BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Prediction model

Volume 11(3)

A method to associate all possible combinations of genetic and environmental factors using GxE landscape plot

Satoshi Nagaie[1], Soichi Ogishima [1], Jun Nakaya[1,2] & Hiroshi Tanaka[1,3]

[1]Dept of Bioclinical Informatics, Tohoku Medical Megabank Organization, Tohoku University; [2]Medical IT Center, School of Medicine, Tohoku University; [3]Dept. of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University; Soichi Ogishima - Email: ogishima@sysmedbio.org; *Corresponding author

Received March 17, 2015; Accepted March 21, 2015; Published March 31, 2015

Abstract:
Genome-wide association studies (GWAS) and linkage analysis has identified many single nucleotide polymorphisms (SNPs) related to disease. There are many unknown SNPs whose minor allele frequencies (MAFs) as low as 0.005 having intermediate effects with odds ratio between 1.5~3.0. Low frequency variants having intermediate effects on disease pathogenesis are believed to have complex interactions with environmental factors called gene-environment interactions (GxE). Hence, we describe a model using 3D Manhattan plot called GxE landscape plot to visualize the association of p-values for gene-environment interactions (GxE). We used the Gene-Environment iNteraction Simulator 2 (GENS2) program to simulate interactions between two genetic loci and one environmental factor in this exercise. The dataset used for training contains disease status, gender, 20 environmental exposures and 100 genotypes for 170 subjects, and p-values were calculated by Cochran-Mantel-Haenszel chi-squared test on known data. Subsequently, we created a 3D GxE landscape plot of negative logarithm of the association of p-values for all the possible combinations of genetic and environmental factors with their hierarchical clustering. Thus, the GxE landscape plot is a valuable model to predict association of p-values for GxE and similarity among genotypes and environments in the context of disease pathogenesis.

Abbreviation: GxE: Gene-environment interactions; GWAS: Genome-wide association study; MAFs: Minor allele frequencies; SNPs: Single nucleotide polymorphisms; EWAS: Environment-wide association study; FDR: False discovery rate; JPT+CHB: HapMap population of Japanese in Tokyo, Japan + Han Chinese in Beijing, China.

Background:
There are two main types of research methodologies that lead to the identification of hundreds of genetic variants associated with disease onset. One of them is GWAS, which has emerged as a powerful and successful tool to identify common human disease alleles by using high-throughput genotyping technology **[1].** GWAS aims to detect common variants with small effects. The other is linkage analysis, which aims to establish linkages between genes using family relationships **[2].** The linkage approach is often used in discovering rare variants with major effects. However, low-frequency variants with intermediate effects have not been captured by GWAS and linkage analysis, due to insufficient frequencies and effect sizes **[3].** These variants are expected to have complex interactions with environmental factors called gene-environment interactions (GxE). GxEs are not mere additive or synergistic interactions, but are complex interactions. For a very specific GxE interaction, the association p-value and risk of disease is high. Le Marchand et al. showed that the relative risk of colorectal cancer was 8.8 for specific combination of environmental factors (smoking and preference for well-done meat) and metabolic enzymes (CYP1A2 or NAT2 rapid or slow metabolizers), unlike conventional methods where the risk is calculated by the product of individual factors **[4].**

As opposed to GWAS where the disease risk is attributed to genetic factors, Butte et al. proposed an environment-wide association study (EWAS) approach **[5].** In addition to using both the environmental risk factors obtained from EWAS and the risk SNPs from GWAS, the combination of SNP (rs13266634) and trans-β- carotene was a significant GxE item correlated to type 2 diabetes. The per-risk-allele effect sizes in subjects with low serum levels of trans-β- carotene were 40% greater than the marginal effect size. EWAS offers an unbiased consideration of environmental and genetic factors that is useful in identification of larger and more relevant effect sizes for disease associations **[6].** However, Butte et al. showed a Manhattan plot for genome-wide genetic factors, but for very limited environmental factors.

