doi: 10.17221/43/2015-CJAS

Using gene networks to identify genes and pathways involved in milk production traits in Polish Holstein dairy cattle

T. Suchocki[1], K. Wojdak-Maksymiec[2], J. Szyda[1]

[1]Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2]Department of Animal Genetics and Breeding, West Pomeranian University of Technology in Szczecin, Szczecin, Poland

ABSTRACT: When analyzing phenotypes undergoing a complex mode of inheritance, it is of great interest to switch the scope from single genes to gene pathways, which form better defined functional units. We used gene networks to search for physiological processes and underlying genes responsible for complex traits recorded in dairy cattle. Major problems addressed included loss of information from multiple single nucleotide polymorphisms (SNPs) located within or close to the same gene, ignoring information on linkage disequilibrium and validation of the obtained gene network. 2601 bulls genotyped by the Illumina BovineSNP50 BeadChip were used. SNP effects were estimated using a mixed model, then underlying gene effects were estimated and tested for significance, subsequently a gene network was constructed and the functional information represented by the network was retrieved. The networks were validated by repeating the above-mentioned analyses after permutation of bulls' pseudophenotypes. Effects of 4345 genes were estimated, what makes 16.4% of all genes mapped to the UMD3.1 reference genome. Assuming the maximum 10% type I error rate, for milk yield 50 different gene ontology (GO) terms and three pathways defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) were significantly overrepresented in the real data as compared to the permuted data sets, while for fat yield nine of the GO terms were significantly overrepresented in the real data network, although none of the KEGG pathways reached the significance level. In turn, for protein yield 28 of the GO terms and six KEGG pathways were significantly overrepresented in the real data. Based on the physiological information we identified sets of loci involved in the determination of milk yield (224 genes), fat yield (72 genes), and protein yield (546 genes). Among the genes some have large effects and have already been reported in previous studies, whereas some others represent novel discoveries and thus most probably genes with medium or small effects on trait variation.

Keywords: cattle; gene networks; GO; GWAS; KEGG; mixed model; SNP; validation

INTRODUCTION

The major motivation for this research was to identify pathways and genes related to the variation of production traits in dairy cattle. This was done by using a novel approach where gene networks are built upon estimates of gene effects, and by using permutations to estimate the relevance of found gene networks.

Genetically, production traits represent complex phenotypes typically determined by several genes with large effects – the so-called major genes or quantitative trait loci (QTL), a number of genes with intermediate effects and a large number of genes, each with a very small effect, cumulatively known as polygenes. A traditional way to identify those genes is to apply genome-wide association analysis (GWAS) (Bolormaa et al. 2010) or genomic selection models

Supported by the Polish National Science Centre (grant No. N N311 609639).

doi: 10.17221/43/2015-CJAS

(VanRaden 2008). However, two major drawbacks of GWAS when applied to traits with a complex mode of inheritance include difficult selection of significant polymorphisms among single nucleotide polymorphisms (SNPs) intercorrelated through linkage disequilibrium (LD) and poor reproducibility of results across methods. As a multiple SNP method, the genomic selection approach accounts for SNP intercorrelation, but suffers from the problem of shrinking SNP estimates to a preimposed, usually normal, distribution. Anyhow, by applying the aforementioned methodology to data sets currently available for research, which are very informative thanks to the large number of SNPs and individuals, it is usually relatively easy to identify major genes. Moreover, in some studies which use very large data sets, comprising thousands of phenotyped individuals and their genotypes, researchers are even able to identify genes with medium effects. However, it is still very difficult to pinpoint genes with small effects. For this purpose some bioinformatics tools can be applied such as the gene network approach, which profits from the information stored in publicly available databases, or methods based on the gene set enrichment analysis (GSEA).

When analyzing phenotypes undergoing a complex mode of inheritance it is of great interest to switch the scope from single genes to gene pathways, which form better defined functional units. Therefore, in our study we applied the gene network analysis in order to search for functional information and genes responsible for milk-, fat-, and protein yields in dairy cattle. Methodologically, two major problems addressed in this study were: (*i*) accounting for information on LD between SNPs linked to the same gene and (*ii*) validation of the obtained gene network. The course of the analyses comprises: (*i*) estimation of SNP effects using the SNP-BLUP model, (*ii*) estimation of gene effects based on SNP effects located in the vicinity of coding regions and significance testing, (*iii*) construction of gene networks using significant genes as a scaffold, (*iv*) construction of an empirical null hypothesis distribution for the network via permutation, (*v*) significance testing for the overrepresentation of gene ontology (GO) terms and pathways defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) underlying the gene network constructed in step (*iii*) as compared to the set of GO terms and KEGG pathways underlying the null distribution constructed in step (*iv*).

MATERIAL AND METHODS

Animal data. A total of 2601 bulls from the Polish Holstein Friesian dairy cattle breed were used in this analysis. The oldest bull was born in 1981 and the youngest animals were born in 2007. This core data was additionally enhanced by the information on pedigree relationship of 2 434 590 individuals and by the information on conventional breeding values of 10 355 individuals.

Three traits undergoing a complex mode of inheritance were selected for the analysis, i.e. milk-, fat-, and protein yields. Milk-, fat-, and protein yields are moderately heritable with heritabilities in the Polish Holstein-Friesian population estimated at 0.33, 0.29, and 0.29, respectively, and have been the most important selection criteria in dairy cattle for over 150 years (Lush et al. 1936). Since all the phenotypes are measured on cows, our analysis utilizes deregressed conventional breeding values of bulls which were estimated based on a random regression test day model (Strabel and Jamrozik 2006). Deregression was performed following the method of Jairath et al. (1998) in order to obtain bulls' pseudophenotypes independent of additive genetic relationship. The means and standard deviations for the pseudophenotypes of the analyzed bulls amounted to 153.1 ± 547.5 kg of milk, 3.6 ± 19.0 kg of fat, and 5.6 ± 15.0 kg of protein per lactation.

Bulls were genotyped by the Illumina BovineSNP50 Genotyping BeadChip, which consists of 54 001 SNPs (version 1) and 54 609 SNPs (version 2). Genotype samples were provided within the frame of the Genomika Polska project and comprised semen probes acquired via a routine semen collection procedure. After genotype preprocessing comprising elimination of SNPs with minor allele frequencies of less than 0.01 and call rate under 90%, 46 267 SNPs were selected for further analysis.

