Microarray Bioinformatics and Its Applications to Clinical Research

A dissertation presented to

the School of Electrical and Information Engineering

of the University of Sydney

in fulfillment of the requirements for the degree of

Doctor of Philosophy

The University of Sydney

by

Ilene Y. Chen

Acknowledgment

This thesis owes its existence to the mercy, support and inspiration of many people. In the first place, having suffering from adult-onset asthma, interstitial cystitis and cold agglutinin disease, I would like to express my deepest sense of appreciation and gratitude to Professors Hong Yan and David Levy for harbouring me these last three years and providing me a place at the University of Sydney to pursue a very meaningful course of research. I am also indebted to Dr. Craig Jin, who has been a source of enthusiasm and encouragement on my research over many years.

In the second place, for contexts concerning biological and medical aspects covered in this thesis, I am very indebted to Dr. Ling-Hong Tseng, Dr. Shian-Sehn Shie, Dr. Wen-Hung Chung and Professor Chyi-Long Lee at Change Gung Memorial Hospital and University of Chang Gung School of Medicine (Taoyuan, Taiwan) as well as Professor Keith Lloyd at University of Alabama School of Medicine (AL, USA). All of them have contributed substantially to this work.

In the third place, I would like to thank Mrs. Inge Rogers and Mr. William Ballinger for their helpful comments and suggestions for the writing of my papers and thesis.

In the fourth place, I would like to thank my swim coach, Hiroto Homma. Hiroto is a master coach at North Sydney Aussie Masters Swimming Club. Although the asthma, interstitial cystitis and cold agglutinin disease I had suffered were untreatable or incurable, Hiroto taught and used "Total Immersion" swimming to increase my overall fitness level. Not only had I completely got rid off the asthma, interstitial cystitis and cold agglutinin disease one and a half years ago, but I can also swim like Michael Phelps at this time.

Finally, I owe a special gratitude to my parents for continuous and unconditional support of all my understandings, the funds to pursue my PhD and many medical expenses.

Microarray Bioinformatics and Its Applications to Clinical Research

Ilene Y Chen

University of Sydney 2009

Abstract

This thesis provides an overview of my research over the past three years into microarray gene expression data analysis in clinical research. This work involves using bioinformatics techniques to study disease-related microarray gene expression data. These bioinformatics methods contain novel features which are of considerable biological interest. Fundamental to this thesis is that the mass of numbers produced by DNA microarrays amounting to hundreds of data points for thousands or tens of thousands of genes are primarily driven by their genetic roots in the human genome. It enables the genome to become a unifying explanation for all of human biology and medicine.

This thesis is composed of three parts. In the first part, I describe the fundamental research strategies in which DNA microarrays have started to affect clinical research. In the second part of the work, I provide a choice of bioinformatics techniques to support the use of genome-based expression profiling as a commonplace microarray platform for minimally invasive diagnostic tests and therapeutic interventions in the study of diseases. In the third part, I employ the bioinformatics techniques presented in the second part to identify the molecular biological mechanisms as well as to pinpoint potential targets for drug discovery and diagnostics.

This work, as described in the three parts, leads to the facility of genome-based expression profiling as a commonplace platform for diagnosis and therapeutic interventions of almost any kind of disease in clinical practice in a straightforward manner.

Microarray Bioinformatics and Its Applications to Clinical Research

Ilene Y Chen

University of Sydney 2009

Table of Contents

List of Tables

List of Figures

Chapter 1
DNA Microarray Technology and Clinical Research

1.1   Introduction

Generally speaking, one may take either a genome-scale experimental or a conventional observational approach to the study of disease at the transcriptional level. A conventional observational study often observes the activities of one gene at any one time, whereas a genome-scale experiment involved planned intervention on all genes suspected of causing the disease under investigation. Nonetheless, the objective of both conventional observational and genome-scale experimental studies is the elucidation of cause-and-effect relationships.

DNA microarrays or gene chips are prominent among the genome-scale experimental technologies [Aitman, 2001]. During the past decade, there has been a dramatic rise in the use of DNA microarrays for clinical research. For almost any disease, clinical investigators can, in principal, start by searching for its genetic roots in the human genome [Wade 2001]. Such progress is a far cry from the conventional observational method of medical discovery with its heavy dependence on luck and inspiration [Wade, 2001]. In the era of post-genomics, clinical investigators should be less dependent on happy accidents of the type that accompanied Alexander Fleming's discovery of penicillin [Fleming, 1944; Wade, 2001]. But instead of waiting for another Fleming, clinical investigators can scan the genome of a pathogen with the aid of DNA microarrays, searching for weak points in their resistance where a drug or vaccine might be brought to bear [Wade, 2001]. For any human disease the DNA microarray technology provides a powerful new starting point and a

means of accelerating clinical research leads from elsewhere.

One of the principal features of DNA microarrays it the mass of numbers, which amount to hundreds of data points from thousands or tens of thousands of genes they produce. As a consequence, the major challenge in the discipline is to analyse and make sense of this immense amount of biological information. The discipline of bioinformatics has come to mean the applications of mathematics, statistics, and information technology in the biological science, and the bioinformatics of microarray is the answer to that challenge [Skeletal, 2001]. As a rule, it is not possible to accomplish a profound microarray experiment without bioinformatics taking part.

Even though there are many proposed techniques in bioinformatics for the analysis of microarray gene expression data, none of these techniques can lead to the discovery of the biology underlying cell function [Fielden and Zacharewski , 2001; Khatri and Drăghici, 2005; Hanai et al., 2006; Werner, 2007]. Particularly, changes in transcriptional levels are assessed by some parametric or non-parametric tests on an individual basis, essentially resulting in long lists of genes that are thought to have significantly changed transcriptional levels [Werner, 2007]. On the other hand, in biology these changes do not occur as independent events as the lists suggest, but in a complex and interdependent manner [Werner, 2007]. To address this problem, this thesis provides better fundamental bioinformatics techniques to account for the biological complexities and interdependencies in the analysis of disease-related microarray gene expression data.

This chapter sets out the fundamentals of DNA microarray technology and research strategies in clinical research. The rest of the chapter is arranged as follows: Section 1.2 described what DNA microarrays are. The methods by which DNA microarrays are produced as described in Section 1.3, whose expression profiling enables us to study the cause of disease (Section 1.4). The concept of clinical research designs and choosing among research strategies are discussed in Sections 1.5, 1.6 and 1.7.

1.2 DNA Microarrays Technologies

DNA microarrays are solid supports onto which the DNA or RNA sequences from thousands or tens of thousands of distinct genes are attached at fixed locations. The supports themselves are glass microscope slides, the size of two side-by-side pinky fingers but can also silicon chips or nylon membranes [Skeletal, 2003]. The DNA is printed, spotted or actually synthesised directly on the glass slide [Skeletal, 2003].

For the past three decades, the standard techniques used to detect specific sequences of DNA or RNA has depended on the use of a DNA probe [Aitman, 2001]. A single-stranded DNA molecule with a known sequence is labeled with a radioactive isotope or a fluorescent tag and then used as a "probe" to detect a fragment of DNA or messenger RNA (mRNA) with the complementary sequence [Aitman, 2001]. For example, if a scientist wanted to know whether gene X is differentially expressed in a particular tissue, he or she would make a radio-labeled DNA probe by using a small piece of gene X, isolate mRNA from the tissue, bind the mRNA to a solid medium and then hybridise the probe to the filter. If gene X is differentially expressed in the tissue, this scientist will see a radioactive signal on the filter.

This procedure is known as northern blot analysis [Alwine et al., 1977]. DNA microarrays have enabled us to do thousands or tens of thousands of northern blot analysis at the same time. DNA microarrays use the same DNA probe detection method but on a much larger scale. Instead of studying one gene at a time, microarrays allow thousands of specific DNA or RNA sequences to be detected at the same time on a small glass or silica slide [Aitman, 2001]. While the principals of specific DNA or RNA detection have remained unchanged, the notably increased scale on which this can be achieved with DNA microarrays has made it possible to examine quantitatively different statistical hypotheses at the same time.

1.3 Making Microarray Platforms

Figure 1.2 Gene expression analysis using an oligonucleotide microarray. Up to half a million distinct oligonucleotides are synthesised on the microarray by photolithography and act as probes in individual "features" on the microarray surface [Brown and Botstein, 1999; Aitman, 2001]. About 30 distinct oligonucleotides, printed as individual features, represent the partial sequence of one gene [Brown and Botstein, 1999; Aitman, 2001]. Fluorescent labelled cDNA derived from a single test sample is hybridised to the microarray, allowing the expression level of up to 15,000 genes to be measured in the test sample [Brown and Botstein, 1999; Aitman, 2001].

In spotted microarrays, cDNA, oligonucleotides, or small fragments of the polymerase chain reaction (PCR) products at the level of particular genes are spotted on the gene chip [Aitman, 2001]. A robot typically carries out the procedure and one or more probes can be used for each gene [Aitman, 2001]. Unlike RNA sin situ oligonucleotide arrays, spotted arrays are "customizable". The user can choose the probes to be spotted consistent with specific experimental needs [Aitman, 2001]. While the technology for manufacturing spotted microarray platforms is now widely accessible, a major disadvantage of this technique is that consistency of spotted and reliable annotation of the DNA on the microarray platforms is hard to achieve [Aitman, 2001].

The photolithography technique, originally borrowed from the semiconductor industry, is limited to a small number of

manufacturers, of whom the best known is Affymetrix, a California Silicon Valley-based biotechnology company that has cornered a large share of the microarray market [Aitman, 2001]. The very high spotted density possible with photolithography (Figure 1.2) means that the expression of up to 15,000 genes can be measured on a single slide (the signal for each being derived from about 30 spots) [Aitman, 2001]. While these commercial arrays have high unit cost, the consistency between arrays is assured by the production process, and microarrays representing most of the genes in the human genome and a range of other genomes can be bought ready for use without the need for development by the user [Aitman, 2001].

