Improved genome-wide associations using a breed-specific 200K SNP chip for German Black Pied (DSN) cattle

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

[1] Humboldt University Berlin, Albrecht Daniel Thaer- Institute for Agricultural and Horticultural Sciences, Animal Breeding Biology and Molecular Genetics, Invalidenstr. 42, 10115 Berlin, Germany; [2]Justus-Liebig-University of Gießen, Institute of Animal Breeding and Genetics, Ludwigstr. 21, 35390 Gießen, Germany; *paula.korkuc@hu-berlin.de

Abstract
The DSN200K chip was specifically designed for German Black Pied cattle (DSN "Deutsches Schwarzbuntes Niederungsrind") to improve genome-wide association analyses (GWAS) in DSN. Here, we performed GWAS with 2,099 DSN cows and 20 milk performance traits using 131,273 SNPs from the DSN200K chip. We compared those GWAS results to results obtained with the Illumina BovineSNP50 BeadChip (Illumina 50K) and *in silico* randomly generated SNP chips. Those random SNP chips had the same number of SNPs as on the DSN200K chip selected from a prefiltered set of 2.6 million sequence variants. We identified 13-15 times more associated SNPs with the DSN200K chip than with the Illumina 50K and 2.7-3.2 times more than with random SNP chips. With the DSN200K chip, we associated DSN-unique SNPs and SNPs with impact on gene transcripts which could potentially be causal mutations. These SNPs were seldom identified with the Illumina 50K or the random SNP chips.

Introduction
German Black Pied cattle are an endangered dual-purpose cattle breed, which is considered to be one of the ancestor populations of the modern Holstein breed. In order to track DSN-specific sequence variants that could help to conserve the specific genetic diversity of DSN and to allow for genetic improvement, we designed a breed-specific chip for DSN cattle (DSN200K chip) with 182,154 sequence variants (Neumann et al., 2021).
Previous genome-wide association analyses (GWAS) in DSN were performed using genotypes from the Illumina BovineSNP50 BeadChip (Illumina 50K) (Korkuć et al., 2021; May et al., 2019; Meier et al., 2020). The Illumina 50K chip was designed based on commercial breeds (Matukumalli et al., 2009). Only 37K out of all 54K SNPs on the Illumina 50K chip were informative in DSN considering a SNP call rate>0.95 and a minor allele frequency (MAF)>0.05 and thus suitable for GWAS. Due to the ascertainment bias of the chip design, DSN-unique SNPs are missing on this chip. The designed DSN200K chip provides not only a higher density of SNPs but includes also DSN-unique SNPs and SNPs from previous GWAS in DSN. Moreover, it contains also SNPs of general interest such as SNPs with impact on gene functions and SNPs from the Illumina 50K chip. Additionally, SNPs were included that track most haplotype blocks found in DSN.
In this study, we performed GWAS with 2,099 DSN cattle and 20 milk performance traits using 131,273 SNPs from the DSN200K chip. We compared those results to results obtained with the lower-density Illumina 50K chip and *in silico* randomly generated SNP chips having the same number of SNPs as the original DSN200K chip in order to test whether only the number of SNPs or also the selection of SNPs on the DSN200K chip improve GWAS results.

Materials & Methods
Genotypes and traits. Genotypes of 2,099 DSN cows from 8 farms in Germany were available from either Illumina 50K (1,473 cows) Korkuć et al., 2021), DSN200K chip (453 cows) or

whole-genome sequencing (173 cows) (Neumann et al., 2021). For GWAS, genotypes were imputed to whole-genome sequencing (WGS) level using 304 sequenced DSN cattle (Neumann et al., 2021) with BEAGLE v5.1 and then reduced to the SNP positions of the DSN200K and the Illumina 50K chip that were also available on the DSN200K chip. SNPs with MAF<0.05 and indels were removed, leading to 131,273 and 33,546 SNPs on the autosomes and chromosome X suitable for analyses on the DSN200K and the Illumina 50K chips, respectively. Milk performance data at 305 days was obtained from "Vereinigte Informationssysteme Tierhaltung w.V." as of February 2021. Traits included milk, fat, and protein yield in kilogram and fat and protein content in percent for the first three lactations and the lactation mean when data for all three lactations was available. Only full lactations with at least 270 days in milk were considered.

