Report for the Meeting of the Dairy Cattle Breeding & Breeding committee – Guelph Oct 2017
Submitted to the World Congress on Genetics Applied to Livestock Production – NewZeeland Feb 2018
In preparation for submission in Journal of BMC Genomics

Discovery of expression quantitative trait loci associated with Johne's disease using both RNA-seq and DNA variants

N. Bissonnette[1], J.-S. Brouard[1], O. Ariel[1], N. Gévry[2], and F. Miglior[3,4]

[1] Sherbrooke Research and Development Centre, Agriculture and Agri-Food Canada, Sherbrooke, Qc, Canada J1M 0C8
Nathalie.Bissonnette@canada.ca (Corresponding Author)
[2] Département de Biologie, Université de Sherbrooke, Qc, Canada J1K 2R1
[3] Center of Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada N1G 2W1
[4] Canadian Dairy Network, Guelph, ON, Canada N1K 1E5

Summary

Johne's disease (**JD**) is a debilitating chronic disease in ruminants caused by Mycobacterium avium ssp. paratuberculosis (**MAP**) which requires the action of gut macrophages for its survival and dissemination. The endemic situation of JD can be in part explained by the lack of genetic resistance to MAP infection in cattle populations. A successful breeding strategy results from a comprehensive understanding of the genetic variability and its link with the biological pathways that impact disease susceptibility/resistance. In this functional genomics study, accurate phenotypic data (e.g. diagnosis records) for JD were used to discriminate 31 MAP-infected (JD(+)) from 30 healthy/resistant (JD(–)) cows. Transcriptomes of primary macrophages from both groups were analysed using the next-generation RNA sequencing (**RNA-Seq**) technology to study genetic variations in susceptible cows that allow MAP to proliferate and escape the normal mycobacterial killing process. DNA genotypes were also identified using two complementary strategies: BovineSNP50 DNAchip and genotyping-by-sequencing methods. More than 87% of the 861K variants found were identified by RNA-seq which identified 99% of the novel discovered single nucleotide variants. Genome-wide association study identified numerous expression quantitative trait loci (**eQTL**) on BTA3, 4, 7, 8, 9, 10, 11, 12, 15, 18, 19, 21, 25, and BTA28, and on mitochondrial genome at $P \leq 0.005$. RNA-Seq is an effective strategy to identify eQTL and thus increases the power to detect functional genetic variants. In the present study, we succeeded to identify eQTL and the regulatory pathways that discriminate MAP infected from healthy/resistant cows. This information is relevant in genetic selection, as it may reduce disease susceptibility. An integration of the findings of this research (genomic information), into the conventional young sire selection and progeny testing program could yield a better, more accurate and rapid genetic improvement of resistance to bovine paratuberculosis in Canadian dairy herds.
Keywords: bovine paratuberculosis, GWAS, RNA-sequencing, expression quantitative trait locus, primary macrophage

Introduction

Johne's disease (**JD**) is a livestock disease caused by a zoonotic pathogen, leading to chronic diarrhoea and ill thrift in adult cattle. This is a slow progressive disease with unpredictable clinical signs, making it difficult to identify infected cows. Diagnosis of the disease is challenging (Fock-Chow-Tho et al., 2017) and vaccines are ineffective in preventing infection (Bannantine et al., 2014, Ghosh et al., 2015). The prevalence of infected farms has been increasing worldwide, and JD (paratuberculosis) is now a global concern.

Enhancing the animals' natural disease resistance by improving the genetics is slow but the results are permanent. With the availability of the full bovine genomic sequence, numerous SNPs are available. A number of genome-wide association studies (**GWAS**) have been carried out to identify quantitative trait loci (**QTL**) associated with JD (van Hulzen et al., 2012, Alpay et al., 2014, Kupper et al., 2014, Zare et al., 2014, Kiser et al., 2017, Sallam et al., 2017). These studies had found evidence for association on multiple and varying chromosomal locations. While numerous QTL were identified, we still have the daunting task of predicting 'genotype-to-phenotype' relationships mainly because markers for many of the traits of importance still explain a relatively small proportion of the JD resistance/susceptibility variance.

