Rare CNVs in the bovine genome are not captured well by 50K density genotyping array SNPs

Y.L. Lee[1*], M. Bosse[1], W. Coppieters[2,3], R.F. Veerkamp[1], L. Karim[3], C. Oget-Ebrad[2], T. Druet[2], M.A.M. Groenen[1], M. Georges[2], A.C. Bouwman[1], C. Charlier[2]

[1] Wageningen University & Research, Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands [2]Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, 4000 Liège, Belgium [3]GIGA Genomics Platform, GIGA Institute, University of Liège, 4000 Liège, Belgium [*]younglim.lee@wur.nl

Abstract
Understanding how genomes, particularly genetic variants, instruct animals to develop and function is crucial for animal breeders. Copy number variants (CNVs), the gain or loss of DNA segments, can affect gene expression and alter phenotypes. However, most large-scale genetic analyses in animal breeding use single nucleotide polymorphism (SNP) genotypes from genotyping arrays, and efforts to incorporate CNVs are limited. In theory, variation from CNVs can be captured by SNPs, if they are in high linkage disequilibrium (LD). We evaluated the LD of CNV-SNP pairs in whole genome sequencing and genotyping array data. Our whole genome sequencing data showed that most CNVs have tagging SNPs. However, CNV-SNP pairs from genotyping array data were mostly in low LD, because most CNV-SNP pairs had unmatching allele frequencies. We conclude that most rare CNVs may not be fully captured by genotyping arrays.

Introduction
Bovine genomes harbour diverse types of genetic variants. Identifying and utilizing impactful genetic variants is of prime interest for animal breeders, given the economic importance of cattle populations. Copy numbre variants (CNVs) are a subset of structural variants, which include deletions (DELs) and duplications (DUPs) of DNA segments larger than 50 bp. Although CNVs are not as abundant as small genetic variants (e.g. SNPs), CNVs contribute disproportionally more to gene expression than SNPs (Chiang *et al.,* 2017; Scott *et al.,* 2021), hinting that they could be associated with phenotypes. Recent reports in farm animals, which unravelled high impact CNVs as causative variants underlying QTL of complex traits, also support the importance of CNVs (Lee et al. 2021; Derks et al. 2018; Kadri et al. 2014).

Despite the functional evidence of CNVs, their utilization in animal breeding is limited. To date, large-scale routine genetic analyses in breeding programmes, such as genomic prediction, have mostly focused on SNPs obtained from genotyping arrays. In theory, if those SNPs are in high LD with CNVs, the variation from CNVs can be captured by SNPs. To this end, we aimed at (i) detecting CNVs in deeply sequenced genomes of Holstein Friesian (HF) cattle, (ii) validating a subset of CNVs in an independent cohort using a direct genotyping approach, and (iii) evaluating pairwise LD in CNV-SNP pairs in whole genome sequencing (WGS) and genotyping data.

Materials & Methods
WGS data. A family cohort (DAMONA) consisting of 266 HF animals (including 127 trios) was sequenced using the Illumina HiSeq 2000 instrument. The data was aligned using BWA mem (version 0.7.9a-r786) (Li 2013) to the bovine reference genome ARS-UCD1.2 (Rosen *et*

al. 2020). All samples had a minimum mean sequencing coverage of 15X, and the mean coverage was 26X.

SNP and CNV discovery in the WGS data. SNP variant calling was done using the GATK workflow (v4.1.7), and recalibrated using the following algorithms: BaseRecalibrator, HaplotypeCaller, GenomicsDBImport, GenotypeGVCF, GatherVcfs, Variant Recalibrator (DePristo *et al.* 2011; McKenna *et al.* 2010; Auwera *et al.* 2013). Variant Quality Score Recalibration (VQSR) at a truth sensitivity filter level of 97.5 was used to remove spurious variants. CNVs were detected using the Smoove pipeline, which utilized split and discordant reads evidence to discover CNVs in each sample, followed by population-wide genotyping (https://github.com/brentp/smoove). Subsequently, we used the fold-coverage change of read depth in CNVs to retain accurately genotyped biallelic CNVs, using Duphold (Pedersen and Quinlan 2019).

Genotyping data. To validate CNV discovery results, we genotyped a subset of CNVs by adding them in the custom part of the EuroGenomics SNP genotyping array, which harbours ~50K SNPs (Boichard *et al*. 2018). Probes targeting breakpoint sequences of 372 CNVs (342 deletions and 30 duplications) that appeared in non-repetitive regions, were added. Genotyping was done for 815 independent HF animals using their ear punch or blood samples.

