Identification of sperm proteins as biomarkers of field fertility in Holstein-Friesian bulls used for artificial insemination

M. B. Rabaglino,1 C. Le Danvic,2 L. Schibler,2 K. Kupisiewicz,3 J. P. Perrier,4 C. M. O’Meara,4 D. A. Kenny,5 S. Fair,6 and P. Lonergan1*

1School of Agriculture and Food Science, University College Dublin, Belfield, Dublin 4, Ireland
2ALLICE, Innovation and Development, 149 rue de Bercy, 75012, Paris, France
3Viking Genetics, Randers, 8960, Denmark
4National Cattle Breeding Centre, Unit K4, M7 Business Park, Naas, County Kildare, W91 WF59, Ireland
5Teagasc Animal and Grassland Research and Innovation Centre, Grange, Dunsany, Meath, C15 PW93, Ireland
6Laboratory of Animal Reproduction, School of Natural Sciences, Biomaterials Research Cluster, Bernal Institute, Faculty of Science and Engineering, University of Limerick, Limerick, V94 T9PX, Ireland

**ABSTRACT**

Despite passing stringent quality control, bulls used in artificial insemination can vary significantly in their fertility, emphasizing the need for reliable markers of sperm quality. This study aimed to identify sperm proteins acting as biomarkers of fertility in 2 different populations of dairy bulls classified based on their field fertility. Semen was collected and cryopreserved from: 54 Holstein bulls located in Ireland, classified according to fertility indexes as low fertility (LF, n = 23), medium fertility (n = 14), or high fertility (HF, n = 17); and 18 Holstein bulls located in Denmark, classified as LF (n = 8) or HF (n = 10). The proteome was measured through liquid chromatography-mass spectrometry and data were analyzed with the R software. Differentially abundant proteins between HF and LF bulls and biomarker proteins were determined through a modified t-test and random forest, respectively, selecting 301 differentially abundant proteins and 34 biomarker proteins. The predictive ability of the 34 biomarkers was evaluated employing support vector machine as the classifier, using their abundance levels in the Irish bulls to train the model and in the Danish bulls for validation. The prediction accuracy was 94.4%, with only one HF bull misclassified, corresponding to the lowest fertility index bull in the HF group. The biomarkers more abundant in sperm of HF bulls enriched axoneme assembly and sperm motility (false discovery rate <0.05), according to functional analysis. In conclusion, a robust model coupled with the application of appropriate bioinformatic tools allowed the identification of functionally relevant sperm proteins predictive of the fertility of Holstein bulls used in artificial insemination.

**Key words:** proteomics, machine learning, bioinformatics, sire, assisted-reproductive technologies

**INTRODUCTION**

In livestock species of economic importance, such as cattle, reproductive efficiency is critical for achieving on-farm profitability (Shalloo et al., 2014). Thus, improving pregnancy per AI is an area of considerable basic and applied research interest. Much of the research effort in this regard has understandably been focused on the female, due in part to the well-described effects of lactation-induced metabolic stress on the quality of the oocyte, embryo, and reproductive tract environment (Walsh et al., 2011). Nonetheless, the fertility of the bull is a major contributor to overall reproductive performance and can have a major effect on productivity and economic return. This is true not only in herds where natural service is predominantly used but even more so where AI is employed, given that semen from elite bulls can be used simultaneously in several countries, frequently reaching tens of thousands of inseminations per year.

Artificial insemination is one of the most extensively used assisted-reproductive technologies, facilitating high selection intensity and the exploitation of a sire’s genetic value through widespread dissemination of his semen. Ejaculates of bulls used for AI are subjected to extensive quality control checks in breeding centers to minimize the chances of adverse pregnancy outcomes after semen is released into the field (Thundathil et al., 1999; Harstine et al., 2018). Semen assessments before cryopreservation, and post-thaw, to evaluate potential fertility include microscopy-based parameters, such as motility, viability and morphology, and other more advanced and objective methods, using computer-assisted sperm analysis and flow cytometry (Harstine et al., 2018).
Nevertheless, despite improvements and modernization of semen evaluation methods, semen from bulls used in AI still exhibits significant variation in field fertility, even after passing stringent quality control checks (Fair and Lonergan, 2018). The challenge is exemplified in the study of Sellem et al. (2015) who correlated bull fertility (153 ejaculates from 19 Holstein bulls) with a combination of sperm functional parameters. At best, these authors could account for no more than 40% of the variation in sire fertility. Taking a similar approach of using flow cytometric and computer-assisted sperm analysis-based parameters, our group recently reported that 47% of the variation in AI bull field fertility could be explained (Bernecic et al., 2021). High-throughput (omics) technologies have emerged as promising tools to determine seminal plasma or sperm components associated with fertility at the molecular level (Klein et al., 2022). Previous studies have employed bioinformatic approaches for the analysis of proteomic (Viana et al., 2018; Kasimanickam et al., 2019), transcriptionomic (Bissonnette et al., 2009; Feugang et al., 2010), or metabolomic (Menezes et al., 2019; Saraf et al., 2020) data from semen samples to identify proteins, mRNAs, or metabolites, respectively, as markers of bull fertility. However, a robust and repeatable set of such biomarkers is still lacking. The challenge of identifying potential biomarkers could benefit from the application of machine-learning approaches, which aim to find patterns in complex data sets such as the proteome (Li et al., 2022). The relevance of these methods can be demonstrated through validations with external, independent data.

We hypothesized that differences in fertility of AI sires can be predicted through the proteomic profile of their sperm. This study aimed to identify sperm proteins behaving as biomarkers of fertility in 2 different populations of dairy bulls whose semen had already passed quality control checks and was used in the field. One of the populations was employed to screen the proteins behaving as biomarkers of fertility in 2 different populations of dairy bulls whose semen had already passed quality control checks and was used in the field. The other was used to validate the sperm proteins as biomarkers of fertility. This study aimed to identify sperm proteins as biomarkers of fertility in 2 different populations of dairy bulls whose semen had already passed quality control checks and was used in the field.

