Animal (2009), **3:3**, pp 393–401 © The Animal Consortium 2008
doi:10.1017/S1751731108003479

A pilot study on transcriptome data analysis of folliculogenesis in pigs

G. Tosser-Klopp[1†a], K.-A. Lê Cao[2,3a], A. Bonnet[1], N. Gobert[1], F. Hatey[1], C. Robert-Granié[2], S. Déjean[3], J. Antic[4], L. Baschet[4] and M. SanCristobal[1]

[1]*Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, UMR444, BP 52627, 31326 Castanet Tolosan Cedex, France;*
[2]*Station d'Amélioration Génétique des Animaux, Institut National de la Recherche Agronomique (UR631), BP 52627, 31326 Castanet Tolosan Cedex, France;*
[3]*Institut de Mathématiques, Université de Toulouse et Centre National de la Recherche Scientifique (UMR5219), 31062 Toulouse Cedex 9, France;*
[4]*Département de Génie Mathématique et Modélisation, INSA, 135, Avenue de Rangueil, 31077 Toulouse Cedex 4, France*

(Received 27 April 2007; Accepted 19 September 2008; First published online 20 November 2008)

Three different stages of pig antral follicles have been studied in a granulosa-cell transcriptome analysis on nylon microarrays (1152 clones). The data have been generated from seven RNA follicle pools and several technical replicates were made. The objective of this paper was to state the feasibility of a transcriptomic protocol for the study of folliculogenesis in the pig. A statistical analysis was chosen, relying on the linear mixed model (LMM) paradigm. Low variability within technical replicates was hence checked with a LMM. Relevant genes that might be involved in the studied process were then selected. For the most significant genes, statistical methods such as principal component analysis and unsupervised hierarchical clustering were applied to assess their relevance, and a random forest analysis proved their predictive value. The selection of genes was consistent with previous studies and also allowed the identification of new genes whose role in pig folliculogenesis will be further investigated.

Keywords: microarray, pig, folliculogenesis, statistical data analysis

Introduction

Enhancement of production efficiency through improved female reproductive performance is of major importance to the pork industry. Direct selection on prolificacy has led to relatively low responses, since this trait presents a low heritability: 0.1 (Bichard and David, 1985). Thus, different types of studies have been undertaken to elucidate and improve the different components of this complex trait. The two main approaches are QTL studies and candidate gene studies (Buske *et al*., 2006). However, until now, no gene explaining a major effect on reproductive performance has been found in pigs, suggesting that this is a complex phenomenon. Possibly for this reason, authors have chosen to study the different components of reproductive performance like ovulation rate (Knox, 2005), perinatal mortality (van der Lende *et al*., 2001) or maternal qualities (Algers and Uvnas-Moberg, 2007) in isolation.

Ovulation rate is determined by a highly dynamic process involving recruitment, development, maturation and atresia of antral follicles. Follicular development depends on hormonal feedback mechanisms between the hypothalamus, the anterior lobe of the pituitary gland and the ovaries, which have also been shown to produce molecules with paracrine and autocrine functions (Foxcroft and Hunter, 1985; Foxcroft *et al*., 1989). Gonadotropins and local factors including steroids, growth factors and other regulatory peptides are known to be involved in the maturation of follicles. Follicular development is thus a complex process and requires the coordinated expression of a large number of genes.

In this study, granulosa cells were chosen as they constitute an important compartment in the mammalian ovarian follicle and are easy to isolate. They actively participate in endocrine function of the ovaries by secreting oestradiol or progesterone in response to FSH or LH stimulation (Duda, 1997).

Microarray analysis is an increasingly developing technique that enables the simultaneous expression screening of thousands of genes. These analyses have already been used for the discovery of genes involved in pig female reproductive performances: Caetano *et al*. (2004) studied whole follicle transcriptome from two pig lines differing in their ovulation rate, Agca *et al*. (2006) focused on early luteinisation process and Whitworth *et al*. (2005) studied early embryonic development.

[a] These authors have contributed equally to this work.
[†] E-mail: gwenola.tosser@toulouse.inra.fr

The current research, however, focuses on the expression profile of granulosa cells that may assist in the identification of the transcriptome involved in late ovarian follicular development and lead to an understanding of the molecular mechanisms involved in the ovulatory process and therefore prolificacy.

In an attempt to validate the biological data in microarray analysis, an important aspect is to identify the sources of variation within the data sets. Some statistical approaches include fixed or mixed effect models that give a full account for sources of variation in the context of statistical design of experiment (Kerr and Churchill, 2001). By accounting for these variations, different effects due to membranes, probes, genes and tissue samples can be estimated and systematically removed by means of the ANOVA approach. The data normalisation and inference for differentially expressed genes have been integrated into a single step in a linear fixed effect model. Although the ANOVA approach seems promising in theory, it is computationally expensive and becomes almost infeasible when the number of genes in the model gets relatively large. Alternatively, a two-step procedure was proposed by Wu *et al.* (2003). However, some of the factors in the model can be better modelled as random effects, such as the variation introduced through the use of multiple membranes for hybridisation. In general, the linear mixed model (LMM) methodology provides both a formal framework and a flexible tool for identifying systematic sources of variation and differential gene expression.

An extension of the ANOVA approach in terms of the mixed effect model is outlined by Wolfinger (Wu *et al.*, 2003), which has been performed on real data sets (Churchill and Oliver, 2001; Jin *et al.*, 2001). This method, using two interconnected sequential LMMs, is applied and extended to take into account the heterogeneity of variance on our data.

The objective of this paper is to state the feasibility of a transcriptomic protocol for the study of folliculogenesis in the pig. A pilot study including few animals but several technical replicates is analysed. The high sensitivity of nylon membranes hybridised with radioactively labelled probes (Bertucci *et al.*, 1999) is an ideal choice as follicular cells are difficult to collect. The first goal was to check the low level of technical variability with LMM analysis and the second was to highlight relevant genes that might be involved in the folliculogenesis process. For this, we selected genes via an LMM analysis and assessed their relevancy with multivariate statistical tools (principal component analysis (PCA) and unsupervised hierarchical clustering (UHC)). The predictive power of the selected genes was then assessed with a random forest (RF) analysis (Breiman, 2001) (http://www.stat.berkeley.edu/users/breiman/RandomForests/). The biological significance of these results is discussed.

