Genome-wide association study identifies candidate markers for bull fertility in Holstein dairy cattle

F. Peñagaricano*, K. A. Weigel† and H. Khatib*

*Department of Animal Sciences, University of Wisconsin-Madison, 1675 Observatory Drive, Madison, WI 53706, USA: †Department of Dairy Science, University of Wisconsin-Madison, 1675 Observatory Drive, Madison, WI 53706, USA

Summary

The decline in the reproductive efficiency of dairy cattle has become a challenging problem worldwide. Female fertility is now taken into account in breeding goals while generally less attention is given to male fertility. The objective of this study was to perform a genome-wide association study in Holstein bulls to identify genetic variants significantly related to sire conception rate (SCR), a new phenotypic evaluation of bull fertility. The analysis included 1755 sires with SCR data and 38 650 single nucleotide polymorphisms (SNPs) spanning the entire bovine genome. Associations between SNPs and SCR were analyzed using a mixed linear model that included a random polygenic effect and SNP genotype either as a linear covariate or as a categorical variable. A multiple testing correction approach was used to account for the correlation between SNPs because of linkage disequilibrium. After genome-wide correction, eight SNPs showed significant association with SCR. Some of these SNPs are located close to or in the middle of genes with functions related to male fertility, such as the sperm acrosome reaction, chromatin remodeling during the spermatogenesis, and the meiotic process during male germ cell maturation. Some SNPs showed marked dominance effects, which provide more evidence for the relevance of non-additive effects in traits closely related to fitness such as fertility. The results could contribute to the identification of genes and pathways associated with male fertility in dairy cattle.

Keywords association study, bull fertility, cattle, single nucleotide polymorphism.

Introduction

Fertility is considered an important economic trait in cattle, yet despite its importance, reproductive efficiency of dairy cattle has decreased dramatically in the past 20 years and is of increasing concern to farmers and the dairy industry (Lucy 2001; Pryce et al. 2004). As a consequence, several countries are now including female fertility in their breeding goals and selection programs (Miglior et al. 2005; Weigel 2006). However, bull fertility has received much less consideration. Nevertheless, some studies have shown that a significant percentage of reproductive failure in dairy cattle is attributable to bull subfertility (e.g., DeJarnette et al. 2004). For this reason, some researchers argue that the fertility of bulls should not be overlooked to solve the emerging crisis of infertility in dairy cattle (Braundmeier & Miller 2001).

It is well documented that bull fertility is influenced by genetic factors. For instance, Druet et al. (2009) reported genetic parameters for several traits related to semen characteristics in Holstein bulls. Semen production traits, such as volume and sperm concentration, were found to have moderate heritabilities (from 0.15 to 0.30), whereas some of the semen quality traits, such as motility and percentage of abnormal sperm, had moderate to high heritabilities (close to 0.60) (Druet et al. 2009). These results suggest that genetic selection may be effective. Furthermore, high-throughput technologies, including massively parallel RNA expression and protein quantification, have revealed numerous differences between the spermatozoa of subfertile and fertile bulls (e.g., Peddinti et al. 2008; Feugang et al. 2010; Gaviraghi et al. 2010). Altogether,
these studies suggest that genetic factors may explain part of the differences in fertility among sires, and hence bull fertility could be improved by genetic means.

There is limited information on genes associated with bull fertility. Using a candidate pathway approach, Khatib et al. (2010) reported that polymorphisms in the FGF2 and STAT5A genes are associated with male fertility. Furthermore, Feugang et al. (2009), using a two-stage genome-wide association study, showed that the IGFB5 gene is involved in the fertilization process in Holstein bulls.

Identifying genomic regions, and preferably individual genes, responsible for genetic variation in bull fertility will enhance the understanding of biological pathways involving this trait and may point to opportunities for improving sire fertility via selective breeding. As such, the objective of this study was to perform a genome-wide association study in Holstein cattle using a high-density SNP array to identify genetic variants significantly related to sire conception rate, a phenotypic evaluation of bull fertility provided to dairy producers.

Materials and methods

Phenotypic and genotypic data

Since August 2008, sire conception rate (SCR), a phenotypic evaluation of bull fertility, has been provided to dairy producers by the Animal Improvement Programs Laboratory of the United States Department of Agriculture (AIPL-USDA). Kuhn & Hutchison (2008) and Kuhn et al. (2008) provided a complete description of the method employed for evaluating bull fertility. In brief, SCR is the expected difference in conception rate of a sire compared with the mean of all other evaluated sires. It should be emphasized that, in contrast to bull evaluations for traits such as milk production and conformation, bull fertility evaluation is intended as phenotypic rather than genetic evaluation, because published estimates reflect both genetic and permanent environmental effects.

