The application of mixed linear models for the estimation of functional effects on bovine stature based on SNP summary statistics from a whole-genome association study

Krzysztof Kotlarz[1], Barbara Kosinska-Selbi[1], Zexi Cai[2], Goutam Sahana[2] and Joanna Szyda[1,3]

[1] Biostatistics Group, Department of Genetics, the 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;

Corresponding author:

Joanna Szyda

Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland

joanna.szyda@upwr.edu.pl

Running title:

KEGG effects on bovine stature based on WGS

Abstract

Genome-Wide Association Studies (GWAS) help identify polymorphic sites or genes linked to phenotypic variance, but a few identified genes / Single Nucleotide Polymorphisms are unlikely to explain a large part of the phenotypic variability of complex traits. In this study, the focus was moved from single loci to functional units, expressed by the metabolic pathways: Kyoto Encyclopaedia of Genes and Genomes (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 KEGG pathway D-amino acid metabolism (map00473) was estimated as significant on stature in two of the analysed breeds and revealed a borderline significance in the third breed. Interestingly, biological evidence exists that described the importance of D-amino acids for growth in experimental organisms as well as in cattle.

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 of 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 locus to functional units, here expressed by the metabolic pathways defined by the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database. This approach allows us to better understand the physiological mechanisms underlying complex phenotypes. For this purpose, we used SNP summary statistics originating from the GWAS conducted for stature and based on whole-genome sequence data of three Nordic dairy cattle breeds.

Results

The effects of 179 KEGG pathways were estimated based on the effects of selected SNPs from a whole-genome sequence-based GWAS of Bouwman et al. (2018), separately for three Nordic cattle breeds - Danish Holstein (DH with 366,877 SNPs), Danish Red Dairy Cattle (DR with 299,723 SNPs), and Finnish Red Dairy Cattle (FR with 396,224 SNPs) (Figure 1). In two breeds, the same pathway - D-amino acid metabolism (map00473) revealed a significant effect on stature with moderate P-values of 0.035 in FR and 0.049 in DH. In DR it also reached a borderline significance of 0.133. Depending on the breed, the effect of map00473 was estimated based on 78 SNPs in DH and FR, and 76 SNPs in DR (Figure 2,

Supplemental Data S1). The differences in SNP counts resulted from the fact that the input SNP panel in Bouwman et al. (2018) was pre-processed separately for each breed, which resulted in breed-specific SNP exclusion.

