Biology of Reproduction, 2024, **111(1)**, 54–62
https://doi.org/10.1093/biolre/ioae052
Advance access publication date 9 April 2024
Research Article

Machine learning to identify endometrial biomarkers predictive of pregnancy success following artificial insemination in dairy cows†

Quinn A. Hoorn[1], Maria B. Rabaglino[2], Thiago F. Amaral[1,3], Tatiane S. Maia[1], Fahong Yu[4], John B. Cole[1,5,6] and Peter J. Hansen[1,*]

[1]Department of Animal Sciences, Donald Henry Barron Reproductive and Perinatal Biology Research Program, and the Genetics Institute, University of Florida, Gainesville, FL, USA
[2]Department of Population Health Science, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands
[3]Genus plc PLC/ABS, Mogi Mirim, São Paulo, Brazil
[4]University of Florida Interdisciplinary Center for Biotechnology Research, Gainesville, FL, USA
[5]URUS Group LP, Madison, WI, USA
[6]Department of Animal Science, North Carolina State University, Raleigh, NC, USA
***Correspondence:** Department of Animal Sciences, Donald Henry Barron Reproductive and Perinatal Biology Research Program, and the Genetics Institute, University of Florida, 2250 SW Shealy Drive, PO Box 110910, Gainesville, FL 32611-0910, USA. E-mail: pjhansen@ufl.edu

†**Grant Support:** Research was supported by a grant from URUS LLP and the L.E. "Red" Larson Endowment. QH was supported by a National Needs Fellowship from USDA NIFA Grant 2021-38420-34067.

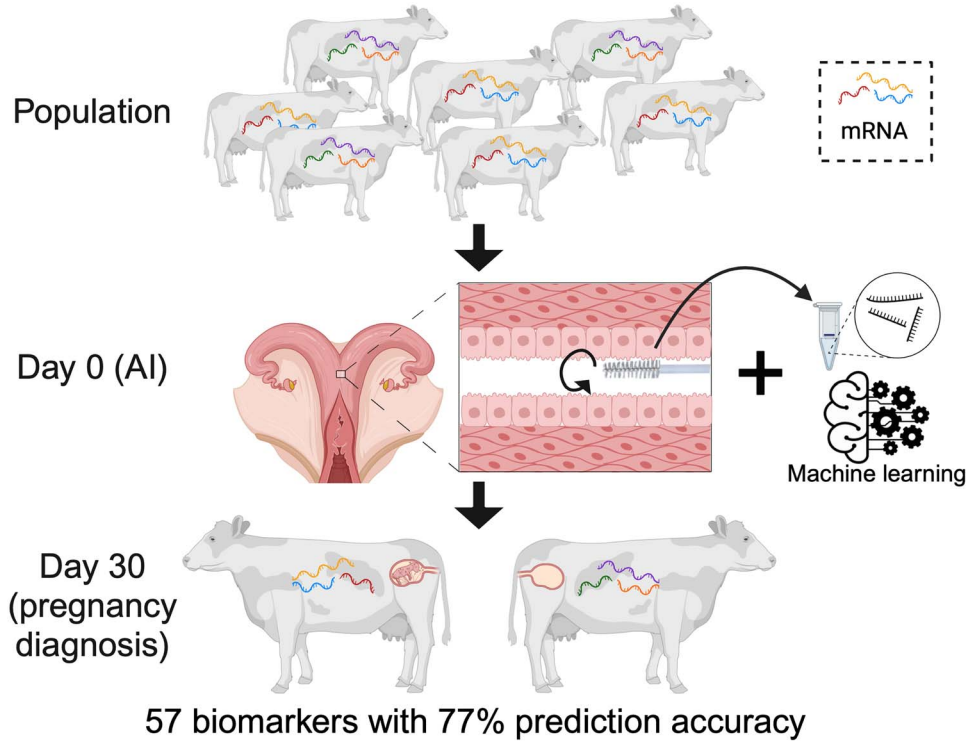
Abstract

The objective was to identify a set of genes whose transcript abundance is predictive of a cow's ability to become pregnant following artificial insemination. Endometrial epithelial cells from the uterine body were collected for RNA sequencing using the cytobrush method from 193 first-service Holstein cows at estrus prior to artificial insemination (day 0). A group of 253 first-service cows not used for cytobrush collection were controls. There was no effect of cytobrush collection on pregnancy outcomes at day 30 or 70 or on pregnancy loss between days 30 and 70. There were 2 upregulated and 214 downregulated genes (false discovery rate < 0.05, absolute fold change >2-fold) for cows pregnant at day 30 versus those that were not pregnant. Functional terms overrepresented in the downregulated genes included those related to immune and inflammatory responses. Machine learning for fertility biomarkers with the R package BORUTA resulted in identification of 57 biomarkers that predicted pregnancy outcome at day 30 with an average accuracy of 77%. Thus, machine learning can identify predictive biomarkers of pregnancy in endometrium with high accuracy. Moreover, sampling of endometrial epithelium using the cytobrush can help understand functional characteristics of the endometrium at artificial insemination without compromising cow fertility. Functional characteristics of the genes comprising the set of biomarkers is indicative that a major determinant of cow fertility, at least for first insemination after calving, is immune status of the uterus, which, in turn, is likely to reflect the previous history of uterine disease.

Summary Sentence
Machine learning of endometrial transcriptome collected via cytobrush identified 57 biomarkers predictive of day 30 pregnancy success with an average accuracy of 77%.

Received: December 3, 2023. **Revised:** February 6, 2024. **Accepted:** March 25, 2024

Graphical Abstract



Key words: biomarkers, cytobrush, endometrium, fertility

Introduction

Sustainability and profitability of the dairy industry depends upon milk production and fertility. In fact, greenhouse gas emissions per unit of production decrease as milk yield or fertility increases [1–3]. Achieving high fertility in a dairy herd is also important from an economic perspective [4, 5]. Reproductive failure is the single most important cause of culling [6], with 32% of lactating cows culled per year in the United States [7]. Development of tools to distinguish phenotypically or genetically fertile females from infertile ones could be used to guide management decisions regarding raising, breeding, and culling. This information could also expand the range of reproductive phenotypes used for genetic selection. The most commonly collected reproductive phenotypes used for genetic selection are daughter pregnancy rate, days open, and conception rate. All of these traits are limited in utility due to low heritabilities, difficulties in measurement, and large impacts of management on phenotype [8]. Development of new reproductive phenotypes that are easy to collect and which possess high heritabilities would be advantageous.

