Understanding the contribution of rare variants to the genetic variation of complex traits, based on bovine genome sequence

Objective

In dairy cattle the approximate number of causal mutations underlying quantitative traits is estimated to 4580. The aim of our study is to identify some of those mutations using Next Generation Sequencing (NGS) technology enhanced with bioinformatics tools. In particular we are going to combine the following: (i) obtain DNA sequence for approximately 1000 bulls in 20-30 preselected genomic regions, based on at least 10 times sequencing coverage, (ii) perform an *in silico* sequence analysis including SNP detection, and (iii) to test which of the novel SNPs are causal mutations for the 8 quantitative traits routinely measured in dairy cattle.

Since the launch of the first commercial NGS platform in 2004 there has been growing interest in the application of DNA sequence information in genetics. A very broad study related to various research possibilities provided by NGS data for the analysis of human genome was described by the 1 000 Genomes Project Consortium (2010). In a context of the analysis of livestock Amaral et al. (2009) described a procedure for SNP selection based on NGS data, using a porcine genome as an example. Moreover, during the last 2 years the application of information from NGS for the analysis of complex traits came into focus, not only in humans (Asimit and Zeggini 2010), but also in animal genetics (Meuwissen and Goddard 2010). For example in a simulation study of Meuwissen and Goddard (2010) incorporation of DNA sequence information, through the knowledge of SNP variation underlying causal mutations, was considered. As pointed out by Goddard at the 2011 Conference on Applied Genomics to Sustainable Livestock Breeding (2-5.05 Melbourne, Australia) using DNA sequence and the incorporation of biological information are undoubtedly the next steps in livestock genomic studies and genetic evaluation.

In view of recent, often yet unpublished, results presented during the International Conference on Quantitative Genetics (Edinburgh 2012) DNA sequence data of hundreds of individuals are crutial for understanding the genetic architecture underlying complex traits. Although it has been commonly agreed that Most of the variation of quantitative traits (approximately 90% of the total heritability) is due to common variants, the rest i.e. “missing heritability” is due to rare variants. Such rare polymorphisms have very low variation in populations and thus cannot be discovered through commercial SNP genotyping panels, but only through dedicated sequencing of a very large number of individuals. Rare variants are very important for the determination of the variation of complex traits, especially under strong selection, which has been imposed on dairy cattle populations for many generations. According to Eyre-Walker (2010): “The genetic variance, in the trait is contributed by mutations at low frequency in the population, unless the mean strength of selection of mutations that affect the trait is very small; each rare mutation tends to contribute more to the variance than each common mutation; the fact that most of the variance in fitness contributed by new nonsynonymous mutations is caused by mutations at very low frequency. These results may explain why most genome-wide association studies have failed to find associations that explain much of the variance.”

Currently, many sequencing project are being initiated in dairy cattle sector. The availability of sequence data will allow our lab to keep up with the up to date research not only for this particular project, but also for other scientific hypotheses and analyses which can be applied to this data in the near future.

Discipline – Contribution

The number of genes underlying milk production traits in cattle

In dairy cattle the approximate number of causal mutations underlying quantitative traits can be estimated as $2N_eL$ (Daetwyler et al. 2008), where N_e is the effective population size and L is the total genome length in Morgans. Assuming N_e=100 (Sorensen et al. 2005) and L=22.9 (Kappes et al. 1997) we end up with 4 580 mutations. Based on the Expressed Sequence Tag analysis Lemay et al. (2009) identified an even larger number of as many as 6 469 genes expressed in bovine mammary glands at different stages of development and estimated that a single QTL covers between 13.9 and 17.1 genes expressed during lactation. Our project aims to the identification of causal mutations with large and intermediate effects on the variation of quantitative traits routinely recorded in dairy cattle. Among all the genes genes a few loci with very strong effects, like e.g. DGAT1 (Grisard et al. 2001) have already been well characterised, but the bulk of genes with intermediate effects still remains unknown.

