A Genetical Genomics Methodology to Identify Genetic Markers of a Bovine Fertility Phenotype Based on CYP19A1 Gene Expression

Nicolas Guillemin, Isabelle Dufort and Marc-André Sirard

Centre de Recherche en Biologie de la Reproduction, Faculté des Sciences de l'Agriculture et de l'Alimentation, Département des Sciences Animales, Pavillon INAF, Université Laval, Québec, Canada.

ABSTRACT: With the decrease in fertility in dairy cow, the interest towards means to control this variable through genetic selection is growing. One of the most important factors controlling follicular maturity and timely ovulation is the aromatase enzyme, which is encoded by the *CYP19A1* gene. The activity of this enzyme is potentially the limiting factor in postpartum fertility. In this study, we developed a methodology, based on genetical genomics, to model the aromatase expression profile from granulosa cell samples. The transcriptomic expression profiles obtained were used to identify 355 genes or isoforms potentially associated with the regulation of aromatase. From those genes, 23,388 single-nucleotide polymorphism (SNPs) in the genome of Holstein cows were identified. Results showed that some SNPs (on *KRT8*, *LHCGR*, *CREB*, *ANXA1*, and on *CYP19A1* itself) were relevant to aromatase expression and the model generated could predict 44% of the observed phenotype. This study demonstrated the value of genetical genomics to generate better biomarkers for the dairy industry.

KEYWORDS: genomics, genetic selection, cow, aromatase, fertility

CITATION: Guillemin et al. A Genetical Genomics Methodology to Identify Genetic Markers of a Bovine Fertility Phenotype Based on CYP19A1 Gene Expression. *Gene Expression to Genetical Genomics* 2015:8 9–18 doi:10.4137/GGG.S24691.

RECEIVED: February 5, 2015. **RESUBMITTED:** March 26, 2015. **ACCEPTED FOR PUBLICATION:** March 30, 2015.

ACADEMIC EDITOR: Shiva Singh, Editor in Chief

TYPE: Original Research

FUNDING: This work was financially supported by Natural Sciences and Engineering Research Council of Canada (CRDPJ 398500-2010). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: marc-andre.sirard@fsaa.ulaval.ca

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Over the last decades, intense selective pressure has been applied to cows for milk production. As a direct or indirect consequence, cows' fertility decreased gradually, creating a major problem for the dairy industry in terms of production costs.[1] Fertility is a complex phenotype, which depends on both genetic and environmental factors.[2] One important fertility factor is the aromatase enzyme, which is directly involved in estradiol production.

The aromatase is an enzyme that belongs to the cytochrome P450 family. It catalyzes the aromatization of testosterone to estrogen mostly in the ovary. The aromatase encoding gene, cytochrome P450, family 19, subfamily A, polypeptide 1 (*CYP19A1*), is expressed in the granulosa cells of recruited follicles with a diameter larger than 4 mm in cows.[3] Expression of the gene is controlled by paracrine and endocrine factors and also by DNA methylation.[4] Aromatase activity is key during follicular development as it controls part of follicular differentiation and the exchange between the ovary and the brain. The estradiol concentration is linked to expression of *CYP19A1*, which encodes the limiting enzyme of the steroidogenic cascade.[5] Estrogens then regulate follicle-stimulating hormone (FSH)–stimulated gene expression in granulosa cells, stimulate follicle growth, gap junction development, cell proliferation, upregulation of FSH and luteinizing hormone (LH) receptors, modulate progesterone secretion, and protect granulosa cells from reactive oxygen–induced apoptosis.[6] The level of aromatase expression is central in the postpartum dairy cow. Indeed, the high lactation content increases the metabolic clearance of estradiol and requires a higher expression of aromatase to maintain the programmed blood surge essential to ovulation.[7] In addition, a previous study using an aromatase inhibitor, letrozole, during growth of the ovulatory follicle delayed ovulation by 24 h,[8] partially mimicking the ovulation delays associated with lower oocyte quality.[9,10]

This study proposed to identify genetic markers in the bovine genome related to the *CYP19A1* gene expression, using a methodology based on genetical genomics. Genetical genomics is a new methodology to analyze and find genetic markers for a complex trait that was developed by Jansen and Nap[11] in *Arabidopsis thaliana*. A year later, the principle of genetical genomics was validated in yeast[12] and has since been extended to mice, maize, rats, poultry, and humans to successfully identify Expression quantitative trait loci (eQTL) for different traits.[13–15] The central principle of genetical genomics consists in dissecting a complex physiological trait of interest into a list of genes for which the expression levels are related to variations of the trait (the genomics part) and then to link the expression-level variations of each gene with genetic marker variations between genomes (the genetics part). In genomics, the global transcriptomic profile of each individual is analyzed and compared to a control profile. This analysis shows transcriptomic variations associated with the complex trait variations and therefore provides a list of genes the expression of which is influenced by the complex trait. In genetics, each individual is genotyped for markers. Then, genetic variations are linked to transcriptomic variations previously identified with association studies. The results reveal the genetic markers that are related to the complex trait of interest and also generate complex biological pathways related to this trait. The results therefore provide explanations on gene architectures and mechanisms, which are typically poorly understood when only genetic studies are performed. Indeed, genetical genomics provides local and distant eQTL for a complex trait, at a scale never observed before, by a precise and high-throughput study with a genetic and a genomic component.[16] With the improvements of new technologies, genetical genomics is becoming a fast and reliable methodology to identify accurate markers for a complex physiological trait. This methodology uses fewer samples than traditional QTL analysis and is therefore suitable for a limited but informative experimental design (with fewer than 100 samples). Moreover, genetical genomics works at the transcriptomic level, which is closer to the final complex trait phenotype than genetic studies. So the information brought by genetical genomics is very useful to explore and understand the complex trait of interest, linking genetics and genomics information. Therefore, the current study was designed to identify genetic markers related to the expression profile of the *CYP19A1* gene, to analyze the accuracy and pertinence of this methodology for fertility assessment, and to integrate those identified markers in selection schemes for dairy cows.

Materials and Methods

Biological samples. Figure 1 represents the overall experimental design of the experiments. Granulosa cell samples (*N* = 83) were obtained from the ovaries of slaughtered Holstein cows. Only follicles with a diameter greater than 8 mm were used. Granulosa cells were washed twice with phosphate-buffered saline (PBS) + EDTA 10 mM and centrifuged at 12,000 g for 2 min at room temperature. A third wash with PBS only was performed followed by a centrifugation at 12,000 g for 2 min. Cell pellets were divided in two: one half for RNA extraction and the other half for DNA extraction.