1.4 Microarray Gene Expression Profiling

One of the principal applications of DNA microarrays is gene expression profiling, in which RNA is isolated from a tissue or cell. The result reveals the expression levels of mRNA from thousands or tens of thousands of genes, in a tissue or cell. It enables the investigation of a complete cell at work at any one time. Through global analysis of gene expression profiling, the function of genes previously identified only by their DNA or RNA sequence is being discovered almost as a matter of routine [Aitman, 2001]. Furthermore, clinical investigators hope to understand and explain the pathogenic mechanisms of almost any kind of diseases at the genetic level and, in time, to develop genome-based treatments or cures.

For individual patients, more precise diagnosis and risk assessment based on gene expression profiling are achievable for almost any kind of diseases, leading to more precise determination of prognosis and more individualised medicine [Aitman, 2001]. In individualised medicine, as it has been called, more and more drugs will be packaged with diagnostic tests designed to identify the patients who will respond best and to screen out those likely to suffer unwanted adverse effects [Wade, 2001]. And as the genetic basis for more diseases is understood, it will be easy to devise tests for susceptibility to each, although treatments or cures will take longer to develop

[Wade, 2001]. Many diagnostic tests could be useful even without an accompanying treatment, if they point to some preventive change in lifestyle [Wade, 2001]. Knowledge of a genetic susceptibility to type 2 diabetes, for example, might help an obese person pay particular attention to health, devise a diet and exercise program.

In a minor way, genome-based diagnostic testing has already begun, although for only a handful of diseases [Wade, 2001]. When a newborn in the United States and Japan is a tender twenty-four hours old, it gets a first taste of medical procedures [Wade, 2001]. A nurse jabs the baby and a drop of blood falls onto a piece of filter paper [Wade, 2001]. The blood is then tested, principally for phenylketonuria, a metabolic disorder caused by a deficiency of the enzyme phenylalanine hydroxylase, which affects 1 in 10,000 live births and leads to mental retardation unless prevented by a special diet [Wade, 2001].

On the other hand, though gene expression profiling has already been proven to be clinically useful in the diagnosis of cancer, most genome-based diagnostic tests are still in the development stages [Aitman, 2001]. Presently, for instance, the procedure used to classify pediatric acute lymphoblastic leukaemia (ALL) subtypes using DNA microarrays is to select a number of diagnostic discriminating genes of the various leukaemia subgroups using Chi-square statistics, and there are marked differences in the number of diagnostic diagnostic discriminating genes for the various leukemia subtypes [Ross et al, 2003; Ross et al, 2004]. This simple technique does not allow the use of genome-based expression profiling as a single standardised microarray platform for diagnosis of the known prognostic subgroups of pediatric acute lymphoblastic leukaemia in clinical practice, since it can be extremely laborious, time consuming and expensive. In addition, such procedures do not address the full potential of genome-scale experiments to alter our understanding of cellular biology by giving, through an inclusive analysis of the entire repertoire of transcripts, a continuing comprehensive window into the biology responsible for the clinical differences among these leukemia subgroups. To direct support to the use of genome-based diagnostic testing as a single standardised microarray platform for the diagnosis

of pediatric acute lymphoblastic leukaemia in clinical practice, we have put forward a better fundamental analysis for transcriptional responses of bone marrow samples from pediatric patients with ALL (Chapter 5, Section 5.1).

1.5 Study Designs

In microarray data analysis of gene expression the expressed genes are referred to as variables, and the values of the measurement recorded for them are referred to as microarray gene expression data. Generally speaking, disease-related data from microarray-based studies may come from different sources, with the two fundamental designs being retrospective and prospective.

1.5.1 Retrospective Studies

Retrospective studies, also called case-control studies, are aetiological investigations in which comparisons are made between individuals who have a particular disease or physiological condition, known as cases, and individuals who do not have the disease or physiological condition, known as controls [Collett, 1991]. At time a distinction is made between the sources from which cases and controls are chosen. In a population based case-control study, all cases of the disease under investigation taking place within a defined geographic area during a specific period of time are ascertained, often through a disease registry or hospital network [Schlesselman, 1982]. The entire case series or a random sample of it is chosen for study, whereas controls are chosen by taking a probability sample of individuals free of the disease being studied in the geographic area from which the case arose [Schlesselman, 1982]. In a hospital-based case-control study, all cases of the disease admitted to a single *hospital or network of hospitals are ascertained during a specified period of time [Schlesselman, 1982]. The entire case series of a* random sample is selected for study. In the hospitals from which the case arose, controls are selected from persons admitted for conditions other than the disease.

In microarray-based clinical studies, gene expression profiles from selected cases and controls are compared to determine differences, if any, in relation to the occurrence of disease. The advantages of a case-control study are that it is relatively inexpensive and it is possible to obtain answers to research questions relatively quickly because the cases are already available.

1.5.2 Prospective Studies

Prospective studies, also called cohort studies, a group of persons of comparable health are enrolled and followed over a certain period to observe the time at which a disease or physiological condition develops. Given the time-to-event data, the scientist then seeks to determine whether there is a statistical relation between exposure and the disease [Le, 1997]. Another form of the prospective study consists of clinical trials [Le, 1997]. A clinical trial is a controlled experiment that often seeks to determine the safety and efficacy of a new drug or device. A controlled clinical trial (Figure 1.3) consists of two periods:

1. During the enrollment period, the time interval $(0,t_1)$, volunteer enter the study sequentially and are randomized to receive either a potentially therapeutic agent or a placebo (or the standard therapeutic treatment) [Le, 1997].
2. Recruited volunteers are then followed to the time of the primary ending event, or to time t_2, where $t_2 \geq t_1$; the time interval (t_1,t_2) is called the follow-up period [Le, 1997]. As an example, a clinical trial may consist of 2 year of enrolment and 2 years of follow-up; no patients are enrolled during the last 2 years.

No new volunteer recruited after
this time point

t_2

Enrollment period

Follow-up period

Study initiation

t_1

Study termination

Figure 1.3 A clinical trial:

1. During the enrollment period, the time interval $(0, t_1)$, volunteer enter the study sequentially and are randomized to receive either a potentially therapeutic agent or a placebo (or the standard therapeutic treatment) [Le, 1997].

2. Recruited volunteers are then followed to the time of the primary ending event, or to

time t_2, where $t_2 \geq t_1$; the time interval (t_1, t_2) is called the follow-up period [Le, 1997]. As an example, a clinical trial may consist of 2 year of enrolment and 2 years of follow-up; no patients are enrolled during the last 2 years.
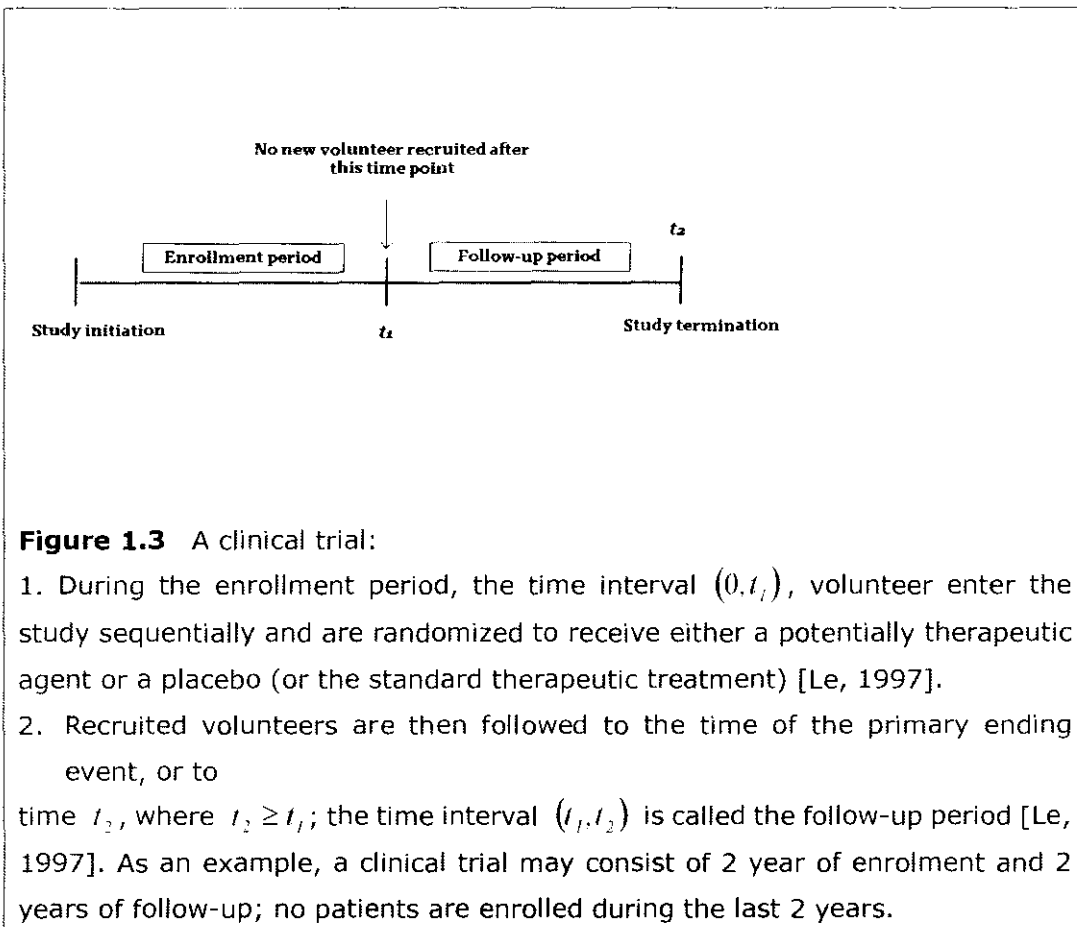
1.6 Exploratory Study

In microarray-based studies of global gene expression, the case-control method of investigation is nonetheless often the research strategy of choice, particularly when initiating an explanatory study of disease etiology. So far, since we have very little a priori knowledge the exact mechanisms of the disease under investigation at the transcriptional level, exploratory study, colloquially called a "fishing expedition", is the one in which multiple hypotheses are proposed for scientific investigation with the help of microarray gene expression profiles from selected cases and controls. The purpose of such studies is to learn enough about possible causes of the disease in question so that one or more specific hypotheses may be proposed and be sufficiently supported to justify a detailed investigation [Schlesselman, 1982].

