Genetic evaluations and genomic characteristics for local cattle using genome sequences, 50K and a specific SNP chip

K. May[1*], M.J. Wolf[1], G.B. Neumann[2], P. Korkuć[2], G.A. Brockmann[2] and S. König[1]

[1]Justus-Liebig University Gießen, Institute for Animal Breeding and Genetics, Ludwigstraße 21 b, 35390 Gießen, Germany; [2]Humboldt University zu Berlin, Albrecht Daniel Thaer Institute for Agricultural and Horticultural Sciences, Animal Breeding Biology and Molecular Genetics, Invalidenstraße 42, 10115 Berlin, Germany; [*]katharina.may@agrar.uni-giessen.de

Abstract

This study compares the performance of the Illumina BovineSNP50 BeadChip (50K chip) with sequence data and a novel designed breed-augmented customized 200K SNP chip (200K chip) in the small local dual-purpose breed German Black Pied cattle (DSN). On such genomic databases, we applied genome-wide association analyses and estimated genetic parameters and breeding values for milk production, health indicator and calving traits. Accuracies of estimated breeding values were calculated using a cross-validation approach. We identified significant SNP marker associations and candidate genes with the specifically designed 200K chip, which were not detected with the commercial 50K chip. The annotated genes (e.g., *MGST1* for 305-day lactation milk yield) were confirmed based on whole-genome-sequence data, indicating breed-specific genomic mechanisms. Heritabilities and accuracies of genomic breeding values increased slightly by using the novel specifically designed 200K SNP chip compared with the 50K SNP chip.

Introduction

Marker density affects the power of genome-wide association studies (GWAS) and genomic estimated breeding value (GEBV) accuracy. The Illumina BovineSNP50 BeadChip with ~54.000 markers (hereafter referred to as 50K chip) is commonly used for GWAS and genetic evaluations in various cattle breeds worldwide. The 50K chip is designed on a reference panel of large commercial populations like Holstein Frisian [HF] and Simmental. The chip works particularly well in breeds closely related to these reference population (Matukumalli et al., 2009). The German Black Pied cattle breed (DSN, German: Deutsches Schwarzbuntes Niederungsrind) is an endangered local dual-purpose breed with a small population size of ~2,500 cattle and considered as the founder breed of HF. However, only ~37,000 markers of the 50K chip are informative in DSN after filtering, which coincides with observations in other small cattle breeds (e.g., Hozé et al., 2013). Moreover, the well described *DGAT1*gene with major effects on milk production traits in HF is one example for genes which are not associated with milk production traits in DSN in GWAS (Korkuć et al., 2021), although DSN and HF are closely genetically related. Hence, we designed a customized 200K SNP chip for DSN including breed-specific variants (Neumann et al., 2021). We hypothesize that a denser breed-specific SNP chip is helpful in GWAS to identify DSN-specific variants for a variety of traits. Wu et al. (2015) showed an increased power to detect significant associations with increasing marker density for mastitis in Danish Holsteins when comparing 50K, 777K and whole-genome sequence (WGS) data.

The objective of this study was to compare three marker densities (50K, 200K and WGS) in GWAS for milk production, health indicator and calving traits in DSN. Furthermore, we compared the accuracies of genomic breeding values obtained using different SNP chip arrays.

Materials & Methods

Phenotypes. We considered first-parity DSN cows from six herds. In total, 2,020 records were available for 305-day milk yield (M305), fat yield (F305) and protein yield (P305), 1,638 records for test-day fat-to-protein ratio (FPR) and somatic cell score (SCS; log-transformed somatic cell count) as health indicators, and 2,606 records for calving ease (CE) and stillbirth incidence (SB). For FPR and SCS, we focused on the challenging early lactation period and included the first test-day record between 5 and 40 days in milk (DIM). The breed (DSN) percentage was larger than 90% for all cows according to an algorithm to clearly differentiate between DSN and HF (Jaeger et al., 2018).