Material and methods

Data

Biological model. Large White × Landrace sows were reared at the Unité Mixte de Recherche SENAH in Saint-Gilles (France) and treated with a progestagene used in cycling animals (Regumate^{TM}; Roussel-Uclaf, Romainville, France, during 18 d, 20 mg/d). Sows were ovariectomised 24 or 96 h after the end of the treatment. These stages correspond to the early and mid-follicular phase. Ovaries were dissected, follicles were extracted and granulosa cells were recovered as individual samples, in 200 μl of MEM121/F12 (v/v) medium (Gasser *et al.*, 1985). For each sample, 5 μl of the granulosa-cell suspension was smeared onto histological slides, fixed for 10 min in methanol–formaldehyde–acetic acid (80 : 15 : 5), and subsequently stained with Feulgen, as previously described (Besnard *et al.*, 1996). The quality of each follicle was assessed by microscopic examination of smears, using classical histological criteria as described by Monniaux (1987). Healthy follicles were classified as small (S), medium-sized (M) and large (L) according to their diameter (1–2, 3–3.5 and 5–6.5 mm, respectively) (Figure 1, adapted from Erickson *et al.* (1985) with kind permission).

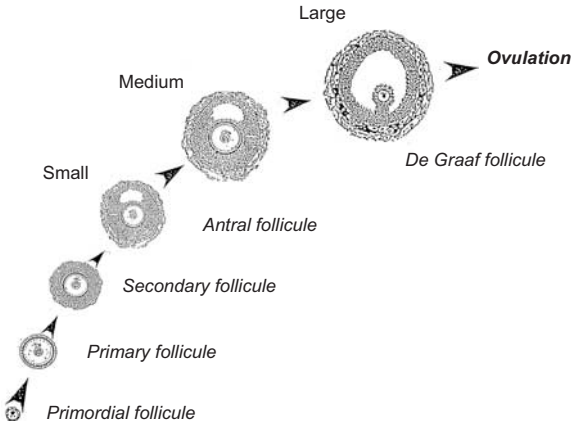

Figure 1 Scheme of folliculogenesis. Follicules grow from a primordial type to De Graaf type prior to ovulation. The follicules of the current analysis (small, medium-sized and large) correspond to Antral and De Graaf types.

Healthy follicles from the same class were pooled before RNA extraction. Two pools of small follicles (about 20 follicles), one pool of medium-sized follicles (about 10 follicles) and four pools of large follicles (five follicles each) were utilised for the microarray analysis. RNAs were extracted from these seven pools and their qualitative and quantitative qualities were checked, as described by Bonnet *et al.* (2006).

Microarray description. Spotting process, hybridisation conditions and image quantification have been detailed by Ferré *et al.* (2007). Briely, PCR products of cDNA inserts were spotted onto nylon membranes. Inserts came from 1152 clones from the AGENAE normalised multi-tissue cDNA library (Bonnet *et al.*, 2008a). Spikes (external genes) and negative controls were also spotted. Single or triplicate membranes were obtained by spotting each amplified insert once or three times, respectively. The GEO platform files for these chips are posted at: http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GPL3970

and http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GPL3971.

The amount of spotted DNA was estimated by hybridising the membranes with a vector probe. Complex probes were obtained by the reverse transcription and ^{33}P labelling of the extracted RNAs and were hybridised onto microarrays. Image analysis of complex and vector hybridisations were performed with BZscan software (Lopez *et al.*, 2004). The *Q* Im Cst values that correspond to constant diameter and conventional pixel integration were used.

Experimental design. Four large, one medium and two small follicle pools were considered. For each follicle pool, two radioactive labellings were performed. Each membrane was exposed 16 h (to avoid saturation of the signal of highly expressed genes) and 28 h (to get some signal from lowly expressed genes). Each probe was hybridised on a single and on a triplicate membrane (except for one named L2172, which was labelled only once and hybridised onto two single membranes), so that four spots were available for each gene (and two spots for L2172), for a given RNA and a given radioactive labelling. Finally, data consisted in (six RNA $\times$ two labellings) + (L2172 RNA $\times$ 1 labelling) = 13 probes, 26 hybridisations and 52 images. The data were posted at http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE5299.

Gene annotation. Clone sequences were annotated with the automatic procedure used in SIGENAE database (http://www.sigenae.org/). Then, Iccare software (Muller *et al.*, 2004) (http://bioinfo.genopole-toulouse.prd.fr/Iccare/), UniGene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene) and TIGR (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gireport.pl?gudb=pig) were used. In case of conflicting results, the annotation is stated as putative.

Methods to validate the biological experiment

Linear mixed model analysis

The data were analysed on the logarithmic scale using an LMM analysis. The pre-processing step that is usually performed before the statistical analysis to remove experimental bias is directly integrated into the LMM analysis.

An LMM is composed of two parts: fixed and random effects. The fixed effects considered are the type of follicle (two levels: L or M/S, as the number of samples in M and S is too small and M and S share biological similarities), the radioactive labelling (two levels), the spot type (five levels: replicates one to three of cDNA library and two groups of reference spots including blank DMSO), the hybridisation rank of the array (three levels), the exposure time (16 or 28 h) and two covariates, namely the logarithm of the background intensity and the logarithm of the vector probe intensity.

The random effects included microarray (14 microarrays), print-tip (32 print-tips by microarray), mRNA within type of follicle (three levels) and residual effects, and all are assumed to be normally distributed random variables with zero means and homogenous variance components σ^2_M, σ^2_P, σ^2_R and σ^2_ε, respectively. The gene effect is also assumed to be normally distributed with zero mean but with heterogeneous variances ($\sigma_{G,h}{}^2$), indexed by the type of follicle (h = L or M/S). All these random effects are assumed to be independent both across their indices (h) and between each other.