An example of an exploratory case-control study was one conducted by Dr. Ling-Hong Tseng and his coauthors (including me), as to the etiology of pelvic organ prolapse (POP) taking place in a certain population of women corresponding to non-prolapse controls [Tseng, 2009]. At the start of this investigation, we did not know the cause of POP since basic science research into its pathogenesis had been very limited [Alperin and Moalli, 2006]. In this research, global gene expression profiles were studied in pelvic connective tissue from women with POP corresponding to non-prolapse controls. The round and uterosacral ligaments were removed from each women, both cases and controls, at the time of a laparoscopic hysterosctomy. Total RNA was then extracted and all labeled samples were hybridised to Human Genome Survey Microarray Version 2.0 (Applied Biosystem, CA, USA). A specific hypothesis which emerged from the finding was that defective mitochondrial translation caused by ribosomal proteins was the molecular etiology of POP [Tseng, 2009].

Irrespective of the terminology employed, the characteristics of exploratory case-control studies are similar to: (1) data concerning the distribution of potentially important characteristics among cases and controls are unavailable, and knowledge of the natural history of the disease is insufficient to justify study of a specific hypothesis; (2) an attempt is made to gather data concerning possible differences on a variety of factors among the cases as compared to the controls in the hope that etiologic clues leading to further study will emerge [Schlesselman, 1982].

While some clinical investigators have disparaged explanatory studied, they are usually the logical first step to understand disease etiology and are to be favoured over a study restricted to a specific hypothesis that is speculative or weakly supported [Schlesselman, 1982]. For example, Dr. Bertha Chen and his coauthors in 2005 reported that elastin metabolism was the etiology of POP resulting in urinary stress incontinence [Chen et al., 2004; Wen et al., 2006; Wen et al., 2007]. But their attempts remained undetermined for two reasons: First, they examined global gene expression profiling of vaginal tissues from women with POP, without examining their connective tissue samples collecting from supportive ligaments (e.g. utero-sacral,

round and uterosacral ligaments). Next, they hypothesised that elastic metabolism is the cause of POP, and chose to examine genes involved in extracellular matrix remodeling only. However, another explanatory study of global gene expression profiles in the round and uterosacral ligaments from women with POP suggests that defective mitochondrial translation caused by ribosomal proteins was the molecular etiology of POP [Tseng, 2009]. Such a study would have been premature and would have diverted energies and research funds from the completion of a comprehensive descriptive investigation.

Unlike an explanatory studies, an analytical study is an investigation designs to test a specific hypothesis concerning the cause of the disease under investigation [Schlesselman, 1982]. The hypothesis must be stated in an explicit way that permits definitive and unbiased assessment. "Obesity causes type 2 diabetes", for example, is a hypothesis of a different degree of clarity than a hypothesis stated as follows: "Obesity is associated with type 2 diabetes". Compared with obese persons who are free of type 2 diabetes, patients with type 2 diabetes will be obese with the progress of hyperglycemia.

1.7 Choosing Among Research Strategies

With the help of the DNA microarrays that monitor the simultaneous activity of thousands or tens of thousands of genes in a cell, it becomes achievable to examine a complete cell at work. A microarray-based experimental method is universally believed to be the most powerful explanatory method, and it is available to most clinical researchers. A patient's past and current biochemical status can be assessed using a single gene chip. In microarray-based clinical research, the choice of research strategies is limited to the cohort or the case-control approach, whereas the case-control method of investigation is usually the research strategy of choice. Furthermore, the exploitation of disease-related global gene expression profiling strengths a case-control study's capability of detecting the underlying etiology.

The case-control method is especially useful for the study of rare diseases [Schlesselman, 1982]. In this situation, a cohort study is extremely inefficient since virtually all of the effort is devoted to the follow-up of individuals who remain free of the disease under investigation [Schlesselman, 1982]. The case-control method is also most appropriate for studying diseases with long latency [Schlesselman, 1982]. A clinical investigator adopting the case-control method studies patients who have already developed the disease under investigation has no need to wait for time to elapse between the possible cause and the appearance of disease symptoms. One can start to search for cases without delay. In other words, a case-control study will almost be less costly and less difficult to mount and complete than a current cohort study. The speed with which the case-control study can be accomplished and its smaller required sample size enables a reduction in costs for gene chips and microarray data processing.

Nonetheless, the case-control method is not suited to the assessment of therapy of prophylaxis of disease. Compared with non-diseased individuals, cases of the study disease would necessarily be expect to have the higher rate of exposure to drugs or other factors than are used for treatment (See Section 1.5.2).

Chapter 2

Analysis of Variant Genes

2.1 Introduction

The main objective of microarray-based clinical studies is intervention on variant genes which contribute to the biological mechanisms of the disease under investigation. Furthermore, if an alternation in a biological process is followed a change in the character of a disease, one may regard this biological process to be a cause. An exposure to a biological process may increase, decrease or have no effect on the onset of disease, and a case-control study is equally well suited to assess any one of these possibilities.

This chapter presents some methods for determining the effect of exposure to variant genes with the disease under investigation, and points out the manner in which an assessment can be made on the basis of a case-control study. The problem of interpreting an apparent effect of exposure to variant genes, in consideration of potential bias from extraneous genetic or environmental factors is addressed in a discussion of confounding. To further characterise the biological process for variant genes associated with the disease under investigation, Entrez Gene [Maglott et al., 2007] is used for the retrieval of gene-specific information and structuring of biological knowledge [Maglott et al., 2007].

2.2 Measure of Association

The concept of association refers to a dependence, which may or may not be casual, between two or more variables [Schlesselman, 1982]. A statistical explication is made in terms of various measures of

association, the correlation coefficient being a typical example. The odds ratio is one other.

Consider two groups of subjects comparable in all respects relevant to the onset of disease, apart from presence or absence of exposure to some variant genes at a specific point or time t_0. Suppose that during a given period of time, from t_0 to t_1, one observes the number of individuals who develop disease and who remain disease free. In biostatistics, if a disease occurs with probability p, then the ratio $p/(1-p)$ is called the odds [Schlesselman, 1982]. If p_1 denotes the rate of disease occurrence among subjects expose to certain variant genes, the odds of disease are $p_1/(1-p_1)$. Similarly, if p_2 denotes the ratio of disease occurrence among subjects unexposed to certain variant genes, the corresponding odds of disease are $p_2/(1-p_2)$. The ratio of the odds of disease in exposed subjects relative to unexposed subjects is called the odds ratio [Collett, 1991]. Using the notation OR, the odds ratio may be written as

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} \tag{2.1}$$

The odds ratio, OR, can be defined in terms of the odds of disease in exposed subjects relative to the odds of disease in the unexposed subjects. In particular, if a disease is rare, p_1 and p_2 are both small and the odds ratio, OR, the odds ratio is approximately equal to p_1/p_2 [Collett, 1991]. To prove this, consider the odds of disease for a subject in the exposed group, $p_1/(1-p_1)$. This can be written as

$$p_1(1-p_1)^{-1} = p_1(1 + p_1 + p_1^2 + ...) \tag{2.2}$$

When p_i is small, p_i^2 and higher-order powers of p_i will be

negligible and therefore $p_i(1 - p_i)^{-1}$ will be approximately equal to p_i

[Collett, 1991]. As a consequence, when the disease is rare, this approximation is likely to be valid in most studies of the association between disease occurrence and variant genes. Furthermore, the odds ratio can easily be computer after fitting a linear logistic regression model leading to the identification of variant genes associated with the disease under investigation, as we will see in Section 3.6.

2.3 Confounding

In biostatistics, the term confounding refers to the effect of an extraneous variable that wholly or partially accounts for the apparent effect of the study exposure that masks an underlying true association [Schlesselman, 1982]. As a result, an apparent association between the exposure and disease may essentially due to another variable [Schlesselman, 1982]. On the other hand, the apparent lack of an association could result from failure of control for the effect of some other factors [Schlesselman, 1982]. To take a simple example, suppose that one wanted to investigate by means of a case-control study the relationship between insulin resistance and type 2 diabetes. In a comparison of cases and controls, one might find that the case group contained a greater proportion of the activation of complement cascade than the control groups since the complement system is known to be a fundamental component of innate immunity [Schifferli, 2005]. At the same time, since the activation of complement cascade may be correlated with insulin resistance. Thus, an apparent increased risk of type 2 diabetes found to be associated with the activation of complement cascade in fact might be due to insulin resistance. As a rule, a confounder's association with the disease under investigation may be either case-and-effect of a non-casual relation resulting from the confounder's association with casual factors other than the study exposure [Schlesselman, 1982].

In microarray-based clinical studies of gene expression, a confounder or confounding gene can be defined as an extraneous gene that satisfies both of two conditions: (1) It is a variant gene for the study disease, and (2) it is associated with another variant gene that is not a direct consequence of that variant gene. For example, if gene A is contributed to disease D, the confounding gene B may not be solely caused by gene A, and gene B shall not always lead to disease D. Figure 2.1 shows path diagram that illustrates in which a gene B is or is not a confounder for an association between gene A and disease D.
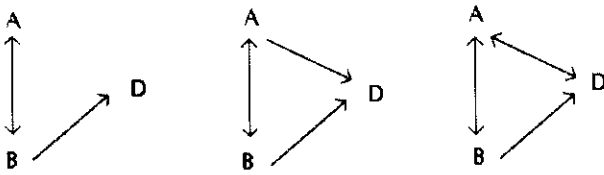
Figure 2.1 (a) Situations in which gene B is a confounder for an association between gene A and disease D.
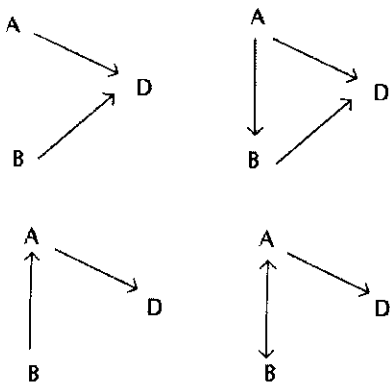
Figure 2.1 (b) Situations in which gene B is not a confounder for an association between gene A and disease D.