RNA samples. Total RNA was extracted from each granulosa cell sample (*N* = 83) and purified using the PicoPure RNA Isolation Kit (Life Technologies). After DNAseI digestion (Qiagen), RNA quality and concentration were checked using a Bioanalyzer (Agilent). Extracted RNA samples were considered to be of good quality when the RNA integrity number was >7.

DNA samples. Genomic DNA was extracted from three different sources of material. First from individual granulosa

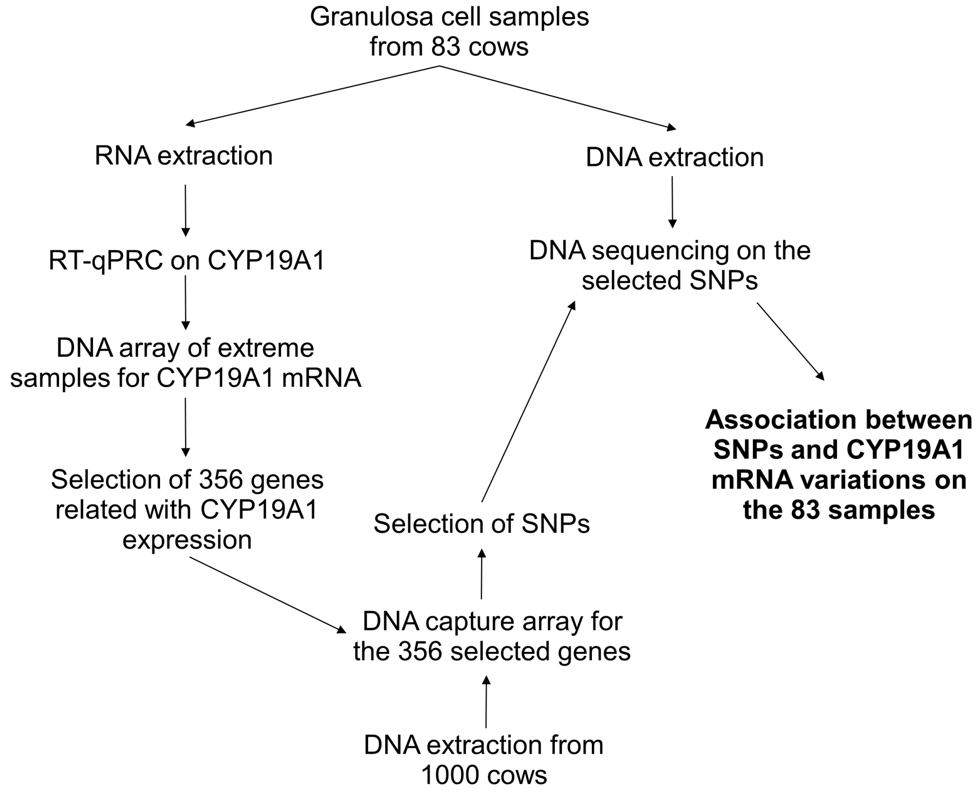
Granulosa cell samples from 83 cows
RNA extraction
DNA extraction
RT-qPRC on CYP19A1
DNA array of extreme samples for CYP19A1 mRNA
Selection of 356 genes related with CYP19A1 expression
DNA sequencing on the selected SNPs
Association between SNPs and CYP19A1 mRNA variations on the 83 samples
Selection of SNPs
DNA capture array for the 356 selected genes
DNA extraction from 1000 cows

Figure 1. Flowchart of the experimental design. RT-qPCR for the selection of the plus/minus aromatase samples, microarrays for the selection of differentially expressed transcripts, and their DNA capture and sequencing for SNPs identification to finally lead to the association of SNPs with the aromatase phenotype.

cell samples ($n = 83$); second, from pooled granulosa cells from 1000 cows; and third, from bull semen ($n = 40$) from one commercial straw. The genomic DNA extractions were performed using the Tissue & Cells genomic prep Mini Spin kit (GE Healthcare). Genomic DNA quality was checked by running the samples on a 0.5% agarose gel to verify DNA integrity, and then quantified using the Nanodrop ND-1000 (Nanodrop Technologies).

CYP19A1 RT-qPCR. The cDNA templates for RT-qPCR were produced from 50 ng of RNA using the qScript flex cDNA synthesis kit with Oligo-dT primers (Quanta Biosciences). The quantitative polymerase chain reaction was then performed using the LightCycler480 (Roche) with the Fast Start SYBR Green I, according to the manufacturer's instructions. Primers for the *CYP19A1* gene were designed with the Primerquest software (Integrated DNA Technologies). Quantities of *CYP19A1* mRNA were normalized to the quantities of three housekeeping genes: *ACTB* (actin, beta), *GADPH* (glyceraldehyde-3-phosphate dehydrogenase), and *CYP11A1* with the package SLqPCR from Bioconductor (http://www.bioconductor.org/packages/2.12/bioc/html/SLqPCR.html), using the methodology of the geometrical mean.[17] The ratio of the abundance of *CYP19A1* mRNA to housekeeping genes mRNA was then transformed by the function $-\log_{10}(x)$.

Microarray. Purified total RNA was amplified using the RiboAmp HS$^{\text{Plus}}$ RNA amplification kit (Life Technologies), labeled with Cy3 and Cy5 using the ULS Fluorescent Labeling Kit (Kreatech), and hybridized on the Agilent-manufactured EmbryoGENE[18] slide in a two-color swap design. A first group, composed of the five samples with the lowest *CYP19A1* mRNA quantity (as determined by RT-qPCR), was hybridized against a second group, composed of the five samples with the highest *CYP19A1* mRNA quantity. One nanogram of total mRNA was used for each sample inside each group.

After 17 h of hybridization at 65°C, the microarray slides were washed for 1 min in gene Expression Wash Buffer 1 (room temperature), 1 min in gene Expression Wash Buffer 2 (42°C), 10 sec in 100% acetonitrile (room temperature), and 30 sec in Stabilization and Drying Solution (Agilent). Slides were scanned with a PowerScanner (Tecan), and features extraction was done with Array-pro6.3 (Media Cybernetics). Intensity files were analyzed with FlexArray 1.6.1.[19] Specifically, raw data were corrected by background subtraction and then normalized within and between each array (Loess and quantile, respectively). Statistical comparison between the two groups (lowest *CYP19A1* mRNA level versus highest *CYP19A1* mRNA level) was done with the Limma algorithm. A significant difference in the expression level of a gene between the two groups was found if the *P*-value was below 0.05 after Bonferroni correction.