Inference was based on the maximum likelihood (ML) and on the restricted estimator maximum likelihood procedures (REML (Patterson and Thompson, 1971)) for the location and the dispersion parameters, respectively. Computations were made using the MIXED procedure of the Statistical Analysis Systems Institute (SAS) software (SAS, 1999).

The 'gene' factor was considered a random effect in the first analysis devoted to quantifying the gene variability, as opposed to technical variability. It was treated as a fixed effect in our second analysis where the aim was to infer differential genes. In this latter case, corrections for multiple testing were performed with the estimation of false discovery rate (FDR (Benjamini and Hochberg, 1995)). The genes with the lowest *P*-value, i.e. that are declared regulated, were kept for further analysis.

Multivariate analyses

We applied PCA and UHC to the above-mentioned selection of the most significant genes, in order to graphically summarise the results given by the LMM analysis. The aim of PCA is to reduce dimensionality, operating a projection of the data from a high-dimensional space (equal to the number of genes) to a lower dimensional one (usually two or three), making the data more accessible to visualisation and analysis (see Raychaudhuri *et al.* (2000) and Yeung and Ruzzo (2001) for more details).

UHC, gathering genes into a reduced number of groups, has been a widely used graphical tool in microarray analysis since Eisen *et al.* (1998). We chose the Euclidian distance in both genes and membranes with the Ward criterion (Seber, 1984).

To quantify the predictive value of the most significant genes from the LMM analysis, a RF analysis (Breiman, 2001) was performed. RF is a classification method that aggregates classification or regression trees to select discriminative (also called predictive) genes. Application of RF on microarray data

Table 1 *Estimates and standard errors of the variance components*

Effects	Variance components	Estimates	Standard errors	Ratio[1]
Gene in LF	$\sigma^2_{G,Large}$	0.66	0.03	
Gene in S/MF	$\sigma^2_{G,Medium/Small}$	0.85	0.04	0.76
Print-tip	σ^2_P	0.03	0.01	0.03
Membrane	σ^2_M	0.01	0.004	0.01
mRNA	σ^2_R	0.07	0.05	0.07
Residual	σ^2_ε	0.13	0.0006	0.13

LF = large follicles; S/MF = small/medium follicles.
[1]Ratio of variance estimates to the estimated total variance, i.e. proportion of variance dispatched in each source of variation.

Table 2 *List of the 29 genes selected in a Fisher test, with a $<10^{-8}$* P*-value*

Clone name	Automatic annotation	HUGO symbol	HUGO gene title	Genbank accession number	Fold change (L/MS)	*P*-value	RF
scag0006.h.12	RS8_HUMAN	RPS8	Ribosomal protein S8	BX666102	−0.63	7.47E−10	
scag0004.h.10	RL28_MOUSE	RPL28	Ribosomal protein L28	BX666563	−0.44	8.90E−08	
scag0010.h.04	scag0010c.h.04_5.1.ss.5	B4GALNT4	Beta-1,4-*N*-acetyl-galactosaminyl transferase 4	BX665057	−0.48	1.17E−09	
scag0010.h.03	X3752795.1.ss.5	*DAG1*	*Dystroglycan 1 (dystrophin-associated glycoprotein 1)*	BX666586	−0.46	5.29E−09	
scag0006.c.10	BTG2_MOUSE	BTG2	BTG family, member 2	BX666261	−0.76	3.82E−12	
scag0006.c.05	RS7_HUMAN	RPS7	Ribosomal protein S7	BX666257	−0.64	5.59E−11	
scag0010.b.01	S111_PIG	S100A11	S100 calcium-binding protein A11	BX665158	−0.54	1.85E−13	
scag0008.b.04	MCM7_HUMAN	MCM7	Minichromosome maintenance complex component 7	BX666389	−0.61	9.08E−13	
scai0001.g.11	VIIC2	RPLP0	Ribosomal protein, large, P0	X91728	−0.67	2.43E−15	*
scag0009.d.04	RS5_HUMAN.1	RPS5	Ribosomal protein S5	BX667032	−0.51	1.96E−15	*
scag0006.b.09	CDK4_PIG	CDK4	Cyclin-dependent kinase 4	BX665322	−0.29	3.11E−08	
scag0001.f.06	RS3_HUMAN.2	RPS3	Ribosomal protein S3	BX667478	−0.39	6.21E−08	
scag0008.d.12	RLA1_HUMAN.1	RPLP1	Ribosomal protein, large, P1	BX666418	−0.67	4.68E−11	
scag0008.d.11	TRI6_HUMAN	WTIP	Wilms tumor 1 interacting protein	BX664960	−0.58	4.39E−08	
scag0008.f.06	RLA1_HUMAN.2	RPLP1	Ribosomal protein, large, P1	BX666436	−0.64	3.93E−10	
scag0007.d.11	RLA1_HUMAN	RPLP1	Ribosomal protein, large, P1	BX666130	−0.72	2.47E−09	
scai0001.g.10	IGFBP3	IGFBP3	Insulin-like growth factor binding protein 3	BG608425	−0.73	1.76E−08	
scag0007.d.07	APEX_CAVPO	*NPTX2*	*Neuronal pentraxin II*	BX666127	−0.94	2.42E−20	*
scai0001.f.07	GPX	GPX3	Glutathione peroxidase 3 (plasma)	AJ783757	−0.92	1.28E−18	*
scai0001.f.05	calpaïn	*CAPNS1*	*Calpain, small subunit 1*	AJ704880	−1.10	9.47E−14	*
scai0001.f.10	Vimentine.1	VIM	Vimentin	AJ704901	−1.11	6.08E−12	
scai0001.e.03	Inhinbine	INHBA	nhibin, beta A (activin A, activin AB alpha polypeptide)		0.70	1.66E−12	
scag0002.e.05	CH10_HUMAN	HSPE1	Heat shock 10 kDa protein 1 (chaperonin 10)	BX667345	0.59	3.19E−11	
scag0007.e.01	TRIC_BOVIN	TNNI3	Troponin I type 3 (cardiac)	BX666133	0.65	1.55E−09	
scag0001.g.02	VK04_VACCC	PLD3	Phospholipase D family, member 3	BX667483	0.45	4.61E−08	
scag0001.e.06	YE63_SCHPO	ABHD4	Abhydrolase domain containing 4	BX667472	0.59	4.31E−12	
scag0004.a.01	LDVR_HUMAN	VLDLR	Very low density lipoprotein receptor	BX665076	0.40	7.75E−09	
scag0007.f.12	X3846497.1.ss.5	*NR5A2*	*Nuclear receptor subfamily 5, group A, member 2*	BX665622	1.22	4.53E−17	*
scag0006.b.05	ID3_RAT	ID3	Inhibitor of DNA binding 3, dominant negative helix-loop-helix protein	BX666246	0.70	9.02E−10	

