BMC Proceedings

PROCEEDINGS **Open Access**

Genetic association analysis for common variants in the Genetic Analysis Workshop 18 data: a Dirichlet regression approach

Osvaldo Espin-Garcia[1,2], Xiaowei Shen[1,2], Xin Qiu[1], Yonathan Brhane[3], Geoffrey Liu[4,5], Wei Xu[1,5*]

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

We propose a genetic association analysis using Dirichlet regression to analyze the Genetic Analysis Workshop 18 data. Clinical variables, arranged in a longitudinal data structure, are employed to fit a multistate transition model in which the transition probabilities are served as a response in the proposed analysis. Furthermore, a gene-based association analysis via penalized regression is implemented using the markers at a single-nucleotide polymorphism level that we previously identified via nonpenalized Dirichlet regression.

Background

Genetic association analyses have had tremendous successes in recent years; however, most of these analyses were based on binary or continuous responses. Thus we propose a multivariate response vector indicating probabilities of transitions to predefined hypertensive states. This enables us to reflect the inherent uncertainty involved in the probability that a patient will transfer to a given state.

An important feature of our approach is the incorporation of prehypertension as an intermediate state. As Winegarden argues, prehypertension blood pressure in young patients helps predict the development of hypertension [1].

Methods

Definition of response

We defined a response summarizing the phenotype information into a vector that will be used in a genetic association analysis. The response is defined as a 3-dimensional vector of probabilities $\gamma = (\gamma_1, \gamma_2, \gamma_3)\,, \sum^{\gamma_j} = 1\,(\mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3)\,, \sum^{\mathrm{y}_\mathrm{j}} = 1,$ such that each element measures the probability of a transition to a blood pressure level (normotensive, prehypertensive, or hypertensive) given the previous level.

The analysis was done without the knowledge of the underlying simulation model and we used the real phenotype data only.

Data quality control

Data quality control was performed in PLINK [2]. We only considered the data from chromosome 3 for analysis. We used a call rate for individuals of 95%, a Hardy-Weinberg disequilibrium test at a significance level of 1×10^{-6}, and a missing rate of 95% for each marker. Markers with a minor allele frequency of at least 5% were retained for analysis. Additionally, all individuals' time points with at least 1 missing clinical variable were excluded.

Multistate transition model

We describe hypertension, our trait of interest, using a 3-state model based on recorded blood pressure levels for each individual at each examination. The states are defined as follows: normal blood pressure (state 1) when the systolic blood pressure is less than 120 mm Hg and diastolic blood pressure is less than 80 mm Hg; prehypertension (state 2) when the blood pressure level is not in state 1, the systolic blood pressure is less than 140 mm Hg, and the diastolic blood pressure is less than 90 mm

* Correspondence: wxu@uhnres.utoronto.ca
[1]Department of Biostatistics, Princess Margaret Cancer Centre, 610 University Ave., Toronto, ON, M5G 2M9, Canada
Full list of author information is available at the end of the article

Hg; and hypertension (state 3) for all other cases. Also, if a patient used antihypertensive medication, the state assigned at that examination is hypertension (state 3) regardless of the recorded blood pressure levels. Once the states are defined, we consider a multistate transition model; it is important to note that all 9 transitions are possible.

Our interest in transition models lies in estimating of the transition probabilities as defined in Kalbfleisch and Lawless [3] which are given by

$$P(S_i\,(t) = j \mid S_i\,(t-1) = l, x_i(t-1)) = \gamma_{ilj}(t), l, j \in \{1, 2, 3\}$$

where $\{S_i\,(r)\,, r = 1, 2, \ldots\}$ and $\{x_i\,(r) = (x_{i1}\,(r)\,, \ldots, x_{ip}\,(r)), r = 1, 2, \ldots\}$ denote the observed state and the covariates for subject i at the r^{th} examination respectively.

This model takes advantage of the longitudinal data structure and the definition of the response follows naturally. To estimate the transition probabilities, we fit a multinomial regression model, based on covariates (gender, smoking status and age) and the state at the previous examination.

To get expressions for $\gamma_{il} = (\gamma_{il1}, \gamma_{il2}, \gamma_{il3})\,, l = 1, 2, 3$, we consider a generalized logit model of the form

$$\log\left(\gamma_{ilj}/\gamma_{ill}\right) = z_{il}\gamma_{lj}, j = 1, 2, 3, j \neq l$$

besides, $1 = \gamma_{ill} + \sum_{j=1,j\neq l}^{3} \gamma_{ilj} = \gamma_{ill}\left(1 + \sum_{j=1,j\neq l}^{3} \exp(z_{il}\gamma_{lj})\right)$, where $z_{il} = (x_i\,(l)\,, time)$ is the observed vector of covariates for subject i plus a categorical variable denoting the effect of examination time in the model (and possible interactions), and γ_{lj} is the vector of coefficients for the corresponding multinomial regression model.

