Analysis of genetic background of quantitative traits related to alcoholism by mixed inheritance and oligogenic models

Joanna Szyda [1,2,*], Przemysław Biecek [3,4], Florian Frommlet [5], Jayanta K. Ghosh [4,6], Małgorzata Bogdan [3,4]

[1] Department of Animal Genetics, Wrocław Agricultural University, Kożuchowska 7, Wrocław, Poland

[2] VIT, Verden, Germany

[3] Institute of Mathematics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, Wrocław, Poland

[4] Department of Statistics, Purdue University, 150 N. University Street, West Lafayette,USA

[5] Department of Statistics, University of Vienna, Brunnerstr. 78, Vienna, Austria

[6] Stat-Math Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta, India

[*] Corresponding author

Email addresses:

JS: szyda@karnet.ar.wroc.pl
PB: cogito@pwr.wroc.pl
FF: florian.frommlet@univie.ac.at
JKG: ghosh@stat.purdue.edu
MB: mbogdan@im.pwr.wroc.pl

Abstract

We present results of the analysis of the genetic background of quantitative traits related to alcoholism. We used a mixed linear model with random polygenic effects to estimate genetic correlations between some of the quantitative traits presumably related to alcoholism and to describe the genetic background for two of those traits: the natural logarithm of the maximal number of drinks in a 24 hour period [ln(maxdrinks)]and one of the electrophysiological traits, ttth3. We used SNPs from Illumina to perform genome scans based on the modified PDT test and identify single genes influencing these traits. We also verified chosen SNPs with the likelihood ratio statistic in the mixed linear model. Apart from looking for genes with significant additive effects we used the modified version of Bayesian Information Criterion proposed in [1] to look for significant epistatic effects. We localized several interesting genome regions which may host genes related to alcoholism but the main conclusion of our research is that the heritability of the analyzed traits can be attributed to a large number of genes with small genetic effects rather than to a few strong quantitative trait loci.

Background

In our research we used the data provided by the Collaborative Study on the Genetics of Alcoholism and analyzed the genetic background of some quantitative traits related to alcoholism. Since the genetic background of alcoholism has not yet been extensively studied and a firm definition of a quantitative assessment of alcoholism is not available one of main goals of our research was to estimate the heritability and genetic correlations between some quantitative traits presumably related to alcoholism. For this purpose we used mixed linear models, which are commonly used to describe the inheritance of quantitative traits in humans, animals and plants (see eg. [2] and [3]). Mixed models allow information on genetic relations to be directly incorporated into the model by using a random polygenic effect, whose covariance matrix expresses coancestry between individuals. This approach allows the model to account for both the influence of many genes with small effects as well as effects of single quantitative trait loci (QTL). In the second part of our research we used two dimensional genome scans based on fixed effects linear models to localize genes related to alcoholism. The main advantage of using such two dimensional scans is the possibility of detecting epistatic effects. Our research suggests that the heritability of alcoholism can be explained by an influence and interactions of many genes with a relatively small individual effects.

Methods

Estimating genetic and residual correlations between traits: The natural logarithm of the maximal number of drinks in a 24 hour period [ln(maxdrinks)] was chosen as the quantitative measure of alcohol dependence. To compute ln(maxdrinks) we set maxdrinks=0.1 for all individuals for whom a true value of maxdrinks was equal to 0. Additionally we analyzed electrophysiological measurements from the Visual Oddball Experiment. Variables ttth and tttd contain data extracetd from the target case of the experiment and correspond to the 'late' time window at 300 to 700 ms following stimulus presentation. The theta band power was in range between 3 to 7 Hz for ttth and 1 to 2.5 Hz for tttd. Variables ntth contain data from the non-target case of the experiment and correspond to the 'early' time window (100 to 300 ms after stimulus presentation) and the theta band power between 3 and 7 Hz. Each measurement was taken at four locations described by

consecutive numbers 1-4 and related to a far frontal left side channel and frontal, central and parietal midline channels correspondingly. To investigate which of the electrophysiological phenotypes can be used as an alternative measure of alcohol dependence we estimated polygenic and residual correlations between selected traits using the following multivariate mixed model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\alpha + \epsilon \ . \tag{1}$$