In this study, a total of 1755 Holstein bulls with SCR data were used. Sire conception rate values from these bulls ranged from −10.66% to +6.80%, and the number of breedings per bull ranged from 303 to 111,402. Sire conception rate data were obtained from seven consecutive evaluations provided by AIPL-USDA between August 2008 and December 2010. For bulls with multiple evaluations, the most recent SCR evaluation of the bull was used in the analysis.

Genotypes of the 1,755 bulls for the Illumina BovineSNP50 Bead Chip were provided by AIPL-USDA. Single nucleotide polymorphisms (SNPs) with minor allele frequencies below 5% were removed. After data editing, 38,650 markers were available.

Statistical analysis

Two different models were used to analyze the association between SNP genotype and SCR. In the first model, SNP genotype was treated as a linear covariate using the following mixed linear model,

$$
SCR_{ijkl} = µ + EVAL_l + βSNP_k + sire_s + e_{ijkl}
$$

where $µ$ is the general mean, $EVAL_l$ is the fixed effect of the $l^{th}$ AIPL-USDA SCR evaluation ($l = 1, 2, ... , 7$), $SNP_k$ is the number of copies of one allele of the SNP (corresponding to 0, 1 or 2 copies) carried by the $i^{th}$ animal ($i = 1, 2, ... , 1755$), $β$ is the regression coefficient for the SNP considered (also known as the allele substitution effect), $sire_s$ represents the random additive genetic effect of the $i^{th}$ sire ($i = 1, 2, ... , 122$) of the $i^{th}$ animal, and $e_{ijkl}$ represents the random residual for each observation.

In the second model, SNP genotype was treated as a categorical variable and the association analysis was performed using the following mixed linear model,

$$
SCR_{ijkl} = µ + EVAL_l + SNP_k + sire_s + e_{ijkl}
$$

where $µ$, $EVAL_l$, $sire_s$, and $e_{ijkl}$ are as described in the first model, and $SNP_k$ is the fixed effect of the $k^{th}$ genotypic class ($k = 1, 2, 3$) for the SNP considered.

Random effects were assumed to follow the multivariate normal distribution,

$$
\begin{pmatrix}
\mathbf{s} \\
\mathbf{e}
\end{pmatrix} \sim N \left( \mathbf{0}, \begin{pmatrix}
\mathbf{A} & \mathbf{0} \\
\mathbf{0} & \mathbf{W}^{-1}\mathbf{σ}_r^2
\end{pmatrix} \right)
$$

where $\mathbf{s}$ and $\mathbf{e}$ are the vectors of sire and residual effects, respectively; $\mathbf{σ}_s^2$ and $\mathbf{σ}_r^2$ are the sire and residual effect variances, respectively; $\mathbf{A}$ represents the matrix of additive relationships between sires in the pedigree (1510 × 1510) and $\mathbf{W}$ is a diagonal matrix of order 1755 with its elements representing reliabilities of SCR values. The $\mathbf{A}$ matrix was calculated based on a five-generation pedigree of sires downloaded from AIPL.

The SNP term was removed for estimating the additive genetic variance. Moreover, for estimating the total variance explained by each SNP, the genotype effect was fitted as a random effect. Associations between the SNPs and SCR were tested using a likelihood ratio test by comparing to a reduced model without the SNP effect. All analyses were performed using the pedigreemm package (Vazquez et al. 2010) of the R language/environment (R Development Core Team 2009).

To correct for multiple testing, we applied the simpleM method (Gao et al. 2008). In this approach, the effective number of independent tests is determined by the degree of linkage disequilibrium (LD) among SNPs using a principal component analysis. First, for each chromosome, a correlation matrix for the SNPs was constructed using the composite LD. Then, eigenvalues were computed through principal component analysis of the composite LD matrix.
Finally, the effective number of SNPs per chromosome was calculated as the number of principal components required to jointly explain 99% of the variance in the SNPs. Because SNPs on different chromosomes are expected to be in linkage equilibrium in the general population, the total effective number of SNPs (Meff) used in this study was obtained by summing all effective number of SNPs for each chromosome. P-values from single SNP association tests were further adjusted for multiple comparisons using the Sidak correction based on Meff: adjusted P-value = 1 – (1 – P-value)^Meff. Manhattan plots of P-values were created using the gap package of the R language/environment (R Development Core Team 2009).

