S5, S6). Globally, the type I error rate is well-controlled in the presence of G-E correlation and for SNPs with low sample compared to the total sample size (Figure S5), but the systematic inflation (resp. deflation) of chi-square statistics observed in a group of individuals when the phenotypic variance differs depending on the exposure (Figure S4) leads to an uncontrolled type I error rate (Figure S5). However, important discrepancies in the significance assessment evaluated as the proportion of SNPs significant using Bonferroni-adjusted  $p$ -values with only one of the two methods (using individuals-level data or the estimation pipeline) can be observed, confirming the impact of these source of bias (Figure S6).

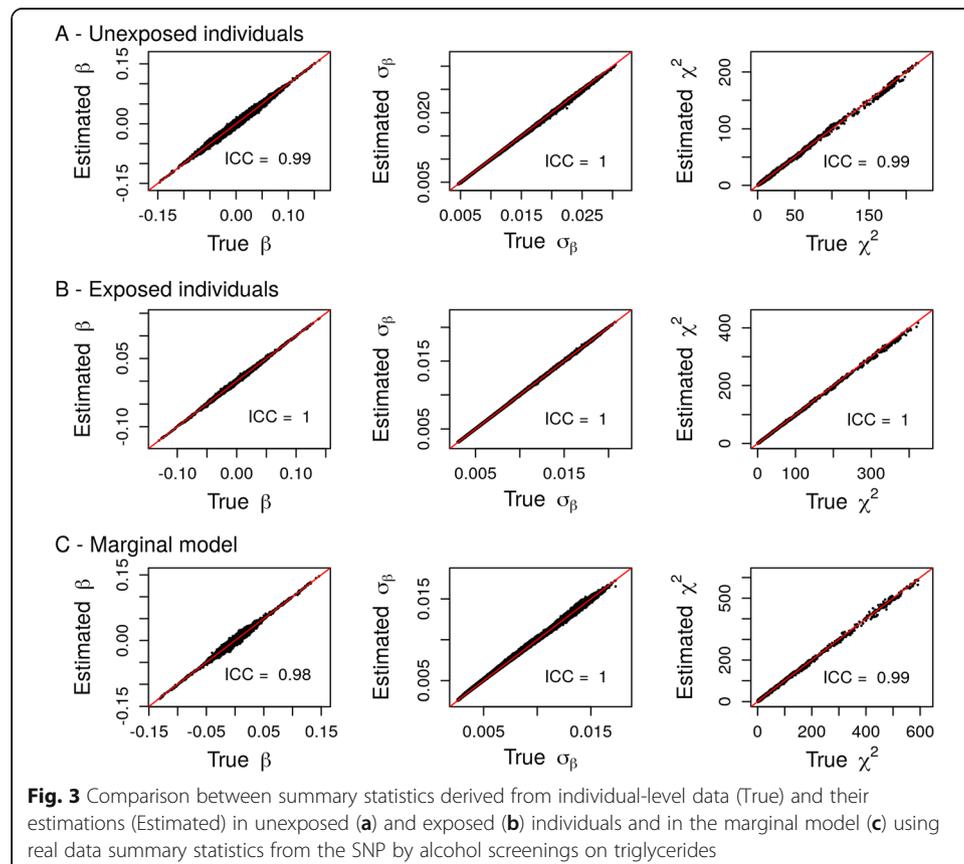
### Real data application

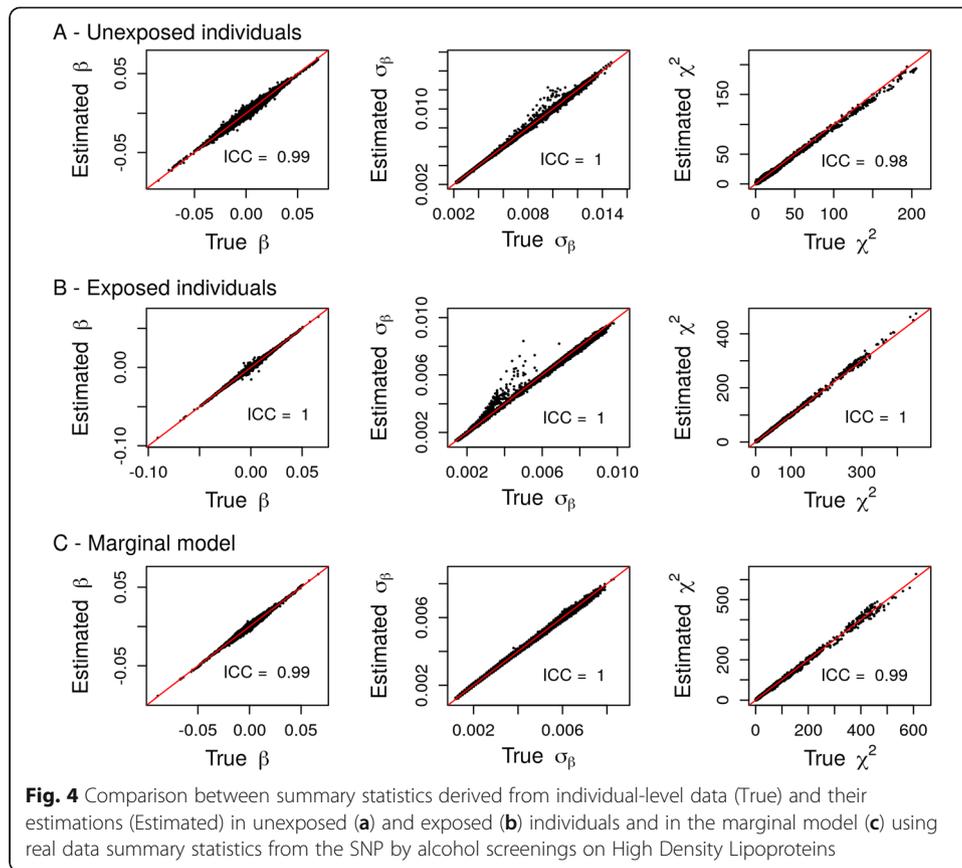
We assessed the accuracy of our estimations using real data from the Gene-Lifestyle Interaction Working Group of the CHARGE consortium [11]. This Working Group recently published genome-wide SNP-by-alcohol interaction screenings [13] using joint tests and focusing on three lipids level: triglycerides (TG), high-density lipoproteins

(HDL), and low-density lipoproteins (LDL). Genome-wide screenings for genetic marginal effects were also performed in unexposed and exposed individuals separately and in the whole sample. Here, we used summary statistics from the genome wide SNP by exposure interaction screenings in individuals from European ancestry and derived marginal summary statistics in unexposed and exposed individuals separately, and in the whole sample. We then compared the inferred summary statistics with the empirical summary statistics derived using individual-level data (Figs. 3, 4, 5). The estimations exhibited high accuracy as demonstrated by the very high ICC between the estimated and true summary statistics (mean ICC = 0.99). Note that some discrepancies are observed for only a very limited number of SNPs (less than 100 out of more than 7 million variants) and do not influence much the ICC, which measures the “agreement” between the true and estimated parameters. Overall, filtering to exclude SNPs with low relative sample size (i.e below the 9th decile of the sample size distribution divided by 1.5) lead to more accurate estimations

**Discussion**

In this work, we aimed at inferring marginal genetic effects in exposed and unexposed individuals separately and in the whole sample using summary statistics of the joint test performed in the context of GxE interaction studies. We analytically derived estimators of marginal genetic effects in the different groups of individuals and in the total sample. We validated the method through simulation studies and real data applications which



