An alternative experimental case–control design for genetic association studies on bovine mastitis

S. Biffani¹, M. Del Corvo², R. Capoferri³, A. Pedretti², M. Luini⁴, J. L. Williams⁵, G. Pagnacco⁶, F. Minvielle⁷ and G. Minozzi²,⁶†

¹Istituto di Biologia e Biotecnologia Agraria, Consiglio Nazionale delle Ricerche, Via Einstein, 26900, Lodi, Italy; ²Parco Tecnologico Padano Srl, Via Einstein, 26900, Lodi, Italy; ³Istituto Sperimentale Italiano ‘Lazzaro Spallanzani’, 26027, Cremona, Italy; ⁴Istituto Zooprofilattico Sperimentale della Lombardia e dell’Emilia Romagna, Via Einstein, 26900, Lodi, Italy; ⁵School of Animal and Veterinary Sciences, Faculty of Sciences, University of Adelaide, Roseworthy SA 5371, Australia; ⁶Department of Veterinary Sciences, University of Milan, Via Celoria 10, 20133, Milan, Italy; ⁷Institut National de la Recherche Agronomique (INRA), UMR 1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy-en-Josas, France

(Received 20 April 2016; Accepted 21 July 2016; First published online 18 August 2016)

The possibility of using genetic control strategies to increase disease resistance to infectious diseases relies on the identification of markers to include in the breeding plans. Possible incomplete exposure of mastitis-free (control) animals, however, is a major issue to find relevant markers in genetic association studies for infectious diseases. Usually, designs based on elite dairy sires are used in association studies, but an epidemiological case–control strategy, based on cows repeatedly field-tested could be an alternative for disease traits. To test this hypothesis, genetic association results obtained in the present work from a cohort of Italian Holstein cows tested for mastitis over time were compared with those from a previous genome-wide scan on Italian Holstein sires genotyped with 50k single nucleotide polymorphisms for de-regressed estimated breeding values for somatic cell counts (SCCs) on Bos taurus autosome (BTA6) and BTA14. A total of 1121 cows were selected for the case–control approach (cases = 550, controls = 571), on a combination of herd level of SCC incidence and of within herd individual level of SCC. The association study was conducted on nine previously identified markers, six on BTA6 and four on BTA14, using the R statistical environment with the ‘qtscore’ function of the GenABEL package, on high/low adjusted linear score as a binomial trait. The results obtained in the cow cohort selected on epidemiological information were in agreement with those obtained from the previous sire genome-wide association study (GWAS). Six out of the nine markers showed significant association, four on BTA14 (rs109146371, rs109234250, rs109421300, rs109162116) and two on BTA6 (rs110527224 and rs42766480). Most importantly, using mastitis as a case study, the current work further validated the alternative use of historical field disease data in case–control designs for genetic analysis of infectious diseases in livestock.

Keywords: experimental design, association study, mastitis, cattle, genetic marker

Implications

Recent developments in livestock genomic research have shown that host genetics has a significant role in the spread of an infectious disease. Experimental plans powerful enough to quantify the genetic variation in disease resistance in most livestock species require field data that, however, is very noisy due to incomplete exposure or imperfect diagnosis. The current work has demonstrated the potential use of individual historical field disease data to set an alternative experimental case–control designs for the study of genetic associations with infectious diseases in livestock aiming to reduce noise in field data, using mastitis in cattle as case study.

Introduction

Control of major infectious disease in livestock remains challenging despite the detailed characterisation of the infectious agents associated with common diseases and the deciphering of their genome sequences (Tomley and Shirley, 2009). Furthermore, recent developments in livestock genetic and genomic research have shown that host genetics had a significant role in the spread of an infectious disease within populations (Springbett et al., 2003; Lively, 2010; Anacleto et al., 2015).

