Gene Expression Signature of Menstrual Cyclic Phase in Normal Cycling Endometrium

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Ling Cen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Adviser: George Vasmatzis
Co-Adviser: Claudia Neuhauser

December 2014

Acknowledgements

First, I would like to thank my advisor, Dr. George Vasmatzis for giving me an opportunity to work on this project in his group. This work could not have been completed without his kind guidance. I also would like to thank my co-advisor, Dr. Claudia Neuhauser for her advising and guidance during the study in the program. Further, I also very much appreciate Dr. Peter Li for serving on my committee and for his helpful discussions. Additionally, I'm grateful to Dr. Y.F. Wong for his generosity in sharing the gene expression data for our analysis. Most importantly, I am deeply indebted to my family and friends, for their continuous love, support, patience, and encouragement during the completion of this work.

Abstract

Gene expression profiling has been widely used in understanding global gene expression alterations in endometrial cancer vs. normal cells. In many microarray-based endometrial cancer studies, comparisons of cancer with normal cells were generally made using heterogeneous samples in terms of menstrual cycle phases, or status of hormonal therapies, etc, which may confound the search for differentially expressed genes playing roles in the progression of endometrial cancer. These studies will consequently fail to uncover genes that are important in endometrial cancer biology. Thus it is fundamentally important to identify a gene signature for discriminating normal endometrial cyclic phases. To this end, gene expression analysis was performed on 29 normal endometrium specimens. Unsupervised analysis demonstrated that gene expression profiles common to secretory endometrium were distinctively different from those of proliferative and atrophic endometrium. Pairwise comparisons further revealed no significant difference in gene expression between proliferative and atrophic endometrium. In addition, using a normal mixture model-based clustering algorithm we were able to identify a gene signature consisting of 35 unique annotated genes that display a switch-like or bimodal expression pattern across all samples. Functional annotation of this gene signature revealed that complement and coagulation cascades and Wnt signaling pathway were significantly enriched. Utility of this gene signature was validated in an independent gene expression data set, where clustered proliferative samples from clustered early, mid, and late-secretory samples were successfully separated. These data suggest that the bimodal gene signature identified in this study could potentially be used to distinguish cyclic phases of the menstrual cycle. Our findings will facilitate future work in understanding the molecular characteristics of endometrial cancers in comparison to normal endometrium.

Abbreviations

A	Atrophic (endometrium)
BI	Bimodality index
cDNA	Complementary DNA
cRNA	Complementary RNA
DAVID	Database for annotation, visualization, and integrated discovery
EST	Expressed sequence tag
GEO	Gene Expression Omnibus
KEGG	Kyoto encyclopedia of genes and genomes
LCM	Laser capture microdissection
LIMMA	Linear models for microarray analysis
MAX	Complete-linkage
MIN	Single-linkage
MM	Mismatch (probe)
MDS	Multidimensional scaling
NCBI	National Center for Biotechnology Institute
P	Proliferative (endometrium)
PM	Perfect match (probe)
PCA	Principal Component Analysis
RMA	Robust Multichip Average
RNA	Ribonucleic acid
S	Secretory (endometrium)

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Problem statement

Gene expression profiling has been widely used in studying global gene expression alterations to characterize cancer vs. normal cells, including endometrial cancer, for the purpose of identifying cancer subtypes, discovering new drug targets and developing novel therapeutic strategies. Normal endometrium is a highly dynamic tissue and undergoes profound histological and structural cyclic changes during menstrual cycle every month, which are ultimately the result of changes in gene expression affected by levels of ovarian steroids. In many microarray-based gene expression studies of endometrial cancers, comparisons of cancer cells with normal cells were generally made using heterogeneous samples in terms of menstrual cycle phases or status of hormonal therapies, etc. Therefore, this may confound the search for differentially expressed genes that may play important roles in the progression of endometrial cancer. These studies will consequently fail to uncover genes that are important in endometrial cancer biology. Thus, it is fundamentally important to identify the gene signature for discriminating endometrial cyclic phases. To this end, we compared the whole-genome expression profiles of tens of human normal endometrium. Our aim is to identify a gene signature with discriminative power to separate endometrial tissue samples into subgroups with respect to their menstrual cycle phases. To demonstrate the gene signature could be used as menstrual cycle phase markers, we performed 2-way

hierarchical clustering analysis on a gene expression data set obtained from public database using the gene signature.

Structure of the thesis

Following the introduction in this chapter, Chapter 2 describes the data sets used in the thesis in details. Chapter 3 describes the methods used in this thesis and details about how they were implemented. Chapter 4 presents the results of our analysis. Finally, Chapter 5 summarizes and discusses our findings.

The endometrial cycle

Cyclic change of endometrium is tightly regulated by ovarian steroid hormones, estrogen and progesterone. Hormonal control of endometrium including endometrial, epithelial, and stromal cells is mediated by estrogen and progesterone receptors, which are proteins located in the nuclei of those cells, through high affinity binding to the hormones [1]. In a typical 28-day cycle, following menstrual shedding (day 1-5) the endometrium regenerates under the stimulation of estrogen, and endometrial thickness increases dramatically as a result of active growth of glands, stromal cells, and blood vessels, which is described as proliferative phase (day 6-14, Figure 1.1). During proliferative phase, the glands become longer, larger, and more coiled following multiple mitoses, which are often seen at higher magnification. The stroma becomes highly vascularized. In addition to tissue proliferation, estrogen is believed to promote the production of estrogen receptors and progesterone receptors. As a result, concentrations of estrogen receptors and progesterone receptors are elevated in both blood and tissue during the proliferative phase. When increased secretion of progesterone inhibits the endometrium proliferation, the secretory transformation initiates under the overall control of estrogen and progesterone, which is known as secretory phase (day 15-28, Figure 1.1). During this phase, the secretory activity is featured by a diversity of structural changes, displaying a different pattern on every day of the cycle.

During the first four days of the secretory phase, infrequent mitoses can still be seen in the glandular epithelium. Endometrial stromal cells show less edema. The glands are not yet in the state of active extracellular secretion. During day 5 and 6 of secretory phase, the newly synthesized intracellular products are secreted into the glandular space. It is characterized by protrusions. The mid and late secretory phase is completely absence of mitoses in the glands. The sharp reduction in the level of estrogen and progesterone leads to the shedding of the endometrium, a phase termed as menstrual phase (Figure 1.1) [2].

The morphological changes described above are often used as characteristics by pathologist to dating endometrial biopsy and categorize them into different cyclic phases. However, there is very low inter-observer agreement, and the histologic endometrial dating is considered inaccurate to guide clinical decision and diagnosis [3]. In fact, the profound changes in physiology, biology, and histology are ultimately the result of altered gene transcription and gene expression pattern, which is believed to precede visually identifiable morphological changes. Therefore, molecular markers will probably be more sensitive than histological characteristics in dating endometrial biopsy and distinguishing cyclic phases.

Microarray and gene expression data

Nowadays the most popular techniques to measure gene expression on a global scale include RNA-sequencing and microarray. While RNA-sequencing is a powerful technique, microarray is still popular because it is cheap, requires fewer resources, and has more mature methods for data analysis. Additionally, tens of thousands of data sets generated by microarray-based experiments are publicly available in databases, such as Gene Expression Omnibus (GEO) Database at National Center for Biotechnology Institute (NCBI).

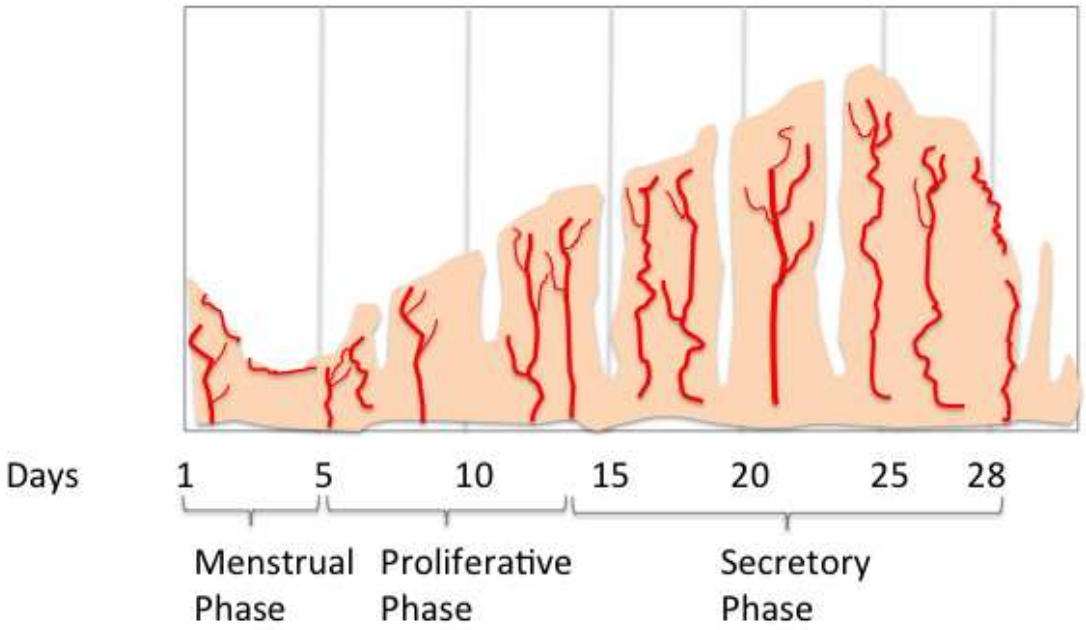

Figure 1.1. The progression of menstrual cycle and cyclic phases

Microarray is a high-throughput technology for measuring the expression levels of tens of thousands of genes simultaneously by hybridization of complementary DNA (cDNA) to a collection of oligonucleotide probes, which are attached to a microscope-sized glass slide [4, 5]. The use of microarrays allows for parallel quantification of gene expression on a global scale in cells and tissues of distinct phenotypes. The application of microarrays facilitates better understanding and classification of diseases as well as identification of novel therapeutic targets from studying basic biology, clinical diagnostics to drug discovery. To be specific, researchers often run microarray experiments to measure the expression level of each known transcript in a set of samples under investigation, or to explore what genes are activated in cells and at what level after stimulation; or to compare the gene expression profiles of treatment versus control groups (e.g. cancer vs. normal); or to identify changes in gene expression under specific conditions.

Two microarray data sets to be analyzed in this thesis are both generated with Affymetrix GeneChip Human Genome U133 Plus 2.0 platform, which is made up with more than 54,000 probe sets, each of which consists of 11-20 probes (25 oligonucleotides long) corresponding to a particular gene or EST (expressed sequence tag). There are two types of probes, perfect match (PM) and mismatch (MM). PM probes are perfectly complementary to a specific region of a gene. MM probes are identical to the PM probes except for the middle (13th) nucleotide in the sequence, which is replaced with its complementary nucleotide. MM probes are informative during computational data analysis to account for non-specific probe binding. Different types of probe sets can be inferred from suffices to the probe set name [6]. Probe sets without suffix are predicted to match a single transcript perfectly; those with "_a" suffix recognize multiple transcript variants from the same gene. Common probe sets with "_s" suffix

recognize multiple transcripts from different genes. Probe sets with "_x" suffix contain some probes that are identical or highly similar to other sequences.

Bimodal expression and biomarker discovery

Genome-wide gene expression analysis with microarrays has been widely used for identifying genes that are differentially expressed between subgroups of clinical samples. A frequent goal in translational research is to look for a gene or gene signature with the discriminative power to separate patients into subgroups with respect to physiological states, prognosis or drug response. Compared to a unimodal distribution, genes with bimodal distributions or switch-like behavior possess a clear advantage to classify patients into high and low expression subgroups. A bimodal distribution is a continuous probability distribution with two different modes (Figure 1.2). The observation and utility of bimodal genes have been described in numerous studies [7, 8]. Therefore, bimodal genes are promising candidates for biomarkers of disease outcome or phenotype.

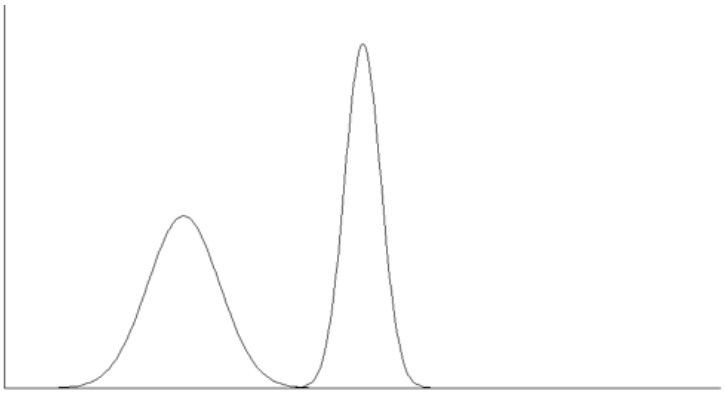

Figure 1.2. An example of bimodal distribution

Chapter 2

Description of the gene expression data sets

In this thesis, two sets of gene expression profiling array data were used: Wong *et al.* [9] and Talbi *et al.* [10]. Both data sets underwent histopathological evaluation for cyclic phase. And both were generated using *Affymetrix GeneChip Human Genome U133 Plus 2.0* platform.

The Wong *et al.* data set consists of 84 clinical samples from Hong Kong Chinese women. Among them, 55 are human microdissected sporadic endometriod endometrial adenocarcinomas and 29 are microdissected normal endometrium specimens. For the scope of this study, the subset of 29 normal endometrium specimens was extracted and used for our analysis. The endometrial cyclic phase of each normal endometrium specimen was determined by histological typing. Among the 29 normal specimens, there are 10 proliferative, 10 secretory, and 9 atrophic (postmenopausal) specimens, respectively. Patient sample IDs and their corresponding cyclic phase are listed in Table 2.1 (information of patient age is unavailable). Thus, these 29 arrays were used as an exploration data set. Each array contains 54675 probe sets corresponding to more than 38500 well-characterized human genes. The extracted data matrix has 54675 rows and 29 columns, where each row represents a probe set and each column is a patient specimen. This expression data set was normalized and log2 transformed.

The Talbi S *et al.* data set consists of 27 samples from normally cycling women undergoing hysterectomy or endometrial biopsy [10]. This data set was used to examine the utility of the gene signature identified from the *Wong et al.* data set. The cyclic phase

of these specimens was assigned through pathological review according to the criteria of Noyes *et al* [2]. Of them, four are proliferative, six ambiguous, and samples of secretory phase are sub-categorized into three early, eight mid, and six late secretory phase. The expression data set and clinical information were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus data repository under the accession number GSE4888 [11]. The downloaded expression data set was normalized and log2 transformed.

