SciForschen
Open HUB for Scientific Research

ISSN 2471-4968

International Journal of Molecular Genetics and Gene Therapy

Research Article | Volume: 1.1 | Open Access

Dynamic Association Mapping based on a Kalman Filter Model using GAW 18 Data Set

Burak Karacaören*

Department of Animal Science, Faculty of Agriculture, Akdeniz University, Antalya, Turkey

***Corresponding author:** Burak Karacaören, Department of Animal Science, Faculty of Agriculture, Akdeniz University, Antalya, Turkey, **E-mail:** burakkaracaoren@akdeniz.edu.tr

Received date: 27 May 2015; **Accepted date:** 6 July 2015; **Published date:** 10 July 2015.

Citation: Karacaören B (2015) Dynamic Association Mapping based on a Kalman Filter Model using GAW 18 Data Set. Int J Mol Genet Gene Ther 1 (1): http://dx.doi.org/10.16966/2471-4968.101

Abstract

Background

Linear mixed model with pedigree information commonly employed to detect and correct for genetic relationship in cross sectional genomics research. Main aim of this study was to dynamic association mapping by using a random walk-Kalman filter approach for analyzing GAW18 dataset. We also extended the model to incorporate stationary process by auto regressive structure for dynamic gene and environmental effects.

Methods

We used random walk model and it is given below

$y_t = \alpha_t + \varepsilon_t, \varepsilon_t \infty N(0, \sigma_e^2); \alpha_{t+1} = \alpha_t + \eta_t, \eta_t \alpha N(0, A\sigma_n^2)$ In (1) the first equation is called the observation equation and the second equation is called the state equation. We assumed that observations, y_t, depends on unobservable quantity, α_t, and our aim was to do statistical inference on α_t (states)

Results

Error, genetic and permanent environmental variance components were predicted as 17.3 (0.0006), 10.9 (0.0007) and 8.2 (0.0006) using genomic relation matrix for and 18.0 (0.0007), 9.2 (0.0008) and 8.4(0.0006) using pedigree relation matrix for diastolic blood pressure. Rs 11711953 from time point 1, 2 and 3 is found to be associated with MAP4 gene.

Conclusions

Probably due to small number of time points the model did not detect all true genomic signals. Genomic relationship matrix gave better inflation factors. Random walk is a non stationary process and in this paper we extended the model for stationary case by tuning Δ parameter. In genomic studies failing to taken into account of longitudinal gene and environmental effects over time may lead to either undetection of true signals and/or may also lead to false positives due to stochastic errors.

Keywords: Association mapping; Random walk; Kalman filter; Gibbs sampling

Background

Mixed model approach with pedigree information commonly employed to detect and correct for genetic relationship in cross sectional genomics research [1]. Since gene expression may change over time repeated measures would be more useful for detecting associated genomic signals [2]. Dynamic association studies commonly use functional mapping approach. However interpreting results from regression coefficients of (non) parametric functions may be difficult biologically. Random regression coefficient model was suggested for dynamic association mapping [3]. However random regression models have limitations for obtaining accurate estimates at the beginning and end of the trajectories [4].

In addition both of the mentioned models (and those others in the literature of dynamic association mapping) needs whole set of observations in order to have predictions: this may also create problem as such to wait for months (if not years) to obtain predictions. In this study we assumed that genomic signal over time could be traced by a random walk- Kalman filter model in state space form to obtain longitudinal residuals. Because of the Kalman filter we do not have to wait for collecting the whole data set to do model evaluations hence estimates becomes available as soon as measurements are taken. And because of longitudinal residuals to employ in association mapping: biological reasoning could also be deduced easily given the signal is genuine.

Recently, we extended the GRAMMAR model of [1] in Bayesian context [5]. In this paper we used [6] model for dynamic association mapping by including dynamic components using random walk-Kalman filter approach to analyze GAW18 dataset. We also extended the model to incorporate stationary process by auto regressive structure for dynamic gene and environmental effects.

Methods

GAW 18 provided simulated phenotypes for 200 replicates from 849 individuals by 3 time points. We used Diastolic Blood Pressures (DBP) phenotype for association mapping. We analyzed 65519 SNPs from chromosome 3 using related 849 individuals for 3 time points of the first replicate.

