ABSTRACT: For traits undergoing a complex mode of inheritance, involving genes of large, moderate and small effects, gene detection based solely on P-values is of limited utility. Such GWAS approach has been broadly widespread, not only in livestock, but most of all in humans, plants and laboratory species. However, it has been demonstrated, that this methodology has serious limitations in terms of power to detect variants of moderate effects as well as in terms of repeatability of results across data sets. Here alternative approaches towards statistical modelling of complex traits are proposed, which utilize functional information on traits to allow for the discovery of genes not only with high, but also moderate effects on functional information on traits to allow for the discovery of genes not only with high, but also moderate effects on functional key processes for phenotype determination. Dairy cattle data sets originating from high-throughput technologies – SNP microarray and whole genome sequence are used as examples.

Keywords: dairy cattle; gene network; NGS

Introduction

Complex traits are determined by, usually a few, genes with large effects - 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 perform a Genome-Wide Association Study (GWAS). When GWAS is applied to the data sets currently available in livestock, which are very informative thanks to the large number of individuals with known (pseudo)phenotypes, deep pedigree and genotypes of many thousands of Single Nucleotide Polymorphisms (SNP) or even whole genome DNA sequence, it is usually relatively easy to identify major genes. However, it is still very difficult to localise genes with small effects. Therefore, recently, in genetic analysis of complex traits focus has been moved from single genes identified via GWAS to genes identified using a functional analysis (Visscher et al. (2012); Evangelou et al. (2013)). For this purpose some bioinformatics tools can be applied such as the so-called “gene network approach”, which profits from the information stored in publicly available databases, or methods based on Gene Set Enrichment Analysis (GSEA).

Data sets

SNP data. A total of 2 601 bulls from the Polish Holstein Friesian dairy cattle breed were used. This core data was additionally enhanced by the information on pedigree relationship and on conventional breeding values of 10 355 individuals. Routinely recorded traits, undergoing a complex mode of inheritance were considered. Since all the phenotypes are measured in cows, our analyses utilise pseudo-phenotypes expressed as deregressed conventional breeding values of the bulls.

The animals were genotyped within the frame of Genomika Polska consortium using both Illumina BovineSNP50 Genotyping BeadChips v.1 and v.2, which consist of 54 001 and 54 609 SNPs, respectively. After genotype pre-processing comprising removing SNPs with minor allele frequency less than 0.01 and call rate under 90%, 46 267 SNPs were selected for further analysis.

Whole genome sequence data. DNA sequences of 32 cows from the Polish Holstein Friesian dairy cattle breed, selected out of the data base, based on ancestry and mastitis prevalence, were obtained by Illumina HiSeq Next Generation Sequencing. The animals formed paternal half-sib groups, comprising a case group – i.e. cows with high incidence of clinical mastitis, and a control group – i.e. animals with no mastitis incidence throughout their production life. The case group consists of 16 individuals born between 2002 and 2005 with the total number of clinical mastitis incidences varying between 6 and 14 throughout the entire production life. The corresponding control group of healthy half-sibs also consists of 16 cows born between 2002 and 2005 with no clinical mastitis recorded throughout their production life.

Raw DNA sequence reads for each cow were edited for quality and then aligned to the UMD3.1 reference genome using BWA (Li and Durbin (2010)). In addition, polymorphic variants (SNP as well as insertions/deletions) were detected using several software for a comparison including Freebayes (Garrison and Marth (2012)).

GWAS results comparison

Motivation. A whole variety of statistical models has been applied to perform GWAS based on SNP chips. These vary from models where each SNP is considered individually to models with effects of all the available SNPs fitted simultaneously. The major drawbacks of GWAS, which are especially severe when analysing traits with complex mode of inheritance, are (i) difficult selection of polymorphisms which are supposed to be associated with the trait among intercorrelated SNPs and (ii) poor reproducibility of results across methods. Therefore, we compared results of different GWAS approaches for complex traits, all applied to the same data set, focussing on...
quantifying the differences between models in defining significant polymorphisms.

**GWAS approaches.** The data set used for GWAS were genotypes of Polish Holstein-Friesian bulls with pseudo-phenotypes for somatic cell score (SCS), milk yield (MY), fat yield (FY), and non-return rate for heifers (NRH).

