GEGEINTOOL: A Computer-Based Tool for Automated Analysis of Gene-Gene Interactions in Large Epidemiological Studies in Cardiovascular Genomics

Oscar Coltell[1]
Department of Computing Languages and Systems. Universitat Jaume I
Castellon 12071, Spain
[1]RETIC «COMBIOMED». ISCIII
Madrid, 28007, Spain

José M. Ordovás
Nutrition and Genomics Laboratory. Tufts University
Boston, 2111, USA

Carmen Saiz
Department of Preventive Medicine. University of Valencia
Valencia, 46010, Spain

Manuel Forner
Department of Mathematics. Universitat Jaume I
Castellon 12071, Spain

Francisco Gabriel
Hospital Clínico Universitario. University of Valencia
Valencia, 46010, Spain

and

Dolores Corella[2]
Department of Preventive Medicine. University of Valencia
Valencia, 46010, Spain
[2]CIBER «Fisiopatología de la Obesidad y Nutrición». ISCIII.
Madrid, 28007, Spain

ABSTRACT

Current methods of data analysis of gene-gene interactions in complex diseases, after taking into account environmental factors using traditional approaches, are inefficient. High-throughput methods of analysis in large scale studies including thousands of subjects and hundreds of SNPs should be implemented. We developed an integrative computer tool, GEGEINTOOL (GEne-GEne INTeraction tOOL), for large-scale analysis of gene-gene interactions, in human studies of complex diseases including a large number of subjects, SNPs, as well as environmental factors. That resource uses standard statistical packages (SPSS, etc.) to build and fit the gene-gene interaction models by means of syntax scripts in predicting one or more continuous or dichotomic phenotypes. Codominant, dominant and recessive genetic interaction models including control for covariates are automatically created for each SNP in order to test the best model. From the standard outputs, GEGEINTOOL extracts a selected set of parameters (regression coefficients, p-values, adjusted means, etc.), and groups them in a single MS Excel Spreadsheet. The tool allows editing the set of filter parameters, filtering the selected results depending on p-values, as well as plotting the selected gene-gene interactions to check consistency. In conclusion, GEGEINTOOL is a useful and friendly tool for exploring and identifying gene-gene interactions in complex diseases.

Keywords: Genetic Epidemiology, Bioinformatics, Gene-Gene Interactions, Cardiovascular Diseases, Statistical Analysis, Genomics, Polymorphisms.

1. INTRODUCTION

Cardiovascular diseases are the first cause of death in the World, claiming 17.1 million lives a year according to the

World Health Organization [1]. In addition to the well known environmental risk factors for these diseases (tobacco smoking, high-saturated fat diet, sedentary lifestyle, etc.), currently, the genetic factors related to these diseases are increasing in relevance due to the recently published results from dozens of Genome-wide Association Analysis (GWAs) [2] [3] [4]. The results from these GWAs have provided a huge amount of information of new genes associated with intermediate (plasma lipid concentrations, blood pressure, inflammatory markers, etc.) and final (stroke, ischemic heart disease, etc.) cardiovascular phenotypes.

The publication of results from these GWAS, makes hundreds of researchers around the world to conduct studies to replicate associations of the main discovered Single Nucleotide Polymorphisms (SNPs) with the phenotypes of interest in their specific population studies. Up to date, these replication studies included a very low number of SNPs (from one to twenty). However, as the number of discovered SNPs is increasing, as well as the genotyping process performance, it is necessary to increase the number of SNPs to be included in replication studies. Moreover, these SNPs may interact with each other increasing or canceling the final effect on the corresponding phenotype. Nevertheless, to analyze these gene-gene interactions in epidemiological studies including thousands of participants and hundreds of SNPs, current methods of data analysis of gene-gene interactions using traditional approaches are inefficient. High-throughput methods of analysis for these cardiovascular epidemiology studies must be implemented

2. OBJECTIVES

Therefore, our aim was to develop an integrative computer tool, GEGEINTOOL (GEne-GEne INTeraction tOOL), for large-scale analysis of gene-gene interactions, in human studies of cardiovascular diseases including a large number of subjects, SNPs, as well as different intermediate (plasma lipid concentrations, blood pressure, inflammatory markers, etc.) and final (stroke, ischemic heart disease, etc.) phenotypes.

3. METHODOLOGY

That tool uses standard statistical packages (SPSS, etc.) to build and fit the gene-gene interaction models by means of syntax scripts in predicting one or more continuous (plasma total cholesterol, plasma LDL-cholesterol, triglycerides, systolic blood pressure, diastolic blood pressure, fasting glucose, etc.) or dichotomic (diabetes, hypertension, cardiovascular diseases, etc.) phenotypes. Codominant, dominant and recessive genetic interaction models including control for covariates are automatically created for each SNP in order to test the best model. Two-way or high order gene-gene interactions can be computed depending on the sample size and the researcher's choice. Environment variables (tobacco smoking, high-saturated fat diet, sedentary lifestyle, etc.), as continuous or as categorical, can be used as control variables for different confounders. From the standard outputs, GEGEINTOOL extracts a selected set of parameters (regression coefficients, p-values, adjusted means, etc.), and groups them in a single MS Excel Spreadsheet. The tool allows editing the set of filter parameters, filtering the selected results depending on p-values, as well as plotting the selected gene-gene interactions to check consistency.