Quality control

We used 849 pedigreed individuals from chromosome 3 with 65519 SNPs for association mapping. We excluded 7229 SNPs due to minor allele frequency <1%, 208 SNPs due to Hardy Weinberg test ($p<0.001$), and 2 SNPs due to missingness test ($p>0.1$) leaving 58080 SNPs in the

analyses [7]. We excluded 44 individuals with too low genotyping leaving 805 individuals in the dataset. Kolmogrow-Smirnow test used to assess normality of the response variables. Time, Sex, smoking status, age and pedigree number was included as a fixed effect in subsequent analyses based on preliminary analyses using correlations between predictions and observations.

Random walk model

We used random walk model and it is given below

$$y_t = \alpha_t + \varepsilon_t, \varepsilon_t \infty N(0, \sigma_e^2)$$
$$\alpha_{t+1} = \alpha_t + \eta_t, \eta_t \infty N(0, \sigma_n^2) \quad (1)$$

In (1) the first equation is called the observation equation and the second equation is called the state equation. We assumed that observations, y_t, depends on unobservable quantity, α_t, and our aim was to do statistical inference on α_t (states). We assumed constant variances for ε_t and η_t as σ_e^2 and σ_n^2 respectively with independent, identically and normally distributed random variables with zero means. We assumed that both gene effects and permanent environmental effects.

For genetic analyses of traits following mixed model is used;

$$\mathbf{y} = \mathbf{X\hat{a}} + \mathbf{Z_a a} + \mathbf{Z_p p} + \mathbf{e} \quad (2)$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{\hat{a}}$ is the vector of fixed effects, $\mathbf{a}$ is the vector of random effects, $\mathbf{p}$ is the vector of random permanent environmental effects, $\mathbf{X}$, $\mathbf{Z_a}$, $\mathbf{Z_p}$ are design matrices and $\mathbf{e}$ is the vector of random residual effects. σ_a^2, σ_p^2, and σ_e^2; are genetic, permanent environment and error variances. $\mathbf{A}$ is the additive genetic relationship matrix for the individuals; $\mathbf{I}$ is an identity matrix. $\mathbf{A}$ was obtained by the coefficient of coancestry matrix using both the genotype and pedigree of individuals.

In the following, we show general assumptions used in KF-RW method, based on Bayesian principles. Proportional joint posterior distribution without constant terms given in (3) using (2) based on following recursive relationship [8];

$$p(\mathbf{\grave{e}_t} | \mathbf{\grave{e}_{(-t)}}, Y_n) \propto p(\mathbf{\grave{e}_t} | \mathbf{\grave{e}_{t-1}}, Y_{t-1}) p(\mathbf{\grave{e}_{t+1}} | \mathbf{\grave{e}_t}, Y_{t-1}) p(Y_t | \mathbf{\grave{e}_t}, Y_{t-1}),$$

$$f(y/b,a,p,\sigma_a^2,\sigma_p^2,\sigma_e^2) \propto (\sigma_e^2)^{-\frac{1}{2}N} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{Xb}-\mathbf{Z_a a}-\mathbf{Z_p p})'(\mathbf{y}-\mathbf{Xb}-\mathbf{Z_a a}-\mathbf{Z_p p})/\sigma_e^2\right)$$

$$\mathrm{x}\,\Delta(\sigma_a^2)^{-\frac{1}{2}N_a} \exp\left(-\frac{1}{2}\mathbf{a_1'A^{-1}a_1}/\sigma_a^2\right)\left[\prod_{t=2}^{T}(\acute{o}_a^2)^{-\frac{1}{2}N_a} \exp\left(-\frac{1}{2}(\mathbf{a_t}-\mathbf{a_{t-1}})'\mathbf{A^{-1}}(\mathbf{a_t}-\mathbf{a_{t-1}})/\sigma_a^2\right)\right]$$

$$\mathrm{x}\,\Delta(\sigma_p^2)^{-\frac{1}{2}N_p} \exp\left(-\frac{1}{2}\mathbf{p_1'p_1}/\sigma_e^2\right)\left[\prod_{t=2}^{T}(\sigma_p^2)^{-\frac{1}{2}N_p} \exp\left(-\frac{1}{2}(\mathbf{p_t}-\mathbf{p_{t-1}})'\mathbf{I^{-1}}(\mathbf{p_t}-\mathbf{p_{t-1}})/\sigma_p^2\right)\right]$$

$$\mathrm{x}(\sigma_e^2)^{-\left(\frac{v_e}{2}+1\right)} \exp\left[-\frac{v_e s_e}{2\sigma_e^2}\right](\sigma_a^2)^{-\left(\frac{v_a}{2}+1\right)} \exp\left[-\frac{v_a s_a}{2\sigma_a^2}\right](\sigma_p^2)^{-\left[\frac{v_p}{2}+1\right]} \exp\left[-\frac{v_p s_p}{2\sigma_p^2}\right] \quad (3)$$