Disease resistance is defined as the ability of the host to exert some degree of control over the pathogen life cycle (Bishop and Stear, 2003; Bishop and Woolliams, 2014). Host genetic variation in infectious disease resistance is present...
for most diseases (Bishop, 2010) and its variability could be used in selection. Therefore, a genetic approach, aiming to improve disease resistance could be applied as a complement to existing control strategies (Davis et al., 2009). Improvements in animal genome sequencing technologies have led to the application of genomic selection in several countries and to the identification of regions of the genome associated to characters of economic interest (e.g. production and/or resistance to diseases in dairy cattle) (Lillehammer et al., 2011). This dual consequence facilitates the incorporation of disease-resistant loci information into livestock breeding (Bishop and Woolliams, 2010).

Infectious diseases in livestock are likely to increase considerably production costs, as is the case for mastitis which may cost up to 2 billion €/year in Europe (http://www.sabre-eu.eu/). Therefore, it is important to select for disease-resistant animals and to decrease steadily the infectious pressure within and across herds through reducing the number of infected animals, which may be achieved by taking advantage of genomic selection programmes applied worldwide.

Experimental plans powerful enough to quantify the genetic variation in disease resistance in most livestock species require field data, as the number of individuals needed in this type of study is usually very high (e.g. Bermingham et al., 2014; Neibergs et al., 2014). Field disease data for genetic association studies are very noisy (Bishop and Woolliams, 2010; Bishop et al., 2012), however, due to incomplete exposure or imperfect diagnosis of infection. Incomplete exposure will occur when individuals test negative because they are sampled from low incidence (LI) herds where they have never encountered the infectious agent. Then, if ones carries out a standard case–control study with equal number of cases and controls balanced by sampling in the same herd, incomplete exposure of the controls in low prevalence herds will lower the power of detection of differences in allele frequencies of the resistant alleles between the two groups. Imperfect diagnosis is associated to the dynamic aspects of individual resistance over the entire time course of the disease, and especially when a recovery is possible, as for mastitis. In this case, the use of a single-point phenotypic measure will be little informative. A new phenotype should be defined to represent the disease status condition, based upon repeated measurements over the entire life cycle of the animal. Ideally, it should consider the dynamic aspects of the individual’s response over time as well as changes in performance with respect to variations in pathogen load in the host population (Doeschl-Wilson et al., 2012). Both these noise factors added to a number of usual environmental sources of variation will decrease the power of the experimental plans, and could be at the basis of the lower heritability estimates, which probably have lowered the expectations raised in the past by the genetic approach to increase disease resistance (Bishop and Woolliams, 2010).

In order to both decrease the noise intrinsic to disease field data and to maximise the expression of individual genetic differences of disease resistance in a case–control study, case (affected) animals should be sampled from herds with LI of infection, whereas control (healthy) individuals should be chosen in herds with high incidence (HI) level (Bishop et al., 2012). Under this setting, incomplete exposure could be overcome from prior classification of herds before sampling individuals. Next, to reduce imperfect diagnosis, all historical information obtained from field data related to the disease should be used synthetically to take into account the dynamic aspects of individual and herd classification variation in time (Bishop et al., 2012).

The purpose of this work was to apply these concepts to the genetic study of resistance to mastitis. The incidence and costs of mastitis, along with the growing public concerns for animal welfare and for the use of antibiotics in farm animals, have made mastitis one of the major diseases of the dairy sector (Thompson-Crispi et al., 2014). The concentration of somatic cells per milliliters of milk, recorded routinely in dairy cattle (somatic cell count (SCC) = cells/ml) can be used as an indicator of intra-mammary infection (Reents et al., 1995). The logarithmic transformation of the value of SCC, called somatic cell score (SCS) or linear score (LS), has a strong genetic correlation (~0.70) with clinical mastitis and has a heritability close to 0.15 (Rupp and Boichard, 2003). Mastitis represents an interesting case study for situations where an endemic disease with systematic industry-wide repeated phenotyping is concerned.

