European Journal of Human Genetics (2011) 19, 1292–1294

www.nature.com/ejhg

SHORT REPORT

CGene: an R package for implementation of causal genetic analyses

Peter J Lipman*[,1,2] and Christoph Lange[1,2]

The excitement over findings from Genome-Wide Association Studies (GWASs) has been tempered by the difficulty in finding the location of the true causal disease susceptibility loci (DSLs), rather than markers that are correlated with the causal variants. In addition, many recent GWASs have studied multiple phenotypes – often highly correlated – making it difficult to understand which associations are causal and which are seemingly causal, induced by phenotypic correlations. In order to identify DSLs, which are required to understand the genetic etiology of the observed associations, statistical methodology has been proposed that distinguishes between a direct effect of a genetic locus on the primary phenotype and an indirect effect induced by the association with the intermediate phenotype that is also correlated with the primary phenotype. However, so far, the application of this important methodology has been challenging, as no user-friendly software implementation exists. The lack of software implementation of this sophisticated methodology has prevented its large-scale use in the genetic community. We have now implemented this statistical approach in a user-friendly and robust R package that has been thoroughly tested. The R package 'CGene' is available for download at http://cran.r-project.org/. The R code is also available at http://people.hsph.harvard.edu/~plipman.

European Journal of Human Genetics (2011) **19,** 1292–1294; doi:10.1038/ejhg.2011.122; published online 6 July 2011

Keywords: causal modeling; statistical genetics; software

INTRODUCTION

The excitement over positive findings from recently published Genome-Wide Association Studies (GWASs) has been tempered by the difficulty in finding the location of the true causal disease susceptibility loci (DSLs), rather than markers that are correlated with the DSL.[1] Complicating the picture is that many recent GWASs have studied multiple phenotypes, which are often highly correlated.[2–4] This has made it very difficult to understand which associations that were discovered by GWASs are causal and which are seemingly causal, induced by phenotypic correlations. The ability to distinguish between causal genetic associations and seemingly causal associations induced by the intermediate phenotypes can provide important clues into the underlying genetic architecture of the disease. In addition, there is much interest in identifying endo-phenotypes or expression profiles that may lie in the 'genetic path' between the marker locus and the phenotype of interest to better understand how the genetic mechanisms influence the complex trait. The recent interest in understanding causal genetic pathways has led to the development of new statistical techniques that look to distinguish between direct and indirect causal genetic mechanisms. For quantitative and binary traits, VanSteenlandt *et al*[5] proposed a regression adjustment procedure that is applied to the quantitative phenotype of interest, adjusting for the potential presence of an association between the endo-phenotype and the test marker locus. Lipman *et al*[6] generalized this regression technique for age-at-onset (survival) phenotypes. The rejection of the null hypothesis of no genetic association by such a modified genetic association test implies a direct causal effect of the marker locus on the quantitative phenotype of interest, that is, an effect through pathways other than that of the intermediate phenotype.

However, so far, no software implementation exists for these causal genetic methods. With the R package 'CGene', we have provided such a tool. The package allows users to implement statistical techniques to understand the causal pathways between genetic markers and a primary outcome when an intermediate phenotype is also associated with both the marker the primary outcome. The package is available for download at http://cran.r-project.org/.[7] The R code is also available at http://people.hsph.harvard.edu/~plipman.

MATERIALS AND METHODS

The functions provided by 'CGene' enable investigators to test if a genetic marker is associated with a primary outcome through pathways other than that of an intermediate secondary phenotype. The functions allow for the primary outcome to be continuous or discrete as described by Vansteenlandt *et al*[5] or the primary outcome may be survival data modeled parametrically or semi-parametrically as described by Lipman *et al*.[6] The R package assumes the situation modeled with the causal directed acyclic graph (DAG) in Figure 1. In Figure 1, X represents the genetic marker, L represents the diagnostic criteria for the secondary phenotype K, T represents the target phenotype, U represents an unmeasured common cause, and P represents factors leading to population stratification (that have been controlled for in the design stage). As currently written, the package assumes population-based data. An example of the scenario modeled by the DAG can be found in respiratory