Table 2.1. Patient sample ID and corresponding cyclic phase

NO.	SAMPLE ID	CYCLIC PHASE
1	187	P
2	188	P
3	212	S
4	214	S
5	215	S
6	163	S
7	165	A
8	168	A
9	169	A
10	170	S
11	171	A
12	206	P
13	160	P
14	161	P
15	211	S
16	184	A
17	162	P
18	166	P
19	172	A
20	175	S
21	176	S
22	177	A
23	178	P
24	179	A
25	181	S
26	182	A
27	183	P
28	185	P
29	189	S

P: proflierative phase

S: secretary phase

A: atrophic phase

Chapter 3

Methods

Data preprocessing

In large-scale microarray experiments, it is a common problem for the presence of noise, which can originate from various sources. Noise from non-specific binding of cRNA fragments to probes is one such source. In Affymetrix GeneChip platform, each Perfect Match probe (PM) is matched to a second probe, Mis-Match probe (MM). MM probes are identical to the PM probes except for the middle (13th) nucleotide in the sequence, which is replaced with its complementary nucleotide. Subtracting the signal for the MM probe from that for the PM probe would show the true signal value. Various preprocessing algorithms have been developed to deal with these artifacts [12].

Among others, GCRMA [13, 14] is a popular preprocessing method of converting *CEL* files directly into expression set using the Robust Multichip Average (RMA) method [15] in combined with additional considering of probe sequence information and GC-content to compute probe affinity to adjust for background correction. Thus, GCRMA is considered as an improved version of RMA. While using the same normalization and expression value summarization steps as RMA, GCRMA uses probe sequence information to adjust for background intensities raised from non-specific binding in Affymetrix data. Usually, the preprocessing includes three steps, background correction, normalization, and summarization.

For background correction step, GCRMA is designed to remove background noise as well as non-specific binding and separate the specific signal from the non-specific signal. GCRMA incorporates probe sequence information to estimate probe affinity to non-specific binding and computes the Affymetrix PM and MM probe affinities from their sequences and MM probe intensities. Specifically, background correction consists of three steps: a) Optical background correction on the PM and MM intensity values; b) Probe intensity adjustment through non-specific binding using affinity information and optical-noise adjusted MM intensities; c) Gene-specific binding correction using probe affinity data.

Normalization is necessary to remove non-biological variations between multiple arrays used in the same experiment, so that multiple arrays can be compared to each other and analyzed together. The normalization used in GCRMA is quantile normalization, which is applied to background-corrected PM probe values.

Once been background-corrected and normalized the probe-level PM values need to be summarized into a single expression measure. This step generates a single expression value for each probe set corresponding to each gene per chip, and ultimately creates an expression matrix to summarize the Affymetrix microarray probe-level data. An expression matrix contains log2 expression values where each row corresponds to a probe set and each column corresponds to an Affymetrix data *CEL* file generated from a single chip.

For data sets used in this thesis, the *CEL* files containing the intensities determined for every oligonucleotide probe on a GeneChip were preprocessed in R using Bioconductor packages GCRMA (version 2.34.0) for background correction, normalization and data summarization. *justGCRMA* function was applied to convert *CEL* files directly into an R expression set object, which was further transformed into expression measures of log base 2 scale for each probe set.

Multidimensional scaling

Multidimensional scaling (MDS) is a useful dimensionality reduction technique that allows researchers to uncover the underlying structure of high-dimensional data sets, such as microarrays. In comparing pairs of objects, MDS helps visualize the level of proximity between individual objects. MDS takes proximities among objects as an input. Proximity is a measurement of the similarity or dissimilarity of a pair of objects. If all pairs of objects were measured in a set, the proximities are represented by a proximity matrix. The output of MDS is a lower-dimensional spatial representation points in a plot. Each point corresponds to one of the objects. The further apart the points in the graph are, the smaller the similarity between the pair of objects are [16]. To represent the distances among the normal endometrial samples in Wong *et al.* data set, a distance matrix was first computed and then MDS was applied to the distance matrix using a function in the R stats package. Finally, the result was plotted for visualization.

Principal components analysis

Principal component analysis (PCA) is a data exploration tool that can reduce the dimensionality of data. It finds a linear projection of high dimensional data into a lower dimensional subspace [17]. PCA transforms a set of correlated variables into uncorrelated variables by decomposing the original data into a set of mutually orthogonal eigenvectors and their weights. Two main methods for performing principal component analysis are eigenvalue decomposition and singular value decomposition. For a given data set stored as a column matrix A, where each column corresponds to a multi-dimensional variable of dimension p and the number of variables is n, the size of the matrix A is $n \times p$. For eigenvalue decomposition, we try to decompose matrix $A^T A$ into:

$$A^T A = W \Sigma W^T$$

, where W is a $p \times p$ matrix whose columns are the eigenvectors or principal components. Σ is a diagonal matrix whose elements on its diagonal represents the importance of each corresponding principal component in the original data. Usually, the columns of W and the diagonals of Σ are ordered in the descending order based on their importance. An alternative and much more popular way of carrying out principal component analysis is to use singular value decomposition. In this method, we try to find the following factorization:

$$A = U\Lambda W^T$$

, where U is a column matrix of size $n \times n$. The column vectors are orthogonal to each other and sometimes called left singular vectors. W is the same matrix W as in eigenvalue decomposition and sometimes also called right singular vectors. Matrix Λ has the same shaped of matrix Σ and its elements on the diagonal is the square roots of those of Σ. PCA is more widely calculated by SVD than eigenvalue decomposition because of its availability, efficiency and applicability [18, 19]. R stats package was used to perform the PCA computation. The first two principal components were plotted for visualization.

Hierarchical Clustering

Clustering is the process to sort different objects into groups in a way that objects in the same group are more similar to each other than to those in other groups. In gene expression data analysis, a common question facing researchers is how to organize the observed data into a meaningful structure. A number of clustering methods have been applied to identify the patterns of gene expression data set. Clustering can be either supervised or unsupervised. Supervised methods use known biological information about specific genes to be functionally related to guide the clustering algorithm. However, most widely used methods are unsupervised, and these methods are usually applied

first to explore structures in the data without any prior knowledge. Although clustering is powerful in data analysis, great caution should be taken in applying these techniques as different methods may place different objects in different clusters when different algorithms or distance metrics were selected subjectively.

Hierarchical clustering is one of the most widely used techniques for the analysis of gene expression data to group experimental samples (usually clinical specimens or cell lines) based on the similarity of their gene expression patterns. The same clustering method could be used to group genes based on the similarity in the pattern with which their expression varied over all the samples [20]. Hierarchical algorithms can be further classified into two categories, agglomerative and divisive. Agglomerative approach starts with all objects as individual clusters, at each iteration merge the most similar pair of clusters. The iteration continues until all objects are in a single cluster. The definition of cluster similarity is discussed in the following section. Divisive approach begins with one cluster including all objects and at each iteration cluster is gradually broken down until only singleton clusters of each individual object remain.

Agglomerative approach is more commonly used thus will only be described in detail here. The procedure of performing agglomerative hierarchical clustering involved several steps. Before any clustering, distance between each pair of objects is computed, and a distance matrix is constructed. During the clustering, initially each object is assigned to its own cluster. Then the algorithm starts iterating by searching the distance matrix for the two most similar objects or clusters, and merges them to produce a new cluster. At each iteration, distances between the new cluster and all other clusters are recomputed. The iteration continues until all objects are in one cluster [21].

A number of clustering methods of hierarchical clustering are available and each of them aims at finding clusters with different characteristics. Single-linkage (MIN) method is also known as minimum or nearest-neighbor method. This technique defines the

distance of two clusters as the minimum of the distance (maximum of the similarity) between members of one cluster and members of another cluster. Complete-linkage (MAX) method is also referred to as maximum or furthest-neighbor method. In complete-linkage method, the distance of two clusters is defined as the maximum of the distance (minimum of the similarity) between members of one cluster and members of another cluster. Thus it is in general less sensitive to noise and outliers compared to other methods. Average-linkage method aims for finding clusters with characteristics somewhere between single-linkage and complete linkage methods. It computes the distance of two clusters using the average values of the distance between all members of one cluster and all members of another cluster. Ward's method assumes that a cluster is represented by its centroid. Cluster membership is assigned by calculating the total variance from the mean of a cluster and joining clusters in such a way that it minimizes possible increase in variance. Thus, this method aims for finding compact and spherical clusters [22].

The key step of the above algorithm is the computation of the distance or similarity between two objects to construct distance or similarity matrix. The way in which distance or similarity measure is carried out between each pair of objects will produce slightly different result in clustering. There are several different definitions of distance measure, which differentiates the various hierarchical clustering techniques. Among others, Euclidean distance is the most commonly used type of distance measure. The distance between object *x* and *y* is defined as

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

, where n is the number of dimensions, and x_i and y_i are the i^{th} components of x and y, respectively. As in the case of maximum distance between two components of x and y is defined as

$$d(x, y) = \max(|x_i - y_i|)$$

In addition, Manhattan distance defines the distance between two objects as the sum of the absolute differences of their Cartesian coordinates, formally as

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

, where n is the number of dimensions, and x_i and y_i are the i^{th} components of x and y, respectively. Other methods for distance measure, such as Minkowski and Canberra are described in more details elsewhere [23, 24]. As for similarity measure, a number of methods are commonly used in microarray gene expression analysis including Pearson's and Spearman's correlation, etc. Pearson's correlation coefficient between two objects is defined as

$$\mathrm{r} = \frac{1}{n-1} \sum_{i=0}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right) \left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

, where $\bar{x}$ and $\bar{y}$ are the sample mean of x and y, respectively, and σ_x and σ_y are the sample standard deviation of x and y, respectively. Pearson's distance for two objects x and y can be defined from their correlation coefficient as

$$d_{x,y} = 1 - \mathrm{r}$$

The Spearman's rank correlation is a nonparametric similarity measure. In calculating the Spearman's rank correlation coefficient, each value in the data matrix is replaced by their rank ordered by their value in each vector. Then Pearson's correlation coefficient is calculated between two rank vectors instead of value vectors [25].

In this work, R stats package allows the computation of distance matrices based on the methods discussed above. For hierarchical clustering different methods for clustering were experimented. The results of hierarchical clustering were visualized graphically as a dendrogram tree and compared. Two-way hierarchical clustering was also conducted for both rows (probe sets) and columns (samples or arrays) of expression data matrices. The results were visualized in a heat map using R package gplots.

Differential gene expression analysis

Analysis of differential gene expression was carried out using LIMMA (linear models for microarray analysis), a software package available from Bioconductor. LIMMA is designed to analyze both simple and complex experiments involving comparisons between many RNA targets at the same time. Essentially it utilizes the expression data for each gene to fit a linear model. Empirical Bayes approach is then used to gather information across genes of different samples (arrays). Summary statistics were computed to describe differences in gene expression among samples [26, 27].

Differentially expressed gene lists generated from each pairwise comparison include only the probe sets that had a fold change value of 2.0 or greater (log2 fold change greater than 1.0). Differentially expressed genes between each pairwise comparison were determined separately with moderated t-test with an adjusted P- value of less than 0.05 by Benjamin-Hochberg multiple hypothesis testing for correcting for false positives in these probe sets. For annotation, all probe sets were first mapped to ENTREZ gene IDs Bioconductor annotation data packages hgu133plus2.db, then the duplicates were determined. Probe sets that mapped to more than one gene were annotated manually.

Definition of bimodality index

In the literature, a number of methods for identification of genes with bimodal distribution have been described [8, 28-30]. Some of these approaches are based on clustering the expression of a gene into two groups and formulating scores for bimodality

from the clustering result. In this work, the bimodality index described by Wang *et al.* [29] was used as criteria to identify bimodal genes in the endometrial gene expression data. It was assumed that for a given gene with bimodal expression pattern, the distribution can be expressed as a mixture of two normal distributions [29]

$$y = \pi N(\mu_1, \sigma) + (1-\pi) N(\mu_2, \sigma)$$

, where y is the normalized expression value, π is the proportion of samples in one component, parameters μ_1 and μ_2 are the means of the expression measurements of the two components, and σ is the assumed common standard deviation. The standardized distance between the two components is defined as [29]

$$\delta = \frac{|\mu 1 - \mu 2|}{\sigma}$$

The bimodality index (BI) is defined as [29]

$$BI = [\pi(1 - \pi)]^{1/2} \cdot \delta$$

In order to accurately estimate the parameters π and δ in the above equation for a given data set, a normal mixture model-based clustering algorithm implemented in R package MCLUST was used [31]. Once the value of π and δ are estimated, the bimodality index can be computed and ranked to identify relevant bimodal gene expression patterns. The value of bimodality index is larger if the two components are balanced in size or if the separation between the two model is larger and easier to be distinguished. Theoretically, bimodality index value of 1.1 or greater is suggested as 'useful' bimodal pattern of expression [29]. In this work, R ClassDiscovery package was used to compute component means μ_1 and μ_2, standard deviation σ, standardized distance δ, π and bimodality index BI for each probe set. Probe sets were then ranked based on value of bimodality index. Distributions of the expression of selected probe sets were estimated using the kernel density function in R.

Chapter 4

Results

In this chapter, experimental analysis of the endometrial gene expression data sets and results are presented. The Wong *et al.* data set consists of 29 endometrial samples (10 proliferative, 10 secretory, and 9 atrophic), and can be viewed as a very large matrix with the gene expression value of 54675 probe sets as its rows and 29 columns representing individual samples (arrays).

Data exploration

To explore the level of similarity of individual samples, we visualized the data set by multidimensional scaling (MDS), a technique for dimensionality reduction. MDS essentially represents the relationship among all samples in terms of their position in two-dimensional Euclidean space. Samples with similar gene expression profiles are placed at closer proximity compared with the dissimilar ones. The coordinates were plotted so that the distances between points reflect the Euclidean distance of the logarithms of expression between the samples. This analysis, which includes the data matrix obtained from the entire 54675 probe sets on arrays, demonstrates distinctively separated cluster for secretory samples (red) from all the other samples in different phases (Figure 4.1). All secretory samples appear on the left side of the plot with only two exceptions (sample 181 and 189), indicating consistent global expression profiles for secretory samples (Figure 4.1). The samples of the other phases seem to have some structures as well: samples are partly mingled but there are obviously groupings.

Atrophic samples (blue) are localized towards the top right of the plot, while proliferative samples (green) spread relatively widely on the right side of the plot (Figure 4.1). Although variable gene expression patterns between individual samples were observed as expected, differential gene expression between secretory and proliferative samples seem to be sufficient to separate the majority into their respective histological subgroups.

Principal component analysis (PCA), another commonly used technique for dimensional reduction, was also applied to the data set containing expression values of all probe sets. Display of the first two components of PCA-transformed data showed that secretory samples (red) are clearly separated from samples of the other two phases with two exceptions (sample 181 and 189). In contrast, proliferative and atrophic samples are less obvious to be separated (Figure 4.2). This indicates a set of gene signature could be identified to discriminate secretory phase from the other two phases, since the majority of secretory samples is already separated on the plane of the first two components. It is somewhat challenging to discriminate between proliferative and atrophic samples. Consistently, both MDS and PCA analysis suggests that the gene expression profiles common to secretory samples are distinctly different from the expression profiles of proliferative and atrophic samples.