SNP selection based on NGS data

The causal mutations are to be represented by SNPs identified via DNA sequence comparison between bulls. Amaral et al. (2009) emphasised that careful preselection of SNPs identified based on the Next Generation Sequencing data is very important in order to remove false positive signals resulting from sequencing errors. In particular, in their study out of the total number of 1 193 814 identified SNPs only 17 489 "survived" various filtering criteria applied. As indicated by the 1 000 Genomes Project Consortium (2010) "for a given total amount of sequencing" the highest number of SNPs is identified when many individuals are sequenced at low coverage, instead of sequencing a smaller number of individuals at great depth. The power analysis results provided in this study show that 320 individuals sequenced at low coverage of 3.56 on average are enough to identify 98.50% of functional SNPs present in the genome.

Prediction of genomic breeding values with and without causal mutations included

In our project the identified SNPs are then going to be incorporated into the genomic evaluation model. Even though a very high density SNP panel of 33 000 SNPs per 1M was assumed by Meuwissen and Goddard (2010) a failure to include causal mutations into the genomic breeding value prediction model resulted in lower accuracy. Although, in the simulations the overall accuracy loss was not substantial and varied between 2% for the heritability of 0.50 and 8% for the heritability of 0.25, it has to be borne in mind that a high density SNP panel assumed in this study was much more informative than high density SNP panels which are now commercially available for cattle (like the Illumina 777K bead chip) and that the heritabilities for most of the traits which are routinely measured in dairy cattle are lower than 0.30 (Interbull). On the other hand decay in accuracy with the increasing generation distance between training and test data sets was much smaller when causal mutations were included into the genomic prediction prediction model. This is a very important result for our study - although our main goal is not a prediction of breeding values, but the search for causal mutations underlying the variation of quantitative traits, we can use the accuracy of predicted genomic values as a sensitive test for the hypothesis whether/which of the considered mutations (SNPs) located within coding regions are the causal ones. Also and association analysis performed by the 1 000 Genomes Project Consortium (2010) showed that using all (previously known and novel) SNPs resulted in a discovery of 20% to 50% more significant eQTL, than a comparable study based on 1 Mb commercial SNP platform.

Significance

At the moment, first results regarding the application of the sequencing technology to dairy cattle are being published. However most of the studies concentrate on a relatively low number of animals, detection of common novel SNPs and their impact of the quality of prediction of genomic breeding values (see e.g. recent contributions on the Annual Conference of the European Association of Animal Production, Bratislava 2012). On the other hand we are aware of two sequencing large projects in human genetics dedicated to the detection of rare variants: (i) sequences of 202 genes for 14002 individuals (Nelson 2012), (ii) whole genome sequences of 1795 for individuals (Jonsson 2012). Because of limited funding our project is going to focus on sequence of predefined genomic regions, but on a very large number of bulls and thus will be the first project in cattle genomics aiming to the detection of rare SNP variants and to the estimation of their effects. Moreover, in comparison to data available for human genetics, our study can profit from very good records on phenotypes and pedigree structure, which are available for dairy cattle populations, but are rarely available in human studies.

Work plan

The experimental design of our study is based on results of projects previously conducted by the group, which cover:

1. Establishing the DNA data base, which currently contains 2 956 Polish Holstein-Friesian bulls → this provides a source of high quality DNA for the current project.

2. Estimates of SNP effects for polymorphisms genotyped using 50K Illumina bead chip platform for 8[1] traits routinely recorded for Polish dairy cattle, based on a data set of 2 461 bulls, calculated within the frame of the MASinBULL project → this enables selection of SNPs with the highest effects on those quantitative traits.

[1] milk yield, protein yield, fat yield, somatic cell score, non return rate for heifers, non return rate for cows, interval between calving and 1st insemination, days open

3. Estimates of genomic breeding values as well as their accuracies for the 8 traits, based on a training data set consisting of 2 461 bulls and SNP genotypes from the 50K Illumina bead chip, calculated within the frame of the MASinBULL project → this provides reference accuracies, which can then be compared with the accuracies obtained using SNPs candidates for causal mutations.