Machine learning is a subfield of artificial intelligence that involves learning from training models to develop predictions of unobserved data [9]. Implementation of this technology in animal reproduction is emerging. For example, machine learning was used to identify 50 genes whose transcript abundance in the endometrium predicted ability of cows to achieve pregnancy after embryo transfer and artificial insemination with 96.1% accuracy [10]. In another study, eight genes were identified whose transcript abundance in the bovine embryo could discriminate between embryos with high and low competence for pregnancy with an accuracy of 85% or greater [11]. In the present study, endometrial samples from the uterine body collected at the time of artificial insemination via cytobrush were used to identify transcriptomic markers of subsequent pregnancy using machine learning. The result was a set of genes whose transcript abundance is predictive of a cow's ability to become pregnant after insemination.

Methods

Ethics

All animal procedures were approved by the University of Florida Institutional Animal Care and Use Committee (IACUC202200000013, approved March 2, 2022).

Cytobrush preparation

Cytobrushes were prepared as previously described [12]. Briefly, trimmed endocervical sampling brushes (Viamed, Miami Lakes, FL, USA) were individually packaged and sterilized by gas autoclave. Prior to collection, the cytobrush was removed from the sealed package and inserted into the tip of a conventional 0.5 cc artificial insemination (AI) pipette. An AI sheath (Agtech Inc., Manhattan, KS, USA) and chemise (Agtech) were placed over the cytobrush for protection.

Cows

The study was performed at North Florida Holsteins in Bell, Florida (29°43′N 82°51′W) from March through May of

2022. A total of 168 first-service, lactating multiparous Holstein cows and 25 first-service, lactating primiparous Holstein cows (193 total) were used for collection of endometrial epithelium from the uterine body using the cytobrush technique. Another 192 first-service multiparous cows and 61 first-service primiparous cows (253 total) were used as controls in which no cytobrush sample was collected. The control group was examined to evaluate whether the cytobrush collection procedure compromised fertility. Lactating, first-service dairy cows were chosen for the study as compared to some other types of cows because fertility is often compromised in this type of animal.

All cows were subjected to a modified DoubleOvsynch ovulation synchronization procedure [13, 14], starting at ~43 days after calving (day −27 relative to insemination) with two injections of prostaglandin F2α on days −3 and −2 relative to insemination. All injections were administered intramuscularly. Insemination was on the morning of day 0, immediately after collection of endometrial epithelium. Control cows were inseminated but not subjected to cytobrush collection. All cytobrush samples were collected by one technician. Artificial insemination was performed by multiple farm personnel according to the farm's standard operating procedures.

Pregnancy outcomes after AI were determined by transrectal ultrasonographic examination of the reproductive tract at day 30 and day 70 after AI by farm veterinarians.

Collection and processing of endometrial epithelium

On day 0, the day of AI, each cow was restrained in a palpation rail after the first milking in the morning. The pipette containing the cytobrush assembly was inserted into the vagina and guided transrectally through the cervix into the uterine body. Once in the uterine body, the sheath covering the cytobrush was retracted slightly to expose the cytobrush. The cytobrush was then rotated five times, the sheath was extended over the cytobrush, and the assembly was removed from the cow. The cytobrush was then removed from the AI pipette and placed in a 2-mL microcentrifuge tube containing 1 mL of diethyl pyrocarbonate–treated Dulbecco phosphate buffered saline and stored on ice until processing (within 2 h of collection). Immediately following cytobrush collection, cows were artificially inseminated with a single straw of frozen–thawed semen by farm technicians.

Cells were collected from cytobrushes as follows. Microcentrifuge tubes containing the cytobrush were vortexed for 1 min. The brush was removed from the tube while scraping any remaining cells on the side of the tube. The brush was then discarded. Samples were centrifuged for 400×g for 7 min. Following centrifugation, the pellet was removed and placed in a cryotube (Electron Microscopy Sciences, Hatfield, PA, USA). Samples were snap frozen in liquid nitrogen and subsequently stored at −80°C.

Statistical analysis of effect of cytobrush on pregnancy rate

Pregnancy rates comparing animals subjected to the cytobrush collection procedure to control animals were analyzed by the GLIMMIX procedure of SAS software v9.4 (Cary, NC, USA) with pregnancy as a binomial variable and with group (i.e., cytobrush vs control) and parity (i.e., 1 vs others) in the model.

Sample homogenization and RNA extraction

Endometrial epithelial cell pellets from the 193 cows collected at day 0 (day of AI) were thawed and transferred to individual 2-mL tubes containing a mixture of 1.4- and 2.8-mm zirconium oxide beads from the Precellys CKMix Tissue Homogenizing Kit (Bertin Corp., Rockville, MD, USA). The samples were supplemented with 350 μL of RNA lysis buffer (RLT) buffer from the Qiagen RNeasy Micro Kit (Qiagen, Germantown, MD, USA) containing 1% (v/v) β-mercaptoethanol (Sigma-Aldrich, St. Louis, MO, USA). Samples were homogenized in the Precellys 24 Tissue Homogenizer (Bertin Corp.) for two 10-s cycles at 6200 rpm. The supernatants were transferred to sterile 2-mL microcentrifuge tubes and centrifuged for 15 min at 21 380×g and 4°C. Supernatants were removed once again and transferred to sterile 2-mL microcentrifuge tubes, and RNA extraction proceeded according to the Qiagen kit protocol. Once extracted, RNA concentration was determined using Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and RNA integrity number (RIN) was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA, USA). The average RIN across the samples was 8.01 with a standard deviation of 0.76.