Pathway analysis. Ingenuity Pathway Analysis (http://www.ingenuity.com) was used to identify molecular pathways related to the differentially expressed genes identified by the transcriptomic array. Genes introduced into the pipeline had a fold change >1.5 (absolute mean) and a *P*-value (after Bonferroni correction) <0.05. The pathways identified were filtered according to their *P*-value calculated by the pipeline.

DNA capture. The Nimblegen Capture Array 2.1M was designed to capture DNA from the selected genes ($N = 351$). The capture was done on pooled DNA from 1000 cows according to the manufacturer's protocol (Roche/Nimblegen) and the Institut de Biologie Intégrative et des Systèmes (Université Laval, Québec, CA). Briefly, DNA was sheared by sonication and adaptors were ligated to the resulting fragments. DNA was amplified by ligation-mediated PCR, purified, and hybridized to the capture array at 42.0°C using the manufacturer's buffer. The array was washed twice at 47.5°C and three more times at room temperature using the manufacturer's buffers. Bound genomic DNA was eluted, purified, and amplified by ligation-mediated PCR. The average length of fragments captured on the array was 200 base and 89.7% of asked bases were captured on the array, which represents 19.4 million bases.

Sequencing. The DNA captured by the Nimblegen array was washed and sequenced by paired-end on one lane of an Illumina HiSeq array at McGill University (Montréal, QC, Canada) according to the manufacturer's instructions. Sequence alignment and SNP analysis were performed by the Institut de Biologie Intégrative et des Systèmes (Université Laval, Québec, QC, Canada). The number of reads was 130 million. The reads were aligned on the BTA3.1 genome and the mean base coverage was 73.61x. The tool Genome Browser (http://emb-bioinfo.fsaa.ulaval.ca/bioinfo/html/index.html), developed by the EmbryoGENE network at Université Laval for the bovine genome, was used to visualize and analyze the SNP data to select those that were the most interesting for this study.

SNP genotyping. The SNP selection was based first on the gene role and function according to pathway analysis, and second on the SNP position inside the gene. SNP genotyping was done by high-resolution melting curve, using the LightCycler480 (Roche) with the Gene Scanning protocol, in respect of the conditions defined by the manufacturer. Primers were designed by the Primer3 tool National Center for Biotechnology Information (NCBI) to generate a DNA fragment containing the SNP. Primer sequences and melting temperatures can be found in Supplementary Table 1. Fifty nanograms of DNA from each granulosa cell sample ($n = 83$) and each bull semen sample ($n = 40$) were used for genotyping.

Statistical analyses. All the statistical analyses were performed by R (2.12.1) (http://www.r-project.org/). Genetic analysis and association/modeling were performed with the library Genetics. The modeling was performed using the linear model function of R (lm function). The model used was the following:

$$Yi = \sum k\beta k\ SNPk(i) + ei + b$$

where *Yi* is the phenotypic value for the sample *i* (the ratio of the CYP19A1 mRNA abundance), *βk* is the linear regression coefficient for the SNP *k*, *SNPk*(*i*) is the genotype of the sample *i* for the SNP *k*, *ei* is the residual for the sample *i*, and *b* is the intercept of the model.

Results

Phenotyping. Aromatase (*CYP19A1*) mRNA was quantified by RT-qPCR in 83 samples as described in Materials and Methods. The *CYP19A1* mRNA ratio from four samples was not determined, and three samples were considered as outliers and removed for the analysis; therefore, 76 samples remained. The quantities of *CYP19A1* mRNA constituted the phenotypic variable to predict, with a ratio (mRNA/housekeeping gene) mean of 1.66 and a coefficient of variation of 28.9%.

To analyze the transcriptomic profiles related to the difference in *CYP19A1* mRNA abundance, two groups were constituted. The first one, "Aromatase Plus", contained the five samples with the highest *CYP19A1* mRNA ratio (mean of 2.99). The second group, "Aromatase Minus", contained the five samples with the lowest *CYP19A1* ratio (mean of 1.07).

Microarray. The transcriptomic analysis contrasting the two groups Aromatase Plus/Aromatase Minus identified 1,805 differentially expressed transcripts with a fold change value >1.5 in absolute mean and an associated *P*-value <0.05 after Bonferroni correction. The *CYP19A1* gene was overexpressed with a fold change of 1.53 and a *P*-value inferior to 0.001 in this contrast.

Pathways. Differentially expressed transcripts were imported in the Ingenuity Pathway Analysis pipeline to identify pathways affected by the *CYP19A1* differential expression pattern. This analysis showed that the differentially expressed genes constituted 10 different networks with different functions (Table 1). Network 1 represented cell death and gene expression, and *CYP19A1*, our main target, was included in this network. The first 10 canonical pathways, in term of significance, are illustrated in Figure 2, which also shows their relative proportion according to the number of genes involved. These pathways included the categories of molecular network involved in the aromatase differential gene expression: metabolism, oxidative stress, immune response, estrogen signaling, apoptosis, and coagulation. Using the IPA software, we were able to analyze the differences in terms of major upregulated functions within those categories associated to high and low aromatase expression (Table 2).

Genotyping of 1000 cows. By filtering the genes based on their accuracy and importance in the pathways, or known to have an important function were also kept (regulators for example), a list of 355 genes related to the aromatase expression profiles in the two groups of samples was generated. Those 355 selected genes were genotyped on a panel of 1000 cows to determine the genetic variability in these genomic regions. The SNPs were filtered and cleaned based on the data quality and SNP frequencies. SNPs with a minor allele frequency (MAF) inferior to 10% were not selected since they are not interesting from a selection scheme point of view. A list of 23,388 SNPs was generated from those 355 genes, which constituted our genetic database.

Table 1. Pathways identified has being affected by the *CYP19A1* differential expression pattern.

ID NETWORK	FUNCTION
1	Cell Death, Gene Expression
2	Lipid Metabolism, Small Molecule Biochemistry, Molecular Transport
3	Cellular Movement, Cellular Development, Tissue Development
4	Cell-To-Cell Signaling and Interaction, Nervous System Development and Function
5	Gene Expression, Cellular Growth and Proliferation, Cellular Development
6	Gene Expression
7	Cell Death, DNA Replication, Recombination, and Repair, Cellular Growth and Proliferation
8	RNA Post-Transcriptional Modification, Cell-To-Cell Signaling and Interaction, Tissue Development
9	Cellular Movement, Cellular Assembly and Organization, Cellular Function and Maintenance
10	Cell Cycle, DNA Replication, Recombination, and Repair, Cellular Assembly and Organization

Notes: This analysis showed that the differentially expressed genes constituted ten different networks with different functions. Networks identified by the pathway analysis. Functions were given by the IPA software.

