European Journal of Human Genetics (2013) 21, 109–111

www.nature.com/ejhg

SHORT REPORT

REGENT: a risk assessment and classification algorithm for genetic and environmental factors

Daniel JM Crouch[1], Graham HM Goddard[1] and Cathryn M Lewis*[,1,2]

The identification of environmental and genetic factors that contribute to disease risk requires appropriate statistical methods and software that can integrate different sources of risk, provide statistical assessment of combined risk factors, and facilitate interpretation of this risk. We have developed an R package, REGENT, to calculate risks conferred by genetic factors and multilevel environmental factors. This is performed at a population level, with the option to also analyse individual-level data. REGENT incorporates variability in risk factors to calculate confidence intervals for risk estimates and to classify the population into different categories of risk based on significant differences from the baseline average member of the population. REGENT is an R package available from CRAN: http://cran.r-project.org/web/packages/REGENT. It will be of value to genetic researchers exploring the utility of the variants detected for their disorder, and to clinical researchers interested in genetic risk studies.
European Journal of Human Genetics (2013) **21,** 109–111; doi:10.1038/ejhg.2012.107; published online 6 June 2012

Keywords: genetic association; risk estimation; SNPs; environment; variability

INTRODUCTION

Genome-wide association studies have shed light on the genetic component for many common human diseases, identifying single-nucleotide polymorphisms (SNPs) which are associated with disorders. Risk is typically distributed across many genetic variants, each conferring a small increase. Although these findings represent a substantial advance in our understanding of genetic disease, the SNPs identified explain only a small proportion of the heritability of most common diseases,[1] and environmental risk factors also have a role. Assessing the distribution of risk for complex disorders in the population and predicting individual-level risk, while accounting for statistical uncertainty in these factors, are of great scientific and medical interest.

We have previously developed a statistical method for genetic risk prediction, which estimates a relative risk of disease from a panel of SNPs, together with a confidence interval (CI).[2] The method further allows us to classify genetic risk profiles into different risk categories, depending on whether the genotype combinations (or profiles) have overlapping CIs. The model defines a baseline multilocus SNP genotype profile, which has average risk. Any genotype profile whose CI overlaps with this baseline risk CI is classified as having average risk, as its risk is not statistically different from baseline. A genotype profile with CI which lies completely below that of the baseline profile CI is classified as being of reduced risk; similarly, profiles with CI completely above the CI for baseline risk are categorised as being of elevated risk. High risk profiles are those whose CIs lie above the CI of the first profile with elevated risk. The categorisation algorithm determines genotype combinations that are of low, average, elevated and high risk using both the estimated level of relative risk and the accuracy with which it is known. For a disorder where SNPs have relative risks (RRs) near one, or where the SNP risk is known with poor accuracy, little discrimination is possible and a large proportion of the population would be categorised as average risk. In contrast, SNPs conferring a higher risk that is accurately determined from large studies yield genotype profiles in which CIs do not overlap with the baseline profile. Discrimination is obtained from genotype profiles classified as having nonaverage risk.

We have designed an R package, REGENT (Risk Estimation for Genetic and Environmental Traits), to implement these methods through calculating the population distribution of disease risk, and categorising individuals according to their RR. REGENT calculates risk from genetic risk factors and also implements an extension of the method to environmental risk factors (manuscript in preparation). The package requires estimates of RR and population frequencies for the input factors, and these can readily be obtained from the genetic and epidemiological literature. Individuals shown to have different risks with statistical confidence are classified into different risk categories, and the overall percentage of the population belonging to each category is determined.

METHODS

Risk assessment in REGENT consists of two steps. First a population level analysis is performed using summary information on genetic and environmental factors in the function REGENTmodel. This establishes the population distribution of risk and determines risk categories. Then the results of the model can be applied to individual-level data risk in the function REGENTpredict, calculating relative risk of disease and assigning genotype profiles to risk categories.