HUGO = human genome organisation; RF = random forest.

The clone name, automatic annotation and Genbank accession numbers are listed in columns 1, 2 and 5, respectively. The HUGO (http://www.gene.ucl.ac.uk/nomenclature/) gene symbol and HUGO gene title are given in columns 3 and 4, in italics when the annotation is putative. The fold change between large and medium + small follicles is given in column 6, with its *P*-value. In the last column, an asterisk is present if the clone was selected in the random forest analysis.

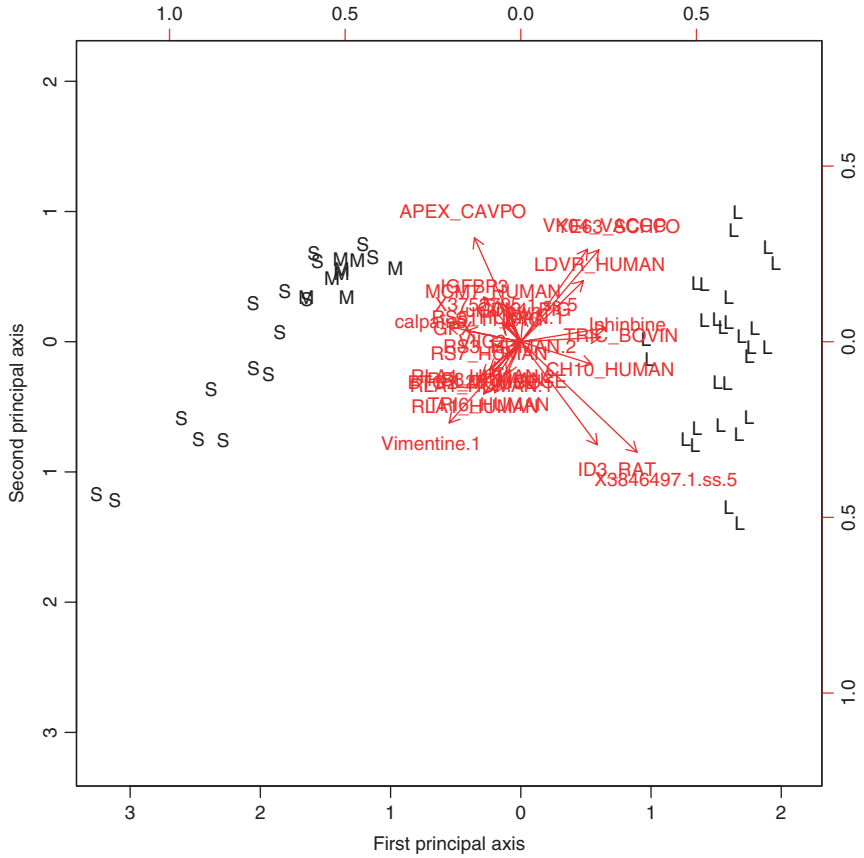

Figure 2 Biplot of the principal component analysis performed on the top-29 differentially expressed genes. The first two principal axes explain 68% and 8% of the total variability.

can be found in Diaz-Uriarte and Alvarez de Andres (2006) and Le Cao *et al.* (2007).

PCA, UHC and RF were performed using the R software (http://www.r-project.org/) (Liaw and Wiener, 2002).

Results

Test of fixed effects

All fixed effects included in the LMM were significant. The intensities obtained at the first hybridisation were globally larger than those obtained at the second. As expected, the intensities obtained at 16 h of exposure were generally lower than those at 24 h. The intensities of the spots increased both with the local background noise and with the intensity of the vector probes, and hence with the quantity of cDNA spotted onto the membrane. The 'type of follicle' effect is not significant, meaning that whatever the type of follicles, the genes are expressed at the same level on average. The 'hybridisation rank of the membrane' effect was significant, due to the significant difference in ranks 0 and 1 for the first hybridisation. The other differences in the rank of hybridisation of the membrane were not significant at the 5% level. The spot type also had a global significant effect due to the difference between the cDNA spots and the reference spots (mainly blank DMSO that have a lower average signal level). No significant difference between the three replicates of the cDNA library was observed.

Variance components

For each random variable, the values of the estimated variance components, the standard error and the ratio (ratio between the variance of the random variable considered and the total variance) are given in Table 1. We showed that 76% of the total variability was due to gene effect, 7% to mRNA extraction effect, 3% to print-tip effect, 1% to the membrane effect and only 13% of the variability remains unexplained by the model. The gene variances in L and S/M follicle $\sigma^2_{G,L}$ and $\sigma^2_{G,S/M}$ were significantly different from 0 and ${\sigma_{G,S/M}}^2$ (equal to 0.85) and was significantly higher than ${\sigma_{G,L}}^2$ (z-test, $P = 0.001$). The variances of print-tips, membranes and especially mRNA effects were very small with moderate standard errors. The low variability of the mRNA effect might be surprising but the number of extractions for each type of follicle (three for S/M follicles and four for L follicles) was low, which explains why the estimation of this variance component was not precise enough.

