Differences Between Haplotype and Single SNP Effects for a Candidate Gene Affecting Fat Yield

L.G. Zielke[*], R.H. Bortfeldt[*], G. Thaller[†] and G.A. Brockmann[*]

Introduction

The availability of the Bovine SNP50 Bead chip of Illumina makes it possible to perform fine scale association analyses with candidate genes and milk composition traits in a fast and cost-effective way. A recent study in humans confirmed BDNF as a candidate gene with a strong association to the fat metabolism *Thorleifsson et al.* (2009). BDNF is a neurotrophic factor and a major regulator of neuronal transmission in the central nervous system. The bovine energy and fat metabolism influences the fat yield in milk, which is an important economical trait and breeding objective. We found BDNF to be highly conserved between human and cattle and chose it as an interesting genomic region to analyse SNP effects and differences between SNP and haplotype association with milk composition traits. It is hypothesised that haplotypes, composed of several SNPs that are inherited together, can capture more genetic variation than these SNPs analysed individually.

Material and methods

Genotypic data. 3,074 Holstein-Friesian (HF) animals (672 dams and 2,402 sires) were genotyped with the Illumina Bovine SNP50 Bead Chip (*Matukumalli et al.*, 2009) containing 54,001 SNPs. Out of 2,402 bulls, 48 were removed for low genotyping (>10% missing SNPs). In total, 8,748 SNPs were removed either because they failed genotyping in more than 10% of the animals (749 SNPs) or due to a minor allele frequency below 1% (7,998 SNPs).

Phenotypic data. Estimated breeding values (EBVs) for milk production traits [milk yield (MY), protein percentage (P%), protein yield (PY), fat percentage (F%) and fat yield (FY)] of all genotyped bulls were available from the national breeding evaluation carried out by VIT Verden (Germany) in August 2009. EBVs were available for the lactations 1 to 3 and the average of them. For the subsequent validation study of significant effects of the target region, 1,467 HF cows belonging to three different herds were genotyped for the most significant SNP located in the most significant haplotype block. Yield deviations for fat kilogram accounting for the herd (H_i), calving season ($H_i \times C_j$) and age of first calving (F_k) effects were calculated using a restricted maximum likelihood (REML) approach with the following model:

$$Y_{ijk} = \mu + H_i + H_i * C_j + F_k + \epsilon_{ijk}$$

Haplotype inference. Based on a more stringently filtered dataset (2,978 HF individuals with <3% missing genotypes, minor allele frequency >5%, <5% missing SNP calls, no missing SNP positions), the haplotypes were derived using the software FASTPHASE (*Scheet and Stephens*, 2006). The program was run with 10 random starts (parameter -T) and 25 iterations (parameter -C).

[†]Christian-Albrecht University Kiel, Olshausenstraße 40, 24098 Kiel, Germany...
[*]Humboldt-University of Berlin, Invalidenstr. 42, 10115 Berlin, Germany

Haplotype block computation. Phased genotyping data was partitioned into haplotype blocks using the solid spine algorithm implemented in the software HAPLOVIEW (*Barrett et al.*, 2005). A block is defined if all intermediate markers within a region are in linkage disequillibrium ($D' > 0.8$) with the first and last marker of that region but not necessarily with each other.

Association analysis. SNP association was tested using the software PLINK (*Purcell et al.*, 2007). The linear regression model estimated the additive effects of the SNP genotypes. The SNP most adjacent to the candidate gene was defined as the centre of the candidate region. All SNPs within 1 Mb up- and downstream of this centre-SNP were used for association analyses. Association analysis within the 2 Mb DGAT1 region, for example, showed that all SNPs were highly significant and hence the most significant DGAT1 SNP was fitted in the model for testing association to account for the allelic dosage of the DGAT1-effect. In the population, family patterns were identified (40 full sibs). To adjust for population stratification, the pairwise population concordance test (PPC) in PLINK was performed based on a IBS-similarity matrix and 124 significant clusters ($p < 10^{-4}$) were identified. Those were fitted in the model using the multi dimensional scaling (MDS) approach. Association analyses were adjusted for multiple testing using the Bonferroni correction. SNP genotype effects were tested for significance using ANOVA and subsequent Tukey-HSD test as implemented in the statistic software R. *Haplotype association* analyses were performed with the blocks spanning the candidate gene region. All haplotypes having a population frequency $>1\%$ were included in the association analyses. The same adjustments as used in the single SNP association analyses were fitted. For generating haplotype effect plots, the phased data generated by FASTPHASE was used. Haplotype effects were tested for significance using a linear model. A Tukey-HSD test was performed to test for differences between the haplotype groups.