LD analyses. Pairwise LD (r^2) for CNV-SNP pairs was calculated, to evaluate whether CNVs had tagging SNPs. This analysis was done separately for the WGS and EuroGenomics array data sets. SNPs located within 100-kb distance from CNV breakpoints (WGS) or CNV probes (array) were considered for LD calculation. Additionally, LD decay was compared between CNV-SNP and SNP-SNP pairs limited to the Eurogenomics array data. This analysis was done for common (allele frequency≥0.05) and rare (allele frequency<0.05) DELs separately. Due to a limited number, DUPs was not considered (n=19). The LD calculation was done using the PLINK software (Purcell *et al.* 2007).

Results

Overall variant discovery in the WGS data set. The SNP variant calling pipeline resulted in 11,030,905 SNPs from the 266 sequenced samples. Using the same WGS data set, we discovered 13,731 CNVs (12,200 deletions and 1,531 duplications). Genotyping accuracy of CNVs is lower than that of SNPs (Chiang *et al.* 2017). Hence, we applied read-depth based filters, resulting in 4,011 accurately genotyped biallelic CNVs (3,827 deletions and 184 duplications).

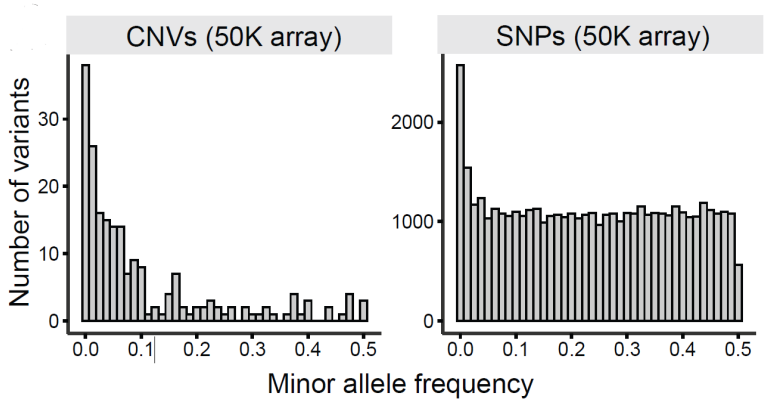

Figure 1. Minor allele frequency spectra of CNVs and SNPs obtained from the EuroGenomics array.

Validation of CNVs using genotyping arrays. We validated a subset of WGS CNVs, using a direct genotyping approach in 815 HF animals. All genotyped samples had a sample call rate > 0.99). Of the 284 CNVs that passed the variant level filter, 229 were segregating (210 deletions and 19 duplications), showing a validation rate of 80% (229/284). The SNP genotyping results showed that 50,342 SNPs were segregating in the population. The minor frequency spectrum of CNVs was skewed towards rare variants, whereas that of SNPs were skewed towards common variants (Figure 1).

LD analyses in the WGS and the genotyping data. To evaluate whether WGS CNVs have tagging SNPs in WGS data, we calculated the LD of CNV-SNP pairs. A majority of biallelic WGS CNVs (97% of DELs and 93% of DUPs) had at least one WGS SNP in LD (r^2>0.8) within a 100-kb distance from the CNV. However, the same analyses based on the EuroGenomics array data revealed that only 15.4% of DELs and 4.5% of DUPs included on the array had a tagging array SNP (r^2>0.8) within a 100-kb distance. Lastly, the LD in the EuroGenomics array data confirmed that the CNV-SNP pairs have low LD than SNP-SNP pairs (Figure 2; analyses was done limited to DELs, due to low numbers of DUPs).

