The application of mixed linear models for the estimation of functional effects on bovine stature based on SNP summary

K. Kotlarz[1], B. Kosinska-Selbi[1], Z. Cai[2], G. Sahana[2] and J. Szyda[1,3*]

[1] Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland; [2] Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark [3] National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland [*]joanna.szyda@upwr.edu.pl

Abstract

Genome-Wide Association Studies (GWAS) help identify polymorphic sites or genes linked to phenotypic variance, but a few genes / SNPs are unlikely to explain the phenotypic variability of complex traits. In the study, the focus was moved from single loci to functional units, expressed by the metabolic pathways (KEGG). Consequently, this study aimed to estimate KEGG effects on stature in three Nordic dairy cattle breeds using SNPs effects from GWAS as the dependent variable. The SNPs were annotated to genes, then the genes to KEGG pathways. The effects of KEGG were estimated separately for each breed using a mixed linear model incorporating the similarity between pathways expressed by common genes. The analysis yielded two conclusions: a single gene in a pathway's component may have a large impact on the pathway's overall effect or the high pathway effect may be due to the influence of many genes constituting the pathway.

Introduction

Genome-Wide Association Studies (GWAS) are very useful for the identification of polymorphic sites, typically Single Nucleotide Polymorphisms (SNPs), or sometimes genes associated with a phenotypic variation or with a disease. Nowadays, the common availability of SNPs obtained based on whole-genome sequencing allows for a very good resolution for the estimation of those associations. However, in the context of phenotypes undergoing a complex mode of inheritance, it is not expected that a few genes / SNPs suffice to explain the variability on a phenotypic level. As a consequence, we often manage to identify loci with a very high effect on the phenotypic variation, but still, a predominant proportion of this variation remains unexplained (Manolio et al. 2009), since it is often due to a combined effect of many loci, each with a moderate or small impact. Therefore, in our study, we moved the focus from individual loci to functional units, here expressed by the metabolic (KEGG) pathways, to better understand the physiological mechanisms underlying complex phenotypes. For this purpose, we use SNP summary statistics originating from the GWAS conducted for stature and based on whole-genome sequence data of three Nordic dairy cattle breeds.

Materials & Methods

Material. The analysed data comprised SNP summary statistics from GWAS performed on 5,062 Danish Holstein bulls, 924 Danish Red Dairy Cattle bulls, and 2,122 Finnish Red Dairy Cattle bulls (Bouwman et al. 2018). The association was calculated for 25.4 million variants imputed with Minimac2 (Fuchsberger et al. 2015) from 630,000 SNPs using the 1000 Bull Genomes reference population from Run4, consisting of 1,147 individuals. SNP additive effects were estimated for deregressed EBVs serving as pseudophenotypes, separately for each breed

with a single SNP, a mixed linear model including a polygenic effect described by a genomic relationship matrix implemented in the EMMAX software (Kang et al. 2010).

Statistical model. Based on their IDs, SNPs were annotated to genes corresponding to the ARS-UCD1.2 reference genome using BIOMART (Smedley et al. 2009), next KEGG manually drawn reference pathways (map) were annotated to genes using the David software (Huang et al. 2009). The effects of KEGG pathways on stature were estimated separately for each breed using the following mixed linear model that accounted for the similarity between pathways:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Zt} + \boldsymbol{e} \quad (1),$$

where $\boldsymbol{y}$ is the vector of absolute values of SNP additive effects on stature from GWAS, $\boldsymbol{\mu}$ represents the general mean, $\boldsymbol{t}$ is the random effect of KEGG pathways with a preimposed normal distribution defined by $N(0, \boldsymbol{V}\sigma_t^2)$, $\boldsymbol{e}$ is a vector of residuals distributed as $N(0, \boldsymbol{I}\sigma_e^2)$, $\boldsymbol{Z}$ is an incidence matrix for $\boldsymbol{t}$. The similarity between KEGGs *i* and *j*, was introduced into the model by incorporating a nondiagonal KEGG covariance matrix $\boldsymbol{V}$. This covariance was expressed by the Jaccard similarity coefficient:

$$J(i,j) = \frac{M}{N}, \quad (2),$$

where *M* represents the number of genes shared between KEGG *i* and *j,* while N represents the total number of genes involved in KEGG *i* and *j*. Variance components were assumed as known, amounting $\sigma_t^2 = 0.3\sigma_y^2$ and $\sigma_e^2 = 1 - \sigma_t^2$. To avoid effect confounding due to high linkage disequilibrium, only one SNP per gene (with the highest effect) was included in $\boldsymbol{y}$.

Solutions. The mixed model equations (Henderson 1984) were used to obtain solutions for $\boldsymbol{\mu}$ and $\boldsymbol{t}$:

$$\begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{t}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}^T\boldsymbol{R}^{-1}\boldsymbol{1} & \boldsymbol{1}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{1} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{1}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}, \text{ where } \boldsymbol{R} = \boldsymbol{I}\hat{\sigma}_e^2 \text{ and } \boldsymbol{G} = \boldsymbol{V}\hat{\sigma}_t^2. \quad (3)$$

To maximise the computational performance of the estimation/prediction process, a custom Python program implementing the NumPy library (Harris et al. 2020) was used. Since all calculations were carried out on a high-performance server, the NumPy library was also used to set the array indexing and order what further improved the computing time compared to a native Python application.

Results

Effects of 180 pathways were estimated. Table 1 summarises the top 10 pathways for each breed. Because NumPy and SciPy APIs are implemented with LAPACK and BLAS, which require Fortran memory layout, all input matrices were transformed to Fortran order to avoid costly transpose. In comparison to a fixed matrix input, this approach results in 10 times faster estimation process.