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA; [2]Department of Medicine, Harvard Medical School, Channing Laboratory, Boston, MA, USA
*Correspondence: Dr PJ Lipman or Dr C Lange, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA. Tel: +443 799 0171; E-mail: pjlipman@fas.harvard.edu
Received 1 February 2011; revised 26 April 2011; accepted 24 May 2011; published online 6 July 2011

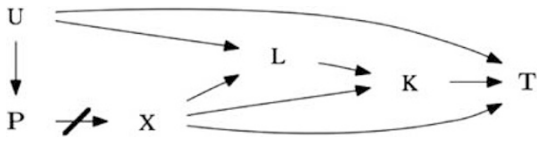

Figure 1 Causal DAG.

genetic epidemiology. There is currently much interest in genetic markers in the 15q25 locus, encompassing the cluster of nicotinic cholinergic receptor genes *CHRNA3*/*CHRNA5*/*CHRNB4*, which have been associated with chronic obstructive pulmonary disease (COPD) and smoking behavior.[3,8] It has been speculated that functional variants in this region may actually be primarily influencing nicotine addiction, as observed in recent GWASs on smoking intensity, and are thus only associated with COPD due to the well-known links between smoking and COPD.[9–11] To determine if the genetic marker influences COPD through pathways other than smoking, the COPD case–control binary variable has the role of target phenotype 'T', smoking, represented as cumulative exposure to tobacco smoke (ie pack-years smoked) or number of cigarettes smoked per day, has the role of secondary phenotype 'K', the well-known diagnostic criteria for smoking, such as gender, has the role of variable 'L'.

In order to test for the direct effect between the marker locus X to the target phenotype T in Figure 1, it is not proper to simply have marker locus X, secondary phenotype K, and diagnostic criteria L as covariates in the regression model (thus checking for the conditional independence of X and T, given K and L). This is because both secondary phenotype K and diagnostic criteria L are colliders in Figure 1. It is well known in causal methodology that having colliders as covariates in a regression model does not 'block' the path of interest, but, in fact, may induce a spurious relationship. Therefore, if we add secondary phenotype K and diagnostic criteria L into the model, the coefficient for the marker locus X variable will not only quantify the direct effect from the marker locus X to the target phenotype T, but will also quantify the 'opened' paths from marker locus X (to secondary phenotype K) to diagnostic criteria L to unmeasured common cause U to target phenotype T. Because of the existence of these colliders, standard regression techniques fail to properly quantify the effect of interest from marker locus X to target phenotype T.[6,12]

To avoid the problems caused by the confounders, we first look to quantify the direct effect of secondary phenotype K on target phenotype T. We then adjust the target phenotype by subtracting out the direct effect of the secondary phenotype. In order to properly quantify the direct effect of the secondary phenotype on the target phenotype, one must model the effect of the secondary phenotype K on target phenotype T while controlling for marker locus X and diagnostic criteria L to block all backdoor paths that could induce spurious associations. After subtracting out the effect of secondary phenotype K on target phenotype T, we then test if the genetic locus X is associated with the adjusted phenotype by running a simple univariate regression. This is possible because there are no open backdoor paths between genetic locus X and adjusted phenotype (and controlling for diagnostic criteria L and secondary phenotype K may induce spurious relationships as described above), thus testing for a direct causal effect, that is, through pathways other than the secondary phenotype. Score tests are used to determine the statistical significance of the parameters from the univariate regression, accounting for the adjustment of the primary phenotype.[6]