SNP effect estimation. The first step of the analysis comprised estimation of effects of SNPs on the selected complex traits. This was done using the following mixed model:

$$y = \mu + \mathbf{Z}g + e \quad (1)$$

where:

y = vector of deregressed conventional breeding values of bulls for milk-, fat-, and protein yields, respectively

μ = general mean

doi: 10.17221/43/2015-CJAS

Z = design matrix for SNP genotypes, which is parameterized as –1, 0, or 1 for a homozygous, a heterozygous, and an alternative homozygous genotype, respectively
g = vector of random additive SNP effects
e = vector of residuals

The covariance structure of the model comprises:

$$g \sim N(0, \mathbf{I}\frac{\sigma_a^2}{46267}) \text{ and } e \sim N(0, \mathbf{D}\hat{\sigma}_e^2)$$

where:
I = identity matrix
$\hat{\sigma}_a^2$ = estimate of total additive genetic variance of a given trait calculated elsewhere for the whole active population of Polish Holstein-Friesian dairy cattle
D = diagonal matrix of the reciprocal of the (effective) number of daughters of each bull
$\hat{\sigma}_e^2$ = estimate of residual variance

The covariance structure of $\boldsymbol{y}$ is as follows:

$$\boldsymbol{y} = \mathbf{ZGR}^{\mathrm{T}} + \mathbf{R}$$

where:
$\mathbf{G} = \mathbf{I}\frac{\sigma_e^2}{46267}$ and $\mathbf{R} = \mathbf{D}\hat{\sigma}_e^2$

The estimation of parameters of the above model was based on solving the mixed model equations (Henderson 1984):

$$\begin{bmatrix}\hat{\mu}\\ \hat{\boldsymbol{g}}\end{bmatrix} = \begin{bmatrix}1^{\mathrm{T}}\mathbf{R}^{-1}1 & 1^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}^{-1}1 & \mathbf{Z}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{bmatrix}^{-1}\begin{bmatrix}1^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{y}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{y}\end{bmatrix}$$

The iteration on data technique was based on the Gauss-Seidel algorithm with residuals update (Legarra and Misztal 2008).

Gene effect estimation. In the next step SNP effect estimates were enhanced by the information on SNP genomic location and LD to form gene effect estimates. For each SNP its genomic location corresponding to the bovine genome build release 68 was retrieved from the Ensembl database (http://www.ensembl.org) and the statistic for pairwise LD was calculated using PLINK (Purcell et al. 2007) based on genotype correlations. Gene effects were then estimated using:

$$t = \frac{\sum_{i=1}^{N_t}|\hat{g}_i|}{\sqrt{\sigma_t^2}} \quad (2)$$

where:
$\hat{g}_i$ = estimate of i^{th} SNP effect
N_t = number of SNPs located within the gene or maximally 1 Kbp from the gene borders
σ_t^2 = variance of a gene effect

The variance of a gene effect σ_t^2 is expressed by:

$$\sigma_t^2 = \sum_{i=1}^{N_t}\sigma_{q_i}^2 + 2\sum_{i=1}^{N_t}\sum_{j=i+1}^{N_t}\sigma_{q_i}\sigma_{q_j} = N_t\,\sigma_q^2 + 2\sum_{i=1}^{N_t}\sum_{j=i+1}^{N_t} r_{q_i q_j}\,\sigma_q^2$$

where:
σ_q^2 = variance of a SNP effect (identical for all SNPs) expressed by (σ_a^2/46267)
r = square root of r^2 (i.e. linkage disequilibrium between SNP i and j)

If only one SNP is located inside the gene, then the following is true:

$$2\sum_{i=1}^{N_t}\sum_{j=i+1}^{N_t}\sigma_{q_i}\sigma_{q_j} = 0$$

Asymptotically, the gene effect statistics (t) follows a standard normal distribution and thus the N(0,1) significance thresholds were used to select significant genes. Since the major issue of the study was focused on genes with small effects, we decided that in a trade-off between type I and type II errors the latter are much more important on the gene selection stage. Consequently, a significance threshold of 0.20 was used to select genes for further analysis based on t.

Network construction and retrieving functional information. Separately for each trait, a network was constructed using significant genes for that trait as a scaffold. A BisoGenet plugin (Martin et al. 2010) for the Cytoscape software (Shannon et al. 2003) was used as a tool to build a gene network based on the information from the SysBiomics database, which integrates such publicly accessible databases as NCBI Entrez Gene, Uniprot, BIND, HPRD, Mint, DIP, BioGRID, and Intact. Networks were constructed using the significant genes as primary input nodes enhanced by additional nodes, representing genes for which known interactions with the input nodes were identified in the databases. A network was enhanced by allowing for maximally one new node between the original input genes. Since no access to information on *Bos taurus* was available through the software, human homologues for the significant genes were identified and the *Homo sapiens* database was used.

Permutation and significance testing. In order to assess how reliable a particular network was in terms of underlying functional information, a null hypothesis distribution of the network was constructed empirically based on permutations. In this study the functional information describing the genetic background of a trait was expressed

doi: 10.17221/43/2015-CJAS

by GO terms (http://geneontology.org) and KEGG pathways (http://www.kegg.jp) underlying genes from the network. The empirical null hypothesis distribution of the network, which reflects the frequency of GO terms and KEGG pathways, was constructed based on permutations. For this purpose, for each trait separately, the following steps were repeated 100 times in order to approximate the distribution: (1) permutation of deregressed breeding values (vector *y*), (2) SNP effect estimation based on the model (1), (3) gene effect estimation and testing using the statistic *t* (2), (4) gene network construction, (5) identification of functional information (GO terms and KEGG pathways) represented by the network.