Gene list. The selection of genes to be genotyped on the 83 studied samples was based on the fold change in the transcriptomic analysis study and the biological relevance with the phenotype as assessed by Ingenuity Pathway Analyzer. The density of known SNPs in our genetic database was also taken into consideration to technically allow genotyping with the high-resolution melting curve. Nine genes were selected: *KRT8* (keratin 8), *CYP19A1* (cytochrome P450, family 19, subfamily A, polypeptide 1), *TBX18* (T-box 18), *CREB1* (cAMP-responsive element binding protein 1), *LHCGR* (luteinizing hormone/choriogonadotropin receptor), *ANXA1* (annexin A1), *GPNMB* (glycoprotein [transmembrane] nmb), *PLXD2* (plexin domain containing 2), and *LOXL4* (lysyl oxidase-like 4) (Table 3).

Genotyping of granulosa cell samples. The nine selected genes were investigated to determine which SNPs inside these genes would be interesting to genotype. The first criterion for SNP selection was based on a technical limitation related to the high melting resolution technique, which was edited by the manufacturer: SNPs should be at least 100–150 bp from each other, and the variation should not be A to T and G to C. The second criterion taken into consideration was the MAF estimated on a Holstein-based population. A minimal MAF of 0.2 was considered, to generate SNPs with a minimal interest for selection by the industry. The third criterion used was the SNP position in the gene and its potential value as a marker: SNPs in regulatory elements or near exon–intron junctions were preferred. Eighteen SNPs were selected for

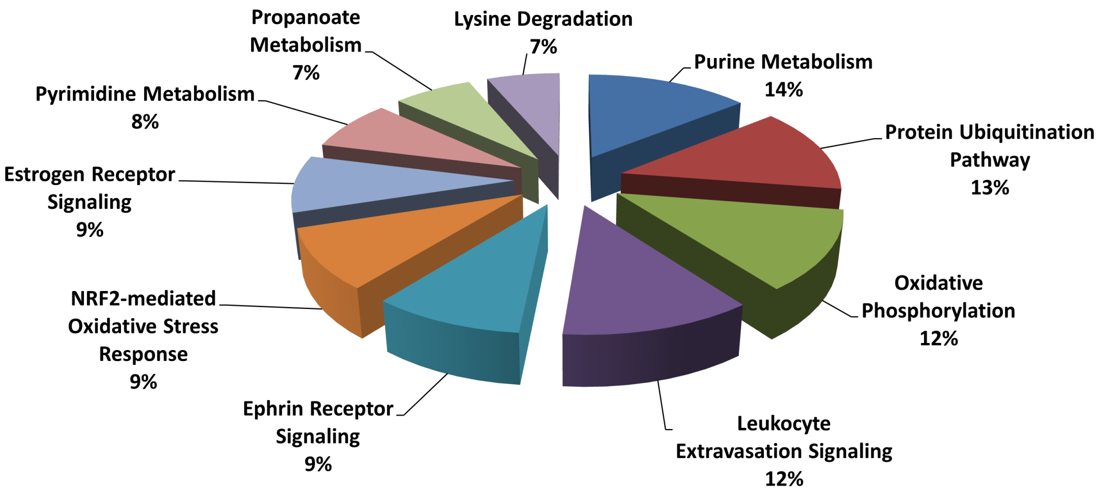

Figure 2. Canonical pathways. The first 10 canonical pathways, in terms of significance, are illustrated here. The percentage represents the relative proportion in the number of genes inside each pathway. These pathways included the categories of molecular network involved in the aromatase differential gene expression.

Table 2. Modification of pathways between fertile and sub-fertile follicles.

	PATHWAY	CHANGE
Metabolism	Xanthine	–
	Adenosine	+
	dGTP	–
	deoxyGTP	+
	Proteasome	+
	Deubiquitination	–
	Ubiquitination	+
	Uracile	+
	Carnitine	–
	Ubiquinol	+
	Pyruvate	+
	Fumarate	+
	Citrate	–
	Acetyl-CoA	+
	ATP	+
Oxydative stress	Heat Shock Proteins	+
	ROS	+
	Antioxydant	+
Immune	Leukocyte extravasation	–
	Chemotraction/Adhesion	–
	Cell retraction	+
	Vascular contraction	+
Estrogen signaling		+
Apoptose		–
Coagulation		+

Notes: This table gives the major biological functions associated with the variation in *CYP19A1* gene expression. Those can be directly or indirectly linked. (+/–) Means the pathway is significantly more or less expressed in fertile cows. Analyses done by the IPA software.

genotyping of our samples (Table 4) based on these criteria. The genotyping of 18 SNPs on 76 samples resulted in 1,368 data sets. Six genotypes from two samples were impossible to determine. These samples were deleted from the studied population for the modeling procedure; therefore, 74 samples remained.

Association/modeling. The phenotypic and genotypic data were analyzed with the Genetic library of R. The linkage disequilibrium (LD) was estimated by the R^2 between SNPs inside a gene. Only KRT8-3/KRT8-5 and LOXL4-1/LOXL4-4 had significant LD values, respectively, at 0.71 and 0.9. The SNP TBX18-2 presented an LD value of 0.35 with PLXDC2-1.

A first iteration gave a model with an R^2 value of 0.07 and a *P*-value of 0.27. No SNPs and no genotypes were significant in this model. After this first iteration, the three SNPs with high LD (KRT8-5, LOXL4-1, and TBX18-2) were removed from the analysis, as their inclusion in the different models slightly decreased the quality of the prediction. The second iteration gave a model with an R^2 of 0.08 and a *P*-value of 0.2. Three SNPs were significant for four different significant genotypes. With this iteration, four samples were found as outliers in the model and were removed for the following step. A third iteration was performed and gave a model with an R^2 of 0.46 and a *P*-value of 6.88e^{-5}. Seven SNPs were significant for nine different genotypes. For this third iteration, 70 samples and 15 SNPs were used. A summary of the different iterations is presented in Table 5.

In this model, we found seven SNPs significantly associated with the aromatase mRNA level: KRT8-3, CYP19A1-1, CREB1-4, CREB1-5 (2 genotypes), LHCGR-3 (2 genotypes), LHCGR-7, and ANXA1-3 (Table 6). The QQ-plot of the last iteration is illustrated in Figure 3 and shows that the model is relevant.

Table 3. Fold change and *P*-value of the mRNA levels for genes targeted in this study and obtained by microarray analysis.