### Materials and Methods

No human or animal subjects were used, so this analysis did not require approval by an Institutional Animal Care and Use Committee or Institutional Review Board.

### Animals

Frozen-thawed sperm from 3 ejaculates collected from Holstein-Friesian bulls located in AI centers in Ireland (n = 54) or Denmark (n = 18) classified as high fertility, medium, or low fertility (Ireland) or high- and low-fertility (Denmark) were pooled within bull and analyzed.

**Irish Bulls.** A panel of Holstein-Friesian bulls (n = 54) was selected from the national population from which cryopreserved semen was used commercially for AI in Ireland. Data on field fertility were obtained from the Irish Cattle Breeding Federation database based on an adjusted sire fertility index (FI; Berry et al., 2011). Sire fertility was defined as pregnancy to a given service identified retrospectively either from a calving event or where a repeat service (or a pregnancy scan) deemed the animal not to be pregnant. These raw data were then adjusted for factors including semen type (frozen, fresh), cow parity, month of service, day of the week when serviced, service number, cow genotype, herd, AI technician, and bull breed. The adjusted sire FI given for each bull was then weighted for the number of service records, resulting in an adjusted pregnancy rate. Holstein-Friesian bulls that had a minimum of 500 inseminations formed the base population (840 bulls), from which low fertility (LF; FI: −0.268 to −0.016, n = 23), medium fertility: (MF; FI: 0.023 to 0.029, n = 14), or high fertility (HF; FI: 0.058 to 0.072, n = 17) were selected. Some of these same bulls were used in previous studies by our group (Bernecic et al., 2021; Donnellan et al., 2021; O’Callaghan et al., 2021; Donnellan et al., 2022).

**Danish Bulls.** Danish bulls were selected from Holstein-Friesian population of 2,220 bulls used for AI in Denmark. Data on field fertility were obtained from Danish National Cattle Register and were expressed as 56-d nonreturn rates, adjusted for the parity of inseminated cow, month of service, herd, year, and season. Eighteen bulls with a minimum of 500 mating were selected for the study based on their 56-d nonreturn rates and semen availability. These bulls were classified as LF (FI: −0.07 to −0.049, n = 8) or HF (FI: 0.062 to 0.114, n = 10).

The mean FI of both populations of bulls was zero. Information about each bull can be found in Supplemental Table S1 (https://figshare.com/articles/dataset/20367708; Rabaglino and Lonergan, 2022a).

### Determination of the Sperm Proteome

**Protein Extraction.** Three straws of semen from each bull were thawed in a water-bath at 37°C for 30 s. The contents were centrifuged at 3,000 × g for 7 min at room temperature to obtain a sperm pellet. After removal of the supernatant, 4 washes were performed to eliminate the extender (1 in water and 3 in PBS, both supplemented with cOmplete Protease Inhibitor cock-
The resulting pellet was resuspended in 200 µL of extraction buffer (4% SDS, 0.2% deoxycholic acid, 50 mM Tris, 100 mM ammonium bicarbonate, 10 mM dithiothreitol, pH 8), boiled for 5 min and incubated for 2 h in an ultrasound bath. Complete lysis was checked by microscopy. After centrifugation to remove cellular debris, the supernatant was centrifuged at 12,000 × g for 15 min at 4°C, and the pellet dried for 5 min under nitrogen flux followed by resuspension in 0.4% SDS. Protein sample content was quantified using the BCA Protein Assay Kit (Thermo Fisher Scientific).

**Nano-Liquid Chromatography-Mass Spectrometry Analysis.** To carry out liquid chromatography-mass spectrometry analysis, 50 µg of sperm protein extracts were concentrated using stacking acrylamide SDS-PAGE gels. The migration was performed for 20 min at 150-V constant. Then, proteins were stained using colloidal Coomassie blue staining R-250 and the stained band was excised. In-gel digestion was performed with an automated protein digestion system, a MassPrep Station (Waters), and the gel pieces were dehydrated. The collected extracts were pooled with a 100% acetonitrile solution until the gel pieces were dehydrated. The collected extracts were pooled to a final volume of 60 µL. Excess of acetonitrile was evaporated at 37°C before analysis. Peptide mixtures were automatically fractionated onto commercial C18 reverse phase column (75 µm × 150 mm, 2-µm particle, PAPMap100 RSLC column, Thermo Fisher Scientific) at 35°C. Trapping was performed for 4 min at 5 µL/min, with 98% H₂O, 2% acetonitrile, and 0.1% formic acid. Elution was performed using 2 solvents, A (0.1% formic acid in water) and B (0.1% formic acid in acetonitrile) at a flow rate of 300 nL/min. Gradient separation was 2 min from 2 to 5% B, 12 min from 5 to 25% B, 2 min from 25 to 80% B, 3 min 80% B. The column was equilibrated for 8 min with 2% buffer B before the next sample analysis. Eluted peptides were electro-sprayed in positive-ion mode at 1.9 kV through a nanoelectrospray ion source heated at 275°C. Full MS scans were acquired in the Orbitrap mass analyzer over m/z 400–1,200 range with a resolution of 70,000 (m/z 200). The target value was 5.00E+5 and the maximum allowed ion accumulation times were 250 ms. Fifteen most intense peaks with charge state between 2 and 5 were fragmented in the HCD collision cell with 27 eV, and tandem mass spectra were acquired with a resolution of 17,500 (m/z 200). The target value was 5.00E+4 and the maximum allowed ion accumulation times were 150 ms. Dynamic exclusion was set to 7 s.

**Data Processing, Protein Identification, and Abundance.** Raw data from the MS/MS were processed and converted into *.mgf peak list format with Proteome Discover 1.4 (Thermo Fisher Scientific). Data were interpreted with Mascot v2.4 (Matrix Science) against the *Bos taurus* database (i.e., 32,284 sequences) fused with the sequences of recombinant trypsin and a list of classical contaminants (118 entries). The following parameters were considered for the search: precursor mass tolerance of 0.2 Da and fragment mass tolerance of 0.2 Da, a maximum of one miss cleavage sites of trypsin, carbamidomethylation (C), oxidation (M), propionamidation (C), and protein N-terminal acetylation set as variable modifications. For each sample, peptides were filtered out according to the cut-off set for proteins hits with 2 or more peptides larger than 9 residues, ion score >15, false discovery rate <1%. Protein identification was validated when at least 2 peptides originated from one protein showed statistically significant identity.