REGENTmodel takes summary statistics routinely reported in research papers for SNPs and environmental factors, and uses these to generate multifactorial risk profiles for a large simulated population of individuals. The risk ratio for each profile is calculated, and an empirical CI is estimated by simulation, based on the standard errors of the input factors. All risks and CIs are scaled by the risk of the multifactorial profile closest to the mean risk of the simulated population, to provide RR with reference to this

[1]Department of Medical and Molecular Genetics, King's College London, London, UK; [2]MRC SGDP Centre, Institute of Psychiatry, King's College London, London, UK
*Correspondence: Dr CM Lewis, Department of Medical and Molecular Genetics, King's College London, 8th Floor Tower Wing, Guy's Hospital, Great Maze Pond, London SE1 9RT, UK. Tel: +1 44 20 7188 2601; Fax: +1 44 20 7188 2585; E-mail: Cathryn.lewis@kcl.ac.uk
Received 12 December 2011; revised 17 April 2012; accepted 24 April 2012; published online 6 June 2012

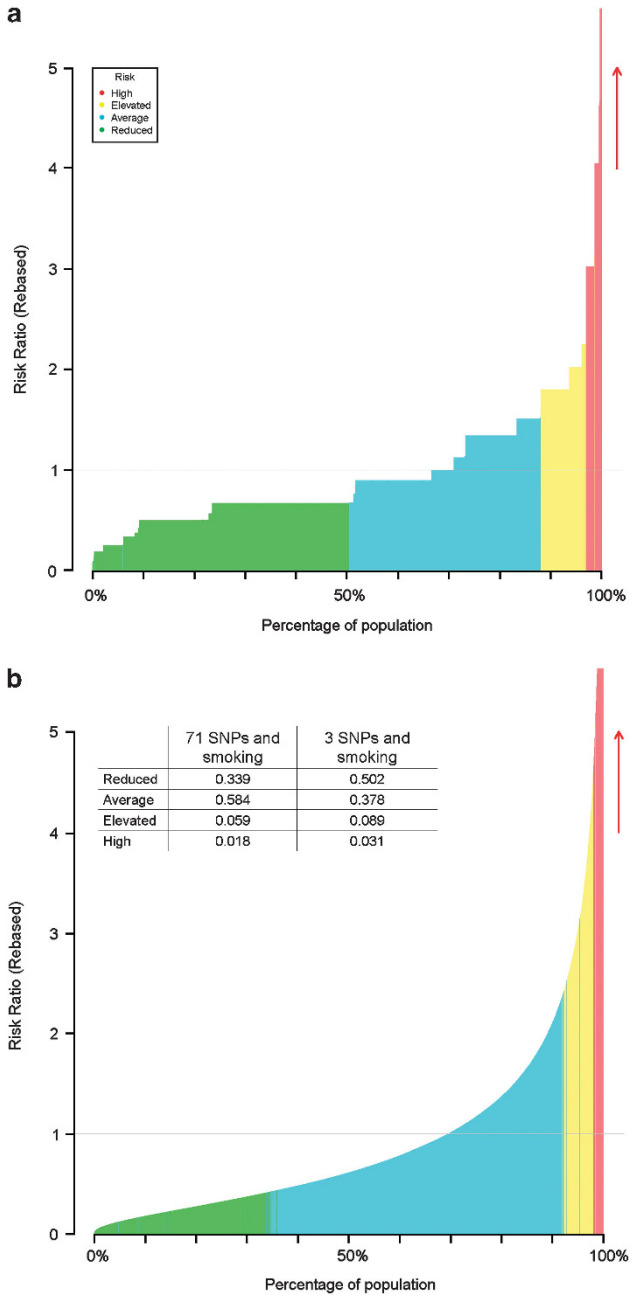

Figure 1 Population risk distribution (REGENTmodel) of Crohn's disease using (**a**) 3 SNPs and (**b**) 71 SNPs, and one environmental factor (smoking). Shading indicates the risk combinations classified as of average risk (blue), reduced risk (green), elevated risk (yellow) and high risk (red), and the proportion of the population in each category is tabulated.

baseline profile (which has RR = 1). The risk categories of low, average, elevated and high risk are then determined, by assessing the overlap of CIs with the baseline profile. REGENTmodel outputs the population level cumulative distribution and risk categories in graphical form (Figure 1). Text output of this distribution, the CI thresholds, the proportion of the population classified in each risk category and the area under the receiver operating characteristic curve (AUC) for the model are also provided. User-defined parameters control features such as CI width (default 95%).