GENE SYMBOL	NAME	CHROMOSOME	FOLD CHANGE	*P*-VALUE
KRT8	Keratin 8	5	–5.68	1.63E-11
CYP19A1	Cytochrome P450. family 19. subfamily A. polypeptide 1	10	1.53	3.00E-04
TBX18	T-box 18	9	–2.86	7.72E-07
ANKDR1	Ankyrin repeat domain 1 (cardiac muscle)	26	–3.85	9.24E-07
CREB1	CAMP responsive element binding protein 1	2	0	1.00E+00
LHCGR	Luteinizing hormone/choriogonadotropin receptor	11	3.1	2.84E-07
ANXA1	Annexin A1	8	–2.76	4.48E-07
GPNMB	Glycoprotein (transmembrane) nmb	4	–4.03	1.89E-09
PLXDC2	Plexin domain containing 2	13	–3.19	1.17E-07
LOXL4	Lysyl oxidase-like 4	26	–4.02	3.39E-09

Note: *CYP19A1* and nine genes were selected based on the fold change in the transcriptomic analysis study and the biological relevance with the phenotype.

A simplified model was built in a fourth iteration which took into account only the significant SNPs identified in the Iteration III. This model gave an R^2 of 0.44 and a *P*-value of 1.54e^{-6}. The parameters are summarized in Table 7. Using the models, we calculated the *CYP19A1* mRNA ratio predicted from SNPs in our data set. The third and fourth iterations gave error rates of 15% and 17%, respectively. Correlations between the measured and predicted phenotypic values were 0.8 and 0.72 for Iterations III and IV, respectively.

Genotyping bull semen. To demonstrate the beneficial use of genetical genomics for the fertility thematic and in order to evaluate our model, we selected 40 bulls with known genetic evaluation. We used the seven SNPs with significance extracted from Iteration IV and ran our model. The model was able to significantly correlate four phenotypic characterizations (Table 8) about health/fertility, calving rate, daughters' fertility, and calving rate.

Discussion

This analysis is the first application of genomic–genetic studies to the ovarian aspect of bovine infertility. The method used here revealed a limited number of highly significant

Table 4. SNPs select for genotyping in our studied population.

SNP	GENE	CHROMOSOME	POSITION	GENE REGION	MAF	REFERENT	VARIANT	BOVINE SNP50
ANKDR1-1	*ANKDR1*	26	12574849	Exon—Read change	0.22	A	G	No
ANXA1-3	*ANXA1*	8	49637458	Exon	0.2	T	C	No
CREB1-4	*CREB1*	2	96295119	1.5 kb from 5′UTR	0.25	T	G	No
CREB1-5	*CREB1*	2	96302726	Near intron/exon junction	0.54	A	G	No
CYP19A1-1	*CYP19A1*	10	59223175	4.7 kb from 5′UTR	0.74	T	C	No
GPNMB-2	*GPNMB*	4	32010219	0.6 kb from 5′UTR	0.54	A	G	Yes
GPNMB-3	*GPNMB*	4	32020662	1.4 kb from 3′UTR	0.33	T	G	No
KRT8-1	*KRT8*	5	27212607	1 kb from 5′UTR	0.42	G	A	No
KRT8-3	*KRT8*	5	27216778	Near intron/exon junction	0.45	T	C	Yes
KRT8-5	*KRT8*	5	27221526	0.4 kb from 3′UTR	0.48	A	G	Yes
LHCGR-3	*LHCGR*	11	30889191	1 kb from 5′UTR	0.53	A	G	No
LHCGR-7	*LHCGR*	11	30865516	Near intron/exon junction	0.39	G	T	No
LOXL4-1	*LOXL4*	26	19228026	4.5 kb from 5′UTR	0.39	C	T	No
LOXL4-4	*LOXL4*	26	19202400	3′UTR	0.52	T	C	No
PLXDC2-1	*PLXDC2*	13	21581875	13 kb from 5′UTR	0.44	G	A	No
PLXDC2-3	*PLXDC2*	13	21808961	3′UTR	0.53	A	C	No
TBX18-1	*TBX18*	9	65570080	9 kb from 5′UTR	0.75	A	G	No
TBX18-2	*TBX18*	9	65597983	Intron	0.36	G	A	No

Notes: The presence/absence of the SNPs in the available BovineSNP50 v2 DNA Analysis BeadChip (Illumina) is indicated. Referent used is the Hereford breed. Variant are indicated for the Holstein breed.

Table 5. Results of the different iterations.

	ITERATION I	ITERATION II	ITERATION III
Model R^2	0.07	0.08	0.46
Model *P*-value	0.27	0.2	6.88E-05
Significant SNP	0	3	7
Significant genotypes	0	4	9

Notes: The phenotypic and genotypic data were analyzed with the Genetic library of R. The Linkage Desequilibrium (LD) was estimated by the R^2 between SNPs inside a gene. Iteration I: all SNPs and samples taken into account. Iteration II: SNPs KRT8-5, TBX18-2 and LOXL4-1 removed. Iteration III: 4 outlier samples removed.

biomarkers associated with aromatase expression and therefore relevant for the infertility phenotype.

Pathway analysis. This analysis showed that the molecular network associated with the aromatase gene expression is complex since 10 different networks are involved. These networks could have direct effects on *CYP19A1* expression (network 1) or indirect effects (other networks). Several factors are involved in the control of the aromatase gene expression (Fig. 3) and many more can be used to assess this expression through a marker function. As a consequence, aromatase expression can be modulated by multiple conditions, from genetic to environmental.

The analysis of the pathways in high and low aromatase follicles revealed some differences between these two groups (Table 2). Estrogenic follicles tended to have a high oxidative metabolism and a better capacity to counter oxidative stress. Estrogenic follicles also induced changes in the immune-related pathways with a decrease in leukocyte signaling. The immune system is present in reproductive organs and is involved in normal reproductive functions as inflammation accompanies ovulation and atresia. Since the follicles selected were not preovulatory (size limitation), lower leukocyte signaling probably means a lower level of atresia.[20]

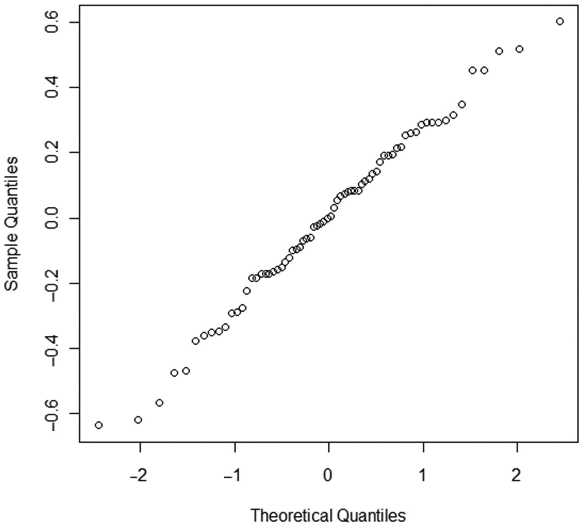

Figure 3. QQ-plot of Iteration III. The QQ-plot show that the model has an R^2 of 0.46 and a *P*-value of $6.88e^{-5}$ and is accurate with the samples and no outliers are present.