Thus, a transition probability matrix (TPM) is defined for each individual as follows

$$\text{TPM}_i = \begin{pmatrix} \dfrac{1}{1 + \sum_{j=2}^{3} \exp(z_{i1}\gamma_{1j})} & \dfrac{\exp(z_{il}\gamma_{12})}{1 + \sum_{j=2}^{3} \exp(z_{i1}\gamma_{1j})} & \dfrac{\exp(z_{il}\gamma_{13})}{1 + \sum_{j=2}^{3} \exp(z_{i1}\gamma_{1j})} \\ \dfrac{\exp(z_{i2}\gamma_{21})}{1 + \sum_{j=1,j\neq 2}^{3} \exp(z_{i2}\gamma_{2j})} & \dfrac{1}{1 + \sum_{j=1,j\neq 2}^{3} \exp(z_{i2}\gamma_{2j})} & \dfrac{\exp(z_{i2}\gamma_{23})}{1 + \sum_{j=1,j\neq 2}^{3} \exp(z_{i2}\gamma_{2j})} \\ \dfrac{\exp(z_{i3}\gamma_{31})}{1 + \sum_{j=1}^{2} \exp(z_{i3}\gamma_{3j})} & \dfrac{\exp(z_{i3}\gamma_{32})}{1 + \sum_{j=1}^{2} \exp(z_{i3}\gamma_{3j})} & \dfrac{1}{1 + \sum_{j=1}^{2} \exp(z_{i3}\gamma_{3j})} \end{pmatrix}$$

Therefore, the response for subject i is a row taken from TPM_i and is determined by conditioning on the patient's last available observed state and covariates.

Dirichlet regression

Once the response is modeled our objective is to determine whether there is an association between it and the genotypes. We assess this association using Dirichlet regression [4], which suits this response structure. The advantage of this approach lies in its tractability in dealing with the proposed response. It also allows a more comprehensive understanding of the genetic effect on the expression of hypertension, and therefore in its possible interpretation. For instance, if a signal was detected for a marker, it would suggest an association between the marker and the transition of blood pressure states jointly rather than a single level. Therefore, the Dirichlet approach is more informative in the sense of explaining the plausibility of each defined state.

To relate the genetic information and the defined response under a Dirichlet regression approach, the likelihood given each individual's vector of covariates, $\mathbf{s}_i$, is

$$L = \prod_{i=1}^{n} \left\{ \Gamma\left(\Lambda(\mathbf{s}_i)\right) \prod_{j=1}^{3} \frac{\gamma_{ij}^{\lambda_j(\mathbf{s}_i)-1}}{\Gamma\left(\lambda_j(textbfs_i)\right)} \right\}$$

where $\lambda_j\,(\mathbf{s}_i) = \lambda_{ij} > 0, \Lambda\,(\mathbf{s}_i) = \Lambda_i = \sum_{j=1}^{3} \lambda_j\,(\mathbf{s}_i)$ and $\Gamma(\cdot)$ is the gamma function.

The parameters, $\lambda_j\,(\mathbf{s}_i)$, are defined in terms of a linear predictor using a logarithm link,

$$\log\left(\lambda_j\,(\mathbf{s}_i)\right) = \log\left(\lambda_{ij}\right) = \sum_{m=1}^{M} \beta_{jm}s_{im} = \mathbf{s}_i\beta_j j = 1, 2, 3$$

where M is the number of covariates included in the model and β_j is the vector of regression coefficients that explains the effects (in log scale) of the covariates on the j^{th} component.

Considering the above, 2 models are analyzed:

Model 1 (M1): $\log\left(\lambda_{ij}\right) = \alpha_j^{M1} + \beta_j^{M1}g_i^k$(base model)

Model 2 (M2): $\log\left(\lambda_{ij}\right) = \alpha_j^{M2} + \beta_j^{M2}g_i^k + \mathbf{FAM}_i\delta_j^{M2}$ (adjusted model)

Here g_i^k represents the number of copies of the minor allele on the k^{th} single-nucleotide polymorphism (SNP) for the i^{th} individual under an additive genetic model; $\mathbf{FAM}_i$ is the i^{th} row of contrast matrix for the pedigree number considered as a categorical variable and $\theta_j^h = (\alpha_j^h, \beta_j^h, \delta^{h^t}_{\;j})^t$ is the vector of regression coefficients on the j^{th} component. (Note $\delta_j^{M1} = 0$).