Library preparation and sequencing

Library preparation and sequencing were performed at the University of Florida Interdisciplinary Center for Biotechnology Research Gene Expression Core and NextGen Sequencing Cores, respectively, using procedures described by Haveman et al. [15]. Briefly, the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Ipswich, MA, USA) was utilized to isolate mRNA from 100 ng of total RNA. The NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs) and SPT Labtech mosquito LV liquid handling instrument (SPT Labtech, Melbourn, UK) were used for RNA library construction with 1/2 of poly A-enriched RNA and 1/5 reaction volume. Double-stranded complementary DNA was synthesized, adaptors were ligated to the samples, and each uniquely barcoded library was enriched by 12 cycles of amplification. A total of 193 barcoded libraries were sized on the Agilent Bioanalyzer and quantified with the Qubit 2.0 Fluorometer. These individual libraries were then pooled in equimolar concentration for sequencing, with a target of 50 million reads per sample with a total of 4 lanes. Illumina NovaSeq 6000-S4 was used to sequence the libraries for 2 × 150 cycles. One lane generated 2.5–3 billion paired-end reads with an average Q30% ≥ 92.5% and Clusters Passing Filter (PF) = 85.4%. The BCL2fastQ function in the Illumina BaseSpace portal was used to generate FastQ files.

Data processing

Reads were trimmed with the cutadapt (v3.4) program [16] to remove adaptors and low-quality bases with a phred-like score <20 and reads <65 bases. The cleaned reads were individually mapped to the *Bos taurus* (ARS-UCD1.3) reference genome from the NCBI database using the read mapper of the STAR package (Spliced Transcripts Alignment to a Reference, v2.7.9a) [17]. Mapped reads underwent further processing with HTSeq (High-Throughput Sequence Analysis in Python, v2.0.3) [18], samtools, and scripts developed in-house at University of Florida Interdisciplinary Center for Biotechnology Research to remove potential PCR duplicates

and to choose and count uniquely mapped reads for gene expression analysis. Read counts were normalized by library size and used for subsequent analyses.

Differentially expressed genes between pregnant and non-pregnant animals

Differentially expressed genes (DEG) between pregnant and non-pregnant animals at day 30 and day 70 after insemination, as well as animals that experienced pregnancy loss between day 30 and day 70, were determined using the edgeR package (v3.42.4) [19] in R software (v.4.2.3). Genes with less than one count per million were filtered out before normalization [20]. Filtered data were normalized through weighted trimmed mean of *M*-values [21]. Next, observation weights were used for robust estimate of the negative binomial dispersion parameter for each gene and for estimating regression parameters. Finally, a negative binomial generalized log-linear model was fit to read counts for each transcript and conduct genewise likelihood ratio tests for the coefficient contrast [22]. The matrix of contrast was built based on the comparisons between the groups. *P* values were adjusted with the Benjamini–Hochberg procedure, and the DEG were defined as those with a false discovery rate (FDR) <0.05 and absolute fold change (FC) >2.

Functional analysis of the downregulated genes from the edgeR analysis of the day 30 pregnant versus non-pregnant comparison was assessed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) [23, 24]. R Software was used to generate the bubble plots from DAVID analysis using the enriched Gene Ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms. In addition, R software was used to generate the volcano plots of the DEG.

Identification of biomarkers

Biomarker genes were identified using the R package BORUTA (v.8.0.0) [25]. This method was chosen based on the outputs of the software BioDiscML [26], which automatizes the steps for model selection. For biomarker identification, the number of transcripts (features) inputted to the BORUTA algorithm was filtered by using those differentially expressed transcripts between pregnant and non-pregnant cows at day 30 with an FDR <0.1. The BORUTA algorithm was then run 10 times up to when the algorithm did not identify new biomarkers. The sparse partial least-squares discriminant analysis, or sPLS-DA method [27], was employed to select the most discriminative combination of biomarkers transcripts from a given number of variables, i.e., the molecular signature providing the most accurate prediction of pregnant and non-pregnant cows at day 30. The evaluation of each combination was done by running a support vector machine algorithm with linear kernels with the caret package (v.6.0-94) [28]. The 193 samples were randomly divided into three parts consisting of 64, 64, and 65 samples. The two parts consisting of 64 samples were used to train the model and the remaining part of 65 samples was used for testing the model. This procedure was repeated 100 times for each combination.

Results

Pregnancy rate

Of the 193 cows that underwent cytobrush collection, 42.0% (81/193) were pregnant at day 30 after insemination. A total of 15 of 81 pregnant cows at day 30 experienced pregnancy loss by day 70 (18.5%). Thus, 66 of 193 cows (34.2%) were pregnant at day 70. Of the 253 control cows, 45.1% (114/253) were pregnant at day 30 after insemination. A total of 21 of 114 pregnant cows at day 30 experienced pregnancy loss by day 70 (18.4%). Therefore, at day 70, 93 of 253 control cows (36.8%) were pregnant. There was no statistical difference between cows subjected to the cytobrush collection and control cows not subjected to collection with regard to pregnancy rate at day 30 ($P = 0.6354$), pregnancy rate at day 70 ($P = 0.6048$), and pregnancy loss ($P = 0.9075$). There were also no significant effects of parity (results not shown).

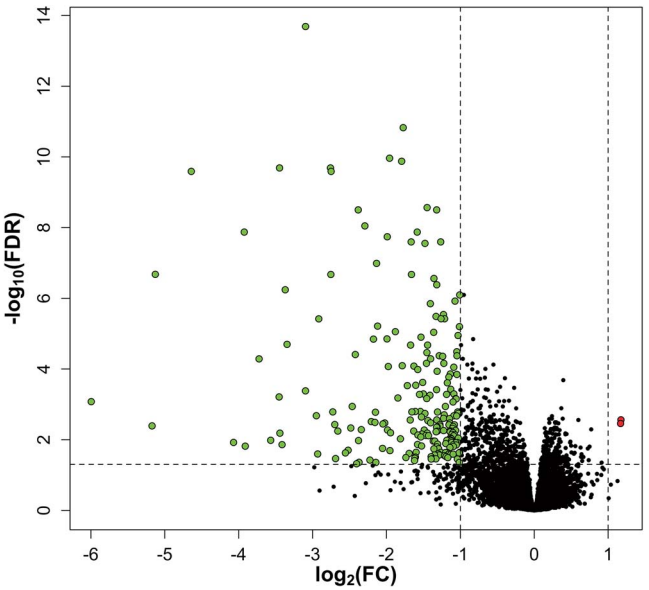


Figure 1. Volcano plot displaying upregulated (shown in red) and downregulated (shown in green) genes (FDR < 0.05 and absolute fold-change >2-fold) when comparing pregnant animals at day 30 after artificial insemination to non-pregnant animals. Non-significant genes are shown in black.