We describe a prediction model using 3D Manhattan plot called GxE landscape plot, in negative logarithm of significance values associated with disease pathogenesis for genotype factors and environmental factors with their hierarchal clustering as a prediction model. The GxE landscape plot enables us to comprehensively visualize negative logarithm of p-values for combinations of associated genetic and environmental factors. This is a useful model to predict similarity of genotypes and environments in the context of association p-values with disease using hierarchical clustering.

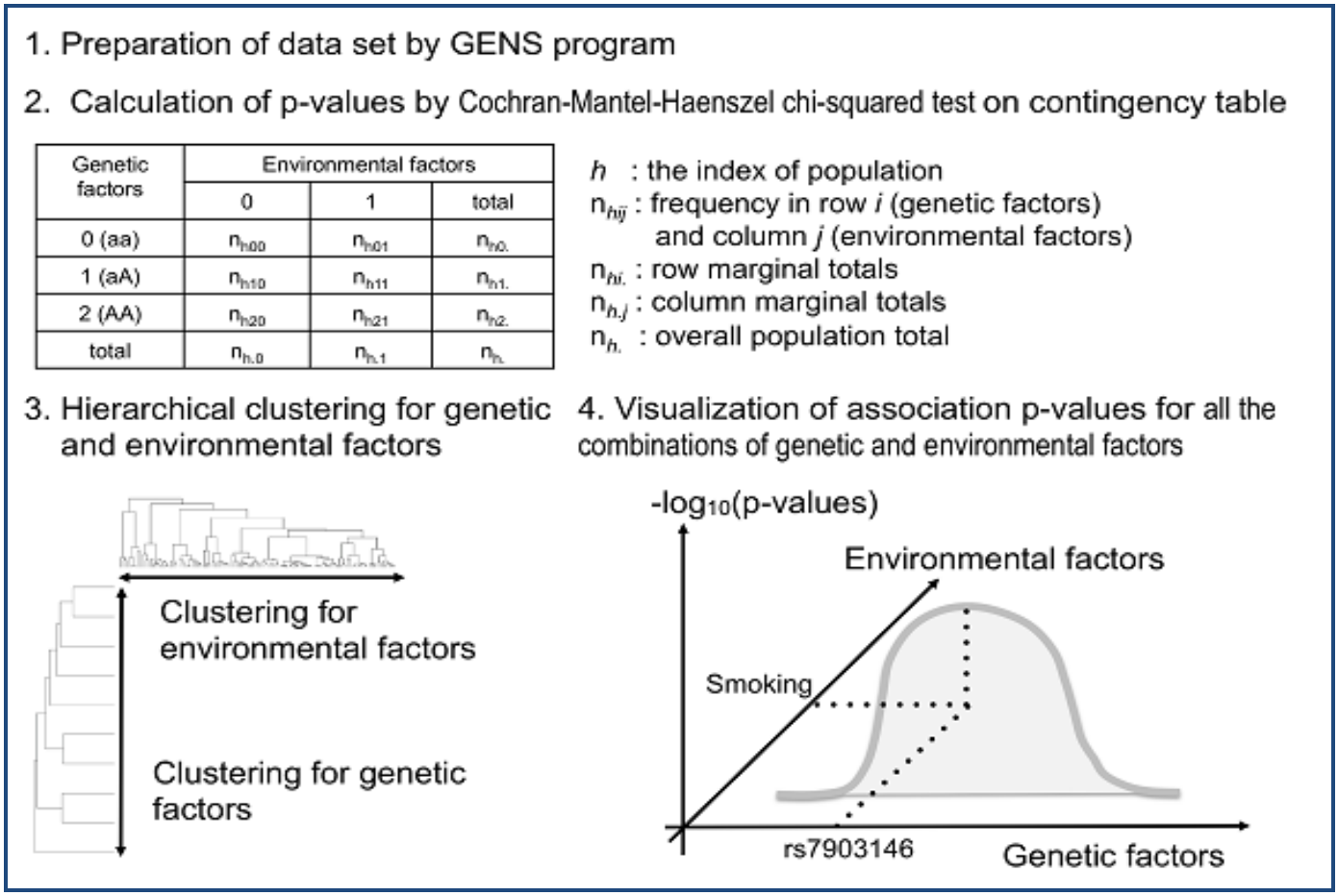

Figure 1: Overview of methodology for creating GxE landscape plot. Chart of the 4 steps that have been used to generate GxE landscape plot with simulated data using GENS2 program: **(1)** preparation of data set by GENS program, **(2)** calculation of p-values by Cochran-Mantel-Haenszel chi-squared test on 3 x 2 x 2 contingency table, **(3)** hierarchical clustering for genetic and environmental factors, **(4)** visualization of association p-values for all the combinations of genetic and environmental factors.