In planning a case-control study, one should regard any known variant genes for the study disease as a potential confounder. The

study design and analysis should be planned to either assess or eliminate the effects of confounder [Schlesselman, 1982]. Genes that are not differentially expressed for the study disease may nonetheless be confounders as the result of unwitting selection biased introduced by the procedures used to obtain cases and controls [Schlesselman, 1982]. A variant gene that is a confounder in the sample may be dealt with by adjusting procedures that rely upon matching (Section 2.5), linear logistic regression model (Section 2.6), conditional or fixed effect logistic regression model (Section 2.7), and cluster analysis with the Mahalanobis distance as a distance measure (Chapter 3), which gives straightforward ways to account for the confounding.

Generally speaking, the confounding effects is a major threat to the validity of inferences made about cause and effect, i.e. internal validity, are the observed effects should be attributed to the confounder. The consequence of ignoring confounders in the analysis of microarray gene expression data is that incorrect molecular signatures underlying the study disease can be obtained.

2.4 Source of Bias

Microarray based case-control studies are vulnerable to misleading association arising from the circumstances in which cases and controls are ascertained and selected for study. Specifically, characteristics or exposures associated with differential diagnosis or selection of subjects can lead to biased estimates of odds ratio [Schlesselman, 1982].

The validity of diagnosis is a major source in which bias may occur. For example, the diagnosis of cervical dysplasia and carcinoma is based on both the objective assessment of well-defined microscopic features and on the more subjective analysis of often subtle parameters to result in a final diagnosis [Jarcoe et al., 2002]. The routine practice of cervical cytology is therefore limited by problems of false negative diagnoses as well as by poor specificity for clinically significant lesions in patients with low-grade cytological abnormalities [Jarcoe et al., 2002].

Differential selection is a further source of potential bias. For example, in a case-control study of complications of tension-free vaginal tape (TVT), some interviews were found to be "keying" on cases who were exposed to the study factor. One nurse in particular was searching out all cases of TVT patients. If all TVT patients are ultimately selected, this causes no problem. If only a fraction of TVT patients are selected, however, the sample is biased toward overexposure among the cases. To avoid such a problem, one must establish precisely and in advance the method by which cases and controls are identified and selected [Mantel & Haenszel, 1959]. One must also carefully train staff to do the field work properly and establish a system of quality control to maintain standards throughout the investigation [Schlesselman, 1982].

2.5 Matching

Matching refers to the pairing of one or more controls to each case on the basis of their "similarity" with respect to selected criteria [[Schlesselman, 1982]. Any characteristics or attribute of an individual, such as age, gender, race, disease stage, personal or family history of disease, may serve as a foundation for pairing. Furthermore, if the cases and controls differ in relation to some variant genes, suggesting an association with the study disease, then the association cannot be explained in terms of case-control differences on the matching variables.

A design with M controls per N cases is known as a $N:M$ matched study, and the subjects that constitute the N cases and the M controls to which the case has been matched as referred to as a matched set [Collett, 1991]. Although the objective of matching is the elimination of biased comparisons between cases and controls, this objective can only be accompanied by an analysis that corresponds to the matched design [Schlesselman, 1982]. Unless the analysis properly account for the matching used in the selection phases of a case-control, the estimated statistic associated with the study disease can be biased as a result of matching [Schlesselman, 1982]. Thus,

matching is only the first of a two-step process that can be used to control for confounding: (1) matched design, followed by (2) matched analysis [Schlesselman, 1982].

While the ultimate objective of matching may be the elimination of biased comparisons, the immediate consequence of matching is the achievement of "balance" in the number of cases and controls that occur at each level of the matching variables [Collett, 1991]. For example, if one pairs on gender, then equal numbers of cases and controls will happen within any gender group. If one pairs on both gender and age, then cases and controls will be balanced on gender within each age group. Similarly, within any specified gender group, cases and controls will be balanced on age.

In microarray-based clinical studies, matching may be used in special circumstance to rule out some particular factor in a postulate casual pathway between exposure and distance. For example, suppose that effect of leptin on type 2 diabetes is mediated by an elevation of serum glucose levels. Then, one might deliberately match on glucose to see if the association of leptin with type 2 diabetes is eliminated or diminished.

2.6 Linear Logistic Regression Model

The first applications of linear logistic regression model to data from clinical research appeared during the early 1960s and were the result of work by Cornfield and his colleagues on the analysis from a study about coronary heart disease [Collett, 1991]. For a number of reasons, this method can be used in the analysis of disease-related microarray gene expression data. First, it gives a methodical way to exploring the relationship between the probability of disease occurrence and some variant genes. Next, it is straightforward to ascertain whether or not there are interactions between variant genes and confounders and estimates of the odds ratio of disease occurrence can be easily obtained from the fitted model.

In a case-control study, the cases are selected from a population or

hospital-network of individuals who have the disease that is being examined, whereas the controls are selected from a corresponding population or hospital-network of individuals without that disease. Information on their exposure to particular variant genes is then sought retrospectively. If the variable that represents whether or not a particular individual is a case in the study is regarded as a binary response variable, and then variant genes are treated as explanatory variables, linear logistic regression model can be used to identify variant genes associated with the study disease. Such model can be used to extent of any association between variant genes and a disease of interest.

In a microarray based case-control study of gene expression, let θ_1 represent the proportion of cases that have been sampled from the disease population, so that θ_1 is the sampling fraction for the cases. The value of θ_1 can be thought as the probability that a disease individual is chosen for inclusion in the study, that is

$\theta_1 = P$ (an individual is in the study | with disease)

Similarly, let θ_2 be the proportion of subjects is the disease-free population who from the controls, so that θ_2 is the sampling fraction for the controls, and

$\theta_2 = P$ (an individual is in the study | without disease)

The probability that an individual in the population develops the disease will depend on the values of k explanatory variables corresponding to levels of gene expression, $x_1, \ldots, x_k$, of k genes. For this reason, the probability that a particular individual has the disease is denoted by $\pi(\mathbf{x})$, where $\mathbf{x} = [x_1, \ldots, x_k]^{\mathrm{T}}$.

If an individual in the case-control study has the disease, this conditional probability is given by

$\pi_0(\mathbf{x}) = \theta_1 \, \pi(\mathbf{x}) \, P$ (with disease| individual is in the study)

To obtain this conditional probability, we use a standard result from probability theory, according to which the probability of an event A, conditional on an event B, is given by $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ [Collett, 1991]. Similarly, $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$. Combining these two formulas we have $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$, which is the standard form of the Bayes Theorem [Collett, 1991]. Using the result

$$\pi(x) = \theta_1 \, \pi_0(x) / P \text{ (individual is diseased)}$$

Subsequently, an individual is included in the study if that individual is diseased and has been selected as a case, or if that person is disease-free and has been selected as a control. As a result, P (individual is in the study) is the sum of P (individual is diseased and is a case) and P (individual is disease-free and is a control). Using the result $P(A \cap B) = P(A|B)P(B)$, P (individual is diseased and is in the study) is equal to

$$P \text{ (individual is in the study | diseased)} \times P \text{ (individual is diseased)}$$

which is $\theta_1 \, \pi_0(x)$ as before. The probability that an individual does not have the disease is $1 - \pi_0(x)$ and therefore P (individual is disease-free and is in the study) is $\theta_2(1 - \pi_0(x))$. Therefore,

$$P \text{ (individual is in the study)} = \theta_1 \, \pi_0(x) + \theta_2(1 - \pi_0(x))$$

and so

$$\pi(x) = \frac{\theta_1 \pi_0(x)}{\theta_1 \pi_0(x) + \theta_2(1 - \pi_0(x))} \qquad (2.3)$$

From this result,

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\theta_1}{\theta_2} \frac{\pi_{_0}(x)}{1-\pi_{_0}(x)} \tag{2.4}$$

and it then follows that

$$\text{logit}[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \log\frac{\theta_1}{\theta_2} + \log\frac{\pi_{_0}(x)}{1-\pi_{_0}(x)} \tag{2.5}$$

If a linear logistic regression model is adopted for the probability that a subject in the population with the levels of gene expression, $x = [x_1, \ldots x_k]^{\text{T}}$, of k genes has the disease, then

$$\text{logit}[\pi_{_0}(x)] = \alpha + \beta_1 x_1 + \ldots \beta_k x_k \tag{2.6}$$

where α and β_i's are regression coefficients. So the probability that

a subject with explanatory variables x is a case in the case-control study is given by

$$\text{logit}[\pi(x)] = \log\frac{\theta_1}{\theta_2} \alpha + \beta_1 x_1 + \ldots \beta_k x_k \tag{2.7}$$

which, since θ_1 and θ_2 do not depend on x, can be written as

$$\text{logit}[\pi(x)] = \alpha^* + \beta_1 x_1 + \ldots \beta_k x_k \tag{2.8}$$

where

$$\alpha^* = \alpha + \log\frac{\theta_1}{\theta_2} \tag{2.9}$$

This model is a linear logistic regression model for the probability that

an individual is a case in the case-control study. To fit this model, a binary response variable is defined that takes the values on when an individual is case, and zero when an individual is a control. On fitting this mode, one could estimate $\pi_{_{ij}}(x)$, the probability that an individual

with altered expression of genes x has the disease. From equation (2.6), this probability depends upon an estimate of α , which from equation (2.9) depends on estimates of α^{*}, $\theta_{_l}$ and $\theta_{_2}$. Though α^{*} can be estimated, the two sampling fractions, $\theta_{_l}$ and $\theta_{_2}$, cannot be estimated unless information is available on the number of diseased and diseased-free persons in the population being sampled. Such information is usually unavailable, and so in general it will not be possible to estimate α^{*}.