Table 6. Parameters of the Iteration III.

	COEFFICIENT	*P*-VALUE	SIGNIFICANCE
Intercept	**2.28**	**0.0004**	***
KRT8-3 C/T	**–0.35**	**0.001**	**
CYP19A1-1 C/T	0.16	0.18	
CYP19A1-1 T/T	**–1.13**	**0.01**	**
KRT8-1 A/G	0.07	0.61	
TBX18-1 G/A	0.40	0.18	
TBX18-1 G/G	0.42	0.16	
ANKDR1-1 A/G	–0.14	0.23	
ANKDR1-1 G/G	0.07	0.78	
CREB1-4 T/T	**–0.19**	**0.05**	*
CREB1-5 G/A	**–0.37**	**0.008**	**
CREB1-5 G/G	**–0.46**	**0.001**	**
LHCGR-3 G/A	**–0.58**	**0.002**	**
LHCGR-3 G/G	**–0.43**	**0.02**	*
LHCGR-7 T/T	**0.97**	**0.000001**	***
ANXA1-3 T/T	**–0.22**	**0.04**	*
GPNMB-2 G/A	–0.24	0.41	
GPNMB-2 G/G	0.04	0.89	
GPNMB-3 T/T	–0.12	0.22	
PLXDC2-1 G/G	0.08	0.43	
PLXDC2-3 C/A	0.05	0.82	
PLXDC2-3 C/C	0.16	0.54	
LOXL4-4 G/G	0.09	0.49	

Notes: Iteration III is the complete model, after deleting the outliers. Seven SNPs were significant for nine different genotypes. Significant parameters are in bold. (0.001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1).

Table 7. Parameters of the Iteration IV.

	COEFFICIENT	*P*-VALUE	SIGNIFICANCE
Intercept	**2.58**	**0.0000**	***
KRT8-3 C/T	**–0.33**	**0.001**	**
CYP19A1-1 C/T	0.05	0.56	
CYP19A1-1 T/T	**–0.79**	**0.04**	*
CREB1-4 T/T	–0.15	0.10	
CREB1-5 G/A	**–0.40**	**0.003**	**
CREB1-5 G/G	**–0.49**	**0.0003**	***
LHCGR-3 G/A	**–0.43**	**0.01**	**
LHCGR-3 G/G	–0.33	0.06	t
LHCGR-7 T/T	**0.89**	**0.000002**	***
ANXA1-3 T/T	**–0.18**	**0.047**	*

Notes: Iteration IV is Iteration III upgraded, with the significant parameters. This model gave a R^2 of 0.44 and a *P*-value of $1.54e^{-6}$. This iteration is easier to use because it requires less SNPs to genotype. Significant parameters are in bold (0.001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1).

Table 8. Parameters obtained on 7 SNPs from the Iteration IV run on 40 bull semen samples.

	COEFFICIENT	*P*-VALUE	SIGNIFICANCE
Daughter's fertility			
Intercept	104.88	8.42E-10	***
KRT8-3 C/T	**12.57**	**0.01**	*
LHCGR-7 T/T	–3.87	0.45	
CYP19A1-1 C/T	–4.42	0.50	
CYP19A1-1 T/T	11.50	0.43	
CREB1-5 G/G	–8.36	0.12	
LHCGR-5 A/G	–4.55	0.65	
LHCGR-5 G/G	–9.92	0.34	
CREB1-4 T/T	0.70	0.89	
ANXA1-3 T/T	1.64	0.74	
Calving			
Intercept	103.42	<2e-16	***
KRT8-3 C/T	0.80	0.49	
LHCGR-7 T/T	0.30	0.82	
CYP19A1-1 C/T	–1.67	0.33	
CYP19A1-1 T/T	–1.14	0.74	
CREB1-5 G/G	–0.42	0.75	
LHCGR-5 A/G	–2.99	0.22	
LHCGR-5 G/G	–4.35	0.09	
CREB1-4 T/T	**2.60**	**0.04**	*
ANXA1-3 T/T	–0.91	0.45	
Daughter's calving			
Intercept	103.79	<2e–16	***
KRT8-3 C/T	1.62	0.36	
LHCGR-7 T/T	–3.58	0.06	.
CYP19A1-1 C/T	–1.40	0.55	
CYP19A1-1 T/T	1.12	0.83	
CREB1-5 G/G	–0.77	0.69	
LHCGR-5 A/G	–4.21	0.25	
LHCGR-5 G/G	–3.28	0.39	
CREB1-4 T/T	2.05	0.26	
ANXA1-3 T/T	0.46	0.80	
Health/Fertility			
Intercept	311.92	0.07	
KRT8-3 C/T	**200.40**	**0.01**	**
LHCGR-7 T/T	–76.12	0.35	
CYP19A1-1 C/T	–75.57	0.45	
CYP19A1-1 T/T	123.71	0.57	
CREB1-5 G/G	–138.69	0.10	
LHCGR-5 A/G	–96.39	0.53	
LHCGR-5 G/G	–136.51	0.39	
CREB1-4 T/T	66.13	0.39	
ANXA1-3 T/T	35.04	0.65	

Notes: Significant parameters are in bold (<0.001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1).

Upregulation of the estrogen receptor is critical for normal granulosa cell development and a defect in this pathway can have important consequences for fertility.[21] Lower apoptosis in fertile follicles is logical, as apoptosis occurs in atretic follicles and is a normal process of follicle selection.[22] The coagulation pathway seems to be important in competent follicles according to this study and recent publications from mouse studies[23] and our group have further identified *SERPINE1* (Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1), *SERPINA5* (serpin peptidase inhibitor, clade A [alpha-1 antiproteinase, antitrypsin], member 5), *F3* (coagulation factor III (thromboplastin, tissue factor), *A2M* (alpha-2-macroglobulin), *PLAUR* (plasminogen activator, urokinase receptor), and *F2R* (coagulation factor II [thrombin] receptor) as a determinant of oocyte quality.[24,25]

Gene selection. In this study, genes with high fold change values, interesting positions in pathways, and good genetic opportunities (availabilities of SNPs and their MAFs) were selected. Complementary studies could be performed to enrich the existing model with other genes to improve the model accuracy. However, the results obtained demonstrated that our gene selection was accurate enough to provide a good model to predict the tested phenotype (R^2 of 0.44 for a P-value of 1.54e^{-6}).