Here $\mathbf{y_{n\times p}} = [y_1, \ldots, y_p]$ is a matrix of phenotypic values for p traits, $\beta = [\beta_1, \ldots, \beta_p]$ is a matrix of fixed effects for p traits, with β_i being a column vector of coeficients corresponding to different values of qualitative predictors comprising sex, smoker status and ethnicity, $\alpha_{\mathbf{n\times p}} = [\alpha_1, \ldots, \alpha_p]$ is a matrix of random additive polygenic effects for each of n individuals and p traits, $\epsilon_{\mathbf{n\times p}}$ is a matrix of random errors, and $\mathbf{X}$, and $\mathbf{Z}$ are corresponding design matrices. The covariance structure corresponding to the model (1) is given by

$$Var[\alpha, \epsilon] = \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix}$$

with

$$\mathbf{G} = \begin{bmatrix} G\sigma^2_{\alpha_1} & \ldots & G\sigma_{\alpha_{1,p}} \\ \ldots & \ldots & \ldots \\ G\sigma_{\alpha_{1,p}} & \ldots & G\sigma^2_{\alpha_p} \end{bmatrix} \ ,$$

where G represents polygenic relationships among individuals expressed by a standard numerator relationship matrix based on the information on coancestry between all available individuals (for definition of G see e.g. [3]) and $\sigma_{\alpha_{i,j}}$ is an additive genetic (co)variance for traits i and j attributed to polygenes. The matrix $\mathbf{R}$ is given by equation

$$\mathbf{R} = \begin{bmatrix} \mathbf{I}\sigma^2_{\epsilon_1} & \ldots & \mathbf{I}\sigma_{\epsilon_{1,p}} \\ \ldots & \ldots & \ldots \\ \mathbf{I}\sigma_{\epsilon_{1,p}} & \ldots & \mathbf{I}\sigma^2_{\epsilon_p} \end{bmatrix} \ ,$$

where $\mathbf{I}$ is the identity matrix and $\sigma_{\epsilon_{i,j}}$ is a residual (co)variance for traits i and j.

To estimate parameters of the model (1) we maximized the restricted maximum likelihood function (REML) defined in [4] using the average information algorithm (AI) of [2].

We considered three following sets of traits:
1) y=[ln(maxdrinks) ttth1 ttth2 ttth3 ttth4]
2) y=[ln(maxdrinks) ttdt1 ttdt2 ttdt3 ttdt4]
3) y=[ln(maxdrinks) ntth1 ntth2 ntth3 ntth4].

Estimating heritability and testing impact of polygenes: A univariate version of model (1) was used to estimate the proportion of observed trait variation, which is due to polygenes (i.e. heritability; h^2):

$$h^2 = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\epsilon} \ .$$

Furthermore, we use the univariate model (1) to test the hypothesis that the polygenic effect has no influence on the trait variation, i.e. $\sigma^2_\alpha = 0$ vs $\sigma^2_\alpha > 0$. For this aim we use the likelihood

ratio statistic:

$$\lambda = -2 \ln \frac{L(M_0)}{L(M_1)} \ . \tag{2}$$

Here $L(M_1)$ and $L(M_0)$ are the maximum values of likelihood functions underlying the unrestricted model (1) and a more parsimonious model, where no polygenic effect is assumed.

Genome scan based on the mixed model: To identify single genes influencing two chosen traits, ln(maxdrinks) and tth3, we performed genome scans based on the modified PDT statistic of [5]. Following the notation of [6] we define:

$$PDT = \frac{(\sum_{i=1}^{N} S_i)^2}{\sum_{i=1}^{N} (S_i^2)} \ , \tag{3}$$

where $S_i = [x_i - E(x_i)][y_i - E(y_i)]$, x_i is the observed number of SNP allele "1", $E(x_i)$ is the expected number of SNP allele "1" for offspring based on parental SNP genotypes, y_i is the trait value and $E(y_i)$ is the the trait value predicted by the above mixed model (1). In the first step we have chosen 20 SNPs yielding the highest value of the PDT statistic. These SNPs were later verified with the likelihood ratio statistic λ based on the likelihood of the model (1) supplied with an additional qualitative effect corresponding to possible genotypes of a given SNP.