SNP chip design and imputation of genotypes. The customized DSN-augmented 200K SNP chip (Axiom® myDesign™ TG Array; hereafter referred as 200K chip) was designed as described by Neumann et al. (2021). Variants for the chip were selected from 20,587,181 sequence variants (SVs) detected from WGS of 304 DSN cattle. The chip contains DSN unique variants, variants associated with important traits of interest in DSN (e.g., disease resistance, milk yield, fertility), variants with low, high or moderate impact on the transcripts, DSN informative variants from the 50K chip as well as 175,537 SNPs and 8,618 indels corresponding to 103,801 haplotype blocks in DSN. In total, 300 DSN cattle were genotyped with the novel designed 200K chip and provided a reference panel for the imputation of 1,797 DSN 50K chip genotypes into 200K genotypes. The filtered WGS data included 16,175,216 SVs of 304 DSN and served as a reference panel for the imputation of 1,797 DSN 200K genotypes to WGS level. The imputation and filtering of SVs were performed in BEAGLE (Browning et al., 2018). Quality control of the three marker densities (50K, 200K and WGS) was performed in PLINK (Purcell et al., 2007). Depending on the trait category, the final number of markers after filtering were ~37.000 SNPs for 50K, ~125.000 for 200K and ~11,600,000 SVs for WGS.

Genome-wide associations and genetic parameter estimations. For the GWAS, we applied a single marker linear mixed model using the “leaving one chromosome out” option in GCTA (Yang et al., 2011) for all three marker densities (50K, 200K, WGS). The statistical model in matrix notation was:

$$\mathbf{y = X\beta + Zu + Ss + e} \qquad (1)$$

where **y** = vector of observations for M305, F305, P305, FPR, SCS, CE and SB ; **β** = vector of fixed effects (herd-calving year-calving season and a linear regression on DIM and age at first calving for M305, F305 and P305; herd, test-day year-season and a linear regression on age at first calving for FPR; herd, test-day year-season, a linear regression on DIM and on fat percentage for SCS; herd-calving year-calving season and sex of calf for CE and SB); **u** = vector of polygenic effects with $\mathbf{u} \sim N\ (0,\ \mathbf{G}\sigma^2_u)$, with **G** denoting the genomic relationship matrix (VanRaden, 2008), and σ^2_u the polygenic variance; **s** = vector for marker effects; **e** = vector of random residuals; and **X**, **Z**, and **S** were incidence matrices for **β**, **u**, and **s**, respectively. We calculated an adjusted Bonferroni-corrected genome-wide significance threshold (*p*Bonf) with $p = 0.05/\mathrm{n_{eff}}$ with $\mathrm{n_{eff}}$ = effective number of independent SNPs/SVs. In addition, we considered a suggestive significance threshold *p*Sug with $p = 1/\mathrm{n_{eff}}$. The -indep-pairwise option in PLINK (Purcell et al., 2007) was used to calculate $\mathrm{n_{eff}}$ by excluding one SV of a SV pair in LD r^2>0.5 in a window size of 5000 SVs, which was shifted in an interval of 500 SVs. SNP effect correlations were calculated genome-wide and chromosome-wide for each trait for all overlapping SNPs/SVs between the three marker densities. Furthermore, model (1) was applied to estimate genetic parameters for all traits with all three marker densities using restricted maximum likelihood (REML) with the --reml function in GCTA.

Gene annotation. Potential candidate genes were queried and assigned to the associated SNPs/SVs using the current gene annotations from ENSEMBL (release 104). A gene was considered as a candidate gene if at least one SNP/SV with $P<p$Sug was located in the respective gene and/or within 150 kb up- and downstream. Physiological functions of candidate genes were studied in the ENSEMBL and KEGG databases.