Methodology:

Overview of methodology for creating GxE landscape plot

There are 4 steps to create the GxE landscape plot **(Figure 1).** (Step 1) Dataset creation. A data set should contain a disease status, gender, environmental exposures and genotypes for each subject. In this study, to prove validity of our prediction model, we generated a data set by using simuGWAS, which was developed in simuPOP **[7]** and GENS2 program **[8].** We iterated this simulation 20 times using GENS2. (Step 2) The association p-values with disease were calculated for all the possible combinations of genetic and environmental factors by using the Cochran-Mantel-Haenszel chi-squared test on 3 x 2 x 2 contingency tables. (Step 3) Clustering analysis was applied for calculated negative logarithm of p-values for genotypes and environmental factors. (Step 4) The GxE landscape plot was generated plotting negative logarithm of significance level for all possible combinations of genotype and environmental factors. In the GxE landscape plot, significant p-values are reflected as peaks, and insignificant p-values as plateaus.

Training dataset creation by simulation of gene-gene and gene-environment interactions

GENS2 is a program based on data with realistic patterns of linkage disequilibrium, and exerts no constraints on the

number of individuals to be simulated or the number of non-predisposing genetic/environmental factors to be considered. GENS2 tool can simulate gene-environment and gene-gene interactions.

We simulated interactions among genetic loci and environmental factors using GENS2. Population data for the genomic model (JPT+CHB chr7) was downloaded from the HapMap3 **[9]** database and converted by simuGWAS for use in the GENS2 program. This population data for the simulation of genomic factors was set to be the initial population. In the simulation process, we utilized 170 sample sizes. Also a gene-environment model was used, with the following parameters: disease predisposing loci (DPL): rs1881690, rs1979600, rs4960568, rs6972501, rs7793905, rs936997; high and low risk allele; dominance parameter; relative risk for high risk homozygote; environmental parameters odds ratio; and disease environmental variable distribution parameters: mean, standard deviation, disease penetrance. If the disease status value is greater than that of 75th percentile, we set the value to 1; otherwise, 0. We removed 23,277 SNPs not possessing the 3 genotypes, out of 75,320 SNPs. In total, 100 SNPs were used for our analysis; 6 SNPs were associated with disease and the other 94 SNPs were randomly selected.

We iterated this simulation 20 times to obtain 20 populations. We assumed that 20 populations were divided into two major population groups, and we made each major population group have a specific gene-environment interaction to cause disease.

Calculation of p-values

To detect SNPs associated with environmental factors, we applied Cochran-Mantel-Haenszel chi-squared test to estimate the statistical significance of gene-environment interaction using the R 3.1.2 statistical software with the Bioconductor package **[10, 11].** For each statistical test, we obtained p-values from multiple hypothetical tests that were adjusted by a false discovery rate (FDR) **[12].**

Creating GxE landscape plot

We created a novel 3D Manhattan plot, called a GxE landscape plot, of the negative logarithm of associated significance values with disease pathogenesis for genotype factors and environmental factors with their hierarchal clustering as a prediction model. Hierarchical clustering was applied on 100 genetic factors (SNPs) and 20 environmental factors, performed using the R statistical software with Pearson's correlation coefficient as a similarity index and using the complete linkage method as an agglomeration. Negative logarithms of p-values were corrected for multiple comparisons by using Benjamini and Hochberg's method.

