



## Development of polymorphic markers in the immune gene complex loci of cattle

K. Bakshy,<sup>1</sup> D. Heimeier,<sup>2</sup> J. C. Schwartz,<sup>2</sup> E. J. Glass,<sup>3</sup> S. Wilkinson,<sup>3</sup> R. A. Skuce,<sup>4</sup> A. R. Allen,<sup>4</sup> J. Young,<sup>1</sup> J. C. McClure,<sup>1</sup> J. B. Cole,<sup>5</sup> D. J. Null,<sup>5</sup> J. A. Hammond,<sup>2</sup> T. P. L. Smith,<sup>6</sup> and D. M. Bickhart<sup>1\*</sup>

<sup>1</sup>Dairy Forage Research Center, USDA-ARS, Madison, WI 53706

<sup>2</sup>The Pirbright Institute, Ash Road, Pirbright, Surrey GU24 0NF, UK

<sup>3</sup>The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush EH25 9RG, Edinburgh, UK

<sup>4</sup>Agri-Food and Biosciences Institute, Stormont, Belfast, Northern Ireland BT4 3SD, UK

<sup>5</sup>Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705

<sup>6</sup>Meat Animal Research Center, USDA-ARS, Clay Center, NE 68933

### ABSTRACT

The addition of cattle health and immunity traits to genomic selection indices holds promise to increase individual animal longevity and productivity, and decrease economic losses from disease. However, highly variable genomic loci that contain multiple immune-related genes were poorly assembled in the first iterations of the cattle reference genome assembly and underrepresented during the development of most commercial genotyping platforms. As a consequence, there is a paucity of genetic markers within these loci that may track haplotypes related to disease susceptibility. By using hierarchical assembly of bacterial artificial chromosome inserts spanning 3 of these immune-related gene regions, we were able to assemble multiple full-length haplotypes of the major histocompatibility complex, the leukocyte receptor complex, and the natural killer cell complex. Using these new assemblies and the recently released ARS-UCD1.2 reference, we aligned whole-genome shotgun reads from 125 sequenced Holstein bulls to discover candidate variants for genetic marker development. We selected 124 SNPs, using heuristic and statistical models to develop a custom genotyping panel. In a proof-of-principle study, we used this custom panel to genotype 1,797 Holstein cows exposed to bovine tuberculosis (bTB) that were the subject of a previous GWAS study using the Illumina BovineHD array. Although we did not identify any significant association of bTB phenotypes with these new genetic markers, 2 markers exhibited substantial effects on bTB phenotypic prediction. The models and parameters trained in this study serve as a guide for

future marker discovery surveys particularly in previously unassembled regions of the cattle genome.

**Key words:** cattle genome reassembly, marker selection, bovine tuberculosis, major histocompatibility complex class

### INTRODUCTION

The selection of sparse maps of genetic variant sites to serve as markers for genomic selection is still a complex task. Originally, variant frequency and spacing in the cattle reference genome were the major criteria used for selecting suitable genetic markers (Matukumalli et al., 2009). Subsequent analysis of sequence variant alleles has used statistical association with phenotypic traits to select sites that show the largest effects on those traits (VanRaden et al., 2017). Both methods have produced genetic marker maps that have been successfully used as a basis for genomic selection in dairy cattle (VanRaden, 2008); however, they are reliant on the accuracy and representative nature of the cattle reference genome. Highly polymorphic and structurally variant regions of the cattle genome, including several that contain genes related to immune responses, could not be accurately assembled using the technologies available at the time (Sanderson et al., 2014; Schwartz et al., 2017). Moreover, the original cattle assembly made use of sequence from 2 different animals, with a minimum tiling path of bacterial artificial chromosome (BAC) clones made from a Line 1 Hereford bull supplemented with approximately 5 to 6× coverage of whole-genome shotgun reads from one of his daughters (Elsik et al., 2009). This approach further constrained the ability to accurately represent haplotypes of the immune complex loci. These polymorphic regions are consequently untracked by markers on the current catalog of commercial cattle genotyping tools and include several

Received October 19, 2020.

Accepted January 18, 2021.

\*Corresponding author: [derek.bickhart@usda.gov](mailto:derek.bickhart@usda.gov)

large immune gene clusters (**IGC**), such as genes in the major histocompatibility complex [**MHC**; on *Bos taurus* chromosome (**bta**) 23 in the 28.3–28.7 megabase pairs (**Mbp**) region], the natural killer complex (**NKC**; bta5: 99.5–99.8 Mbp) and the leukocyte receptor complex (**LRC**; bta 18: 63.1–63.4 Mbp).

There are substantial distances between markers on the Illumina BovineHD array (Matukumalli et al., 2009) that span the NK C (largest gap size: 200 kb), the MHC class I region (largest gap size: 50 kb), and the LRC (largest gap size: 100 kb with many genes missing in the assembly). We previously demonstrated that the LRC and NK C loci in the UMD3.1 reference genome were poorly assembled (Sanderson et al., 2014; Schwartz et al., 2017), which likely contributes to their underrepresentation in genotyping assays. Furthermore, markers were based on coordinates from the UMD3.1 reference genome assembly (Zimin et al., 2009), which only contains a pseudohaploid representation of sequence in these regions that may not reflect the structural polymorphisms of alternative haplotypes. We hypothesized that if alleles of genes, or indeed novel genes, in these regions were involved in animal health traits, their effects could not be assessed unless additional genetic markers were included. The gene-dense, polymorphic, and repetitive nature of these regions suggested that original genetic marker design was limited by the incomplete nature of the UMD3.1 reference genome assembly (Zimin et al., 2009), so we first sought to sequence and assemble haplotypes of these regions to better characterize their genetic content.

Genome assembly methods and techniques have advanced substantially in the time since the release of the first commercial cattle genotyping chips (Bickhart et al., 2017; Rosen et al., 2020), and our hierarchical assembly approach is only one option for future surveys of candidate genetic markers in polymorphic genomic regions. Improvements in assembly algorithms (Koren et al., 2017; Kolmogorov et al., 2019) and decreases in sequencing costs have accelerated the rate at which new genome assemblies can be published for new species or individuals of a species with a reference genome. Furthermore, use of heterozygous parental crosses has been shown to accurately assemble parental haplotypes into contiguous chromosome scaffolds (Koren et al., 2018), thereby providing unprecedented views into the structure of structurally polymorphic regions such as the IGC regions (Low et al., 2020). However, these approaches have substantial logistical prerequisites such as the generation of hybrid offspring from lineages with sufficient sequence divergence. This limits the applicability of such methods for marker discovery, particularly when crossing 2 individuals of a breed that has a low ancestral population size, such as Holstein cattle

(Hayes et al., 2003). Targeted approaches, such as our hierarchical assembly of BAC insert sequence, are the most efficient means of assessing the breadth of diversity of IGC regions. Recent improvements in targeted sequencing, such as ReadFish (Payne et al., 2020), are especially promising; however, we note that such methods will benefit from the use of our assembled contigs in filtering reads belonging to structurally diverse IGC regions such as the MHC locus.

Previous studies have identified several distinct haplotypes of the MHC locus segregating in cattle populations, suggesting that there is substantial genetic diversity in the locus (Codner et al., 2012; Vasoya et al., 2016). These IGC in cattle have fundamental roles in the innate and adaptive immune system, but the extent to which different alleles and haplotypes influence differential outcomes to infection by different pathogens is not yet understood, such as *Mycobacterium bovis* in bovine tuberculosis (**bTB**).