Differentially expressed genes between pregnant and non-pregnant cows

A summary of expression of all genes are presented in Supplementary File S1. Using an FDR of <0.05 and absolute FC >2 as a cutoff, there were 216 DEG (2 upregulated genes and 214 downregulated genes) comparing cows pregnant versus nonpregnant at day 30 after AI as shown in Figure 1. For pregnancy at day 70 after AI, there were 148 DEG (2 upregulated genes and 146 downregulated genes). A total of 137 DEG (1 upregulated gene and 136 downregulated genes) for pregnancy at day 70 were also DEG for pregnancy at day 30. A comparison of gene expression between cows that were pregnant at both days 30 and 70 versus those that were pregnant at day 30 but not at day 70 revealed only three DEG, with all being downregulated in cows that were pregnant at both days (FDR < 0.05 and absolute FC > 2). Two of these genes, actin gamma 2 (*ACTG2*) and desmin (*DES*), were among the sets of DEG at day 30 and day 70.

Functional characteristics of downregulated DEG (FDR < 0.05 and absolute FC > 2) for pregnancy at day 30 were analyzed by DAVID. Inflammatory and immune response, chemokine-mediated signaling pathways, and neutrophil chemotaxis were overrepresented in the Gene Ontology

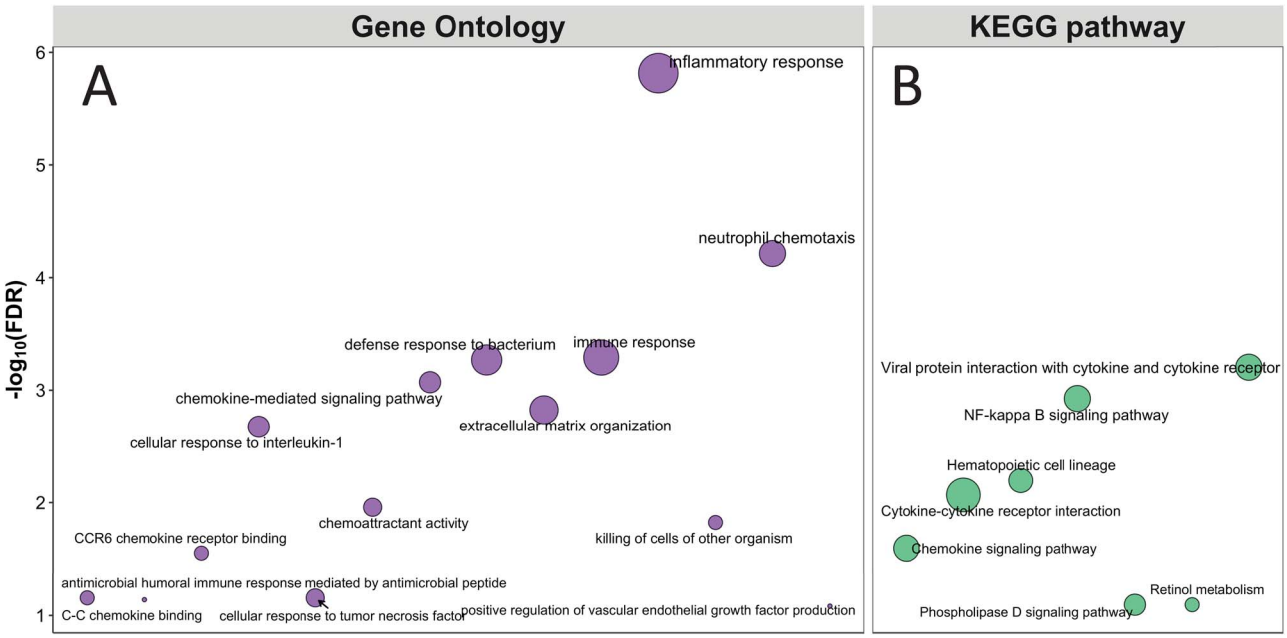


Figure 2. Bubble plots showing functional annotations overrepresented in the genes downregulated in pregnant cows at day 30 (FDR <0.05 and absolute fold-change >2-fold) using Gene Ontology (A) and KEGG pathway (B). The size of the bubble is proportional to the number of genes enriched in the term.

results (Figure 2A). Hematopoietic cell lineage, NF-kappa B signaling, chemokine signaling, and retinol metabolism were overrepresented in the KEGG pathway results (Figure 2B).

Fertility biomarkers

Each run of the BORUTA algorithm generated between 33 and 40 biomarkers. The specific biomarkers were not the same in each run and there was a total of 87 potential unique biomarkers identified. The 87 transcripts were analyzed by the sPLS-DA method to identify the most discriminative combination of transcripts when the number of biomarkers were set at from 10 to 87. In total, 77 combinations of biomarkers were analyzed. Using this approach, the most discriminative set of biomarkers were a set of 57 transcripts whose identity is shown in Table 1. A total of 28 biomarker genes were upregulated in cows that became pregnant at day 30 while 29 biomarker genes were downregulated. Distribution of samples based on the whole transcriptome can be seen in the principal component analysis (PCA) plot in Figure 3A, and then according to the expression of the 57 biomarkers in Figure 3B. The PCA plot is a type of scatterplot where samples are clustered based on their similarity. The separation of samples belonging to cows that resulted pregnant or not was clearly improved according to the expression of the biomarker genes (Figure 3B) compared to the expression of the whole transcriptome (Figure 3A). The average accuracy for prediction of pregnancy at day 30 using 100 rounds of prediction was 77% (range: 65% to 88%; average sensitivity and specificity was 82% and 69%, respectively). All the parameters describing the biomarkers' performance are shown in Table 2.

Discussion

Advances in 'omics technologies, where data on thousands of phenotypes can be generated in a single assay, and machine learning techniques, in which computers are used to develop predictive algorithms, have created the possibility of development of new biomarkers for economically important traits in livestock production. One trait for which biomarkers would have utility in the dairy industry is female fertility. Identification of females that are fertile or infertile would allow new approaches for improving reproductive function and profitability of dairy operations. Females that are infertile could be culled or assigned to reproductive management programs tailored for less-fertile animals. Moreover, discovery of genetic variants associated with expression of the biomarkers (so-called expression QTL, [29]) could be used to increase accuracy of genetic estimates of reproductive traits. Here, we have shown the use of RNA-sequencing in combination with machine learning to identify 57 biomarkers of fertility in lactating dairy cows with an average sensitivity and specificity of prediction of 82% and 69%, respectively. Functional characteristics of the genes comprising the set of biomarkers are indicative that a major determinant of cow fertility, at least for first insemination after calving, is immune status of the uterus, which, in turn, is likely to reflect the previous history of uterine disease.