The JD causative pathogen, *Mycobacteria avium* spp. *paratuberculosis* (MAP), is an obligatory mycobacteria requisitioning macrophage for its multiplication and survival. The hypothesis of the present study is that genetic variations affecting macrophage activity in susceptible cows would allow MAP to proliferate and escape the normal mycobacterial killing process. Under the assumption that a better understanding of the transcriptome can lead to the identification of functional genetic variants, the primary goal of our study was to use JD case and control macrophages for identifying biomarkers associated with JD.

The association between a genetic variant at a genomic locus and a trait is not directly informative but expressed QTL is informative with respect to the mechanism whereby the variant influence the phenotype. Nowadays, RNA-sequencing (**RNA-seq**) strategy is becoming increasingly affordable and increases the power to detect subtle pathway activity changes. The originality of this research is to perform a GWAS using RNA-Seq and DNA genotypes to provide high-resolution genomic analysis. These findings provide a comprehensive strategy to unravel the relationship between genotype and phenotype in the context of host-pathogen interaction.

Materials and Methods

Animal selection, JD diagnosis, and Differentiation of monocyte-derived macrophages

Diagnosis of the commercial dairy farms positive for JD was performed as described (Fock-Chow-Tho et al., 2017). Twelve cows analyzed previously (Ariel et al., 2017) and 49 additional cows were selected for the RNA-seq analysis. The cows were divided into two groups: 30 JD negative [JD(–)] cows tested using serum ELISA and fecal PCR or culture and 31 JD positive [JD(+)] confirmed using both tests. Monocyte isolation, differentiation, macrophage cultured and MAP infection were previously described (Ariel et al., 2017).

RNA-seq data processing and bioinformatics analysis

Since 6 distinct libraries (4 infection time points and 2 controls) were sequenced for each of the 12 cows previously analyzed (Ariel et al., 2017), sequence data were pooled for each cow containing roughly 360 million paired-end (PE) reads per cow. Sequence files from the 49 additional cows contained a minimum of 60 million PE reads. Variants were identified from RNA-seq data using the Genome Analysis Toolkit (GATK, (McKenna et al., 2010)). All analysis were runned on the MP2 computing cluster of Compute Canada at the Université de Sherbrooke. Reads were firstly mapped against the UMD_3.1.1/BosTau8 assembly using the STAR aligner (version 2.4.0j, (Dobin et al., 2013). Next, duplicated reads were marked with Picard tools. The recommended Split'N'Trim and indel realigments steps were also performed (GATK, version 3.3-0-g37228af). At the variant calling step, we moved to the joint variant discovery workflow (GVCF mode). As recommend by the GATK BestPractices, we used the HaplotypeCaller algorithm to call variants.

DNA genotyping using Bovine SNP50 BeadChip and Genotyping-by-sequencing

From the 61 individuals analyzed by RNA-seq, 51 were also genotyped using blood cells' DNA. Two DNA genotyping methods were used: the commercial Illumina BovineSNP50 BeadChip (Zoetis, Kalamazoo, MI) and the Genotyping-By-Sequencing technique. Bioinformatic treatment of the SNP50 genotyping data has been previously described (Brouard et al., 2017). GBS sequence data (100-nt fastq files) were processed using the Fast-GBS pipeline (Torkamaneh and Belzile, 2015). The variants were called with PLATYPUS (Rimmer et al., 2014) if they had a minimum read depth of 2 and a minimum genotype quality score of 5.

Variant filtering, imputation, annotation, and functional analysis

In all cases, variants were retained if they mapped to the 29 bovine autosomes, the mitochondrial genome or the X chromosome. Note that RNA-Seq and GBS genotypes supported by read depth values < 4 or genotype quality score < 20 were filtered out (i.e. changed to missing data). Other variants were eliminated from downstream analysis : those with Minor allele frequencies (MAF) < 0.025, those with a Call rate < 0.2, those harboring two or more alternative alleles and those that are out of Hardy-Weinberg equilibrium (values < 0.001). SVS (Golden Helix) and the BCFtools (Li et al., 2009) were used to obtain high quality variants. To prepare the final dataset used for GWAS, variants were combined according to the procedure summarized in Supplementary Figure 1. Imputation of the missing data in the Single Nucleotide Variant merged dataset was performed using FImpute v2.2 (Sargolzaei et al., 2014). Known variants detected in our study were annotated with a bovine reference VCF file (SNPdb-version 150) downloaded from the NCBI ftp site.