Genomic breeding value estimation. Breeding values were estimated in two consecutive runs: i) 50K chip vs. 200K chip using 1,900 DSN and ii) 200K chip vs. 10 random selected 200K out of WGS data (random200K) using 1,980 DSN. Phenotypic data for M305 and F305 were pre-corrected for fixed effects according to model (1) in R. Afterwards, we applied an animal model for genomic breeding value estimations using pre-corrected phenotypes (residuals) in BLUPF90 (Misztal et al., 2002). The statistical model was:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{g} + \mathbf{e} \qquad (2)$$

where $\mathbf{y}$ = vector for residuals for M305 and F305; $\boldsymbol{\mu}$ = overall mean; $\mathbf{g}$ = vector of additive genetic effects with $\sim N\,(0, G\sigma^2_g)$, with $\mathbf{G}$ denoting the genomic relationship matrix (VanRaden, 2008), and σ^2_g the genomic variance; $\mathbf{e}$ = residual effects with $\sim N\,(0, \sigma^2_e)$, and σ^2_e residual variance. For cross-validation, we divided the dataset in a validation set (20% of cows) and training set (80% of cows). We selected the cows randomly in a procedure with 10 replicates using replacements. The accuracy was calculated for each of the 10 replicates as the correlation between the direct GEBV and residuals. Afterwards, the mean accuracy was calculated based on the 10 replicates for M305 and F305 for 50K, 200K and random200K.

Results

The numbers of associated SNPs/SVs and genes for the three marker densities per trait are given in Table 1. For M305, F305 and P305, we identified 23 SNPs with 50K, 84 SNPs with 200K and 950 SVs with WGS with $P<p$Bonf/pSug. For M305, 14.8% (4/27) of genes (*KLF3, LIMCH1, NSUN7, RALGAPA2*) overlapped for 50K and 200K. The genes with the largest number of SNPs/SVs located in a gene for M305 were *MGST1* (BTA5), *ADGRL3* (BTA6) and *GNAL* (BTA24), exclusively detected with 200K and WGS, but not with 50K (Figure 1).

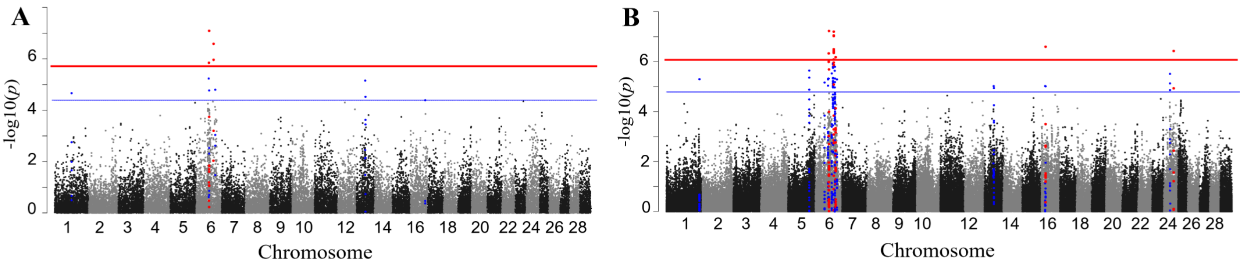

Figure 1. Manhattan-plots for 305-day lactation milk yield (M305); A) 50K Illumina BeadChip and B) novel designed DSN 200K SNP chip; markers above *p*Bonf (red line) are highlighted in red; markers above *p*Sug (blue line) are highlighted in blue; markers in a distance of 125kb up- and downstream of associated markers are highlighted, too.

For SCS, we identified 62 SVs with WGS but no SNP reached pSug for the 50K and 200K chips. For the calving traits CE and SB, we identified 11 genes with 50K and 23 genes with 200K. Among them, seven common genes were annotated with both SNP chips. Genome-wide and chromosome-wide SNP effect correlations between 50K and 200K were ≥ 0.99 ($p \leq 0.001$) for all traits, and ranged from 0.62 to 0.81 ($p \leq 0.001$) between WGS with 50K and 200K. The estimated heritabilities were quite similar for the three marker densities (Table 1).