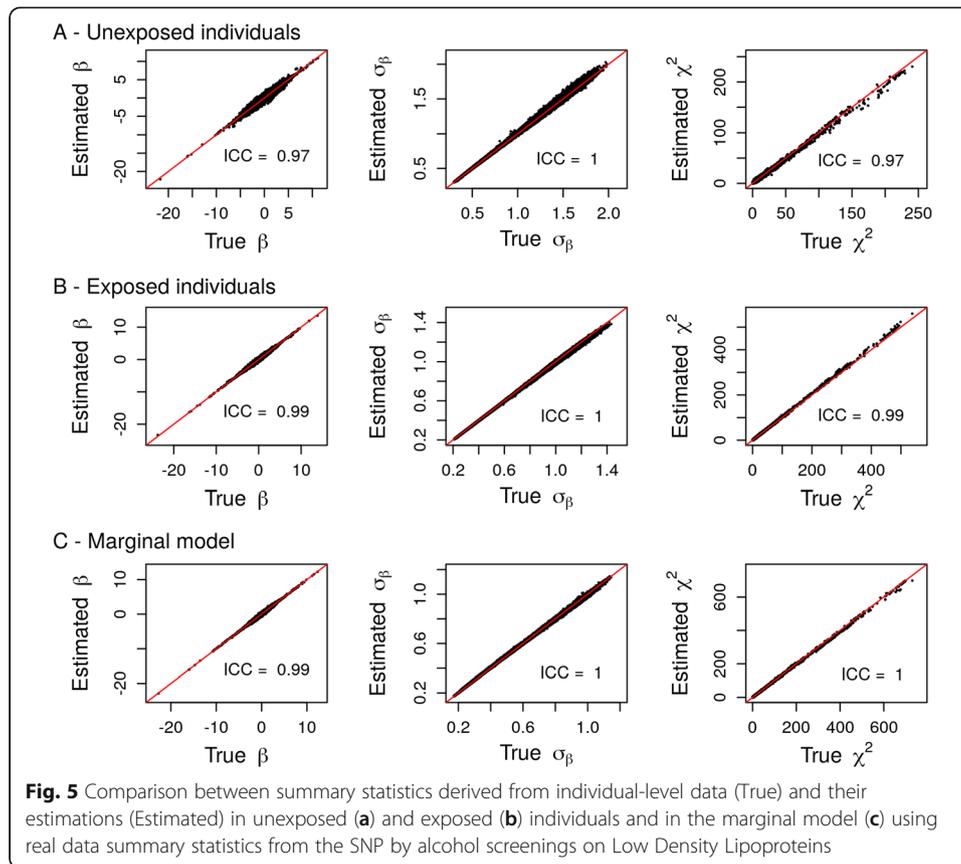


both highlighted the accuracy of our estimations. Notably, this method can be applied without loss of accuracy to quantitative and binary traits.

As demonstrated by our simulation studies, differences between true and estimated parameters are observed for SNPs with the lowest sample sizes. This is explained by a different proportion of exposed individuals for this particular SNP and in the whole sample. Also, our method also provides basic estimates of the expected sample size in the groups of exposed and unexposed individuals. For the same reason, these estimates could be biased for SNPs with low sample size compared to the total sample size. Consequently, we implemented a procedure to filter out variants with low relative sample size to minimize this potential bias.

Our estimations rely on the independence between genotypes and exposures. Relaxing this assumption leads to biased estimations of the marginal effect size standard deviation in the marginal model, but does not impact the accuracy of the estimations in the stratified models. As correlation between SNPs and the exposure cannot be retrieved using summary statistics from the joint model, although this assumption may not hold only for a very limited number of SNPs, existing literature may be helpful to identify variants which should be discarded from the analyses because of existing correlation with the considered exposure. The correlations between genotypes and exposures are expected to be low, resulting in little overall impact, as observed when validating our estimators using real data from the Gene-Lifestyle Interaction Working Group.

Finally, we evaluated our estimations in the case of exposure-dependent phenotypic variance. Although our simulations showed clear impacts on the estimations in the



stratified models, we noted that the error increased with the magnitude of this difference. In real data applications, such differences in phenotypic variance are expected to be small and should consequently have only a limited impact on the estimations in each exposure stratum. Application to real data sets confirmed this notion as our estimations were highly concordant with real data.