Our interest in these models lies in the potential genetic effect of each marker on the proposed response. To assess this, Wald statistics were used to test the null hypothesis of no association between each SNP and the response, $\text{H}_0 : \beta = 0$ $(\text{vs.H}_A : \text{not}\,\text{H}_0)$, $\beta = (\beta_1, \beta_2, \beta_3)$.

Gene-based association

Once we identify significant SNPs through the genetic association analysis as described above, we proceed to perform the analysis at a gene level. To achieve this, we propose a penalized regression. Including all the markers simultaneously, this penalized method aims to select those SNPs with higher association. The analysis is done on those candidate genes that contain at least 1 significant marker that has already been determined.

Variable selection on the SNPs is assessed via a penalized likelihood of the form

$$pl\,(\eta; Y, G, c, \kappa) = l\,(\eta; Y, G) - c\kappa \sum\nolimits_{l=1}^{p} \sqrt{k}\|\eta_{.l}\|_2 - (1-c)\kappa \sum\nolimits_{l=1}^{p} \|\eta_{.l}\|_1$$

where $l\,(\eta; Y, G)$ represents the log-likelihood of a dirichlet distributed sample with response matrix $G = (g_1^t, \ldots, g_n^t)^t$ $G = (g_1^t, \ldots, g_n^t)^t$ (or $G : \mathbf{FAM}$ for $\mathbf{M2}$) is the design matrix, $g_i = (1, g_i^1, \ldots, g_i^p)$; p denotes the number of markers considered for variable selection; k is the number of states; η is the regression coefficients vector; c and λ are parameters for the penalized regression; and $\|\eta_{.l}\|_2 = \left(\sum\nolimits_{j=1}^{k} \eta_{lj}^2\right)^{1/2}$ and are the penalty norms. It is important to note that when $c = 1$ we have a ridge regression penalty, whereas when $c = 0$ we have a lasso penalty. We implement the variable selection for the penalized dirichlet regression using R code provided on the Statistical Genetics and Genomics Laboratory at the University of Pennsylvania webpage [5].

Results

Data quality results

The Genetic Analysis Workshop 18 (GAW18) data consists of 855 individuals with genotype and phenotype information. As a result of missing data, transition probabilities are estimated for only 835 individuals. Of these, 43 are removed because of low call rate. The overall call rate for the remaining 792 individuals is 99.82%.

The Genome Wide Association Study (GWAS) data includes 65,519 SNPs for chromosome 3, of which 59 are excluded because it is not possible to reliably obtain position information for these markers. The remaining 65,460 SNPs are considered for data quality control. Because of a low genotyping rate, 114 markers are removed; none are excluded by the Hardy-Weinberg equilibrium test; and 13,011 markers are removed because of low minor allele frequency. The remaining 52,357 markers are considered for analysis.

Analysis results

The parameter estimates for the transition models are obtained using R [6]. To test examination time effect; likelihood ratio tests are performed in which the null model considers only the available clinical variables. Table 1 presents the final transition models.

Table 1 Selected transition models

Transition	Model
$1 \rightarrow j$	$\log\left(\gamma_{1j}/\gamma_{11}\right) = \gamma_{1j0} + \left(\gamma_{1j1}x_{sex} + \gamma_{1j2}x_{smoke} + \gamma_{1j3}x_{age}\right) * time\; j = 2, 3$
$2 \rightarrow j$	$\log\left(\gamma_{2j}/\gamma_{22}\right) = \gamma_{2j0} + \gamma_{2j1}x_{sex} + \gamma_{2j2}x_{smoke} + \gamma_{2j3}x_{age} + time\; j = 1, 3$
$3 \rightarrow j$	$\log\left(\gamma_{3\mathrm{j}}/\gamma_{33}\right) = \gamma_{3l0} + \gamma_{3j1}x_{sex} + \gamma_{3j2}x_{smoke} + \gamma_{3j3}x_{age}\, j = 1, 2$

After the response is estimated, models M1 and M2 are fit using R [7] for each available SNP. Figure 1 displays the Manhattan plots for the p values that result from testing the null hypotheses of no association between the markers and the response. The graphs show that only 1 marker under M2 is significant at the standard significance level for GWAS (5×10^{-8}). Interestingly, the same marker is the most significant marker under M1, although it is not significant at the standard threshold. This suggests that the adjustment for family incorporated in M2 accounts for the family structure in the data. Also, the proposed methodology demonstrates consistency in that the same marker proves to be the most significant under both models. Table 2 summarizes these findings.