In a previous study, a genome-wide scan for mastitis resistance was performed on a panel of Italian Holstein top sires using de-regressed estimated breeding values (EBVs) for SCCs. Results showed two chromosomal regions associated to resistance/susceptibility located on chromosome 6 (Bos taurus autosome (BTA6)) and on chromosome 14 (BTA14) (Minozzi et al., 2011) and further details on the seven significant single nucleotide polymorphisms (SNPs) identified are given in this paper. The operational objective of the present study was to check if these SNPs could be detected through the epidemiologically based design described above using a cohort of Italian Holstein cows selected as case and control individuals, respectively, from herds with low and high mastitis incidence.

Furthermore, the effect of the above-mentioned seven SNPs as well as the association with two other markers located in DGAT1 (Winter et al., 2002), were discussed in relation with possible candidate genes.

Material and methods

Phenotypic data

Data were provided by the Lombardy Breeders Association (Milano, Italy) and comprised 980 725 monthly test-day of milk (kg/day) and SCCs from 29 535 primiparous (n = 10 964) and pluriparous (n = 18 571) Holstein cows. Records were collected from 179 herds and included calvings from 1998 to 2013. Only the first four lactations were used and records were retained if between 10 and 400 days in milk (DIM). A minimum of 10 test-days/cow were required.
and if a cow had more than two lactations only the last 22 test-days were retained (e.g. if a cow had three calvings only the last two calvings were retained). After editing the data set included 146 160 records from 7662 cows: 67 062 records for first parity (45.9%), 63 115 records for second parity (43.2%), 13 375 records for third parity (9.2%) and 2608 records for fourth parity (1.8%). To obtain a distribution that is closer to normal, SCC data were transformed to LS as follows (Ali and Shook, 1980):

\[
\text{Linear Scores (LS)} = \log_{10} \frac{\text{SCC} + 3}{100}
\]

Edited data were used to estimate for each cow a LS value adjusted for a set of fixed and random effects fitting the following repeatability mixed model:

\[
Y_{iijlmnopqrs} = \mu + T_{Di} + \text{HERD}_j + \text{CDIM}_l + Y_{Sm} + \text{PARITY}_n + \text{AGE}_o + b_p(\text{MILK}) + \text{PE}_q + \text{SIRE}_r + \epsilon_{iijlmnopqrs}
\]

where \(Y_{iijlmnopqrs}\) is the test-day record of LS, \(\mu\) the overall mean of edited records, \(T_{Di}\) the fixed effect of test-day, \(\text{HERD}_j\) the fixed effect of the herd, \(\text{CDIM}_l\) the fixed effect for class of DIM (defined as 1 class for every 30 days, resulting in 14 classes), \(Y_{Sm}\) the fixed effect of year and season of calving, \(\text{PARITY}_n\) the fixed effect for the parity effect, \(\text{AGE}_o\) the fixed effect of age at calving and \(b_p\) the regression coefficient associated with the fixed effect of milk yield (MILK). The random environmental effect between consecutive test-day records within each lactation of a cow was included in the term \(PE_q\). Population stratification was accounted for by including an uncorrelated sire (SIRE) random effect and \(\epsilon_{iijlmnopqrs}\) the residual error. The R software (http://cran.r-project.org) and the lmer function from lme4 package (Bates et al., 2015) were used for all the data editing and analyses.

Next, a two-step procedure was used to identify HI v. LI herds and control and case animals, respectively, within the two herd types. Herd classification was based on the percentage of animals that showed individual high or low mean SCC values. A value of 200 000 cells/ml was used as a threshold to identify a possible case of mastitis or suggest an ongoing bacterial infection (Bradley and Green, 2005; Madouasse et al., 2012).

Using historical SCC herd data, herds with >30% of the animals averaging above 200 000 cells/ml were classified as HI herds. Then, herds with <30% of animals above the threshold, on average, were classified as LI herds. Of the entire data set, 69 and 50 herds were classified as HI and LI herds, respectively.