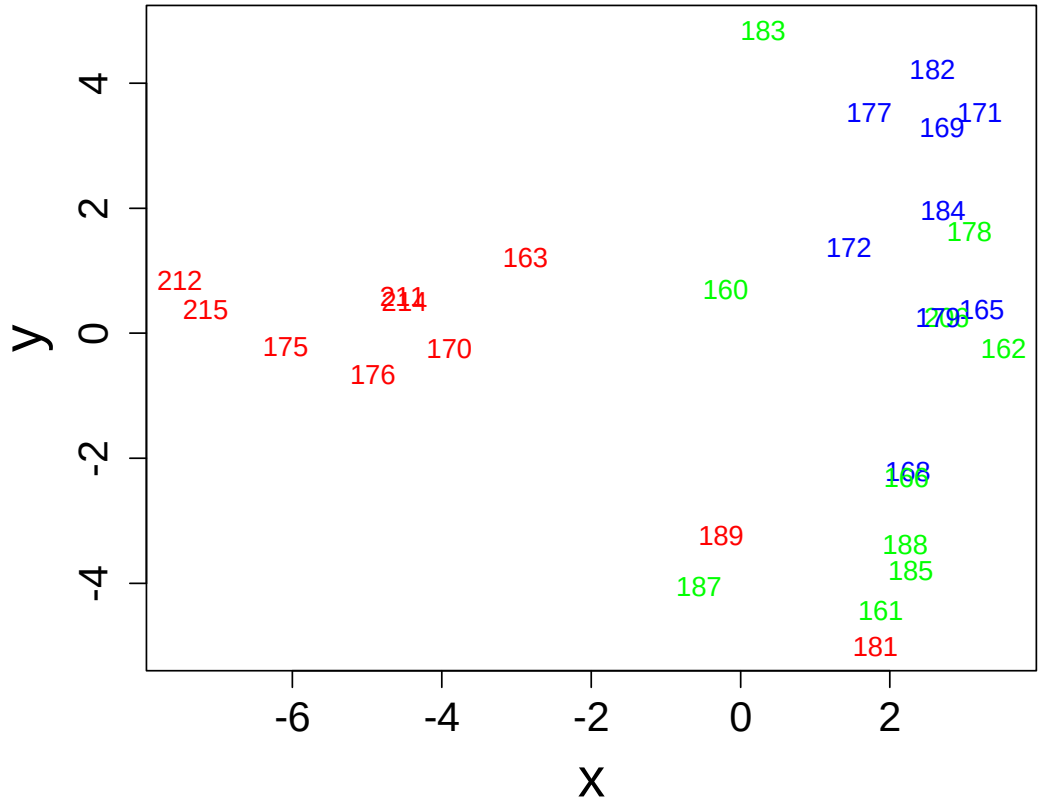

Figure 4.1. Multidimensional scaling analysis using the gene expression data of 29 endometrial samples generated using all probe sets. A two-dimensional projection shows each individual sample (labeled with sample ID) plotted onto an arbitrarily scaled x-y plane. Colors represent different cyclic phases: red, secretory; green, proliferative and blue, atrophic.

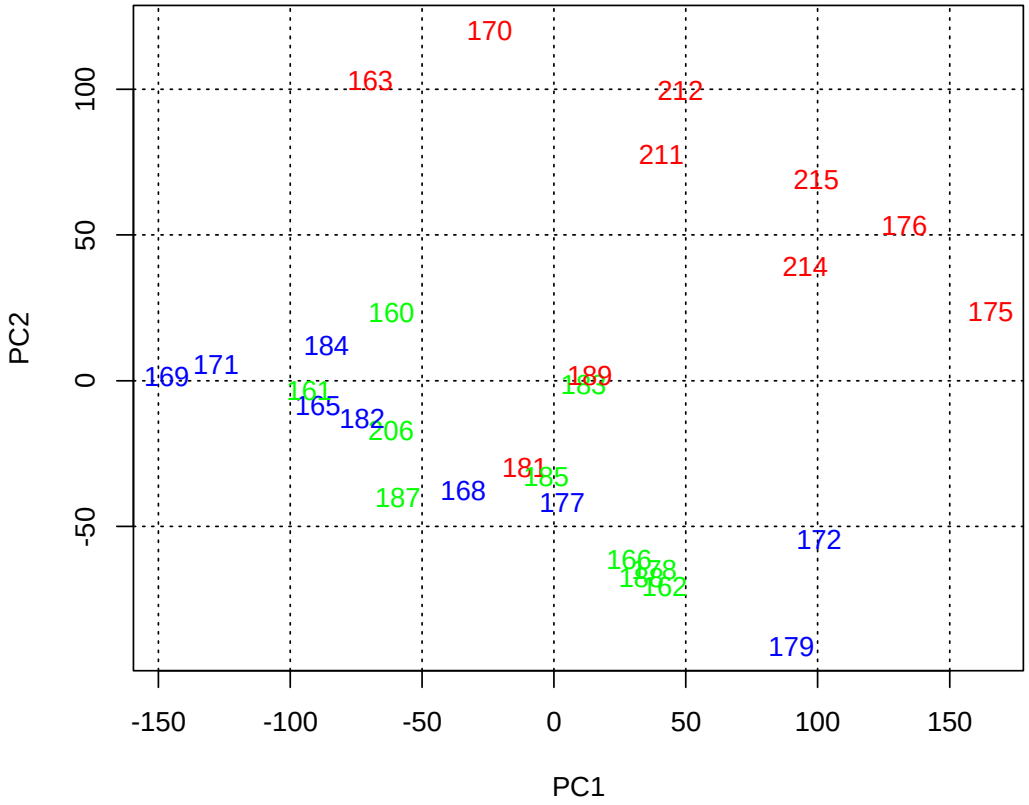

Figure 4.2. Principal component analysis. First and second components of the endometrial expression data using all probe sets are displayed. Sample is shown as sample ID. Colors represent different cyclic phases: red, secretory; green, proliferative and blue, atrophic.

Unsupervised hierarchical clustering

Next, unsupervised hierarchical clustering was conducted to explore if normal cycling endometrial tissues could be organized into meaningful structures on the basis of similarity of gene expression profiles. As shown in a dendrogram tree, hierarchical clustering using all 54675 probe sets clearly identified two major clusters, where 6 of 10 secretary samples self-clustered as one major branch, while the other branch is mingled with proliferative and atrophic samples (Figure 4.3A). As some probe sets have relatively small variations in expression values across patient samples or subgroups, which would be less informative for clustering analysis, standard deviation (SD) of the log2 expression values of each probe set across 29 samples was computed and probe sets were ranked based on their SD values. Varying number of top-ranked probe sets was experimentally tested using hierarchical clustering analysis. As shown in Figure 4.3B, hierarchical clustering with 200 top-ranked probe sets resulted a dendrogram tree with one branch primarily consists of secretory samples (with exception of two proliferative samples), and the other branch is still mingled with proliferative and atrophic samples. Further, the branch that contains only the proliferative and atrophic samples is composed of two sub-branches. One sub-branch contains mostly proliferative samples (5 proliferative and 3 atrophic samples). The other sub-branch contains 6 atrophic and 3 proliferative samples. Clustering with a subset of top-rank probe sets showed a superior performance in self-clustering of secretory samples compared to clustering with all probe sets.

In addition to clustering samples, we also performed hierarchical clustering of probe sets to explore if the expression of any sets of genes were regulated in a similar way across different samples. As demonstrated in a heat map, a two-way hierarchical clustering with 200 top-ranked probe sets identified two major gene clusters with distinctive expression patterns across different phases of endometrial samples (Figure

4.4). Cluster 1 contains a cluster of genes that were selectively up-regulated in self-clustered secretory samples, thus may representing genes involved in active regulation in secretary phase of menstrual cycle. These included, among others, progestagen-associated endometrial protein (PAEP), chemokine ligand 14 (CXCL14), glutathione peroxidase 3 (GPX3), complement factor D (CFD), complement component 3 (C3), 4A (C4A), and metallothionein 1M (MT1M). Genes in cluster 2A were only up-regulated in a cluster of samples consisting proliferative samples and two secretory outliers (sample ID 181 and 189). These genes included topoisomerase II alpha (TOP2A), insulin-like growth factor 1 (IGF1), matrix metallopeptidases, MMP11, MMP26, etc. Majority of the genes in cluster 2B were increasingly expressed in atrophic and a subset of proliferative samples. These included, among others, early growth response 1 (EGR1), WNT inhibitory factor 1 (WIF1), and secreted frizzled related protein 4 (SFRP4). These indicate the expression of individual cluster or sub-cluster of genes was regulated by a similar pattern in cyclic phase-specific endometrial sample.

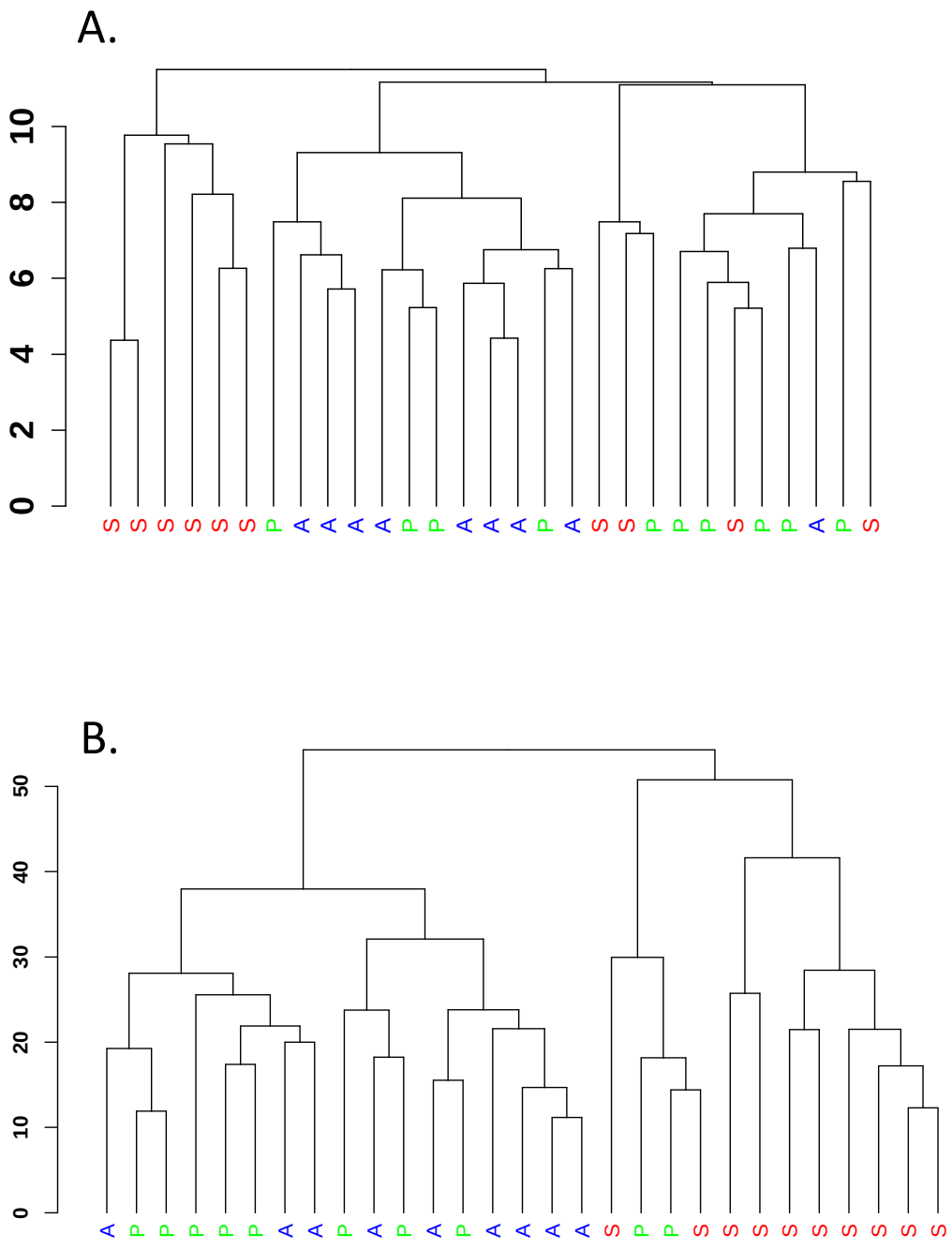

Figure 4.3. Dendrogram of unsupervised hierarchical clustering using Euclidean distance and complete link with all 54675 probe sets (A) and 200 top-ranked (by SD) probe sets (B). A, P, S represents atrophic, proliferative, and secretory endometrium, respectively.

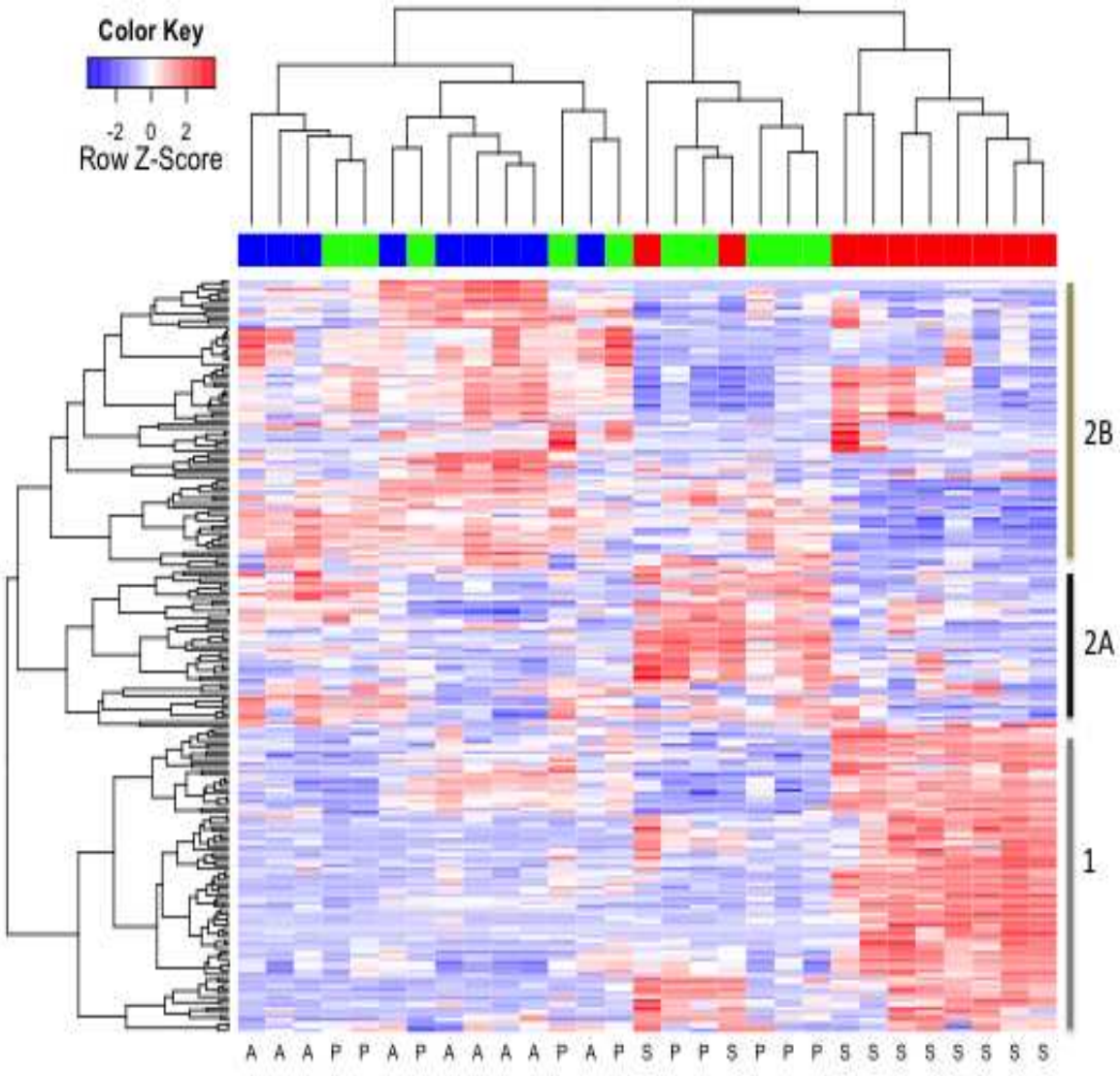

Figure 4.4. Heat map of unsupervised two-way hierarchical clustering using 200 top-ranked probe sets. Gene expression values of different samples are shown in columns and probe sets in rows. Probe sets were clustered using Pearson correlation and samples were clustered using Euclidean distance. Red indicates high expression, white intermediate and blue low expression.