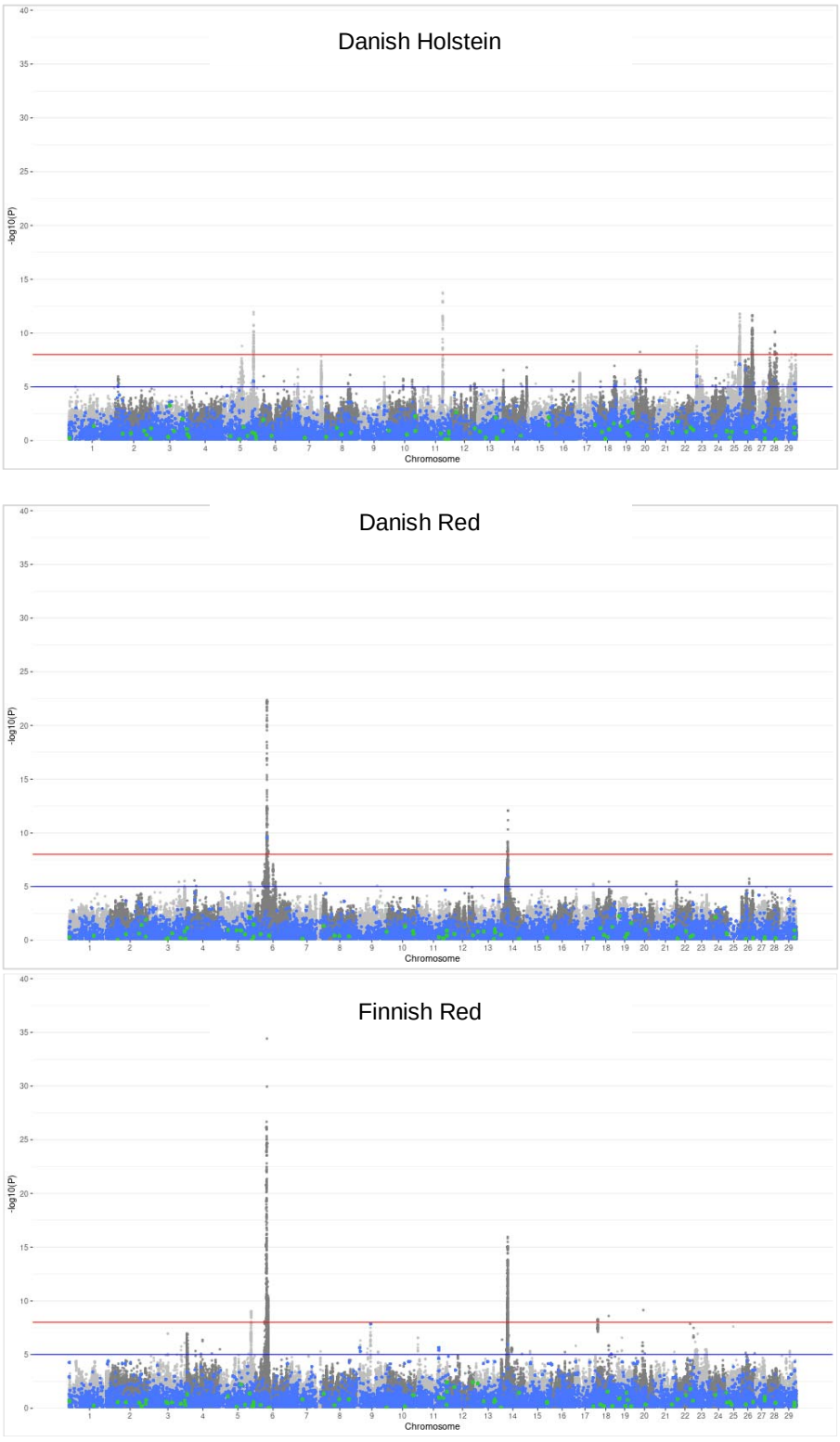

Figure 1. SNP significance from the whole-genome sequencing study of Bouwman et al. (2018). Blue dots correspond to genic SNPs used for the estimation of KEGG pathway effects

in model (1), green dots correspond to SNPs marking genes constituting the map00473 pathway, and gray SNPs are the remainder.