The SNP effects were estimated by using four different models. Single SNP models comprising (M1) $y = \mu + X\beta + \epsilon$ and (M2) $y = \mu + X\beta + Z\alpha + \epsilon$. Variance and covariance components of those models were estimated using the ASReml3 software (Gilmour et al. (2009)). In the above models $y$ represents a vector of deregressed breeding values for MY, FY, SCS or NRH; $\beta$ denotes a vector of fixed SNP effects, $\alpha \sim N(0, A\sigma^2_\alpha)$ is a vector of random polygenic effect of a bull with $A$ being a pedigree based relationship matrix and $\sigma^2_\alpha$ representing the estimate of total additive genetic variance, $\epsilon \sim N(0, D\sigma^2_\epsilon)$ is a vector of residuals where $D$ is a diagonal matrix weighted by the effective daughter contribution for each bull and $\sigma^2_\epsilon$ denotes a residual variance, $\mu$ denotes an overall mean. The CAR score regression (M3) proposed by Zuber and Strimmer (2011) and defined as $\omega = P^{1/2}P_{y\beta}$, where $P$ denotes the empirical correlation matrix among SNP genotypes and $P_{y\beta}$ is the marginal correlation vector between deregressed breeding values and SNPs, was selected to represent a nonparametric approach. The CAR criterion was evaluated using the R package CARE. The genomic selection model (M4) $y = \mu + Zg + \epsilon$ is equivalent to the so called SNP-BLUP model. Here $g \sim N\left(0, \frac{\sigma^2_\epsilon}{N_{SNP}}\right)$ is a vector of random additive SNP effects with $N_{SNP}=46267$ being the number of SNP considered. In case of M1, M2, and M4 the Wald test was used to obtain the nominal type I error corresponding to a standard normal distribution. For M1 and M2, SNPs were selected as significant when P values associated with their estimates, subjected to Bonferroni correction for the number of SNP tested, exceeded the 0.001 threshold. For M3 the null distribution of the empirical CAR scores was defined as

$$Beta\left(\hat{\omega}_i^2, \frac{\hat{\omega}_i^2}{2}, \frac{\hat{\omega}_i^2(N-2)}{2}\right),$$

where $\hat{\omega}_i$ is the CAR score for i-th SNP and the threshold was $P \leq 0.001$. For M4, in order to account for the amount of empirical genetic variation explained, out of the potentially available variation imposed by the model, the P value threshold was calculated separately for each trait as a product: $0.001 \times r$, where $r$ denotes the ratio of standard deviations of the empirical to the theoretical distributions and was equal to 0.031, 0.043, 0.093, and 1.224 for SCS, FY, NRH and MY respectively.

**Results.** The numbers of SNPs selected as significant by different models are presented in Table 1. For MY, FY, and, SCS the simplest model M1 always selected a very large number of SNPs ranging between 2 242 (SCS) and 3 398 (MY), widely exceeding the number of SNPs selected by M2-M4. Although models 2 and 3 markedly differ in modelling and hypothesis testing assumptions, they both select a very similar and low number of SNPs for each trait. Except for FY the genomic selection model M4 selected larger number of SNPs than models M2 and M3. For MY and FY respectively 27 and 34 SNPs were common for all the three models, but no common polymorphisms were identified for SCS and NRH. More mutual SNPs existed between M4 and M2 than between M4 and M3.

<table>
<thead>
<tr>
<th>Trait</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>fat yield</td>
<td>2435</td>
<td>72</td>
<td>49</td>
<td>65</td>
</tr>
<tr>
<td>milk yield</td>
<td>3398</td>
<td>66</td>
<td>86</td>
<td>176</td>
</tr>
<tr>
<td>somatic cell score</td>
<td>2242</td>
<td>4</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>non-return rate for heifers</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>36</td>
</tr>
</tbody>
</table>

**Figure 1:** The percentage of significant SNPs common between models

**Discussion.** Here we demonstrated that even for the same data set, GWAS results heavily depend on the methodology applied. For traits with moderate heritability and known major genes, in this study represented by MY and FY, there is a group of SNPs which is identified as significant regardless of the model used. However, for SCS and NRH, where there are no known genes of strong effects, no common polymorphisms were identified across models. Such result indicates that each method, apart from false positives, presumably selects different true positive associations depending on the inheritance information incorporation into statistical modelling.
Selecting physiological processes underlying quantitative traits

Motivation. Following Eric Lander’s statement: “Biology emerges from pathways, not from single genes” the major motivation for further research was to identify genes with major and moderate effects, but most of all to identify functional pathways responsible for the variation of complex traits of economic importance in dairy cattle. However, as demonstrated above, in deciphering the genetic background underlying complex traits we cannot solely rely on GWAS.