Supervised analysis of gene expression profiles

In parallel, the gene expression profiles were subjected to pairwise comparisons, which were made between different cycle phases, specifically, secretory (S) vs. proliferative (P), atrophic (A) vs. secretory (S), and proliferative (P) vs. atrophic (A). The pairwise comparisons were carried out through linear model fitting using the software package *LIMMA* implemented for the R computing environment [32]. Differentially expressed probe sets between each pairwise comparison were determined separately with moderated t-test. Only probe sets with at least a 2.0 fold change and an adjusted p-value of less than 0.05 by Benjamini-Hochberg test for adjusting false discovery rate were considered. Probe sets were ranked based on their adjusted P-values.

Comparison of secretory vs. proliferative endometrium (S_P) identified 406 probe sets to be significantly differentially expressed ($P < 0.05$). Among them, about two-thirds of the probe sets were up-regulated and one-third were down-regulated in secretory compared to proliferative samples. Under a more stringent condition, secretory vs. proliferative endometrium comparison identified a total of 86 probe sets (fold change > 4.0 and $P < 0.01$ by Benjamini-Hochberg test) (Table 4.1). The majority of these probe sets displays up-regulated expression in secretory compared to proliferative samples. On the other hand, comparison of atrophic vs. secretory endometrium (A_S) revealed 1071 differentially expressed probe sets to be significant ($P < 0.05$), where approximately two-thirds were increasingly expressed and one-third was decreasingly expressed in atrophic compared to secretory samples. Surprisingly, the expression profiles of proliferative and atrophic samples (P_A) are highly similar as no probe set was found to be differentially expressed under the criteria as described above. Relationship of these differentially expressed probe sets is shown in a Venn diagram (Figure 4.5), where 326 probe sets are shared between the S_P and A_S comparisons. Consistent with the findings in the unsupervised analysis, these suggest the greatest

significant difference in the expression profiles was between secretory and the other two phases, whereas the least between proliferative and atrophic phase.

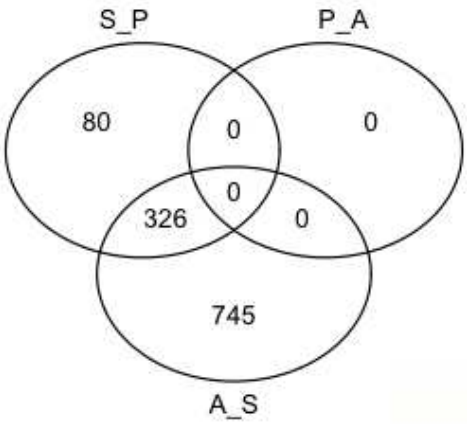

Figure 4.5. Relationship of differentially expressed probe sets between each pairwise comparisons. S_P: secretory vs. proliferative, A_S: atrophic vs. secretory, P_A: proliferative vs. atrophic.

Functional annotation analysis

To inspect the biological significance of differentially expressed probe sets, we performed GO term and KEGG pathway analysis using the DAVID bioinformatics tool [33] to identify the KEGG pathways that are most significantly enriched within the top ranked probe sets. This analysis was performed separately with the lists of probe set obtained from pairwise comparisons of different cyclic phases.

Among the 406 probe sets that were differentially expressed between secretory and proliferative samples, 283 were up-regulated and 123 were down-regulated in secretory compared to proliferative endometrium. Functional annotation analysis of these probe sets identified a few biological processes that were significantly different between secretary and proliferative endometrium. These include a series of immune responses,

such as acute inflammatory response and adaptive immune response (Table 4.2). In addition, the KEGG pathway, complement and coagulation cascades, was significantly enriched (adj. $P = 2.3\times10^{-3}$). Enriched genes in the complement and coagulation cascades pathway are listed in Table 4.3.

Comparison of atrophic and secretory endometrium (A_S) identified 1071 differentially expressed probe sets ($P < 0.05$). Among them the expression of 657 probe sets were elevated and 414 were decreasingly expressed in atrophic endometrium as compared to secretory endometrium. Functional annotation analysis identified Wnt signaling pathway was significantly enriched in the up-regulated genes ($P = 1.6\times10^{-3}$) and ECM-receptor interaction ($P = 2.6\times10^{-4}$) and complement and coagulation cascades ($P = 3.1\times10^{-2}$) are significantly enriched among the down-regulated genes. Enriched genes of these pathways are listed in Table 4.4 and Table 4.5, respectively. These data provide an overview of the biological processes and pathways that are involved in the cyclic phases of menstrual cycle. These data demonstrate the activation of complement and coagulation cascades in the secretory phase but not in the proliferative and atrophic phases.

Table 4.1. Differentially expressed genes between secretory and proliferative endometrium *

NO.	PROBE ID	GENE SYMBOL	LogFC	AveExpr	Adj.P-VALUE
1	218002_s_at	CXCL14	6.86	5.27	3.02E-03
2	241031_at	C2CD4A	6.12	5.56	1.39E-03
3	222484_s_at	CXCL14	5.84	6.11	8.07E-03
4	204602_at	DKK1	5.11	4.84	8.11E-04
5	205799_s_at	SLC3A1	4.68	5.95	4.90E-03
6	207254_at	SLC15A1	4.54	4.05	1.26E-03
7	205713_s_at	COMP	4.51	3.94	9.27E-03
8	213524_s_at	G0S2	4.28	6.87	2.58E-04
9	229638_at	IRX3	4.22	8.35	6.01E-03
10	207802_at	CRISP3	4.20	4.21	8.44E-03
11	205382_s_at	CFD	3.95	8.04	4.21E-03
12	214450_at	CTSW	3.84	4.89	2.16E-03
13	39248_at	AQP3	3.65	9.09	6.44E-03
14	243713_at	SLC1A1	3.63	4.58	7.04E-03
15	209875_s_at	SPP1	3.60	8.20	6.26E-03
16	202238_s_at	NNMT	3.58	4.24	2.23E-03
17	204388_s_at	MAOA	3.58	6.13	6.82E-03
18	210164_at	GZMB	3.57	6.12	5.75E-03
19	208451_s_at	C4A, C4B	3.56	7.79	3.55E-03
20	229254_at	MFSD4	3.53	4.85	3.59E-03
21	217546_at	MT1M	3.49	3.76	8.44E-03
22	215223_s_at	SOD2	3.47	5.62	9.00E-03
23	229004_at	ADAMTS15	3.23	5.94	5.16E-03
24	203946_s_at	ARG2	3.22	5.13	2.06E-03
25	224840_at	FKBP5	3.16	6.14	5.67E-03
26	212741_at	MAOA	3.15	6.25	5.41E-03
27	230084_at	SLC30A2	3.13	3.93	2.16E-03
28	204745_x_at	MT1G	3.10	7.82	4.21E-03
29	209283_at	CRYAB	3.10	5.84	2.31E-04
30	216841_s_at	SOD2	3.09	6.10	9.13E-03
31	217165_x_at	MT1F	3.07	8.80	8.39E-03
32	228486_at	SLC44A1	3.03	5.19	8.73E-03
33	203973_s_at	CEBPD	2.95	10.70	2.57E-03
34	214428_x_at	C4A	2.89	9.26	6.82E-03
35	206461_x_at	MT1H	2.79	8.78	5.23E-03
36	213629_x_at	MT1F	2.78	9.48	8.44E-03
37	242874_at	ENSG00000260711	2.78	4.93	2.17E-03
38	203836_s_at	MAP3K5	2.71	6.41	2.55E-03
39	218960_at	TMPRSS4	2.68	5.86	5.71E-03
40	211417_x_at	GGT1	2.68	5.97	4.26E-03
41	208581_x_at	MT1X	2.61	8.82	7.69E-03
42	200986_at	SERPING1	2.60	9.74	6.96E-03
43	213637_at	DDX52	2.57	6.98	2.07E-03
44	212859_x_at	MT1E	2.56	8.84	8.81E-03
45	202856_s_at	SLC16A3	2.55	5.24	7.92E-03

46	209919_x_at	GGT1	2.53	5.94	6.32E-03
47	204389_at	MAOA	2.51	5.17	8.44E-03
48	215603_x_at	GGT2, GGT1	2.51	5.29	3.49E-03
49	208284_x_at	GGT1	2.50	5.80	4.21E-03
50	238063_at	TMEM154	2.46	4.91	4.32E-03
51	33323_r_at	SFN	2.43	10.06	5.92E-03
52	211456_x_at	MT1HL1	2.42	8.89	6.96E-03
53	212834_at	DDX52	2.36	6.95	4.23E-03
54	218880_at	FOSL2	2.30	7.47	3.32E-03
55	212185_x_at	MT2A	2.27	11.14	4.28E-03
56	211416_x_at	GGTLC1	2.27	4.06	8.44E-03
57	210524_x_at	MT1F	2.26	9.73	9.95E-03
58	218001_at	MRPS2	2.25	6.58	6.94E-03
59	1553986_at	RASEF	2.24	5.81	4.17E-03
60	216336_x_at	MT1E	2.20	7.76	8.44E-03
61	33322_i_at	SFN	2.17	10.78	5.71E-03
62	207131_x_at	GGT1	2.16	5.76	9.27E-03
63	230537_at	AA401256	2.16	3.73	8.81E-03
64	1568736_s_at	BC030096	2.13	3.17	1.39E-03
65	205098_at	CCR1	2.11	3.97	7.07E-03
66	208869_s_at	GABARAPL1	2.06	7.27	4.32E-03
67	224374_s_at	EMILIN2	2.03	7.88	5.23E-03
68	214889_at	FAM149A	2.01	4.87	8.44E-03
69	235048_at	FAM169A	-2.17	5.43	6.84E-03
70	224480_s_at	AGPAT9	-2.19	4.58	2.06E-03
71	235079_at	ZNF704	-2.20	8.78	1.26E-03
72	224428_s_at	CDCA7	-2.25	8.33	7.80E-03
73	214247_s_at	DKK3	-2.29	9.83	4.32E-03
74	230943_at	SOX17	-2.41	8.81	2.07E-03
75	218718_at	PDGFC	-2.61	6.78	5.92E-03
76	202037_s_at	SFRP1	-2.93	8.42	4.90E-03
77	206622_at	TRH	-2.96	7.01	9.44E-03
78	210319_x_at	MSX2	-2.97	6.76	9.13E-03
79	218824_at	PNMAL1	-2.98	6.95	1.39E-03
80	238066_at	RBP7	-3.24	6.84	1.78E-03
81	203296_s_at	ATP1A2	-3.28	5.74	1.39E-03
82	229281_at	NPAS3	-3.40	7.05	2.57E-03
83	223475_at	CRISPLD1	-3.53	6.02	5.36E-03
84	210809_s_at	POSTN	-4.01	6.88	1.52E-03
85	204051_s_at	SFRP4	-4.26	10.90	1.39E-03
86	204052_s_at	SFRP4	-4.42	10.36	1.78E-03

* Only probe sets with logFC (log2 fold change) greater than 2.0 and adjusted P-value less than 0.01 are included. Probe sets were ranked based on logFC.