**Bioinformatic Analysis of the Proteomic Data**

**Determination of Differentially Abundant Proteins and Identification of Putative Biomarker Proteins in the Sperm of Bulls Located in Ireland.** All bioinformatic procedures were carried out using the R software (R Core Team, 2020). Data on protein abundance were log10 transformed, filtered to retain proteins present in more than 14 samples (smaller class), and normalized with the Combat method (Leek et al., 2021) to control for the potential effect of different diluents. Data from the 2,156 proteins retained after filtering (out of 2,742) were analyzed with the limma package (Smyth, 2005). Briefly, an empirical Bayes method was applied to calculate a moderated t statistic for differential abundance for each protein by performing a linear model fit on the data. Empirical
Bayes moderated the standard errors of the estimated log fold changes to produce more stable estimates. A contrast was done between HF versus LF bulls, to determine those differentially abundant sperm proteins (DAP, \(P < 0.05\)) between bulls at the extremes of the FI, which were visualized by a volcano plot. A volcano plot is a type of scatterplot that shows statistical significance (\(P\)-value) for the statistical comparison between 2 groups on the y-axis versus the magnitude of change between the 2 groups on the x-axis. Each dot represents a protein; those on the right (in red) are proteins upregulated (i.e., more abundant in the HF bulls), whereas those on the left (in green) are downregulated (i.e., more abundant in the LF bulls). The remainder (in black) are not differentially abundant between HF versus LF bulls. However, a \(P > 0.05\) and < 0.1 was considered a tendency.

Biomarker proteins were selected through a wrapper algorithm around random forest using the Boruta package (Kursa and Rudnicki, 2010). Briefly, the algorithm adds shadow features to the data to train a random forest classifier, evaluating the importance of each feature (protein) through a Mean Decrease Accuracy. Selected features are those with a higher score than the best of their shadow features. The algorithm was set to run a maximum of 100,000 interactions, to select important features with a \(P < 0.0001\).

Sample distribution and clustering, according to the abundance of the DAP and biomarker proteins, were evaluated through a principal component analysis (PCA) and a hierarchical clustering dendrogram, respectively, using internal packages of R. Briefly, a PCA is a dimension reduction technique that captures sample variability in several components (corresponding to the number of samples, 40 in this case), with component 1 being the one that explains most of the variability. The PCA plot is a type of scatterplot showing the result of the PCA, where samples are clustered based on their similarity. Hierarchical clustering is an algorithm that groups samples according to their similarities (in protein abundance levels for this study). A Spearman Rank Correlation was employed as the similarity metric and complete linkage as the clustering method. Plotting and coloring of the dendrogram were done with the weighted gene co-expression network analysis (WGCNA) package (Langfelder and Horvath, 2008).

The lists of DAP and biomarkers more or less abundant in the HF bulls compared with the LF bulls were interrogated for enriched functional terms (false discovery rate <0.05) with ShinyGO v0.75 (Ge et al., 2020). Enriched terms were visualized through lollipop plots using the ggplot2 package (Wickham, 2016) and networks downloaded from the ShinyGo software.

**Validation of the Biomarker Proteins With an External Population.** Data from the sperm of bulls located in Denmark were used to validate the selected biomarkers. The training data consisted of the abundance data of the biomarker proteins in sperm from the bulls located in Ireland. In contrast, the abundance of those proteins in the sperm of bulls located in Denmark was employed to test the model. The testing data were normalized with the training data through an add-on batch effect adjustment with the bapred package (Hornung et al., 2017). The classifier was a support vector machine with linear kernels, employing the leave-one-out cross-validation (LOOCV) method as the internal control, and applied with the kernlab package (Karatziogou et al., 2004), through the caret package (Kuhn, 2008).

**RESULTS**

There were 301 DAP in the sperm from HF and LF bulls located in Ireland, of which 106 and 195 proteins were more abundant in the HF or LF bulls, respectively (Figure 1A and Supplemental Table S2; https://figshare.com/articles/dataset/20367708; Rabaglino and Lonergan, 2022b), which separated samples from each group in the PCA plot (Figure 1B). This plot shows that samples belonging to the HF and LF bulls cluster apart in the first component, which captures the main source of variability among the samples, explaining 18.9% of overall variability when considering the 40 components. Samples from LF bulls are more spread in both the first and second component than samples from HF bulls, reflecting the largest standard deviation (SD) in the FI of LF bulls than of HF bulls (mean FI ± SD for: HF bulls: 0.065 ± 0.004; LF bulls: −0.064 ± 0.059). Those DAP which were more abundant in HF bulls enriched 15 biological terms mainly related to the axoneme and cilium assembly, and ribonucleotide binding (Figure 2A). However, the 195 DAP that were more abundant in the sperm of LF bulls enriched 221 terms involved mainly in metabolism, proteosome and cell-cell recognition, including binding of sperm to the zona pellucida (Figure 2B).

Application of the machine learning method selected 34 DAP after 27,061 iterations (Table 1). According to the abundance levels of these proteins, samples from HF and LF bulls clustered apart (Figure 3A). When samples from the MF bulls were considered, most of them tended to cluster with those from LF bulls, with only 2 clustering with the HF bulls, as can be
Figure 1. Differentially abundant proteins (DAP) between the sperm of high fertility (HF) and low fertility (LF) bulls. (A) The volcano plot shows the DAP more abundant in HF bulls (106 red dots) or LF bulls (195 green dots). (B) Principal component (PC) analysis plot depicting distribution of samples from HF and LF bulls according to the DAP. FDR = false discovery rate; FC = fold change.
appreciated in the hierarchical clustering (Figure 3B). Accordingly, the PCA plot showed, in the first component, that although samples from MF bulls clustered in between samples from HF and LF bulls, most of them were on the right side of the plot, together with samples from the LF bulls (Figure 3C). Functional analysis of the 19 biomarker proteins more abundant in HF Irish bulls revealed enrichment of ontological terms involved mainly in the axoneme and sperm motility (Figure 4). These terms were enriched because of the following proteins (coded by the following genes): E1BMD1 (CFAP43, cilia and flagella associated protein 43); E1BAJ3 (CCDC40, coiled-coil domain containing 43); and F1N5R7 (DNAH7, dynein axonemal heavy chain 7). The 15 biomarker proteins more abundant in LF Irish bulls enriched only 2 terms: hexose and monosaccharide metabolic process.