Material, pathway construction, validation and testing of functional features. The same data set of Polish Holstein-Friesian bulls genotyped with the Illumina BovineSNP50 BeadChip with traits comprising milk-, fat-, and protein yields (PY) were considered in this study.

The first step of the analysis comprised estimation of effects of SNPs using M4 as defined above.

In the next step SNP effect estimates were enhanced by the information on their genomic location and linkage disequilibrium to form gene effect estimates. For each SNP its genomic location was retrieved from the Ensembl data base and the linkage disequilibrium was calculated using PLINK (Purcell et al. (2007)). Gene effects were estimated using: 
\[ t = \frac{\hat{g}_i}{\sigma^2_i}, \]
where \( \hat{g}_i \) denotes the estimate of i-th SNP effect, \( N_i \) is the number of SNPs located within the gene or maximum 1Kbp distance from gene borders, and \( \sigma^2_i \) represents the variance of a gene effect expressed by: 
\[ \sigma^2_i = N_i \sigma^2_q + 2 \sum_{j=1}^{N_i} \sum_{j \neq i} r_{ij}^2 \sigma^2_q, \]
with \( \sigma^2_q = \frac{\sigma^2_q}{46267} \)
representing the variance of a SNP effect identical for all SNPs. Asymptotically, \( t \) follows a standard normal distribution which was used to select significant genes. Since the major issue of the study was focused on genes with medium and small effects we decided that in a trade-off between a type I and a type II errors the latter is much more important at this stage. Consequently, a significance threshold of 0.20 was used to select genes for further analysis.

Separately for each trait, significant genes were used as a scaffold to construct a gene network. A Bisogenet plugin (Martin et al. (2010)) for the Cytoscape software (Shannon et al. (2003)) was used as a tool for building a gene network based on the information from publicly accessible data bases. Networks were constructed using the significant genes as primary input nodes and enhancing the network by additional nodes representing genes for which known interactions with the input nodes were identified in the data bases. 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 used. Functional information represented by the gene network was categorised by GO terms and KEGG pathways. In order to assess how reliable a particular network was a null hypothesis distribution of the network was constructed empirically based on permutations. For this purpose, for each trait separately, the following steps were repeated 100 times in order to approximate the null distribution:
1) permutation of deregressed breeding values (vector \( \gamma \));
2) SNP effect estimation based on M4;
3) gene effect estimation and testing using the statistic \( t \);
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 identified in the original network was based on the Odds Ratio statistic: 
\[ OR = \frac{C_o + 0.5}{C_p + 0.5}, \]
where \( C_o \) represents the number of times a given GO term or a KEGG pathway was identified within the original (\( N_o \)) and permuted (\( N_p \)) networks and \( N_o \) represents the total number of different GO terms or KEGG pathways identified for the original (\( N_o \)) and permuted (\( N_p \)) networks. The underlying hypotheses are defined in terms of the probability of a given functional feature occurring in the original (\( P_o \)) and permuted (\( P_p \)) data as \( H_0 : P_o = P_p \) and \( H_1 : P_o \neq P_p \). The natural logarithm 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 Bonferroni correction.

Results. Effects of 4 345 genes were estimated. Due to a relatively low SNP density of the 50K chip, 87% of these estimates were based on a single SNP. The remaining estimates were based on 2 to 6 SNPs. Seven and nine genes were identified as significant on MY and FY 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 DGAT1 gene. All genes significant for MY were also significant for FY. For FY six significant genes were observed, each located on a different chromosome, with the highest effect of AP1B1 with 0.10 kg of protein per lactation.