Table 4.2. Enrished biological processes in secretory vs. proliferative comparison

TERM	Adj. P-VALUE
Defense response	0.005
Response to wounding	0.005
Regulation of epithelial cell proliferation	0.018
Acute inflammatory response	0.023
Complement activation	0.029
Locomotory behavior	0.026
Activation of plasma proteins involved in acute inflammatory response	0.024
Regulation of cell proliferation	0.022
Complement activation, classical pathway	0.026
Protein processing	0.026
Innate immune response	0.024
Inflammatory response	0.026
Humoral immune response mediated by circulating immunoglobulin	0.025
Lymphocyte mediated immunity	0.027
Positive regulation of immune system process	0.032
Protein maturation	0.031
Response to organic substance	0.034
Immunoglobulin mediated immune response	0.033
Adaptive immune response	0.035
Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.035
Immune response	0.035
B cell mediated immunity	0.035
Humoral immune response	0.036
Vascular process in circulatory system	0.038
Behavior	0.040

Table 4.3. Enriched genes in complement and coagulation cascade pathway in secretory vs. proliferative comparison

PROBE ID	GENE SYMBOL	GENE NAME
1555950_a_at , 201926_s_at	CD55	CD55 molecule, decay accelerating factor for complement
200983_x_at	CD59	CD59 molecule, complement regulatory protein
205870_at	BDKRB2	bradykinin receptor B2
212067_s_at	C1R	complement component 1, r subcomponent
208747_s_at	C1S	complement component 1, s subcomponent
203052_at	C2	complement component 2
214428_x_at, 208451_s_at	C4A	complement component 4 alpha
205382_s_at	CFD	complement factor D (adipsin)
200986_at	SERPING1	serpin peptidase inhibitor, clade G (C1 inhibitor), member 1
203887_s_at	THBD	thrombomodulin

Table 4.4. Enriched genes in Wnt signaling pathway in atrophic vs. secretory comparison

Gene Symbol	Gene Name
SMAD4	SMAD family member 4
SOX17	SRY (sex determining region Y)-box 17
WIF1	WNT inhibitory factor 1
AXIN2	Axin 2
FZD1	Frizzled homolog 1 (Drosophila)
FZD10	Frizzled homolog 10 (Drosophila)
FZD2	Frizzled homolog 2 (Drosophila)
FZD7	Frizzled homolog 7 (Drosophila)
LRP6	Low density lipoprotein receptor-related protein 6
LEF1	Lymphoid enhancer-binding factor 1
MMP7	Matrix metallopeptidase 7
PLCB1	Phaspholipase C, beta 1 (phosphoinositide-specific)
SFRP1	Secreted frizzled-related protein 1
SFRP4	Secreted frizzled-related protein 4
TCF7L1	Transcription factor 7-like 1 (T-cell specific, HMG-box)

Table 4.5. Enriched genes in ECM-receptor interaction pathway in atrophic vs. secretory comparison

Gene Symbol	Gene Name
CD44	CD44 molecule
COMP	Cartilage oligomeric matrix protein
COL4A1	Collagen, type IV, alpha 1
COL4A2	Collagen, type IV, alpha 2
COL5A2	Collagen, type V, alpha 2
FN1	Fibronectin 1
ITGA5	Integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
LAMB1	Laminin, beta 1
LAMB3	Laminin, beta 3
LAMC3	Laminin, gamma 3
SPP1	Secreted phsphoprotein 1
SDC3	Syndecan 3
THBS2	Thrombospondin 2

Identification of bimodal genes

Compared to a unimodal distribution, genes with bimodal distributions or switch-like behavior have a clear advantage to classify samples into high and low expression subgroups. A bimodal distribution is a continuous probability distribution with two different modes. In the unsupervised analysis, we noticed switch-like behavior of two most dynamically expressed genes, PAEP (progestagen-associated endometrial protein, SD = 4.49) and CXCL14 (chemokine ligand 14, SD = 4.19). Both have been reported being functionally involved in endometrium cycling [34-36]. Distribution of PAEP expression showed a profile with distinctive two peaks roughly at 2.6 and 12.0, respectively, and covers an overall range of approximately 20.0, so as CXCL14 (Figure 4.6A). Figure 4.6B plotted the gene expression values of these two genes, where we can visually separate secretory samples (in red with relatively high expression values) from proliferative and atrophic samples (in green and blue, respectively with relatively low expression values) to certain extent, although with two secretory samples (sample 181 and 189) being exceptions. These suggest that genes with bimodal expression, similar to PAEP and CXCL14 are likely relevant to distinguish endometrial cycle phases.

In order to systematically develop a comprehensive list of bimodal or switch-like genes that are markedly changed during the process of endometrial menstrual cycle, the proposed two-component mixture model was applied to the data using the MCLUST algorithm. As in large-scale microarray experiments, it is a common problem for the presence of noise originating from various sources. In order to remove some noises and reduce dimensionality of data matrix prior to the analysis, we computed standard deviation (SD) for each probe set across all samples and ranked the probe sets based on SD values. Only probe sets with a SD value of 1.5 or greater (n = 1683) were subjected to MCLUST algorithm.

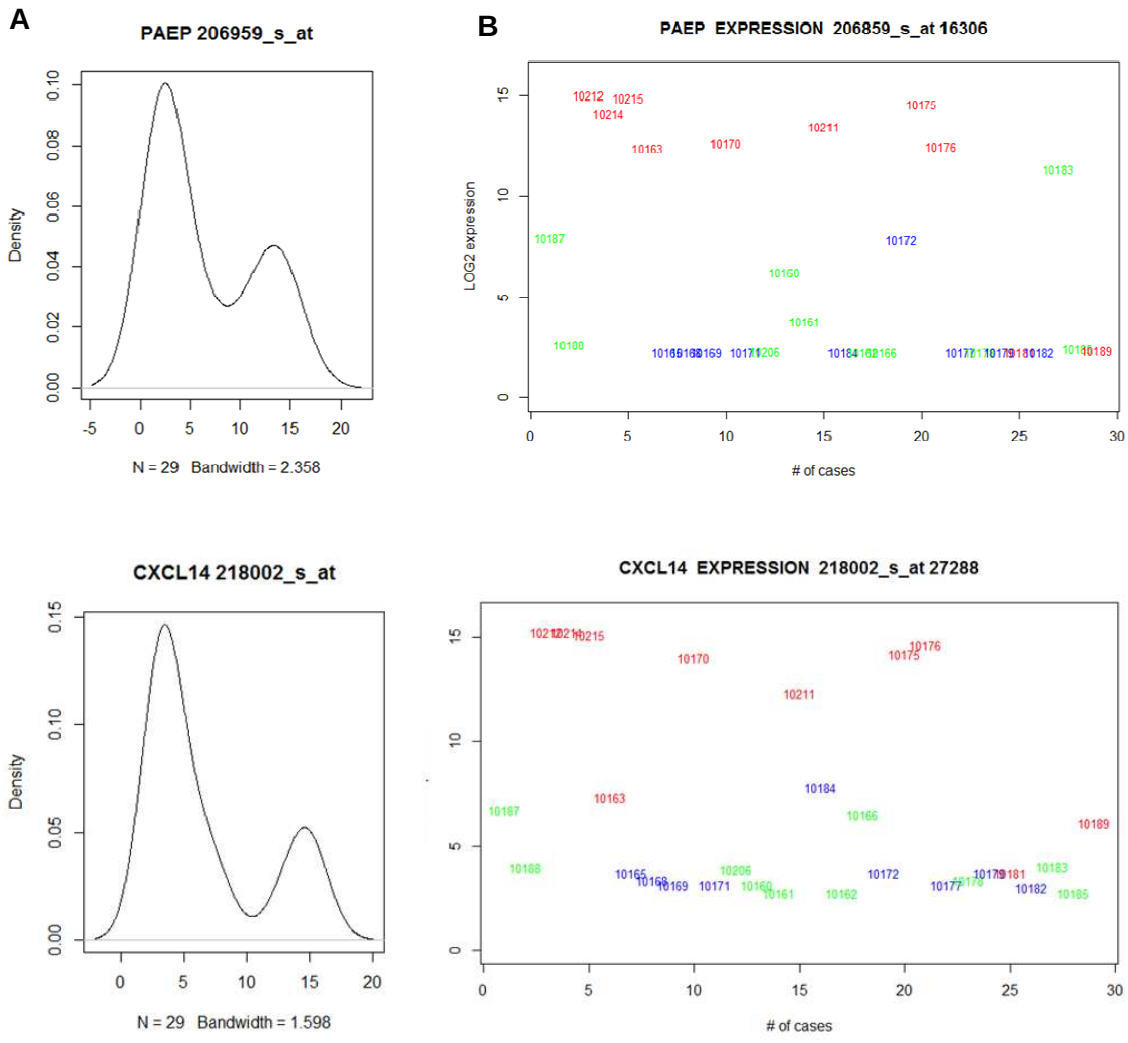

Figure 4.6. Density and expression plots of PAEP and CXCL14. A, Density plot of PAEP and CXCL14; B, Relative expression values of PAEP and CXCL14 across 29 normal endometrial samples. Each sample is labeled with its corresponding identification number. Colors represent different cycle phase: red, secretory; blue, atrophic; green, proliferative.

For each probe set, parameters μ_1, μ_2, σ, δ, and π were estimated from the expression data set across all samples. These parameters were then used to compute the bimodal index. Results were summarized in a table containing values of the above six parameters for each of 1683 probe sets. To identify a list of tens of genes that are applicable to molecular measurements, several filtering conditions were applied (some subjectivity in the selection of filtering criteria was involved). First, a cutoff of BI equal or greater than1.5 as suggested by Wang et al. was applied [29]. As π is the proportion of samples in one group ranging from 0.0 to 1.0 and there are almost equal numbers of proliferative, secretory, and atrophic samples in the data set, when π is close to either end of the range the power is much weaker to distinguish endometrial phases. According to manual inspection of their distribution, probe sets with π values out of the range from 0.20 to 0.80 were filtered out. Further, the larger the separation between the two components (δ) the easier they could be distinguished. δ value equal or greater than 4.0 as a cutoff identified 263 bimodal probe sets (Table 4.6). In order to get an impression of the distribution of these probe sets, we examined density plot of the log2 expression values. Density plots for selected top ranked genes are shown in Figure 4.7. The bimodality index identifies these genes have apparently visible bimodal expression distribution. Density plots for these top genes all show a larger group with lower expression and a smaller group with higher expression.

In order to identify a robust gene signature to distinguish menstrual cycle phases, we checked if any of the 263 bimodal probe sets was also differentially expressed between secretory and proliferative endometrium in the supervised analysis. As shown above, under a more stringent condition, secretory vs. proliferative endometrium comparison identified a total of 86 probe sets (fold change > 4.0 and $P < 0.01$ by Benjamini-Hochberg test) (Table 4.1). The majority of these probe sets displays up-regulated expression in secretory compared to proliferative samples. In merging 263

bimodal probe sets and 86 differentially expressed probe sets, we identified 43 probe sets corresponding to 35 unique genes common to both list of probe sets (Table 4.7). Approximately three quarters of these genes were up-regulated, and a quarter were down-regulated in secretory as compared to proliferative endometrium.

To examine if these genes could be used as a gene signature to distinguish endometrial cyclic phase markers, we performed a two-way hierarchical clustering analysis on an independent data set (see Talbi *et al.* data set in Chapter 2) using these genes. As different from the analyzed data set used above, Talbi data set assessed samples into proliferative (P), early (ES), mid (MS), and late-secretary (LS) cyclic phases. As shown in a heat map in Figure 4.8, hierarchical clustering separated samples into two major clusters. The left branch consists of all four proliferative samples (P) as one sub-cluster and all three early secretory samples (ES) as another sub-cluster. The rest secretory samples form the right branch, where one sub-branch is almost exclusively composed of mid secretory samples (MS) with only two exceptions (LS), and the other sub-branch contains three late secretory samples (Figure 4.8). As for probe sets, hierarchical clustering identified two major clusters. The smaller cluster consisting 9 probe sets was dramatically up-regulated in proliferative samples, moderated up-regulated in selected ES samples compared to MS and LS samples. These 9 probe sets correspond to genes SFRP4, SOX17, ATP1A2, TRH, MSX2, POSTN, PDGFC, and PNMAL1. In contrast, expression of genes in the other big cluster was gradually increased from P to ES, and even more in MS, while much reduced in LS samples. Genes in this big cluster showed a continuous gene expression pattern throughout menstrual cycle phases. The expression patterns of ES samples across these genes were somewhere in between those of P and MS samples (Figure 4.8). Thus, this gene signature was able to separate clustered proliferative endometrium from clustered early, mid, and late-secretory samples to a different extent.

Table 4.6. Identified bimodal genes with their estimated parameters [#]

NO.	PROBE ID	GENE SYMBOL	μ1	μ2	σ	δ	π	BI
1	207254_at	SLC15A1	3.06	9.62	0.54	12.23	0.72	5.46
2	236761_at	LHFPL3	2.28	8.09	0.58	10.10	0.66	4.80
3	203815_at	GSTT1	2.89	7.50	0.62	7.46	0.52	3.73
4	238103_at	LOC100505989	2.40	6.75	0.53	8.24	0.72	3.68
5	230673_at	PKHD1L1	3.21	9.91	0.80	8.41	0.76	3.60
6	209728_at	HLA-DRB4	3.09	9.21	0.78	7.87	0.72	3.52
7	238275_at	HAP1	2.36	7.36	0.64	7.88	0.76	3.37
8	204818_at	HSD17B2	2.36	6.91	0.65	6.95	0.62	3.37
9	207802_at	CRISP3	2.46	9.48	0.97	7.24	0.69	3.36
10	239336_at	THBS1	2.77	9.06	0.95	6.63	0.66	3.15
11	241031_at	C2CD4A	3.46	11.44	1.24	6.46	0.62	3.13
12	218002_s_at	CXCL14	4.15	14.39	1.44	7.13	0.76	3.05
13	1563077_at	LOC100289058	2.81	6.21	0.55	6.13	0.62	2.97
14	1568736_s_at	NA	2.34	5.55	0.52	6.20	0.66	2.94
15	206859_s_at	PAEP	3.04	13.24	1.68	6.07	0.68	2.84
16	238032_at	NA	3.23	6.42	0.57	5.64	0.52	2.82
17	206391_at	RARRES1	5.05	9.37	0.76	5.71	0.59	2.81
18	213791_at	PENK	2.92	9.86	1.06	6.55	0.76	2.81
19	202376_at	SERPINA3	3.00	6.40	0.61	5.60	0.55	2.78
20	1554663_a_at	NUMA1	2.38	5.57	0.51	6.21	0.72	2.78
21	229177_at	C16orf89	3.83	8.78	0.82	6.02	0.71	2.72
22	204745_x_at	MT1G	8.09	12.58	0.71	6.29	0.76	2.70
23	204602_at	DKK1	2.96	10.11	1.28	5.59	0.64	2.68
24	202238_s_at	NNMT	3.98	9.12	0.87	5.90	0.72	2.64
25	204846_at	CP	4.12	9.24	0.98	5.23	0.47	2.61
26	205844_at	VNN1	3.94	9.53	1.09	5.14	0.52	2.56
27	231181_at	NA	3.58	10.31	1.13	5.97	0.76	2.56
28	1555867_at	GNG4	2.78	6.55	0.69	5.49	0.69	2.54
29	230084_at	SLC30A2	3.43	8.54	0.97	5.29	0.65	2.52
30	205591_at	OLFM1	3.15	8.69	1.05	5.28	0.65	2.51
31	219580_s_at	TMC5	5.42	9.49	0.81	5.03	0.56	2.50
32	1565484_x_at	EGFR	2.84	6.44	0.72	4.97	0.53	2.48
33	231773_at	ANGPTL1	2.83	7.26	0.89	4.99	0.58	2.46
34	1554771_at	NA	2.72	6.18	0.65	5.29	0.69	2.46
35	204137_at	GPR137B	4.74	8.41	0.75	4.91	0.51	2.46
36	205654_at	C4BPA	3.28	8.94	1.13	4.99	0.59	2.45
37	215223_s_at	NA	5.36	10.26	0.93	5.25	0.68	2.45
38	228097_at	MYLIP	3.18	6.54	0.63	5.37	0.71	2.44
39	242579_at	BMPR1B	5.24	8.53	0.67	4.92	0.42	2.43
40	206461_x_at	MT1H	8.90	13.07	0.74	5.61	0.75	2.43
41	205433_at	BCHE	2.51	7.56	0.93	5.43	0.73	2.41
42	213992_at	COL4A6	2.77	6.25	0.70	4.99	0.63	2.40
43	203887_s_at	THBD	3.43	7.08	0.71	5.14	0.68	2.39
44	1553179_at	ADAMTS19	3.09	6.30	0.67	4.81	0.55	2.39
45	205890_s_at	NA	2.85	6.52	0.71	5.17	0.69	2.39
46	204378_at	BCAS1	3.93	6.70	0.58	4.76	0.49	2.38
47	224840_at	FKBP5	4.33	9.50	1.06	4.86	0.61	2.37
48	205259_at	NR3C2	3.53	7.76	0.77	5.52	0.76	2.37
49	219463_at	LAMP5	3.46	7.04	0.76	4.72	0.51	2.36