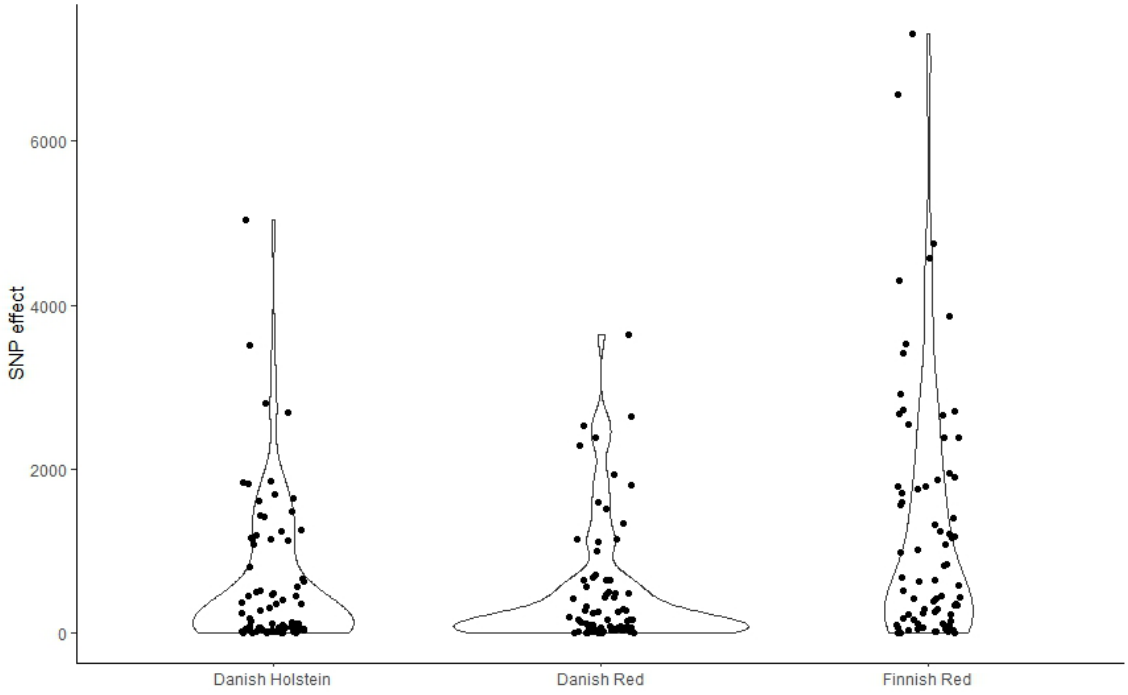

Figure 2. Estimated effects of SNPs marking genes from the map00473 pathway.

Additionally, the pathway responsible for the metabolism of terpenoids and polyketides (map01059) was significant (P=0.041) in DH, while the synthesis and degradation of ketone bodies pathway (map00072) and the pathway of biosynthesis of various plant secondary metabolites (map00999) were significant in DR with P=0.037 and P=0.047 respectively.

Discussion

While interpreting KEGG pathway 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 genome-wide association study (GWAS). On the other hand, the high pathway effect may be due to the combined effects 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 from the pathway.

In the case of our data – none of the genes harbouring the most significant SNPs in GWAS performed by Bouwman et al. (2018) was the component of the D-amino acid metabolism pathway, which therefore leads to the conclusion that the whole pathway is a significant component of the genetic determination of stature. Biologically, an outstanding pattern of our study was that the pathway associated with the metabolism of D-amino acids in all three breeds is significant for two breeds and on the border of claimed significance in the third breed. Although D-amino acids do not occur in naturally translated proteins, the link between D-amino acids metabolism and growth has long been recognised. Experimentally, a supplementation of mice with D-amino acids resulted in increased weight observed with the increased concentration of D-Phenylalanine and D-Tryptophan in the diet (Friedman and Levin, 2012). Moreover, D'Aniello (2007) reported that, in the pituitary gland, D-aspartic acid stimulates the secretion of the growth hormone in rats. In cattle, a supplementation of food with synthetic amino acids is a very common practice with commercial diet supplements containing a mixture of naturally occurring L-versions as well as not naturally occurring D-version. Campbell et al. (1996) observed that D-amino acids are somewhat less efficiently metabolised than their naturally occurring synonyms. Since methionine is often the first limiting amino acid for growth in cattle (Richardson and Hatfield, 1978) individuals that possess a more efficient mechanism of D-amino acid metabolism are expected to grow better which may result in higher stature in adults.

Another metabolic pathway demonstrating potential importance on stature is the synthesis and degradation of the ketone bodies pathway (map00072) that was significant in DR. It has been demonstrated that ketone bodies metabolism is related to growth on the whole organism (mainly through the *SLC16A6* gene as reported by Kichaev et al. (2019) and Karanth et al. (2019)) as well as on the single-cell level (Kolb et al. 2021). Moreover, although the other significant pathway of biosynthesis of various plant secondary metabolites does not directly

relate to animal metabolism, it can be hypothesised that genes playing a role in the biochemical processing of metabolites originating from plants lead to higher feed efficiency in cattle and furthermore influence animals growth, but experimental evidence is lacking.

Still, our results demonstrate that considering higher-order components of biological systems, such as metabolic pathways, provides a valuable insight into the basis of the variation of complex phenotypes, that may be missed by conventional GWAS analyses and should be used as an enhancement thereof.

Materials and 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 mixed linear model including an additive polygenic effect with a covariance described by a genomic relationship matrix. The model was implemented via 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 Bioconductor BioMart tool version 3.14 (Smedley et al. 2009) and then genes were annotated to KEGG reference pathways (map) using the David software version 6.8 (Huang et al. 2007). 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:

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

where $\boldsymbol{y}$ is the vector of absolute values of SNP additive effects on stature estimated in GWAS of Bouwman et al. (2018), $\boldsymbol{\mu}$ 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 $\boldsymbol{t}$. Note that if multiple SNPs were identified within a gene only one SNP with the highest effect was included in $\boldsymbol{y}$, so that each gene is represented by a single variant. 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 $\sigma_e^2 = 0.7\sigma_y^2$.