Similarly, the odds of disease for a subject with exposure to differentially expressed genes $x_{_1}$ relative to someone whose values are $x_{_0}$ is given by

$$OR = \frac{\pi_{_{ij}}(x_{_1}) / (1 - \pi_{_{ij}}(x_{_1}))}{\pi_{_{ij}}(x_{_0}) / (1 - \pi_{_{ij}}(x_{_0}))}$$

(2.10)

From equation (2.7),

$$\frac{\pi_{_{ij}}(x_{_j})}{1 - \pi_{_{ij}}(x_{_j})} = \frac{\theta_{_2}}{\theta_{_l}} \frac{\pi(x_{_j})}{1 - \pi(x_{_j})}$$

(2.11)
for $j = 0,1$, and so

$$OR = \frac{\pi(x_{_1}) / (1 - \pi(x_{_1}))}{\pi(x_{_0}) / (1 - \pi(x_{_0}))}$$

(2.12)

The odds ratio is independent of α , and so it can be estimated from microarray gene expression profiles from a case-control study. In practice, we can find the maximum likelihood estimators for α and the estimated odds ratio using a standard statistical package, such as SPSS, SAS, STATA, Splus or R. As we will see in Section 4.2, we have used this method to examine microarray gene expression profiling of skeletal muscle response to endurance exercise training.

2.7 Conditional or Fixed Effect Logistic Regression Model

In section 2.4, we explained how a matched case-control study can be used for the purpose of controlling the effects of confounding variables. As in modeling microarray gene expression data from an unmatched case-control study, the probability that an individual is diseased will be assumed to depend on the expression levels of p genes that have been measured on that individual with the help of gene chips. The expression levels of p genes may represent risk genes and those potential confounders that have not been used in the matching process. In addition, the probability that a particular individual is diseased may also depend on the values of the matching variables that define the matching set in which an individual occurs [Collett, 1991]. To take the matching into consideration in the analysis of variant genes from a matched case-control study, a conditional likelihood or fixed effect likelihood is constructed. The augmentation used to construct the relevant likelihood is:

Assume a $1:M$ matched case-control study, where the t-th matched

set contains one case and M controls. Let D and $\bar{D}$ denote the

presence and absence of the disease under investigation, respectively. Assume that the presence of the disease under investigation when comparing with controls depend on the expression levels, $x_{1t},\ldots,x_{pMt}$,

of p genes. Let $\mathbf{Y}$ be the vector of joint levels of gene expression,

i.e. $\mathbf{Y} = [x_{1u},\dots,x_{p\backslash u}]'$. Let $P(D|Y_{u})$ denote the probability that the *i*-th

individual has the disease at the *t*-th matched set has the vector of

joint levels of gene expression, Y_{u}. Similarly, $P\left(\bar{D}|Y\right)= 1-P(D|Y_{u})$

denote the probability that the *i*-th individual corresponds to a control
at the *t*-th matched set has the vector of joint levels of gene

expression, Y_{u}. We consider the linear logistic regression model as

follows:

$$\log \frac{P(D|Y_{u})}{1 - P(D|Y_{u})} = \alpha_{t} + \beta_{1}x_{1u} + \dots + \beta_{p}x_{pu} \tag{2.13}$$

Therefore,

$$P(D|Y_{u}) = \frac{\exp(\alpha_{t} + \beta_{1}x_{1u} + \dots + \beta_{p}x_{pu})}{1 + \exp(\alpha_{t} + \beta_{1}x_{1u} + \dots + \beta_{p}x_{pu})} \tag{2.14}$$

Since an individual can be either a case, with the disease, or a control,

without the disease, the likelihood of $\alpha_{t}, \beta_{1},\dots,\beta_{p}$, is given by

$$L = L(\alpha_{t}, \beta_{1},\dots,\beta_{p} \mid data)$$

$$= \prod_{t-1}^{M}\binom{M}{A_{u}}P(D|Y_{u})^{t_{u}}[1 - P(D|Y_{u})]^{M-t_{u}}$$

$$= \prod_{t-1}^{M}\binom{M}{A_{u}}\frac{\exp(\alpha_{t} + \beta_{1}x_{1u} + \dots + \beta_{p}x_{pu})^{t_{u}}}{[1 + \exp(\alpha_{t} + \beta_{1}x_{1u} + \dots + \beta_{p}x_{pu})]^{M}} \tag{2.15}$$

where $\binom{M}{A_{u}}$ denotes the combination of the *i*-th subject has the

disease at the *t*-th matched set has the vector of joint levels of gene

expression, Y_{u}. The conditional log-likelihood is therefore

$$\log L = \sum_M \{(D|Y_{tt})(\alpha_t + \beta_1 X_{1tt} + \dots + \beta_p X_{ptt})$$

$$- M \log[1 + \exp(\alpha_t + \beta_1 X_{1tt} + \dots + \beta_p X_{ptt})] + \log\left(\frac{M}{D|Y_{tt}}\right)\}$$

(2.16)

We can estimate $\beta_1, \dots, \beta_p$ using a standard statistical package, such as SPSS, SAS, STATA, Splus or R. FORTRAN code for analysis data from matched case-control studies is also given by Smith et al. [1981]. And the process of estimating $\beta_1, \dots, \beta_p$ is termed as conditional or fixed effect logistic regression modeling.

The proof of the model is given to show that the effects on levels of gene expression are considered within and among various experimental criteria, respectively. To do this, let us consider the conditional probability given by equation (2.15). Let $x^*_{1tt}, \dots, x^*_{ptt}$ and $x^{**}_{1tt}, \dots, x^{**}_{ptt}$ denote the levels of the gene expression when the i-th individual with and without the disease, respectively. When the i-th individual at the t-th matched set has the disease, we have

$$\frac{P(D|Y_{tt})}{P\left(D|Y_{tt}\right)} = \alpha_t + \beta_1 x^*_{1tt} + \dots + \beta_p x^*_{ptt}$$

.

When the i-th individual at the t-th matched set does not have the disease, we have

$$\frac{P(D|Y_{tt})}{P\left(D|Y_{tt}\right)} = \alpha_t + \beta_1 X^{**}_{1tt} + \dots + \beta_p X^{**}_{ptt}$$

Then, the conditional likelihood in equation (2.15) becomes

$$\left[1 + \sum_M \exp(\beta_1(x^*_{1tt} - x^{**}_{1tt}) + \dots + \beta_p(x^*_{ptt} - x^{**}_{ptt}))\right]^{-1}$$

(2.17)

In particular, the values $x^*_{1tt}, \dots, x^*_{ptt}, \dots, x^{**}_{1tt}, \dots, x^{**}_{ptt}$ correspond to the

effects of gene expression measured within each experimental criterion. As a consequence, this conditional likelihood considers the effects of gene expression within each experimental criterion primarily; then, it judges the effects of gene expression measured among various experimental criteria. As we will see in Section 6.2.1, we have used this method to examine microarray gene expression profiling of bladder epithelial cells from patients with interstitial cystitis corresponding to their age- and gender- matched controls.

2.8 Entrez Gene Database

In sections 2.5 and 2.7, the linear and conditional logistic regression models are given to identify variant genes associated with the disease under investigation on an individual basis, essentially resulting in lists of genes that are found to have significantly changed transcriptional levels. On the other hand, in biology these changes do not occur as individual events as the lists suggest, but in a complex and interdependent manner [Werner, 2007]. In order to characterise the *biological processes for variant genes associated with the disease under investigation*, Entrez Gene [Maglott et al., 2007] is used for the retrieval of gene-specific information and structuring of biological knowledge.

What is Entrez Gene? Entrez Gene (http://www.pubmedcentral.nih.gov/redirect3.cgi?&&auth=0DkKdLs ftR-nRijlEdyHjbWpRkGcHu6apdX-9BkbT&reftype=extlink&artid=539 985&iid=17832&jid=4&FROM=Article%7CFront%20Matter&TO=Ext ernal%7CLink%7CURI&article-id=539985&journal-id=4&rendering-t ype=normal&&www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) is the gene-specific database at the National Centre for Biotechnology Information (NCBI), a division of the National Library of Medicine, located on the campus of the United States National Institute of Health (NIH) in Bethesda, MD, USA [Maglott, 2005]. It provides unique integer identifier for genes and other loci for a subset of model organisms [Maglott, 2005]. It includes up-to-date gene-specific information for genes in sequenced genomes, that have an active research community to contribute gene-specific information, or that

are scheduled for intense sequence analysis [Maglott, 2005].

Table 2.1 recapitulates the gene specific information that can be retrieved through the Entrez Gene database, how the data are revealed, and some aspects of how those data are processed [Maglott, 2005]. For example, Gene Ontology (GO) terms provide links of the functions, processes and components for all completely sequenced genes in model organisms [Maglott, 2005]. These connections are designed to provide keywords and linked to make Entrez Gene an effective starting place to retrieve information of interest [Maglott, 2005].

Table 2.1 Gene specific information that can be retrieved through Entrez Gene, how the data are revealed, and some aspects of how those data are processed [Maglott, 2005].

