Segregation of the structural variant responds for mastitis resistance in re-sequencing Nordic Holstein animals

Z. Cai[1*] and G. Sahana[1]

[1] Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark; [*]zexi.cai@qgg.au.dk

Abstract

An important source of genome sequence variations, structural variants (SVs), are largely ignored due to the difficulty to precisely calling these variants. In this study, we used whole genome sequence (WGS) data to capture the SVs in various breeds. We have identified 84,561 SVs in 567 WGS animals. When compared the called SVs with the major quantitative trait locus for mastitis resistance on chromosome 6 near to *GC* gene, we found a duplication (DUP) is highly similar to previous reported copy number variant. Even though the linkage disequilibrium between the lead SNP for mastitis resistance and the called DUP is not complete, the SVs profile we obtained still showed the potential for identifying functional variants.

Introduction

Structural variants (SVs) are referred to the variants which are large in size (>50bp). The type of SVs include insertion (INS), deletion (DEL), duplication (DUP), copy number variants (CNV), inversion (INV), and translocations. Generally, such variants are difficult to capture using short-read sequencing due to the length of the sequence reads ranging 30bp-250bp. Therefore, SVs are not a typically included when using whole-genome sequence (WGS) data to obtain high density variants map. Recently, more and more reports have proven SVs are important variants that usually have high impact on the phenotypes (Mahmoud et al., 2019) and have the potential to pinpoint the candidate genes in the GWAS hits. Therefore, calling for SVs for reference panel will be informative. Even though using short-read sequencing with moderate coverage to identify SVs is less powerful than long read sequencing and deep sequencing, the choice of the method could partially overcome the disadvantage of the short reads and achieve high accuracy of SVs calling.

In this study, we collected previously generated WGS dataset with moderate sequencing depth to obtain the SVs in various cattle breeds. Smoove (https://github.com/brentp/smoove) pipeline was used to call SVs, which perform the calling, filtering and annotation of SVs. The potential present of SVs around the known mastitis resistance quantitative trait locus (QTL) on chromosome 6 in Holstein cattle was investigated.

Materials & Methods

Animals. In total, WGS from 567 animals were used in this study, which include 123 Holstein (HOL), 60 Jersey (JER), 175 Nordic Red Dairy cattle (RDC) and 209 from other breeds. DNA was extracted from semen samples using standard procedures at Aarhus University, Denmark. Sequencing was done using Illumina sequencing platform using shot-gun pair-end strategy with approximated 10 fold coverage.

Mapping and Structural variants calling. The raw reads were subjected to Trimmomatic 0.38 (Bolger et al., 2014) to remove the adapter sequence and trim the low quality base. Then the clean reads were mapped to the cattle reference genome ARS-UCD1.2 (Rosen et al.,

2020) using *bwa mem* (Ma et al., 2020) with parameter 'M' to mark shorter split hits as secondary. BaseRecalibrator was applied to the raw alignment bam files using GATK 3.8 (McKenna et al., 2010). We applied the smoove (https://github.com/brentp/smoove) pipeline to call SVs using the base recalibrated bam files with default parameter. The statistics of the SVs in Holstein, Jersey and Nordic Red was calculated using PLINK v1.9 (Purcell et al., 2007).

Locuszoom plot. We reused previous GWAS result (Cai et al., 2018) of mastitis resistance in Danish Holstein as an example to show the potential function of structural variants. We applied the locuszoom (Pruim et al., 2010) to visualize the surrounding region of the lead SNP. The regulatory elements were retrieved from previous study (Kern et al., 2021) and severed as additional track.

Results and discussion

The statistics of the structural variants identified.

We have identified 84,561 SVs, including 9,416 DUP, 59,709 DEL, 3,982 INV, 7,471 pairs of generic breakends (BND) across different chromosomes, and 3,983 BNDs within the same chromosome which could indicated large deletions. Furthermore, there are 66,031, 51,695 and 79,221 segregating SVs in HOL, JER and RDC respectively (Table 1). In addition, 45,894 SVs are shared by these three breeds, and RDC have largest number of breed-specific SVs (Figure 1). Besides, HOL and RDC also share a large number of SVs that are not presented in JER. The number of the SVs is affected by the number of samples and also the population, which is consistent with previous study (Kommadath et al., 2019).