Testing of the hypotheses regarding the significance of a particular GO term or a KEGG pathway was based on the Odds Ratio (OR) statistics. It involved comparing the number of times the given feature (GO or KEGG) was represented by a gene within the network resulting from original (unpermuted) data with the number of times this feature was represented in the network resulting from permuted data. The underlying assumption is that permuted data sets represent empirically derived H_0 distribution of GO/KEGG:

$$OR = \frac{\frac{C_o + 0.5}{(N_o - C_o) + 0.5}}{\frac{C_p + 0.5}{(N_p - C_p) + 0.5}} \quad (3)$$

where:

C_x = number of times a given GO term or a KEGG pathway was observed among genes building the original (C_o) and permuted (C_p) networks

N_x = total number of occurrences observed by the features in the original (N_o) and permuted (N_p) networks

The underlying hypotheses can be defined in terms of the probability to observe a functional feature (GO term or KEGG pathway) in the original (P_o) and permuted (P_p) data sets, respectively: $H_0 : P_0 = P_p$ and $H_1 : P_0 \neq P_p$. The natural logarithm

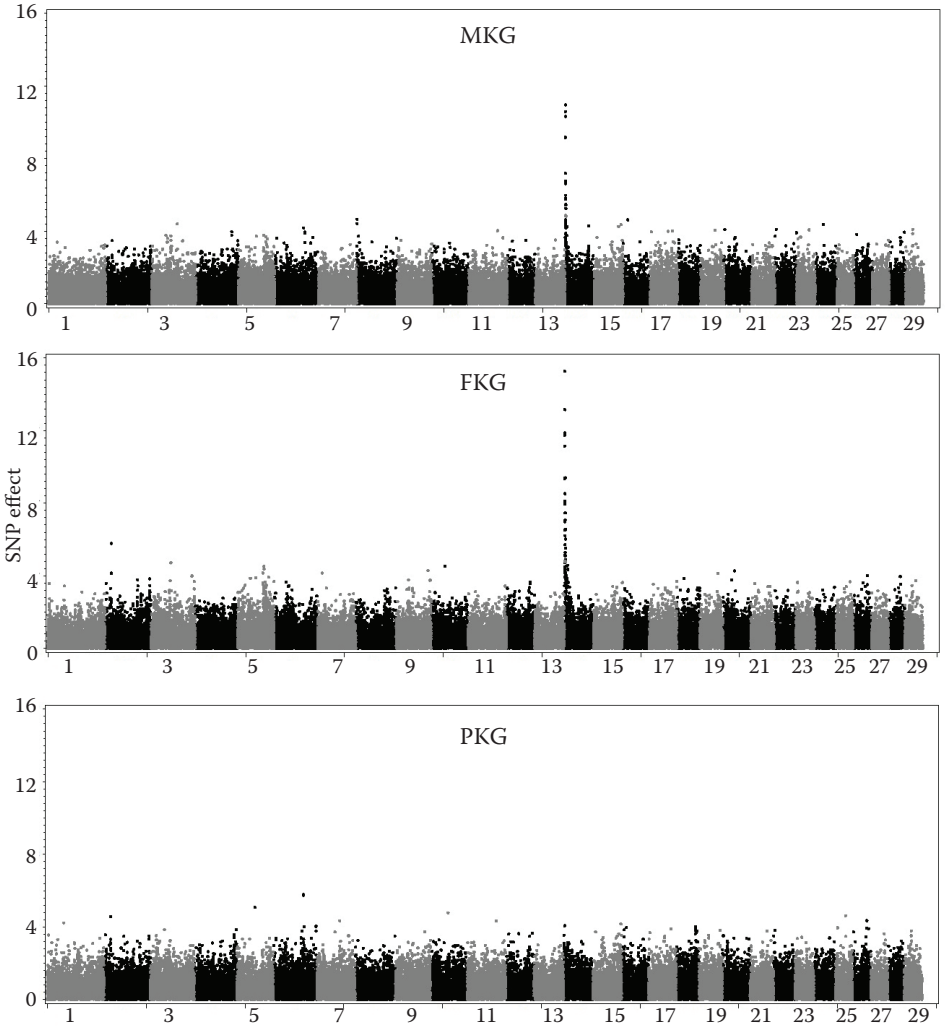

Figure 1. Standardized additive effects of single nucleotide polymorphisms (SNPs). Absolute values of 46 267 SNPs for milk yield (MKG), fat yield (FKG), and protein yield (PKG)

Czech J. Anim. Sci., 61, 2016 (11): 526–538

doi: 10.17221/43/2015-CJAS

Table 1. Significant genes selected based on single nucleotide polymorphism (SNP) estimates in the real data set

Gene symbol	Gene full name	Gene ID	Human homologs ID	KEGG	Gene position on UMD3.1	Number of SNPs in gene	Gene additive effect	*P*-value
Milk yield								
DGAT1	*diacylglycerol O-acyltransferase 1*	282609	8694	bta00561 bta00830 bta01100 bta04975	14:1795351-1804562	1	7.52	0.00047
LOC786966/PLEC	*plectin*	786966	5339	bta04812	14:2054917-2088261	1	7.37	0.00059
C8orf33	*chromosome 8 open reading frame 33*	513585	65265	–	14:1487366-1489409	1	6.29	0.00335
MAF1	*MAF1 homolog (S. cerevisiae)*	512498	84232	–	14:1921784-1924818	1	4.94	0.02138
MAPK15	*mitogen-activated protein kinase 15*	512125	225689	bta01000 bta01001	14:2235034-2240765	1	3.95	0.06600
RHPN1	*rhophilin, Rho GTPase binding protein 1*	618397	114822	–	14:2462544-2471434	1	2.90	0.17846
LY6D	*lymphocyte antigen 6 complex, locus D*	618714	8581	bta00537	14:2801383-2803020	1	2.79	0.19241
Fat yield								
DGAT1	*diacylglycerol O-acyltransferase 1*	282609	8694	bta00561 bta00830 bta01100 bta04975	14:1795351-1804562	1	0.39	< 0.00001
LOC786966/PLEC	*plectin*	786966	5339	bta04812	14:2054917-2088261	1	0.31	0.00022
MAF1	*MAF1 homolog (S. cerevisiae)*	512498	84232	–	14:1921784-1924818	1	0.28	0.00093
C8orf33	*chromosome 8 open reading frame 33*	513585	65265	–	14:1487366-1489409	1	0.22	0.00883
RHPN1	*rhophilin, Rho GTPase binding protein 1*	618397	114822	–	14:2462544-2471434	1	0.18	0.03115
LY6D	*lymphocyte antigen 6 complex, locus D*	618714	8581	bta00537	14:2801383-2803020	1	0.17	0.03786
MAPK15	*mitogen-activated protein kinase 15*	512125	225689	bta01000 bta01001	14:2235034-2240765	1	0.17	0.04407
AGO2	*argonaute RISC catalytic component 2*	404130	27161	bta03019 bta03036	14:4085146-4168483	1	0.14	0.09502
GML	*glycosylphosphatidylinositol anchored molecule like*	767955	2765	–	14:2715416-2742638	1	0.11	0.19682
Protein yield								
AP1B1	*adaptor-related protein complex 1, beta 1 subunit*	506192	162	bta00001	17:70734993-70783359	2	0.09	0.13446
HEPHL1	*hephaestin-like 1*	519580	341208	–	29:653016-744427	1	0.09	0.14508
LHX8	*LIM homeobox 8*	512385	431707	bta03000	3:69890357-69916293	1	0.09	0.16909

doi: 10.17221/43/2015-CJAS

Table 1 to be continued.