**Results and discussion**

Associations between SNPs and SCR were analyzed using the SNP allele count as a linear covariate and using the SNP genotype as a categorical variable. The first analysis has more power to detect significant SNPs with additive effects, while the second approach allows the detections of significant markers that show some degree of dominance (e.g. complete dominance or overdominance). Using SNP genotype as a linear covariate, five SNPs (located on chromosomes 5, 25, and 29) showed significant associations with SCR based on genome-wise adjusted P-value below 0.10 (Table 1). When SNP genotype was analyzed as a categorical variable, six SNPs (located on chromosomes 2, 5, 18, and 25) showed significant associations with SCR using a genome-wise significance level of 10% (Table 2). Three SNPs were significant in both models. Figure 1 shows the Manhattan plots of the P-values obtained from both types of analyses.

Twenty-four percent of the phenotypic variance was explained by the additive genetic component, and the proportion of the total variance explained by each significant SNP ranged from 1.9 to 7.4% (Tables 1 & 2). It should be noted that these percentages can be overestimated, especially when the effect of the SNP is small and the sample size is not large enough, a phenomenon called the Beavis effect (Xu 2003). Therefore, these estimations should be treated with caution.

The most significant SNP (ARS-BFGL-NGS-116417) associated with bull fertility using the SNP genotype as a linear covariate (adjusted P-value = 0.001, Table 1) or as a categorical variable (adjusted P-value = 0.005, Table 2) is located on chromosome 5, approximately 27 kb upstream of the LOC784935 gene. CPB/p300 homolog family member gene (cph-1)-like. cph-1 has key functions in the spermatogenesis of Caenorhabditis elegans (Luitjens et al. 2000). CPB-1 protein is present in the germ line just prior to overt spermatogenesis and is essential for successful progression through meiosis. Furthermore, silencing of this gene in spermatocytes using RNA-mediated interference causes failure of these cells to progress through meiotic divisions and, hence, to produce functional sperm (Luitjens et al. 2000). Given that ARS-BFGL-NGS-116417 and LOC784935 are only 27 kb apart and that strong LD can extend up to 50 kb in Holstein cattle (Kim & Kirkpatrick 2009), our findings suggest LOC784935 as a strong candidate gene for male fertility.

The SNP Hapmap44380-BTA-46707, located on chromosome 2, showed a significant association with SCR (adjusted P-value = 0.008) when SNP genotype was analyzed as a categorical variable (Table 2). This SNP showed a clear dominance effect, with a marked difference in the effect of one homozygous genotype (coded as 2) versus the other two genotypes (which presented almost the same effect). This SNP is located in intron 5 of the dynein, cytoplasmic 1, intermediate chain 2 (DYNCI12) gene.
Cytoplasmic dyneins are the principal microtubule motor proteins of the cell and are involved in many essential cellular processes, including spindle-pole organization and nuclear migration during mitosis and meiosis, the positioning and functioning of cellular organelles and the transport of vesicles along microtubules (Pfister et al. 2006). Interestingly, in Drosophila, some cytoplasmic dyneins have been found to be associated with proper postmeiotic spermatid development and male fertility (Li et al. 2004). The exact function of DYNC1I2 is not well understood, and our findings provide a foundation for further research to unravel the possible functions of this gene with respect to male fertility.

The marker Hapmap38225-BTA-43804, located on chromosome 18, showed a significant association with bull fertility (adjusted $P$-value $= 0.048$) using the SNP genotype as a categorical variable (Table 2). This SNP showed a clear overdominance effect, with a remarkable superiority of the heterozygous genotype. This marker is located approximately 16 kb upstream of the zinc finger protein 541 (ZNF541) gene. Interestingly, ZNF541 has been implicated in chromatin remodeling during

![Manhattan plots](image)

Figure 1 Manhattan plots of $P$-values for the genome-wide association study using SNP genotype as a linear covariate (a) or as a categorical variable (b). Dotted horizontal lines indicate the genome-wise significance level of 10%.
spermatogenesis. The ZNF541 gene encodes a nuclear protein expressed at specific stages of spermiogenesis and has a potential role in chromatin remodeling during the period of development from round to elongated spermatids (Choi et al. 2008). Furthermore, the ZNF541 protein, which contains specific motifs and domains potentially involved in DNA binding and chromatin reorganization, is localized in testicular spermatogenic cells in the mouse (Choi et al. 2008). Overall, our results provide further evidence for the possible role of ZNF541 in spermatogenesis and, hence, in male fertility.