Genome scan based on the modified version of Bayesian Information Criterion: To further investigate the genetic background of ln(maxdrinks) and ttth3 we performed a genome scan based on the modified version of Bayesian Information Criterion (mBIC) proposed in [1] and [7]. In brief, the method relies on choosing the multiple regression (or ANOVA) model which best describes the trait. The predictor variables corresponding to additive, dominance and epistatic effects are defined by SNPs genotypes as in [8]. To decide which terms should be included in the model a sequence of two dimensional (i.e. simultaneously including two SNPs) genome scans is performed. The modified version of Bayesian Information Criterion is used to decide on the number of terms included. This criterion suggests choosing the model for which

$$mBIC = n \log RSS + (p+q) \log n + 2pL + 2qU \ , \tag{4}$$

obtains a minimum. Here RSS is the residual sum of squares from regression, n is the sample size, p is the number of additive and dominance terms present in the model, q is the number of interaction terms, L is the penalty for including an additive or a dominance term and U is the penalty for including an interaction term. In our experiment we used 4720 SNPs and in such case formulas from [9] suggest using penalties $L = 8.35$ and $U = 16.55$. These additional penalties take into account multiplicity testing problem and guarantee that under standard model assumptions and the considered sample sizes ($n \geq 700$) the probability of the type I error does not exceed 0.03.

We used mBIC to analyze residuals resulting from fitting linear models describing the dependence of ttth3 and ln(maxdrinks) on two different sets of covariates. In the first step we used the same set of covariates as the one used for the mixed model, i.e. sex, ethnicity and a smoker status. In the second analysis we replaced the variable smoker status with an age at interview. We verified our findings with the repeated analysis based on 700 randomly chosen individuals.

Results

Genetic correlations between selected traits: We observed that traits ttth1, ttth2 and ttth3 have relatively strong negative genetic correlations with ln(maxdrinks) ($r_{\alpha_{i,j}}$=-0.2, -0.34, -0.46 respectively) , while traits ttdt1, ttdt2 and ttdt3 are genetically positively correlated with ln(maxdrinks) ($r_{\alpha_{i,j}}$=0.38, 0.32, 0.42 respectively). Traits ntth1, ntth2 and ntth3 are not genetically correlated with ln(maxdrinks) ($r_{\alpha_{i,j}}$=0.03, -0.09, -0.02 respectively). Our analysis shows also that for each of the considered electrophysiological phenotypes the observations taken at far frontal left side channel, frontal midline channel and central midline channel are strongly correlated, both genetically and residually. The observations taken at parietal midline channel exhibit much smaller correlation with measurements taken at other locations and in case of ttth and ttdt show also much smaller correlation (both genetical and residual) with ln(maxdrinks). Based on the estimated correlations we decided to use ttth3 as the second quantitative trait in further analysis since it has the strongest genetic correlation and relatively low residual correlation with ln(maxdrinks).

Heritability and significance of polygenic effects: For ln(maxdrinks) the polygenic and residual variances estimated by the univariate model (1) amount respectively to 0.145 and 0.833, which result in a relatively low heritability of 0.148. Corresponding estimates for ttth3 are: $\hat{\sigma}^2_\alpha = 0.255$, $\hat{\sigma}^2_\epsilon = 0.462$, indicating a moderate heritability of 0.356.

Genome scan based on PDT: In the table below we give results for some SNPs which have relatively small p-values for both PDT and likelihood ratio statistics.

HERE TABLE 1.

The reported p-values both for λ and PDT are much smaller for ttth3 than for ln(maxdrinks). Thus the evidence of single gene trait determination through QTL is much stronger for ttth3. The p-values observed for ln(maxdrinks) only allow vague speculation about the possible impact of selected genome sites on the trait.

Genome scans based on mBIC : When using the covariates as in the mixed model above mBIC finds no significant effects. This agrees with the reported results of single genome scans since none of the signals found by PDT is significant when we adjust for multiple testing. Replacing the smoker status covariate by the age at interview dramatically changes the results of mBIC scans. The corresponding scan for ttth3 finds a significant additive effect at SNP rs1019374 on chromosome 3. The p-values corresponding to this effect in a simple regression model are equal to $2.29 \cdot 10^{-7}$ for the full set of data and to $4.82 \cdot 10^{-6}$ for the randomly chosen subsample of 700 individuals. For ln(maxdrinks) we obtain 13 significant additive signals and 27 significant epistatic effects scattered over all chromosomes with the exception of chromosomes 1, 16, 21 and 22. Of these signals only two, on chromosomes 3 and 10, are strong enough to exceed mBIC threshold for the smaller sample of 700 individuals. These signals, as well as two strongest interactions of the additive-additive type and the corresponding p-values are reported in Table 2. All other signals detected by mBIC in the large sample, together with the corresponding p-values in both samples are reported on the webpage http://neuron.im.pwr.wroc.pl/cogito/GAW.html.