Gene symbol	Gene full name	Gene ID	Human homologs ID	KEGG	Gene position on UMD3.1	Number of SNPs in gene	Gene additive effect	*P*-value
FBP2	*fructose-1,6-bisphosphatase 2*	514066	8789	bta00010 bta00030 bta00051 bta01100 bta01200 bta04152 bta04910	8:82396095-82438817	1	0.09	0.17416
TANC2	*tetratricopeptide repeat, ankyrin repeat and coiled-coil containing 2*	529494	26115	–	19:48089631-48406633	1	0.08	0.18452
DHX34	*DEAH (Asp-Glu-Ala-His) box polypeptide 34*	506965	9704	bta01000 bta03019	18:54786976-54810524	1	0.08	0.18670

KEGG = Kyoto Encyclopedia of Genes and Genomes

transformation of OR divided by its standard error follows a standard normal distribution, which allows for assessing corresponding *P*-values. In order to circumvent the problem of testing multiple GO terms and KEGG pathways, nominal *P*-values were subjected to the Bonferroni correction.

GSEA. Additionally, for the genes identified as significant in (2) we performed a GSEA as implemented in the KOBAS software (Mao et al. 2005). In order to be compatible with aforementioned analyses, *Homo sapiens* genome was chosen as the baseline data set.

RESULTS

SNP effects. Figure 1 shows Manhattan plots of additive SNP effect estimates along the genome for the three considered traits, rescaled to the standard normal distribution and expressed as absolute values. For milk yield a SNP with the highest estimate accounts for 7.60 kg of milk per lactation, for fat yield it accounts for 0.41 kg of fat per lactation, both attributed to the same SNP located on BTA14 within an intron of *DGAT1*, and for protein yield the SNP with the highest effect accounting for 0.13 kg of protein per lactation is located on BTA06 260 bp downstream of *Alpha-S2-casein Casocidin-1*.

Gene effects. Effects of 4345 genes were estimated, but due to a relatively low SNP density of the Bovine SNP50 BeadChip 87% of these estimates were based on a single SNP. The remainder consisted of estimates based on 2 to 6 SNPs. Figure 2 shows histograms of estimated gene effects and underlying normal densities for the analyzed traits. For milk-, fat-, and protein yield the empirical standard deviations of gene effects were 0.33, 0.36, and 0.28, respectively. Based on *t*, seven and nine genes were identified as significant for milk- and fat yield, respectively, all located on BTA14, with the highest effects of 7.52 kg of milk and 0.39 kg of fat per lactation, both attributed to the *DGAT1* gene. All genes significant for milk yield were also significant for fat yield. For protein yield six significant genes were observed, each located on a different chromosome, the most significant was *AP1B1* with the effect of 0.09 kg of protein per lactation (Table 1).

Gene networks and functional information. Milk yield was described by the network of 98 genes, representing 1115 various GO terms and 130 various KEGG pathways. Assuming the maximum 20% type I error rate corrected for multiple testing, 50 different GO terms and three KEGG pathways

doi: 10.17221/43/2015-CJAS

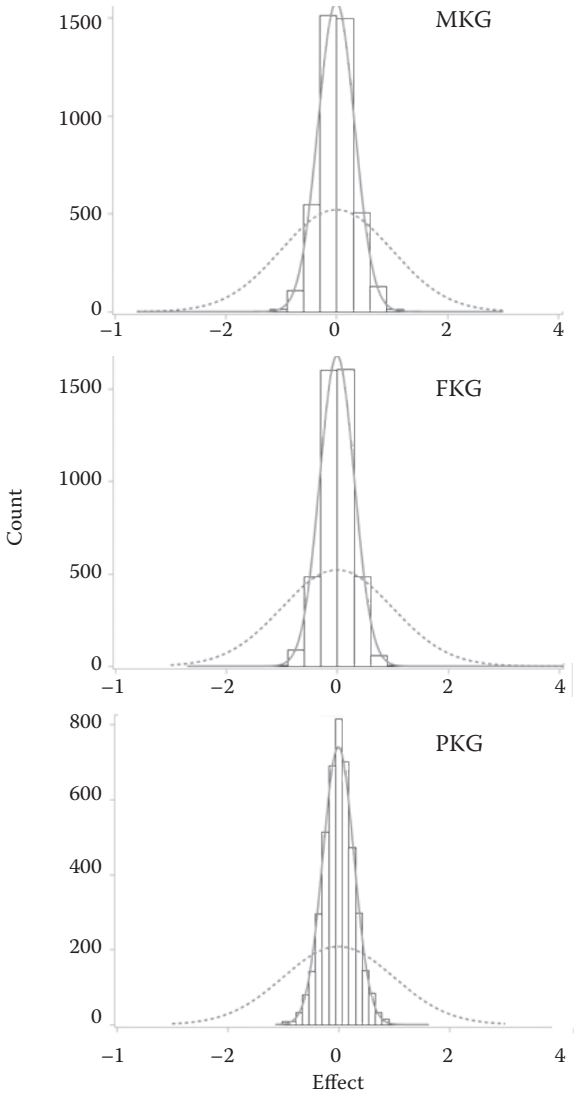

Figure 2. Gene effects distributions. Histogram of 4345 gene effects for milk yield (MKG), fat yield (FKG), and protein yield (PKG) with underlying normal density with empirical standard deviations (solid line) and standard normal densities (dashed line)

were significantly overrepresented in the real data as compared to the permuted data sets. The most significant among the 50 GO terms with *P*-values < 10^{-5} comprised: the condensin complex (GO:0000796), intercellular canaliculus (GO:0046581), and mitotic chromosome condensation (GO:0007076). Significant KEGG pathways comprised:

(1) **Arrhythmogenic right ventricular cardiomyopathy** (bta05412; *P*-value = 0.04464), this pathway also reached the significance level with false discovery rate (FDR) of 0.10972 in GSEA as estimated by KOBAS software;

(2) **Dilated cardiomyopathy** (bta05414; *P*-value = 0.07862), this pathway was not significant in GSEA (FDR = 0.16281);

(3) **Tight junction** (bta04530; *P*-value = 0.09542), this pathway was the second most significant one in GSEA, revealing FDR = 0.06847.