Genes of the model. In the final iteration (Iteration IV), we found six SNPs significantly related to *CYP19A1* expression on five genes: *KRT8*, *CYP19A1*, *CREB1*, *LHCGR*, and *ANXA1*. Of course, the presence of an SNP in the *CYP19A1* gene itself was not surprising. Its position, at 4.7 kb from the 5′UTR (Table 4), could directly affect *CYP19A1* gene expression regulation since the genotype T/T is relevant to the model. In a study on oocyte competence, the gene *KRT8* was previously identified as being relevant to fertility.[26] This gene was also found to be underexpressed in preovulatory granulosa cells.[23] In the present study, this gene had the highest negative fold change and was the top candidate to analyze. The SNP selected on KRT8-3 is positioned at 16 bp from an exon–intron junction (Table 4) and could therefore play a role in the regulation of *KRT8* mRNA splicing. The genotype C/T was significantly associated with *CYP19A1* mRNA abundance.

The transcription factor *CREB* is associated with the response to gonadotropins (LH and FSH) and requires *CREB1* for its synthesis; *CREB1* was analyzed in this study and pathway analysis revealed that it is an important node linking several genes with differential fold change, notably *CYP19A1*. Two SNPs of *CREB1* were identified: CREB1-4 and CREB1-5. The former appeared in the third iteration (with all genes in the model) but not in the fourth (with only genes related to *CYP19A1* expression). Its position at 1.5 kb from the 5′UTR could explain its role in the regulation of *CREB1* expression. Because of its absence in the fourth iteration, this SNP was not the most interesting one from this gene. The SNP CREB1-5 was identified in this study as being significantly related to

CYP19A1 expression. Its position near an intron–exon junction (less than 150 bp) could explain its role in the regulation of *CREB1* expression (Table 4). Since *CREB* is a transcription factor related to *CYP19A1*, we can propose that the SNP CREB1-5 has a trans effect on *CYP19A1* expression regulation. The two different genotypes (G/A and G/G) had a significant effect on *CYP19A1* mRNA abundance.

The ovarian LH receptor is *LHCGR*. Stimulation of *LHCGR* by LH provokes the aromatization of androgens to estradiol by the action of *CYP19A1* in granulosa cells from dominant follicles. This receptor is therefore important for fertility and is the target of several studies.[27,28] The two SNPs studied in this gene, LHCGR-3 and LHCGR-7, were significantly associated with *CYP19A1* mRNA abundance. The SNP LHCGR-3 is located 1 kb from the 5′UTR and could potentially have a functional role in expression regulation (Table 4). Two genotypes are relevant: G/A and G/G but only the G/G genotype had a significant association with aromatase mRNA levels in the third and fourth iteration. The genotype G/A only had a tendency for association with aromatase mRNA in the fourth iteration, which makes it less interesting. The SNP LHCGR-7 is situated 45 bp from an intron–exon junction (Table 4), so it could affect the splicing of *LHCGR* mRNA. There are seven isoforms of this receptor in bovine granulosa cells[29] and as LHCGR is in the same pathway as *CYP19A1*, SNPs inside *LHCGR* may affect *CYP19A1* in a trans effect.

The gene *ANXA1* codes for an anti-inflammatory protein. The SNP identified in this gene, ANXA1-3, is situated in an exon (Table 4). The triplet containing this SNP codes for isoleucine, with the T referent. If the genotype changes for a C, the amino acid changes for a threonine (mistranslation mutation). Those two amino acids have different properties: the isoleucine is apolar, contrary to the threonine. The protein-folding process, and the final protein properties, could be then affected, explaining why this SNP had a significant effect on our phenotype. As the inflammatory process is important in ovulation as discussed above,[30] the SNP in *ANXA1* could act in trans, in a pathway correlated with *CYP19A1* mRNA abundance. Only the genotype T/T is significantly relevant to the mRNA abundance.

Bull analysis. Samples from 40 bulls were used to assess the possible association of the seven SNPs with known fertility phenotypes. The model was able to significantly correlate four phenotypes: health/fertility, calving rate, and the daughters' fertility and calving rate. As the markers were related to an enzyme regulating follicular function, it makes sense that the SNPs were associated with fertility more than with calving. Although calving ease could be associated with the amount of estradiol at the time of parturition, the main effect observed for this phenotype is associated with calf size and therefore not directly related to ovarian function.

Accuracy of the methodology. With this methodology, we can study precise regions of genes related to the target phenotype, according to their relevance with the gene function (exon, exon–intron junction, UTR regions). The SNPs identified (KRT8-3, CYP19A1-1, ANXA1-3, CREB1-5, LHCGR-3 and LHCGR-7) belonged to these different categories (Table 4) and could identify causative SNPs inside a known QTL. For example, the SNP marker KRT8-3 is inside a QTL for easy calving and present on the BovineSNP50 v2 DNA Analysis BeadChip (Illumina) (Table 4). The SNP ANXA1-3, which is also inside a QTL for easy calving, is not on the bovine chip. The majority of our SNPs was not on the commercially available bovine chips and therefore could not be used at this time to assess fertility. This study demonstrated that this methodology is able to enrich the current knowledge in genetic markers for complex phenotypes. We were able to finely dissect the complex trait in some SNP effects and identify eQTL with cis and trans effects. These SNPs could be effective markers and could avoid some LD issues between generations. Used as genetic markers for selection, the identified SNPs could increase accuracy of selection after a few generations. The last iteration (Iteration IV) showed that seven SNPs were enough for a prediction test for *CYP19A1* mRNA abundance.

The generated model was able to predict 44% of the phenotype variability (Iteration IV), which is a good result suitable for use by the dairy industry. Taking the sum of all the coefficients of the model for an animal, we can directly calculate a breeding value for the animal. The error rate in the prediction was satisfactory, at 17%.

This methodology presents the advantage of having more observations than variables (74 samples and 18 SNPs in this study), contrary to genetic association studies, where the main problem is to have much more variables than observations (millions of variables versus only hundreds of samples). Consequently, this methodology allows us to bypass one of the major problems in genetic analysis and is simpler to do.

Moreover, the experimental scheme does not require a complex design with related animals on two or three generations. Unrelated and unknown animals were used in this study.

Conclusion

This study, based on a genomic–genetic approach, was able to identify biomarkers related to a complex phenotype, demonstrating its relevance to find genetic targets for such phenotypes.

Because of its simple design, this methodology is easy to use in genomic laboratories, for lower costs than a complex genetic association study. Therefore, the genetical genomics approach is a good complement to genetic association studies, as it is possible to zoom inside a QTL of interest previously identified by genetic studies to identify the causative SNPs.