Genome-wide association analysis. Multiple linear regression models implemented in R were used to test the additive effect of each SNP. For each trait *Y*, the model included covariates for population stratification *ps,* farm *f*, sire *s*, birth year *by*, birth season *bs*, calving year *cy*, calving season *cs*, and age at first calving in days *ac*, together with the SNP genotype *gt* and the residual error *e:* $Y = ps + f^* + s^* + by^* + bs^* + cy^* + cs^* + gt + e$ (1)
Population stratification *ps* was estimated using the pairwise population concordance test in PLINK v1.9 which is based on a pairwise identity-by-state matrix. The p-value cut-off of 0.0001 resulted in 65 clusters of relatedness. Covariates marked with an asterisk "*" were only included in the model when the difference in Akaike information criterion was ≤ -10 between the null model ($Y = ps$) and the null model extended with one of the covariates ($Y = ps +$ *covariate x*). To reduce the number of false positives in GWAS, p-values of each trait were corrected to an inflation factor of λ=1.3 on autosomes and chromosome X, separately. The number of independent SNPs (n= 95,140) was estimated using LD-pruning SNPs with r^2>0.8 using PLINK v1.9 and used to adjust p-values for multiple testing with Bonferroni correction. After correction, SNPs were considered as significant when p<0.05 or suggestive when p<0.1. Associated SNPs within ± 5 Mb across all traits were combined to quantitative trait loci (QTLs).

In silico random SNP chips. For the design of the DSN200K chip, a set of 3.1 million quality-controlled sequence variants was used (Neumann et al., 2021) which were filtered for MAF<0.05 and indels. From the left 2.6 million SNPs, 100 random SNP chips with 131,273 SNP positions were generated *in silico* by keeping all 33,546 informative SNPs from the Illumina 50K chip and selecting additional 97,727 SNPs randomly and uniformly across all chromosomes. The SNP distribution across chromosomes on the original and random SNP chips was similar. The mean distance between adjacent SNPs was 19,990 ± 25,439 SD for the original and 19,972 ± 26,148 SD for the randomly generated SNP chips. GWAS was repeated with those 100 random SNP chips as described above.

SNP annotation and haplotype blocks. The SNP selection categories for the design of the DSN200K chip and the estimation of haplotype blocks were thoroughly described in Neumann et al. (2021). SNP categories included SNPs from previous GWAS in DSN (Korkuć et al., 2021; May et al., 2019; Meier et al., 2020; Wolf et al., 2021), SNPs with high, moderate or low impact on gene transcripts (e.g. missense mutations) as estimated with Ensembl Variant Effect Predictor, and DSN-unique SNPs.

Results

GWAS with the DSN200K chip identified a total of 163 (131 significantly and 32 suggestively) associated SNPs (Table 1). Those SNPs covered 81 out of possible 87,266 haplotype blocks and could be grouped into 25 QTLs (61 haplotype blocks; 13 QTLs when considering only

significant SNPs). The identified associated SNPs were mainly SNPs that were added to the DSN200K chip as significant SNPs from previous GWAS results. But additional SNPs with high, moderate or low impact on the function of genes and DSN-unique SNPs that are important to identify breed-specific QTLs in DSN were also found. Many of the associated SNPs that had been found already in previous GWAS in DSN, had a potential impact on the gene transcript and/or were DSN-unique.

Only 13 (9 significantly and 4 suggestively) associated SNPs were identified when using the 33,546 SNPs from the Illumina 50K chip which were also present on the DSN200K chip. These were 13-15 times less associated SNPs than with the DSN200K chip while using only 4 times less SNPs. With regard to QTLs, 11 out of 25 would not have been detected when considering only SNPs from the Illumina 50K chip (5 out of 13 QTLs when considering only significant SNPs). Only 12 haplotype blocks were covered by those SNPs (8 when considering only significant SNPs).

GWAS with the randomly generated SNPs chips identified on average 60 ± 8 SD associated SNPs (41 ± 7 SD when considering only significant SNPs) (Table 1). This was around 2.7-3.2 times less SNPs than with the breed-specific DSN200K chip. Also, on average 1.6-1.7 times less haplotype blocks were covered. With regard to QTLs, with the randomly generated SNP chips on average 17 ± 1 SD out of 25 QTLs that were found with the original DSN200K chip could be identified as well (11 ± 1 SD out of the 13 QTLs when considering only significant SNPs). Almost no SNPs with impact on the gene transcripts or unique to DSN were identified with the randomly generated SNP chips.

The QTL with the highest significance using the DSN200K chip was found on chromosome 5 for milk fat content in lactation 1 (Figure 1). The top SNP of this QTL (5:93,516,066, rs211210569) is located in the gene *MGST1* (Microsomal Glutathione S-Transferase 1). This region was found to be as well associated with milk fat content in Holstein (Littlejohn et al., 2016). This QTL would have been missed when considering only significant SNPs from the Illumina 50K chip, but would be likely found with any SNP chip of higher density.

Table 1. Comparison of GWAS results using different SNP chips.