Discussion:

Figure 2 shows a prediction model with a 3D Manhattan plot called GxE landscape plot showing negative logarithm of p-values for all the possible combinations of genotypes and environmental factors. The genotypes and environmental factors were hierarchically clustered. The genetic and environmental factors were roughly clustered into two groups in combination of genetic and environmental factors depicted as two peaks surrounded by yellow ellipse in **Figure 2 (A)** and **(B).** We iterated GENE2 simulation 20 times to obtain 20 populations. The 20 populations were divided into two major population groups. This is simulated to make each major population group with specific gene-environment interactions to cause the disease.

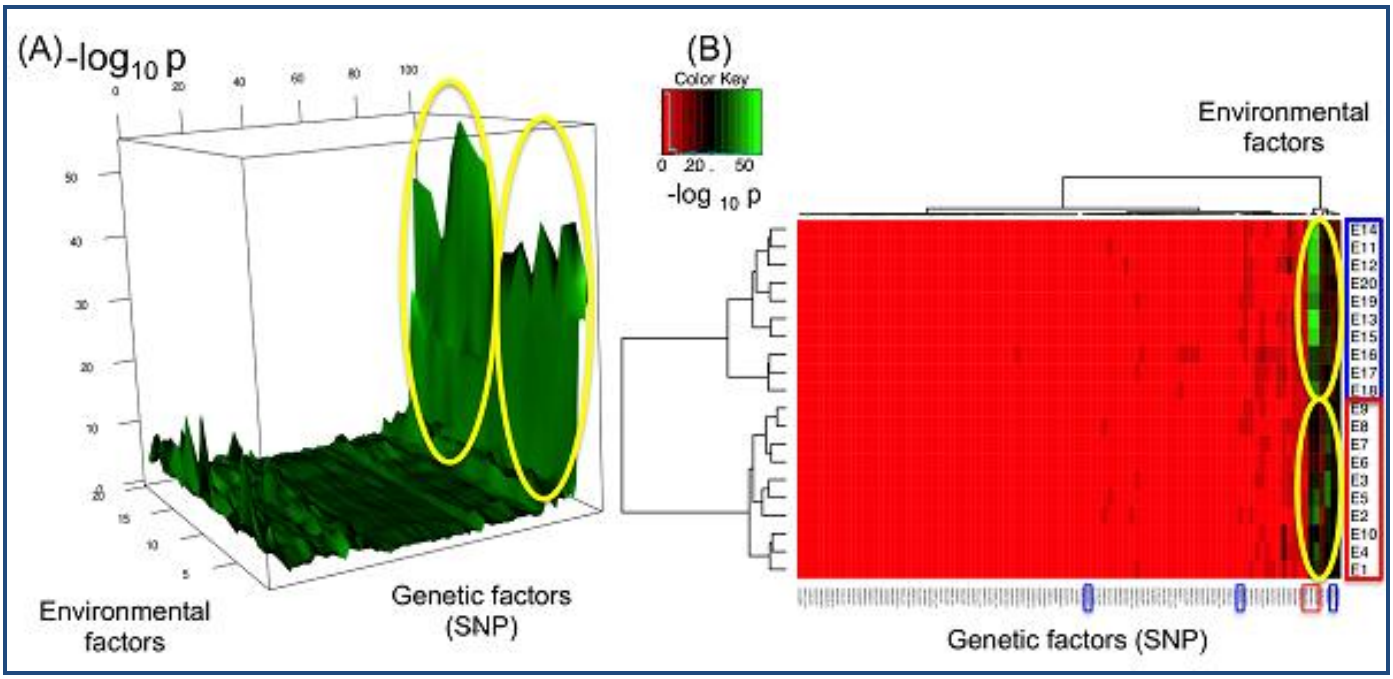

Figure 2: GxE landscape plot. 3D GxE landscape plot of negative logarithm of the association p-values for all the possible combinations of genetic and environmental factors with their hierarchical clustering. Genetic and environmental factors were roughly clustered into two groups in combination of genetic and environmental factors, depicted as two peaks surrounded by yellow ellipse. **(A)** GxE landscape plot with hierarchical clustering for genetic and environmental factors. **(B)** Hierarchical clustering of genetic factors (SNPs) and 20 environmental factors. The data are shown in a table format, in which rows represent

individual environmental factors and columns represent individual SNPs. The color in each cell reflects the negative logarithm of the association p-values for combinations of genetic and environmental factors. As for genetic factors, except for 2 SNPs, other SNPs were clustered into two groups surrounded by blue and red squares. As for environmental factors, all the environmental factors were also clustered into two groups clearly surrounded by blue and red squares.