Bovine tuberculosis is a systemic disease that causes severe economic losses to UK (Allen et al., 2018) and, to a lesser extent, US dairy farmers (for a review, see le Roex et al., 2013). The causal agent, *Mycobacterium bovis*, infects susceptible species directly via respiratory aerosols or potentially indirectly via a contaminated environment and establishes the hallmark granulomas in the lung and lymphatic tissue. *Mycobacterium bovis* is difficult to eradicate, as it can infect wildlife reservoirs, such as badgers, brush-tailed possum, and white-tailed deer, that come into contact with domestic cattle (le Roex et al., 2013). Previous surveys on the genetic basis for bTB infection have revealed a heritability for disease incidence (Allen et al., 2010; Bermingham et al., 2011; Raphaka et al., 2017); however, these case-control studies found that individual marker association testing was a poor predictor of case status owing to the likely polygenic nature of bTB resistance. The intracellular nature of the pathogen suggests that resistance to the disease may be influenced by the cytotoxic arm of cellular immunity, namely CD8 T cells and natural killer (**NK**) cells. Cattle are known to have a highly diverse and polymorphic NK cell receptor repertoire and MHC antigen presentation system (Ellis and Hammond, 2014; Sanderson et al., 2014; Allan et al., 2015; Schwartz et al., 2017; Gibson et al., 2020), which makes the genomic regions encoding these genes highly likely to influence variation in disease manifestation. However, the polymorphic nature of these regions also contributed to the aforementioned fact that these regions were misassembled in prior reference assembly versions (Ellis and Hammond, 2014; Sanderson et al., 2014; Allan et al., 2015; Schwartz et al., 2017).

This study sought to identify new genetic markers within 3 IGC fundamental to the recognition and

control of intracellular pathogen infections, the MHC class I, the NKC, and the LRC, using simple scoring metrics and machine learning models. We then created and tested a custom genotype panel that could explain more of the genetic variance in bTB incidence in dairy cattle, to act as a proof-of-principle study for the utility of markers within highly variable immune gene complexes.

## MATERIALS AND METHODS

### Reassembly of IGC Regions and Identification of Candidate Genetic Markers

Those BAC clones that contained inserts relevant for sequencing were identified through comparative alignment of BAC-end sequence reads (GenBank Accessions: AJ698510:AJ698674) to the UMD3.1 reference genome assembly (Zimin et al., 2009). At least one BAC-end read needed to align to previously identified IGC genomic regions for the clone to be selected for follow-up assembly. Using this criterion, 40 clones were selected for targeted resequencing and assembly (Supplemental Table S1, <https://doi.org/10.6084/m9.figshare.14067410.v1>). The BAC clones from the RPCI42 (Holstein) and CHORI240 (Hereford) collections were provided by the CHORI BACPAC service (<https://bacpacresources.org/>), and BAC inserts (average length was approximately 200 kb) were sequenced to an average depth of 40× coverage using a PacBio RS II. Read length N50 values for each library ranged from 9 to 10 kb. Reads were assembled into contigs via smrtanalysis in smrtportal v1.3 software (<https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/>) using default settings and 200 kb as expected genome size. Contigs were polished using Quiver (version packaged in smrtportal v1.3). Contigs from the same IGC regions were compared using minimap2 (Li, 2016) alignments and equivalent regions on the ARS-UCD1.2 (Rosen et al., 2020) reference assembly as determined from those same alignments. Contigs that had >95% nucleotide identity to another contig, or the haplotype on the ARS-UCD1.2 reference, were considered redundant and were removed from subsequent alignment and analysis. Assembled, nonredundant contigs can be downloaded from NCBI GenBank BankIt (MT145922-MT145940 accessions).

### Variant Discovery and Initial Marker Selection

To discover variant sites on these contigs that were segregating in the Holstein breed, we generated paired-end Illumina sequence reads using a HiSeq X sequencer

from 125 Holstein bulls that were predicted to contain novel haplotypes exclusive to each other by our inverse weight selection algorithm (Bickhart et al., 2016) as assessed by SNP genotype data. Sequence data were provided by the Cooperative Dairy Cattle DNA Repository (available for research projects on request). We aligned these paired-end reads to a concatenated reference consisting of the ARS-UCD1.2 assembly (Rosen et al., 2020) and our assembled IGC contigs. The new contigs containing novel IGC haplotypes were added as unplaced scaffolds at the end of the ARS-UCD1.2 assembly before alignment to avoid ambiguous alignment of sequence reads from homologous regions of the reference. Alignments were performed using BWA MEM (version 0.7.17) and variants were called using the Samtools (version 1.6) mpileup pipeline using default parameters (Supplemental Table S2, <https://doi.org/10.6084/m9.figshare.14067404.v1>). The INDEL calls were filtered due to concerns with false positive calls within assembled contigs due to usage of the Quiver algorithm as a polishing step (Watson and Warr, 2019). Instead, alternative read mapping statistics were used as proxies to detect the alignment ambiguity around marker sites. These statistics were included as an extension to the spacing equation derived by Matukumalli et al. (2009) to select candidate markers in a first pass trial:

$$\text{Score} = \text{Max} \left( \frac{\text{GMS}_{\text{up}}}{100}, \frac{\text{GMS}_{\text{down}}}{100} \right) \times \frac{\text{QS}}{1,000} \times \text{MAF} \\ \times \left[ (E - S) - |2a - (E + S)| \right],$$

where  $\text{GMS}_{\text{up/down}}$  represents the phred-scaled alignment quality score generated by BWA MEM (Li and Durbin, 2009) upstream and downstream of the SNP, respectively; QS represents the variant call format (The 1000 Genomes Project Consortium, 2010) quality score; MAF represents the minor allele frequency; and the bracketed terms are the marker spacing terms defined in the previous study (Matukumalli et al., 2009). Briefly, each SNP position ( $a$ ) in the target region's start ( $S$ ) and end ( $E$ ) boundaries is evaluated for its position relative to the center of the region. Variants were assigned scores in a recursive fashion until at least 6 SNP markers covered the haplotype, with the highest scoring variant sites in each contig being selected for Agena custom assay (Neogen) design. The implementation of this algorithm can be found on GitHub ([https://github.com/njdbickhart/perl\\_toolchain/](https://github.com/njdbickhart/perl_toolchain/)). Due to the complex nature of these regions, final marker location and suitability was confirmed on our reference haplotypes manually.

**Table 1.** Marker covariate descriptions

Name	Bidirectional <sup>1</sup>	Description
percent_GC	Yes	The percentage of bases in the flanking region that were composed of G and C bases.
percent_N	Yes	The percentage of bases in the flanking region that were N bases (indicative of gaps or 4-fold variant sites)
Minor.Allele.Freq	No	The allele frequency of the alternate base (estimated from the alignment of 125 Holstein whole genome sequence data sets)
VCF_QUAL	No	The phred-scaled ( $-10 \times \log_{10} P$ ) probability of no variant site at the region. Higher values indicate higher confidence in variant site prediction.
percent_IUPAC	Yes	The percentage of bases in the flanking region that were composed of IUPAC alternative base codes (i.e., R is the equivalent of all purine bases). This indicates the presence of other variant sites in the flanking regions.
MapQ	Yes	The phred-scaled probability that a read maps to more than one location in the reference genome. Higher values indicate higher confidence in unique alignment.