HERE TABLE 2

Discussion

We investigated the genetic background of quantitative traits related to alcoholism using several different methods differing in parameterisation of the genetic components (only polygenes, polygenes and a single QTL, multiple QTL with interactions). The estimated heritabilities based on polygenic mixed models did not differ between uni- and multivariate approach indicating that variances and covariances of traits can be well separated by our multivariate models. On the other hand, we observed that the choice of covariates has a big impact on results of genome scans. In particular, covariate age at interview explained large part of variation of ln(maxdrinks) and helped to increase the power of detection of genetic effects.

Conclusions

Although all of the quantitative traits analyzed are to some extend determined genetically, they considerably differ in the proportion of genetic variation within the observed (phenotypic) variability. The results of our analysis suggest that among various electrophysiological measurements taken, only a few can be considered as good indicators of alcoholism. We also conclude that the genetic component of the variation of the analyzed traits is mainly determined by many interacting genes with relatively small effects, rather than by a few strong QTLs.

Authors' contributions

JS carried out mixed linear model analysis and genome scans based on PDT statistic and participated in project coordination and drafting the paper. PB prepared data for analysis and participated in both genome scans. FF wrote programs for genome scan based on mBIC. JKG participated in genome scans based on mBIC and project coordination. MB participated in genome scans based on mBIC, project coordination and drafting the paper. All authors read and approved the final manuscript.

REFERENCES

1. Bogdan M, Ghosh JK, Doerge RW: **Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci.** *Genetics* 2004, **167**:989-999.

2. Gilmour AR, Thompson R, Cullis BR: **AI, an efficient algorithm for REML estimation in linear mixed models.** *J. Dairy Sci.* 1995, **51**:1440-1450.

3. Henderson CR: **Applications of linear models in animal breeding.** 1984, University of Guelph Press, Guelph.

4. Patterson HD, Thompson R: **Recovery of interblock information when block sizes are unequal.** *Biometrika* 1971, **58** :545-54.

5. Monks SA, Kaplan NL: **Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus.** *Am. J. Hum. Genet.* 2000, **66** :576-592.

6. Lange C, DeMeo DL, Laird NM: **Power and design considerations for a general class of family-based association tests: quantitative traits.** *Am. J. Hum. Genet.* 2002, **71** :1330-1341.

7. Baierl A, Bogdan M, Frommlet F, Futschik A. **On locating multiple interacting quantitative trait loci in intercross design.**, submitted.

8. Kao C-H, Zeng Z-B **Modeling epistasis of quantitative trait loci using Cockerham's model.** *Genetics* 2002, **160**:1243-1261.

9. Biecek P, Bogdan M, Doerge RW, Ghosh JK **Locating multiple interacting quantitative trait loci with the modified version of Bayesian Information Criterion.** In preparation.

Tables

Table 1 Results of genome scans based on the PDT and the likelihood ratio statistics.

ln(maxdrinks)			ttth3		
	nominal p-value			nominal p-value	
SNP	PDT	λ	SNP	PDT	λ
rs624228(chr3)	0.0033	0.086	rs1022092(chr6)	0.0001	0.0004
rs1549114(chr3)	0.014	0.079	rs716493(chr1)	0.0006	0.0004
rs1989749(chr14)	0.015	0.011	rs240153(chr6)	0.0007	0.00003

Table 2 Results of genome scans for ln(maxdrinks) based on mBIC. p_1 and p_2 are p-values in the large and the small sample correspondingly.

additive			interactions		
SNP	p_1	p_2	SNPs	p_1	p_2
rs224136 (chr10)	$6 \cdot 8^{-10}$	$1.2 \cdot 10^{-6}$	rs17114 (chr3)-rs891674 (chr4)	$6 \cdot 10^{-12}$	$5.1 \cdot 10^{-8}$
rs765695 (chr3)	$4.7 \cdot 10^{-7}$	$2.9 \cdot 10^{-7}$	rs2137289 (chr18)-rs2377473 (chr20)	$1 \cdot 10^{-12}$	$4.6 \cdot 8^{-10}$