Construction of the machine learning model using the abundance levels of the 34 biomarker proteins identified in the samples from HF and LF bulls in the Irish population led to a prediction of the HF or LF bulls in the Danish validation population with 94.4% accuracy: all 8 LF samples were correctly classified as such, whereas 9 out of 10 HF samples were correctly classified as HF. The only HF sample misclassified corresponded to the bull with the lowest FI among the HF bulls (bull 64). This high accuracy could not be reached if the model was built using the abundance levels of all 301 DAP; although all the 8 LF samples were predicted as such, only 3 of the 10 HF samples were correctly classified.
(61% accuracy). Selection of the 301 DAP were based on the raw P-value and, therefore, it is possible that several of those proteins were considered DAP when in fact they were not differentially abundant (i.e., false positive). Thus, a further selection of important sperm proteins for fertility through machine learning tools improved pinpointing those that would behave as such in an external independent population. This fact was reflected when the hierarchical clustering was performed on all 27 HF and all 31 LF samples from bulls in both populations. Considering the abundance of the 301 DAP, 3 LF samples clustered with the HF samples but 10 HF samples clustered with the LF samples (Figure 5A). However, only 3 HF samples clustered with the LF samples when the abundances of the 34 biomarkers were used to construct the hierarchical clustering (Figure 5B).

**DISCUSSION**

Genomically assisted selection has revolutionized how the dairy industry identifies and selects bulls for use in AI. However, even though it has increased the reliability of breeding values and shortened the generation interval, the reliable prediction of the field fertility of individual bulls is still challenging. This was never more important, given the widespread use of first season sires internationally before data are available on their field fertility. Identification of bulls of compromised fertility before their semen is released into the field is critical to avoid the economic effect of low pregnancy rates, particularly in seasonal systems of production with short breeding seasons where compact calving patterns are essential (Shalloo et al., 2014).

A reliable in vitro test, or a combination of tests predictive of bull field fertility, would enable AI companies to identify and eliminate bulls with potential LF before their widespread use in the field, thus improving overall reproductive and productive efficiency. Most AI centers use pre- and post-thawing assessment of the sperm through traditional microscopy-based methods, more modern computational approaches such as computer-assisted sperm analysis, or assays using fluorescent staining and high sample throughput, such

| UniProt | Symbol | Gene ID | Name | Fold change | P-value |
|---------|--------|---------|------|-------------|---------|
| O77834 | PRDX6  | 282438  | Peroxiredoxin 6 | 6.3552 | 0.0003 |
| A6QP39 | MSLN   | 516237  | Mesothelin | 3.7983 | 0.0032 |
| A4IF7 | ARF5   | 511918  | ADP ribosylation factor 5 | 3.6376 | 0.0052 |
| A6QP01 | SELENOT | 783831  | Selenoprotein T | 3.5342 | 0.0063 |
| O46385 | SVIL   | 281506  | Supervillin | 3.3720 | 0.0070 |
| Q2KI2 | MS4A5  | 777599  | Membrane spanning 4-domains A5 | 2.6080 | 0.0074 |
| P43896 | TSFM   | 281551  | Elongation factor Ts, mitochondrial | 2.3838 | 0.0019 |
| F1MB7 | TGM1   | 407997  | Transglutaminase 1 | 2.2742 | 0.0129 |
| A2V2D8 | RUSC2  | 507809  | RUN and SH3 domain containing 2 | 2.0932 | 0.0099 |
| E1BM8 | SBNO1  | 540582  | Strawberry notch homolog 1 | 2.0070 | 0.0298 |
| A8E4P2 | FARSB  | 788792  | Phosphodynamyl-tRNA synthetase subunit beta | 1.9099 | 0.0297 |
| Q72PH2 | TMEM143 | 505862  | Transmembrane protein 143 | 1.7967 | 0.0400 |
| Q08DF4 | DNM1   | 508794  | Dynamin 1 | 1.7513 | 0.0038 |
| Q07336 | ALDH6A1 | 327692  | Alddehyde dehydrogenase 6 family member A1 | 1.6952 | 0.0002 |
| F1MIH9 | KPN4A  | 540719  | Karyopherin subunit alpha 3 | 1.6944 | 0.0302 |
| E1B16 | FBXL13 | 101965084 | F-box and leucine rich repeat protein 13 | 1.5639 | 0.0035 |
| E1BMD1 | CFAP43 | 518801  | Cilia and flagella associated protein 43 | 1.2611 | 0.0042 |
| E1BAJ3 | CCTD20 | 529788  | Coiled-coil domain containing 40 | 1.1488 | 0.0019 |
| F1N57R | DNAH7  | 517781  | Dynein axonemal heavy chain 7 | 1.0550 | 0.0462 |
| E1BHC8 | ESPN   | 104976499 | Espin | 1.6091 | 0.0042 |
| Q148H0 | APOO   | 513847  | Apolipoprotein O | 1.6257 | 0.0075 |
| Q59B7 | CLC1   | 515646  | Chloride intracellular channel 1 | 1.3651 | 0.0018 |
| P16368 | TIMP2  | 283093  | TIMP metallopeptidase inhibitor 2 | 1.3830 | 0.0335 |
| E1BPS1 | ANKRD42 | 522004  | Ankyrin repeat domain 42 | 1.2695 | 0.0053 |
| G5EM8 | LOC532207 | 532207  | Epitidymis-specific alpha-mannosidase-like | 1.2573 | 0.0023 |
| A6QP36 | LMAN2  | 790870  | Lectin, mannose binding 2 | 1.2636 | 0.0088 |
| P23932 | SPADH1 | 282373  | Spermadhesin 1 | 2.8615 | 0.0261 |
| F1MQN0 | PPP2R1B | 100138150 | Protein phosphatase 2 scaffold subunit Abeta | 3.2250 | 0.0028 |
| Q3TH0 | LCTMT1 | 618021  | Leucine carboxyl methyltransferase 1 | 3.3412 | 0.0063 |
| G3X7H8 | LOC617406 | 617406  | Serpin peptidase inhibitor, clade B (ovalbumin), member 6-like | 3.5103 | 0.0118 |
| F1MQ88 | FHDC1  | 784913  | FH2 domain containing 1 | 3.8545 | 0.0063 |
| Q58DL1 | ARG2   | 518752  | Arginase 2 | 3.9494 | 0.0001 |
| F1NIE9 | LEPR   | 497205  | Leptin receptor | 4.7861 | 0.0020 |
| A5PJ5 | GET3   | 504586  | Guided entry of tail-anchored proteins factor 3, ATPase | 6.0164 | 0.0031 |
as flow cytometry (Harstine et al., 2018). Nevertheless, in terms of pregnancy rate after AI, fertility in the field can vary by up to 20 to 30%, despite the semen being deemed acceptable after stringent quality control measures (Donnellan et al., 2022). Thus, there is still a need to identify sperm biomarkers that can reliably and repeatedly predict bull fertility. The use of omics technologies combined with traditional bioinformatic approaches, such as regression analysis for hypothesis testing, could determine potentially relevant molecules from the vast data generated through these techniques. Furthermore, the application of machine learning tools could help identify the most important molecules, as these methods are not limited to the scores from the statistical test. In this sense, a random forest classifier is a powerful wrapper method for feature selection, particularly when applied with the Boruta algorithm (Kumar and Shaikh, 2017). In this study, we have employed both statistical and machine learning methods to select sperm proteins characterizing bulls classified according to their fertility in the field. We used the protein data of the sperm from bulls classified in both extremes (HF and LF) to avoid confusion in the data that could be introduced by bulls classified in the middle (MF). Semen collected from these animals had satisfactorily passed all the quality control measures carried out before and after thawing. In other words, animals classified as LF were not infertile, but, in the overall population, semen from LF bulls achieved a lower pregnancy rate than those from HF bulls.