Table 1. The statistics of structural variants in Holstein, Jersey and Nordic Red cattle.

Breed	Number				Total
	DUP	DEL	INV	BND	
Holstein	6,449	39,619	3,192	16,771	66,031
Jersey	4,747	31,486	2,574	12,888	51,695
Nordic Red	7,412	49,238	3,595	18,976	79,221

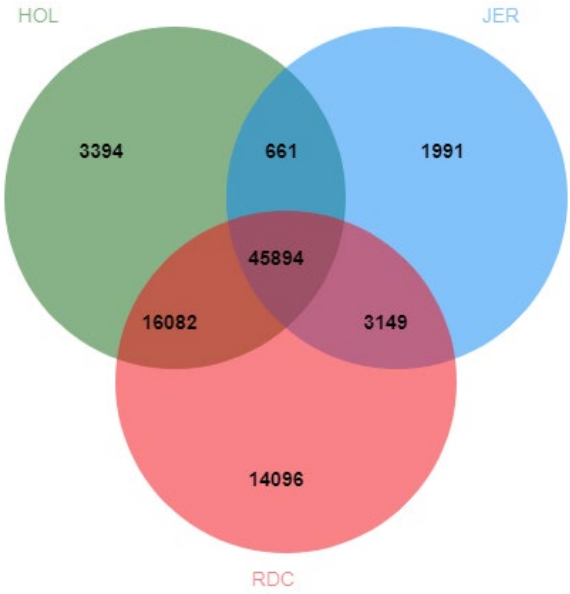

Figure 1. The venn plot shows the overlap of structural variants in Holstein (HOL), Jersey (JER) and Nordic Red Dairy cattle (RDC)

The QTL of mastitis resistance of chromosome 6.

Previous studies have shown that there is a major QTL located on chromosome 6 near to *GC* gene for mastitis resistance. In our GWAS, we identified a SNP 6:86997761 with $-\log_{10}$ (p value) equal to 10.51 as the lead SNP (Figure 2). In the SVs call, we identify 105 SVs located within the surrounding 1 Mb region (1 Mb up and 1 Mb down) of the lead SNP. We extracted the genotype of the lead SNP from the WGS animals and checking the LD pattern with the SVs. We only find 5 SVs are in LD with the lead SNP but none of them was in complete LD. These are a 863 bp DEL ($r^2 = 0.22$) start from 87,034,646, a 2,013 bp DEL ($r^2 = 0.53$) start from 87,065,781, a 272 bp DEL ($r^2 = 0.54$) start from 86,997,761, an 11,781 DUP ($r^2 = 0.52$) start from 86,949,652 and an 11,236 bp DUP ($r^2 = 0.46$) start from 86,950,196.

A previous study has identified a CNV ranging from 86,949,653 ~ 86,961,428 as a potential candidate mutation for the mastitis QTL (Lee et al., 2021). In our case, we have a similar DUP ranging from 86,949,652 ~ 86,961,433. The annotation from smoove showed the start and end point could be imprecise, so it is highly possible these two mutations are the same or the region undergo similar evolutionary event. The flanking area of the lead SNPs and these five SVs are overlapped with active regulatory elements (Figure 2). Even though, the available evidence did not confirm the causality, the SVs profile show the potential for wider scanning for causal variants for QTL whole genome wide and their functional validation.