MY was described by the network of 98 genes representing 1 115 various GO terms and 130 various KEGG pathways. Assuming the maximum 10% type I error rate corrected for multiple testing, 50 different GO terms and three KEGG pathways 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 less than 0.00001 comprised: condensin complex (GO:0000796), intercellular canaliculus (GO:0046581), and mitotic chromosome condensation (GO:0007076). Significant KEGG pathways comprised: (i) Arrhythmogenic right ventricular cardiomyopathy (bta05412, \( P=0.04464 \)) with 5 and 259 hits in the original and permuted gene sets respectively, (ii) Dilated
that mouse phenotypes - clonic seizures and abnormal brain wave pattern are genomically similar traits to human epilepsy. Our study was focussed on a within-species phenotype comparison by quantifying functional similarities between traits routinely recorded in dairy cattle using the functional information gathered through gene networks.

**Material and methods.** Again, the same data set of Polish Holstein-Friesian bulls, their genotypes and pseudogenotypes comprising deregressed proofs for MY, FY, PY, SCS, NRH, non-return rate for cows (NRC), stature (STA) and body size score (SIZ) were used.

For each trait significant genes were selected using the t test statistic. Further on, gene networks were constructed for each trait, as described in the previous section.

In order to quantify the functional similarities between traits, the sets of genes composing each network and the sets of GO terms associated with significant genes were summarised in a design matrix, which was then used for calculating similarity scores between traits. Two measures were used to quantify similarity between pairs of traits by comparing the sets of genes underlying networks for each trait or by comparing sets of GO terms related to genes which effects were estimated as significant in GWAS. The cosine similarity between traits i and j is given by: $\cos = \frac{N_{ij}}{N_i + N_j}$, where $N_{ij}$ represents the number of times a feature (i.e. gene or GO term) was significant for both traits, $N_i(N_j)$ is the number of times a feature was significant for trait i(j). Spatially, the metric represents an angle between two vectors of features. The second of applied measures was the Jaccard similarity coefficient, defined as the quotient between the intersection and the union of the pairwise compared variables: $Jac = \frac{N_{ij}}{N_i + N_j + N_{ij}}$. In addition Pearson correlation coefficients were calculated between SNP or gene effect estimates for each pair of traits.

**Results.** The obtained networks consisted of 98 genes for MY, 97 genes for FY, 44 genes for PY, as many as 1 255 genes for SCS, and the smallest network was observed for STA with 26 genes. No significant genes were obtained for NRH, NRJ and SIZ and consequently no networks were generated for those traits. For each of the significant gene the corresponding GO terms were retrieved.

Similarities between traits based on gene and GO term sets underlying the gene networks were calculated using two different measures – the cosine and the Jaccard coefficients, were very consistent. While comparing sets of genes constituting a gene network for each trait the highest similarity of 0.455 was observed between MY and FY and no similarity, expressed by metrics equal to 0, was observed between PY and STA. Considering sets of GO terms the highest similarity score of 0.622 was also calculated for MY and FY, while the lowest score of 0.049 was attributed to PY and FY (Figure 2). Pearson correlation coefficients calculated between 4 345 estimates of gene effects were
highest for PY and MY (0.762) and lowest (-0.011) for PY and SCS, whereas correlation between 46,267 SNP estimates varied between 0.779 for PY-MY and 0.025 for PY-STA (Figure 3).

**Figure 2: Similarity measures between traits calculated based on gene and GO term sets**

**Figure 3: Pearson correlation coefficients between traits calculated based on gene and SNP effect estimates**

**Discussion.** It is interesting that for many trait-pair combinations polygenic based information expressed by Pearson correlation coefficients is not consistent with the functional similarity measures. Especially, all of the trait pair comparisons involving PY indicated high polygenic similarity, but low functional similarity. The discrepancy indicates that an infinitesimal model reflects an averaged correlation, which is due to selected genes of high effects on the traits. However, it fails to reveal the functional background underlying the traits, which is due to a cumulative composition of many genes involved in metabolic pathways – which appear to be different between PY and FY/MY.