The software program SnpEff (Cingolani et al., 2012) was used to predict the functional or biologically interpretable pathways and networks associated with significant variants. Functional analysis (pathways and networks) was performed using BovineMINE (Elsik et al., 2016). Additional information can be found in Supplementary Material and Methods.

Results and Discussion

RNA-Seq and DNA-derived variants

For each cow, the data from uninfected macrophages and *in vitro* MAP-infected macrophages collected at different post-infection (pi) time points (1 hpi, 4 hpi, 8 hpi, and 24 hpi) were combined. Overall, we sequenced the whole primary macrophage transcriptome from 30 JD(–) and 31 JD(+) cows. Among them, 56 cows were at the same time genotyped using 2 other methods: genotyping-by-sequencing (GBS) and Illumina's BovineSNP50 BeadChip (SNP50). Using the approach summarized in Supplementary Figure 1, we identified 861,423 single nucleotide variants (SNV), including 763,722 RNA-seq, 58,939 GBS, and 46,890 SNP50 variants. Among them, 107,781 were indels, essentially detected by RNA-seq. By comparing the corresponding genotypes from markers shared by two genotyping methods (e.g. RNA-seq and SNP50), we validated the accuracy of RNA-Seq and GBS variant calls. The common SNV identified by the respective method are shown in Figure 1. A larger number of variants were identified by RNA-seq (87.8%). Interestingly, a larger amount of SNV identified by GBS (4,223) were found in the RNA-Seq dataset compared to the 3,586 SNP50 markers.

Among the SNV identified, we obtained 20.6% of novel SNV. Most of the 177,447 novel sites (99.8%) were identified by RNA-seq. Only 335 and 248 SNP were identified by GBS and SNP50, respectively.

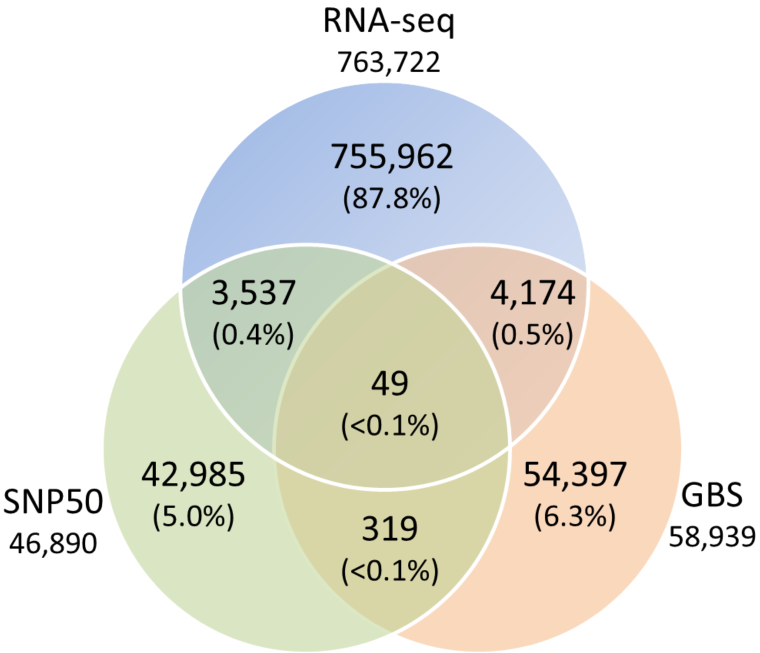


Figure 1. Comparison of SNVs identified in the transcriptomes of GSE98363 (n=12) and the transcriptome (RNA-seq) of 49 additional cow's derived macrophages, and SNVs identified in the genome of the common individuals (n=61). The percentage of SNVs over the total number of identified variants is also indicated in parentheses.