Results and discussion

36 SNPs were found in the candidate 2 Mb region of the BDNF gene. After fitting the model described above and corrections for multiple testing, significant associations for all EBV milk composition traits were detectable. Six SNPs were found to be significantly ($p < 0.05$) associated with the average EBV for fat yield over the first three lactations. The most significant SNP ($p < 10^{-4}$) was *Hapmap53286-rs29015961*(Figure b). The difference between the two homozygous classes AA and GG was ~10 units of the EBV. The genotype GG, associated with the lowest mean value for EBV milk fat yield, had a low frequency of ~6% in the population. All genotypes were significantly different from each other ($p < 0.05$). This genotype explains ~1.7% of the total genetic variance. The mode of inheritance is additive ($a \approx 5$ units of EBV for fat yield over the first three lactations) (Figure b).

Table 1: Validation of SNP *Hapmap53286-rs29015961* in a HF cow population. [a]

Genotyp	Genotype Frequency	No. of Animals	Mean + SE	ANOVA
L1	-	1398	2.25, 1.17	0.015
GG	4.58	64	-5.89, 5.46	-
AG	34.12	477	-1.29, 1.99	-
AA	61.30	857	4.83, 1.49	-
L2	-	1299	1.72, 1.50	0.041
GG	4.31	56	-10.82, 7.24	-
AG	35.79	465	-1.31, 2.51	-
AA	59.89	778	4.34, 1.94	-
Average (L1,L2)	-	1239	3.36, 1.20	0.0075
GG	4.44	55	-6.33, 4.87	-
AG	35.51	440	-0.33, 2.00	-
AA	60.05	744	6.26, 1.56	-

[a] Shown are the p-values of the ANOVA (α = 0.05) from assocication with yield deviations for fat kilogram.SE = standard error of the mean, L = lactation.

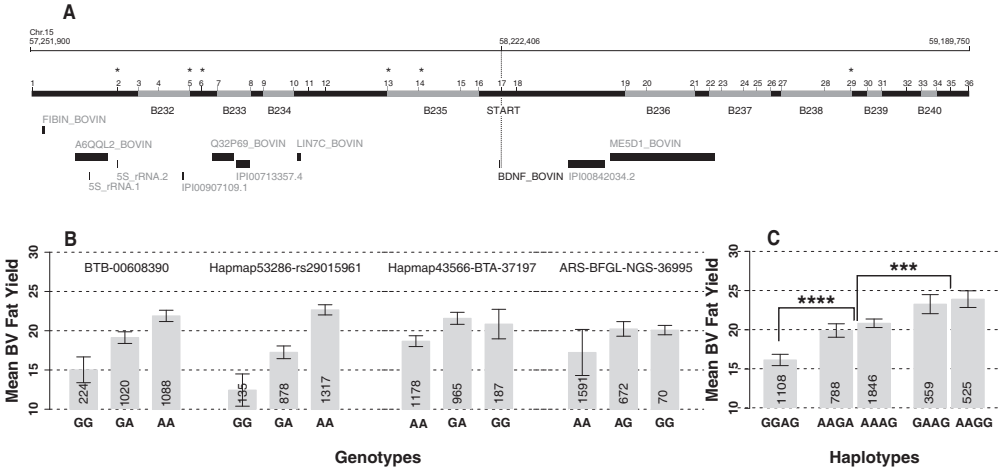

Figure 1: BDNF region with locations of SNPs, partially with significant associations for EBV fat yield (* $p < 0.05$). Haplotype blocks (horizontal grey bars) and adjacent genes are shown underneath. (B) Genotype effect plots of four SNPs associated with the EBV for fat yield. Two SNPs, *BTB-00608390*, and *Hapmap53286-rs29015961* are significant for this trait ($p < 0.05$). (C) Effect plot of five haplotypes in thre three groups; parentheses indicate significant differences in the EBV between the groups (* $p < 10^{-3}$, **** $< 10^{-4}$).**