		Number of SNPs				Number of haplotype blocks
		Categories[1]				
Chip	**Threshold**	**Previous GWAS**	**VEP impact**	**DSN-unique**	**Total[2]**	**Total[2]**
DSN200K	**MAF>0.05**	1,898	40,120	7,295	131,273	87,266
	p<0.1	99	16	8	163	81
	p<0.05	83	10	8	131	61
Illumina 50K	**MAF>0.05**	134	1,016	0	33,546	26,615
	p<0.1	8	0	0	13	12
	p<0.05	6	0	0	9	8
Random SNP chips[3]	**MAF>0.05**	19 ± 8	2,491 ± 36	294 ± 17	131,273 ± 0	56,382 ± 98
	p<0.1	12 ± 2	1 ± 1	0 ± 1	60 ± 8	49 ± 6
	p<0.05	9 ± 2	0 ± 1	0 ± 1	41 ± 7	35 ± 5

[1] SNPs can be assigned to multiple categories; [2] Each SNP or haplotype block was counted only once;
[3] Values are listed including their ± standard deviations.

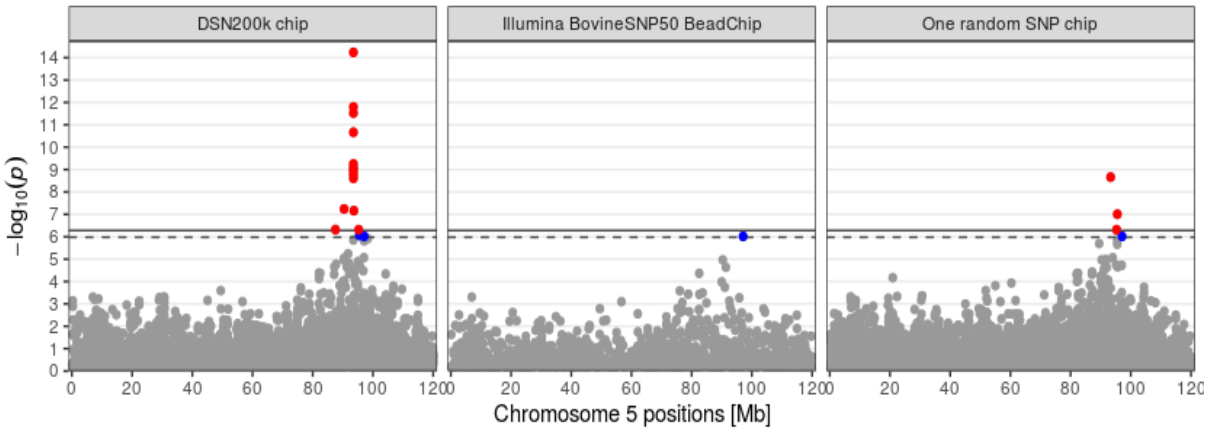

Figure 1. QTL on chromosome 5 for milk fat content in lactation 1 using different SNP chips (red=significant SNPs, blue=suggestive SNPs).

Discussion

Higher density SNP chips (DSN200K and random SNP chips with 137K informative SNPs) identified as expected remarkably more associated SNPs that could be grouped also into more QTLs compared to the Illumina 50K chip. The higher number of associated SNPs in each QTL region improved also their certainty. Furthermore, the carefully designed breed-specific DSN200K chip was better at identifying associated SNPs and QTLs with milk performance in DSN than the randomly generated SNP chips. This was expected, as this chip already contained previously identified SNPs associated with milk performance in a smaller population of 1,490 DSN cows (Korkuć et al., 2021). With regard to SNP categories, SNPs having impact on gene transcripts or breed-specific SNPs were underrepresented on the randomly generated SNPs chips. Because these functional SNPs have a high potential for being causal for economically important phenotypes, they should be included on breed-specific SNP chips. Our findings support the fact that adding significant SNPs from GWAS and SNPs with functional impact or unique to the investigated breed to a custom SNP chip is beneficial for the target breed for future GWAS and genomic breeding. In conclusion, we obtained not only quantitatively but also qualitatively best results with the breed-specific DSN200K chip compared to the lower density Illumina 50K chip and to the randomly generated SNP chips.

References

Korkuć, P., Arends, D., May, K., König, S., Brockmann, G. A. (2021). Front. Genet. 12:275, doi:10.3389/fgene.2021.640039

Littlejohn, M. D., Tiplady, K., Fink, T. A., Lehnert, K., Lopdell, T., et al. (2016). Sci. Rep. 6(1):1-14, doi:10.1038/srep25376

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., et al. (2009). PLoS One 4(4):e5350. doi:10.1371/journal.pone.0005350.

May, K., Scheper, C., Brügemann, K., Yin, T., Strube, C., et al. (2019). BMC Genomics 20(1):1-15. doi:10.1186/s12864-019-5659-4

Meier, S., Arends, D., Korkuć, P., Neumann, G. B., Brockmann, G. A. (2020). J. Dairy Sci. 103(11):10289-10298, doi:10.3168/jds.2020-18209

Neumann, G. B., Korkuć, P., Arends, D., Wolf, M. J., May, K., et al. (2021). BMC Genomics 22(1):1-13. doi:10.1186/s12864-021-08237-2

Wolf, M. J., Yin, T., Neumann, G. B., Korkuć, P., Brockmann, G. A., et al. (2021). Genes, 12(8):1163. doi:10.3390/genes12081163