Remarkably, these markers were identified in samples of endometrial epithelium collected at a time and position removed from the presence of the embryo. In particular, the endometrial epithelium was harvested from the body of the uterus at day 0, the day of timed artificial insemination, which is 4 to 5 days before the embryo enters the uterus [30]. Once entering the uterus, the early embryo resides in the uterine horn and not the uterine body [31]. Endometrium was collected from the uterine body because it was deemed less likely to result in damage to the endometrium than if endometrial cells were collected in the uterine horn. Indeed, there was no difference in pregnancy rate between cows subjected to cytobrush sampling and those that were not. It was also reasoned that gene expression in the uterine body

Table 1. The list of biomarkers and corresponding absolute fold change (FC) between pregnant and non-pregnant cows at day 30 after insemination[a]

Gene ID	Gene definition	FC	FDR
NM_206970.1	triggering receptor expressed on myeloid cells 1 (TREM1)	−15.1866	1.34E-08
XM_025001090.1	PREDICTED: tyrosine-protein phosphatase non-receptor type substrate 1-like (LOC104973826), transcript variant X2	−5.2136	3.16E-09
XM_024975574.1	PREDICTED: aldehyde oxidase 2 (AOX2), transcript variant X2	−2.7869	1.82E-03
LOC101905517	—	−2.5686	4.80E-04
XR_804895.3	PREDICTED: uncharacterized LOC104971235 (LOC104971235), ncRNA	−2.1806	3.70E-03
XM_024995833.1	PREDICTED: extracellular matrix protein 2 (ECM2), transcript variant X1	−2.1266	2.11E-03
NM_001076164.2	protein phosphatase 1 regulatory subunit 3C (PPP1R3C)	−2.0674	3.26E-05
XM_024999645.1	PREDICTED: uncharacterized LOC112448856 (LOC112448856), transcript variant X1	−2.0369	1.79E-03
XM_003585891.5	PREDICTED: interferon induced protein 44 like (IFI44L)	−2.0120	8.03E-07
XM_024981599.1	PREDICTED: Rh family C glycoprotein (RHCG), transcript variant X1	−1.9515	1.02E-02
XR_817181.3	PREDICTED: uncharacterized LOC104970645 (LOC104970645), transcript variant X2, ncRNA	−1.8639	7.40E-03
NM_001194964.1	C-C motif chemokine receptor 10 (CCR10)	−1.7853	7.03E-05
NM_001098140.1	Src homology 2 domain containing E (SHE)	−1.7666	2.75E-02
XM_024993411.1	PREDICTED: POTE ankyrin domain family member G-like (LOC618219)	−1.7087	7.79E-02
XM_003586816.5	PREDICTED: KIAA1755 (KIAA1755)	−1.6621	6.18E-03
XR_003036567.1	PREDICTED: U6 spliceosomal RNA (LOC112448180), ncRNA	−1.6554	2.70E-02
NM_001142485.2	SPEG complex locus (SPEG)	−1.5996	6.34E-04
XM_002697051.6	PREDICTED: collagen type VII alpha 1 chain (COL7A1)	−1.5893	1.76E-04
NM_174209.2	uveal autoantigen with coiled-coil domains and ankyrin repeats (UACA)	−1.5719	5.11E-02
XM_015461355.2	PREDICTED: tetraspanin 32 (TSPAN32), transcript variant X8	−1.4820	3.91E-02
NM_174730.2	PYD and CARD domain containing (PYCARD)	−1.4669	7.53E-05
NM_001034668.3	MHC class II antigen (BLA-DQB)	−1.4177	4.40E-04
XM_010804458.2	PREDICTED: kelch domain containing 10 (KLHDC10), transcript variant X3	−1.3995	2.52E-02
NM_181014.2	monoamine oxidase A (MAOA)	−1.3769	9.46E-03
NM_174676.3	RAS p21 protein activator 3 (RASA3)	−1.3168	1.82E-04
XR_003038123.1	PREDICTED: uncharacterized LOC112449547 (LOC112449547), ncRNA	−1.2637	1.30E-02
NM_001078092.1	WW domain containing oxidoreductase (WWOX)	−1.2163	1.51E-02
NM_001098912.1	fumarylacetoacetate hydrolase (FAH)	−1.1497	7.20E-03
NM_001206262.1	5′,3′-nucleotidase, mitochondrial (NT5M); nuclear gene for mitochondrial product	−1.1465	6.17E-02
NM_001103095.1	fat storage inducing transmembrane protein 2 (FITM2)	1.0984	1.51E-02
XM_015475466.2	PREDICTED: malectin (MLEC), transcript variant X1	1.1084	6.47E-03
NM_174135.3	prolyl 4-hydroxylase subunit beta (P4HB)	1.2339	4.63E-03
NM_001046091.2	protein disulfide isomerase family A member 5 (PDIA5)	1.2378	1.91E-02
NM_174333.3	protein disulfide isomerase family A member 3 (PDIA3)	1.2541	3.44E-03
NM_001081609.2	transmembrane p24 trafficking protein 3 (TMED3)	1.2610	7.04E-03
NM_001192963.1	solute carrier family 2 member 13 (SLC2A13)	1.2682	5.35E-03
NM_001099121.2	DnaJ heat shock protein family (Hsp40) member C10 (DNAJC10)	1.2891	1.29E-03
NM_174700.2	heat shock protein 90 beta family member 1 (HSP90B1)	1.3063	2.54E-03
NM_001038048.2	nucleoporin 50 (NUP50)	1.3132	2.07E-04
NM_001035394.1	LIM and SH3 protein 1 (LASP1)	1.3397	6.63E-03
XM_005223170.3	PREDICTED: RAB7A, member RAS oncogene family (RAB7A), transcript variant X1	1.3402	7.04E-03
NM_001192439.2	solute carrier family 2 member 10 (SLC2A10)	1.3427	1.71E-02
XM_024987334.1	PREDICTED: heterogeneous nuclear ribonucleoprotein U like 2 (HNRNPUL2), transcript variant X1	1.3531	2.77E-02
XM_024980497.1	PREDICTED: MAX network transcriptional repressor (MNT), transcript variant X1	1.3538	6.89E-03
NM_001099391.2	golgi associated RAB2 interactor family member 4 (GARIN4)	1.3953	7.96E-02
XM_005215930.3	PREDICTED: nucleobindin 2 (NUCB2), transcript variant X1	1.3999	1.20E-02
NM_001244612.1	insulin like growth factor 1 receptor (IGF1R)	1.4040	1.13E-02
XM_010818457.3	PREDICTED: histone H2B type 1 (LOC104975676)	1.4171	6.59E-02
NM_001076290.2	TATA-box binding protein like 1 (TBPL1)	1.4183	2.30E-02
XM_005214293.4	PREDICTED: ADP ribosylation factor like GTPase 5B (ARL5B), transcript variant X1	1.4703	1.52E-02
XM_024976678.1	PREDICTED: family with sequence similarity 177 member B (FAM177B), transcript variant X1	1.4726	7.97E-02
XM_025000381.1	PREDICTED: La ribonucleoprotein domain family member 4B (LARP4B), transcript variant X6	1.5128	2.77E-03
NM_001105035.1	actin like 6A (ACTL6A)	1.5446	1.67E-02
NM_001075467.2	homer scaffold protein 2 (HOMER2)	1.5584	2.58E-02
XM_024978019.1	PREDICTED: nucleobindin 1 (NUCB1), transcript variant X2	1.5860	3.08E-02
XM_005223936.4	PREDICTED: myelin basic protein (MBP), transcript variant X3	1.7064	5.10E-03
LOC788672	-	1.7102	1.10E-02