Last line of (3) are product of density of scaled inverted chi-square distributions assumed prior for variance parameters. Δ is assumed to be 1 for random walk model. After algebraic manipulations conditional distributions could be written as following,

$$b_t | \sigma_a^2, \sigma_p^2, \sigma_e^2, a_t, p_t, y_t \sim N\left(\left(X'X\right)^{-1} X'(y - Z_a a - Z_p p_t), \left(X'X\right)^{-1} \sigma_e^2\right)$$

$$a_t | \sigma_a^2, \sigma_p^2, \sigma_e^2, b_t, p_t, y_t \sim$$

$$N\left(\left(\frac{1}{\sigma_e^2}\mathbf{Z'_{a_t}Z_{a_t}} + \frac{2}{\sigma_a^2}\mathbf{A^{-1}}\right)^{-1}\left(\frac{1}{\sigma_e^2}\mathbf{Z'}(\mathbf{y_t} - \mathbf{X_a b}_t - \mathbf{Z_{p_t} p_t}) + \frac{1}{\sigma_a^2}\mathbf{A^{-1}}a_{t+1}\right), \left(\frac{1}{\sigma_e^2}\mathbf{Z'_{a_t}Z_{a_t}} + \frac{2}{\sigma_a^2}\mathbf{A^{-1}}\right)^{-1}\right)$$

$$p_t | \sigma_a^2, \sigma_p^2, \sigma_e^2, a_t, b_t, y_t \sim$$

$$N\left(\left(\frac{1}{\sigma_e^2}\mathbf{Z'_{p_t}Z_{p_t}} + \frac{2}{\sigma_p^2}\mathbf{I^{-1}}\right)^{-1}\left(\frac{1}{\sigma_e^2}\mathbf{Z'}(\mathbf{y_t} - \mathbf{X_a b}_t - \mathbf{Z_{p_t} a_t}) + \frac{1}{\sigma_p^2}\mathbf{I^{-1}}a_{t+1}\right), \left(\frac{1}{\sigma_e^2}\mathbf{Z'_{p_t}Z_{p_t}} + \frac{2}{\sigma_p^2}\mathbf{I^{-1}}\right)^{-1}\right)$$

$$\sigma_a^2 | \sigma_p^2, \sigma_e^2, p_t, a_t, b_t, y_t \sim \frac{(Q_a + v_a s_a)}{\chi^2_{DF}}$$

$$\sigma_p^2 | \sigma_a^2, \sigma_e^2, p_t, a_t, b_t, y_t \sim \frac{(Q_p + v_p s_p)}{\chi^2_{DF}}$$

$$\sigma_e^2 | \sigma_a^2, \sigma_p^2, p_t, a_t, b_t, y_t \sim \frac{(Q_e + v_e s_e)}{\chi^2_{DF}}$$

where in the last two line Q_a, Q_p and Q_e stands for quadratic form of the respective error terms and *DF* degrees of freedoms. We ran the model with 10,000 iterations using a 5000-iteration burn-in period for DBP. To reduce auto-correlation, we sampled every tenth iteration. We tried different parameters of inverse Wishart prior distributions to obtain residuals.

We used a mixed model to perform genome-wide association analyses [9,7] using R software [10]:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \quad (4)$$

where $\mathbf{y}$ contains the residuals or random effects from (3), *b* designates the fixed effects (SNP), $\mathbf{X}$ and is incidence matrices, and $\mathbf{e}$ is a vector containing residuals and assumed normally distributed with $I\sigma_e^2$. $\mathbf{I}$ is an identity matrix, σ_e^2 is the residual variance. We used a false discovery threshold of 5 % to detect a genomic signal in association mapping. We also used cross sectional GRAMMAR [1] approach by each time points for comparison purposes. We estimated heritability of DBP as 0.299 and 0.259 using genomic coancestry matrix [9] and pedigree information respectively.

Results

Analyses were performed without knowledge of the underlying simulation model. However, we used the GAW18 answers in discussing the results.

We confirmed the normality using Kolmogrow-Smirnov test. However since we employed Bayesian residuals: all response variables transformed to be normally distributed (P > 0.01). Time, Sex, smoking status, age and pedigree number was included as a fixed effect in subsequent analyses based on preliminary analyses using correlations between predictions and observations. We found that correlations between predictions and observations were highest up to 0.15. Error, genetic and permanent environmental variance components were predicted as 17.3 (0.0006), 10.9 (0.0007) and 8.2 (0.0006) using genomic relation matrix and 18.0 (0.0007), 9.2 (0.0008) and 8.4(0.0006) using pedigree relation matrix for DBP. DBP was simulated with 0.317 heritability whereas genomic kinship estimates were found to be closer to its true value (Tables 1-3).