Among the various molecules that can be measured through high-throughput techniques, protein presence or abundance have been related to male fertility in several species (Druart et al., 2019; Griffin et al., 2020; Leahy et al., 2020; Mills et al., 2020), and may represent a better source than transcripts for finding biomarkers, as sperm are transcriptionally silent. Several reports have employed proteomics in sperm, seminal plasma, or both, from bulls with divergent fertility (Peddinti et al., 2008; D’Amours et al., 2010; Soggiu et al., 2013; Somashekar et al., 2017; Kasimanickam et al., 2019). In a recently published review, Klein et al. (2022) identified the common, more abundant proteins across several studies, in HF or LF bulls. Among the 28 DAP they

![Figure 3. Distribution of samples according to the abundance levels of the 34 sperm proteins selected as biomarkers. Hierarchical clustering of the samples from (A) high fertility (HF) and low fertility (LF) bulls (red and blue colors, respectively) and (B) including the medium fertility (MF) bulls (green color). (C) Principal component (PC) analysis plot for samples from HF, MF, and LF bulls.](image-url)
identified in sperm, 23 are present among the sperm proteins detected in the current study. Six of the 23 proteins were differentially abundant between HF or LF bulls \((P < 0.1)\), and 2 are consistent with the results found by other authors: CCT5 (chaperonin containing TCP1 subunit 5; \(P = 0.005\)) and PEBP4 (phosphatidylethanolamine binding protein 4; \(P = 0.06\)), which were more abundant in LF or HF bulls, respectively. The abundance of CCT5 in sperm decreased with fertility in bulls from the Irish population, whereas it tended to decrease in bulls from the Danish population (Figure 6A), consistent with previous findings (D’Amours et al., 2010; Kasimanickam et al., 2019). This protein is a chaperonin involved in protein folding, particularly of actin and tubulin (Sternlicht et al., 1993). Cytoplasmatic CCT in the spermatid should be discarded in residual bodies at spermiation, at least in rats (Souès et al., 2003). Therefore, the higher abundance of this protein in sperm samples could indicate an incomplete process during spermatogenesis, potentially comprising fertility.

Apart from CCT5 and PEBP4, the other 21 proteins in common with our study and that of Klein et al. (2022) were not differentially abundant or did not follow the same trend in our study, most likely due to the many factors involved in fertility and differences in animal and sperm handling across the experiments. Among these factors is the different diluents used during the cryopreservation of sperm; nonetheless, it is worth noting that we applied a bioinformatic approach to remove the potential effect of the diluent, as it has been shown that it can affect sperm parameters (Murphy et al., 2018).