**Using whole a genome sequence**

**Motivation.** A serious drawback of the previously described approach of complex trait dissection is a very poor resolution of the underlying genomic structure covered by the 50K SNP chip. Thus, a natural extension of the analysis is the utilisation of a whole genome sequence of individuals, which captures all existing variation. Unfortunately, up to now, very large data sets sequenced with enough coverage allowing for reliable polymorphism detection are still quite expensive. So in this the last section we present a preliminary application of the analysis to a qualitative trait – clinical mastitis.

**Material and methods.** The case and control groups of Polish Holstein-Friesian cows with whole genome sequence were compared. For each of the identified SNPs the Odds Ratio statistic: $OR = \frac{N_{A, \text{case}}}{N_{B, \text{case}}} \div \frac{N_{A, \text{control}}}{N_{B, \text{control}}}$, where, $N_x$ represents the number SNP alleles coded A or B in case or control group respectively. The underlying hypotheses and testing procedure was the same as described above, with an exception that no Bonferroni correction of P values was performed. Processing of the raw data were performed at the Poznan Supercomputing – Networking Center.

**Results.** The average number of SNP detected per cow was 5,505,166±663,223 which varied between 2,063,811 and 6,117,976 (Figure 4) corresponding to 0.21% of the whole genome sequence. Across all cows 15,188,516 SNPs were identified. Most of the SNPs were bi-allelic, but 0.29% of all SNPs were tri- or tetra-allelic. The largest number of multiallelic SNP was observed on BTA14, which contained 5,875 tri- and 47 tetra-allelic SNPs. The least number of SNPs with more than 2 alleles was detected on BTA27 with 657 and 2 tri- and tetra-allelic SNPs respectively.

**Figure 4: The number of detected SNPs and average coverage per cow. Black (grey) bars represent SNP number in the control (case) group, striped bars represent coverage.**

The accuracy of the aligned sequence expressed as an average coverage of the whole genome by sequence reads markedly varied among individuals, between the lowest coverage of only 5 and the highest of 17 (Figure 4, 5). Unfortunately, a high correlation of 0.72 was calculated between coverage and the number of SNPs detected.

Using a nominal P value of 0.001 as a threshold 359 SNPs appeared to be significantly associated with mastitis. The SNPs were distributed across 11 different chromosomes, but especially noteworthy is that a very large number of 313 SNPs was located within the telomeric region of BTA20 (Figure 6).
Discussion. Results of our study clearly point out at telomeric regions of BTA20 as important in risk of clinical mastitis. Telomeres are distinctive structures of a repetitive DNA sequence and associated proteins, which allow cells to distinguish chromosome ends from DNA double-strand breaks and suppress DNA damages. Maintaining telomeric intactness and securing cellular longevity depends on three molecular processes: telomeric capping by shelterin proteins, sequence elongation by telomerase and repair of strand lesions by DNA repair pathways (Hohensinner et al. (2011)). Structural changes of the telomeric ends result from aberrations in the machinery that maintains the 3-D configuration of telomeres. Alterations of telomeric sequences and chromatin assembly might explain the chromosome instability (Georgin-Lavialle et al. (2010)). Most tissues and organs including peripheral blood cells, lymphocytes and epithelial cells show significant telomere shortening in result of intense proliferation. Particularly the immune system is uniquely sensitive to the effects of telomere shortening because its cells adaptively respond to immune challenges with massive proliferation and contraction, and its competence depends strictly on cell renewal (Andrews et al. (2010)). Cellular proliferation is a key component of an effective immune response thus telomere dynamics are critical in preserving immune function. Therefore, telomeric sequences in leukocytes and in udder epithelial cells are likely to play a significant role in resistance to mastitis.

Conclusions.
Presented results indicate that often a strict reliance on P-values may lead to loss of important information. A more “fuzzy” approach towards interpreting significance is recommended. For example in mastitis data, the nominal P values used to select “significant” polymorphisms were highly non-conservative and many of the selected polymorphisms represent false positives, but looking at a clearly non-random distribution of selected SNPs along a genome, it became evident that telomeric regions of BTA20 play an important role in determining mastitis resistance. Moreover, although the pathway approach incorporated in the analysis of bull data needs some fine tuning, still we argue that it is a promising way to identify the physiological processes and underlying genes responsible for traits routinely recorded and evaluated in dairy cattle. This would not be possible by relying solely on SNPs selected as significant via GWAS.

Literature Cited