Genomics 105 (2015) 69–75

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Detection of genomic loci associated with environmental variables using generalized linear mixed models

Stéphane Lobréaux *, Christelle Melodelima *

Laboratoire d'Ecologie Alpine, UMR CNRS 5553, Université Joseph Fourier, BP53 38041, Grenoble, France

ARTICLE INFO

Article history:
Received 7 February 2014
Accepted 5 December 2014
Available online 11 December 2014

Keywords:
GLMM
Adaptation
SNP
Genome scan

ABSTRACT

We tested the use of Generalized Linear Mixed Models to detect associations between genetic loci and environmental variables, taking into account the population structure of sampled individuals. We used a simulation approach to generate datasets under demographically and selectively explicit models. These datasets were used to analyze and optimize GLMM capacity to detect the association between markers and selective coefficients as environmental data in terms of false and true positive rates. Different sampling strategies were tested, maximizing the number of populations sampled, sites sampled per population, or individuals sampled per site, and the effect of different selective intensities on the efficiency of the method was determined. Finally, we apply these models to an *Arabidopsis thaliana* SNP dataset from different accessions, looking for loci associated with spring minimal temperature. We identified 25 regions that exhibit unusual correlations with the climatic variable and contain genes with functions related to temperature stress.

1. Introduction

The recent evolution in sequencing technologies now allows affordable high-resolution genotyping of many individuals [1]. The discovery of genome-wide single-nucleotide polymorphisms (SNPs) may be based on full genome resequencing [2–4] or targeted sequencing of genomic regions such as exon or sequences adjacent to a restriction enzyme cleavage site [5,6]. The large genetic variation data sets produced with extensive genome coverage offer the possibility to analyze and compare polymorphisms between individuals. Genome-wide association studies (GWAS) have been used to identify the association between genetic loci and phenotypes such as traits or diseases from SNP data using statistical modeling [7,8]. Hundreds of genes implicated, for example, in human diseases [9,10] or important plant traits have been discovered [11]. Next-generation sequencing technologies (NGS) open up the possibility to detect many rare variants to improve such association mapping strategy [12]. Another approach is to scan for genomic loci associated with environmental variables, such as climatic data, for example, that could be involved in the adaptation to specific conditions of the species of interest [13]. There is a growing interest in identifying factors that influence adaptation in species [14]. Genetic variability evolves as a result of environmental factors imposing selective pressure on part of the genome, generating changes in allele frequency in such regions. Looking for loci associated with environmental variables requires statistical modeling as GWAS to find an allele distribution fitting for the chosen variable [13]. In both association study approaches, it is essential to take into account the population structure of the sampled individuals [15,16]. Indeed, hierarchical population structure can cause bias in loci detection, generating an excess of false positives [15]. The main problem is that without corrections for the effect of population structure, the underlying null distribution may be insufficient to account for demographic history. In GWAS, Generalized Linear Mixed Models have been shown to efficiently take into account the population structure [16]. Studies in maize, potato, and *Arabidopsis* revealed that such mixed models led to fewer false positives and were a powerful method [17–19]. GLMMs are an extension of a generalized linear models providing a more flexible approach for analyzing non-normal data when random effects are present [20,21]. The most common types of random effects are variation among individuals, genotypes, species, and geographical regions or time periods. For instance, individuals within clusters (for example, at distances of only some meters) are genetically more similar than distant individuals between clusters. GLMMs can more efficiently account for population structure in comparison with previously used methods [22,23]. Using such models to detect the correlation between allele distribution and environmental data can be performed on samples collected in different sites, without acquiring population data such as allele frequency at the sampling sites. GLMMs are fast, while some Bayesian approaches, for example, require extensive computing time and are therefore more difficult to apply to very large genetic variability data sets [13,24].

In this study, we tested the use of GLMMs to detect associations between genetic loci and environmental variables. Then we applied these models to analyze an *A. thaliana* SNP data set from different accessions

* Corresponding authors. Fax: +33 476514279 15.
E-mail addresses: stephane.lobreaux@ujf-grenoble.fr (S. Lobréaux), christelle.melo-de-lima@ujf-grenoble.fr (C. Melodelima).

http://dx.doi.org/10.1016/j.ygeno.2014.12.001
0888-7543/© 2014 Elsevier Inc. All rights reserved.

[2] and to identify 25 regions that exhibit unusual correlations with climatic variables and contain 18 genes with functions related to temperature stress.

2. Materials and methods

2.1. Simulations

Genotype simulations were performed using the GenomePop2 program [25], allowing the simulation of chromosomes and their evolution under different scenarios. All simulations were performed under a biallelic model, with independent SNPs, with population sizes of 100 individuals, and with a mutation rate of 10^{-4}. To reproduce a situation where different sampling sites are located in different geographical regions, clusters that mimic geographical regions were defined. Ten clusters were run in parallel. In each cluster, 20 populations were simulated, and a migration rate of 1% was applied in an island model. For each cluster, a 20-step gradient of selective coefficient was defined, and each step value was applied to one population in the cluster. The gradient names 0.01, 0.02, 0.03, 0.05, 0.1, and 0.2 corresponded to selective coefficients in a range, respectively, of −0.01 to 0.01, −0.02 to 0.02, −0.03 to 0.03, −0.05 to 0.05, −0.1 to 0.1, and −0.2 to 0.2. For SNPs to which selective pressure was applied, the initial allele frequency in the population was set to 0.5. For neutral SNPs, the initial frequency was set as a random value between 0.1 and 0.9 in each cluster. Each simulation run consisted of 1000 generations, and 25 runs were performed. Simulations were repeated eight times for each cluster with new initial allele frequency for neutral SNPs. A total of 2.10^5 SNPs under selective pressure and 2.10^6 neutral SNPs were simulated. In all subsequent analysis, at least 2000 SNPs under selective pressure and 20,000 neutral SNPs were sampled randomly and independently from the simulated data set. The selective coefficient applied to an individual was assimilated to the environmental variable.