Differential expression

The 29 most differentially expressed clones are listed in Table 2, with a proposed annotation. These were assessed based on Fisher F tests specifying gene as a fixed effect in

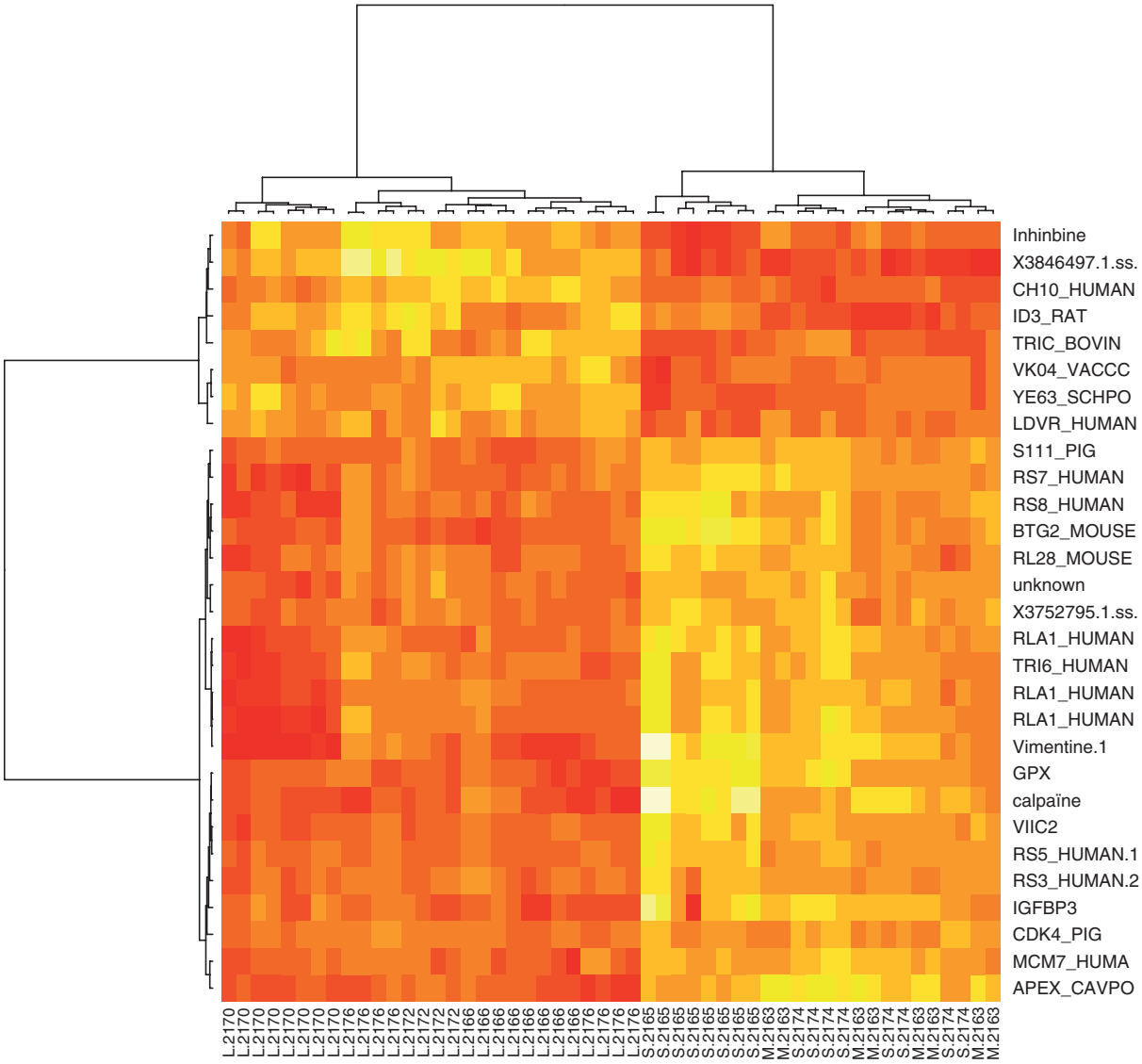

Figure 3 Heat map display of unsupervised hierarchical clustering results (Ward method and Euclidian distance) of the top-29 most significant genes. Genes are displayed in lines and membranes in columns. The light (respectively, dark) colour represents up- (respectively, down-) regulated genes.

the LMM, and a significance threshold of *P*-value $<10^{-8}$ corresponding to a FDR $<4\times10^{-6}$. This threshold was chosen because biologists are interested in a very small subset of genes that can be easily visualised with graphical and statistical tools. Among the 29 differentially expressed genes, 21 are down-regulated in L follicles, whereas eight are over-expressed. Terminal follicular growth (L follicle stage as opposed to S/M stages) is associated with decreased expression of genes implicated in protein translation (nine clones are subunits of ribosomal proteins: RPS8, RPL28, RPS7, RPLPO, RPS5, RPS3 and RPLP1), ion binding (S100A11) and cell shape (VIM, CAPNS1 and DAG1) and the over-expression of genes implicated in lipid metabolism and steroidogenesis (NR5A2, VLDL and INHBA).

Multivariate analyses

A PCA was performed on these 29 most significant genes. The scree graph of eigenvalues (not shown) showed a large drop between the first- and the second-ordered eigenvalues. Indeed, the first principal axis contains most of the total variance (68%) and the second only 8%. Figure 2 shows that the first axis actually separates L from S/M follicles. Genes such as CH10_HUMAN (HSPE1) or Inhibine (INHBA), represented as vectors pointing towards the L follicles, are all up-regulated in the L follicles. The calpaïne (CAPNS1), on the other hand, is up-regulated in the M and S follicles. The second axis in Figure 2 tends to separate a group of S follicles that are technical replicates of the follicle number S.2165, from another group with replicates M.2163 and S.2174. This suggests that these latter follicles have closer expressions for the 29 most significant genes than the two S follicles S.2165 and S.2174 themselves.