To model the risk conferred by SNPs, users must provide (1) allelic or genotype RR, (2) the case–control sample sizes used to estimate these risks and (3) allele frequencies. These values are readily available in the research literature, for example, from meta-analyses of genome-wide association studies. For environmental factors included in the model, the user must provide (1) population exposure rates, (2) odds ratios and (3) standard errors for each level of exposure. An estimate is also required for the disease prevalence, together with a measure of its precision. REGENT can model genotype-specific risks, without being restricted to an allele-based model, and it allows for multilevel environmental risk factors. The model is multiplicative across risk factors, and therefore requires independence within and between the genetic and environmental risk factors included. Users must ensure this assumption is not violated with their chosen risk factors. For genetic risk factors, multiple SNPs within a gene are generally associated with disease. The user should select either the single SNP with strongest evidence of association or a subset of SNPs, which are not in linkage disequilibrium, and which contribute independently to disease risk (eg, three high risk SNPs in *NOD2* conferring risk of Crohn's Disease). Care must be taken when including environmental factors which themselves have a genetic component: for example, if obesity is included as an environmental risk factor in a model for type 2 diabetes (T2D), then a SNP in the *FTO* gene should not also be included, as it is associated with both T2D and obesity.[3]

REGENTpredict takes an individual's data of genotype and environmental factors. It calculates the relative risk and the absolute risk of disease, and then uses the results from REGENTmodel to determine the individual's risk category (low, average, elevated or high). The detailed level of output provided by REGENT will enable researchers to extract information relevant to their disorder. In addition to the built-in risk categorisation algorithm, REGENT output can be used for prespecified risk benchmarks. For example, what proportion of the population (or which of the individuals tested) has relative risk estimates above two? Or, for the 5% of the population at highest risk, what are the risk estimates and CIs?

System requirements

REGENT is computationally tractable for a personal computing system, with running time dependant on user-specified parameters and the number of risk factors modelled. REGENTmodel takes ~225 s for 10 SNPs using the default parameters on dual core 2.13Ghz processors (R running on single processor) and 2GB RAM. Analysis of 71 SNPs takes 8 h on this system. Running time will increase less than linearly with the number of risk factors, as a large proportion of the multifactorial profiles will be rare, and do not appear in the population simulations. Excess RAM can be utilised to provide gains in running time by increasing the number of multifactorial profiles held in memory during CI simulation. R 64-bit should be used when over 4GB RAM (2GB on 32-bit Windows systems) is required. REGENTpredict is not computationally intensive.

Application

REGENT was tested on SNPs associated with Crohn's Disease[4] and an environmental risk factor of smoking, which confers a twofold increased risk.[5] Analyses were run for three SNPs (in genes *NOD2*, *IL23R*, and *ATG16L1*) and then for all 71 SNPs. Allelic RRs were used for all SNPs except for *NOD2*, where a genotype-specific RR is appropriate.[6] Three SNPs and one environmental factor give 54 ($=3^3\times 2$) risk-factor profiles. The population distribution has a characteristic stepped output with some genotype–environment combinations occurring at >20% frequency (Figure 1). These factors enable 8.9% of the population to be classified as at elevated risk (ie, at a risk level significantly different from the baseline individual with RR = 1), and 3.1% classified at high risk. Using 71 SNPs gives $3^{71}\times 2 > 10^{34}$ risk profiles, and the population risk distribution appears continuous with a small proportion of the population at much increased RR. The blended colours at risk category boundaries occur because CI limits do not increase monotonically with increasing risk: they depend on factors such as allele frequency, exposure prevalence and estimate precision. With 71 SNPs modelled, the proportions of the population classified at an elevated and high risk drop to 5.9 and 1.8%. This decrease in model resolution is due to the low risk SNPs included (31 SNPs have RR below 1.15). For these SNPs, their signal for association with Crohn's disease in the model may be outweighed by the noise added by the imprecision of their RR estimate.