Subsequently, individuals to be included in the association study were selected from these herds, identifying animals with high adjusted LS (case animals) in LI herds and animals with low adjusted LS (control animals) in the HI herds. An adjusted LS threshold equal to 0 was used to cluster animals in case and control groups. If a cow from a LI herd had a positive adjusted LS value, she was included in the ‘case’ cluster. If a cow from a HI herd had a negative adjusted LS value, she was included in the ‘control’ cluster. The adjusted LS practically represents the actual performance of any cow, that is, adjusted for non-genetic effects. A more stringent threshold could have been used but it would have reduced the sample size, eventually affecting the genome-wide association study (GWAS) statistical power. Cows with the most extreme LS values from all 119 herds were included as much as possible but selecting half and full sibs within herd was avoided.

**DNA extraction and genotyping**

DNA extraction was carried out on all blood samples using the commercial kit QIAamp DNA Mini kit and blood (Qiagen, Hilden, Germany) using the specific protocol (DNA purification from blood or body fluids). The DNA extracted was quantified by Nanodrop (NanoDrop Technologies, Wilmington, DE, USA). Only samples with a quantity of DNA between 5 and 40 ng/µl and A260/A280 values between 1.7 and 1.9 were considered suitable for analysis. Nine markers were studied in the present work. Seven had been identified in the previous GWAS analysis and are shown in Table 1 (Minozzi et al., 2011), and marker rs109234250 and marker rs109162116 were added because they are located within the DGAT1 gene.

### Table 1 Results of the association analysis of nine single nucleotide polymorphism (SNP) on BTA6 and BTA14 tested in Holstein cattle

<table>
<thead>
<tr>
<th>Markers</th>
<th>BTA</th>
<th>Position</th>
<th>A1</th>
<th>A2</th>
<th>n</th>
<th>effB</th>
<th>se_effB</th>
<th>P-value</th>
<th>Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs110527224</td>
<td>6</td>
<td>88592295</td>
<td>G</td>
<td>A</td>
<td>1099</td>
<td>-0.16</td>
<td>0.04</td>
<td>2.03E-04</td>
<td>0.0018234</td>
</tr>
<tr>
<td>rs42766480</td>
<td>6</td>
<td>88891318</td>
<td>T</td>
<td>C</td>
<td>1111</td>
<td>-0.15</td>
<td>0.04</td>
<td>3.74E-04</td>
<td>0.0033620</td>
</tr>
<tr>
<td>rs29022799</td>
<td>6</td>
<td>88421804</td>
<td>T</td>
<td>C</td>
<td>1103</td>
<td>-0.14</td>
<td>0.05</td>
<td>6.00E-03</td>
<td>0.0539800</td>
</tr>
<tr>
<td>rs109571486</td>
<td>6</td>
<td>88486150</td>
<td>C</td>
<td>G</td>
<td>1114</td>
<td>-0.12</td>
<td>0.05</td>
<td>1.60E-02</td>
<td>0.1448384</td>
</tr>
<tr>
<td>rs29010419</td>
<td>6</td>
<td>88132026</td>
<td>G</td>
<td>A</td>
<td>1117</td>
<td>-0.09</td>
<td>0.05</td>
<td>3.55E-02</td>
<td>0.3191040</td>
</tr>
<tr>
<td>rs109146371</td>
<td>14</td>
<td>16513111</td>
<td>T</td>
<td>C</td>
<td>1102</td>
<td>-0.17</td>
<td>0.05</td>
<td>9.34E-04</td>
<td>0.0084051</td>
</tr>
<tr>
<td>rs109234250</td>
<td>14</td>
<td>1802265</td>
<td>G</td>
<td>A</td>
<td>1102</td>
<td>-0.16</td>
<td>0.05</td>
<td>1.70E-03</td>
<td>0.0153054</td>
</tr>
<tr>
<td>rs109421300</td>
<td>14</td>
<td>1801116</td>
<td>T</td>
<td>C</td>
<td>1110</td>
<td>-0.15</td>
<td>0.05</td>
<td>2.43E-03</td>
<td>0.0218324</td>
</tr>
<tr>
<td>rs109162116</td>
<td>14</td>
<td>1804647</td>
<td>T</td>
<td>C</td>
<td>1112</td>
<td>-0.15</td>
<td>0.05</td>
<td>3.09E-03</td>
<td>0.0277911</td>
</tr>
</tbody>
</table>