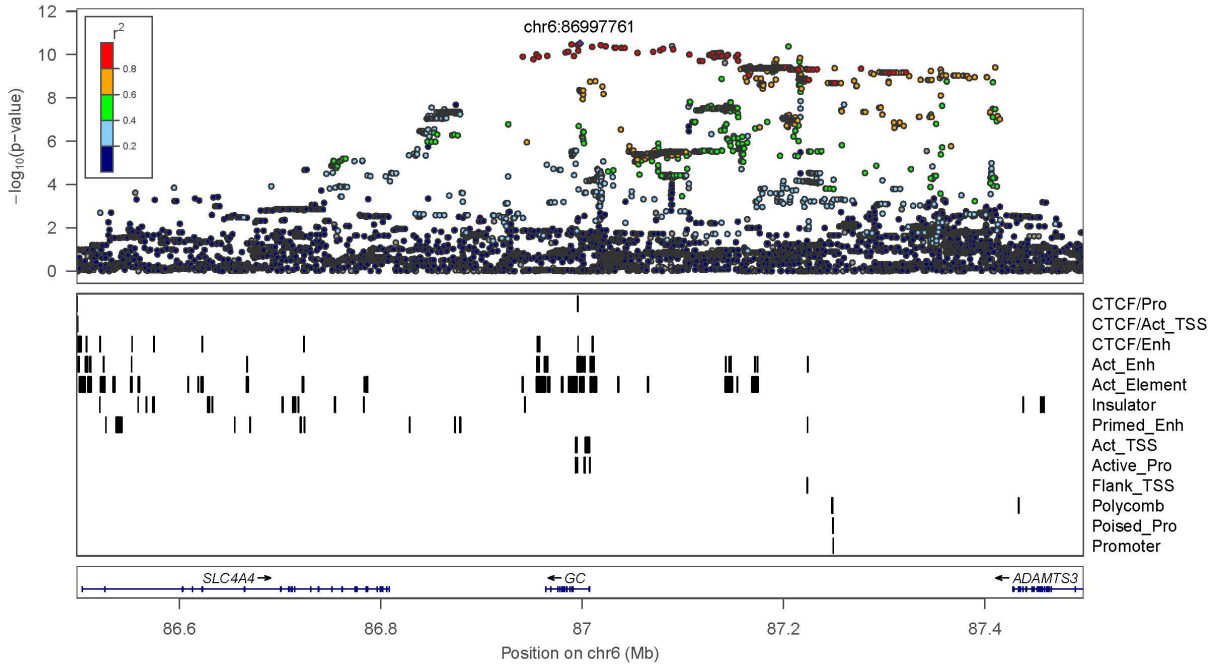

Figure 2. Locuszoom plot showed the lead SNP, LD pattern of the flanking SNPs, the flanking genes (with exon) and the predicted regulatory element from liver tissue.

Acknowledgements

We are grateful to the Nordic Cattle Genetic Evaluation (NAV, Aarhus, Denmark) for providing the phenotypic data used in this study and Viking Genetics (Randers, Denmark) for providing samples for genotyping.

References

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Bioinformatics 30(15):2114-2120. doi:10.1093/bioinformatics/btu170.

Cai, Z., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018. BMC Genomics 19(1):656. doi:10.1186/s12864-018-5050-x.

Kern, C., Y. Wang, X. Xu, Z. Pan, M. Halstead, et al. 2021. Nature Communications 12(1):1821. doi:10.1038/s41467-021-22100-8.

Kommadath, A., J. R. Grant, K. Krivushin, A. M. Butty, C. F. Baes, et al. 2019. GigaScience 8(6). doi:10.1093/gigascience/giz073.

Lee, Y. L., H. Takeda, G. Costa Monteiro Moreira, L. Karim, E. Mullaart, et al. 2021. PLoS Genet 17(7):e1009331. doi:10.1371/journal.pgen.1009331.

Ma, K. C., T. D. Mortimer, M. A. Duckett, A. L. Hicks, N. E. Wheeler, et al. 2020. arXiv preprint arXiv:1303.3997. doi:10.1101/2020.03.24.006650.

Mahmoud, M., N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, et al. 2019. Genome Biology 20(1):246. doi:10.1186/s13059-019-1828-7.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, et al. 2010. Genome research 20(9):1297-1303.

Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, et al. 2010. Bioinformatics 26(18):2336-2337. doi:10.1093/bioinformatics/btq419.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, et al. 2007. The American journal of human genetics 81(3):559-575.

Rosen, B. D., D. M. Bickhart, R. D. Schnabel, S. Koren, C. G. Elsik, et al. 2020. Gigascience 9(3):giaa021. doi:10.1093/gigascience/giaa021.