Acknowledgment

The authors thank Isabelle Laflamme for her technical assistance with granulosa cell RNA extraction.

Author Contributions

Designed and conducted the experiments, analyzed and interpreted the data, and drafted the manuscript: NG. Conceived and designed the experiments and revised the manuscript: ID. Oversaw the project and revised and approved the manuscript: MAS. All authors reviewed and approved of the final manuscript.

Supplementary Data

Supplementary table 1. RT-qPCR primer sequences, annealing temperature and accession number.

REFERENCES

1. Kommadath A, Mulder HA, de Wit AA, et al. Gene expression patterns in anterior pituitary associated with quantitative measure of oestrous behaviour in dairy cows. *Animal*. 2010;4(8):1297–1307.
2. Friggens NC, Disenhaus C, Petit HV. Nutritional sub-fertility in the dairy cow: towards improved reproductive management through a better biological understanding. *Animal*. 2010;4(7):1197–1213.
3. Lenz S, Pohland R, Becker F, Vanselow J. Expression of the bovine aromatase cytochrome P450 gene (Cyp19) is primarily regulated by promoter 2 in bovine follicles and by promoter 1.1 in corpora lutea. *Mol Reprod Dev*. 2004;67(4):406–413.
4. Vanselow J, Pohland R, Furbass R. Promoter-2–derived Cyp19 expression in bovine granulosa cells coincides with gene-specific DNA hypo-methylation. *Mol Cell Endocrinol*. 2005;233(1–2):57–64.
5. Simpson ER, Clyne C, Rubin G, et al. Aromatase—a brief overview. *Annu Rev Physiol*. 2002;64:93–127.
6. Manikkam M, Bao B, Rosenfeld CS, et al. Expression of the bovine oestrogen receptor-beta (bER beta) messenger ribonucleic acid (mRNA) during the first ovarian follicular wave and lack of change in the expression of bER beta mRNA of second wave follicles after LH infusion into cows. *Anim Reprod Sci*. 2001;67(3–4):159–169.
7. Sangsritavong S, Combs DK, Sartori R, Armentano LE, Wiltbank MC. High feed intake increases liver blood flow and metabolism of progesterone and estradiol-17beta in dairy cattle. *J Dairy Sci*. 2002;85(11):2831–2842.
8. Yapura J, Mapletoft RJ, Pierson RA, Singh J, Adams GP. Aromatase inhibitor treatment with an intravaginal device and its effect on pre-ovulatory ovarian follicles in a bovine model. *Reprod Biol Endocrinol*. 2013;11:97.
9. Dias FC, Costa E, Adams GP, et al. Effect of duration of the growing phase of ovulatory follicles on oocyte competence in superstimulated cattle. *Reprod Fertil Dev*. 2013;25(3):523–530.
10. Nivet AL, Bunel A, Labrecque R, et al. FSH withdrawal improves developmental competence of oocytes in the bovine model. *Reproduction*. 2012;143(2):165–171.
11. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet*. 2001;17(7):388–391.
12. Brem R, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science*. 2002;296(5568):752–755.
13. de Koning DJ, Cabrera CP, Haley CS. Genetical genomics: combining gene expression with marker genotypes in poultry. *Poult Sci*. 2007;86(7):1501–1509.
14. de Koning DJ, Haley CS. Genetical genomics in humans and model organisms. *Trends Genet*. 2005;21(7):377–381.
15. Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. *Nature*. 2004;430(7001):743–747.
16. Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. *Hum Mol Genet*. 2005;14(suppl 2):R163–R169.
17. Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. 2002;3(7):RESEARCH0034.
18. Robert C, Nieminen J, Dufort I, et al. Combining resources to obtain a comprehensive survey of the bovine embryo transcriptome through deep sequencing and microarrays. *Mol Reprod Dev*. 2011;78(9):651–664.
19. Blazejczyk M, Miron M, Nadon R. *FlexArray: A Statistical Data Analysis Software for Gene Expression Microarrays*. Canada: Genome Quebec; 2007.
20. Zhang YM, Rao CV, Lei ZM. Macrophages in human reproductive tissues contain luteinizing hormone/chorionic gonadotropin receptors. *Am J Reprod Immunol*. 2003;49(2):93–100.
21. Couse JF, Yates MM, Deroo BJ, Korach KS. Estrogen receptor-beta is critical to granulosa cell differentiation and the ovulatory response to gonadotropins. *Endocrinology*. 2005;146(8):3247–3262.
22. Blondin P, Dufour M, Sirard MA. Analysis of atresia in bovine follicles using different methods: flow cytometry, enzyme-linked immunosorbent assay, and classic histology. *Biol Reprod*. 1996;54(3):631–637.
23. Nagaraja AK, Middlebrook BS, Rajanahally S, et al. Defective gonadotropin-dependent ovarian folliculogenesis and granulosa cell gene expression in inhibin-deficient mice. *Endocrinology*. 2010;151(10):4994–5006.
24. Gilbert I, Robert C, Vigneault C, Blondin P, Sirard MA. Impact of the LH surge on granulosa cell transcript levels as markers of oocyte developmental competence in cattle. *Reproduction*. 2012;143(6):735–747.
25. Nivet AL, Vigneault C, Blondin P, Sirard MA. Changes in granulosa cells' gene expression associated with increased oocyte competence in bovine. *Reproduction*. 2013;145(6):555–565.
26. Torner H, Ghanem N, Ambros C, et al. Molecular and subcellular characterisation of oocytes screened for their developmental competence based on glucose-6–phosphate dehydrogenase activity. *Reproduction*. 2008;135(2):197–212.
27. Eppig JJ. Oocyte control of ovarian follicular development and function in mammals. *Reproduction*. 2001;122(6):829–838.
28. Mihm M, Baker PJ, Fleming LM, Monteiro AM, O'Shaughnessy PJ. Differentiation of the bovine dominant follicle from the cohort upregulates mRNA expression for new tissue development genes. *Reproduction*. 2008;135(2):253–265.
29. Robert C, Gagne D, Lussier JG, Bousquet D, Barnes FL, Sirard MA. Presence of LH receptor mRNA in granulosa cells as a potential marker of oocyte developmental competence and characterization of the bovine splicing isoforms. *Reproduction*. 2003;125(3):437–446.
30. Casey OM, Morris DG, Powell R, Sreenan JM, Fitzpatrick R. Analysis of gene expression in non-regressed and regressed bovine corpus luteum tissue using a customized ovarian cDNA array. *Theriogenology*. 2005;64(9):1963–1976.