The *Hapmap53286-rs29015961* SNP belongs to a haplotype block consisting of four SNPs covering ~187 Kb with an average D' of 0.95. This haplotype block is significantly associated with the average EBV of fat yield ($p < 10^{-4}$). In the block, five haplotypes occur each having a frequency of more than 1% in the bull population (Figure c). The effects of the five haplotypes cluster in three different groups: group I = {GGAG}, group II = {AAGA, AAAG} and group III = {GAAG,AAGG}. The difference in the mean EBV is significant between all groups ($p < 10^{-3}$) but not within the groups. The haplotype occuring most frequently in the

population belongs to group II. They show a medium EBV for milk fat yield of ~21 units (Figure c). As the three haplotype groups clearly have different effects, there must be more than one SNP responsible for the underlying genetic variation. For this reason, the patterns of inheritance of all four SNPs belonging to the same haplotype block were analysed separately. The two SNPs *ARS-BFGL-NGS-36995* and *Hapmap43566-BTA-37197* were not significantly associated to the EBV of fat yield when analysed in the SNP association separately (Figure c). However, it can be seen that their pattern of inheritance is different from the first two significant SNPs in the block (distance: 123,370 bp). Out of 135 homozygous low fat yield carriers of the genotype GG at the SNP *Hapmap53286-rs29015961* and 70 animals of the AA genotype at the SNP *ARS-BFGL-NGS-36995*, only 25 animals have a haplotype that harbours the low alleles at both SNPs. This implies that these two low effect alleles are rarely inherited together.

The results obtained in the bull population were validated in an independent HF cow population. The genotypes for the SNP *Hapmap53286-rs29015961* occur with frequencies of GG = 4.6%, AG = 34.1% and AA = 61.30% (Table) and are consistent with the frequencies in the bull population. Yield deviations (fat kilogram) were used for the validation and accounted for environmental effects, hence the differences between the means reflect the genetic difference. There are significant ($p < 0.05$) differences between the three genotypes when associated with the average yield deviations of lactation one and two as well as with both lactations indivudally. A subsequent Tukey-HSD test provided evidence for significant differences between the genotype classes AG and AA when associated to the average yield deviation of lactation one and two ($p < 0.05$). However, genotype GG showed no significant difference to the other genotypes, which could be due to the low number of animals or the small phenotypic effect size.

Conclusion

The candidate gene approach reduces the stringency of multiple testing by including prior knowledge of the position and thus can be more powerful than the genome-wide approach for the identification of significant association with the performance. This is particularly valuable for loci with small effects. According to the infinitesimal model, many genes with small effects are expected to contribute to milk fat yield. In addition to the advantage of the candidate gene approach, the analysis of haplotypes makes it possible to capture more of the underlying genetic variation than single SNPs alone. The additonal information gained by analysing haplotypes could contribute to increased accuracy of positonal genetic effects in the genomic selection approach, which currently only uses additive effects of single SNPs. Finally, the refined knowledge of the underlying genetic variation at a locus can be used for further fine-mapping studies to facilitate the search for the causal variants.

Acknowledgements
This project was supported by the German Federal Ministry for Eduaction and Research (BMBF). GenoTrack, Ref-number: O315 134B.

References

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly (2005), *Bioinformatics*, *21*(2), 263–265.

Matukumalli, L. K., et al. (2009), *PLoS One*, *4*(4), e5350.

Purcell, S., et al. (2007), *Am J Hum Genet*, *81*(3), 559–575.

Scheet, P., and M. Stephens (2006), *Am J Hum Genet*, *78*(4), 629–644.

Thorleifsson, G., et al. (2009), *Nat Genet*, *41*(1), 18–24.