REGENT is available for download from http://cran.r-project.org/web/packages/REGENT. A user manual is included, describing in detail the modelling procedure and the assumptions of the model.

DISCUSSION

Applying our knowledge of risk factors that contribute to disease requires appropriate statistical methods and software that can integrate different sources of information, provide statistical assessment of combined risk, and allow interpretation of risk conferred. REGENT achieves all these aims in a flexible framework of risk assessment. The package will be of value to genetic researchers interested in exploring the utility of the variants detected for their disorder, and to clinical researchers performing the first generation of genetic risk assessment studies.

For any model, it is essential to understand underlying assumptions and data input requirements. The model applied in REGENT assumes that all genetic and environmental risk factors contribute independently to disease risk, with no gene–gene, gene–environment or environment–environment interaction present (although such interactions may be modelled through a multilevel risk factor in the environmental component of the model). REGENT allows the user to input parameter estimates from different studies, which have tested relevant risk factors. This approach enables the user to select the most appropriate study for each risk factor, for example, using genetic estimates from large meta-analyses of genome-wide association studies, and environmental estimates from large well-designed case–control or cohort studies. However, the method has limitations. The model provides no internal assessment of independence of the risk factors. Most genetic studies analyse single SNPs, and test for independent signals of association within a short genetic region, but not across regions. Our model follows the protocol of these studies by using each SNP as an independent predictor of disease status, which allows REGENT input parameters to be extracted from standard GWAS output. However, a joint analysis of SNPs from a GWA study may give risk estimates different from the single SNP analysis.[7] Such challenges could be circumvented by analysing genetic and environmental risk factors in large cohort studies to determine a good prediction model. In practice, few suitable studies are available and REGENT provides an *in silico* solution to combine genetic and environmental risk factors.

REGENT has two major applications of interest to genetic researchers. Firstly, it enables them to explore the population-level implications of genetic findings using output of the distribution of RRs in the population and the proportion of the population classified into different risk categories. Secondly, it can be used to estimate RRs of disease for specific individuals from their profile of genetic and environmental risk factors. For this analysis, it is crucial that studies chosen to provide model-input parameters are appropriate for the individuals to be analysed so that risk estimates are applicable.

The genetic RR estimation model used in REGENT is similar to that used by direct-to-consumer genetic testing companies such as 23andme and deCODEme, but it extends the functionality of their models by allowing environmental factors to be included and by calculating CIs for risk estimates. REGENT's risk categorisation algorithm provides a statistically valid measure of an individual's disease risk relative to a baseline member of the population, but also allows users to extract risk estimation relevant to their disorder. For example, REGENT can give information on the 10%, 5% or 1% of population at highest risk for whom specific interventions may be relevant. Any application of genetic prediction model through REGENT will need to be validated. Future applications of these methods may determine appropriate access to therapeutic interventions or lifestyle-modification programmes, ensuring that the decision-making process is based on statistically significant increases in risk.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Institute and Faculty of Actuaries, by Guy's and St Thomas' Charity and by a Medical Research Council Capacity Building PhD Studentship to GHMG. We thank Dr Paola Forabosco for helpful discussions.

1 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
2 Goddard GH, Lewis CM: Risk categorization for complex disorders according to genotype relative risk and precision in parameter estimates. *Genet Epidemiol* 2010; **34**: 624–632.
3 Frayling TM, Timpson NJ, Weedon MN *et al*: A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007; **316**: 889–894.
4 Franke A, McGovern DP, Barrett JC *et al*: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010; **42**: 1118–1125.
5 Calkins BM: A meta-analysis of the role of smoking in inflammatory bowel disease. *Dig Dis Sci* 1989; **34**: 1841–1854.
6 Lewis CM, Whitwell SC, Forbes A, Sanderson J, Mathew CG, Marteau TM: Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn disease. *J Med Genet* 2007; **44**: 689–694.
7 Stringer S, Wray NR, Kahn RS, Derks EM: Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS One* 2011; **6**: e27964.