<sup>1</sup>Indicates if this covariate is assessed by collecting statistics on the upstream (Superscript: X5) and downstream (X3) 100 bases that immediately flank the variant site.

## Second Round Marker Selection

To improve the success rate of a second round of marker selections, we tested the performance of 3 distinct machine learning classifiers (i.e., logistic regression, decision tree, and random forest) using a 10-fold cross validation method. These analyses were performed in R (v3.6.1, <https://www.r-project.org/>) and source code to reproduce these analyses is available in the following GitHub repositories ([https://github.com/bkiranmayee/My\\_Labnotes/blob/master/IGC/glm.Rmd](https://github.com/bkiranmayee/My_Labnotes/blob/master/IGC/glm.Rmd) and [https://github.com/bkiranmayee/My\\_Labnotes/blob/master/IGC/decision.trees.Rmd](https://github.com/bkiranmayee/My_Labnotes/blob/master/IGC/decision.trees.Rmd)). The classifier models were trained on a random selection of 70% (i.e., training set = 48 SNP ID) of the original 67 marker selections and evaluated on the remaining 30% (i.e., testing set = 19 SNP ID) subset of this data set in all cases. In the training stage of all the 3 classifiers, we started with a full model that included all 10

covariates derived from 100 bp flanking each marker site (see Table 1) as independent variables and the category (i.e., pass or fail) as the dependent variable. Each classifier method was evaluated according to the following performance metrics: accuracy (percentage of the correct predictions over total predictions), sensitivity [true positives (TP) divided by TP + false negatives (FN)], specificity [true negatives (TN) divided by TN + false positives (FP)], precision (TP divided by TP + FP) and Cohen's kappa (or Kappa; accuracy divided by expected accuracy) of its predicted outcomes. These metrics, as well as the confusion matrix obtained after predictions performed on the testing set, are available at Table 2 and Supplemental Table S3 (<https://doi.org/10.6084/m9.figshare.14067401.v1>).

Logistic regression was performed using the glmStepAIC method of the caret R package (Kuhn, 2008) to choose an optimal model based on Akaike information criterion (AIC) by stepwise elimination or addition of

**Table 2.** Performance on train and test stage of 3 different machine learning classifiers (i.e., logistic regression, decision tree and random forest) that were tested to classify SNP\_ID in pass or fail<sup>1</sup>

Performance	Logistic regression	Decision tree	Random forest
Training set			
Accuracy	0.52	0.50	0.54
AccuracySD	0.22	0.25	0.24
Kappa	0.01	0.01	0.06
KappaSD	0.44	0.50	0.48
Testing set			
Accuracy	0.68	0.53	0.63
Kappa	0.35	0.06	0.27
Sensitivity	0.44	0.56	0.78
Specificity	0.90	0.50	0.50
Precision	0.80	0.50	0.58
Confusion matrix			
Fail-Fail	4	5	7
Fail-Pass	1	5	5
Pass-Pass	5	4	2
Pass-Fail	9	5	5

<sup>1</sup>Full description of the models is available in Supplemental Tables 3 (<https://doi.org/10.6084/m9.figshare.14067401.v1>) and 4 (<https://doi.org/10.6084/m9.figshare.14067407.v1>).



covariates. Our final logistic regression model was fitted with 4 variables that had produced the model with the minimum AIC, specifically the X5.MapQ, Minor.Allele.Freq, X3.MapQ, and VCF\_QUAL covariates (Table 1). We then trained 7 decision tree models with the following features: (1) all the variables, (2) only QUAL, (3) X5 and X3 MapQ, (4) X5 and X3 MapQ and MAF, (5) X5 and X3 MapQ and QUAL, (6) X3 MapQ and QUAL, and (7) X5 MapQ and QUAL (see description of terms in Table 1). The complexity parameter was tuned in each decision tree model and the optimal model that produced the highest accuracy was selected. Lastly, we trained 2 random forest models using the R base package `randomForest` (<https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>): (1) using all the variables, and (2) using only marker composition and MAF (analysis results and random forest parameters are listed in Supplemental Table S4, <https://doi.org/10.6084/m9.figshare.14067407.v1>). The models were trained with different number of trees and number of variables available for splitting at each tree node (using the `mtry` parameter of the function). Test set accuracy was used to select the optimal model, which was the model that included all features for classification with 1,000 trees as it had an overall accuracy and out of bag error equal to 63.2 and 52%, respectively. Feature importance was measured as a percentage decrease in the Gini index, which was then scaled from 0 to 100 to constitute a variable importance score.

### GWAS of Genetic Markers with bTB Phenotypes

DNA stocks and Illumina BovineHD genotypes from 1,797 Holstein cattle used in prior bTB surveys (Bermingham et al., 2014; Wilkinson et al., 2017) were used in this analysis. The test herds were originally assessed for bTB incidence through the use of a single intradermal comparative tuberculin test (SICTT) as described previously (Bermingham et al., 2014). The SICTT-positive cattle were subjected to postmortem inspection and were classified as having visible granuloma lesions, consistent with bTB or granuloma lesions not visible on postmortem inspection. Specialist mycobacterial culture was attempted on SICTT-positive cattle samples. These phenotypic measures were condensed into a binary trait, with animals separated into cases and controls as previously (Bermingham et al., 2014); cases ( $n = 1,083$ ) were culture-confirmed, SICTT-positive cattle, and controls ( $n = 460$ ) were derived from a separate pool of equally exposed but repeatedly SICTT-negative and apparently noninfected herd-mates.

A generalized linear mixed model implemented in the GMMAT (v 1.1.1) R package (Chen et al., 2016) was used to assess the effect and significance of each cus-

tom marker on the cases of each phenotype. The linear model was assessed using the `glm.wald` function of the GMMAT package using the following terms:

$$y = \mathbf{X}_i\boldsymbol{\alpha} + G_i\beta + b_i,$$

where  $y$  represents the binary bTB case status,  $\mathbf{X}$  is a row vector of covariates (including herd, age, year, and season) for the  $i$ th animal,  $\boldsymbol{\alpha}$  is the column vector of fixed covariate effects,  $G$  is the genotype of variant  $n$  for the  $i$ th animal, and  $\beta$  is the genotype effect. Finally,  $b$  is the random effects for each animal. To create genotype files suitable for the GMMAT package, animal genotype text files were converted using GEMMA (Zhou and Stephens, 2012). Manhattan plots of the  $-\log_{10}P$  values for each marker and qqplots were generated using the qqman R package (Turner, 2018). Eigenvalues and principal components were generated from genotype files using the `—pca` option of plink v1.90. The first 2 principal components were plotted using the ggplot2 package in R (<https://www.r-project.org/>). The genomic inflation factor ( $\pi$ ) was defined as the median of chi-squared statistical tests on the  $P$ -values of each marker divided by the median of the expected chi-squared distribution (van den Berg et al., 2019). Both chi-squared tests assumed one degree of freedom in expected and observed  $P$ -values.