In total, the three significant pathways consist of 224 genes.

For fat yield a gene network consisted of 114 genes, which represented 1513 various GO terms and 147 KEGG pathways. Fourteen of the GO terms were significantly overrepresented in the real data network (as compared to the artificial networks generated by permutations) and one KEGG pathway reached the 10% significance level after the correction for multiple testing. The most highly significant GO terms (*P*-value < 10^{-5}) were represented by the negative regulation of translation involved in gene silencing by miRNA (GO:0035278) and cytoplasmic mRNA processing body (GO:0000932). The significant pathway was **RNA degradation** (bta03018; *P*-value = 0.04594). This pathway comprises 72 genes and was also on the border of significance in GSEA with FDR = 0.19183.

The network obtained for protein yield consisted of 44 genes assigned to 660 GO terms and 75 KEGG pathways. A total of 28 of the GO terms and six KEGG pathways were significantly overrepresented in the real data. GO terms with *P*-values < 10^{-5} comprised: antigen processing and presentation of the exogenous peptide antigen via major histocompatibility complex (MHC) class II (GO:0019886), clathrin adaptor complex (GO:0030131), clathrin-coated vesicle membrane (GO:0030665), post-Golgi vesicle-mediated transport (GO:0006892), regulation of defense response to virus by virus (GO:0050690), and trans-Golgi network membrane (GO:0032588). Significant KEGG pathways were:

(1) **Lysosome** (bta04142; *P*-value < 10^{-5}), this pathway was also estimated as significant in GSEA (FDR = 0.03352);

(2) **Cell cycle** (bta04110; *P*-value = 0.00005), the pathway was also significant in GSEA (FDR = 0.00001);

(3) **Pentose phosphate pathway** (bta00030; *P*-value = 0.00588), this pathway was not significant in GSEA;

(4) **Endocytosis** (bta04144; *P*-value = 0.00848), significant in GSEA with FDR of 0.00265;

doi: 10.17221/43/2015-CJAS

Table 2. KEGG pathways significant at a maximum 0.1 level after a multiple testing correction pathways also significant in gene set enrichment analysis are marked in bold

KEGG symbol	KEGG description	Count original	Count permuted	95% CI for Odds Ratio	*P*-value
Milk yield					
bta05412	arrhythmogenic right ventricular cardiomyopathy	5	259	2.2–11.9	0.04464
bta05414	dilated cardiomyopathy	5	276	2.0–11.2	0.07862
bta04530	**tight junction**	**8**	**592**	**1.7–6.9**	**0.09542**
Fat yield					
bta03018	RNA degradation	6	48	2.1–10.9	0.04594
Protein yield					
bta04142	**lysosome**	**5**	**64**	**8.8–51.7**	**0.00001**
bta04110	**cell cycle**	**9**	**413**	**3.0–11.4**	**0.00005**
bta00030	pentose phosphate pathway	1	8	7.5–245.4	0.00588
bta04144	endocytosis	**8**	**489**	**2.2–8.8**	**0.00848**
bta00051	fructose and mannose metabolism	1	9	6.8–216.6	0.00884
bta04721	**synaptic vesicle cycle**	**2**	**58**	**2.9–37.5**	**0.07768**

KEGG = Kyoto Encyclopedia of Genes and Genome, CI = confidence intervals

(5) **Fructose and mannose metabolism** (bta00051; *P*-value = 0.00884);
(6) **Synaptic vesicle cycle** (bta04721; *P*-value = 0.07768).

Altogether the pathways cover 546 genes.

The summary of KEGG pathways significant for the analyzed traits is presented in Table 2 and the list of significant GO terms is given in Supplementary Table S1.

DISCUSSION AND CONCLUSION

Methodologically, some concepts utilized in our study have also been recently dealt with by other authors. Pathway analysis involving genes important in dairy cattle production was a part of the study by Cochran et al. (2013). Xiao et al. (2014) used the real data set to generate background data for hypothesis testing, which is conceptually similar to our approach; however, it was based on data resampling, not on permutation. In agreement with our study, but in the context of gene expression analysis, Zhe et al. (2013) and Verbanck et al. (2013) indicate that the incorporation of pathway information, which represents biologically functional groups of genes, improves interpretability of results. Regarding the hypothesis testing, the concept of gene grouping based on biological knowledge, in order to profit from the information contained in correlations between genes, was also considered by Huang and Lin (2013). The permutation approach applied to gene network analysis was also incorporated by Zhou et al. (2013). In that case the authors directly permuted interactions between genes stored in the Human Interactome Resource database.

We are well aware that the obtained gene networks and then the resulting final list of genes can be influenced by the initial set of information used for network construction. The so-called hub genes, which reveal many more edges in networks than the majority of genes, have a strong influence on network physiological interpretation (Zhou et al. 2013). We used the BisoGenet software (Martin et al. 2010), which combines multiple sources of biological information for network construction, and by performing network validation through permutations. The most severe drawback of our study was the poor resolution of the available SNP panel. Out of approximately 20 000 genes identified in cattle only 6000 are marked by SNPs (Michelizzi et al. 2011) on the Illumina BovineSNP50 Genotyping BeadChip, but after data editing in our study we estimated effects of only 4345 genes. Consequently, effects of many genes could not be estimated due to a lack of SNPs located in the vicinity of those genes and thus could be represented only indirectly through network information stored in biological