50	226690_at	ADCYAP1R1	3.32	7.51	0.80	5.22	0.72	2.36
51	209555_s_at	CD36	2.93	7.57	0.88	5.27	0.72	2.36
52	212671_s_at	NA	3.04	7.40	0.83	5.27	0.72	2.35
53	1552507_at	KCNE4	2.56	5.60	0.59	5.16	0.71	2.35
54	227475_at	FOXQ1	4.98	10.09	0.98	5.21	0.72	2.34
55	201242_s_at	ATP1B1	8.98	12.42	0.60	5.76	0.21	2.34
56	202870_s_at	CDC20	2.65	6.59	0.78	5.06	0.69	2.33
57	204051_s_at	SFRP4	8.44	13.58	0.96	5.35	0.26	2.33
58	219649_at	ALG6	3.88	6.83	0.63	4.73	0.42	2.33
59	209270_at	LAMB3	3.90	7.93	0.77	5.22	0.73	2.33
60	212834_at	DDX52	6.61	10.76	0.77	5.37	0.75	2.33
61	203951_at	CNN1	4.35	7.89	0.71	5.01	0.31	2.33
62	223672_at	SGIP1	2.97	7.03	0.78	5.18	0.72	2.32
63	1568768_s_at	BRE-AS1	2.58	7.73	0.98	5.26	0.74	2.31
64	205883_at	ZBTB16	2.76	7.86	1.10	4.65	0.56	2.31
65	221477_s_at	NA	4.56	7.93	0.72	4.67	0.58	2.30
66	205799_s_at	SLC3A1	3.78	10.50	1.43	4.69	0.60	2.30
67	209570_s_at	NSG1	4.38	7.73	0.73	4.60	0.49	2.30
68	244726_at	NA	7.61	10.77	0.69	4.59	0.48	2.30
69	211143_x_at	NR4A1	7.72	11.15	0.64	5.34	0.76	2.29
70	219260_s_at	ELP5	3.71	7.03	0.72	4.60	0.56	2.29
71	202237_at	NNMT	7.51	12.28	0.95	5.02	0.71	2.27
72	222484_s_at	CXCL14	5.49	14.17	1.65	5.27	0.75	2.27
73	202575_at	CRABP2	5.19	8.98	0.82	4.59	0.58	2.27
74	214567_s_at	NA	5.30	9.28	0.86	4.62	0.59	2.27
75	231063_at	NA	3.82	9.29	1.15	4.76	0.35	2.26
76	209283_at	CRYAB	4.92	8.83	0.80	4.89	0.70	2.24
77	228055_at	NAPSB	2.97	6.67	0.81	4.56	0.60	2.24
78	228325_at	SPIDR	3.16	6.39	0.69	4.69	0.65	2.23
79	218960_at	TMPRSS4	4.99	8.68	0.83	4.46	0.52	2.23
80	204052_s_at	SFRP4	6.62	12.76	1.18	5.21	0.24	2.23
81	1565483_at	EGFR	4.83	8.79	0.88	4.48	0.44	2.23
82	226612_at	UBE2QL1	3.66	7.06	0.68	4.97	0.73	2.22
83	244876_at	NA	2.50	5.67	0.64	4.99	0.73	2.22
84	202952_s_at	ADAM12	2.93	7.01	0.85	4.80	0.69	2.22
85	218880_at	FOSL2	6.78	10.09	0.69	4.77	0.69	2.22
86	205242_at	CXCL13	2.76	7.91	1.12	4.60	0.64	2.22
87	234032_at	NA	3.16	6.34	0.72	4.44	0.54	2.21
88	239178_at	FGF9	3.67	7.42	0.73	5.11	0.75	2.21
89	203180_at	ALDH1A3	4.75	10.06	1.16	4.59	0.64	2.20
90	210029_at	IDO1	4.53	8.94	1.00	4.40	0.49	2.20
91	201289_at	CYR61	6.76	12.65	1.16	5.09	0.25	2.20
92	227463_at	ACE	3.72	7.21	0.72	4.85	0.71	2.19
93	233241_at	PLK1S1	4.37	7.40	0.64	4.72	0.69	2.19
94	205413_at	MPPED2	2.72	7.19	0.92	4.85	0.28	2.18
95	1568647_at	LOC100505851	2.63	5.80	0.65	4.87	0.28	2.18
96	1568648_a_at	LOC100505851	3.98	7.84	0.77	5.03	0.24	2.16
97	1568611_at	NA	3.66	7.10	0.71	4.86	0.73	2.16
98	214234_s_at	CYP3A5	2.74	6.88	0.82	5.06	0.76	2.16
99	244444_at	PKD1L2	3.51	8.03	1.02	4.43	0.61	2.16
100	203571_s_at	ADIRF	3.51	6.39	0.65	4.43	0.61	2.16
101	243395_at	NA	4.85	8.99	0.94	4.40	0.60	2.16
102	225834_at	NA	3.24	6.41	0.67	4.70	0.70	2.16
103	230378_at	SCGB3A1	4.96	8.96	0.89	4.48	0.63	2.16

104	242064_at	SDK2	5.31	9.22	0.90	4.35	0.57	2.15
105	241595_at	NA	2.83	5.53	0.62	4.31	0.49	2.15
106	240253_at	NA	3.68	7.53	0.88	4.38	0.60	2.15
107	241916_at	NA	4.00	7.34	0.72	4.62	0.68	2.15
108	236373_at	NA	3.15	6.54	0.70	4.82	0.73	2.14
109	1555786_s_at	LINC00520	2.87	6.43	0.83	4.27	0.49	2.14
110	229254_at	MFSD4	4.28	9.14	1.02	4.74	0.72	2.13
111	206622_at	TRH	2.98	8.79	1.34	4.33	0.41	2.13
112	224339_s_at	ANGPTL1	2.69	5.55	0.66	4.32	0.59	2.13
113	228143_at	CP	4.60	8.90	1.00	4.30	0.42	2.12
114	210809_s_at	POSTN	3.63	9.02	1.21	4.44	0.35	2.12
115	213131_at	OLFM1	5.56	10.69	1.19	4.31	0.59	2.12
116	211456_x_at	MT1HL1	9.89	13.24	0.68	4.95	0.76	2.12
117	228377_at	KLHL14	3.72	7.17	0.75	4.62	0.70	2.11
118	239568_at	PLEKHH2	2.43	5.73	0.78	4.23	0.56	2.10
119	221872_at	RARRES1	2.82	7.16	1.01	4.29	0.61	2.10
120	230147_at	F2RL2	3.36	7.01	0.75	4.89	0.76	2.10
121	206366_x_at	XCL1	5.28	9.03	0.88	4.25	0.59	2.09
122	225987_at	STEAP4	3.36	8.33	1.17	4.25	0.60	2.08
123	1558605_at	NA	3.63	7.14	0.75	4.67	0.73	2.08
124	222378_at	NA	3.39	7.36	0.93	4.28	0.62	2.08
125	205656_at	PCDH17	5.95	9.83	0.86	4.51	0.69	2.08
126	230943_at	SOX17	7.87	11.69	0.81	4.74	0.26	2.08
127	212859_x_at	MT1E	8.98	12.70	0.77	4.84	0.76	2.07
128	236901_at	NA	2.67	6.41	0.79	4.72	0.74	2.07
129	229839_at	SCARA5	3.38	10.06	1.46	4.58	0.71	2.07
130	1556474_a_at	FLJ38379	3.07	7.65	1.08	4.23	0.60	2.07
131	227641_at	FBXL16	3.95	7.16	0.78	4.14	0.52	2.07
132	202953_at	C1QB	6.09	9.01	0.70	4.16	0.57	2.06
133	223423_at	GPR160	3.80	7.25	0.83	4.15	0.44	2.06
134	220794_at	GREM2	3.80	7.99	1.02	4.12	0.50	2.06
135	236264_at	LPHN3	3.17	6.12	0.65	4.54	0.71	2.06
136	242324_x_at	CCBE1	3.59	6.28	0.65	4.13	0.54	2.06
137	236420_s_at	ANO4	4.32	7.01	0.65	4.14	0.45	2.06
138	202833_s_at	SERPINA1	4.68	9.54	1.05	4.63	0.73	2.05
139	1554485_s_at	TMEM37	4.04	8.08	0.90	4.49	0.70	2.05
140	205470_s_at	KLK11	3.97	8.09	0.99	4.15	0.57	2.05
141	1556097_at	HOMER2	3.43	7.27	0.86	4.47	0.70	2.05
142	213790_at	ADAM12	4.18	8.59	1.07	4.13	0.55	2.05
143	204388_s_at	MAOA	5.51	10.77	1.10	4.76	0.76	2.05
144	227058_at	MEDAG	4.90	8.82	0.94	4.16	0.59	2.05
145	217767_at	C3	5.94	11.77	1.36	4.30	0.35	2.05
146	203296_s_at	ATP1A2	3.01	7.40	1.07	4.12	0.44	2.04
147	232481_s_at	SLITRK6	3.98	8.75	1.11	4.29	0.35	2.04
148	1559528_at	LOC100129917	3.13	6.19	0.71	4.30	0.66	2.04
149	210346_s_at	CLK4	7.49	10.23	0.65	4.22	0.37	2.04
150	231982_at	C19orf77	3.30	6.28	0.65	4.55	0.72	2.03
151	228218_at	LSAMP	2.69	6.26	0.88	4.07	0.50	2.03
152	202920_at	ANK2	3.21	6.42	0.78	4.11	0.59	2.02
153	204794_at	DUSP2	4.36	8.26	0.90	4.35	0.31	2.02
154	229542_at	C20orf85	3.36	6.75	0.83	4.10	0.59	2.02
155	229569_at	NA	3.55	7.53	0.99	4.04	0.49	2.02
156	204712_at	WIF1	3.87	11.32	1.60	4.66	0.75	2.02
157	239006_at	SLC26A7	4.66	9.87	1.28	4.07	0.43	2.02

158	242907_at	GBP2	4.91	8.08	0.79	4.04	0.53	2.02
159	209792_s_at	KLK10	3.83	7.20	0.77	4.39	0.70	2.01
160	205266_at	LIF	5.00	8.25	0.77	4.24	0.66	2.01
161	219181_at	LIPG	2.99	6.83	0.82	4.70	0.76	2.01
162	214595_at	KCNG1	3.00	5.90	0.68	4.26	0.67	2.01
163	240935_at	NA	2.90	5.84	0.73	4.04	0.55	2.01
164	1559663_at	NA	2.75	5.91	0.78	4.02	0.51	2.01
165	203789_s_at	SEMA3C	4.37	7.90	0.79	4.49	0.28	2.01
166	228004_at	LINC00261	4.26	7.87	0.82	4.39	0.70	2.01
167	213524_s_at	G0S2	5.80	10.34	1.05	4.32	0.69	2.00
168	231969_at	STOX2	4.69	7.62	0.72	4.09	0.60	2.00
169	239726_at	ANK3	5.16	8.09	0.68	4.34	0.69	2.00
170	219478_at	WFDC1	4.58	8.61	0.99	4.07	0.59	2.00
171	227884_at	TAF15	3.49	6.21	0.67	4.05	0.42	2.00
172	228692_at	PREX2	4.46	7.33	0.69	4.13	0.62	2.00
173	213880_at	LGR5	3.86	9.72	1.37	4.28	0.68	2.00
174	205382_s_at	CFD	6.58	12.01	1.24	4.39	0.71	1.98
175	203946_s_at	ARG2	3.51	7.76	1.03	4.14	0.64	1.98
176	240509_s_at	GREM2	2.75	5.94	0.74	4.27	0.69	1.98
177	204748_at	PTGS2	2.79	5.81	0.69	4.35	0.71	1.98
178	213637_at	DDX52	7.07	10.47	0.75	4.51	0.74	1.98
179	242874_at	NA	4.53	8.23	0.88	4.20	0.67	1.98
180	205765_at	CYP3A5	3.22	7.71	1.01	4.43	0.73	1.98
181	32625_at	NPR1	5.28	8.69	0.78	4.40	0.72	1.97
182	208581_x_at	MT1X	9.09	12.79	0.80	4.61	0.76	1.97
183	243713_at	NA	3.73	8.98	1.18	4.46	0.73	1.97
184	203417_at	MFAP2	5.86	9.71	0.91	4.24	0.31	1.97
185	218824_at	PNMAL1	4.79	9.29	1.00	4.48	0.26	1.95
186	238584_at	IQCA1	4.35	8.08	0.93	4.04	0.37	1.95
187	218718_at	PDGFC	5.63	9.11	0.85	4.08	0.35	1.94
188	1568638_a_at	IDO2	2.87	7.78	1.08	4.53	0.76	1.94
189	217165_x_at	MT1F	8.80	13.20	0.98	4.51	0.76	1.93
190	231172_at	C9orf117	2.87	6.08	0.74	4.34	0.73	1.93
191	206268_at	LEFTY1	3.25	7.78	1.04	4.34	0.73	1.92
192	222314_x_at	EGOT	3.22	7.10	0.88	4.41	0.74	1.92
193	1555938_x_at	VIM	3.95	7.54	0.88	4.09	0.33	1.92
194	229659_s_at	PIGR	5.75	9.62	0.90	4.29	0.28	1.92
195	206010_at	HABP2	3.51	7.78	1.06	4.04	0.65	1.92
196	206012_at	LEFTY2	4.37	9.63	1.28	4.10	0.68	1.92
197	203304_at	BAMBI	5.73	9.12	0.83	4.08	0.67	1.91
198	201243_s_at	ATP1B1	7.36	11.29	0.95	4.14	0.31	1.91
199	228174_at	SCAI	4.60	7.49	0.71	4.06	0.67	1.91
200	226334_s_at	AHSA2	7.53	10.65	0.76	4.08	0.32	1.91
201	237719_x_at	RGS7BP	2.88	6.10	0.78	4.15	0.70	1.91
202	203186_s_at	S100A4	7.74	10.99	0.73	4.44	0.76	1.91
203	220014_at	PRR16	3.78	7.14	0.80	4.21	0.72	1.90
204	1555950_a_at	CD55	7.46	11.64	1.04	4.01	0.66	1.90
205	235849_at	SCARA5	3.38	9.87	1.58	4.12	0.69	1.90
206	228697_at	HINT3	4.35	7.77	0.81	4.24	0.72	1.89
207	201926_s_at	CD55	7.42	11.45	1.00	4.02	0.67	1.89
208	1556054_at	NA	3.78	6.82	0.72	4.22	0.73	1.88
209	238063_at	TMEM154	4.79	8.21	0.85	4.02	0.68	1.87
210	205495_s_at	GNLY	4.25	11.79	1.78	4.24	0.26	1.87
211	213212_x_at	NA	6.89	10.16	0.79	4.12	0.29	1.87