The heat map resulting from the hierarchical clustering of the top-29 genes is plotted in Figure 3. Two clusters appeared on the membranes, clearly separating L from S/M follicles. Replicates from the same follicle were clustered

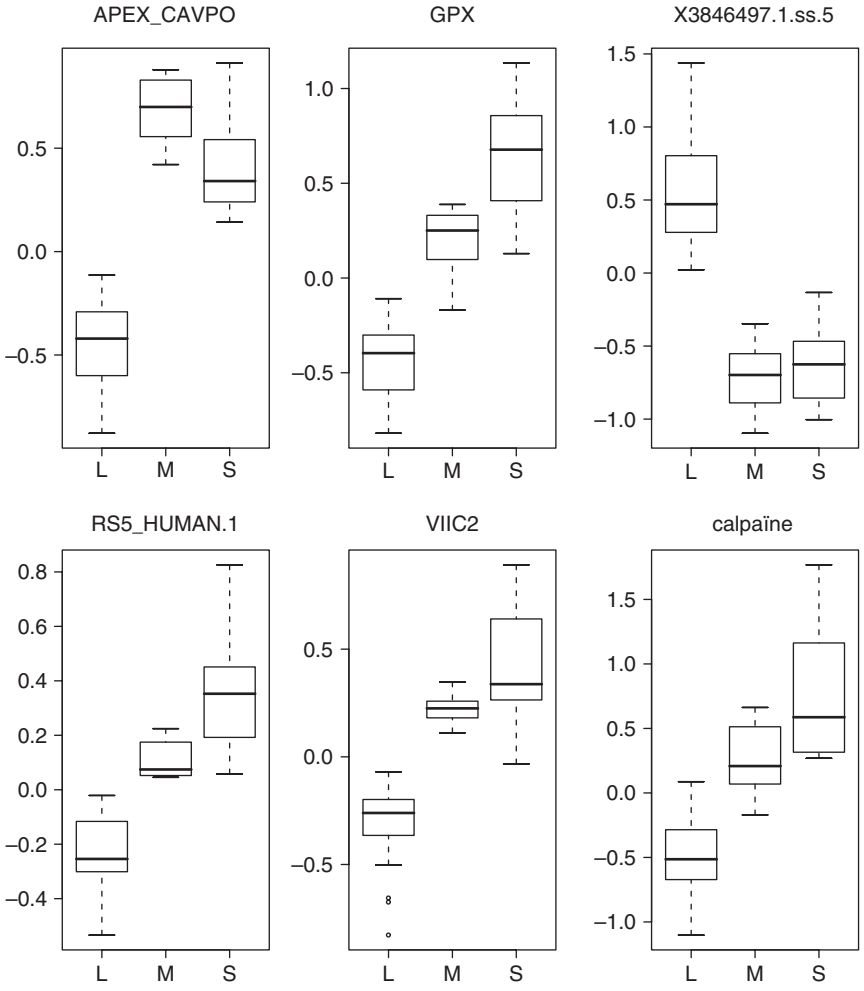

Figure 4 Boxplots of some of the most discriminant (or predictive) genes from the random forest analysis, also in the top-29 list.

together, except for S.2163 and L.2176. This shows that the technical replicates are similar compared to animal replicates or intrinsic gene variation. Note that the same conclusions were drawn with PCA. Two clusters on the gene clustering separated up-regulated from down-regulated genes on the L follicles.

Predictive value of the most differentially expressed genes
RF analysis does not require a fine-tuning of the parameters to achieve the best performance. We chose 5000 trees and a number of input variables tried at each split set to the default value in R (equal to $\sqrt{p}$, where p is the total number of genes). The estimate of the classification error rate was 7.69%. The 28 L follicles and the 16 S follicles were all classified without error in their respective classes. Only four out of eight M follicles were correctly classified.

The discrimination power of each gene in the forest is given by two importance measures proposed in the R package. Both measures were similar (not shown). The top genes hold useful information concerning their discrimination properties, as shown in Figure 4. APEX_CAVPO (NPTX2), VIIC2 (RPLPO), RS5_HUMAN.1 (RPS5), GPX (GPX3) and calpaïn (CAPNS1) are all down-regulated in L while the clone X3846497.1.ss.5 (NR5A2) is up-regulated in L. These boxplots also show that there is no clear discrimination that can be made between S and M. The regulation of these six genes has been investigated by real-time quantitative PCR and all the regulations were confirmed (Bonnet *et al.*, 2008b) and personal communication.

Discussion and conclusion

The experiment described in this paper was designed to quantify the repeatability of one microarray technique, namely nylon membranes of cDNA from a normalised multi-tissue library. Results showed that the variability due to the technical replicates is very low compared to the global variability. The LMM allowed to apprehend and to quantify the part of variability due to various sources of variations. The part of variability due to genetic diversity represents 7%, technique measures 4% and genes 76%.

Even if the variability between animals had a low estimate, it seems difficult to generalise this result because of the limited number of animals. A further study is being planned that involves more genes (a higher density microarray), more animals and fewer technical replicates. This

would allow a better confidence in the expressed genes, even if the results obtained in this pilot study are already satisfactory from a biological point of view.

It is difficult to separate S from M follicles, whereas L follicles differ greatly from S and M in their expression. The graphical displays (PCA and UHC) highlighted the particularity of this data set. This difference between S/M and L follicles can be correlated to the time of apparition of LH receptors (May and Schomberg, 1984), which is a major event in folliculogenesis.