doi: 10.17221/43/2015-CJAS

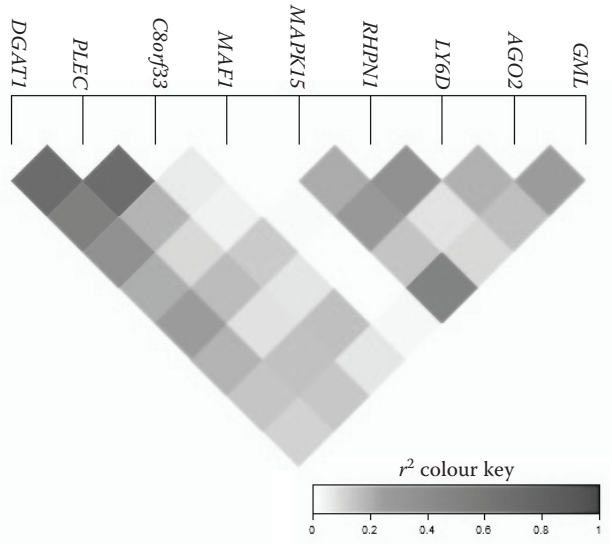

Figure 3. Pairwise r^2 coefficients between single nucleotide polymorphisms (SNPs) representing genes significant for fat yield on BTA14

databases explored by BisoGenet software. Therefore, a denser SNP panel is recommended for future applications. On the other hand, the Bovine SNP50 BeadChip microarray from Illumina is the most common platform used for genotyping dairy cattle and thus allows for the composition of data sets with large numbers of individuals, which is a problem for more expensive and less common high density microarrays. Furthermore, even though LD between SNPs is partially accounted for by a simultaneous estimation of all SNP effects in one model and by incorporating pairwise r^2 into the estimator of gene variance, the scaffold gene selection may still suffer from a nonzero difference between true LD and LD accounted for by the estimation methodology. One possible way to improve the performance of significant SNP/gene selection could be the incorporation of a Bayesian estimation (as discussed by Gianola et al. 2009). On the other hand, the practical application of such models is much more demanding computationally and will make permutation of data impossible, especially in the context of dense SNP arrays as proposed above. We calculated pairwise r^2 coefficients between SNPs which mark genes significant for fat yield on BTA14 (Figure 3). The selected genes generally remain in low LD below 0.50, except of a cluster *DGAT1–PLEC–C8orf33* with r^2 varying between 0.52 and 0.71.

The approach of screening significant pathways for the physiological processes and underlying genes needs to be enhanced by the candidate gene approach. For example, the gene with a very strong effect on milk- and fat yields, *DGAT1*, was identified in the original network for the traits, but the KEGG pathway, in which *DGAT1* participates, was not significant as a whole. Another well-known cluster of major genes, the casein gene cluster, was selected both from original networks and through pathways. Several other genes, members of significant pathways, were already reported by other studies – as indicated in Supplementary Table S2.

Functionally, the associations between individual pathways and milk yield traits can be explained in many different ways. When analyzing milk yield against the Arrhythmogenic right ventricular cardiomyopathy pathway (bta05412), the genes involved include integrin genes. β1-integrins are found on mammary epithelial cells (MEC) (Nemir et al. 2000), which play a key role in the mammary gland development and lactation (Naylor et al. 2005). This pathway also includes calcium channel subunits, which are crucial to calcium metabolism and affect milk secretion. The mammary gland can extract large quantities of Ca^{2+} from plasma during lactation, to ensure sufficient Ca^{2+} concentration in milk, and thus supports the calcification of teeth and bones in the calf (Horst et al. 1997; Shennan and Peaker 2000; Neville 2005). The link between milk production and the Dilated cardiomyopathy pathway (bta05414) might have a similar background: the pathway contains integrin genes and also genes related to calcium metabolism. In the case of the Tight junctions (TJs) pathway (bta04530), it is known that TJs exist as macromolecular complexes composed of several types of membrane proteins, cytoskeletal proteins, and signalling molecules. Many of these components are regulated during mammary gland development and pregnancy cycles (Nguyen and Neville 1998). In secretory glandular tissues, such as the mammary gland, TJs create the variable barrier regulating paracellular movement of molecules through epithelial sheets, thereby maintaining tissue homeostasis. Moreover, TJ appears to be closely associated with milk secretion. An increase in TJ permeability is accompanied by a decrease in the milk secretion rate, and conversely, a decrease in TJ is accompanied by an increase in the milk secretion rate (Nguyen and Neville 1998). Several studies have also indicated that the mammary alveolar TJs are impermeable during lactation

doi: 10.17221/43/2015-CJAS

and therefore allow milk to be secreted from the apical membrane without the leakage of milk components from the lumen into the blood serum via paracellular pathways in goats and cows (Linzell and Peaker 1973; Stelwagen et al. 1998a, b). As far as fat yield is concerned, it is difficult to indicate unequivocally which particular pathway plays a key role. The pathway RNA degradation (bta03018) is highly unspecific, since RNA degradation might regulate the expression of multiple genes involved in various physiological mechanisms related to fat secretion into the milk. RNA synthesis and degradation are key steps in the regulation of gene expression in all living organisms, because RNA degradation is ubiquitous in all cells.