## RESULTS AND DISCUSSION

A hierarchical assembly strategy was chosen to assemble alternative haplotypes for IGC regions. A total of 40 BAC clones from the CHORI-240 (17 clones) and RPCI-42 (23 clones) libraries were selected, based on alignments of their BAC-end sequences to coordinates of the UMD3.1 reference that should have contained the MHC I (chr23:28,300,000–28,750,000), LRC (chr18:63,100,000–63,400,000), and NKC (chr5:99,500,000–99,850,000) gene clusters. The PacBio RSII sequence of the insert of each BAC clone was assembled separately into 40 separate sets of contigs, with 33 clones assembling into single contigs of sizes within the range of the expected BAC clone insert sizes (~170–250 kb). Assembled clones were then aligned to the ARS-UCD1.2 reference (Rosen et al., 2020) to confirm their location and to remove sequence that was redundant with the reference. This last step was necessary as the CHORI-240 library was created using DNA extracted from L1 Domino who was the sire of the reference animal, L1 Dominette (Elsik et al., 2009). A total of 19 nonredundant contigs (consisting of 3.15 Mbp of total sequence) were used in subsequent alignment and variant calling using a samtools mpileup workflow (Li et al., 2009) and the sequence data from 125 Holstein bulls. A total

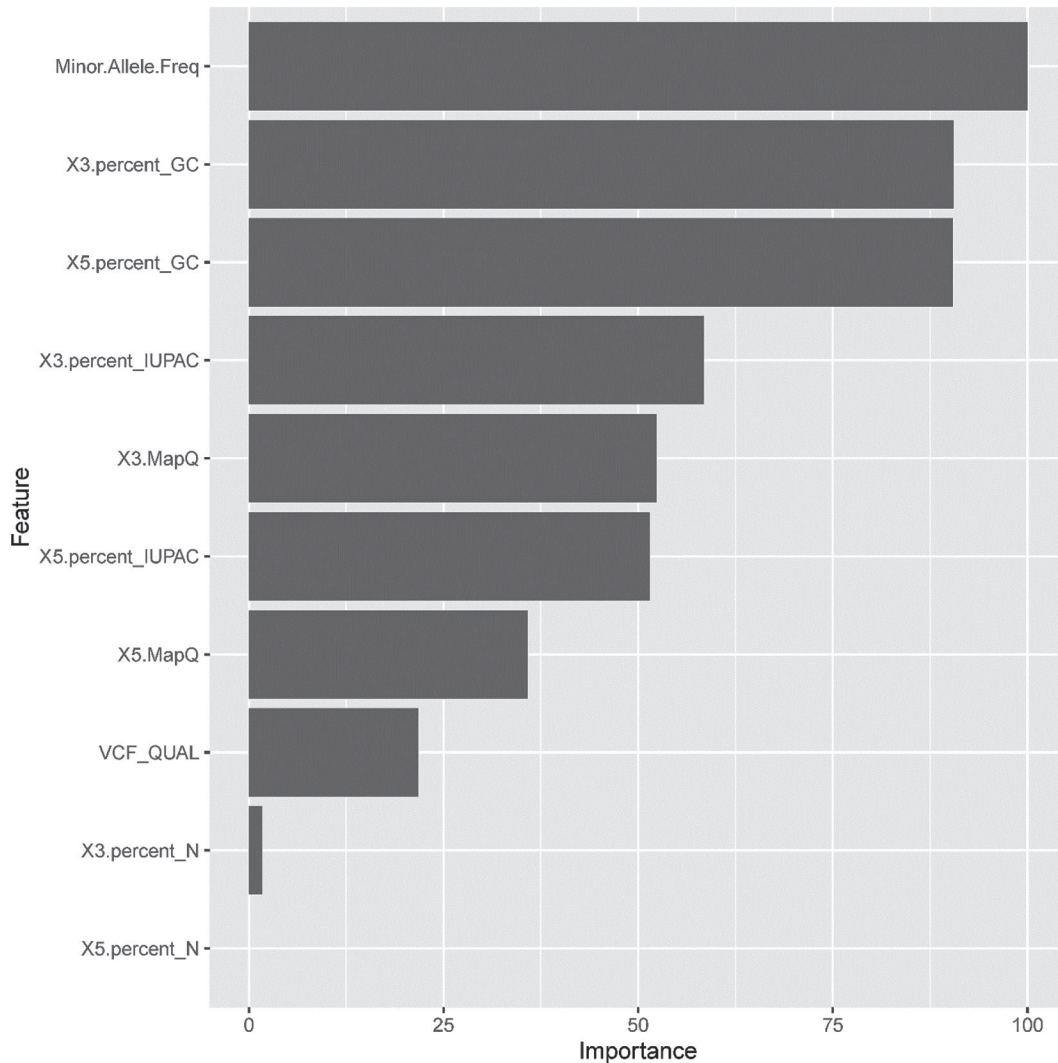
of 54,555 raw SNP variants were identified within our alternative haplotype contigs (31,054 SNP; 57% of the total) and IGC regions present on the ARS-UCD1.2 reference (23,501; 43%). We found that all IGC regions had high degrees of sequence alignment ambiguity as measured by BWA mapping quality (MapQ) scores. To select only sites that could be unambiguously mapped in sequence data, we filtered variant sites that did not have at least one 36 bp flanking region with an average read MapQ greater than 80. This resulted in a final list of 341 SNP sites (149 of which were present on assembled alternative haplotypes) that were used for candidate marker selection. Using an adaptation of a previously developed marker spacing equation (Matukumalli et al., 2009), we selected an initial 67 markers (33 from assembled alternative haplotypes) from this list for custom genotyping.

We assessed marker viability by genotyping a cohort of 1,797 Holstein cattle that were used in previous bTB association studies (Bermingham et al., 2014; Wilkinson et al., 2017). Case and control DNA samples collected from test herds were genotyped using the custom markers. Of an initial 67 marker selections, only 40 markers (59.7%) passed design quality control and had call rates greater than 80% when used in downstream panel genotyping. This was despite the use of variant site information (both SNP and INDEL variants) from the 125 sequenced Holstein bulls in the design of the marker probe sequence. To improve the success rate of a second round of marker selections, we trained 3 different types of binary classifier models using the failure status of the original 67 marker selections as a training set. Features included the general statistics of the candidate marker site as well as the composition of flanking sequence that would be used in primer design (Table 1). A random forest model that included all features was found to have the highest sensitivity and specificity at 0.78 and 0.50, respectively. A benefit to using random forest models is the ability to identify the importance of each feature in the final set of decision tree forests. Our variable importance analysis of random forest features found that the MAF of the marker and the GC percentage of flanking sequence were the most important features discovered by the binary classifier (Figure 1). This was supported by the independent selection of these 2 features as the decision criteria in the best decision tree model (Figure 2). We hypothesized that flanking sequence MapQ scores would play a larger role in successful marker region design due to the repetitive nature of the targeted regions and the potential for off-target probe hybridization. However, flanking sequence MapQ scores were only the fifth and seventh most important features in the model (Table 2). It is

possible that probe binding specificity due to increased GC content plays a larger role in genotyping rate than flanking sequence uniqueness. This would confirm prior observations of increased signal intensity in Illumina beadchip arrays for markers in GC-rich regions of the genome (Diskin et al., 2008). Using the random forest model categorization and manual selection of equally spaced sites, we selected an additional 57 candidate marker sites, of which 44 markers (77%) had call rates greater than 80% in custom panel genotyping.