The different types of statistical analyses corroborate several studies on the regulation of known genes during follicular growth. For example, inhibin alpha has already been described as up-regulated (Guthrie *et al.*, 1994; Garrett *et al.*, 2000) and the over-expression of genes implicated in lipid metabolism is consistent with the increased steroidogenic activity of granulosa cells during terminal follicular development. Moreover, a decrease in protein synthesis may be associated with a decreased cellular growth rate. This is in agreement with previous studies describing a decrease in the percentage of proliferating granulosa cells during the final stages of follicular development in pigs and other species (Hirshfield, 1986; Fricke *et al.*, 1996; Pisselet *et al.*, 2000). Hierarchical clustering of the genes screened in this study allows a global view of the regulation occurring during follicular growth and is the first step in a study using a tool such as a gene ontology analysis.

The RF analysis performed well, although the number of samples was very small. The internal importance measures were relevant for the biological aim. The selected genes were indeed discriminative for the three classes. RF underlined six discriminative genes, which are of significant interest. Among them, three had already been selected by other experiments on cellular models: VIIC2 had been found in a differential hybridisation screening between control and FSH-treated pig granulosa cells (Tosser-Klopp *et al.*, 1997), GPX and calpaïn had been found in suppression subtractive experiments between control and FSH-treated pig granulosa cells (Bonnet *et al.*, 2006) and may play a role during folliculogenesis. NR5A2 transcript has been shown to be up-regulated in bovine dominant follicles, compared to small follicles (Fayad *et al.*, 2004), which is consistent with our findings but has not been described yet in pig ovary. NPTX2 transcript regulation has never been described in granulosa cells. However, NPTX3, from the same family, has been shown in a knockout approach (Varani *et al.*, 2002) to play key roles in the ovulation process and NPTX2 has been shown to be down-regulated in the endometrium from women with endometriosis (Kao *et al.*, 2003). It may thus play a role in folliculogenesis. RPS5 has already been described as a down-regulated gene during the differentiation process of murine erythroleukemia cells (Vizirianakis *et al.*, 1999) and it may also be a marker of development of ovarian follicles. All these hypotheses need further biological investigations, as *in situ* hybridisation to detail the expression pattern of these six discriminant genes.

To conclude, we have validated the nylon microarray technique on our biological model. This will lead to more ambitious studies, including more genes and more samples.

Acknowledgements

We thank the Génopole de Toulouse-Midi Pyrénées for technical support, and especially Cécile Donnadieu-Tonon. We thank the SIGENAE team (www.sigenae.org) for the help in annotation and submitting data to the GEO database. We thank Hélène Quesnel and Hervé Demay for providing ovaries. We are grateful to Hervé Lefebvre, Claire Rogel-Gaillard and Hélène Bergès for their help in generating microarrays. We also thank an anonymous referee for his helpful suggestions. This work was in part funded by the ACI IMP BIO.

References

Agca C, Ries JE, Kolath SJ, Kim JH, Forrester LJ, Antoniou E, Whitworth KM, Mathialagan N, Springer GK, Prather RS and Lucy MC 2006. Luteinization of porcine preovulatory follicles leads to systematic changes in follicular gene expression. Reproduction 132, 133–145.

Algers B and Uvnas-Moberg K 2007. Maternal behavior in pigs. Hormones and Behavior 52, 78–85.

Benjamini Y and Hochberg Y 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Bertucci F, Bernard K, Loriod B, Chang YC, Granjeaud S, Birnbaum D, Nguyen C, Peck K and Jordan BR 1999. Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. Human Molecular Genetics 8, 1715–1722.

Besnard N, Pisselet C, Monniaux D, Locatelli A, Benne F, Gasser F, Hatey F and Monget P 1996. Expression of messenger ribonucleic acids of insulin-like growth factor binding protein-2, -4, and -5 in the ovine ovary: localization and changes during growth and atresia of antral follicles. Biology of Reproduction 55, 1356–1367.

Bichard M and David PJ 1985. Effectiveness of genetic selection for prolificacy in pigs. Journal of Reproduction and Fertility. Supplement 33, 127–138.

Bonnet A, Frappart PO, Dehais P, Tosser-Klopp G and Hatey F 2006. Identification of differential gene expression in *in vitro* FSH treated pig granulosa cells using suppression subtractive hybridization. Reproductive Biology and Endocrinology 4, 3510.1186/1477-7827-4-35.

Bonnet A, Iannuccelli E, Hugot K, Benne F, Bonaldo MF, Soares MB, Hatey F and Tosser-Klopp G 2008a. A pig multi-tissue normalised cDNA library: large-scale sequencing, cluster analysis and 9 K micro-array resource generation. BMC Genomics 9, 1710.1186/1471-2164-9-17.

Bonnet A, Le Cao KA, San Cristobal M, Benne F, Robert-Granié C, Law-So G, Fabre S, Besse P, De Billy E, Quesnel H, Hatey F and Tosser-Klopp G 2008b. *In vivo* gene expression in granulosa cells during pig terminal follicular development. Reproduction 136, 211–224.

Breiman L 2001. Random forests. Machine Learning 45, 5–32.

Buske B, Sternstein I and Brockmann G 2006. QTL and candidate genes for fecundity in sows. Animal Reproduction Science 95, 167–183.

Caetano AR, Johnson RK, Ford JJ and Pomp D 2004. Microarray profiling for differential gene expression in ovaries and ovarian follicles of pigs selected for increased ovulation rate. Genetics 168, 1529–1537.

Churchill GA and Oliver B 2001. Sex, flies and microarrays. Nature Genetics 29, 355–356.

Diaz-Uriarte R and Alvarez de Andres S 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 310.1186/1471-2105-7-3.

Duda M 1997. The influence of FSH, LH and testosterone on steroidsecretion by two subpopulations of porcine granulosa cells. Journal of Physiology and Pharmacology 48, 89–96.

Eisen MB, Spellman PT, Brown PO and Botstein D 1998. Cluster analysis and display of genome-wide expression patterns. The Proceedings of the

National Academy of Sciences of the United States of America 95, 14863–14868.

Erickson GF, Magoffin DA, Dyer CA and Hofeditz C 1985. The ovarian androgen producing cells: a review of structure/function relationships. Endocrine Reviews 6, 371–399.