212	204619_s_at	VCAN	8.26	11.79	0.78	4.53	0.22	1.87
213	215775_at	THBS1	2.51	5.56	0.75	4.05	0.69	1.87
214	210524_x_at	NA	9.51	12.64	0.72	4.35	0.76	1.86
215	207828_s_at	CENPF	3.69	7.36	0.85	4.34	0.76	1.86
216	209542_x_at	IGF1	7.18	11.87	1.13	4.16	0.27	1.85
217	1570259_at	LIMS1	3.10	6.06	0.71	4.19	0.74	1.84
218	216248_s_at	NR4A2	5.61	10.05	1.07	4.17	0.73	1.84
219	233090_at	NA	2.74	5.73	0.74	4.04	0.71	1.84
220	205316_at	SLC15A2	4.63	8.44	0.93	4.07	0.72	1.82
221	238123_at	GABRQ	3.02	7.73	1.15	4.11	0.73	1.82
222	211748_x_at	PTGDS	9.43	13.19	0.91	4.13	0.26	1.81
223	218009_s_at	PRC1	5.07	8.93	0.92	4.18	0.75	1.81
224	211577_s_at	IGF1	6.91	11.70	1.19	4.02	0.28	1.81
225	205934_at	PLCL1	3.66	7.86	1.02	4.13	0.74	1.81
226	212867_at	NCOA2	5.84	9.72	0.90	4.32	0.22	1.80
227	229638_at	IRX3	8.05	13.10	1.21	4.18	0.75	1.80
228	235049_at	ADCY1	3.15	6.13	0.74	4.05	0.73	1.80
229	33304_at	ISG20	5.93	9.49	0.88	4.05	0.73	1.80
230	225777_at	SAPCD2	4.04	7.06	0.75	4.03	0.73	1.79
231	230987_at	NA	4.13	7.65	0.87	4.04	0.73	1.79
232	227404_s_at	EGR1	8.49	13.12	1.14	4.07	0.26	1.78
233	213629_x_at	MT1F	10.37	14.08	0.91	4.08	0.74	1.78
234	219993_at	SOX17	6.35	9.93	0.87	4.11	0.25	1.78
235	210319_x_at	MSX2	3.45	8.54	1.18	4.32	0.21	1.75
236	37145_at	GNLY	3.41	10.66	1.80	4.02	0.25	1.75

[#] Probe sets were ranked based on BI values.

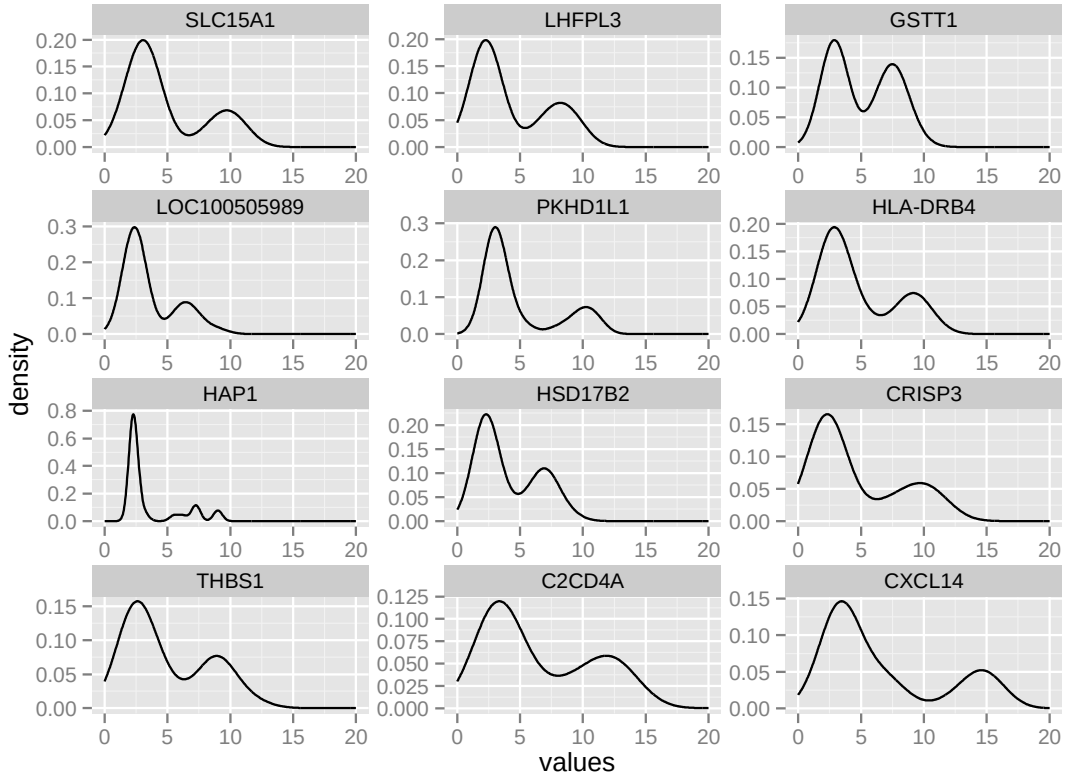

Figure 4.7. Density plots of top probe sets ranked by bimodality index

Table 4.7. Identified 43 probe sets as gene signature of endometrial cyclic phase

PROBE ID	GENE SYMBOL	GENE NAME
202238_s_at	NNMT	Nicotinamide N-methyltransferase
203296_s_at	ATP1A2	ATPase, Na+/K+ Transporting, Alpha 2 Polypeptide
203946_s_at	ARG2	Arginase 2
204051_s_at, 204052_s_at	SFRP4	Secreted frizzled-related protein 4
204388_s_at	MAOA	Monoamine oxidase A
204602_at	DKK1	Dickkopf WNT signaling pathway inhibitor 1
204745_x_at	MT1G	Metallothionein 1G
205382_s_at	CFD	Complement factor D (adipsin)
205799_s_at	SLC3A1	Solute carrier family 3 member 1
206461_x_at	MT1H	Metallothionein 1H
206622_at	TRH	Thyrotropin-releasing hormone
207254_at	SLC15A1	Solute carrier family 15 (oligopeptide transporter), member 1
207802_at	CRISP3	Cysteine-rich secretory protein 3
208581_x_at	MT1X	Metallothionein 1X
209283_at	CRYAB	Crystallin, alpha B
210319_x_at	MSX2	Msh homeobox 2
210524_x_at, 213629_x_at, 217165_x_at	MT1F	Metallothionein 1F
210809_s_at	POSTN	Periostin, osteoblast specific factor
211456_x_at	MT1HL1	Metallothionein 1H-like 1
212834_at, 213637_at	DDX52	DEAD (Asp-Glu-Ala-Asp) box polypeptide 52
212859_x_at	MT1E	Metallothionein 1E
213524_s_at	G0S2	G0/G1switch 2
215223_s_at	SOD2	Superoxide dismutase 2, mitochondrial
218002_s_at, 222484_s_at	CXCL14	Chemokine (C-X-C motif) ligand 14
218718_at	PDGFC	Platelet derived growth factor C
218824_at	PNMAL1	Paraneoplastic Ma antigen family-like 1
218880_at	FOSL2	FOS-like antigen 2
218960_at	TMPRSS4	Transmembrane protease, serine 4
224840_at	FKBP5	FK506 binding protein 5
229254_at	MFSD4	Major facilitator superfamily domain containing 4
229638_at	IRX3	Iroquois homeobox 3
230084_at	SLC30A2	Solute carrier family 30 (zinc transporter), member 2
230943_at	SOX17	SRY (sex determining region Y)-box 17
238063_at	TMEM154	Transmembrane protein 154
241031_at	C2CD4A	C2 calcium-dependent domain containing 4A

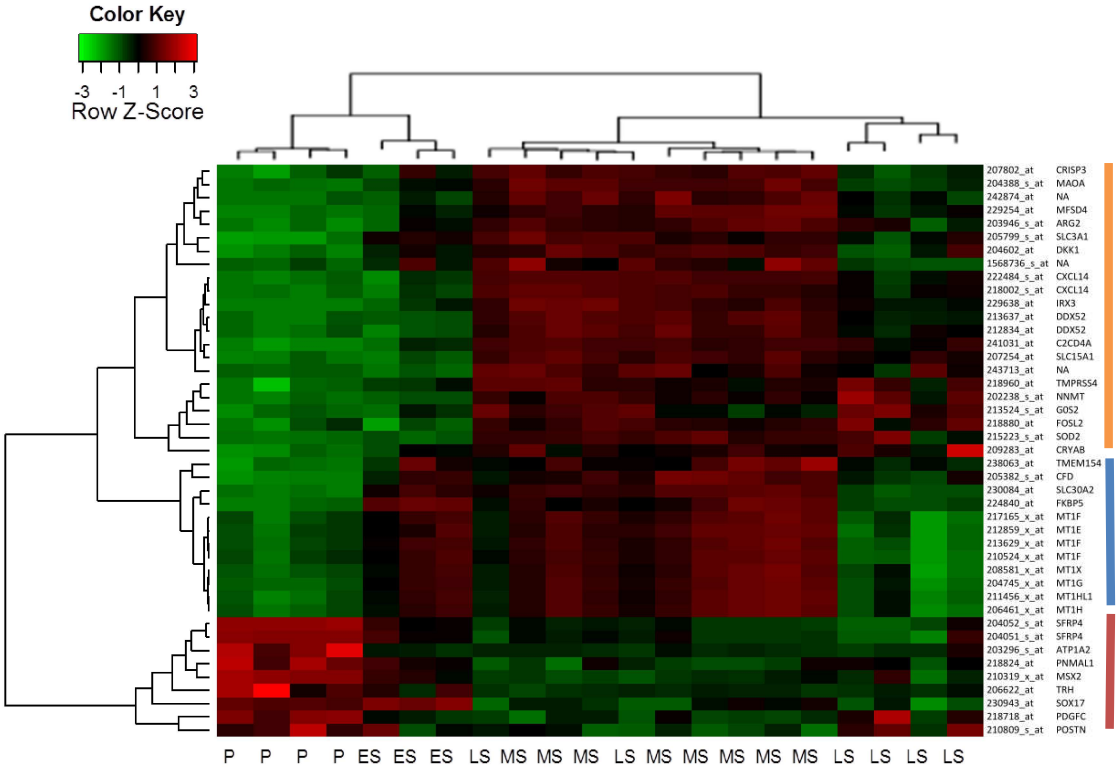

Figure 4.8. Two-way hierarchical clustering of an independent data set using 43 probe sets. Gene expression values of different samples are shown in columns and probe sets in rows. Red indicates high expression, black intermediate and green low expression. ES, early secretory; MS, mid secretory; LS, late secretory; P, proliferative

Chapter 5

Discussions and Conclusions

Several studies have been done to investigate whole-genome transcriptional profiles of normal endometrial menstrual cycle phases using microarray technologies [10, 37, 38] . However, these studies all used bulk collection of heterogeneous tumor samples which may contain contaminations from intervening stromal cells and infiltrating lymphocytes. In this study, we utilized a microarray gene expression data set generated by Wong *et al.* using laser microdissected (LCM) endometrial tissues. LCM is a powerful technique to accurately dissect pure population of specific cells from clinical specimens. Difference in gene expression patterns between microdissected vs. bulk dissected endometrial cancer tissues was demonstrated earlier [39], highlighting a confounding role of tissue dissection method play in interpreting the results of gene expression profiling data.

The Wong data set was originally used to investigate the development and progression of endometrial cancers through comparing gene expression profiles of endometrial cancers and normal endometrial tissues. In this study, our primary goal is to identify a relevant gene signature with the discriminative power to separate patient samples into subgroups with respect to menstrual cyclic phases. To achieve this, we investigated the global gene expression profiles of 29 normal endometrium samples, a subset of Wong data set. This data subset contains almost equal number of proliferative (10), secretory (10), and atrophic (postmenopausal, 9) endometrial samples, which provides composite representation of dynamic endometrial menstrual cycle (Table 2.1).

In our analysis, dimension reduction techniques, multidimensional scaling and principal component analysis were used to explore hidden structure in the data set. And unsupervised and supervised hierarchical clustering was used to cluster samples into meaningful structures based on similarities in their gene expression profiles. Interestingly, all these approaches, although using different list of genes, consistently demonstrated that global gene expression profiles of secretory endometrium are highly similar to each other, but significantly different from proliferative and atrophic endometrium (Figure 4.1, 4.2, 4.3 and 4.4). These suggest that endometrial tissue in each menstrual cycle phase harbors a unique gene expression signature, which could potentially be used as markers to distinguish endometrial cyclic phase in clinical samples. In clinical practice, menstrual cyclic phase of clinical endometrial samples is routinely determined by histological assessment, where errors or ambiguity might occur. There were two exceptional secretory samples (sample 181 and 189), which were consistently observed to be segregated away from the other secretory samples, and clustered together with proliferative samples in PCA, MDS, as well as hierarchical clustering analysis (Figure 4.1, 4.2, and 4.4). This may indicate histological dating errors in these two samples. It further highlights the potential of using gene expression profiles in conjunction with histological assessment in clinical practice to achieve better accuracy in dating endometrial tissues.

As far as for gene expression, unsupervised analysis using 200 probe sets with the most varying expression levels across samples, detected three major patterns (Figure 4.4). Cluster 1 was highly expressed in almost all secretory samples but not in samples of other phases. These included, among others, progestagen-associated endometrial protein (PAEP), chemokine ligand 14 (CXCL14), glutathione peroxidase 3 (GPX3), and complement factor D. Cluster 2A was only up-regulated in proliferative samples and two secretory outliers (sample 181 and 189), as well as a few atrophic samples. These included topoisomerase II alpha (TOP2A), insulin-like growth factor 1 (IGF1), matrix

metallopeptidase, MMP11 and MMP26. In contrast, expression of cluster 2B was somewhat heterogeneous across the samples (Figure 4.4). Majority of the genes in cluster 2B were increasingly expressed in atrophic and a subset of proliferative samples. These included, among others, early growth response 1 (EGR1), WNT inhibitory factor 1 (WIF1), and secreted frizzled related protein 4 (SFRP4). These indicate the expression of individual cluster or sub-cluster of genes was regulated in a similar pattern for cyclic phase-specific endometrial sample.