Table 1. Number of associated markers (SNPs/SVs) and identified genes (G) in the GWAS and heritabilities (h^2) with standard error (SE) for the three marker densities

	50K		200K		WGS	
Trait	No. SNPs/G	h^2	No. SNPs/G	h^2	No. SVs/G	h^2
M305	11/11	0.40 (0.04)	64/20	0.39 (0.04)	724/48	0.41 (0.04)
F305	8/10	0.33 (0.04)	11/8	0.33 (0.04)	115/20	0.34 (0.04)
P305	4/6	0.34 (0.04)	9/7	0.33 (0.04)	111/11	0.35 (0.04)
FPR	4/4	0.13 (0.04)	10/6	0.14 (0.04)	52/12	0.13 (0.04)
SCS	0/0	0.11 (0.04)	0/0	0.12 (0.03)	62/2	0.13 (0.04)
CE	0/0	0.03 (0.02)	1/1	0.03 (0.02)	5/2	0.04 (0.02)
SB	1/1	0.04 (0.02)	2/1	0.05 (0.02)	10/8	0.05 (0.02)

M305/F305/P305 = 305-day lactation yield for milk, fat and protein; FPR = fat-to-protein ratio; SCS = somatic cell score; CE = calving ease; SB = stillbirth

For GEBV run (i), the mean accuracy was 0.46 for M305 and 0.36 for F305 with the 50K chip, and 0.47 and 0.37 with the 200K chip, respectively. When comparing the 200K chip with random200K in run (ii), mean accuracy was 0.46 for M305 and 0.34 for F305 with the 200K chip, and ranged from 0.446 to 0.450 for M305 and from 0.320 to 0.336 for the 10 random200K.

Discussion

Similar heritabilities and GEBV accuracies for different marker densities and traits showed a limited advantage when using the customized 200K chip compared to the 50K chip in genetic evaluations for the endangered DSN breed. Similarly, Erbe et al. (2012) indicated a small increase in genomic prediction accuracies of 0.01 in HF and 0.03 in Jerseys when using imputed high-density SNP panels compared to the Illumina 50K. However, since DSN and HF are closely genetically related, a breed-specific SNP chip might be more advantageous for small breeds that are distantly related with the large commercial cattle populations HF, Simmental and Brown Swiss. In agreement with the results by Wu et al. (2015), the power to detect candidate genes was improved when using the specifically designed 200K chip and WGS data compared to 50K. This might be due to stronger linkage disequilibrium between markers and quantitative trait nucleotides (QTN) for higher marker densities. SNP effect correlations were close to one between 50K and 200K, but moderate to high between WGS with both SNP chips. Hence, WGS is suitable to detect breed-specific variants affecting traits of interest in small-sized endangered cattle breeds, which can be integrated into genomic prediction models to improve prediction accuracies.

References

Browning, B.L. (2018) Am. J. Hum. Genet. 103:338-348. doi.org/10.1016/j.ajhg.2018.07.015

Erbe M. et al. (2012) J. Dairy Sci. 95:4114-4129. https://doi.org/10.3168/jds.2011-5019

Hozé C. et al. (2013) Sel. Evol. 45:33. https://doi.org/10.1186/1297-9686-45-33

Jaeger M., Scheper C., König S., Brügemann K. (2018) Züchtungskunde 90, pp. 262-279.

Korkuć P. et al. (2021) Front. Genet. 12:275. https://doi.org/10.3389/fgene.2021.640039

Matukumalli L.K. et al. (2009) PloS ONE 4:e5350. doi.org/10.1371/journal.pone.0005350

Misztal I., Tsuruta S., Strabel T. et al. (2002) Proc. of the 7th WCGALP, Montpellier, France.

Neumann G.B. et al. (2021) BMC Genomics 22:905. doi.org/10.1186/s12864-021-08237-2

Purcell, S. et al. (2007) Am. J. Hum. Genet. 81:559-575. https://doi.org/10.1086/519795

VanRaden, P.M. (2008) J. Dairy Sci. 91:4414-4423. https://doi.org/10.3168/jds.2007-0980

Wu X. (2015) Genet. Sel. Evol. 47:50. https://doi.org/10.1186/s12711-015-0129-1

Yang J. et al. (2011) Am. J. Hum. Genet. 88:76-82. https://doi.org/10.1016/j.ajhg.2010.11.011