Marker = accession number of SNP as in the bovine 50 K SNP Chip panel; BTA = Bos taurus autosome; Position = position (bp) on Bovine Genome Assembly UMD 3.1.1.; n = number of animals tested for the specific marker; effB = effect of the minor allele (B allele); se_effB = standard error of the effect of the minor allele; P-value = raw significance level of association; Bonferroni = Bonferroni-corrected significance level.
The KBiosciences Competitive Allele-Specific PCR SNP genotyping system (KASPAR™; KBioscience Ltd, Hoddesdon, UK) was used to genotype the nine SNPs. KASPAR™ assays were developed to target nine genome-specific SNPs on BTA6 and BTA14, details of SNP position and chromosome are shown in Table 1. All assay primer sets were designed using PrimerPicker (KBioscience Ltd.) with default parameters. The genotyping assay was carried out on genomic DNA and read on a Real Time ABI7900 by means of an allelic discrimination assay according to the protocol specified by the manufacturer.

Association analysis
Association analysis was conducted using the R statistical environment with the ‘qtscore’ function of the package GenABEL (Aulchenko et al., 2007). The response variable, namely the adjusted LS value, was analysed by fitting a single-SNP regression model. The significance of an association between a given SNP and the response was measured using a likelihood-based score test (Schaid et al., 2002). In order to deal with multiple testing a Bonferroni correction was applied to uncorrected P-values. Bonferroni correction is used to change the threshold value (α) = 0.05 against which the P-value is compared, into α/k (0.05/k) where k is the number of statistical tests performed.

Results

Results of the phenotypic analysis
Step 1: herd classification (high and low incidence herds). On average, HI herds (n = 69) had 40.7% of animals with high LS, with values ranging from 32.2% to 57.5%. In LI herds (n = 50), the mean percentage of animals with high LS was 22.2%, with herd means ranging from 12.2% to 29.9%.

Step 2: selection of case–control individuals (case and control animals). Control animals were selected within HI herds on adjusted LS values. After adjusting LS using model 1, its mean value in the control group was −1.11, ranging from −2.24 to −0.10. Their observed LS before adjustment ranged from 0.8 to 5 and the mean was 2.23. Case animals were selected from LI herds and showed a mean adjusted LS value of 1.27, ranging from 0.50 to 3.44 (Table 2), whereas their observed unadjusted LS values ranged from 2.09 to 6.7 with mean equal to 3.95. The distribution of the adjusted LS values of selected cows in HI and LI herds is shown in Figure 1.

Results of the Genotypic and association analysis
A total of nine SNPs were genotyped on 1121 samples (550 cases and 571 controls). Genotype efficiency ranged between 99.9% and 98.8%. Each SNP was considered significant at a threshold of α ⩽ 0.05 after Bonferroni correction for the nine tests.

The results of the association analysis on five SNPs located on BTA6 and four SNPs located on BTA14 are shown in Table 1. Six out of the nine markers showed significant association with SCC expressed as adjusted LS. The association was significant for the four SNPs located on BTA14, with P-values ranging from 0.008 to 0.028 after Bonferroni correction, with the same minor allele showing the equal effect on the phenotype (−0.15) and both showing a similar standard error of the effect 0.05.

Only two out of the five markers on BTA6 were in association with the trait, namely markers rs110527224 and rs42766480 with P-values of 0.002 and 0.003, respectively (Table 1).