Marker quality assessments conducted with plink (version 1.9) (Purcell et al., 2007) revealed additional discrepancies in the custom markers that required additional filtering. Despite the efforts to assemble and include additional representative haplotypes in our original variant discovery survey, we identified 12 markers that were monomorphic or had extremely high (>50%) heterozygosity in the test herd. These markers were removed from downstream association testing as they likely represented variants within repetitive regions (multimapping) or were tracking undiscovered structural variants in the test herd. After this last round of filtering, we identified 72 custom genetic markers for an association analysis within previously untracked immune gene regions in the cattle reference genome (Supplemental Table S5, <https://doi.org/10.6084/m9.figshare.14067413.v1>). BovineHD genotypes on the test herd were subject to linkage disequilibrium filtering using the following parameters: window size = 10 SNP, step size = 5 and variance inflation factor,  $\lambda = 4$  in plink 1.9. The 72 custom markers were added to the filtered BovineHD data set giving a final marker count of 187,273 for all 1,797 animals in the test herd. We estimated the linkage disequilibrium between the custom markers and the filtered BovineHD marker set using the `—r2` flag in plink 1.9 with default settings. We identified only one BovineHD marker (BovineHD2300007989) that had an  $r^2$  greater than 0.5 (value = 0.539) with one of our MHC custom markers (MHC\_154399). This suggests that the majority of our custom markers do segregate independently in the genotyped population, and that the markers themselves track novel haplotypes or alleles of the assembled IGC.

Similar to a previous survey (Bermingham et al., 2014), one marker (BovineHD0300013035; different from the SNP identified in that survey) achieved suggestive significance (Figure 3), but had a small effect size (0.44). All markers (187,273 in total; including the BovineHD markers) had small predicted effects on the phenotype (ranges from  $-0.95$  to  $0.63$ ). This is similar to the findings of a previous study (Raphaka et al., 2017), and the 72 custom IGC markers had similar, smaller effect sizes ( $-0.28$  to  $0.63$ ; Table 3; Supple-



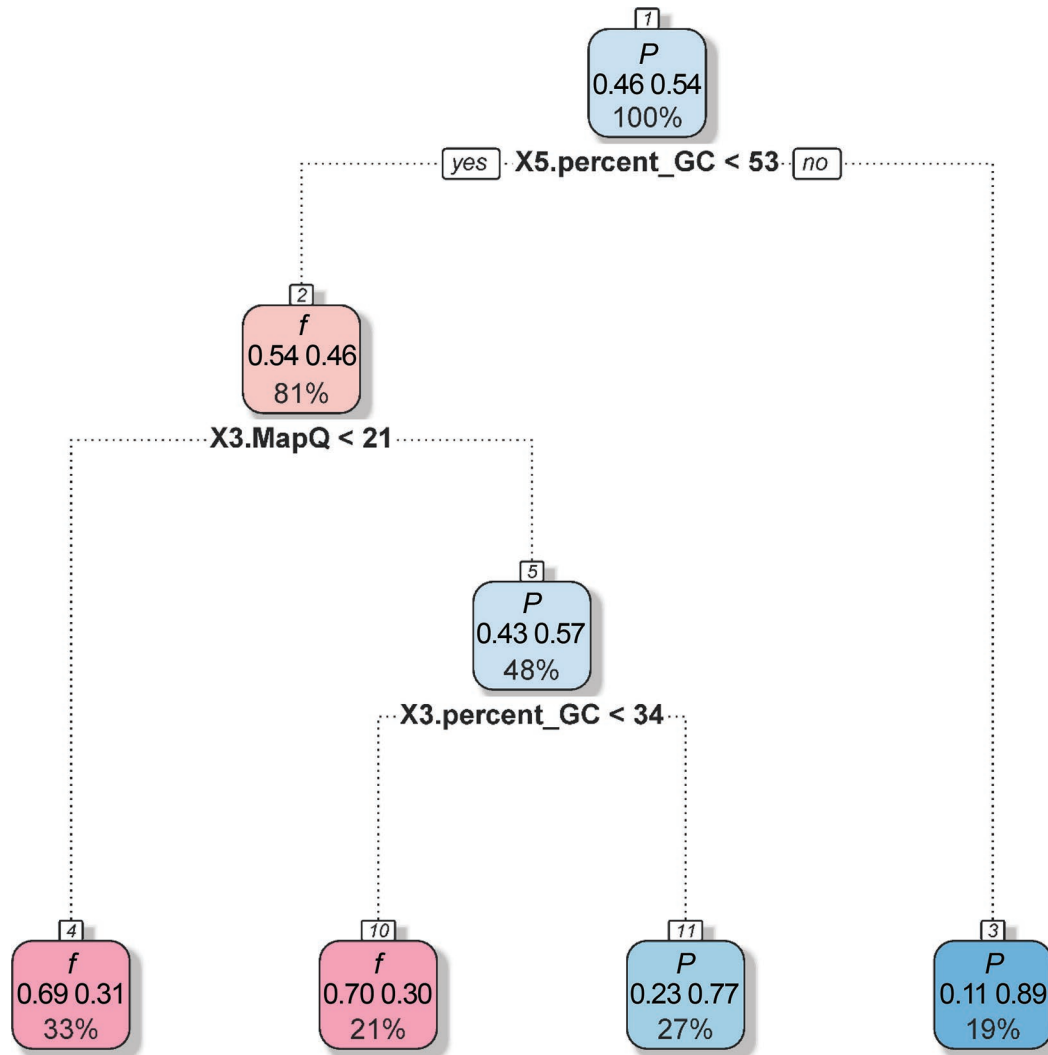
**Figure 1.** Feature selection by random forest classifier. Mean decrease in Gini was used to calculate the variable importance score (VIMP: Importance) scaled from 0 to 100.

mental Table S5) to those identified for the BovineHD markers. As found in the previous study (Wilkinson et al., 2017), a principal components analysis did not identify substantial population substructure in our data set that could be associated with the phenotype (Figure 4). The genomic inflation factor ( $\pi$ ) value of 1.012 suggested little deviation in observed test statistics from the expected, which also suggests that population substructure had little influence on the association analysis. This is further reflected in a Q-Q plot of expected and observed  $P$ -values that also show little deviation from expected values (Figure 5). Although our new IGC markers did not achieve genome-wide significance, predicted effect sizes suggest that they may still contribute information in bTB ge-

nostic selection models. However, we acknowledge that our methods may have missed additional structural diversity in IGC regions within the surveyed Holstein population. Detection of individual structural variants could be accomplished by future surveys using the latest in low-error, long-read sequencing technologies (Wenger et al., 2019). To promote their use in other studies, we have made variant site information freely available with this publication.

## CONCLUSIONS

We report the first suite of suitable genetic markers for genotyping within important immune gene complex loci in the cattle genome, derived from assembled IGC



**Figure 2.** The final tree model, which presents complexity parameter and overall accuracy equal to 0.08 and 50%, respectively (see Supplemental Table S4, <https://doi.org/10.6084/m9.figshare.14067407.v1>).