Overall, an advantage of exposure-stratified models is that they allow for a comparison between genetic effects in each group of individuals. This different way of quantifying GxE interactions makes the interpretation more intuitive compared to the joint test by comparing genetic effects between the two groups. In addition, exposure-stratified summary statistics can also be used to apply further analyses such as biological pathways [16] or heritability-based [17–19] analyses. Results from those analyses in each group could then be compared and help better understanding the genetic architecture of the trait. These strategies could also highlight different genetic mechanisms induced by the exposure, opening new path towards public health prevention policies or the identification of potential drug targets.

## Conclusion

In this work, we derived accurate estimations of the marginal genetic effects in unexposed and exposed individuals separately and in the whole sample in the context of genome-wide GxE interaction screenings using the joint test. This method can not only lead to a more intuitive understanding of GxE interactions but also be used to perform additional studies that can guide further functional analyses. We implemented j2s, a Python3 script to easily apply this method, available at <https://gitlab.pasteur.fr/statistical-genetics/j2s>.

## Availability and requirements

**Project name:** j2s

**Project home page:** <https://gitlab.pasteur.fr/statistical-genetics/j2s>

**Operating systems:** Linux

**Programming language:** Python3

**Other requirements:** None

**License:** MIT

**Any restrictions to use by non-academics:** None

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12859-020-03569-4>.

**Additional file 1: Figure S1.** Comparison between summary statistics derived from individual-level data (True) and their estimations (Estimated) in unexposed (A) and exposed (B) individuals and in the marginal model (C) using simulated data in the case of a binary phenotype. **Figure S2.** Comparison between summary statistics derived from individual-level data (True) and their estimations (Estimated) in unexposed (A) and exposed (B) individuals and in the marginal model (C) using simulated data in the case of a quantitative phenotype when relaxing the genotype-environment independence. **Figure S3.** Comparison between summary statistics derived from individual-level data (True) and their estimations (Estimated) in unexposed (A) and exposed (B) individuals and in the marginal model (C) using simulated data in the case of differences between the proportion of exposed individuals for the SNP and the proportion of exposed individuals in the whole sample. **Figure S4.** Comparison between summary statistics derived from individual-level data (True) and their estimations (Estimated) in unexposed (A) and exposed (B) individuals and in the marginal model (C) using simulated data in the case of different phenotypic variance conditionally on the exposition. **Figure S5.** Comparison of the Type I error rate evaluated between summary obtained using individual-level data (blue) and summary statistics estimated using the pipeline (orange) with respect to the different quintiles of the different sources of bias: G-E correlation(A), misspecification of the proportion of exposed individuals (B) and different phenotypic variance in the strata of the exposure (C). Type I error rate were evaluated for the marginal model in all individuals (left), in unexposed individuals only (middle) and in exposed individuals only (right). The dashed line represents the nominal significance threshold (5%). **Figure S6.** Proportion of SNPs with discordant significance results between summary statistics obtained using individual-level data and summary statistics estimated using our pipeline with respect to the different quintiles of the different sources of bias: G-E correlation(A), misspecification of the proportion of exposed individuals (B) and different phenotypic variance in the strata of the exposure (C). Type I error rate were evaluated for the marginal model in all individuals (left), in unexposed individuals only (middle) and in exposed individuals only (right).

## Abbreviations

GLIWG: Gene-Lifestyle Interaction Working Group; GxE: Gene-Environment; ICC: Intraclass Correlation Coefficient; SNP: Single Nucleotide Polymorphism

## Acknowledgements

The authors acknowledge all the people involved in the generation and sharing of the data from the CHARGE consortium.

## Authors' contributions

VL and HA designed the study; VL developed the script, VL and TM performed simulation studies and evaluated the accuracy of estimations in real data applications; PSdV, ARB, MFF and YJS generated individual-level data analysis used in the study, DCR, AM and HA supervised the study; VL and HA wrote the manuscript. All authors contributed to the improvement of the manuscript, agreed to be responsible for the accuracy and integrity of this work and provided final approval of the manuscript.