2.2. Detection of potentially adaptive loci using generalized linear mixed models

GLMMs were used to detect SNPs correlated to environmental variables. The general form of the model (in matrix notation) is

$$Y = X\beta + Z\gamma + \varepsilon, \tag{1}$$

where Y is a column vector, the outcome variable; X is a matrix of the predictor variables; β is a vector of fixed-effects regression coefficients; Z is a design matrix for the random effects (the random complement to the fixed X); γ is a vector of random effect (the random complement to the fixed β); and ε is a vector of the residual, the part of Y that is not explained by the model $X\beta + Z\gamma$.

In GLMMs, environmental variables are introduced as fixed effect, while geographical proximity is modeled by random effect. In the model, the matrix $Z\gamma$ models the part of the genetic variation that cannot be explained by the environmental pressures.

A GLMM model with a logit link and binomial error distribution was estimated between each SNP and the selected environmental variable. The likelihood ratio (LR) and the Wald tests were used to evaluate each model's performance [22]. For both tests, the null hypothesis corresponds to no correlation between a particular SNP and an environmental variable. The Wald test for GLMMs tests the null hypothesis of no effect by scaling parameter estimates by their estimated standard errors and comparing the resulting test statistic to zero [26]. The LR test determines the contribution of a single (random or fixed) factor by comparing the fit (measured as the deviance, that is, minus two times the log-LR) for models with and without the factor, namely, nested models. All GLMM models were calculated using the R package lme4.

2.3. Statistical analysis of false and true positive rates

McNemar tests and p value corrections have been applied to test the effect of the different thresholds of the Wald and the LR tests on false and true positive rates. An analysis of variance has been applied to test the sampling effect (e.g., the number of sampled populations, sampled sites per population, and sampled individuals per site). Friedman and Wilcoxon paired tests have been used to test the effect of the different selection intensities on the rate of true positives.

2.4. *Arabidopsis thaliana* SNP dataset

Eighty *A. thaliana* ecotypes sampled across Europe have been submitted for genomic DNA sequencing in the frame of the 1001 Genomes Project (http://1001genomes.org) by Cao et al. [2]. Sequence reads were mapped against the Columbia ecotype reference genome to detect genetic polymorphisms [2]. From this data set, we retained 78 ecotypes, removing two samples for which data quality was lower. SNPs were filtered according to the following criteria: only biallelic sites were retained, positions for which genotyping data were not available for more than two ecotypes which were discarded, and a minimal coverage threshold of 5 was applied. The resulting filtered data correspond to a high density of 1 SNP/125 bp. The 948,330 SNPs of the filtered data set were distributed across the five *A. thaliana* chromosomes as follows: chr1 (247,859 SNPs), chr2 (146,826), chr3 (179,642), chr4 (153,061), and chr5 (220,942).

The climatic data were extracted from the WorldClim database (http://www.worldclim.org) with a spatial resolution of 30 arcsec. We used the R package raster for that purpose, according to the GPS coordinates of each sampling site (http://1001genomes.org/projects/MPICao2010/index.html). The average minimal temperature was calculated over the period from April to June, corresponding to spring in which plants are under vegetation in all sampled sites.

The detection of SNP markers potentially associated with minimal temperature was conducted as described above using GLMM.

3. Results and discussion

3.1. Simulations

3.1.1. Allele frequency and selection intensities

We evaluated the effect of the selective intensities on allele frequencies of our simulation. For this purpose, we generated a plot of allele frequencies under the different intensities of selection (Fig. 1). A clear structuring of allele frequencies was observed in clusters when the gradient of selection intensity was strong (gradient = 0.05–0.2). For a gradient of selection of 0.03 and 0.02, the structuring of allele frequencies in the cluster was present, although it was lower than that under strong selection. Finally, with a weaker gradient of selection (0.01), the structuring was not discernible.

3.1.2. Influence of the Wald and the LR tests on true and false positives

False positives are a major concern when using GLMMs in association studies. A strategy adopted by Joost et al. [22] was to jointly use the Wald and the LR test results to reach a higher stringency. In the frame of our study, we have investigated the influence of the two tests over the false positive and false negative rates obtained when using GLMMs. Results are presented for the gradient 0.05 which selection intensities values were assimilated to the environmental variable, but similar results were obtained for all selection intensities. For each test, four different thresholds 5.10^{-5}, 5.10^{-4}, 5.10^{-3}, and 5.10^{-2} were used.

The false positive rate was strongly influenced by the LR test threshold, but not by the Wald test threshold; differences of 14.7% (CI, [11.8; 17.57]) and 0.6% (CI, [0.3; 0.9]) between thresholds were obtained for the LR and the Wald tests, respectively. In large-scale genomic experiments involving thousands of statistical tests