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 1.19.5 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 which further improved the computing time compared to a native Python application. Each element of $\hat{\boldsymbol{t}}$ was assessed for significance (H_0: $\hat{t}_i \leq 0$ vs. H_1: $\hat{t}_i > 0$) by calculating the probability of obtaining a more extreme value from the $N(0, \sigma_t^2)$ density function.

Since 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 transposing. In comparison to a fixed matrix input, this approach results in a ten times faster estimation process.

Competing Interest

The authors declare no competing interests.

Data Access

Accession codes are available at https://doi.org/10.1038/s41588-018-0056-5.

Acknowledgment

The genome-wide association studies were part of the Center for Genomic Selection in Animals and Plants (GenSAP) financed by Innovation Fund Denmark (Grant: 0603-00519B). The 1000 Bull Genomes Project is acknowledged for sharing sequence data for imputation. We thank prof. Bernt Guldbrandtsen for fruitful discussions on pathway modelling.

Calculations have been carried out using resources provided by Wroclaw Centre for Networking and Supercomputing.

References

Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, Sahana G, Govignon-Gion A, Boitard S, Dolezal M. et al. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet* **50:** 362-367. doi:10.1038/s41588-018-0056-5.

Campbell CG, Titgemeyer EC, St-Jean G. 1996. Efficiency of D- vs L-methionine utilization by growing steers. *J Anim Sci* **74:** 2482-2487. doi:10.2527/1996.74102482x.

D'Aniello A. 2006. D-Aspartic acid: an endogenous amino acid with an important neuroendocrine role. *Brain Res Rev* **53:** 215-234. doi:10.1016/j.brainresrev.2006.08.005.

184 Friedman M, Levin CE. 2012. Nutritional and medicinal aspects of D-amino acids. *Amino*
185 *Acids* **42:** 1553-1582. doi: 10.1007/s00726-011-0915-1.
186 Fuchsberger C, Abecasis GR, Hinds DA. 2015. minimac2: faster genotype imputation.
187 *Bioinformatics* **31:** 782-784. doi:10.1093/bioinformatics/btu704.
188 Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E,
189 Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature* **585:** 357-
190 362. doi: 10.1038/s41586-020-2649-2.
191 Henderson CR 1984. University of Guelph.
192 Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler
193 MW, Lane HC, Lempicki RA. 2007. The DAVID Gene Functional Classification Tool: a
194 novel biological module-centric algorithm to functionally analyze large gene lists. *Genome*
195 *Biol* **8:** R183. doi:10.1186/gb-2007-8-9-r183.
196 Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010.
197 Variance component model to account for sample structure in genome-wide association
198 studies. *Nat Genet* **42:** 348-354. doi:10.1038/ng.548.
199 Karanth S, Schlegel A. 2019. The Monocarboxylate Transporter SLC16A6 Regulates Adult
200 Length in Zebrafish and Is Associated With Height in Humans. *Front Physiol* **9:** 1936.
201 doi:10.3389/fphys.2018.01936.
202 Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price
203 AL. 2019. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J*
204 *Hum Genet* **104:** 65-75. doi:10.1016/j.ajhg.2018.11.008.
205 Kolb H, Kempf K, Röhling M, Lenzen-Schulte M, Schloot NC, Martin S. 2021. Ketone
206 bodies: from enemy to friend and guardian angel. *BMC Med* **19:** 313. doi:10.1186/s12916-
207 021-02185-0.

208 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI,
209 Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of
210 complex diseases. *Nature* **461:** 747-753. doi:10.1038/nature08494.
211 Richardson CR, Hatfield EE. 1978. The limiting amino acids in growing cattle. *J Anim Sci*
212 **46:** 740-745. doi:10.2527/jas1978.463740x.
213 Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. 2009.
214 BioMart--biological queries made easy. *BMC Genomics* **10:** 22. Doi:10.1186/1471-2164-10-
215 22.