Enrichment of variants in Functional Categories

To predict their effect (i.e. regulatory functions on known genes), the variants (SNP and indel) were annotated using SNPeff. The level of impact (high, low, or moderate effect) obtained using the respective genotyping methods is given in Table 1 and also described in Supplementary Figure 2. Interestingly, although both DNA genotyping methods identified similar amount ~50K of genotypes, the GBS identified almost twice the number of SNP having a predicted functional effect. While the distribution of SNP at different positions is similar for both DNA genotyping methods (Supplementary Figure 2), the mutational profile of SNV is different (Supplementary Table 1).

Table 1. Summary statistics of the identified variants using the respective methods.

Genotyping methods (counts)	RNA-seq	GBS	SNP50
Variants processed			
SNP	763,722	58,939	46,890
Insertions	55,819	1,057	0
Deletions	48,246	2,230	0
Effects by impact			
high	8,100	14	6
low	31,400	946	624
moderate	25,468	499	261

modifier	989,354	63,771	52,740

[1]Genetic variant annotation and effect prediction was performed using SNPEff

Genome Wide Association Study (GWAS) and pathway analysis of the significant SNV

Genome-wide association analysis was performed using the 857,707 SNV called from the 61 cows which includes the imputed RNA-seq, GBS, and SNP50 datasets from 30 JD(–) and 31 JD(+) cows. Numerous eQTL were identified on BTA3, 4, 7, 8, 9, 10, 11, 12, 15, 18, 19, 21, 25, and BTA28 and on mitochondrial genome (Figure 2). Analysis for identifying functional interpretable network biomarkers is necessarily more difficult for complex traits. However GWAS empowered significant molecular biomarkers identification enabling the association of pathways in the context of JD resistance/susceptibility. Network and pathway analysis using BovineMine from JD(+/–) macrophages reveals interesting cue regarding pathways that deserved investigation (e.g. STAT transcription factor).

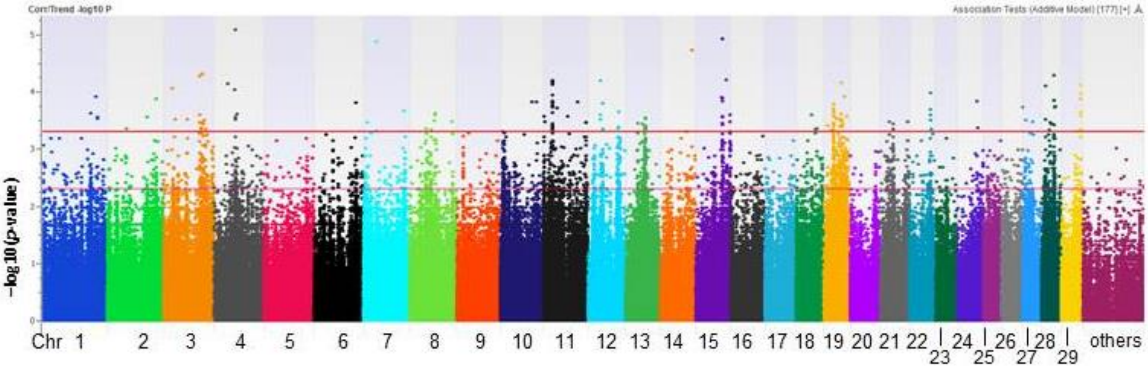


Figure 2. Manhattan plot of SNP associated with Johne's disease.
The –log$_{10}$ of the P-value is plotted.

Conclusion

In the present study, we have identified functional biomarkers associated to JD that potentially explain the presence of putative dysfunctional allele(s) of these specialized phagocyte immune cells, which allow MAP to niche and proliferate. These functional SNP are part eQTL. Information on favorable or detrimental eQTLs will eventually be used in genetic selection allowing beneficial allele (s) to disseminate in the progeny which would be able to efficiently prime host immune reaction. The originality of this research was to use RNA-Seq to provide high-resolution genomic analysis of macrophages for identifying novel mutations and transcripts associated with bovine paratuberculosis.

Funding

This study was funded by a Collaborative Research and Development grant from the Dairy Cattle Genetics Research and Development (DairyGen) Council of Canada, the Natural Sciences and Engineering Research Council of Canada (NSERC) (Project CRDPJ 441719), and Agri-Food and Agriculture Canada (Project AAFC J0000-79). Special thanks are due to the veterinarians and dairy farmers for their valuable collaboration.