Once significant markers were identified, a gene-level association analysis is performed using the penalized regression described above for different levels of c (0, 0.3, 0.5, 0.7, and 1). The analysis is conducted utilizing both the GWAS and the dosage imputed genotypes (GENO) information as the explanatory variables. Figure 2 shows the penalized regression results for the gene containing the significant SNP (rs12492830) for $c = 0.5$ only. This level of c is a blended penalty function, equally weighting the ridge and lasso penalties. Table 3 shows the results for different levels of c under M2 for gene *PCCB*.

Discussion

The present work implements a multistate transition model that conveniently accommodates the longitudinal data structure. Whether the information contained by the available clinical variables is sufficient for predicting the hypertensive state is debatable, however.

Although the adjusted model (M2) is an improvement over the base model (M1), neither of the described models accounts for correlation between individuals nor heteroscedasticity. One way to possibly overcome this is to incorporate a latent variable into the model. Such an extension follows.

Model 3 (M3):$\log\left(\lambda_{ij}\right) = \alpha_j^{M3} + \beta_j^{M3} g_i^k + u_i$ where u_i is the i^{th} element of a vector u that follows a $\text{MVN}(0, \mathbf{K})$ distribution; here $\mathbf{K}$ is twice the estimated kinship matrix. In this case, however, the estimation of the parameters of interest, β_j, is not straightforward. Further research of this methodology is warranted.

With respect to the penalized regression, to avoid an arbitrary selection of c,a cross-validation method could be implemented.

Conclusions

We propose a methodology that conveniently uses the longitudinal data structure to define a probabilistic outcome, which, we believe, explains hypertension in a more suitable way. Dirichlet regression provides an interesting

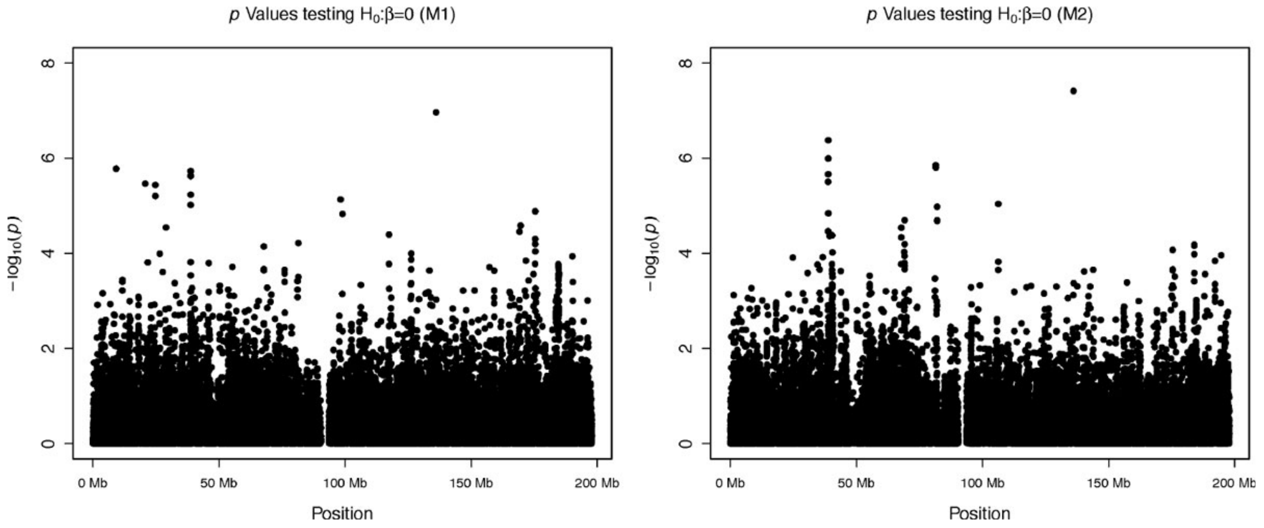

Figure 1 Manhattan plots for genetic association analysis.

Table 2 Association analysis results

SNP	Gene	MA	MAF (%)	*p* Value (M1)	*p* Value (M2)
rs12492830	*PCCB*	C	7.22	1.1×10^{-7}	3.9×10^{-8}

MA, minor allele *MAF*, minor allele frequency.