Sperm proteins have great potential as predictive biomarkers of fertility. Nevertheless, although each new study provides increased clarity, the complex nature of male fertility, and the multifactorial nature of the causes of subfertility, means that the selection of a single marker of male fertility is unlikely. In contrast, the combined abundance levels of many proteins can lead to a reliable and repeatable indicator of fertility. In this study, biomarker proteins were selected in one of the bull populations (Ireland) and validated in the other (Denmark), classifying the sperm samples correctly from this latter population with 94.4% accuracy. Among the proteins selected as potential biomarkers, CFAP43, CCDC40, and DNAH7 may have a major role in fertility, as they are involved with axoneme assembly and cilium movement, including sperm motility and regulation of the beating frequency. Indeed, the abundance of these proteins increased with increasing fertility in the sperm from bulls in the Irish population. For bulls in the Danish population, the abundance of DNAH7 increased, whereas CCDC40 tended to increase in the sperm of HF bulls compared with LF bulls. The abundance of CFAP43 followed the same pattern, but it was not different \((P = 0.1; \text{Figure 6 B–D})\). Although these proteins have not previously been associated with fertility in cattle, several studies in other species have found an association between mutations in the coding genes and infertility. Two independent studies have reported that mutations in \(CFAP43\) and \(CFAP44\) are...
Figure 5. Hierarchical clustering of sperm samples from high fertility (HF) and low fertility (LF) bulls from 2 different populations. Samples were clustered according to the abundance levels of (A) 301 differentially abundant proteins (DAP) between sperm of HF and low fertility LF bulls (red and blue colors, respectively), or (B) 34 DAP selected as biomarkers. Sperm samples obtained from bulls located in Ireland or Denmark are represented in gray or black colors, respectively.
Figure 6. Box plots showing the abundance levels for selected proteins in the sperm of bulls with high fertility (HF; red boxes), medium fertility (MF; green boxes), and low fertility (LF; blue boxes). The top or bottom plots correspond to data from bulls from an Irish or Danish population, respectively. CCT5: chaperonin containing TCP1 subunit 5; CFAP43: cilia and flagella associated protein 43; CCDC40: coiled-coil domain containing 40; and DNAH7: dynein axonemal heavy chain 7. The solid or dotted horizontal bars show differences ($P < 0.05$) or a tendency toward a difference ($P > 0.05$ and $< 0.1$), respectively, in protein abundance between groups after ANOVA. Upper and lower edges of boxes represent the third and first quartiles, respectively. The midlines show the median, or second quartile. The upper and lower whiskers extend from the third and first quartile to the maximum and minimum values, respectively. Dots represent extreme values, i.e., higher or lower than the third or first quartiles, respectively, plus 1.5 times the interquartile range (difference between the third and first quartile).
related to male infertility in men and mice because of axonemal disorganization in the sperm tail, which can lead to 100% of immotile, morphologically abnormal sperm if the mutation is biallelic (Tang et al., 2017; Coutton et al., 2018). Mutations in \textit{CCDC39} and \textit{CCDC40} are involved in primary ciliary dyskinesia, characterized by abnormal ciliary motility because of axonemal disorganization (Blanchon et al., 2012; Liu et al., 2021). Finally, mutations in \textit{DNAH7}, which is part of the ciliary or flagellar axonemal inner dynein arm-heavy chain (Neesen et al., 1997), are a cause of idiopathic asthenozoospermia in men (sperm with low motility) given the defects in the structure and function of the sperm flagellum (Wang et al., 2020; Wei et al., 2021). The sperm of bulls employed in this study and classified as LF passed the commercial quality control parameters, including progressive motility. This would suggest that the issues leading to LF were post-insemination events, for example, sperm transport in the female reproductive tract, ability to fertilize the oocyte, or to sustain the embryo.

**CONCLUSIONS**

The development of an accurate and robust bull fertility prediction model is challenging. Ultimately, the combination of state-of-the-art genetic, physiological, immunological, and molecular approaches, combined with in vitro sperm functional and bioinformatic analyses will likely lead to a better prediction of field fertility. In this study, we screened the sperm proteome in bulls with high- or low-field fertility to identify potential biomarker proteins. Application of bioinformatics and machine learning methods identified 34 sperm proteins, the abundance of which could predict fertility in an independent population of bulls with 94% accuracy. Some of these proteins, more abundant in HF bulls, were involved in sperm motility. Thus, this study introduces a set of proteins that could help to discriminate the semen from bulls with undesirable performance when used in AI.

**ACKNOWLEDGMENTS**

This project was funded by Science Foundation Ireland (Dublin, Ireland; Project 16/1A/4474). MBL was supported by an H2020-MSCA-Individual Fellowship (Proposal 101021311). The authors have not stated any conflicts of interest.

**REFERENCES**

Bernecic, N. C., E. Donnellan, E. O’Callaghan, K. Kupischwicz, C. O’Meara, K. Weldon, P. Lonergan, D. A. Kenny, and S. Fair. 2021. Comprehensive functional analysis reveals that acrosome integrity and viability are key variables distinguishing artificial insemination bulls of varying fertility. \textit{J. Dairy Sci.} 104:11226–11241. https://doi.org/10.3168/jds.2021-20319.

Berry, D. P., R. D. Evans, and S. Mc Parland. 2011. Evaluation of bull fertility in dairy and beef cattle using cow field data. \textit{Theriogenology} 75:172–181. https://doi.org/10.1016/j.theriogenology.2010.08.002.

Bissonnette, N., J. P. Levesque-Sergerie, C. Thalbault, and G. Boissonneault. 2009. Spermatozoal transcriptome profiling for bull sperm motility: a potential tool to evaluate semen quality. \textit{Reproduction} 138:65–80. https://doi.org/10.1530/REP-08-0503.

Blanchon, S., M. Legendre, B. Copin, P. Duquesnoy, G. Montantin, E. Kott, F. Dastot, L. Jeanson, M. Cachanado, A. Rousseau, J. F. Papon, N. Beydon, J. Brouard, C. Breton, A. Deschildre, J. Desir, H. Dolfusi, B. Leheup, A. Tamalet, C. Thuret, A. M. Vojtek, D. Escalier, A. Coate, J. de Blic, A. Clement, E. Escudier, and S. Amelem. 2012. Delineation of CCDC39/CCDC40 mutation spectrum and associated phenotypes in primary ciliary dyskinesia. \textit{J. Med. Genet.} 49:410–416. https://doi.org/10.1136/jmedgenet-2012-100867.