Subcategory	Revealed	Remarks
Nomenclature		
Gene symbols and full descriptions	Report, Table	Sources: External authorities, GenBank, Publications. 'LOC'+GeneID designation assigned if none of the above officially accepted nomenclature has precedence
Protein names	Report, Table	Often same as the gene name, but may be edited to make orthologs' names uniform
Gene structure and sequence		
Gene structure	Report, Table	Based on annotation of the Reference sequence
Reference sequences	Report, Table, Links	The accessions are shown in the report page; the sequences are retrieved from Nucleotide or Protein
Related sequences	Report, Links	Based on cDNA or protein comparison, best genomic placement and curation. Accessions are shown in the report page; sequences are retrieved from Nucleotide or Protein

Genomic position		
By sequence	Report, Links	Genomic annotation
By independent maps	Links	Shared markers or reports of cytogenetic position
Citations		
Not annotated	Links	Sources: external authorities, RefSeq curation
Annotated	Report, Links	Sources: external databases, GeneRIFs
Functional annotation		
Domain content	Report, Links	Conserved Domain Database (CDD)
GO terms	Report	GO Consortium
Pathways and interactions	Report, Links	Kyoto Encyclopedia of Genes and Genomes (KEGG)
Disease and other phenotypes	Report, Links	External authorities such as OMIM, RefSeq curation
Homology		
By Gene	Links	HomoloGene
By Protein	Links	COG
Conserved segments	Links	Map Viewer
Expression		
ESTs	Links	UniGene
External resources	Links	External resource is named and used to anchor a Link. The expression data are available at that source
Arrays	Links	GEO
Related information		
Integrated by Gene staff	Links	May be displayed on report as well as from Links menu
External sources	Links	LinkOut choice in Links menu

In addition, Entrez Gene can serve as a directory to gene-specific information for database outside of NCBI [Maglott, 2005] For instance, Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway is a collection of acknowledged biochemical pathway representing our knowledge of the molecular interactions and reaction networks for metabolisms, genetic information processing, environmental information processing, cellular processes, human diseases and drug development [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

In Chapters 3, 4, and 5, Entrez Gene and KEGG pathway are used to characterise biological processes for variant genes associated with the disease or physiological condition under investigation.

Chapter 3

Clustering of Gene Expression Profiles

3.1 Introduction

Cluster analysis is a genetic term that attempts to determine whether or not a given dataset contains distinct groups, and, if so, to determine the groups [Chen et al., 2007; Chen et al., 2007; Chen et al., 2008]. Two general references on the subjects of cluster analysis are given by Hartigan [1975] and Gordon [1981].

In microarray based clinical studies of gene expression, beyond identification of variant genes associated with the disease under investigation, another general problem facing clinical investigators is how to arrange measured gene expression profiles into biological meaningful groups, if any, according to similarity of their expression patterns. Cluster analysis is an explanatory data analysis tool for resolving this problem.

There are many possible methods of performing cluster analysis, and the relative merits of the alternative approaches have been discussed extensively [Shannon, 2003; Armstrong and van de Wiel, 2004; Wu and Dewey, 2006]. Recommendations for clustering methods have focused on the selection between supervised and unsupervised algorithms, between hierarchical and non-hierarchical methods, and among various linkage options, but inadequate consideration has been given to the substance of the method of computing the distance matrix. Given that all clustering procedures are based on some form of distance measures, the method of computing this matrix can have an important effect on the clustering results. In this chapter, the

applicability of the Mahalanobis distance as a distance measure is discussed in the context of grouping disease-related microarray gene expression profiles, whose application enables us to study the etiology of cancers and other diseases are discussed in Chapters 5 and 6.

3.2 Distance Measures

Cluster analysis was one of the first methods used to impose order on microarray gene expression data [Eisen et al., 1998]. Its purpose is to arrange genes into groups in order that the degree of association is strong between genes of the same cluster and weak between genes of different clusters. This method always starts from the assignment of a set of n genes to groups on the basis of the measurements of distance between the genes, as measured on a set of p samples. Let $d(A,B)$ denote the distance between data points A and B in the p-dimensional space, all distance measures have the following three properties:

1. symmetric, i.e. $d(A,B) = d(B,A)$;
2. non-negative, i.e. $d(A,B) \geq 0$;
3. identificatin, i.d. $d(A,A) = 0$

It is generally considered desirable for the distance measure to be a metric:

$d(A,C) \leq d(A,B) + d(B,C)$, which is called the triangular inequality,

or an ultrametric:

$d(A,C) \leq max(d(A,B), d(B,C))$

but need not be either [Venables and Ripley, 1995].

At present, cluster analysis with the corelation coefficient (e.g. Pearson correlation coefficient, Spearman correlation coefficient, and standard correlatin coefficient) as a distance measure has been commonly used in the analysis of microarray gene expression data. Although the correlation coefficient is scale invariant, it does not take

the covariance between genes into consideration. However, changes in expression levels of genes do not occur as individual events, but participate in a complex and interdependent manner due to gene co-expression and co-regulation, giving rise to the working of lively organisms and their parts. Unless the distance measure properly takes the covariance between genes into account, the organised profiling can be disturbed as a result of ignoring interdependencies.

To account for the correlation between genes, a proper distance measure can be used is the Mahalanobis distance defined as follows:

$$d^2_{ij} = (\mathbf{x}_i - \mathbf{x}_j)' \, S^{-1}(\mathbf{x}_i - \mathbf{x}_j)$$
(3.1)

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two data points, S is the $p \times p$ covariance matrix of X, and X is assumed to be of full rank so that S^{-1} exists [Mimmack, 2001]. The Mahalanobis distance incorporates covariance between gene expression levels as well as differences in variances [Mimmack et al., 2001]. It gives less weight to genes with high variance and to high correlated genes, so that all characteristics are treated as equally important [Mimmack et al., 2001]. Thus far, the Mahalanobis distance has not explicitly been commonly used in the organisation of microarray gene expression data.

3.3 Hierarchical Clustering

This section discusses the most frequently used explanatory data analysis tool for gene expression data: hierarchical clustering. This is a methodology that arranges the expression profiles into a tree-like diagram, known as a dendrogram, so that similar profiles appear close together in the dendrogram and dissimilar profiles are farther apart [Stekel, 2003].

This technique has become popular for four reasons:

1. It simplifies the mass of numbers generated by DNA microarrays.

2. The analysis exhibits clusters of genes sharing similar expression patterns.
3. The results can be viewed graphically in a hierarchical way using a tree-like dendrogram.
4. The technique is familiar to most biologists through its application in sequence and phylogenetic analysis [Stekel, 2003].

Hierarchical clustering begins from a distance matrix computed between the genes to be clustered, on the basis of their expression patterns. Furthermore, differences between the methods arise from the variety of ways in which distance between a cluster and a single gene, or between two clusters, can be defined [Everitt, 1994]. The distance between two clusters could, for example, be defined as the distance between their closest genes, leading to single linkage clustering. Another possibility is to define inter-group distance as that which is between the most remote pairs of genes, one from each cluster. This leads to complete linkage clustering. A measure that uses more than a single inter-gene distance is average linkage, where the distance between two clusters is defined as the average of the distance between all pairs of genes where members of a pair are in different clusters. Ward's linkage is distinct from all other linkage options since it uses an analysis of variance approach to calculate the distances between clusters [Tseng et al., 2009]. Briefly, this method attempts to minimize the sum of squares of any two clusters that can be formed at each step.

As we will see in Chapters 5 and 6, the Mahalanobis distance in the Ward's linkage hierarchical clustering is used to investigate the aetiology of various types of cancer and other diseases with the help of DNA microarrays.

3.4 K-means clustering

K-means clustering is described in Sneath and Sokol [1973] and differs from hierarchical clustering. Concisely, this algorithm has six steps:

1. Choose the number of clusters, denoted as k.
2. Randomly assign each gene expression profile to one of the k clusters.
3. Calculate the centroid of each of the k clusters.
4. For each profile in turn, calculate the distance between it and the centroid of each of the k clusters.
5. If that profile is closest to a cluster different from the one in which it currently belongs, move the profile to the new cluster and update the centroids of both clusters.
6. Go to step 4 and repeat until no profiles change cluster membership [Stekel, 2003].

As we will see in Chapters 5 and 6, the Mahalanobis distance in the K-means clustering algorithm is used to examine the aetiology of various types of cancer and other diseases with the help of DNA microarrays.

Chapter 4

DNA Microarrays and Endurance Training –induced Physiological Responses

4-1 Introduction

Endurance training was initially developed by Dr. Alios Mader in the 1980s, and it has had considerable influence on the training of athletes in all endurance sports, including competitive swimming [Hollman et al, 1986]. Coaches design training programs around the idea of pushing athletes to the limit of their pain tolerance and then motivating them to go beyond it [Maglischo, 2002; Chen et al,2009]. Programs had athletes swimming faster, farther, or with less rest in training than they or their competitors had ever swum before [Maglischo, 2002; Chen et al, 2009]. These training programs involve targeting each of the major phases of the energy system and other aspects of physical conditioning, such as muscle strength, flexibility and power, with specific training procedures designed to develop each to their optimum potential [Maglischo, 2002; Chen et al,2009].

The Australian Institute of Sport (AIS) is one of the major institutes for swimming in Australia and trains approximately one-third of swimmers for the Australian national swimming team [Colwin, 2002]. The AIS uses several testing protocols to assess various components of training loads in international-level swimmers, including the measurements of heart rate, oxygen consumption and blood lactate concentration [Colwin, 2002; Maglischo, 2002]. To do this, for instance, a sample of blood is collected from the ear or fingertip from each elite swimmer after each swim in order to measure the amount of lactate acid [Maglischo, 2002; Chen et al, 2009]. Such blood testing involves swimming a series of repeated at progressively faster speeds during the training of high-rate swimmers [Maglischo, 2002].

However, athletes may not like such blood testing since it can increase the possibility of carrying the AIDS, hepatitis B, hepatitis B and many other viruses from one person to another [Maglischo, 2002]. Furthermore, it would be unnecessary to do so since the quantities that are dealt with are merely the amount of lactate acid; consequently, an integrated knowledge of the athletes' physiological conditions being studied would be quite superficial. We need to develop some better fundamental tools to examine athletes' physiological conditions for endurance training.