Danish Holstein			Danish Red Dairy Cattle			Finnish Red Dairy Cattle		
KEGG ID	KEGG name	Higher metabolic hierarchy	KEGG ID	KEGG name	Higher metabolic hierarchy	KEGG ID	KEGG name	Higher metabolic hierarchy
00966	**Glucosinolate biosynthesis**	Biosynthesis of other secondary metabolites	**00253**	**Tetracycline biosynthesis**	Metabolism of terpenoids and polyketides	**00253**	**Tetracycline biosynthesis**	Metabolism of terpenoids and polyketides
00642	**Ethylbenzene degradation**	Xenobiotics biodegradation and metabolism	**00642**	**Ethylbenzene degradation**	Xenobiotics biodegradation and metabolism	**00945**	**Stilbenoid, diarylheptanoid and gingerol biosynthesis**	Biosynthesis of other secondary metabolites
00333	Prodigiosin biosynthesis	Biosynthesis of other secondary metabolites	**00945**	**Stilbenoid, diarylheptanoid and gingerol biosynthesis**	Biosynthesis of other secondary metabolites	**01051**	**Biosynthesis of ansamycins**	Metabolism of terpenoids and polyketides
00281	Geraniol degradation	Metabolism of terpenoids and polyketides	**01051**	**Biosynthesis of ansamycins**	Metabolism of terpenoids and polyketides	**00642**	**Ethylbenzene degradation**	Xenobiotics biodegradation and metabolism
00363	Bisphenol degradation	Xenobiotics biodegradation and metabolism	**00364**	**Fluorobenzoate degradation**	Xenobiotics biodegradation and metabolism	**00966**	**Glucosinolate biosynthesis**	Biosynthesis of other secondary metabolites
00253	**Tetracycline biosynthesis**	Metabolism of terpenoids and polyketides	**00966**	**Glucosinolate biosynthesis**	Biosynthesis of other secondary metabolites	**00364**	**Fluorobenzoate degradation**	Xenobiotics biodegradation and metabolism
00626	Naphthalene degradation	Xenobiotics biodegradation and metabolism	**00903**	**Limonene and pinene degradation**	Metabolism of terpenoids and polyketides	**04916**	Melanogenesis	-
00121	Secondary bile acid biosynthesis	Lipid metabolism	**01040**	**Biosynthesis of unsaturated fatty acids**	Lipid metabolism	**00903**	**Limonene and pinene degradation**	Metabolism of terpenoids and polyketides
00572	Arabinogalactan biosynthesis	Glycan biosynthesis and metabolism	00571	Lipoarabinomannan biosynthesis	Glycan biosynthesis and metabolism	00791	Atrazine degradation	Xenobiotics biodegradation and metabolism
00650	Butanoate metabolism	Carbohydrate metabolism	**04916**	Melanogenesis	-	**01040**	**Biosynthesis of unsaturated fatty acids**	Lipid metabolism

Table 1. The top 10 KEGG pathways, based on their effect on stature estimated by model (1).

Discussion

While interpreting KEGG pathways effects two scenarios emerge. On the one hand, the overall high effect of a pathway may be driven by a high effect of a single gene that is this pathway's component – a situation that could have been detected in a conventional GWAS. On the other hand, the high pathway effect may be due to the influence of many genes constituting this pathway – a situation that may easily be missed in GWAS due to the small or moderate effects of particular genes. In our study, an emerging pattern is the high effect of pathways associated with the metabolism of xenobiotics (map00363, map00364, map00626, map00642, map00791). Already over 20 years ago Cheriathundam et al. (1998) indicated that the Growth hormone regulates the metabolism of enzymes from the cytochrome family which are responsible for the metabolism of a large number of xenobiotics. Furthermore, tetracycline (map00253) induces expression of the SHOX gene, which results in altered human growth (see e.g. Marchini et al. 2007). Bile acid synthesis (map00121) disorder in humans is known to influence weight and height (see e.g. Heubi and Setchell 2020).

References

Bouwman A.C., Daetwyler H.D., Chamberlain A.J., Ponce C.H., Sargolzaei M., *et al*. (2018) Nat. Genet. 50(3):362–367.

Cheriathundam E., Doi S.q., Knapp J.R., Jasser M.Z., Kopchick J.J. (1998) Biochem. Pharmacol. 55(9): 1481–1487.

Fuchsberger C., Abecasis G.R., Hinds D.A. (2015) Bioinformatics. 31(5):782–784.

Harris C.R., Millman K.J., van der Walt S.J., Gommers R., Virtanen P., *et al.* (2020) Nature, 585:357–362.

Henderson C.R. (1984) University of Guelph.

Heubi J., Setchell K. (2020) J. Pediatr. Gastroenterol. Nutr. 70(4):423–429.

Huang D.W., Sherman B.T., Lempicki R.A. (2009) Nature Protoc. 4(1):44–57.

Kang H.M., Sul J.H., Service S.K., Zaitlen N.A., Kong S.Y., *et al*. (2010) Nat. Genet. 42(4):348–354.

Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., *et al*. (2009) Nature. 461(7265):747–753.

Marchini A., Häcker B., Marttila T., Hesse V., Emons J., *et al*. (2007) Hum. Mol. Genet. 16(24):3081–3087.

Smedley D., Haider S., Ballester B., Holland R., London D., *et al*. (2009) BMC Genom. 10:.22