We implemented the GEGEINTOOL in a real study to validate the tool and to compute the difficulties and limitations. The study in which the computer tool was tested was the PREDIMED Study [5]. The PREDIMED study (PREvención con DIeta MEDiterránea) is a parallel, multi-center, controlled, randomized clinical trial aimed at assessing the effects of the TMD on the primary prevention of cardiovascular disease. The trial is currently taking place with more than 7000 high-cardiovascular risk participants assigned to 3 intervention groups: (1) a traditional Mediterranean Diet (TMD) with virgin olive oil (TDM + VOO); (2) a TMD with mixed nuts (TMD + nuts); and (3) a low-fat diet. This study, started in October 2003, is carried out in several regions of Spain. The Institutional Review Board of the recruitment centers approved the study protocol and participants signed an informed consent. From October 2003 to March 2004, a total of 930 asymptomatic subjects at high risk for CHD, aged 55–80 years, were initially selected in 10 Spanish Primary Care Centers. They fulfilled at least one of the two following criteria: type 2 diabetes mellitus or three or more cardiovascular risk factors (smoking, hypertension, dislipemia, obesity, or family history of cardiovascular disease). Exclusion criteria were: history of cardiovascular disease; severe chronic illness; drug or alcohol addiction; history of food allergy or intolerance to olive oil or nuts; and any condition that may impair participation in the study. Participants' eligibility was based on the review of clinical records and a screening visit in the Primary Care Center by the physician.

The baseline examination included the administration of: a 14-item questionnaire, an extension of a questionnaire designed to assess the degree of adherence to the Mediterranean diet, the Minnesota Leisure Time Physical Activity questionnaire; and a 47-item general questionnaire assessing life-style, health conditions, socio-demographic variables, medical diagnoses, and medication use as previously reported [5] [6]. At baseline anthropometric data were also measured by standardized procedures. Fasting blood samples were obtained at baseline for each participant and serum glucose, cholesterol, and triglyceride concentrations were measured using standard enzymatic reagents (Trinder, Bayer Diagnostics, Tarrytown, NY, USA) adapted to a Cobas Mira automated analyzer (Hoffmann-La Roche, Basel, Switzerland). HDL cholesterol was quantified after

precipitation with phosphotungstic acid and magnesium chloride. LDL cholesterol was calculated by the Friedewald formula. Other biochemical determinations were carried out as previously detailed [5] [6].

Genomic DNA was extracted from buffy-coat with the MagNaPure LC DNA Isolation kit (ROCHE Diagnostics). We determined 100 SNPs in different genes related to cardiovascular diseases using 7900HT Sequence Detection system (Applied Biosystems) by fluorescent allelic discrimination TaqMan assays and OpenArray platforms. Quality control procedures were applied.

4. RESULTS

We implemented our tool to test gene-gene (epistasis) interactions in the PREDIMED study. The mean age of the participants was 67.7 years and prevalence of diabetes, hypertension and dislipemias was high as this is a high cardiovascular risk population. We selected 100 SNPs based on the literature y tested the Hardy-Weinberg equilibrium. Once checked this equilibrium we identified the minor allele for each polymorphism in order to create three variables for the three models of inheritance: additive, codominant and dominant. The 100 selected SNPs were not in linkage disequilibrium and were considered relevant tag SNPs associated with one or more of the selected intermediate cardiovascular risk phenotypes (plasma lipids, fasting glucose, blood pressure, inflammation markers and anthropometric measurements). Figure 1 shows the GEGEINTOOL workflow.

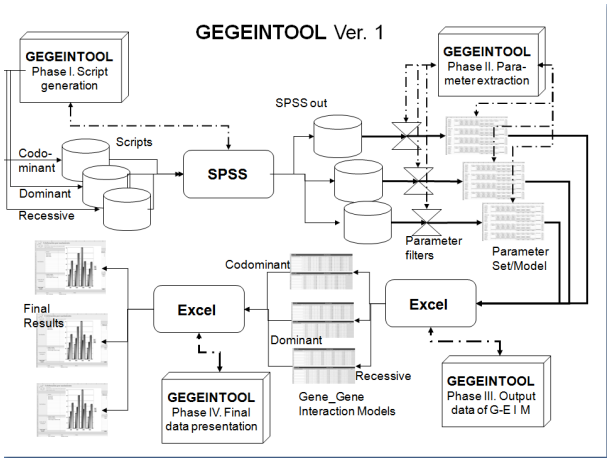

Figure 1. GEGEINTOOL Workflow

We first fitted first order gene-gene interaction models involving two SNPs. To compute these models the computer tool has two options: a) To test every combination of the 100 SNPs two by two; or b) To select only the biologically related SNPs determining the same phenotypical trait. The first option involves more computation time and has the limitation of increasing false positive results because of the multiple comparisons. The advantage is the complete screening of all the analyzed variants. In term of correction of the P values for multiple comparisons, the GEGEINTOOL allows to customize the selected P value (P<0.05; P<0.01; p>0.001; P<0.0001, etc) in order to consider a gene-gene interaction term statistically significant. Currently there is not complete agreement among researcher on the best algorithm to correct the P value [8] [9] [10], thus the GEGEINTOOL shows all P values and the researcher can select by order the best ones.