References

Alpay, F., Y. Zare, M. H. Kamalludin, X. Huang, X. Shi, G. E. Shook, M. T. Collins, and B. W. Kirkpatrick. 2014. Genome-wide association study of susceptibility to infection by Mycobacterium avium subspecies paratuberculosis in Holstein cattle. PLoS One 9(12):e111704.

Ariel, O., N. Bisseonnette, A. E. Ibeagha, G. Fecteau, and N. Gévry. 2017. Mycobacterium avium ssp. paratuberculosis Infection Induces an Immune Tolerance Phenotype in Primary Bovine Macrophages from Johne's Disease Cows. In preparation.

Bannantine, J. P., M. E. Hines, 2nd, L. E. Bermudez, A. M. Talaat, S. Sreevatsan, J. R. Stabel, Y. F. Chang, P. M. Coussens, R. G. Barletta, W. C. Davis, D. M. Collins, Y. T. Grohn, and V. Kapur. 2014. A rational framework for evaluating the next generation of vaccines against Mycobacterium avium subspecies paratuberculosis. Frontiers in cellular and infection microbiology 4:126.

Brouard, J. S., B. Boyle, E. M. Ibeagha-Awemu, and N. Bissonnette. 2017. Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC Genet 18(1):32.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80-92.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15-21.

Elsik, C. G., D. R. Unni, C. M. Diesh, A. Tayal, M. L. Emery, H. N. Nguyen, and D. E. Hagen. 2016. Bovine Genome Database: new tools for gleaning function from the Bos taurus genome. Nucleic Acids Res 44(D1):D834-839.

Fock-Chow-Tho, D., E. Topp, E. A. Ibeagha-Awemu, and N. Bissonnette. 2017. Comparison of commercial DNA extraction kits and quantitative PCR systems for better sensitivity in detecting the causative agent of paratuberculosis in dairy cow fecal samples. J Dairy Sci 100(1):572-581.

Ghosh, P., D. C. Shippy, and A. M. Talaat. 2015. Superior protection elicited by live-attenuated vaccines in the murine model of paratuberculosis. Vaccine 33(51):7262-7270.

Kiser, J. N., S. N. White, K. A. Johnson, J. L. Hoff, J. F. Taylor, and H. L. Neibergs. 2017. Identification of loci associated with susceptibility to Mycobacterium avium subspecies paratuberculosis (Map) tissue infection in cattle. J. Anim. Sci. 95(3):1080-1091.

Kupper, J., H. Brandt, K. Donat, and G. Erhardt. 2014. Phenotype definition is a main point in genome-wide association studies for bovine Mycobacterium avium ssp. paratuberculosis infection status. Animal 8(10):1586-1593.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297-1303.

Rimmer, A., H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, W. G. S. Consortium, A. O. M. Wilkie, G. McVean, and G. Lunter. 2014. Integrating mapping-, assembly- and haplotypebased approaches for calling variants in clinical sequencing applications. Nature genetics 46(8):912-918.

Sallam, A. M., Y. Zare, F. Alpay, G. E. Shook, M. T. Collins, S. Alsheikh, M. Sharaby, and B. W. Kirkpatrick. 2017. An across-breed genome wide association analysis of susceptibility to paratuberculosis in dairy cattle. J Dairy Res 84(1):61-67.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478.

Torkamaneh, D. and F. Belzile. 2015. Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. PLoS One 10(7):e0131533.

van Hulzen, K. J. E., G. C. B. Schopen, J. A. M. van Arendonk, M. Nielen, A. P. Koets, C. Schrooten, and H. C. M. Heuven. 2012. Genome-wide association study to identify chromosomal regions associated with antibody response to Mycobacterium avium subspecies paratuberculosis in milk of Dutch Holstein-Friesians. J. Dairy Sci. 95(5):2740-2748.

Zare, Y., G. E. Shook, M. T. Collins, and B. W. Kirkpatrick. 2014. Genome-Wide Association Analysis and Genomic Prediction of Mycobacterium avium Subspecies paratuberculosis Infection in US Jersey Cattle. PLoS One 9(2):e88380.