Protein yield has been found to be associated with the Lysosome pathway (bta04142). The lysosomal system is involved in numerous physiological processes such as degradation of endogenous and exogenous macromolecules (proteins, lipids, polysaccharides, and nucleic acids), cytoplasmic formations (mitochondria, peroxisomes, Golgi complex) that have performed their functions, tissue regression (post-lactation mammary gland), hormone secretion regulation (proinsulin to insulin), etc. The lysosomal system is also involved in a number of pathologic processes, such as inflammation, allergic reactions, ischemia, hypoxia, as well as lysosomal diseases. During mammary gland involution the extracellular matrix and the alveolar basement membrane are degraded. The alveoli lose their structural integrity and massive death of MEC is observed. In bovine mammary glands the loss of the MEC population begins after the peak of lactation, when the dynamic equilibrium between mitosis and apoptosis is shifted towards apoptosis. However, the most dynamic induction of MEC apoptosis is associated with the beginning of the dry period (Wilde et al. 1997, 1999). Apoptosis is a physiological mechanism of cell loss that depends on both preexisting proteins and *de novo* protein synthesis. The process of apoptosis is integral to normal mammary gland development. The link between protein yield and the cell cycle pathway (bta04110) may result from the function of cyclin D1, which is involved both in the normal development and malignant transformation of mammary epithelium (Sicinski et al. 1995; Neuman et al. 1997). The pentose phosphate pathway (bta00030) is yet another process related to protein yield. This pathway is highly active in the cytoplasm of the liver, adipose tissue, mammary gland, and the adrenal cortex. The pathway includes transketolases (TK), which play an important role in the system of substrate rearrangement between pentose shunt and glycolysis, permitting the cell to adapt to a variety of metabolic conditions (Sax et al. 1996). Its presence is necessary for the production of NADPH, especially in tissues actively engaged in biosyntheses, such as mammary glands (Kochetov and Sevostyanova 2010). In the case of a link between protein yield and the endocytosis pathway (bta04144) it is known that endocytic mechanisms serve many important cellular functions, including the uptake of extracellular nutrients, regulation of cell-surface receptor expression, maintenance of cell polarity, and antigen presentation (Mukherjee et al. 1997; Clague 1998). This pathway includes genes known to be associated with milk production traits in dairy cows, such as members of the growth factor family (*TGF*, *EGF*, *IGF*) and their receptors. The presence of mRNA for *EGF*, *TGF-α*, and *AR* suggests that these growth factors may be important in mammogenesis in pubertal heifers and during pregnancy, especially during proliferation and differentiation of the MEC. A role of IGF-I in mammary duct development has been postulated based on several observations, as IGF-I can stimulate the proliferation of MEC in organ culture at low concentrations (Richert and Wood 1999). Moreover, *IGF-I*, *IGF-II*, and the *IGF-IR* are expressed within both the epithelial and stromal compartments of the virgin mammary gland (Richert and Wood 1999; Berry et al. 2001). The endocytosis pathway also includes the cytokine and chemokine receptor genes (*CCR, CXCR, TRAF, IL2R*) related to innate and adaptive immunity, as well as the major histocompatibility complex BoLA class I (Behl et al. 2012). The role of the fructose and mannose metabolism pathway (bta00051) in protein yield may be related to the fact that carbohydrates are the most important source of energy. In addition to the *FBP2* identified as a scaffold gene in our study, the link with bta00051 can result from the function of hexokinases (HK2), which are a key control point in glycolysis and are expressed throughout mammary gland development. HK2 has a specific role in the mammary gland as a consequence of the increased energy production associated with lactation (Kaselonis et al. 1999).

Based on 4345 genes, through the validated KEGG pathway selection, we identified sets of genes functionally involved in milk yield (224 genes),

doi: 10.17221/43/2015-CJAS

fat yield (72 genes), and protein yield (546 genes) (Supplementary Table S2).

Just a few of the loci represented genes with high effects already reported by previous studies, such as e.g. 13, 4, and 27 of genes listed in the review by Ogorevc et al. (2009). Moreover, for milk and fat yields none of the scaffold genes were represented in the set and out of the six scaffold genes for protein yield only two (*AP1B1*, *FBP2*) were present in the set. Those findings illustrate the key difference between a traditional GWAS approach, which is focused on genes with high effects on phenotypic variation, and our approach, which aims to discover genes with medium and small effects on trait variation.

The incorporation of data available for other species, for which more functional information on genes has been encoded, through gene networks and the following identification of significant pathways and GO terms is a promising way to diminish the fraction of missing heritability of complex phenotypes measured in dairy cattle. However, even though large phenotyped cohorts of animals are easy to obtain for the Holstein-Friesian breed, the bottleneck is the availability of dense marker maps from the genetic perspective and network validation from the statistical perspective. The former is soon going to be resolved thanks to ongoing whole genome sequencing and imputation projects carried out in cattle, the latter can be done using empirical methods provided sufficient computational power is available. Another critical aspect of gene network based studies using human homologues of cattle genes is that the functional information for humans has been encoded with more focus on human related phenotypes and may lack some of the information related to traits of interest for dairy cattle. Thanks to the current Functional Annotation of Animal Genomes (FAANG) initiative towards functional annotation of animal genomes focusing on livestock, this problem can be overcome in the future.

Acknowledgement. We would like to thank the Genomika Polska (formerly MASinBULL) consortium which provided the data set used in the analysis.

REFERENCES

Behl J.D., Verma N.K., Tyagi N., Mishra P., Behl R., Joshi B.K. (2012): The major histocompatibility complex in bovines: a review. ISRN Veterinary Science, 2012, Article ID 872710.

Berry S.D., McFadden T.B., Pearson R.E., Akers R.M. (2001): A local increase in the mammary IGF-I : IGFBP-3 ratio mediates the mammogenic effects of estrogen and growth hormone. Domestic Animal Endocrinology, 21, 39–53.

Bolormaa S., Pryce J.E., Hayes B.J., Goddard M.E. (2010): Multivariate analysis of a genome-wide association study in dairy cattle. Journal of Dairy Science, 93, 3818–3833.

Braun R., Buetow K. (2011): Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. PLoS Genetics, 7, e1002101.

Clague M.J. (1998): Molecular aspects of the endocytic pathway. Biochemical Journal, 336, 271–282.

Cochran S.D., Cole J.B., Null D.J., Hansen P.J. (2103): Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. BMC Genetics, 14: 49.

Gianola D., de los Campos G., Hill W.G., Manfredi E., Fernando R. (2009): Additive genetic variability and the Bayesian alphabet. Genetics, 183, 347–363.

Henderson C.R. (1984): Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Canada.

Horst R.L., Goff J.P., Reinhardt T.A. (1997): Calcium and vitamin D metabolism during lactation. Journal of Mammary Gland Biology and Neoplasia, 2, 253–263.

Huang Y.T., Lin X. (2013): Gene set analysis using variance component tests. BMC Bioinformatics, 14: 210.

Jairath L., Dekkers J.C.M., Schaeffer L.R., Liu Z., Burnside E.B., Kolstad B. (1998): Genetic evaluation for herd life in Canada. Journal of Dairy Science, 81, 550–562.

Kaselonis G.L., McCabe R.E., Gray S.M. (1999): Expression of hexokinase 1 and hexokinase 2 in mammary tissue of nonlactating and lactating rats: evaluation by RT-PCR. Molecular Genetics and Metabolism, 68, 371–374.

Kochetov G.A., Sevostyanova I.A. (2010): Functional non-equivalence of transketolase active centers. IUBMB Life, 62, 797–802.

Legarra A., Misztal I. (2008): Technical note: computing strategies in genome-wide selection. Journal of Dairy Science, 91, 360–366.

Linzell J.L., Peaker M. (1973): Changes in mammary gland permeability at the onset of lactation in the goat: an effect on tight junctions? The Journal of Physiology, 230, 13–14.