Discussion
When relying on the identification of resistant/susceptible animals under field conditions, the identification of significant markers may be difficult and validation is not always successful. Moreover, access to correct phenotypes and correction for environmental noise in disease field data may be somewhat limited when all animals are not routinely
tested. However, for endemic diseases with phenotypes recorded on a regular basis, and if individual longitudinal data are available, as for mastitis, alternative approaches to reduce incomplete exposure can be applied. The synthetic phenotype for SCC developed in this work takes into account the dynamic aspects of the individual’s response or recovery over time as well as environmental and production effects and incomplete exposure through the prior step of classification into HI v. LI herds.

It is interesting to note that the present study with a cohort of 1121 cows selected in a two-step approach and a previous GWAS on a cohort of 1200 elite sires identified six (out of seven possible) common markers associated to mastitis. This convergence of results, however, is not always observed. For example, several GWAS on bovine paratuberculosis identified different chromosomal regions associated with infection with Mycobacterium avium. paratuberculosis (Gonda et al., 2007; Settles et al., 2009; Minozzi et al., 2010). Furthermore, these results extend and generalizes a previous validation of markers on BTA6 between 85.5 and 88.1 Mb, which were identified in GWAs on elite bulls and validated using logarithm transformed SCC measured in three lactations of cows from three German dairy farms similarly managed (Abdel-Shafy et al., 2014).

The SNPs on BTA6 that were found in the Italian cow population are located in a region where several quantitative trait loci (QTLs) have been identified for mastitis traits (Lund et al., 2008; Nilsen et al., 2009; Sodeland et al., 2011; Meredith et al., 2013; Sahana et al., 2013; Abdel-Shafy et al., 2014), and a similar association of SCS with rs42766480 has been identified in a study on American Holstein cattle (Cole et al., 2011). In addition, the analysis of EBV of progeny tested Nordic Holstein sires identified a region around 88.97 Mb on BTA6 of interest for several mastitis-related traits (Sahana et al., 2013). Likewise, QTL for clinical mastitis phenotypic data in a Norwegian Red cattle population have been identified in a region of chromosome BTA6 (Nilsen et al., 2009) with a granddaughter design. This region on chromosome 6 includes the casein gene cluster and several genes thought to be involved in mastitis as the immunoglobulin J chain that is located at 87.75 Mb, close to the SNPs identified (Sahana et al., 2013) and the Mucin 7 gene located on BTA6 from 87 573 430 to 87 584 044 bp (Gene ID: 101905278). In particular, the Mucin 7 gene encodes an antimicrobial peptide that is involved in the mucosal immune system (Liu et al., 2000) making it a good candidate.

Both markers have been found to be significant in a genome scan for clinical mastitis in Nordic Holstein cattle, with effects showing same direction and standard error (Sahana et al., 2013). GWAS studies on bovine mastitis-related traits have also reported QTL affecting susceptibility to mastitis on chromosome 14 (Schulman et al., 2004; Lund et al., 2007; Tiezzi et al., 2015) by using different populations and experimental plans.

In conclusion, this study has shown that results obtained through a GWAS on EBV for SCC from a population of elite sires were confirmed in a case–control approach with SCC field data from cows selectively chosen using individual and within herd epidemiological information. Most importantly, the current work has demonstrated the potential use of historical field disease data to set case–control designs for the study of genetic associations with infectious diseases in livestock, but further work would be needed to apply this procedure to other diseases. Of course, genetic studies for other infectious diseases may be more limited than for mastitis because availability of single or repeated individual phenotypic data may be lacking. Finally, while genomic information is widely accessible today, the most limiting step in extending the epidemiological approach to other diseases is the availability of longitudinal phenotypic data.

**Acknowledgements**

This work was funded by the MASTFIELD project no. 1745 of the Lombardy Region (Agricultural regional research program 2010–12). Massimo Battaglia and Lucio Zanini and ARAL Association are acknowledged for providing test-day records of all the animals used for this work.

**References**


Meredith BK, Berry DP, Kearney F, Finlay EK, Fahay AG, Bradley DG and Lynn DJ 2013. A genome-wide association study for somatic cell score using the Illumina high-density bovine beadchip identifies several novel QTL potentially related to mastitis susceptibility. Frontiers in Genetics 6, 229.