All SNPs (6 SNPs) except 2 SNPs were clustered into two groups for genetic factors; 4 SNPs out of 6 SNPs (66.6% of SNPs) were correctly clustered on GxE landscape plot. Clustering of genetic factors is considered to reflect linkage disequilibrium among genes. It is observed that all the environmental factors were clustered into two groups clearly; 20 environmental factors out of 20 environmental factors (100% of environmental factors) were correctly clustered on GxE landscape plot. Thus, our model can predict underlying cluster of populations with high accuracy (100% for environmental factors, 66.6% for genetic factors). It showed that the model predicts similarity of genetic factors and environmental factors associated with the disease. It should be noted that the p-values obtained in this study are based on the pretext that the confounding factors between the genotype and the environment are removed. The model GxE landscape plot allows us to clearly visualize the distinctive features of the populations, and also allows us to detect novel genetic and environmental factor interactions.

The Butte's EWAS approach is a comprehensive testing and screening manner for gene-environment interactions **[5]**, but is a limited visualization for environmental factors based on Manhattan plot. On the other hand, this method is a comprehensive visualization of association p-values for all the possible genomic and environmental factors. The GxE landscape plot enables us to overview landscape of association significance level for GxE.

There is a limitation in the process of data analysis at the 75th percentile. It should be noted that the value is set at 1 if the disease status value is greater than the value of 75th and 0 otherwise for the statistical test. However, for a more accurate representation of the data, it would be more desirable to use dispersion values instead. Dispersion values provide more detailed information about the disease status.

The slope of the landscape is equivalent to the rate of change of the p-values. It is possible to infer the disease pathogenesis using the change of slope in the landscape. For example, if the peak of the landscape is steep, and the others low; it can be recommended for the subject to refrain from partaking in the environmental factor that results in high peaks.

Conclusion:
GWAS and linkage analysis have identified many SNPs related to several diseases. However, there remain many unknown SNPs whose MAFs are low having intermediate effects. Low frequency variants having intermediate effects on disease pathogenesis are believed to have complex GxE. We describe a model using 3D Manhattan plot called GxE landscape plot to visualize association p-values for GxE. We used GENS2 to simulate interactions between two genetic loci and one environmental factor, and p-values were calculated by Cochran-Mantel-Haenszel chi-squared test on simulation data. We thus created a 3D GxE landscape plot of negative logarithm of the association p-values for all the possible combinations of genetic and environmental factors with their hierarchical clustering. The GxE landscape plot is a valuable model to predict similarity among genotypes and environments in the context of association p-values with disease pathogenesis.

Acknowledgment:
This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References:
[1] Welter *et al. Nucleic Acids Res*. 2014 **42:** D1001 [PMID: 24316577]
[2] Risch *et al. Science* 1996 **273**: 1516 [PMID: 8801636]
[3] Manolio *et al. Nature* 2009 **461**: 747 [PMID: 19812666]
[4] Marchand *et al. Cancer Epidemiol Biomarkers Prev*. 2001 **10**: 1259 [PMID: 11751443]
[5] Patel *et al. PLoS One* 2010 **5:** e10746 [PMID: 20505766]
[6] Patel *et al. Hum Genet*. 2013 **132:** 495 [PMID: 23334806]
[7] Peng *et al. Bioinformatics*. 2005 **21**: 3686 [PMID: 16020469]
[8] Pinelli *et al. BMC Bioinformatics*. 2012 **13:** 132 [PMID: 22698142]
[9] The International HapMap3 Consortium Nature. 2010 **467:** 52 [PMID: 20811451]
[10] http://www.r-project.org
[11] http://www.bioconductor.org
[12] Benjamini *et al. J Royal Stat Soc Ser*. 1995 **289**-300 [PMID: N/A]

Edited by P Kangueane
Citation: ***Nagaie et al.*** Bioinformation 11(3): 161-164 (2015)