The functions in the R package 'CGene' allow for the marker genotypes to be in a matrix, where rows represent subjects and each column is a marker. The intermediate phenotype K is modeled using any family of a generalized linear model (GLM), as chosen by the investigator. The diagnostic criteria L may also be in a matrix, where each row represents a subject and each column is a different variable. When primary outcome T is continuous, the investigator may choose to model it using any family of a GLM. When T is survival data, separate functions exist for modeling it parametrically or semi-parametrically. When investigators use a parametric model for the primary outcome, any family allowed by the 'Survreg' function in the Survival package may be chosen. A separate function allows for investigators to model the primary outcome with a Cox Proportional Hazards model (semi-parametric). The package has been carefully tested and debugged. To assure its correctness, the empirical significance level was estimated under a variety of scenarios. The causal DAG as in Figure 1 was simulated under the null hypothesis of no direct effect from genetic marker X to primary phenotype T. Simulations showed that the proper type-1 α level was maintained under realistic scenarios (where genotype–phenotype effect sizes were roughly 1% and phenotype–phenotype effect sizes were between 5% and 10%) when the secondary phenotype K was either binary or continuous, and the primary phenotype was either a binary, continuous, or time-to-event outcome. We do note here that the methodology has an important limitation that the seemingly causal associations between the genetic marker and the target phenotype may be driven by a separate causal genetic locus that is correlated with the genetic marker tested.

RESULTS

Each function outputs a single *P*-value for each marker (column of X), testing whether there is a direct effect of marker X on primary outcome T through pathways other than that of secondary phenotype K. The function also outputs an effect size estimate, which is the regression coefficient of the genetic marker on the primary outcome, after the primary outcome has been adjusted for the effect of the secondary phenotype. This estimate is appropriate for population-based data, assuming no population substructure.

DISCUSSION

We have developed an R package 'CGene' to implement the genetic causal methodology developed by Vansteelandt *et al*[5] and Lipman *et al*.[6] The R package is currently available for download at http://cran.r-project.org/.[7] The functions provided by 'CGene' allow the researcher to test if a genetic marker is associated with a primary outcome, accounting for the presence of an intermediate secondary phenotype. The rejection of the null hypothesis of no genetic association by these methods implies a direct causal effect of the marker locus on the quantitative phenotype of interest, that is, an effect through pathways other than that of the intermediate phenotype. The functions allow for a wide range of models using different families of GLMs for the intermediate phenotype and a wide range of models for the primary outcome, which may be continuous, discrete, or survival data. The 'CGene' library provides the genetics community with a valuable set of tools for the identification of DSLs and ultimately the pathways underlying complex phenotypes and diseases.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by RO1MH087590 and R01MH081862. We thank the reviewers for their insightful comments. We also acknowledge the help of Dr Edwin K Silverman and Dr Mateusz Siedlinski.

1 Manolio TA: Genome-wide association studies and disease risk assessment. *N Engl J Med* 2010; **363**: 166–176.
2 Boezen HM: Genome-wide association studies: what do they teach us about asthma and chronic obstructive pulmonary disease? *Proc Am Thorac Soc* 2009; **6**: 701–703.
3 Pillai SG, Ge D, Zhu G *et al*: A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009; **5**: e1000421.
4 Wang J, Spitz MR, Amos CI *et al*: Mediating effects of smoking and chronic obstructive pulmonary disease on the relation between the CHRNA5-A3 genetic locus and lung cancer risk. *Cancer* 2010; **116**: 3458–3462.
5 Vansteelandt S, Goetgeluk S, Lutz S *et al*: On the adjustment for covariates in genetic association studies: a novel, simple principle to infer direct causal effects. *Genet Epidemiol* 2009; **33**: 394–405.
6 Lipman PJ, Liu KY, Muehlschlegel JD *et al*: Inferring genetic causal effects on survival data with associated endo-phenotypes. *Genet Epidemiol* 2011; **35**: 119–124.

7 R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0 2005.
8 Young RP, Hopkins RJ, Hay BA *et al*: Lung cancer gene associated with COPD: triple whammy or possible confounding effect? *Eur Respir J* 2008; **32**: 1158–1164.
9 Liu JZ, Tozzi F, Waterworth DM *et al*: Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; **42**: 436–440.
10 Thorgeirsson TE, Gudbjartsson DF, Surakka Ida *et al*: Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* 2010; **42**: 448–453.
11 Tobacco and Genetics Consortium: Genome-wide meta-analyses identify mul- tiple loci associated with smoking behavior. *Nat Genet* 2010; **42**: 441–447.
12 Rothman K, Greenland S, Lash T: *Modern Epidemiology*. Lippincott, Williams & Wilkins: Philadelphia, PA, 2008; 185–186.