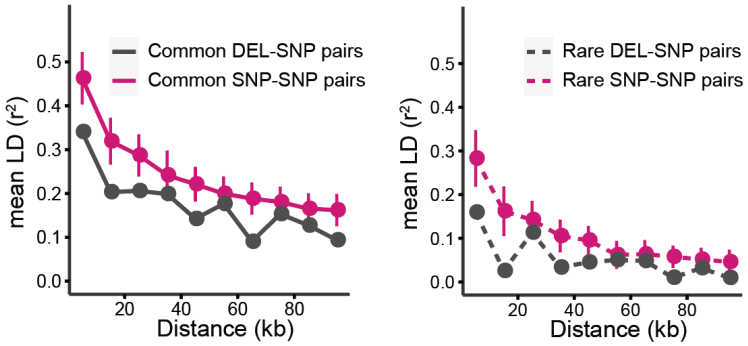

Figure 2. LD decay patterns in DEL-SNP and SNP-SNP pairs

Discussion

A large percentage of the CNV in our WGS data set were well tagged by at least one WGS SNP. In a human study, only ~80% of CNVs had at least one tagging SNPs, based on low-pass WGS data (4-7X coverage; Sudmant et al. 2015). The higher LD in our data set is likely due to (i) high relatedness among the animal samples, as opposed to the non-related human samples in Sudmant et al. (2015), and (ii) deeper WGS data in our study. The deeper WGS contributed to more accurate genotyping of CNVs, which may have resulted in higher LD between CNV and SNP genotypes.

The LD between CNV and SNP in our array genotyped data set was low. The frequency spectra of CNVs and SNPs were discordant: CNVs were skewed towards rare variants, whereas SNPs were enriched for common variants. Thus, only a small fraction of CNVs, with high allele frequencies, was tagged by SNPs in the EuroGenomics array. Furthermore, the generally low LD in CNV-SNP pairs compared to SNP-SNP pairs confirmed previous findings based on the Illumina BovineHD Genotyping BeadChip data (Figure 2; Lee et al. 2020).

Taking these results together, we conclude that most CNVs have tagging SNPs, which can be identified in deep WGS data. However, such tagging SNPs may not be present in genotyping arrays, as arrays tend to harbour common SNPs. This, in turn, suggests that a small fraction of CNVs, segregating at high frequency, may be tagged well by SNPs in the genotyping arrays. However, most CNVs are rare, and hence not properly captured with the set of SNPs in genotyping arrays. To mitigate this gap, we propose to utilize the WGS CNV catalogue to prioritize likely functional CNVs, particularly rare ones, and directly genotype them.

Funding
This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri and Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The HF WGS data set was funded by the DAMONA ERC advanced grant to MG.

References
Auwera G.A. Van der, Carneiro M.O., Hartl C., Poplin R., Angel G. del, *et al.* 2013 Curr Protoc Bioinformatics 11(1110): 11.10.1–11.10.33. doi:10.1002/0471250953.bi1110s43
Boichard D., Boussaha M., Capitan A., Rocha D., Sanchez M.P. *et al*. 2018. In 11th World Congress on Genetics Applied to Livestock Production, 11(675) pp. 1–6.
Chiang C., Scott A.J., Davis J.R., Tsang E.K., Li X. *et al.* 2017. Nat Genet 49: 692–699. Doi: 10.1038/ng.3834
DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R. *et al.* 2011. Nat Genet 43: 491–8. http://dx.doi.org/10.1038/ng.806.
Derks M.F.L., Lopes M.S., Bosse M., Madsen O., Dibbits B. *et al.* 2018. PLoS Genet 14: 1–20. http://dx.doi.org/10.1371/journal.pgen.1007661.
Kadri N.K., Sahana G., Charlier C., Iso-Touru T., Guldbrandtsen B. *et al.* 2014. PLoS Genet 10: 1-11 doi:10.1371/journal.pgen.1004049
Lee Y.-L., Bosse M., Mullaart E., Groenen M.A.M., Veerkamp R.F. *et al.* 2020. BMC Genomics 21: 1–15. doi: 10.1186/s12864-020-6496-1
Lee Y.-L., Takeda H., Moreira G.C.M., Karim L., Mullaart E. *et al.* 2021. PLoS Genet 17: 1–27. http://dx.doi.org/10.1371/journal.pgen.1009331.
Li H. 2013. arXiv Prepr arXiv 00: 3. http://arxiv.org/abs/1303.3997.
McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K. *et al.* 2010. Genome Res 20: 1297–1303. doi:10.1101/gr.107524.110
Pedersen B.S. andQuinlan A.R.. 2019. Gigascience 1–5. doi:10.1093/gigascience/giz040
Purcell S., Neale B., Todd-brown K., Thomas L., Ferreira M.A.R. *et al.* 2007. Am J Hum Genet 81: 559–575. doi:10.1086/519795
Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G. *et al.* 2020. Gigascience 9: 1–9. doi:10.1093/gigascience/giaa021
Scott A.J., Chiang C., Hall I.M.. 2021. Genome Res. 2021. 31: 2249-2257. doi:10.1101/gr.275488.121
Sudmant P.H., Rausch T., Gardner E.J., Handsaker R.E., Abyzov A. *et al.* 2015. Nature 526: 75–81. doi:10.1038/nature15394