Two SNPs located on chromosome 25, ARS-BFGL-NGS-4009 and ARS-BFGL-NGS-31020, showed associations with bull fertility in both statistical analyses. The associations were more significant using SNP genotype as a linear covariate. In this case, SNPs showed adjusted $P$-values equal to 0.013 and 0.010 for ARS-BFGL-NGS-4009 and ARS-BFGL-NGS-31020 respectively (Table 1). Although these two SNPs are more than 3.2 Mb apart, they showed remarkable high LD ($r^2 = 0.48$), and therefore it is possible that they represent the same signal. The marker ARS-BFGL-NGS-31020 is located approximately 60 kb from the LOC521021 gene, a member of the ectonucleotide pyrophosphatase/phosphodiesterase family. This gene family is involved in sphingolipid metabolism (Massé et al. 2010), and this is the first study that reports an association between LOC521021 and male fertility. ARS-BFGL-NGS-4009 is located in the last intron of the calcium channel, voltage-dependent, T type, alpha 1H subunit (CACNA1H) gene. This gene encodes a protein in the voltage-dependent calcium channel complex (Cribbs et al. 1998). Calcium channels mediate the influx of calcium ions into the cell upon membrane polarization. Calcium entry through low-voltage-activated calcium channels (also called T-type calcium channels) plays an essential role in various physiological events, including thalamic oscillation, muscle contraction, and hormone secretion (Bootman et al. 2001). Importantly, calcium flux across these channels also plays a central role during the acrosome reaction (Bootman et al. 2001). The acrosome reaction allows spermatozoa to penetrate the zona pellucida and fuse with the oocyte membrane (Brucker & Lipford 1995). This process is a crucial step during gamete interaction and, thus, fertilization. Therefore, our findings may provide more evidence of the association between CACNA1H and male fertility.

The marker ARS-BFGL-NGS-13272 showed a significant association with bull fertility (adjusted $P$-value = 0.010) when SNP genotype was used as a categorical variable in the model (Table 2). This SNP showed a marked dominance effect. ARS-BFGL-NGS-13272 is located on chromosome 25 and showed moderate LD ($r^2 = 0.30$) with ARS-BFGL-NGS-4009 and ARS-BFGL-NGS-31020. This SNP is sited in intron 4 of the ROGDI gene, for which functions are not well documented. However, ROGDI could be associated with an increase in the rate or extent of cell proliferation (Tulin et al. 2002). This is the first study that reports a possible relationship between ROGDI and male fertility. Additional studies are needed to confirm and characterize this relationship.

Another genetic marker located on chromosome 25, ARS-BFGL-NGS-13853, showed an association with bull fertility when SNP genotype was coded as a linear covariate (Table 1). This SNP did not show remarkable LD with the other significant markers located on chromosome 25. Although the statistical evidence for the association with SCR was not very strong (adjusted $P$-value = 0.099), this marker is located in a genomic region possibly related to male fertility. ARS-BFGL-NGS-13853 is located in the

---

**Table 2** Genetic markers associated with bull fertility using SNP genotype as a categorical variable in the analysis.

<table>
<thead>
<tr>
<th>Marker</th>
<th>Chr.</th>
<th>Position</th>
<th>Nearest gene (distance)</th>
<th>$N$ (genotype)</th>
<th>$\sigma^2_{\text{SNP}}/\sigma^2_{\text{TOTAL}}$</th>
<th>Genotype effect ± SE</th>
<th>Raw $P$-value</th>
<th>Adjusted $P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hapmap44380-BTA-46707</td>
<td>2</td>
<td>24837034</td>
<td>DYNC112 (within)</td>
<td>1037 (0)</td>
<td>7.0%</td>
<td>1.00 ± 0.02</td>
<td>6.35E-07</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>618 (1)</td>
<td></td>
<td>1.02 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARS-BFGL-NGS-116417</td>
<td>5</td>
<td>112775479</td>
<td>LOC784935 (27 kb)</td>
<td>41 (0)</td>
<td>7.4%</td>
<td>2.04 ± 0.04</td>
<td>4.45E-07</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>482 (1)</td>
<td></td>
<td>1.23 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1232 (2)</td>
<td></td>
<td>0.74 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hapmap38225-BTA-43804</td>
<td>18</td>
<td>54965977</td>
<td>ZNF541 (16 kb)</td>
<td>1355 (0)</td>
<td>2.7%</td>
<td>0.77 ± 0.02</td>
<td>4.00E-06</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>364 (1)</td>
<td></td>
<td>1.38 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>36 (2)</td>
<td></td>
<td>0.80 ± 0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARS-BFGL-NGS-4009</td>
<td>25</td>
<td>983759</td>
<td>CACNA1H (within)</td>
<td>812 (0)</td>
<td>3.0%</td>
<td>1.17 ± 0.02</td>
<td>5.02E-06</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>763 (1)</td>
<td></td>
<td>0.85 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>180 (2)</td>
<td></td>
<td>0.37 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARS-BFGL-NGS-13272</td>
<td>25</td>
<td>3898682</td>
<td>ROGDI (within)</td>
<td>905 (0)</td>
<td>4.8%</td>
<td>1.08 ± 0.02</td>
<td>8.46E-07</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>686 (1)</td>
<td></td>
<td>0.96 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164 (2)</td>
<td></td>
<td>0.15 ± 0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARS-BFGL-NGS-31020</td>
<td>25</td>
<td>4229011</td>
<td>LOC521021 (60 kb)</td>
<td>837 (0)</td>
<td>2.6%</td>
<td>1.18 ± 0.02</td>
<td>5.28E-06</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>729 (1)</td>
<td></td>
<td>0.80 ± 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>189 (2)</td>
<td></td>
<td>0.45 ± 0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LOC617302 gene, which is also known as testisin-like gene. Apparently, this is a pseudogene. Interestingly, this locus is upstream of another gene, protease, serine, 21 (PRSS21), also known as testisin. Testisin encodes a cell-surface anchored serine protease, which is a member of the trypsin family of serine proteases (Hooper et al. 1999). The protein localizes in pre-meiotic testicular germ cells and is likely involved in proteolytic reactions that are associated with male germ cell maturation (Hooper et al. 1999). Our study provides more evidence for the involvement of this gene in male fertility.