[a]Note that a gene defined as a biomarker might not be necessarily a differentially expressed gene according to the definition used (false discovery rate < 0.05 and fold change > |2|).

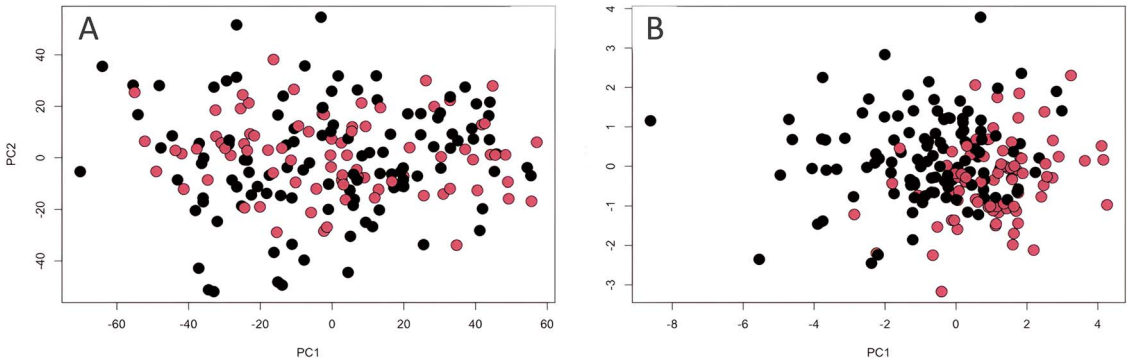


Figure 3. Principal component analysis (PCA) plots showing pregnant (red) and non-pregnant (black) animals at day 30 of pregnancy according to the expression of the whole transcriptome (A) and the biomarker genes (B), i.e., before (A) and after (B) separation by the identified biomarkers.

Table 2. Evaluation metrics corresponding to the classification of pregnancy status at day 30 based on the expression of the 57 biomarker genes

	Average	Range
Accuracy	0.77	0.65–0.88
Kappa	0.51	0.27–0.76
Sensitivity	0.82	0.56–0.98
Specificity	0.69	0.48–0.92
Positive predictive value	0.78	0.64–0.94
Negative predictive value	0.74	0.52–0.95
Precision	0.78	0.64–0.94
Recall	0.82	0.56–0.98
F1	0.80	0.67–0.89
Prevalence	0.57	0.46–0.69
Detection rate	0.47	0.34–0.58
Detection prevalence	0.60	0.42–0.76

is an accurate reflection of gene expression elsewhere in the uterus. A recent experiment indicates that expression of genes by the endometrium is mostly uniform throughout the length of the uterus early in the estrous cycle. When sampled at day 2 after ovulation, expression of 99.8% of genes examined did not differ between endometrium collected in the uterine body, middle uterine horn, or upper uterine horn [32].

The endometrium at day 0 is under hormonal control characterized by the predominant actions of estradiol whereas progesterone starts to be the major hormone affecting the endometrium when the embryo is present [33]. Perhaps other markers of fertility could be identified if endometrial tissue was sampled later in the estrous cycle. Day 0 was chosen for the current experiment because, in terms of practical applications, this is a stage of the estrous cycle when a female is most likely to be accessible to a technician collecting endometrial samples.

Most of the DEG associated with pregnancy success were downregulated in females that became pregnant after insemination. Functional ontology analysis was indicative that many of these genes were overrepresented in pathways associated with immune function and inflammation. It is proposed, therefore, that an important determinant of whether a cow became pregnant after insemination is the status of immune activation in the uterus, with cows with less active inflammation or immune reactivity being predisposed to fertility. Such a hypothesis is consistent with Maillo et al. [34], where genes associated with inflammation and the immune system were downregulated at day 3 of pregnancy in oviductal isthmus tissue of heifers in which multiple embryos were transferred endoscopically into the oviduct. There is also a well-established negative association between prior incidence of disease, of both uterine and extra-uterine origin, with cow fertility [35]. Prior occurrence of metritis early postpartum can have carryover actions that reduce pregnancy rate after embryo transfer later in the postpartum period [36]. Uterine infection is a particular problem of the postpartum cow and especially the postpartum dairy cow [37, 38]. It is, possible, therefore, that the markers discovered here will be less useful for bovine females where the incidence of uterine inflammatory disease is low.