## Funding

This work was supported by the HL118305 grant from the NHLBI. HA was also supported by R21HG007687 from NHGRI. PSdV was supported by American Heart Association grant number 18CDA34110116. This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute in the Center for Research in Genomics and Global Health (CRGGH—Z01HG200362). CRGGH is also supported by National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), Center for Information Technology, and the Office of the Director at the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

## Availability of data and materials

The data that support the findings of this study are part of the The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium and are available from the authors upon reasonable request.

## Ethics approval and consent to participate

Not applicable

**Consent for publication**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Department of Computational Biology, USR 3756 CNRS, Institut Pasteur, Paris, France. <sup>2</sup>Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. <sup>3</sup>Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. <sup>4</sup>Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. <sup>5</sup>Division of Biostatistics, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA. <sup>6</sup>Center for Human Genetics Research, Massachusetts General Hospital, Boston, MA 02114, USA. <sup>7</sup>Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA.

Received: 2 August 2019 Accepted: 28 May 2020

Published online: 18 June 2020

**References**

1. Aschard H. A perspective on interaction effects in genetic association studies. *Genet Epidemiol.* 2016;40(8):678–88.
2. Frost HR, Shen L, Saykin AJ, Williams SM, Moore JH, Alzheimer's Disease Neuroimaging I. Identifying significant gene-environment interactions using a combination of screening testing and hierarchical false discovery rate control. *Genet Epidemiol.* 2016;40(7):544–57.
3. Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet Epidemiol.* 2012;36(3):183–94.
4. Figueiredo JC, Hsu L, Hutter CM, Lin Y, Campbell PT, Baron JA, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet.* 2014;10(4):e1004228.
5. Hutter CM, Chang-Claude J, Slattery ML, Pflugeisen BM, Lin Y, Duggan D, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer Res.* 2012;72(8):2036–44.
6. Kallberg H, Padyukov L, Plenge RM, Ronnelid J, Gregersen PK, van der Helm-van Mil AH, et al. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *Am J Hum Genet.* 2007;80(5):867–75.
7. Tyrrell J, Wood AR, Ames RM, Yaghootkar H, Beaumont RN, Jones SE, et al. Gene-obesogenic environment interactions in the UK biobank study. *Int J Epidemiol.* 2017;46(2):559–75.
8. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered.* 2007;63(2):111–9.
9. Manning AK, LaValley M, Liu CT, Rice K, An P, Liu Y, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients. *Genet Epidemiol.* 2011;35(1):11–8.
10. 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(3):201–10.
11. Rao DC, Sung YJ, Winkler TW, Schwander K, Borecki I, Cupples LA, et al. Multiancestry Study of Gene-Lifestyle Interactions for Cardiovascular Traits in 610 475 Individuals From 124 Cohorts: Design and Rationale. *Circ Cardiovasc Genet.* 2017;10(3):e001649.
12. Bentley AR, Sung YJ, Brown MR, Winkler TW, Kraja AT, Ntalla I, et al. Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat Genet.* 2019;51(4):636–48.
13. de Vries PS, Brown MR, Bentley AR, Sung YJ, Winkler TW, Ntalla I, et al. Multiancestry genome-wide association study of lipid levels incorporating gene-alcohol interactions. *Am J Epidemiol.* 2019;188(6):1033–54.
14. Feitosa MF, Kraja AT, Chasman DI, Sung YJ, Winkler TW, Ntalla I, et al. Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries. *PLoS One.* 2018;13(6):e0198166.
15. Sung YJ, Winkler TW, de Las Fuentes L, Bentley AR, Brown MR, Kraja AT, et al. A large-scale multi-ancestry genome-wide study accounting for smoking behavior identifies multiple significant loci for blood pressure. *Am J Hum Genet.* 2018; 102(3):375–400.
16. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet.* 2010; 11(12):843–54.
17. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47(3):291–5.
18. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet.* 2015;47(11):1228–35.
19. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet.* 2018;50(4):621–9.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.