Lush J.L., Holbert J.C., Willham O.S. (1936): Genetic history of the Holstein-Friesian cattle in the United States. Journal of Heredity, 27, 61–72.

Mao X., Cai T., Olyarchuk J.G., Wei L. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics, 21, 3787–3793.

doi: 10.17221/43/2015-CJAS

Martin A., Ochagavia M.E., Rabasa L.C., Miranda J., Fernandez-de-Cossio J., Bringas R. (2010): BisoGenet: a new tool for gene network building, visualization and analysis. BMC Bioinformatics, 10: 91.

Meredith B.K., Kearney F.J., Finlay E.K., Bradley D.G., Fahey A.G., Berry D.P., Lynn D.J. (2012): Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. BMC Genetics, 13: 21.

Michelizzi V.N., Wu X., Dodson M.V., Michal J.J., Zambrano-Varon J., McLean D.J., Jiang Z. (2011): A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to water buffalo. International Journal of Biological Sciences, 7, 18–27.

Mukherjee S., Ghosh R.N., Maxfield F.R. (1997): Endocytosis. Physiological Reviews, 77, 759–803.

Naylor M.J., Li N., Cheung J., Lowe E.T., Lambert E., Marlow R., Wang P., Schatzmann F., Wintermantel T., Schuetz G., Clarke A.R., Mueller U., Hynes N.E., Streuli C.H. (2005): Ablation of beta1 integrin in mammary epithelium reveals a key role for integrin in glandular morphogenesis and differentiation. The Journal of Cell Biology, 171, 717–728.

Nemir M., Bhattacharyya D., Li X., Singh K., Mukherjee A.B., Mukherjee B.B. (2000): Targeted inhibition of osteopontin expression in the mammary gland causes abnormal morphogenesis and lactation deficiency. The Journal of Biological Chemistry, 275, 969–976.

Neuman E., Ladha M.H., Lin N., Upton T.M., Miller S.J., DiRenzo J., Pestell R.G., Hinds P.W., Dowdy S.F., Brown M., Ewen M.E. (1997): Cyclin D1 stimulation of estrogen receptor transcriptional activity independent of cdk4. Molecular and Cellular Biology, 17, 5338–5347.

Neville M.C. (2005): Calcium secretion into milk. Journal of Mammary Gland Biology and Neoplasia, 10, 119–128.

Nguyen D.A., Neville M.C. (1998): Tight junction regulation in the mammary gland. Journal of Mammary Gland Biology and Neoplasia, 3, 233–246.

Ogorevc J., Kunej T., Razpet A., Dovc P. (2009): Database of cattle candidate genes and genetic markers for milk production and mastitis. Animal Genetics, 40, 832–851.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., Sham P.C. (2007): PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81, 559–575.

Qanbari S., Pimentel E.C., Tetens J., Thaller G., Lichtner P., Sharifi A.R., Simianer H. (2010): A genome-wide scan for signatures of recent selection in Holstein cattle. Animal Genetics, 41, 377–389.

Richert M.M., Wood T.L. (1999): The insulin-like growth factors and IGF type I receptor during postnatal growth of the murine mammary gland: sites of mRNA expression and potential functions. Endocrinology, 140, 454–461.

Sax C.M., Salamon C., Kays W.T., Guo J., Yu F.X., Cuthbertson R.A., Piatigorsky J. (1996): Transketolase is a major protein in the mouse cornea. The Journal of Biological Chemistry, 271, 33568–33574.

Shannon P., Markiel P., Ozier O., Baliga N.S., Wang J.W., Ramage D., Amin N., Schwikowski B., Ideker T. (2003): Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13, 2498–2504.

Shennan D.B., Peaker M. (2000): Transport of milk constituents by the mammary gland. Physiological Reviews, 80, 925–951.

Sicinski P., Donaher J.L., Parker S.B., Li T., Fazeli A., Gardner H., Haslam S.Z., Bronson R.T., Elledge S.J., Weinberg R.A. (1995): Cyclin D1 provides a link between development and oncogenesis in the retina and breast. Cell, 82, 621–630.

Stelwagen K., McLaren R.D., Turner S.A., McFadden H.A., Prosser C.G. (1998a): No evidence for basolateral secretion of milk protein in the mammary gland of lactating goats. Journal of Dairy Science, 81, 434–437.

Stelwagen K., van Espen D.C., Verkerk G.A., McFadden H.A., Farr V.C. (1998b): Elevated plasma cortisol reduces permeability of mammary tight junctions in the lactating bovine mammary epithelium. Journal of Endocrinology, 159, 173–178.

Strabel T., Jamrozik J. (2006): Genetic analysis of milk production traits of Polish Black and White cattle using large-scale random regression test-day models. Journal of Dairy Science, 89, 3152–3163.

VanRaden P.M. (2008): Efficient methods to compute genomic predictions. Journal of Dairy Science, 91, 4414–4423.

Verbanck M., Le S., Pages J. (2013): A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. BMC Bioinformatics, 14: 42.

Wilde C.J., Addey C.V.P., Fering D. (1997): Programmed cell death in bovine mammary tissue during lactation and involution. Experimental Physiology, 82, 943–953.

Wilde C.J., Knight C.H., Flint D.J. (1999): Control of milk secretion and apoptosis during mammary gland involution. Journal of Mammary Gland Biology and Neoplasia, 4, 129–136.

Xiao Y., Hsiao T.H., Suresh U., Chen H.I.H., Wu X., Wolf S.E., Chen Y. (2014): A novel significance score for gene selection and ranking. Bioinformatics, 30, 801–807.

Zhe S., Naqvi S.A.Z., Yang Y., Qi Y. (2013): Joint network and node selection for pathway-based genomic data analysis. Bioinformatics, 29, 1987–1996.

doi: 10.17221/43/2015-CJAS

Zhou X., Chen P., Wei Q., Shen X., Chen X. (2013): Human interactome resource and gene set linkage analysis for the functional interpretation of biologically meaningful gene sets. Bioinformatics, 29, 2024–2031.

Received: 2015–06–09
Accepted after corrections: 2016–06–20

Corresponding Author

prof. dr. Joanna Szyda, Wroclaw University of Environmental and Life Sciences, Department of Genetics, Biostatistics Group, Kożuchowska 7, 51-631 Wrocław, Poland
Phone: +480 713 205 957, e-mail: joanna.szyda@upwr.edu.pl