In contrast to the identification of many DEG associated with pregnancy outcomes at days 30 and 70 of pregnancy, there were few DEG associated with pregnancy loss in pregnant females that occurred between days 30 and 70. Such a result suggests that biological causes of pregnancy loss during this period are largely independent of status of the uterus at the time of insemination. This inference is somewhat surprising because pregnancy loss has been reported to be higher in cows that previously experienced reproductive tract inflammation or disease in some [39], but not all [40], studies. Perhaps, the failure to find many DEG associated with pregnancy loss was because uterine inflammation is not as important a determinant of pregnancy loss as for establishment of pregnancy. Also, there were fewer observations of pregnancy loss than pregnancy rate at day 30, which could have contributed to the lower number of DEG for the latter trait.

An important, but unresolved, question is whether a female that is competent to establish pregnancy after insemination at a specific estrus is inherently fertile, or if fertility status is inconsistent from estrous cycle to estrous cycle. This question has relevance for the effectiveness of the markers for predicting future fertility as well as present fertility. Repeatability estimates of traits used to estimate dairy cow fertility are low. The repeatabilities of daughter pregnancy rate, heifer conception rate, and cow conception rate in US Holsteins are 0.13, 0.12, and 0.07, respectively [41, 42]. Silva et al. [43] estimated that the repeatability of fertility traits like days open, calving interval, and daughter pregnancy rate ranged from 0.06 to 0.12 in Brazilian Holstein cattle. On the other hand, studies in which populations of bovine females were subjected to repeated rounds of insemination or embryo transfer have identified populations of cows that consistently become pregnant [44–46]. Markers which identify these females will be

particularly valuable for use in reproductive management and genetic selection.

In summary, this study illustrated that machine learning, a developing artificial intelligence technology, is capable of discovering biomarkers useful for identifying dairy cows with high likelihood of establishing pregnancy after artificial insemination. As the name implies, machine learning requires constant learning to improve the accuracy of the system [9]. While a 77% accuracy in use of the biomarkers to predict pregnancy success is promising, collection and analysis of additional animals will improve the accuracy of prediction and extend results to wider populations of cows besides lactating dairy cows. An additional outcome of the study was the finding that cytobrush sampling of the uterine endometrium can be performed without compromising cow fertility, at least when done at day 0 relative to insemination. Thus, the technique can be useful for sampling endometrium in cows that have been inseminated without risking the pregnancy.

Acknowledgment

The authors thank everyone at North Florida Holsteins, including Don Bennink, John Karanja, and Adalid (Angel) Diaz Reyes, for their support and assistance with this project. The authors also thank Yanping Zhang at the University of Florida Interdisciplinary Center for Biotechnology Research for overseeing RNA sequencing, Cecilia Rocha and Mario Binelli at the University of Florida for assistance with methodology for the cytobrush, and Camila Cuéllar and Maura McGraw at the University of Florida for technical assistance.

Supplementary material

Supplementary material is available at *BIOLRE* online.

Author contributions

QH, MBR, JBC, and PJH contributed to the experimental design. QH, TA, and TM collected and processed samples. MBR, QH, and PJH performed the data analysis. QH wrote the initial draft of the manuscript, which was reviewed, revised, and approved by all authors.

Conflict of Interest: URUS Group LLC has an interest in identifying markers of fertility. Other authors declare no conflict of interest.

Data availability

Data have been deposited in the Gene Expression Omnibus of the National Center for Biotechnology Information [47] and are accessible through GEO Series Accession No. GSE248266 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE248266).

References

1. Garnsworthy PC. The environmental impact of fertility in dairy cows: a modelling approach to predict methane and ammonia emissions. *Anim Feed Sci Technol* 2004; **112**:211–223.
2. Knapp JR, Laur GL, Vadas PA, Weiss WP, Tricarico JM. Invited review: Enteric methane in dairy cattle production: quantifying the opportunities and impact of reducing emissions. *J Dairy Sci* 2014; **97**:3231–3261.
3. Diavão J, Silva AS, Sguizzato ALL, da Silva CS, Tomich TR, Pereira LGR. How does reproduction account for dairy farm sustainability? *Anim Reprod* 2023; **20**:e20230066.
4. De Vries A. Economic value of pregnancy in dairy cattle. *J Dairy Sci* 2006; **89**:3876–3885.
5. Li M, Reed KF, Lauber MR, Fricke PM, Cabrera VE. A stochastic animal life cycle simulation model for a whole dairy farm system model: assessing the value of combined heifer and lactating dairy cow reproductive management programs. *J Dairy Sci* 2023; **106**: 3246–3267.
6. Tsuruta S, Lourenco DA, Misztal I, Lawlor TJ. Genotype by environment interactions on culling rates and 305-day milk yield of Holstein cows in 3 US regions. *J Dairy Sci* 2015; **98**:5796–5805.
7. Pinedo PJ, De Vries A, Webb DW. Dynamics of culling risk with disposal codes reported by dairy herd improvement dairy herds. *J Dairy Sci* 2010; **93**:2250–2261.
8. Ma L, Cole JB, Da Y, VanRaden PM. Symposium review: genetics, genome-wide association study, and genetic improvement of dairy fertility traits. *J Dairy Sci* 2019; **102**:3735–3743.
9. Morota G, Ventura RV, Silva FF, Koyama M, Fernando SC. Big data analytics and precision animal agriculture symposium: machine learning and data mining advance predictive big data analysis in precision animal agriculture. *J Anim Sci* 2018; **96**: 1540–1550.
10. Rabaglino MB, Kadarmideen HN. Machine learning approach to integrated endometrial transcriptomic datasets reveals biomarkers predicting uterine receptivity in cattle at seven days after estrous. *Sci Rep* 2020; **10**:16981.
11. Rabaglino MB, Salilew-Wondim D, Zolini A, Tesfaye D, Hoelker M, Lonergan P, Hansen PJ. Machine-learning methods applied to integrated transcriptomic data from bovine blastocysts and elongating conceptuses to identify genes predictive of embryonic competence. *FASEB J* 2023; **37**:e22809.
12. Cardoso B, Oliveira ML, Pugliesi G, Batista E, Binelli M. Cytobrush: a tool for sequential evaluation of gene expression in bovine endometrium. *Reprod Domest Anim* 2017; **52**:1153–1157.
13. Ayres H, Ferreira RM, Cunha AP, Araújo RR, Wiltbank MC. Double-Ovsynch in high-producing dairy cows: effects on progesterone concentrations and ovulation to GnRH treatments. *Theriogenology* 2013; **79**:159–164.
14. Borchardt S, Pohl A, Carvalho PD, Fricke PM, Heuwieser W. Short communication: effect of adding a second prostaglandin $F_{2\alpha}$ injection during the Ovsynch protocol on luteal regression and fertility in lactating dairy cows: a meta-analysis. *J Dairy Sci* 2018; **101**:8566–8571.
15. Haveman NJ, Zhou M, Callaham J, Strickland HF, Houze D, Manning-Roach S, Newsham G, Paul AL, Ferl RJ. Utilizing the KSC fixation tube to conduct human-tended plant biology experiments on a suborbital spaceflight. *Life* 2022; **12**:1871.
16. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011; **17**:10–12.
17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**:15–21.
18. Anders S, Pyl PT, Huber W. HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics* 2015; **31**: 166–169.
19. Robinson M, McCarthy D, Smyth G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**:139–140.
20. Rau A, Gallopin M, Celeux G, Jaffrézic F. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 2013; **29**:2146–2152.
21. Robinson M, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010; **11**:R25.
22. McCarthy D, Chen Y, Smyth G. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 2012; **40**:4288–4297.
23. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**:44–57.

24. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* 2022; **50**:W216–W221.
25. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010; **36**:1–13.
26. Leclercq M, Vittrant B, Martin-Magniette ML, Scott Boyer MP, Perin O, Bergeron A, Fradet Y, Droit A. Large-scale automatic feature selection for biomarker discovery in high-dimensional OMICs data. *Front Genet* 2019; **10**:452.
27. Lê Cao K, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform* 2011; **12**:253.
28. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008; **28**:1–26.
29. Shi XM (ed.). *eQTL Analysis. Methods and Protocols*. New York, NY: Humana Press; 2020.
30. Hackett AJ, Durnford R, Mapletoft RJ, Marcus GJ. Location and status of embryos in the genital tract of superovulated cows 4 to 6 days after insemination. *Theriogenology* 1993; **40**:1147–1153.
31. Sponchiado M, Gomes NS, Fontes PK, Martins T, Del Collado M, Pastore AA, Pugliesi G, Nogueira MFG, Binelli M. Pre-hatching embryo-dependent and -independent programming of endometrial function in cattle. *PloS One* 2017; **12**:e0175954.
32. Hoorn QA, Rabaglino MB, Maia TS, Sagheer M, Fuego D, Jiang Z, Hansen PJ. Transcriptomic profiling of the bovine endosalpinx and endometrium to identify putative embryokines. *Physiol Genomics* 2023; **55**:557–564.
33. Pohler KG, Geary TW, Atkins JA, Perry GA, Jinks EM, Smith MF. Follicular determinants of pregnancy establishment and maintenance. *Cell Tissue Res* 2012; **349**:649–664.
34. Maillo V, Gaora PÓ, Forde N, Besenfelder U, Havlicek V, Burns GW, Spencer TE, Gutierrez-Adan A, Lonergan P, Rizos D. Oviduct-embryo interactions in cattle: two-way traffic or a one-way street? *Biol Reprod* 2015; **92**:144.
35. Ribeiro ES, Gomes G, Greco LF, Cerri RLA, Vieira-Neto A, Monteiro PLJ Jr, Lima FS, Bisinotto RS, Thatcher WW, Santos JEP. Carryover effect of postpartum inflammatory diseases on developmental biology and fertility in lactating dairy cows. *J Dairy Sci* 2016; **99**:2201–2220.
36. Estrada-Cortés E, Ortiz WG, Chebel RC, Jannaman EA, Moss JI, de Castro FC, Zolini AM, Staples CR, Hansen PJ. Embryo and cow factors affecting pregnancy per embryo transfer for multiple-service, lactating Holstein recipients. *Transl Anim Sci* 2019; **3**:60–65.
37. Hanzen C, Laurent Y, Ward WR. Comparison of reproductive performance in Belgian dairy and beef cattle. *Theriogenology* 1994; **41**:1099–1114.
38. Sheldon IM. The postpartum uterus. *Vet Clin North Am Food Anim Pract* 2004; **20**:569–591.
39. Moraes JGN, Silva PRB, Mendonça LGD, Okada CTC, Chebel RC. Risk factors for purulent vaginal discharge and its association with reproductive performance of lactating Jersey cows. *J Dairy Sci* 2021; **104**:12816–12829.
40. Diaz-Lundahl S, Garmo RT, Gillund P, Klem TB, Waldmann A, Krogenæs AK. Prevalence, risk factors, and effects on fertility of cytological endometritis at the time of insemination in Norwegian red cows. *J Dairy Sci* 2021; **104**:6961–6974.
41. VanRaden PM, Sanders AH, Tooker ME, Miller RH, Norman HD, Kuhn MT, Wiggans GR. Development of a national genetic evaluation for cow fertility. *J Dairy Sci* 2004; **87**:2285–2292.
42. Kuhn MT, Hutchison JL, Wiggans GR. Characterization of Holstein heifer fertility in the United States. *J Dairy Sci* 2006; **89**:4907–4920.
43. Silva HT, Lopes PS, Carvalheira J, Silva DA, Silva AA, Silva FF, Veroneze R, Thompson G, Costa CN. Autoregressive model for genetic evaluation of longitudinal reproductive traits in Brazilian Holstein cattle. *Reprod Domest Anim* 2021; **56**:391–399.
44. McMillan WH, Donnison MJ. Understanding maternal contributions to fertility in recipient cattle: development of herds with contrasting pregnancy rates. *Anim Reprod Sci* 1999; **57**:127–140.
45. Parr MH, Mullen MP, Crowe MA, Roche JF, Lonergan P, Evans AC, Diskin MG. Relationship between pregnancy per artificial insemination and early luteal concentrations of progesterone and establishment of repeatability estimates for these traits in Holstein-Friesian heifers. *J Dairy Sci* 2012; **95**:2390–2396.
46. Geary TW, Burns GW, Moraes JG, Moss JI, Denicol AC, Dobbs KB, Ortega MS, Hansen PJ, Wehrman ME, Neibergs H, O'Neil E, Behura S, *et al.* Identification of beef heifers with superior uterine capacity for pregnancy. *Biol Reprod* 2016; **95**:47.
47. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30**:207–210.