Fayad T, Levesque V, Sirois J, Silversides DW and Lussier JG 2004. Gene expression profiling of differentially expressed genes in granulosa cells of bovine dominant follicles using suppression subtractive hybridization. Biology of Reproduction 70, 523–533.

Ferré PJ, Liaubet L, Concordet D, Sancristobal M, Uro-Coste E, Tosser-Klopp G, Bonnet A, Toutain PL, Hatey F and Lefebvre HP 2007. Longitudinal analysis of gene expression in porcine skeletal muscle after post-injection local injury. Pharmaceutical Research 24, 1480–1489.

Foxcroft GR and Hunter MG 1985. Basic physiology of follicular maturation in the pig. Journal of Reproduction and Fertility. Supplement 33, 1–19.

Foxcroft GR, Hunter MG and Grant SA 1989. The physiology of follicular maturation in the pig. Acta Physiologica Polonica 40, 53–63.

Fricke PM, Ford JJ, Reynolds LP and Redmer DA 1996. Growth and cellular proliferation of antral follicles throughout the follicular phase of the estrous cycle in Meishan gilts. Biology of Reproduction 54, 879–887.

Garrett WM, Mack SO, Rohan RM and Guthrie HD 2000. *In situ* analysis of the changes in expression of ovarian inhibin subunit mRNAs during follicle recruitment after ovulation in pigs. Journal of Reproduction and Fertility 118, 235–242.

Gasser F, Mulsant P and Gillois M 1985. Long-term multiplication of the Chinese hamster ovary (CHO) cell line in a serum-free medium. In Vitro Cellular and Developmental Biology 21, 588–592.

Guthrie HD, Barber JA, Leighton JK and Hammond JM 1994. Steroidogenic cytochrome P450 enzyme messenger ribonucleic acids and follicular fluid steroids in individual follicles during preovulatory maturation in the pig. Biology of Reproduction 51, 465–471.

Hirshfield AN 1986. Patterns of [3H] thymidine incorporation differ in immature rats and mature, cycling rats. Biology of Reproduction 34, 229–235.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G and Gibson G 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nature Genetics 29, 389–395.

Kao LC, Germeyer A, Tulac S, Lobo S, Yang JP, Taylor RN, Osteen K, Lessey BA and Giudice LC 2003. Expression profiling of endometrium from women with endometriosis reveals candidate genes for disease-based implantation failure and infertility. Endocrinology 144, 2870–2881.

Kerr MK and Churchill GA 2001. Experimental design for gene expression microarrays. Biostatistics 2, 183–201.

Knox RV 2005. Recruitment and selection of ovarian follicles for determination of ovulation rate in the pig. Domestic Animal Endocrinology 29, 385–397.

Le Cao KA, Goncalves O, Besse P and Gadat S 2007. Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology 6, Article 29.

Liaw A and Wiener M 2002. Classification and regression by random Forest. The Newspaper of R Project 2, 18–22.

Lopez F, Rougemont J, Loriod B, Bourgeois A, Loi L, Bertucci F, Hingamp P, Houlgatte R and Granjeaud S 2004. Feature extraction and signal processing for nylon DNA microarrays. BMC Genomics 5, 38.

May JV and Schomberg DW 1984. Developmental coordination of luteinizing hormone/human chorionic gonadotropin (hCG) receptors and acute hCG responsiveness in cultured and freshly harvested porcine granulosa cells. Endocrinology 114, 153–163.

Monniaux D 1987. Short-term effects of FSH *in vitro* on granulosa cells of individual sheep follicles. Journal of Reproduction and Fertility 79, 505–515.

Muller C, Denis M, Gentzbittel L and Faraut T 2004. The Iccare web server: an attempt to merge sequence and mapping information for plant and animal species. Nucleic Acids Research 32, W429–W434.

Patterson HD and Thompson R 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.

Pisselet C, Clement F and Monniaux D 2000. Fraction of proliferating cells in granulosa during terminal follicular development in high and low prolific sheep breeds. Reproduction, Nutrition, Development 40, 295–304.

Raychaudhuri S, Stuart JM and Altman RB 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pacific Symposium on Biocomputing 5, 452–463.

Seber G 1984. Multivariate observations. Wiley, New York, NY.

Statistical Analysis Systems Institute 1999. SAS/STAT Software, version 8. SAS Institute Inc., Cary, NC.

Tosser-Klopp G, Benne F, Bonnet A, Mulsant P, Gasser F and Hatey F 1997. A first catalog of genes involved in pig ovarian follicular differentiation. Mammalian Genome 8, 250–254.

van der Lende T, Knol EF and Leenhouwers JI 2001. Prenatal development as a predisposing factor for perinatal losses in pigs. Reproduction 58(suppl.), 247–261.

Varani S, Elvin JA, Yan C, DeMayo J, DeMayo FJ, Horton HF, Byrne MC and Matzuk MM 2002. Knockout of pentraxin 3, a downstream target of growth differentiation factor-9, causes female subfertility. Molecular Endocrinology 16, 1154–1167.

Vizirianakis IS, Pappas IS, Gougoumas D and Tsiftsoglou AS 1999. Expression of ribosomal protein S5 cloned gene during differentiation and apoptosis in murine erythroleukemia (MEL) cells. Oncology Research 11, 409–419.

Whitworth KM, Agca C, Kim JG, Patel RV, Springer GK, Bivens N, Forrester LJ, Mathialagan N, Green JA and Prather RS 2005. Transcriptional profiling of pig embryogenesis by using a 15k member unigene set specific for pig reproductive tissues and embryos. Biology of Reproduction 72, 1437–1451.

Wu H, Kerr MK, Cui X and Churchill GA 2003. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. In The analysis of gene expression data: methods and software (ed. G Parmigiani, ES Garrett, RA Irizarry and SL Zeger). Springer, New York, NY.

Yeung KY and Ruzzo WL 2001. Principal component analysis for clustering gene expression data. Bioinformatics 17, 763–774.