The SNP BTB-01354898, located on chromosome 29, showed an association with SCR when coded as a linear covariant, with a genome-wide adjusted P-value equal to 0.041 (Table 1). This SNP is sited in a genomic region where there are no reported genes. However, McClure et al. (2010) reported the presence of a quantitative trait locus (QTL) in this region that was significantly associated with scrotal circumference in Angus cattle. Scrotal circumference is positively correlated with sperm production and is used by farmers and veterinarians to predict bull fertility (Steifien 1997). Thus, our finding provides more evidence for the presence of one or more genes that affect bull fertility in this region of chromosome 29.

It is important to note that we have identified three SNPs with remarkable non-additive effects. Hapmap44380-BTA-46707 and ARS-BFGL-NGS-13272, located in DYNC1I2 and ROGDI respectively, showed marked dominance effects, whereas Hapmap38225-BTA-43804, located near ZNF541, showed a clear overdominance effect. Previous studies also have shown important dominance effects in genes affecting bull fertility (Khatib et al. 2010). It is known that non-additive genetic effects are important for traits closely related to fitness, such as fertility (Falconer & Mackay 1996). Our study provides more evidence of this phenomenon.

Multiple testing is a challenging concern in genetic association studies using large numbers of SNPs because of the high levels of LD; therefore tests performed on single SNPs are not independent. In this case, Bonferroni or Šidák corrections, popular approaches for controlling the experiment-wise error rate, would be overly conservative, resulting in a failure to detect true positive signals. A valid adjustment for multiple testing must account for the correlation between tests. In this study we applied the simpleM method (Gao et al. 2008), which was specifically developed to calculate the number of informative SNPs being tested in a genome-wide association study using principal component analysis. We estimated the effective number of tests (Meff = 12 236) as the number of principal components necessary to account for 99% of the variance in the SNP data. Importantly, it is well documented that the simpleM method is an effective and accurate approach for controlling the experiment-wise error rate in genome-wide association studies (Gao et al. 2010; Johnson et al. 2010).

Conclusions

In this study, we have identified a panel of SNPs that have significant effects on bull fertility, measured as sire conception rate, in Holstein dairy cattle. To the best of our knowledge, this is the first study that reports a genome-wide association for male fertility in cattle using a continuous trait measured in a large population. Interestingly, some of the significant SNPs were located in the vicinity of or in the middle of genes with roles in male fertility, such as the sperm acrosome reaction, chromatin remodeling during the spermatogenesis or the meiotic process during male germ cell maturation. These results could contribute to the identification of genes and pathways associated with male fertility in dairy cattle and subsequent use of marker-assisted selection for male fertility in commercial breeding schemes.

Acknowledgements

The authors thank scientists at the USDA-ARS Beltsville Agricultural Research Center (Beltsville, MD) for providing genotypic and phenotypic data. Also, we extend our thanks to Bill Taylor and the Condor Project staff at the University of Wisconsin for providing access to a large computer cluster that made the analyses possible. This study was supported by USDA Hatch grant No. WIS-142-PRJ16[H from the University of Wisconsin-Madison and by the Robert Draper Technology Innovation Funding, Graduate School, University of Wisconsin. Kent Weigel acknowledges the National Association of Animal Breeders (Colombia, MO). Special thanks are extended to Morris Soller for his continuous support and inspiration.

Conflicts of interest

The authors have declared no potential conflicts.

References