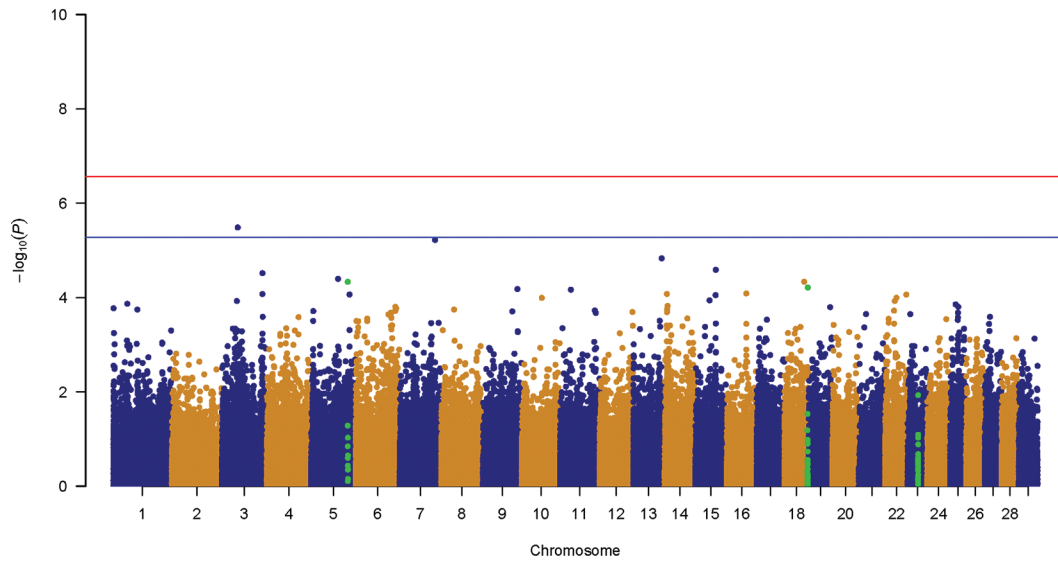
haplotypes using linear and random forest models. Our approach highlights the need for hierarchical approaches for marker development in otherwise polymorphic regions of the genome that exist in more than one allelic state. We tested the association of our new custom markers in a case-control study of bTB incidence as a proof-of-principle test. Although we did identify 2 markers with moderate effects on phenotype prediction (ARS-PIRBRIGHT-18\_63417698 and ARS-PIR-

BRIGHT-5\_99190989), the effect sizes were within the range of other BovineHD markers for each phenotype and were not statistically significant. We also found that individual SNP did not account for a considerable proportion of the genetic variance underlying the trait, which is consistent with earlier findings (Raphaka et al., 2017). This is to be expected for complex polygenic traits with relatively low heritabilities, such as bTB resistance and susceptibility. The custom markers iden-

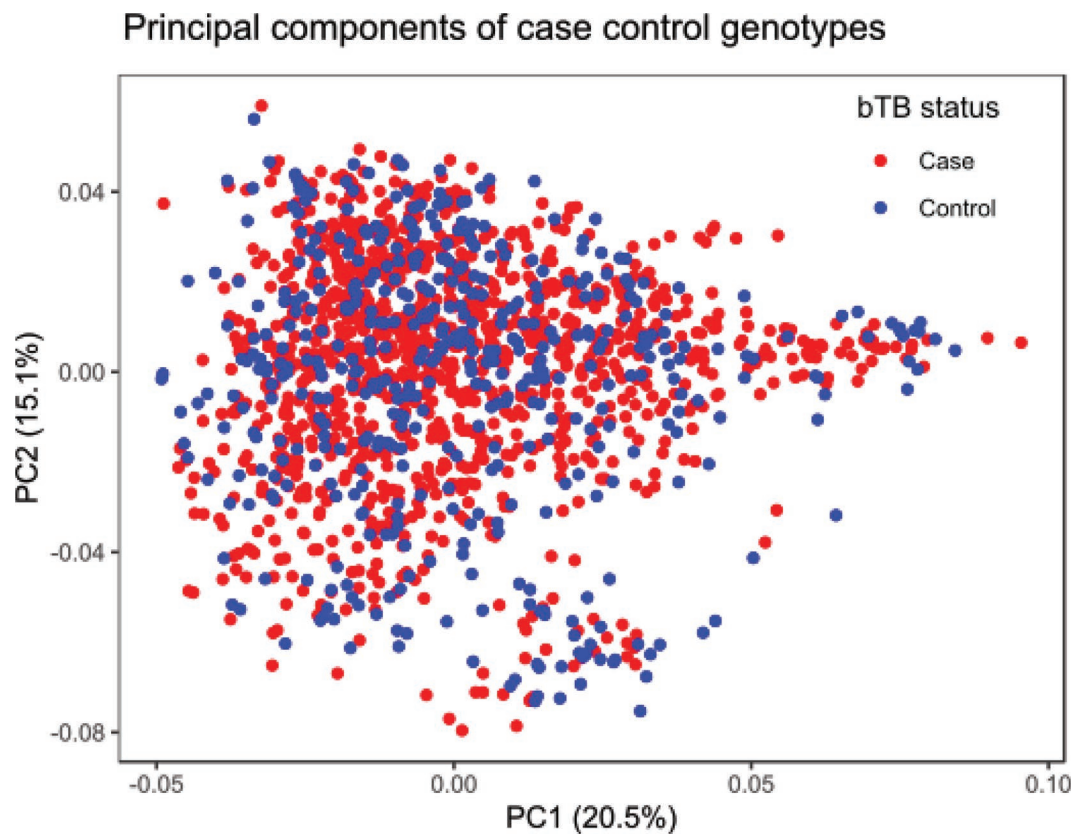
**Table 3.** The 2 custom SNP with the largest effects on bovine tuberculosis case-control status

Item	ARS-PIRBRIGHT-5_99190989	ARS-PIRBRIGHT-18_63417698
Beta	0.635	0.431
<i>P</i> -value	$4.63 \times 10^{-5}$	$6.14 \times 10^{-5}$

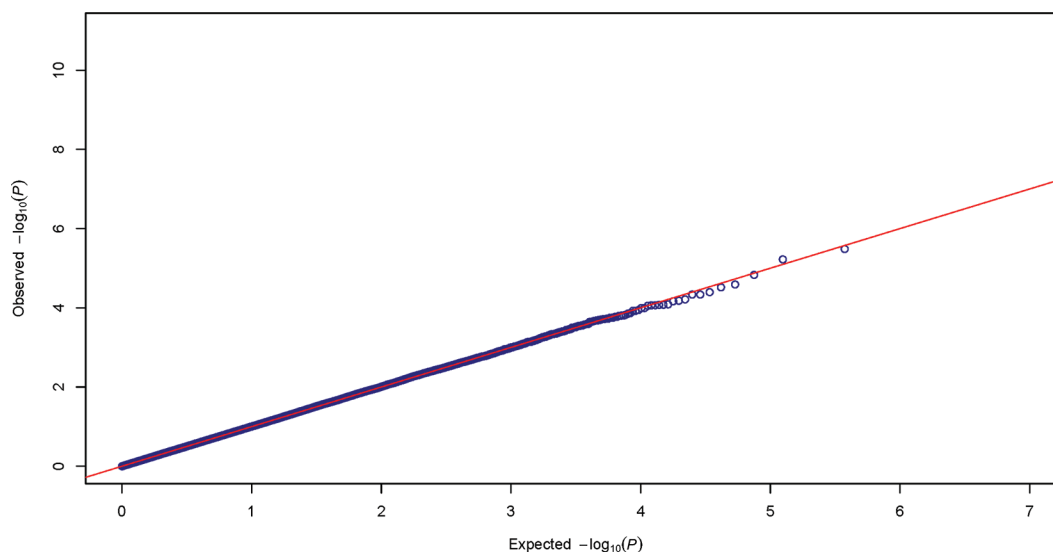




**Figure 3.** Manhattan plots of the phenotype derived from postmortem and skin-test observations. The 72 novel markers are shown as bright green points. The red line indicates genome-wide significance ( $-\log_{10}(P) > 8$ ) and the blue line indicates suggestive associations ( $-\log_{10}(P) > 5$ ). Manhattan plots were generated in the qqman package with the 72 novel custom markers displayed in bright green.