4. Annotations of the most significant SNPs from the 50K Illumina bead chip on the reference cattle genome sequence, obtained within the frame of the grant N N311 524940 financed by the Ministry of Science and Education → this gives information on genes exhibiting the highest probability of harbouring causal mutations responsible for the genetic variation of the considered traits, i.e. genes which are going to be subjected to sequencing and SNP analysis within this project.

Note that our material is going to cover less of the DNA sequence than most of the studies based on the Next Generation Sequencing (NGS) technology, but many more individuals will be sequenced. Since we are interested in discovering rare variants, a large number of individual genomes is essential for determining sequence variation. Our experimental design matches, albeit on a smaller scale, an "exon data set" sequenced within the 1 000 Genomes Project (1 000 Genomes Project Consortium 2010). Out of the three experimental designs generated and analysed in the project the "exon data", consisting of 697 individuals sequenced at 8 140 exons (equivalent to 845 Gb), revealed the highest percentage of novel SNPs.

The particular steps of the analyses within the frame of the current project comprise:

1. Selection of genomic regions and animals to be sequenced.

   Genomic regions will be selected based on the results of previous studies (MASinBULL and N N311 524940) considering: (i) SNP effects' estimates - i.e. high effects on the 8 traits analysed, and (ii) their genomic location - i.e. annotation to known coding or regulatory regions. Bulls will be selected based on: (i) phenotypic information - i.e. bulls must have conventional breeding values available for each of the 8 traits, breeding values should be accurate - i.e. estimated by the high number of effective daughter contributions and representative for the population - i.e. cover high, middle and low values, (ii) pedigree information - i.e. choosing bulls with the lowest relationship possible in order to maximise DNA diversity, (iii) - genotypic information - i.e. possibly highest heterozygosity in selected genomic regions expressed by SNP genotypes from the 50K Illumina bead chip.

2. Preparation of DNA and sequencing.

3. The analysis of DNA sequence data, comprising:
   a. editing sequence files,
   b. sequence assembly to the reference genome,
   c. estimation of the most probable genotype for each sequenced nucleotide and each individual,

d. estimation of MAF for each nucleotide (possibly with the incorporation of information on genotype probability uncertainty) and detection of novel SNPs,
e. classification of identified SNPs into known (i.e. represented in SNP data bases and/or on the commercial SNP genotyping arrays) and novel polymorphisms,
f. estimation of pairwise linkage disequilibrium between novel and common SNPs (SNPs from the Illumina 50K array) located in the physical neighbourhood of the sequenced genomic regions,
g. the analysis of DNA sequence diversity in the proximity and within coding regions.

As it has been demonstrated by Kim et al. (2011), when estimating MAF and using SNP genotype information in association analysis, it is very important to take uncertainty in genotype calling into account. Therefore, in our project we wish to consider uncertainty in SNP selection (step 4d). Furthermore, while comparing the most probable genotype obtained based on NGS technology generated with 8x sequencing depth with the corresponding true genotype available from other studies, Kim et al. (2011) observed that "in many individuals the highly supported genotype based on [NGS] differs from the [true] genotype". In step 4e of our project, after uncertainty in genotype calling has been taken into account, we plan to estimate the false discovery rate by comparing SNP genotyped derived by two independent technologies. The genotype uncertainty can also be incorporated into the genomic breeding value estimation model into a SNP design matrix. Instead of the custom, ambiguous (-1), 0, 1 coding, function of actual probabilities can be used. It is important however, that in order to maintain correspondence to the custom allele coding the original genotype probabilities need to be transformed from the [0, 1] probability values to the [(-1), 1] scale. This is a novel approach, which to our knowledge has not been used so far.

The design of our project also allows for the analysis of DNA sequence diversity. It makes it possible to investigate whether diversity calculated for the cattle genome is similar to what was obtained in humans. 1 000 Genomes Project Consortium (2010) observed that the diversity decreases with the decreasing physical distance from the centre of a gene. Such comparison is especially interesting since both populations markedly differ in terms of population history and selection intensity. It can be further hypothesised that those cattle genes which are important for selection are much less diverse than genes which are selection neutral, so that for genes responsible for economically important traits the loss of sequence diversity may stretch across longer distances.