Discussion

Assumptions regarding evolution of gene and permanent environmental effects over time might be important. Certain degree of autoregressive structure might be more realistic compared with a random walk model. We simply tuned the model based on restrictions of parameters space in (3) using Δ . We considered two extreme cases for deviation from random walk using Δ=0.1 and Δ=0.9. Random walk assumes that gene effects could change slowly in both up and down directions over time, Δ=1.0,. Autoregressive structure for both gene and environmental effects could be introduced by tuning Δ to obtain stationary distributions. Here the time series will be distributed around the mean trajectory.

Error, genetic and permanent environmental variance components were predicted as 10.26 (0.0006), 1.98 (0.0001) and 1.98 (0.0003) using Δ=0.1 and 14.2 (0.0002), 6.4 (0.0009) and 5.5 (0.0006) using Δ=0.9 for DBP. Heritability were predicted as 0.299, 0.139 and 0.246 using Δ=1.0, Δ=0.1 and Δ=0.9 .The random walk gave better results compared with autoregressive structures (DBP was simulated with 0.317). We hypothesis that: increasing the time points should decrease the genomic inflation factors [7] due to accumulation of information regarding both relatedness and substructure over time. We employed both genomic and pedigree based relationship matrix in the mixed model (3). Genomic relationship matrix found to give lower genomic inflation factor as 1.40, 1.33, and 1.63 compared with pedigree based relationship matrix 1.57, 1.83, and 1.59 over three time points for DBP. Due to small number of time points (t=3) still we obtained high level of genomic inflation factors ($\lambda > 1$). Table 1 and Table 2 shows that both genomic relationship and pedigree relationship matrix detected mostly different set of SNPs for different time points.

However both small sampling size and small number of time points may lead to false positives and false negatives. This may be true especially for very first time point: genomic relationship matrix detected 154 SNPs at 5 % False Discovery Rate (FDR) (134 and 56 SNPs detected for time points 2 and 3 respectively at 5 % FDR) and pedigree relationship matrix detected 216 SNPs at 5 % FDR (300 and 96 SNPs detected for time points 2 and 3 respectively at 5 % FDR). Due to smaller genomic inflation factors we investigated results of genomic relationship matrix for causal SNPs. rs11711953 from time point 1, 2 and 3 is found to be associated with MAP4 gene.

We used GRAMMAR approach to analyze each time points (and average of them) cross sectionally (Table 3). However we did not detect any genomic signals after multiple hypothesis corrections. Although there was signals from time point 2 by rs1948722 at the vicinity of ARHGEF3 ($p<0.00012$), the SNP became non significant after multiple hypothesis correction by FDR. Magnitude of GRAMMAR p values (Table 3) found to be larger compared with the p values of random walk models (Tables 1,2). This clearly shows that longitudinal gene and environmental effects over time needs to be taken into account by proper methodology. Otherwise since the genomic signals will be contaminated by stochastic errors this may lead to either undetection of the signals or may also lead to false positives. Random walk is a non stationary process and in this paper we extended the model for stationary case by tuning Δ parameter. However theoretical and empirical dynamic association studies are needed if non stationary assumption is useful or not for dynamics of gene and permanent environmental effects.

Conclusions

Genomic relationship matrix gave better inflation factors and estimates of heritability compared with pedigree information. The random walk model may be useful for long time series in practice due to its recursive structure from Kalman filter. When the longitudinal observations available (daily or monthly for example) the model could predict the on-line genomic signals sequentially due to the Kalman Filter. In genomic studies failing to taken into account of longitudinal gene and environmental effects over time may lead to either undetection of true signals and/ or may also lead to false positives due to stochastic errors.