Report for the Meeting of the Dairy Cattle Breeding & Breeding committee – Guelph Oct 2017
Submitted to the World Congress on Genetics Applied to Livestock Production – NewZeeland Feb 2018
In preparation for submission in Journal of BMC Genomics

Supplementary Tables

Supplementary Table 1 Summary statistics[1] of the identified variants using the respective methods

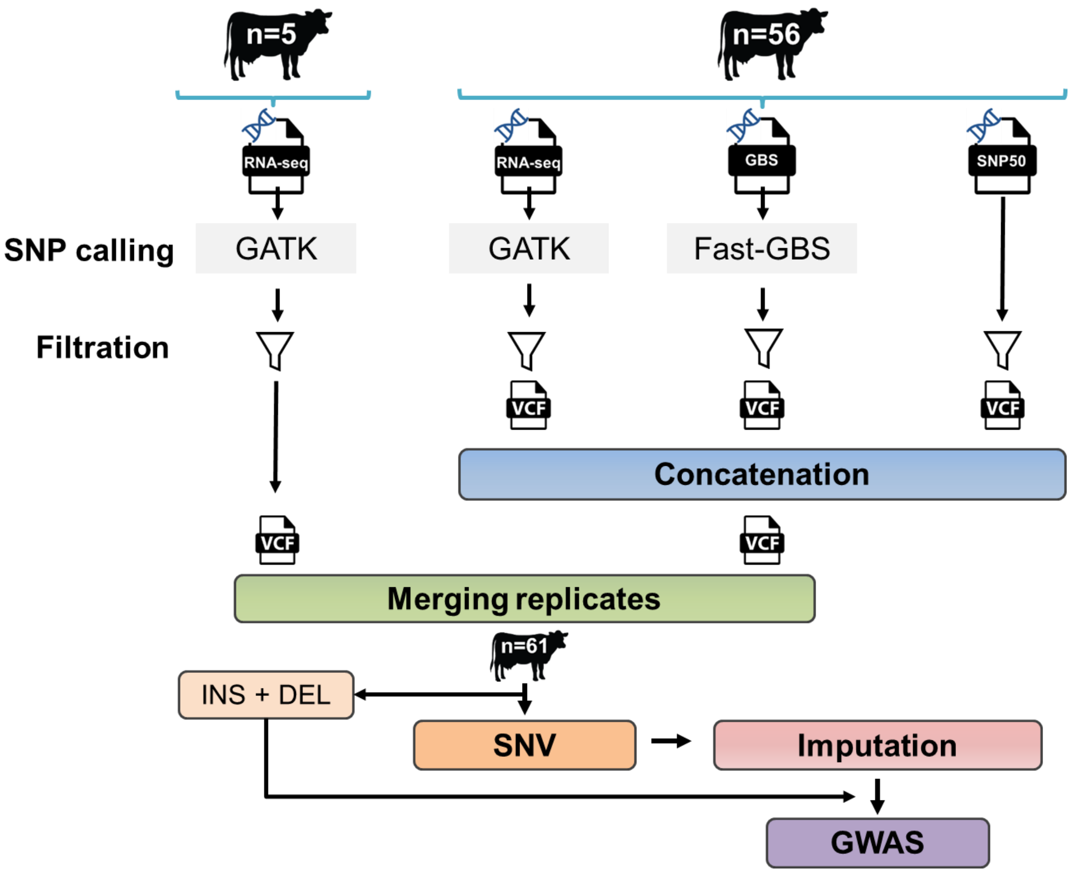
Genotyping methods (counts)	RNA-seq	GBS	SNP50
Variants processed			
SNPs	659,667	55,004	46,890
Insertions	55,819	1,705	0
Deletions	48,246	2,230	0
Categories			
MISSENSE	23,550	500	262
NONSENSE	331	3	
SILENT	28,884	847	578
Effects by impact			
HIGH	8,100	14	6
LOW	31,400	946	624
MODERATE	25,468	499	261
MODIFIER	989,354	63,771	52,740
Effects by type and region			
3_prime_UTR_variant	25,230	260	309
5_prime_UTR_premature_start_codon	527	15	8
5_prime_UTR_truncation	1		
5_prime_UTR_variant	4,286	64	47
bidirectional_gene_fusion	1		
conservative_inframe_deletion	247		
conservative_inframe_insertion	816		
disruptive_inframe_deletion	522		
disruptive_inframe_insertion	593		
downstream_gene_variant	152,781	4,127	2,505
exon_loss_variant	1		
frameshift_variant	6,967		
gene_fusion	1		
initiator_codon_variant	3		
intergenic_region	227,759	35,547	31,699
intragenic_variant			
intron_variant	50,691	20,265	15,843
missense_variant	23,508	499	261