**Figure 4.** Genotype principal components (PC) plot; bTB = bovine tuberculosis.



**Figure 5.** Quantile-quantile plots of  $P$ -values from the marker-phenotype association analysis. The close correspondence of the observed with the expected values suggests the absence of notable population stratification.

tified in this survey may be helpful for improving the accuracy of estimated breeding values for these traits in the future (Banos et al., 2017).

### ACKNOWLEDGMENTS

Hammond, Heimeier, and Schwartz were supported by United Kingdom Research and Innovation, Biotechnology and Biological Sciences Research Council (UKRI-BBSRC) funding awards BB/M027155/1, BBS/E/I/00007031, BBS/E/I/00007038, BBS/E/I/00007039, BBS/OS/GC/000015B, and BBS/OS/GC/200016. Glass was supported by UKRI-BBSRC funding awards BB/J004227/1, BB/J004235/1, and BB/P013740; Glass, Skuce, and Allen were also supported by UKRI-BBSRC BB/E018386/1, BB/E018335/1 and 2, and BB/L004054/1; Glass was also supported by UKRI-BBSRC award BB/M027155/1 and BB/P013740/1. Wilkinson was supported by UKRI-BBSRC BB/L004054/1. We gratefully acknowledge the Agri-Food and Biosciences Institute (AFBI, Northern Ireland) who collected and provided samples in the form of phenotyped bTB case/control samples for use within this project. Bickhart, Bakshy, McClure, and Null were supported by appropriated projects 5090-31000-026-00-D, Investigating Microbial, Digestive, and Animal Factors to Increase Dairy Cow Performance and Nutrient Use Efficiency, and 8042-31000-001-00-D, Enhancing Genetic Merit of Ruminants Through Improved Genome Assembly, Annotation, and Selection, of the Agricultural Research Service (ARS) of the USDA. Cole and Null were supported by appropriated

project 8042-31000-002-00-D, “Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals of ARS-USDA. Cole was also partially supported by the grant “Reducing Mastitis in the Dairy Cow by Increasing the Prevalence of Beneficial Polymorphisms in Genes Associated with Mastitis Resistance” from the Minnesota Agricultural Experiment Station Rapid Agricultural Response Fund. Smith was supported by appropriated project 3040-31000-100-00-D, “Developing a Systems Biology Approach to Enhance Efficiency and Sustainability of Beef and Lamb Production,” of ARS-USDA. Bickhart, Bakshy, Young, and Smith were supported by USDA NIFA grant number 2015-67015-22970, “US-UK Collaborative project: “Reassembly of cattle immune gene clusters for quantitative analysis.” Sequence data used as background in initial variant selection was provided by the Cooperative Dairy DNA Repository on request. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer. The authors have not stated any conflicts of interest.








### REFERENCES

- Allan, A. J., N. D. Sanderson, S. Gubbins, S. A. Ellis, and J. A. Hammond. 2015. Cattle NK Cell Heterogeneity and the influence of MHC Class I. *J. Immunol.* 195:2199–2206. <https://doi.org/10.4049/jimmunol.1500227>.
- Allen, A. R., G. Minozzi, E. J. Glass, R. A. Skuce, S. W. J. McDowell, J. A. Woolliams, and S. C. Bishop. 2010. Bovine tuberculosis: The