Taking into account that in a gene*gene interaction the effect of one genetic variant in determining a phenotype is modified by the second variant. It is interesting to depict the magnitude of effects and to test the consistency between the different models of inheritance. Our tool conducts figures for the selected gene-gene interactions. In addition it computes both crude and adjusted gene-gene interaction means for continuous variables. The main variables to adjust for are gender, clinical conditions, age, tobacco smoking, diet and physical activity. GEGEINTOOL allows comparing the P values before and after adjustment for a step by step control for covariates.

In addition to the testing of first order gene-gene interactions, GEGEINTOOL is able to conduct higher order gene-gene interaction models. We have computed second and third order interaction models in our data set and we have observed the advantages and the limitations of these models.

When implementing the GEGEINTOOL in the PREDIMED Study, we first analyzed the 100 SNPs in 3000 participants randomly selected. This sample constitutes the training sample and then we have the rest of the participants as an internal replication sample to check the consistency of the first identified gene-gene interactions.

After having applied the GEGEINTOOL to the search of statistically significant gene-gene interactions in the PREDIMED study in determining intermediate cardiovascular phenotypes, we have obtained dozens of statistically significant first order interaction terms among SNPs related with the dependent phenotype. Due to limitations in sample size, we have obtained less statistically significant second and third order gene-gene interactions. The most interesting gene-gene interactions are selected for new genotyping of the involved SNPs and conducting replication in the other PREDIMED participants. Moreover, we are also studying how the gene-gene interactions found are homogeneous or heterogeneous across the different strata of clinical conditions (diabetes, hypertension), gender age-groups or environmental factors (smokers, drinkers, categories of Mediterranean diet adherence, etc). GEGEINTOOL is a

easy tool that allows a wide range of customization of biomedical analysis depending on the requirement of the specific epidemiological study and the measured variables of interest.

5. CONCLUSIONS

In conclusion, GEGEINTOOL is a very useful and friendly tool for exploring and identifying gene-gene interactions in cardiovascular diseases for biomedical researchers using standard statistical packages of statistical analysis.

ACKNOWLEDGEMENTS

This work has been partially funded by the following grants: GEWIMICS (SAF2009-12304, MICINN), RETIC COMBIOMED (RD07/0067/0006, ISCIII-FIS), BEST/2010/211 and BEST/2010/032 (GVA), ACOMP/2011/145 (GVA), CS2010-AP-111 (GVA), CNIC06, CIBER "Fisiopatología de la Obesidad y Nutrición" (ISCIII-FIS). CIBERobn is an initiative of ISCIII and fondos Europeos para el desarrollo regional (FEDER).

REFERENCES

[1] The World Health Organization. Cardiovascular diseases. World Heart Day 2010. http://www.who.int/cardiovascular_diseases/en/..

[2] Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466 (7307):707-13.

[3] Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011 6;43(4):333-8.

[4] Corella D, Ordovas JM. Nutrigenomics in cardiovascular medicine. Circ Cardiovasc Genet. 2009;2(6):637-51.

[5] Martínez-González MA, Corella D, Salas-Salvadó J, Ros E, Covas MI, Fiol M, et al. Cohort Profile: design and methods of the PREDIMED study. Int J Epidemiol. 2010 Dec 20 (in press).

[6] Estruch R, Martínez-González MA, Corella D, Salas-Salvadó J, Ruiz-Gutiérrez V, Covas MI, et al. Effects of a Mediterranean-style diet on cardiovascular risk factors: a randomized trial. Ann Intern Med. 2006;145(1):1-11.

[7] Corella D, Carrasco P, Fitó M, Martínez-González MA, Salas-Salvadó J, Arós F, Lapetra J .Gene-environment interactions of CETP gene variation in a high cardiovascular risk Mediterranean population. J Lipid Res. 2010;51:2798-807.

[8] Li J, Horstman B, Chen Y. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. Bioinformatics. 2011;27(13):i222-i229.

[9] Chen M, Cho J, Zhao H. Detecting epistatic SNPs associated with complex diseases via a Bayesian classification tree search method. Ann Hum Genet. 2011;75(1):112-21.

[10] Wan X, Yang C, Yang Q, Xue H, Tang NL, Yu W. Detecting two-locus associations allowing for interactions in genome-wide association studies. Bioinformatics. 2010;26(20):2517-25.