The observation of significant up-regulation of gene expression in only a subgroup of samples motivated us to look for switch-like or bimodally expressed genes. Bimodal distribution of gene expression have been observed in alternative mode within physiological or disease states such as diabetes, congestive heart failure, Alzheimer's disease, breast cancer, hypertension, obesity, and skeletal muscle tissue etc. [8, 40]. Thus, genes with bimodal distribution of expression process more robust power to distinguish endometrium menstrual cyclic phase. In this study, we demonstrated that the gene signature consisting 37 unique genes not only display a robust bimodal distribution in their gene expression across all samples, but also highly differentially expressed between secretory vs. proliferative phases (Table 4.7). Among these 37 genes, three quarters of them were highly up-regulated, and a quarter was down-regulated in secretory as compared to proliferative endometrium. The dramatic increased expression level of a large amount of transcripts in secretory endometrium is probably due to the influence of steroid hormones, estrogen and progesterone acting on the response element in the promoter regions of targeted genes [41].

Although it is unable to directly compare with others' report, as in the Wong data set secretory phase was not further subdivided into early, mid, and late secretory phase as some other data sets did, some observations in our analysis were still in consistency with previous reports. For example, we identified CXCL14 to be the most up-regulated

gene (> 100-fold) in S vs. P (Table 4.1). A previous study showed a 61-fold increase in CXCL14 expression, the most among others, in MS vs. ES [10]. CXCL14 has been known as a chemokine to recruit monocytes and may be other cell types to endometrium during the endometrial implantation window. Besides, numerous roles of CXCL14 have been described in cancers, including chemotactic factor for dendritic cells, potent inhibitor of angiogenesis, target for epigenetic silencing, and mediator of cancer cell mobility [42-44]. These all indicate important and complicated roles that CXCL14 play under both normal physiological and disease conditions.

Our study reports differences in expression levels of metallothioneins between phases of menstrual cycle. We showed that a number of metallothionein family members, MT1E, MT1F, MT1G, MT1H, MT1X, and MT1HL1 were significantly up-regulated in secretory phase endometrium (Table 4.1 and 4.7). This is in agreement with the previous observations in comparing MS vs. ES [10] and secretory vs. proliferative endometrium [41], respectively. Metallothioneins are a family of cysteine-rich heavy-metal binding proteins that express ubiquitously to protect cells against heavy metal toxicity and harmful reactive oxygen species [45]. Increased expression of these genes during the menstrual cycle probably plays a role in protecting the embryo from heavy metals and free radicals. It is worth noting that elevated metallothionein expression also has been observed in endometrial carcinomas [46, 47].

Among the down-regulated genes, Wnt signaling inhibitor SFRP4 was the most highly down-regulated (nearly 20-fold) in S vs. P. Significant decrease in SFRP4 expression has been observed in MS vs. ES previously [10]. Another Wnt inhibitor DKK1 was up-regulated more than 30-fold in S vs. P, which is known as an induced effect by progesterone. Furthermore, SOX17, a transcription regulator of Wnt signaling, was down-regulated 4-fold in secretory endometrium (Table 4.1). As the Wnt family consists of more than 20 secreted glycoproteins, the balance of these proteins determines the net

effect of Wnt signaling in regulating tissue remodeling, cellular proliferation, and differentiation during different phases of the menstrual cycle.

The current study used a model that assumes both components are normally distributed with equal variances in the bimodal distribution. As the two components of the bimodal distribution often do not follow normal distributions and are unequally distributed in reality, future study using models that allow for other distributions and unequal variances is worth pursuing for performance comparison. Methods using such models were described previously [48-50].

As shown in this study and many by others, the gene expression profiles of normal endometrium is highly dynamic during the menstrual cycle due to changing levels of ovarian steroids. In microarray-based gene expression studies of endometrial cancers, comparisons of cancer versus normal tissues were generally made using heterogeneous samples in terms of menstrual cycle phases or status of hormonal therapies, etc. Therefore, this may confound the search for differentially expressed genes that may play important roles in the progression of endometrial cancer.

As an exploration for this issue, we assessed expression of the signature genes identified above in an endometrial cancer gene expression data set, consisting 30 cancers and 28 surrounding normal samples, that we generated using Affymetrix GeneChip U133 plus 2.0 platform (unpublished data). For example, we observed high expression of NNMT, nicotinamide N-methyltransferase, in both normal and endometrial cancer samples (Figure 5.1). As NNMT was shown to be increasingly expressed in secretory phase of menstrual cycle in our analysis, this may confound the comparison of endometrial cancer and normal tissues. And knowing individual sample's cyclic phase is helpful to interpret results of specific gene expression comparison between normal and cancer endometrial samples.

In conclusion, our study provided a comprehensive overview of gene expression profiles of different cyclic phases of menstrual cycle. We identified a clinically manageable panel of signature genes that could potentially serve as a predictor to determine the menstrual cyclic phase of individual endometrium specimens. In addition, our study highlights the potential confounding effects of these cyclic genes on detecting differentially expressed genes in cancer versus normal endometrial tissues. Thus, it is recommended to determine the status of cyclic phase for each sample in identifying novel genes responsible for cancer aggressiveness, especially for endometrial cancer patients under 45 years old who still have normal menstrual cycle. For future work, further validation of the gene signature identified in this study is worth pursuing using a larger data set and quantitative real-time PCR on clinical specimens.

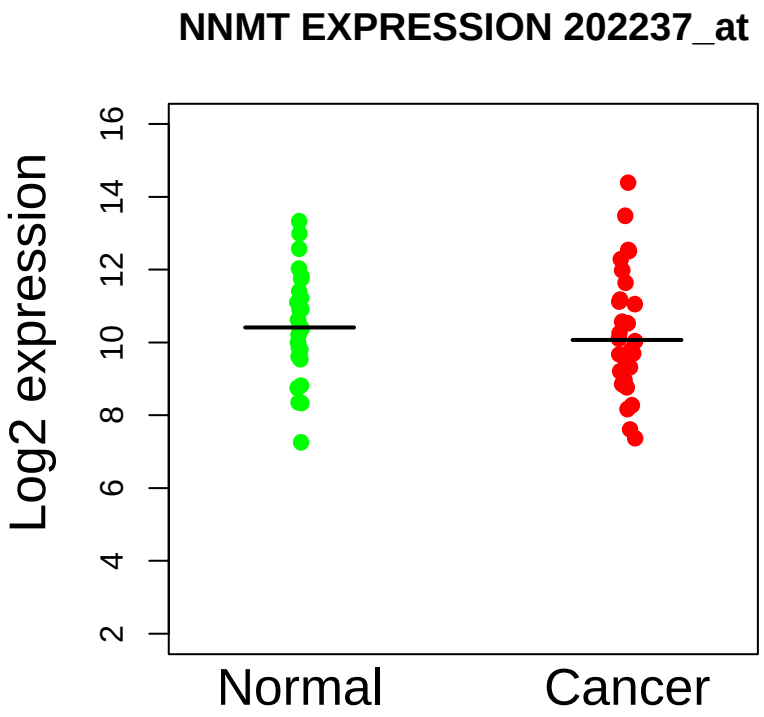

Figure 5.1. Expression of cyclic gene NNMT in a data set consisting endometrial cancer and surrounding normal tissues. Log2 expression value of normal (green) and cancer (red) samples were plotted. Black bars indicate the mean of each group.

Bibliography

1. Ferenczy, A., *The Endometrial Cycle*, in *Gynecology and Obstetrics*2004.
2. Noyes, R.W., A.T. Hertig, and J. Rock, *Dating the endometrial biopsy.* American journal of obstetrics and gynecology, 1975. **122**(2): p. 262-3.
3. Murray, M.J., et al., *A critical analysis of the accuracy, reproducibility, and clinical utility of histologic endometrial dating in fertile women.* Fertility and Sterility, 2004. **81**(5): p. 1333-1343.
4. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
5. Lockhart, D.J. and E.A. Winzeler, *Genomics, gene expression and DNA arrays.* Nature, 2000. **405**(6788): p. 827-36.
6. Affymetrix, *Technical Note: Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Array*s, 2003.
7. Paliwal, S., et al., *MAPK-mediated bimodal gene expression and adaptive gradient sensing in yeast.* Nature, 2007. **446**(7131): p. 46-51.
8. Ertel, A. and A. Tozeren, *Switch-like genes populate cell communication pathways and are enriched for extracellular proteins.* BMC Genomics, 2008. **9**(1): p. 3.
9. Wong, Y.F., et al., *Identification of molecular markers and signaling pathway in endometrial cancer in Hong Kong Chinese women by genome-wide gene expression profiling.* Oncogene, 2007. **26**(13): p. 1971-82.
10. Talbi, S., et al., *Molecular phenotyping of human endometrium distinguishes menstrual cycle phases and underlying biological processes in normo-ovulatory women.* Endocrinology, 2006. **147**(3): p. 1097-1121.
11. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic acids research, 2002. **30**(1): p. 207-10.
12. McCall, M.N. and A. Almudevar, *Affymetrix GeneChip microarray preprocessing for multivariate analyses.* Briefings in Bioinformatics, 2012. **13**(5): p. 536-546.
13. Wu, Z., et al., *A Model-Based Background Adjustment for Oligonucleotide Expression Arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.
14. Wu, Z., et al., *gcrma: Background Adjustment Using Sequence Information*, 2004.

15. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-264.
16. Cox, T.F. and M.A.A. Cox, *Multidimensional scaling*. Second ed2010.
17. Wall, M.E., A. Rechtsteiner, and L.M. Rocha, *Singular Value Decomposition and Principal Component Analysis*, in *A Practical Approach to Microarray Data Analysis*2003, Springer. p. 91-109.
18. Raychaudhuri, S., J.M. Stuart, and R.B. Altman, *Principal components analysis to summarize microarray experiments: application to sporulation time series.* Pac Symp Biocomput, 2000: p. 455-66.
19. Quackenbush, J., *Computational analysis of microarray data.* Nat Rev Genet, 2001. **2**(6): p. 418-427.
20. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences, 1998. **95**(25): p. 14863-14868.
21. Steinbach, M., G. Karypis, and V. Kumar, *A comparison of document clustering techniques*, 2000, Unversity of Minnesota: Minneapolis.
22. Tan, P., M. Steinbach, and V. Kumar, *Introduction to Data Mining*2005: Addison-Wesley.
23. Kruskal, J.B., *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.* Psychometrika, 1964. **29**(1): p. 1-27.
24. Emran, S.M. and N. Ye, *Robustness of Chi-square and Canberra distance metrics for computer intrusion detection.* Quality and Reliability Engineering International, 2002. **18**(1): p. 19-28.
25. Fulekar, M.H., *Bioinformatics: Applications in Life and Environmental Sciences*2009: Springer.
26. Smyth, G.K., *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**(1): p. 1544-6115.
27. Smyth, G.K., *limma: Linear Models for Microarray Data*, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*2005. p. 397-420.
28. Teschendorff, A.E., et al., *PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer.* Bioinformatics, 2006. **22**(18): p. 2269-2275.

29. Wang, J., et al., *The Bimodality Index: A Criterion for Discovering and Ranking Bimodal Signatures from Cancer Gene Expression Profiling Data.* Cancer Informatics, 2009. **7**: p. 199-216.
30. Hartigan, J.A. and P.M. Hartigan, *The Dip Test of Unimodality.* 1985(1): p. 70-84.
31. Fraley, C., et al., *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, 2012, University of Washington: Seattle.
32. Smyth, G.K., J. Michaud, and H.S. Scott, *Use of within-array replicate spots for assessing differential expression in microarray experiments.* Bioinformatics, 2005. **21**(9): p. 2067-75.
33. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat. Protocols, 2008. **4**(1): p. 44-57.
34. Joshi, S.G., et al., *Serum Levels of a Progestagen-Associated Endometrial Protein during the Menstrual Cycle and Pregnancy.* The Journal of Clinical Endocrinology & Metabolism, 1982. **55**(4): p. 642-648.
35. Julkunen, M., et al., *Secretory Endometrium Synthesizes Placental Protein 14.* Endocrinology, 1986. **118**(5): p. 1782-1786.
36. Mokhtar, N.M., et al., *Progestin regulates chemokine (C-X-C motif) ligand 14 transcript level in human endometrium.* Molecular Human Reproduction, 2010. **16**(3): p. 170-177.
37. Ponnampalam, A.P., et al., *Molecular classification of human endometrial cycle stages by transcriptional profiling.* Molecular Human Reproduction, 2004. **10**(12): p. 879-93.
38. Petracco, R.G., et al., *Global gene expression profiling of proliferative phase endometrium reveals distinct functional subdivisions.* Reproductive sciences, 2012. **19**(10): p. 1138-45.
39. Sugiyama, Y., et al., *Microdissection Is Essential for Gene Expression Profiling of Clinically Resected Cancer Tissues.* American Journal of Clinical Pathology, 2002. **117**(1): p. 109-116.
40. Mason, C., et al., *Bimodal distribution of RNA expression levels in human skeletal muscle tissue.* BMC Genomics, 2011. **12**(1): p. 98.
41. Borthwick, J.M., et al., *Determination of the transcript profile of human endometrium.* Molecular Human Reproduction, 2003. **9**(1): p. 19-33.

42. Tessema, M., et al., *Re-expression of CXCL14, a common target for epigenetic silencing in lung cancer, induces tumor necrosis.* Oncogene, 2010. **29**(37): p. 5159-70.
43. Shellenberger, T.D., et al., *BRAK/CXCL14 is a potent inhibitor of angiogenesis and a chemotactic factor for immature dendritic cells.* Cancer Research, 2004. **64**(22): p. 8262-70.
44. Pelicano, H., et al., *Mitochondrial dysfunction and reactive oxygen species imbalance promote breast cancer cell motility through a CXCL14-mediated mechanism.* Cancer Research, 2009. **69**(6): p. 2375-83.
45. Thornalley, P.J. and M. Vasak, *Possible role for metallothionein in protection against radiation-induced oxidative stress. Kinetics and mechanism of its reaction with superoxide and hydroxyl radicals.* Biochimica et biophysica acta, 1985. **827**(1): p. 36-44.
46. McCluggage, W.G., et al., *High metallothionein expression is associated with features predictive of aggressive behaviour in endometrial carcinoma.* Histopathology, 1999. **34**(1): p. 51-5.
47. Ioachim, E.E., et al., *Immunohistochemical localization of metallothionein in endometrial lesions.* The Journal of pathology, 2000. **191**(3): p. 269-73.
48. Lim, T.O., et al., *Bimodality in blood glucose distribution: is it universal?* Diabetes care, 2002. **25**(12): p. 2212-7.
49. Fan, J., et al., *Bimodality of 2-h plasma glucose distributions in whites: the Rancho Bernardo study.* Diabetes care, 2005. **28**(6): p. 1451-6.
50. McLachlan, G.J., R.W. Bean, and D. Peel, *A mixture model-based approach to the clustering of microarray expression data.* Bioinformatics, 2002. **18**(3): p. 413-22.