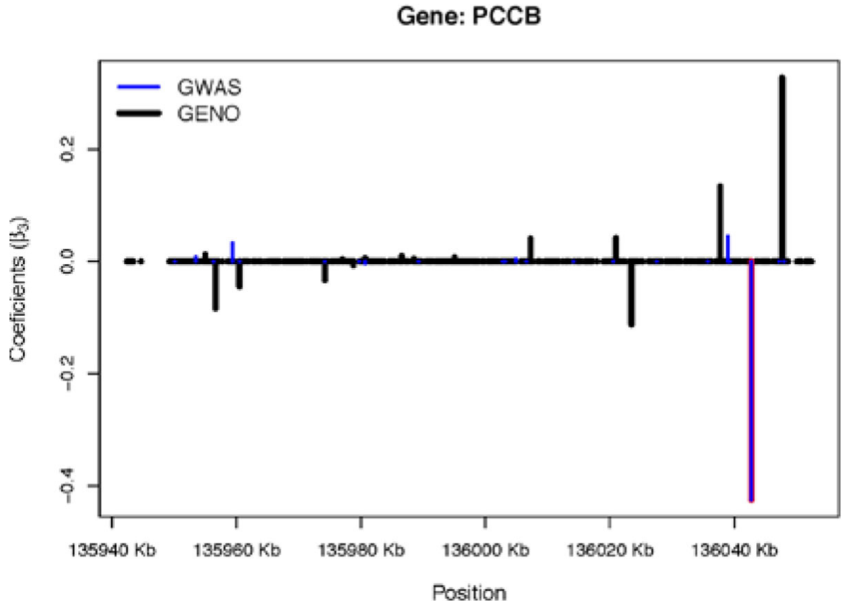

Figure 2 Penalized regression results for M2 only.

Table 3 Comparison of penalized regression under different levels of c

			No. of parameters selected (iterations for convergence)				
Data	# SNPs	c =	0	0.3	0.5	0.7	1
GWAS	22		10 (260)	10 (148)	8 (135)	7 (163)	7 (174)
GENO	607		42 (427)	35 (199)	24 (176)	25 (136)	19 (115)

approach that, along with other more common responses, can be successfully used in genetic association analysis. Our model finds a statistically significant marker at the standard significant level for GWAS, which is noteworthy, considering that it is often difficult to find association. Moreover, when the penalized method is used on the GENO data we are able to find significant markers in addition to those have already found using GWAS data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

OEG and WX designed the overall study; OEG conceived the study, conducted statistical analyses, and wrote the manuscript; XS, XQ, and YB helped develop the study; GL revised the clinical aspects of the study. All authors read and approved the final manuscript.

Acknowledgements

In a personal communication, 2012, Daniel Merico (TCAG/Sick Kids) provided Gene annotation files. OEG thanks Ian Jones for his valuable support in the manuscript review. Thanks to the reviewers for their comments and suggestions. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.
This article has been published as part of *BMC Proceedings* Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

Authors' details

[1]Department of Biostatistics, Princess Margaret Cancer Centre, 610 University Ave., Toronto, ON, M5G 2M9, Canada. [2]Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada. [3]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 60 Murray Street, Toronto, ON, M5T 3L9, Canada. [4]Ontario Cancer Institute/Princess Margaret Cancer Centre, 610 University Ave., Toronto, ON, M5G 2M9, Canada. [5]Dalla Lana School of Public Health, University of Toronto, 155 College St., Toronto, ON M5T 3M7, Canada.

Published: 17 June 2014

References

1. Winegarden CR: **From "Prehypertension" to Hypertension? Additional Evidence.** *Ann Epidemiol* 2005, **15**:720-725.
2. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, *et al*: **PLINK: a toolset for whole-genome association and population-based linkage analysis.** *Am J Hum Genet* 2007, **81**:559-575.
3. Kalbfleisch JD, Lawless JF: **The analysis of panel data under a Markov assumption.** *J Am Statist Assoc* 1985, **80**:863-871.
4. Campbell G, Mosimann JE: **Multivariate methods for proportional shape.** *ASA Proceedings of the Section on Statistical Graphics* Washington, DC; 1987, 10-17.
5. Chen J, Li H: **Variable selection for Dirichlet-multinomial regression for identifying covariates that are associated with microbiomes.** *Ann Appl Stat* 2013, **7**:418-442.
6. Venables WN, Ripley BD: **Modern Applied Statistics with S.** New York, Springer;, 4 2002.
7. Maier MJ: **DirichletReg: Dirichlet Regression in R.** [http://dirichletreg.r-forge.r-project.org], Ver. 0.4-0.

doi:10.1186/1753-6561-8-S1-S70
Cite this article as: Espin-Garcia *et al.*: Genetic association analysis for common variants in the Genetic Analysis Workshop 18 data: a Dirichlet regression approach. *BMC Proceedings* 2014 **8**(Suppl 1):S70.