non_coding_transcript_exon_variant	1,450	49	34
non_coding_transcript_variant	20		
splice_acceptor_variant	574	1	2
splice_donor_variant	532	9	3
splice_region_variant	3,426	105	59
start_lost	33	1	1
stop_gained	767	3	
stop_lost			
stop_retained_variant	26		2
synonymous_variant	28,857	847	576
upstream_gene_variant	73,623	3,547	2,345

[1] Genetic variant annotation and effect prediction was performed using SNPEff

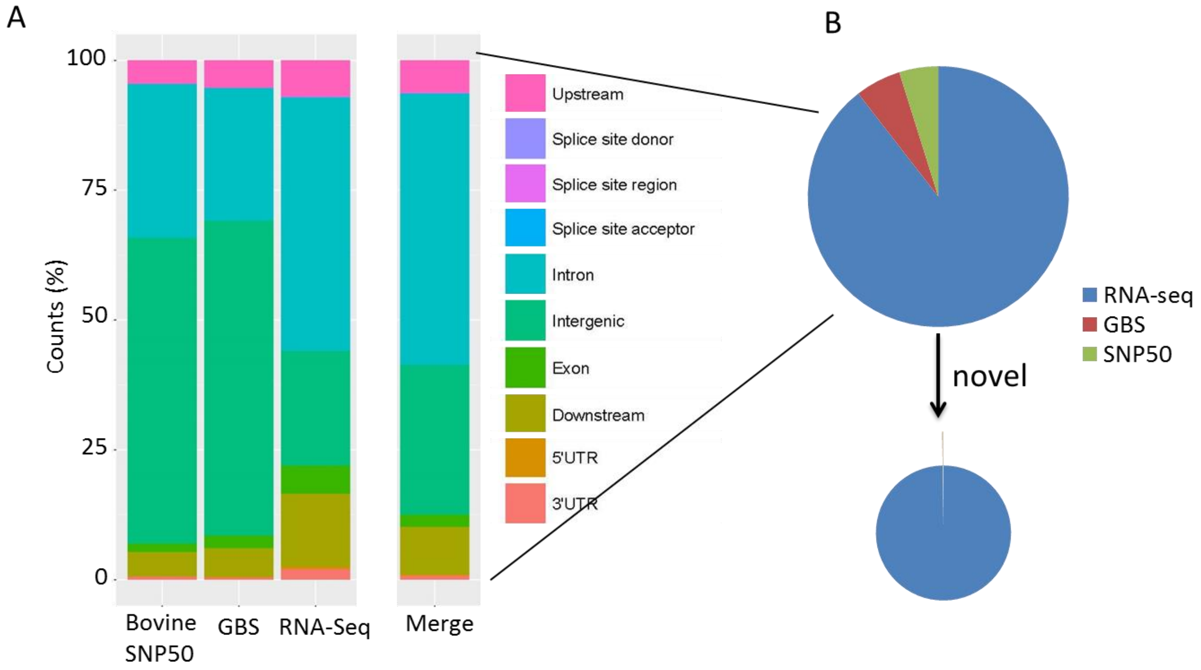
Supplementary Figures

Supplementary Figure 1



Supplementary Figure 1. Summary of the bioinformatics treatment of the RNA-seq and DNA genotypes.

Supplementary Figure 2



Supplementary Figure 2. Characteristics of SNP identified from Bovine SNP50, GBS, and RNA-Seq datasets. (A) The genomic distribution of SNV identified using the respective genotyping methods suggests a high enrichment of RNA-seq SNV in introns and exons. The total 857,707 SNV is merged. (B) Novel SNV. Among the 865,618 SNV, 177,447 novel sites were found, which were mostly identified by RNA-seq analysis.