Coutton, C. A., S. Vargas, A. Amir-Yekta, Z. E. Kherras, S. F. Ben Mustapha, P. Le Tanno, C. Wamberge-Legrand, T. Kararuzene, C. T. Gonzalez, S. Crouzy, H. Ninomi, V. Mitchell, L. Halouni, O. Marrakech, M. Makni, H. Latarou, M. Kharrar, J. F. Deleuze, A. Boland, S. Hennebicq, V. Satre, S. P. Jouk, N. Thierry-Mieg, B. Conne, D. Dacheux, N. Landrein, A. Schmitt, L. Stouvenel, P. Loes, E. El Khouri, S. P. Bottari, J. Faure, J. P. Wolf, K. Pernet-Gallay, J. Escoffier, H. Gourabi, D. R. Robinson, S. Nef, E. Duliou, R. Zouari, M. Bonhivers, A. Toue, C. Arnault, and P. F. Ray. 2018. Mutations in \textit{CCDC39} and \textit{CCDC40} cause male infertility and flagellum defects in Trypanosoma and human. \textit{Nat. Commun.} 9:686. https://doi.org/10.1038/s41467-017-0279-2.

D’Amours, O., G. Frenette, M. Fortier, P. Leclerc, and R. Sullivan. 2018. Proteomic comparison of detergent-extracted sperm proteins from bulls with different fertility indexes. \textit{Reproduction} 139:545–556. https://doi.org/10.1530/REP-18-0375.

Donnellan, E. M., P. Lonergan, K. G. Meade, and S. Fair. 2022. An ex vivo assessment of differential sperm transport in the female reproductive tract between high and low fertility bulls. \textit{Theriogenology} 181:42–49. https://doi.org/10.1016/j.theriogenology.2022.01.011.

Donnellan, E. M., M. B. O’Brien, K. G. Meade, and S. Fair. 2021. Comparison of the uterine inflammatory response to frozen-thawed sperm from high and low fertility bulls. \textit{Theriogenology} 176:26–34. https://doi.org/10.1016/j.theriogenology.2021.09.012.

Druart, X., J. P. Rickard, G. Tsikis, and S. P. de Graaf. 2019. Seminal plasma proteins as markers of sperm fertility. \textit{Theriogenology} 137:30–35. https://doi.org/10.1016/j.theriogenology.2019.05.004.

Fair, S., and P. Lonergan. 2018. Review: Understanding the causes of variation in reproductive wastage among bulls. Animal 12(s1):s53–s62. https://doi.org/10.1177/1751731118000964.

Feugang, J. M., N. Rodriguez-Osorio, A. Kaya, H. Wang, G. Page, G. C. Ostermeier, E. K. Topper, and E. Memili. 2010. Transcriptome analysis of bull spermatozoa: Implications for male fertility. Reprod. Biomed. Online 21:312–324. https://doi.org/10.1016/j.rbmo.2010.06.022.

Ge, S. X., D. Jung, and R. Yao. 2020. ShinyGO: A graphical gene-set enrichment tool for animals and plants. Bioinformatics 36:2628–2629. https://doi.org/10.1093/bioinformatics/btz931.

Griffin, R. A., A. Swegen, M. Baker, R. J. Aitken, D. A. Sorkett-Byrne, A. Silva Rodriguez, F. E. Martin-Can, B. Nixon, F. J. Pena, M. Delehedde, N. Sergeant, and Z. Gibb. 2020. Mass spectrometry reveals distinct proteomic profiles in high- and low-quality stallion spermatozoa. \textit{Reproduction} 160:695–707. https://doi.org/10.1530/REP-20-0284.

Harstine, B. R., M. D. Ull, and J. M. DeJarnette. 2018. Review: Integrating a semen quality control program and sire fertility at a large artificial insemination organization. Animal 12(s1):1–12. https://doi.org/10.1177/1751731118000319.
Hornung, R., D. Causier, C. Bernau, and A. Boulesteix. 2017. Improving cross-study prediction through addon batch effect adjustment or addon normalization. Bioinformatics 33:397–404.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis. 2004. kernlab - An S4 package for kernel methods in R. J. Stat. Softw. 11:1–20. https://doi.org/10.18637/jss.v011.i09

Kasimanickam, R. K., V. R. Kasimanickam, A. Arangasamy, and J. P. Kastelic. 2019. Sperm and seminal plasma proteomics of high-versus low-fertility Holstein bulls. Theriogenology 126:41–48. https://doi.org/10.1016/j.theriogenology.2018.11.032

Klein, E. K., A. Sweeney, A. J. Gunn, C. P. Stephen, R. J. Aitken, and Z. Gibb. 2022. The future of assessing bull fertility: Can the ‘omics fields identify usable biomarkers? Biol. Reprod. 106:854–864. https://doi.org/10.1093/biolreio/biaa031

Kuhn. M. 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28:5. https://doi.org/10.18637/jss.v028.i05

Kumar, S., and T. Shaikh. 2017. Empirical evaluation of the performance of feature selection approaches on Random Forest. Pages 227–231 in 2017 International Conference on Computer and Applications.

Kursa, M. B., and W. R. Rudnicki. 2010. Feature selection with the Boruta package. J. Stat. Softw. 36:1–13. https://doi.org/10.18637/jss.v036.i11

Lam, K. P., and S. Horvath. 2008. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9:559. https://doi.org/10.1186/1471-2105-9-559

Lehky, T., J. P. Rickard, T. Pini, B. M. Gadella, and S. P. Graaf. 2020. Quantitative proteomic analysis of seminal plasma, sperm membrane proteins, and seminal extracellular vesicles suggests vesicular mechanisms aid in the removal and addition of proteins to the ram sperm membrane. Proteomics 20:1900289. https://doi.org/10.1002/pmic.201900289

Leek, J., J. W. Johnson, H. Parkr, E. Fertig, A. Jaffe, Y. Zhang, J. D. Storey, and L. Torres. 2021. svase: Surrogate variable analysis. R package version 3.4.2.