During the past six decades, research in the sport and exercise sciences has developed at an ever-increasing rate. In the period from 1950 to the early 1970s the motivate for most scientific work was primarily at the organ and whole-body level [Colwin, 2002]. Many studies revealed the substance of body size, muscle strength, flexibility, power, cardiac and respiratory fitness, demonstrating how these attributes could be improved through athletic training [Colwin, 2002]. In the 1980s and 1990s, advancements in medical instrumentation and technology shifted focus on examining the effects of exercise and training at the cellular level [Baldwin, 2002]. This period saw a great deal of interest in using emerging medical technologies to study muscle and blood [Colwin, 2002].

The next few decades of the 21st century will see application of DNA microarray technology to sport and exercise science [Baldwin, 2002]. At present, most attention is directed toward addressing the biological complexities of cancer, asthma, diabetes, obesity, aging and other diseases. In this chapter, we target such issues as identification of genetically gifted athletes, and improved diagnosis of training loads.

The primary theme of this chapter is to show how DNA microarrays can usefully assess athletes' physiological conditions for endurance training at the transcriptional levels. As we will see in Section 4.2, linear logistic regression model is used to examine gene expression profiling of vastus lateralis muscle biopsies from subjects following a 50-minute endurance training course at 20-week before and after undergoing the endurance training course, respectively.

The main theme of this chapter is to show how DNA microarrays can usefully assess athletes' physiological conditions for endurance training at the transcriptional levels. As we will see in Section 4.2, linear logistic regression model is used to examine global gene expression profiling of vastus lateralis muscle biopsies from subjects following a 50-minute endurance training course at 20-week before and after undergoing the endurance training course, respectively.

4-2 Microarray Data Analysis

In this section, we examined microarray gene expression profiling of vastus lateralis muscle biopsies from subjects following a 50-minute endurance training course at 20-week before and after undergoing the endurance training course. The data used here was collected from Gene Expression Omnibus under platform accession no GSE1718, and this microarray experiment was performed by Teran-Garcia et al [2005]. Here we use linear logistic regression model to identify variant genes following a 50-minute endurance training course at 20-week before and after undergoing the endurance training course. This enables us to characterise 109 distinct genes that were differentially expressed in vastus lateralis biopsies from subjects following a 50-minute endurance training course at 20-week before and after undergoing the endurance training course (Table 4.1).

The results were very interesting. First, complex metabolic reactions, including carbohydrate, lipid and energy metabolisms, were found to have a strong relationship with endurance training –induced vastus lateralis muscle contraction. This muscle contraction requires the release of energy, which is made available from ribosomes. The ribosomes have sites that adapt transfer-RNA (tRNA), which identify molecular components that occur between macromolecular entities in vastus lateralis muscle cells. Next, based on the mapping of the resultant variant genes associated with endurance training –induced vastus lateralis muscle contraction onto the KEGG reference pathways of axon guidance such as regulation of actin cytoskeleton, mitogen-activated protein kinase (MAPK) signaling transduction and peroxisome-activated receptor (PPAR) signaling transduction, we are

able to identify the transcriptional regulation of endurance training –induced vastus lateralis muscle contraction status, including the level of contractile force, tissue strength, plasticity and energy homeostasis. Thus, the mapping of a gene onto the KEGG pathway mimics its actual transcriptional. A more insightful explanation is provided as follows.

Table 4.1 Variant genes in vastus lateralis muscle biopsies from subjects following a 50-minute endurance training course at 20-week after undergoing the endurance training course. Variant genes are prioritized relating to the following

criteria: the estimated odds ratio, $\hat{OR}$, greater than or equal to unity.

Gene Symbol	Gene Ontology
ELAVL1	mRNA stabilisation
CPT1B	lipid metabolism
CPT2	lipid metabolism
ACADM	lipid metabolism
ACADVL	lipid metabolism
ACAT1	lipid metabolism
ACE	metabolism
EFNB1	axon guidance
EFNB2	axon guidance
EPHB2	axon guidance
EPOR	small GTPases mediated signaling transduction
FABP3	transport
FABP4	cytokine production
ACADM	lipid metabolism
ACADVL	lipid metabolism
ACAT1	lipid metabolism
ACE	metabolism
ACTB	cell motility
ACTN1	focal adhesion formation
ACTN2	focal adhesion formation
ADAM8 (CD156)	cell-cell adhesion
ADORA1	nervous system development
ADORA2A	central nervous system development;
ADRA1B	G protein coupled receptor signaling pathway
AGTR2	G protein coupled receptor signaling pathway

ALDOB	carbohydrate metabolism
ALDOC	carbohydrate metabolism
CD34	cell adhesion
CD36	lipid metabolism
CDH5 (CD144)	cell adhesion
COL4A1	phosphate transport
COX4I1	electron transport
COX5B	electron transport;
CPT1B	lipid metabolism
CPT2	lipid metabolism
CSK	T cell activation
CYCS	caspase activation
DCI	lipid metabolism
DES	cytoskeleton biosynthesis and organization
DMD	cytoskeleton anchoring
ECH1	lipid metabolism
ECHS1	lipid metabolism
EDN3	neural crest cell migration
EFNB1	axon guidance
EFNB2	axon guidance
ELAVL1	mRNA stabilization
EPOR	small GTPases mediated signaling transduction
FABP3	regulation of cell proliferation
FABP4	cytokine production
FASN	lipid metabolism
FGFR2	amino acid phosphorylation; cell growth
FGFR4	amino acid phosphorylation
FLT1	TGF-beta receptor signaling pathway
FLT4	TGF-beta receptor signaling pathway
FN1	acute phase response
FST	TGF-beta receptor signaling pathway
GHR	skeletal development
HIF1A	neural crest cell migration
HK1	glycolysis
HPRT1	nucleotide metabolism
IGF1	Ras protein signal transduction
IGF2	skeletal development
IGF2R (CD222)	receptor-mediated endocytosis

IGFBP1	regulation of cell growth
IGFBP3	regulation of signal transduction
IL6	cell surface receptor linked signal transduction
IL6ST	cell surface receptor linked signal transduction
INSR	carbohydrate metabolism
ITGA8	cell-cell adhesion
ITGAV (CD51)	cell adhesion
ITGB1	cell adhesion
JUN	leading edge cell differentiation
K-ALPHA-1	microtubule based movement
LAMA4	metabolism
LAMB1	neurite development
LDHA	carbohydrate metabolism
LDHB	carbohydrate metabolism
LPL	lipid metabolism
MB	oxygen transport
MDH2	carbohydrate metabolism
MEF2A	muscle development
MGST3	lipid metabolism
MYF5	striated muscle contraction
MYH1	striated muscle contraction
MYH3	striated muscle contraction
MYH4	striated muscle contraction on
MYH6	striated muscle contraction
MYH7	striated muscle contraction
MYOD1	striated muscle contraction
MT-CO1	electron transport
PDGFA	actin cytoskeleton organization and biogenesis
PDGFB	actin cytoskeleton organization and biogenesis
PDHA2	glycolysis
PDK1	carbohydrate metabolism
PLAU	blood coagulation
PPARA	lipid metabolism
PPARG (NR1C3)	lipid metabolism
RB1	striated muscle differentiation
RPL13A	protein biosynthesis
RPS29	ribosome biosynthesis
SCP2	lipid metabolism

SLC27A4	lipid metabolism
SOD2	superoxide metabolism
THBS3	cell motility
TNNC1	regulation of muscle contraction
TTN	carbohydrate metabolism
VCAM1 (CD106)	cell-cell adhesion;
VCL	metabolism
VIM	cell motility

Regulation of the level of contraction force. The contraction of vastus lateralis muscle is supported by the nervous systems. Examining the mapping onto the pathways representing the regulation of axon guidance (Figure 4.1) and actin cytoskeleton regulation (Figure 4.2), respectively, indicates that developing neuron extend axons were found to be guided by Ephrin-B (EPHB1 and EPHB2) along specific pathways to regulate cytoskeleton dynamics in axonal growth cones in response to contraction stimulated by the assembly of focal adhesions and actin stress fibers in vastus lateralis muscle cells.

Figure 4.1 Reference KEGG pathway representing the regulation of axon guidance [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Figure 4.2 Reference KEGG pathway representing the regulation of actin cytoskeleton [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006;

Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Ras proteins of small GTPases domains are strong candidates to transmit guidance signals in the growth cone in response to stress. They regulate muscular cellular processes where filaments actin plays a central role. Both Ras and ATP proteins are identified as responsible for the generation of mechanical forces upon the hydrolysis, consisting with their structural similarity of the catalytic region of GTPases and ATPases domains [Kosztin et al., 2002]. The force transmission mechanisms is based on an irreversible structural change, produced by the hydrolysis, which trigger thermal switching between force-generating substates through change in the configurational space of the proteins [Kosztin et al., 2002]. Studies of vertebrate growth cone suggest common mechanisms that regulate growth cone behaviours and axon branching [Dickson, 2002]. These include reformation of actin and microtubules, affect of axon guidance factors, actions of actin regulatory proteins and dynamic changes in intracellular Ca2+ signaling. These guidance systems are contrived to encourage as astonishing varied set of neuronal circuits underlying axon guidance in response to the contraction stimulated by the assembly of focal adhesions and actin stress fibers.

Regulation of tissue strength and plasticity. The mitogen receptors are known to regulate synaptic strength and plasticity in adult nervous system via a cell-dependent manner [Dent et al., 2003]. Inspecting the mapping on the MAPK signaling pathway indicates that neutrophins signal through Trk receptor domains of platelet derived growth factor receptors and fibroblast growth factor receptors to regulate cell survival and proliferation. The whole process involved the fate of neural precursors, axon and dendrite growth and patterning, the expression and activity of functionally important proteins through vastus lateralis muscle contraction process [Dent et al., 2003]. Thus, the classical MAPK signaling transduction system regulates the strength and plasticity of vastus lateralis muscle tissues.