- genetic basis of host susceptibility. *Proc. Biol. Sci.* 277:2737–2745. <https://doi.org/10.1098/rspb.2010.0830>.
- Allen, A. R., R. A. Skuce, and A. W. Byrne. 2018. Bovine tuberculosis in Britain and Ireland – A perfect storm? The confluence of potential ecological and epidemiological impediments to controlling a chronic infectious disease. *Front. Vet. Sci.* 5:109. <https://doi.org/10.3389/fvets.2018.00109>.
- Banos, G., M. Winters, R. Mrode, A. P. Mitchell, S. C. Bishop, J. A. Woolliams, and M. P. Coffey. 2017. Genetic evaluation for bovine tuberculosis resistance in dairy cattle. *J. Dairy Sci.* 100:1272–1281. <https://doi.org/10.3168/jds.2016.11897>.
- Bermingham, M. L., S. C. Bishop, J. A. Woolliams, R. Pong-Wong, A. R. Allen, S. H. McBride, J. J. Ryder, D. M. Wright, R. A. Skuce, S. W. McDowell, and E. J. Glass. 2014. Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. *Heredity* 112:543–551. <https://doi.org/10.1038/hdy.2013.137>.
- Bermingham, M. L., S. Brotherstone, D. P. Berry, S. J. More, M. Good, A. R. Cromie, I. M. White, I. M. Higgins, M. Coffey, S. H. Downs, E. J. Glass, S. C. Bishop, A. P. Mitchell, R. S. Clifton-Hadley, and J. A. Woolliams. 2011. Evidence for genetic variance in resistance to tuberculosis in Great Britain and Irish Holstein-Friesian populations. *BMC Proc.* 5(Suppl 4):S15. <https://doi.org/10.1186/1753-5561-5-S4-S15>.
- Bickhart, D. M., J. L. Hutchison, D. J. Null, P. M. VanRaden, and J. B. Cole. 2016. Reducing animal sequencing redundancy by preferentially selecting animals with low-frequency haplotypes. *J. Dairy Sci.* 99:5526–5534. <https://doi.org/10.3168/jds.2015.10347>.
- Bickhart, D. M., B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisà, F. A. Ponce de León, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, and T. P. L. Smith. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* 49:643–650. <https://doi.org/10.1038/ng.3802>.
- Chen, H., C. Wang, M. P. Conomos, A. M. Stilp, Z. Li, T. Sofer, A. A. Szpiro, W. Chen, J. M. Brehm, J. C. Celedón, S. Redline, G. J. Papanicolaou, T. A. Thornton, C. C. Laurie, K. Rice, and X. Lin. 2016. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* 98:653–666. <https://doi.org/10.1016/j.ajhg.2016.02.012>.
- Codner, G. F., J. Birch, J. A. Hammond, and S. A. Ellis. 2012. Constraints on haplotype structure and variable gene frequencies suggest a functional hierarchy within cattle MHC class I. *Immunogenetics* 64:435–445. <https://doi.org/10.1007/s00251-012-0612-6>.
- Diskin, S. J., M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* 36:e126. <https://doi.org/10.1093/nar/gkn556>.
- Ellis, S. A., and J. A. Hammond. 2014. The functional significance of cattle major histocompatibility complex class I genetic diversity. *Annu. Rev. Anim. Biosci.* 2:285–306. <https://doi.org/10.1146/annurev-animal-022513-114234>.
- Elsik, C. G., R. L. Tellam, K. C. Worley, R. A. Gibbs, D. M. Muzny, G. M. Weinstock, et al. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* 324:522–528. <https://doi.org/10.1126/science.1169588>.
- Gibson, M. S., A. J. Allan, N. D. Sanderson, J. Birch, S. Gubbins, S. A. Ellis, and J. A. Hammond. 2020. Two lineages of KLRA with contrasting transcription patterns have been conserved at a single locus during ruminant speciation. *J. Immunol.* 204:2455–2463. <https://doi.org/10.4049/jimmunol.1801363>.
- Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13:635–643. <https://doi.org/10.1101/gr.387103>.
- Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37:540–546. <https://doi.org/10.1038/s41587-019-0072-8>.
- Koren, S., A. Rhie, B. P. Walenz, A. T. Diltthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J. L. Williams, T. P. L. Smith, and A. M. Phillippy. 2018. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 36:1174–1182. <https://doi.org/10.1038/nbt.4277>.
- Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. 2017. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–736. <https://doi.org/10.1101/gr.215087.116>.
- Kuhn, M. 2008. Building predictive models in R using the caret package. *J. Stat. Softw.* 28:1–26. <https://doi.org/10.18637/jss.v028.i05>.
- le Roex, N., P. D. van Helden, A. P. Koets, and E. G. Hoal. 2013. Bovine TB in livestock and wildlife: What's in the genes? *Physiol. Genomics* 45:631–637. <https://doi.org/10.1152/physiolgenomics.00061.2013>.
- Li, H. 2016. Minimap and minimap: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32:2103–2110. <https://doi.org/10.1093/bioinformatics/btw152>.
- Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
- Low, W. Y., R. Tearle, R. Liu, S. Koren, A. Rhie, D. M. Bickhart, B. D. Rosen, Z. N. Kronenberg, S. B. Kingan, E. Tseng, F. Thibaud-Nissen, F. J. Martin, K. Billis, J. Ghurye, A. R. Hastie, J. Lee, A. W. C. Pang, M. P. Heaton, A. M. Phillippy, S. Hiendleder, T. P. L. Smith, and J. L. Williams. 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat. Commun.* 11:2071. <https://doi.org/10.1038/s41467-020-15848-y>.
- Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4:e5350. <https://doi.org/10.1371/journal.pone.0005350>.
- Payne, A., N. Holmes, T. Clarke, R. Munro, B. J. Debebe, and M. Loose. 2020. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* Online ahead of print. <https://doi.org/10.1038/s41587-020-00746-x>.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559–575. <https://doi.org/10.1086/519795>.
- Raphaka, K., O. Matika, E. Sánchez-Molano, R. Mrode, M. P. Coffey, V. Riggio, E. J. Glass, J. A. Woolliams, S. C. Bishop, and G. Banos. 2017. Genomic regions underlying susceptibility to bovine tuberculosis in Holstein-Friesian cattle. *BMC Genet.* 18:27. <https://doi.org/10.1186/s12863-017-0493-7>.
- Rosen, B. D., D. M. Bickhart, R. D. Schnabel, S. Koren, C. G. Elsik, E. Tseng, T. N. Rowan, W. Y. Low, A. Zimin, C. Couldrey, R. Hall, W. Li, A. Rhie, J. Ghurye, S. D. McKay, F. Thibaud-Nissen, J. Hoffman, B. M. Murdoch, W. M. Snelling, T. G. McDanel, J. A. Hammond, J. C. Schwartz, W. Nandolo, D. E. Hagen, C. Dreischer, S. J. Schultheiss, S. G. Schroeder, A. M. Phillippy, J. B. Cole, C. P. Van Tassell, G. Liu, T. P. L. Smith, and J. F. Medrano. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9:giaa021. <https://doi.org/10.1093/gigascience/giaa021>.
- Sanderson, N. D., P. J. Norman, L. A. Guethlein, S. A. Ellis, C. Williams, M. Breen, S. D. E. Park, D. A. Magee, F. Babrzadeh, A. Warry, M. Watson, D. G. Bradley, D. E. MacHugh, P. Parham, and J. A. Hammond. 2014. Definition of the cattle killer cell Ig-like receptor gene family: Comparison with aurochs and human coun-

- terparts. *J. Immunol.* 193:6016–6030. <https://doi.org/10.4049/jimmunol.1401980>.
- Schwartz, J. C., M. S. Gibson, D. Heimeier, S. Koren, A. M. Phillippy, D. M. Bickhart, T. P. L. Smith, J. F. Medrano, and J. A. Hammond. 2017. The evolution of the natural killer complex; a comparison between mammals using new high-quality genome assemblies and targeted annotation. *Immunogenetics* 69:255–269. <https://doi.org/10.1007/s00251-017-0973-y>.
- The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073. <https://doi.org/10.1038/nature09534>.
- Turner, S. D. 2018. qqman: An R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* 3:731. <https://doi.org/10.21105/joss.00731>.
- van den Berg, S., J. Vandenplas, F. A. van Eeuwijk, M. S. Lopes, and R. F. Veerkamp. 2019. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. *J. Anim. Breed. Genet.* 136:418–429. <https://doi.org/10.1111/jbg.12419>.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>.
- VanRaden, P. M., M. E. Tooker, J. R. O’Connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol.* 49:32. <https://doi.org/10.1186/s12711-017-0307-4>.
- Vasoya, D., A. Law, P. Motta, M. Yu, A. Muwonge, E. Cook, X. Li, K. Bryson, A. MacCallam, T. Sitt, P. Toye, B. Bronsvort, M. Watson, W. I. Morrison, and T. Connelley. 2016. Rapid identification of bovine MHCI haplotypes in genetically divergent cattle populations using next-generation sequencing. *Immunogenetics* 68:765–781. <https://doi.org/10.1007/s00251-016-0945-7>.
- Watson, M., and A. Warr. 2019. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* 37:124–126. <https://doi.org/10.1038/s41587-018-0004-z>.
- Wenger, A. M., P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, and M. W. Hunkapiller. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37:1155–1162. <https://doi.org/10.1038/s41587-019-0217-9>.
- Wilkinson, S., S. C. Bishop, A. R. Allen, S. H. McBride, R. A. Skuce, M. Bermingham, J. A. Woolliams, and E. J. Glass. 2017. Fine-mapping host genetic variation underlying outcomes to *Mycobacterium bovis* infection in dairy cows. *BMC Genomics* 18:477. <https://doi.org/10.1186/s12864-017-3836-x>.
- Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–824. <https://doi.org/10.1038/ng.2310>.
- Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* 10:R42. <https://doi.org/10.1186/gb-2009-10-4-r42>.

## ORCIDS

- J. C. Schwartz  <https://orcid.org/0000-0003-2057-1831>
- R. A. Skuce  <https://orcid.org/0000-0001-5177-9198>
- J. Young  <https://orcid.org/0000-0001-7255-3315>
- J. B. Cole  <https://orcid.org/0000-0003-1242-4401>
- J. A. Hammond  <https://orcid.org/0000-0002-2213-3248>
- T. P. L. Smith  <https://orcid.org/0000-0003-1611-6828>
- D. M. Bickhart  <https://orcid.org/0000-0003-2223-9285>