Li, K., L. Li, Y. Xu, and J. Yang. 2022. Erratum to: Machine learning meets omics: applications and perspectives. Brief. Bioinform. 23:bjab560. https://doi.org/10.1093/bib/bjab560

Liu, L., K. Zhou, Y. Song, and X. Liu. 2021. CCDC40 mutation as a cause of infertility in a Chinese family with primary ciliary dyskinesia. Medicine (Baltimore) 100:e28275. https://doi.org/10.1097/MD.00000000000028275

Meneses, E. B., A. L. C. Velho, F. Santos, T. Dinh, A. Kaya, E. Topkara, A. A. Moura, and E. Memili. 2019. Uncovering sperm membrane associated with fertility. BMC Syst. Biol. 2:19. https://doi.org/10.1186/s12864-019-0559-2

Peddinti, D. B., Nanduri, A. Kaya, J. M. Feugang, S. C. Burgess, and E. Ang. 2008. An intensive proteomic analysis of bovine spermatozoa of varying fertility rates and identification of biomarkers associated with fertility. BMC Syst. Biol. 2:19. https://doi.org/10.1186/1752-0509-2-19.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rabaglino, M., and P. Lonergan. 2022a. Table S1. Details about the bulls used in Project 16/IA/4174 (Science Foundation Ireland). figshare. Dataset. https://figshare.com/articles/dataset/20607708

Rabaglino, M., and P. Lonergan. 2022b. Table S2. List of differentially abundant proteins in the sperm of high fertility bulls compared to low fertility bulls. figshare. Dataset. https://figshare.com/articles/dataset/20367708

Sarf, K. K., A. Kumarasen, M. Dasgupta, G. Kartihkeyan, T. S. K. Prasad, P. K. Modi, K. Ramesha, S. Jayakumar, and A. Manimaran. 2020. Metabolomic fingerprinting of bull spermatozoa for identification of fertility signature metabolites. Mol. Reprod. Dev. 87:692–703. https://doi.org/10.1002/mrd.23354

Sellem, E. M. L. Broekhuysse, L. Chevrier, S. Camugli, E. Schmitt, L. Schibler, and E. P. Koenen. 2015. Use of combinations of in vitro quality assessments to predict fertility of bovine semen. Theriogenology 84:1447–1454.e5. https://doi.org/10.1016/j.theriogenology.2015.07.035

Shallow, L., A. Cromie, and N. McHugh. 2014. Effect of fertility on the economics of pasture-based dairy systems. Animal 8(Suppl 1):222–231. https://doi.org/10.1111/anj.12033

Smyth, G. K. 2005. Limma: Linear models for microarray data. Pages 507–520 in Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. V. C. R. Gentleman, S. Dudoit, R. Irizarry, W. Huber, ed. Springer.

Soggiu, A., C. Piras, H. A. Hussein, M. De Canio, A. Gaviraghi, A. Galli, A. Urbani, L. Bonizzi, and P. Roncada. 2013. Unravelling the bull fertility proteome. Mol. Biosyst. 9:1188–1195. https://doi.org/10.1039/c3mb2549a

Somasekhar, L., S. Selvaraju, S. Partiphan, S. K. Patil, B. K. Binsla, M. M. Venkataswamy, S. Karthik Bhat, and J. P. Ravindra. 2017. Comparative sperm protein profiling in bulls differing in fertility and identification of phosphatidylethanolamine-binding protein 4, a potential fertility marker. Andrology 5:1032–1051. https://doi.org/10.1111/andr.12404

Soues, S., M. Kann, J. Fouquet, and R. Melki. 2003. The cytosolic chaperonin CCT associates to cytoplasmic microtubular structures during mammalian spermiogenesis and to heterochromatin in germine and somatic cells. Exp. Cell Res. 288:363–373. https://doi.org/10.1006/sicy.2002.4829.

Sternlicht, H., G. Furr, M. Sternlicht, J. Driscoll, K. Willison, and M. Yaffe. 1993. The 3C-like polypeptide 1 complex is a chaperonin for tubulin and actin in vivo. Proc. Natl. Acad. Sci. USA 90:9422–9426. https://doi.org/10.1073/pnas.90.20.9422

Tang, S., X. Wang, W. Li, X. Yang, Z. Li, W. Liu, C. Li, Z. Zhu, L. Wang, J. Wang, L. Zhang, X. Sun, E. Zhi, H. Wang, H. Li, J. Lin, Y. Luo, J. Wang, Yang, and F. Zhang. 2017. Biallelic mutations in CFAP43 and CFAP44 cause male infertility with multiple morphological abnormalities of the sperm flagella. Am. J. Hum. Genet. 100:854–864. https://doi.org/10.1016/j.ajhg.2017.04.012

Thundathil, J., J. Gil, A. Januskauskas, B. Larsson, L. Soderquist, R. Mapleton, and H. Rodriguez-Martinez. 1999. Relationship between the proportion of capacitated spermatozoa present in frozen-thawed bull semen and fertility with artificial insemination. Int. J. Androl. 22:366–373. https://doi.org/10.1046/j.1365-2605.1999.00194.x

Viana, A. G. A., A. M. A. Martins, A. H. Pontes, W. Fontes, M. S. Castro, C. A. O. Ricart, M. V. Sousa, A. Kaya, E. Topper, E. Memili, and A. A. Moura. 2018. Proteomic landscape of seminal plasma associated with dairy bull fertility. Sci. Rep. 8:16.1632. https://doi.org/10.1038/s41598-018-34152-w

Walsh, S. W., E. J. Williams, and A. C. Evans. 2011. A review of the causes of poor fertility in high milk producing dairy cows. Anim. Reprod. Sci. 123:127–138. https://doi.org/10.1016/j.anireprosci.2010.12.001

Wang, Y. Y., C. C. Ke, Y. L. Chen, Y. H. Lin, J. S. Yu, W. C. Ku, M. K. O’Bryan, and Y. H. Lin. 2020. Deficiency of the Tbc1d21 gene causes male infertility with morphological abnormalities of the sperm mitochondria and flagellum in mice. PLOS Genet. 16:e1009020. https://doi.org/10.1371/journal.pgen.1009020.
Wei, X., Y. Sha, Z. Wei, X. Zhu, F. He, X. Zhang, W. Liu, Y. Wang, and Z. Lu. 2021. Bi-allelic mutations in DNAH7 cause asthenozoospermia by impairing the integrality of axoneme structure. Acta Biochim. Biophys. Sin. (Shanghai) 53:1300–1309. https://doi.org/10.1093/abbs/gmab113.

Wickham, H. 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.