Acknowledgements

The Genetic Analysis Workshops are supported by National Institutes of Health (NIH) grant R01 GM031575 from the National Institute of General Medical Sciences. This work was supported by Research Fund

Top SNPs for Time Point 1		Top SNPs for Time Point 2		Top SNPs for Time Point 3	
rs6763824	2.28E-09	rs11711953	3.72E-07	rs7619263	8.41E-08
rs11716779	5.18E-09	rs11706549	3.82E-07	rs6794386	8.58E-08
rs1665982	6.11E-09	rs17468698	5.45E-07	rs6444444	3.65E-07
rs11711953	6.11E-09	rs7631950	1.04E-06	rs11711953	3.74E-07
rs11706549	7.06E-09	rs2133601	1.15E-06	rs11706549	3.97E-07
rs34448818	7.53E-09	rs13063042	1.21E-06	rs7621594	1.13E-06
rs13072132	7.64E-09	rs11130143	1.22E-06	rs902119	1.82E-06
rs17785248	8.38E-09	rs11130146	1.22E-06	rs35836	1.88E-06
rs319680	2.19E-08	rs11716582	1.22E-06	rs9861506	2.42E-06
rs7621236	1.96E-07	rs1403579	1.22E-06	rs902118	2.79E-06

Table 1: Top 10 SNPs and correspondent raw p values obtained using random walk model from genomic relationship matrix for first replicate of DBP

Top SNPs for Time Point 1		Top SNPs for Time Point 2		Top SNPs for Time Point 3	
rs17468698	5.90E-09	rs2001665	2.36E-08	rs1917098	5.24E-07
rs6763824	3.41E-07	rs11711953	2.56E-08	rs1320260	1.33E-06
rs1515592	6.44E-07	rs11706549	2.81E-08	rs7616403	1.79E-06
rs6774170	1.39E-06	rs861375	3.93E-08	rs3924267	2.40E-06
rs7621236	1.49E-06	rs836852	4.08E-08	rs1818553	2.78E-06
rs4858842	1.53E-06	rs2589601	9.44E-08	rs7651173	3.13E-06
rs12635772	1.63E-06	rs13067751	5.30E-07	rs4974272	3.21E-06
rs7429162	1.77E-06	rs2371997	7.81E-07	rs7632426	3.39E-06
rs11716779	1.96E-06	rs9858561	1.43E-06	rs9814548	3.59E-06
rs7624474	1.96E-06	rs13323469	2.19E-06	rs11710809	3.59E-06

Table 2: Top 10 SNPs and correspondent raw p values obtained using random walk model from pedigree relationship matrix for first replicate of DBP

Top SNPs for Time Point 1		Top SNPs for Time Point 2		Top SNPs for Time Point 3	
rs7634891	7.21E-05	rs11919819	3.78E-05	rs17295091	1.16E-05
rs3821750	0.000439	rs1515441	4.76E-05	rs7634891	2.59E-05
rs4681625	0.000526	rs6778909	7.90E-05	rs765006	6.08E-05
rs6771925	0.000831	rs7653527	8.59E-05	rs4858283	8.32E-05
rs4685707	0.000832	rs1949587	0.000107	rs12632218	0.000222
rs13323544	0.000969	rs1948722	0.000123	rs2196550	0.000301
rs4234379	0.000981	rs1515442	0.000195	rs9877517	0.000309
rs1012583	0.000991	rs6775757	0.000205	rs1386948	0.000311
rs13322784	0.001168	rs9880343	0.000214	rs1389660	0.000349
rs9813330	0.001757	rs17066303	0.000309	rs6764110	0.000352

Table 3: Top 10 SNPs and correspondent raw p values obtained using GRAMMAR for first replicate of DBP

of the Akdeniz University Project Number 106. The author wishes to acknowledge useful discussions with Dr Luc Janss about auto regressive structures.

Competing Interests

There are no competing interests.

References

1. Aulchenko SY, de Koning DJ, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177: 577-585.
2. Wu R, Lin M (2006) Functional mapping - how to map and study the genetic architecture of dynamic complex traits. Nat Rev Genet 7: 229-237.
3. Macgregor S, Knott SA, White I, Visscher PM (2005) Quantitative trait locus analysis of longitudinal quantitative trait data in complex pedigrees. Genetics 171: 1365-1376.
4. Faro EL, Cardoso VL, Albuquerque LG (2008) Variance component estimates applying random regression models for test-day milk yield in Caracu heifers (Bos taurus Artiodactyla, Bovidae). Genet Mol Biol 31.
5. Karacaören B (2014) Admixture mapping of growth related traits in F_2 mice dataset using ancestry informative markers. J Bioinform Comput Biol 12.
6. Karacaören B (2015) A Bayesian random walk approach for mapping dynamic quantitative trait. J applied nonlinear dynamics.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. Am J Hum Genet 81: 559-575.
8. West M, Harrison J (1997) Bayesian forecasting and dynamic models. Springer.
9. Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen 4: 250-255.
10. R Development Core Team (2014) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.