4. Estimating effects of all (rare and common) variants in a genomewise association study for the 8 traits and all sequenced individuals using statistical tests especialy desidned for rare variants.

5. Estimating effects of a “standard” set of SNPs currently used in genomic evaluation programmes in Poland and of an “extended” set of SNP including novel variants using a data subset of sequenced bulls as a training data set.
6. Predicting genomic breeding values for the remainder of sequenced bulls (test data set) based on a “standard” and on an “extended” sets of SNPs.

Methodology

Experimental design

Since sequencing of a whole genome with a reasonable sequencing depth is still very expensive, e.g. 46 000 USD for humans and even more for cattle and working with a shallow sequencing depth prone to technical errors (Metzker 2010), we decided to take an intermediate approach of sequencing a set of preselected genomic regions based on a reasonable sequencing depth and a large number of animals. Moreover, by considering many individuals we have a better chance to pick rare variants. Depending on the current price for sequencing services and the level of funding we aim to sequence at least 1 000 individuals with the minimum sequencing depth of 10.

Selection of genomic regions

Based on results of the currently running project at least 20 genomic regions spanning 1Mbp each will be selected for sequencing. The choice of regions is based on the effects of SNPs marking the region (as estimated within the frame of the MASinBULL project) and the information on physical locations of exons (as estimated within the frame of grant No. N N311 524940).

Prediction of genomic breeding values

The genomic breeding values will be predicted based on the following mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}\,,$$

where **y** represents a vector of deregressed conventional breeding values, **X** is a design matrix for fixed effects, **b** is a vector of fixed effects, which in the current model comprises only a general mean, **Z** is a 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** is a vector of random additive SNP effects, and **e** is a vector of residuals with $\mathbf{e} \sim \mathrm{N}\left(0, \mathbf{D}\hat{\sigma}_e^2\right)$ with **D** being a diagonal matrix containing the reciprocal of effective daughter contributions on the diagonal. The covariance structure of **g** is $\mathbf{g} \sim \mathrm{N}\left(0, \mathbf{I}\frac{\hat{\sigma}_a^2}{\mathrm{N}_{\mathrm{snp}}}\right)$, with **I** being an identity matrix and $\hat{\sigma}_a^2$ representing the additive genetic variance of a given trait. The estimation of parameters of the model will be based on solving the mixed model equations:

$\begin{bmatrix}\overset{\wedge}{\mathbf{b}}\\ \overset{\wedge}{\mathbf{g}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y}\end{bmatrix}$, with **R** represented by $\mathbf{D}\hat{\sigma}_e^2$ and **G** represented by $\frac{\hat{\sigma}_a^2}{N_{snp}}$. The genomic breeding value is defined as the sum of additive effects of SNPs estimated from the above model: $\hat{\mathbf{a}} = \mathbf{X}\hat{\mathbf{b}} + \mathbf{Z}\hat{\mathbf{g}}$ **.**

Estimation of the accuracy of genomic breeding values

The accuracy of genomic breeding values will be estimated based on the following model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}^{*}\mathbf{a} + \mathbf{e}$ **,**

where, $\mathbf{Z*}$ represents a design matrix for **a** - a vector of genomic breeding values of bulls distributed as $\boldsymbol{a}{\sim}N\left(0, \mathbf{A}_g\hat{\sigma}_a^2\right)$ with $\boldsymbol{A}_g$ defined as $\mathbf{ZZ^T}\frac{\mathbf{1}}{\boldsymbol{p}_{het}^{b}}$, with $\boldsymbol{p}_{het}^{b}$ representing the sum over all SNPs of heterozygous genotype frequencies in the base population estimated following (VanRaden, 2008). The accuracy of bulls' genomic breeding values are then given by: $\left\{\left(\mathbf{A}_{\boldsymbol{g}} - \frac{\hat{\sigma}_e^2}{\hat{\sigma}_a^2}\boldsymbol{C}^{\mathbf{22}}\right)\mathbf{A}_{\boldsymbol{g}}^{-\mathbf{1}}\right\}$ **,** where $\mathbf{C^{22}}$represents the part of inverse of the coefficient matrix from the mixed model equations corresponding to:

$$\begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}^{*}\\ \mathbf{Z}^{*T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{*T}\mathbf{R}^{-1}\mathbf{Z}^{*} + \mathbf{A}_g^{-1}\dfrac{\hat{\sigma}_e^2}{\hat{\sigma}_a^2}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{C}^{11} & \mathbf{C}^{12}\\ \mathbf{C}^{21} & \mathbf{C}^{22}\end{bmatrix}.$$

Estimation of linkage disequilibrium

Pairwise linkage disequilibrium (LD) is expressed as a squared correlation coefficient between allele counts observed at two SNPs and will calculated using the PLINK software (Purcell et al., 2007).

Analysis of DNA sequence data

The statistical methodology and software for the estimation of the most probable genotype and for the analysis of DNA sequence diversity is going to be developed and implemented within the framework of the project.

Results

Major results comprise a list of genes with a major and intermediate effects on quantitative traits routinely recorded in dairy cattle, accompanied with a corresponding list of SNPs which can be used as markers for the genes. Some of the SNPs may even represent the causal mutations, which will markedly increase the accuracy of the breeding value prediction in dairy cattle, since by now it mainly utilises linkage disequilibrium between SNPs and causal mutations. Knowledge of the genes will substantially improve understanding of the biology of complex traits.

The dissemination of intermediate and final project results will be done by:

- presentation on international conferences;
- discussing results and methodology on dedicated, national seminars;
- publication in scientific journals;
- information on current project development on a dedicated website.

Literature references

Amaral AJ, Megens HJ, Kerstens HHD, Heuven HCM, Dibbits B, Crooijmans RPMA, den Dunnen JT and Groenen MAM (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. BMC Genomics 10: 374.

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annual Reviews of Genetics 44: 293-308.

Daetwyler HD, Villanueva B and Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3: e3395.

Eyre-Walker A (2010) Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. PNAS 107:1752–1756.

Grisard B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelman R, Georges M, Snell R, Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with a major effect on milk yield and composition. Genome Research 12: 222-231.

Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S, Stefansson H, Sulem P, Gudbjartsson D, Maloney J, Hoyte K, Gustafson A, Liu Y, Lu Y, Bhangale T, Graham RR, Huttenlocher J, Bjornsdottir G, Andreassen OA, Jönsson EG, Palotie A, Behrens TW, Magnusson OT, Kong A, Thorsteinsdottir U, Watts RJ, Stefansson K. (2012) A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488:96-99.

Kappes M, Keele JW, Stone RT, McGraw RA, Sonstegard TS, Smith TPL, Lopez-Corrales NL, Beattie CW (1997) A second-generation linkage map of the bovine genome. Genome Research 7: 235-249.

Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, Jorgensen T, Hansen T, Pedersen O, Wang J, Nielsen R (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics 12: 231.

Lemay DG, Lynn DJ, Martin WF, Neville MC, Casey TM, Rincon G, Kriventseva EV, Barris WC, Hinrichs AS, Molenaar AJ, Pollard KS, Maqbool NJ, Singh K, Murney R, Zdobnov EM, Tellam RL, Medrano JF, German JB, Rijnkels M (2009) The bovine lactation genome: insights into the evolution of mammalian milk. Genome Biology 2009, 10:R43.

Meuwissen T and Goddard M (2010) Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics 185: 623–631.

Metzker ML (2010) Sequencing technologies — the next generation. Nature Reviews. Genetics 11: 31-46.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zöllner S, Whittaker JC, Chissoe SL, Novembre J and Mooser V (2012) An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science 337:100-104.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81: 559–575.

Sorensen AC, Sorensen MK and Berg P (2005) Inbreeding in Danish dairy cattle breeds. Journal of Dairy Science 88: 1865–1872.

The 1 000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

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