Regulation of energy homeostasis. PPAR is a member of the nuclear receptor superfamily, which is activated by various hydrophobic compounds [Takada and Koto, 2005]. PPAR has three

isoforms, PPARalpha, PPARgamma and PPARdelta, which regulate homeostasis, cell proliferation, cell differentiation and associated hypolipidemia, atherosclerosis, diabetes and obesity [Takahashi et al., 2005]. According to the mapping on the PPAR signaling pathway (Figure 4.3), PPARalpha and PPARgamma are activated in response to the stimulus of vastus lateralis muscle contraction. PPARalpha plays a role in the clearance of circulating in lipid metabolism in liver and skeletal muscle [Takahashi et al., 2005]. PPARgamma promotes adipocyte differentiation to enhance blood glucose uptake [Takahashi et al., 2005]. Thus, PPARalpha and PPARgamma act at crucial nodes of the regulatory network, which regulate energy homeostasis in response to the contraction stimulated by the assembly of focal adhesions and actin stress fibers.
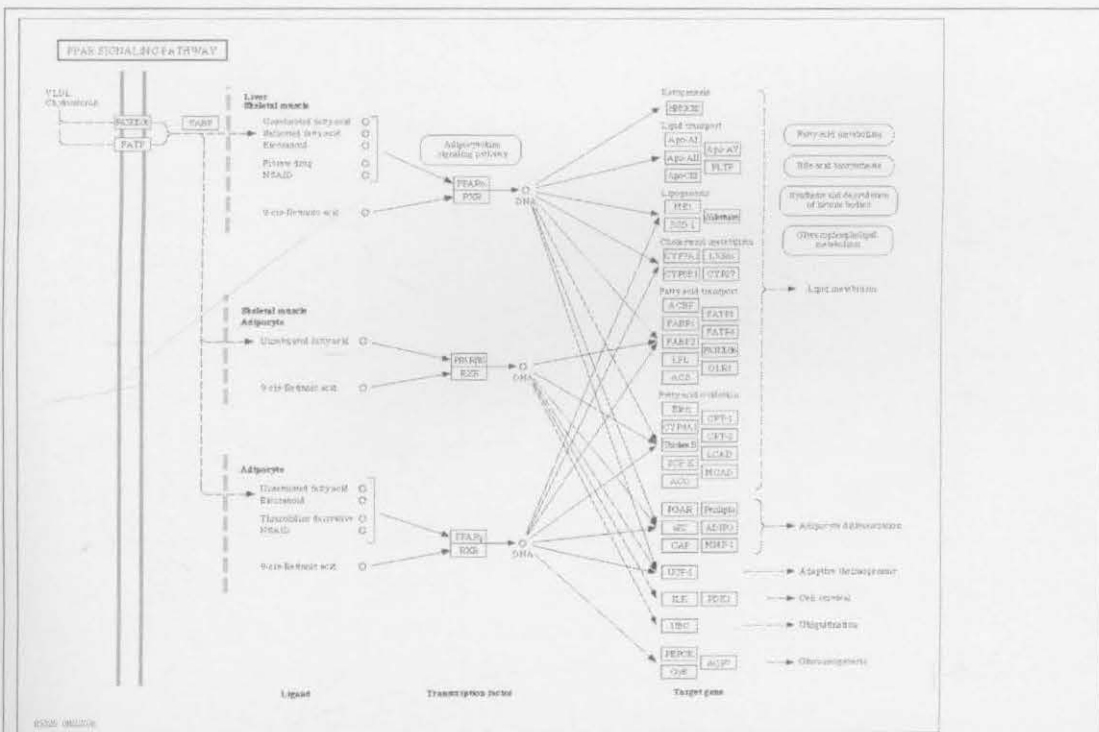
Figure 4.3 The KEGG reference pathway representing the regulation of PPAR signaling transduction [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Reactive oxygen species excess. There is an indication of the presence of reactive oxygen species (ROS) excess (oxidative stress) during endurance training –induced vastus lateralis muscle contraction status. To date, it is well known that ROS are a family of

molecules and its derivatives produced in all aerobic cells, which are yielded from the metabolisms of molecular oxygen [Waris and Ahsah, 2006]. On the basis of our present observations this endurance training course has led to the deleterious effects of oxygen from the metabolic reduction of the highly reactive and toxic species, promoting endothelial damage or dysfunction and atherosclerosis [Waris and Ahsah, 2006].

4-3 Conclusion

We very often need to examine athletes' physiological conditions for endurance training from an initial state to a final state through a succession of many intermediate states. Unfortunately, none of the physiological assessment tools in current use, including the measurements of heart rate, oxygen consumption and blood lactate concentration, can precisely assess the changes in athletes' physiological conditions for endurance training. Nonetheless, if the state of athletes' physiological conditions can be assessed with the help of DNA microarrays, our approach described here will make available the recognition of athletes' physiological conditions. To do this, a natural way is to first scan and inspect variant genes with endurance training −induced skeletal muscle response and then to analyse the details of interest. In this chapter, we first identify variant genes with endurance training −induced vastus lateralis muscle response using the linear logistic regression model. Next, we map variant genes onto the KEGG pathway to attain a linkage between key molecules and biochemical pathways with endurance training −induced skeletal muscle response in a cause-effect format. A study of the gene expression model relating to endurance training −induced skeletal muscle response exhibits the presence of carbohydrate, lipid and energy metabolisms, the transcriptional regulations of endurance training −induced skeletal muscle response, and the presence of oxidative stress causing endothelial damage or dysfunction and atherosclerosis.

Finally, we have seen that the approach described here can supply general tools to assess athletes' physiological conditions for

endurance training, and this has made possible a revolution in the outlook of athletic training.

Chapter 5

DNA Microarrays for the Study of Cancer

5.1 Introduction

In 1970 President Richard Nixon declared the "war on cancer" in the United States [Barnes, 2005]. The war continues to this day, costing billions of dollars each year in clinical research and treatment, whereas overall cancer rates have not shown any change in their continuous climb for the past hundred and fifty years [Barnes, 2005]. Actually, rates of some types of cancer have decreased in the last fifty years, but other types of cancer have increased significantly, pushing the overall cancer rates upward [Barnes, 2005]. Part of this growth correlates with the increased in the aging population over the past fifty years, since the majority of cancers show up in the old ago groups [Tomatis, 1990]. However, the climbing rates, along with changing patterns of cancer, can also be attributed to changes within our environmental and life types that began with industrialisation [Barnes, 2005].

The term cancer refers to a class of disease in which a group of cells lose control of replication at the cellular level and therefore lose their ability to specialise because of the loss of the normal replication of the nucleic acids in the DNA. Almost every tissue on the body can spawn malignancies and some can yield several types [Waris and Ahsan, 2006]. Cancer cells possess a very insidious properly enabling them to migrate from the site where they originate and form masses at distinct sites in the body [Waris and Ahsan, 2006]. Cancer progression is a stepwise process where the initiated cells, nodules, polyp or the papilloma evolve further and become progressively more malignant [Waris and Ahsan, 2006]. The genes implicated in

malignancy are often modified forms of human genes [Waris and Ahsan, 2006]. All result from cells that have accumulated a set of genetic changes that allow them to invade the body's usual strict constraints on growth and proliferation [Wade, 2001].

The treatment of patients with cancer depends on establishing precise diagnoses with a combination of clinical and pathological information [Ramaswamy et al., 2001]. However, clinical information can be incomplete or misleading [Ramaswamy et al., 2001]. Furthermore, *there is a broad spectrum in cancer morphology and many tumours are atypical or lack morphologic features that are useful for* differential diagnosis [Ramaswamy et al., 2001]. These problems have caused diagnostic confusion, promoting calls for mandatory second opinions in all surginal pathology cases [Tomaszewski, 1999]. In the aggregate, there are limitations that may hinder patient care and confound the results of clinical trials [Golub et al., 1999].

Among genome's first applications in cancer may be the use of DNA microarrays to monitor the simultaneous activity of variant genes in tumour cells. Although substantial progress toward commercialisation has occurred in some cases, most microarray-based diagnostic tests of cancer are still at the development stages [Aitman, 2001]. Presently, for instance, the procedure used to classify childhood ALL subtypes with the help of gene chips is the selection of a number of discriminating genes for the various ALL subtypes using chi-square statistics, and there are marked differences in the number of diagnostic discriminating genes for the various ALL subtypes [Ross et al., 2003; Ross et al., 2004]. This simple technique, however, does not support the use of genome-based expression profiling as a commonplace microarray platform for the diagnosis of the known prognostic subgroups of pediatric ALL in clinical practice, since it can be extremely laborious, time-consuming and expensive. In addition, such procedures do not address the full potential of genome-based experiments to alter our understanding of cellular biology by offering, through an inclusive analysis of the entire repertoire of transcripts, a continuing comprehensive window into the biology underlying the clinical differences among these leukaemia subtypes.

To address these challenge, the Mahalanobis distance in K-means or hierarchical clustering algorithms have been used to examine transcriptional responses in several types of cancer, including childhood ALL Waldenström's macroglobulinemia, bladder cancer, prostate cancer, serous ovarian cancer and breast cancer. The variant genes can be exhibited for further interpretation of toxic mechanisms responsible for the progression of tumour cells.

5.2 Microarray Data Analysis

In this section, the Mahalanobis distance in K-means or hierarchical clustering algorithms is used to classify microarray gene expression profiles in childhood acute lymphoblastic leukaemia, Waldenström's macroglobulinemia, bladder cancer, prostate cancer, serous ovarian cancer and breast cancer. Heat maps are used to represent the measured fluorescence intensities on the 2-dimensional grids. Cells with intensities of zero are coloured black, increasing positive intensities with reds of increasing intensity, and increasing negative intensities with greens of increasing intensities. This visual representation preserves all the quantitative information, but transmits to our brain by ways of a much higher-bandwidth channel than the "number-reading" channel [Eisen et al., 1998].

5.2.1 Childhood ALL

Childhood ALL is a heterogeneous disease with prognostic subtypes that differ markedly in their cellular and molecular characteristics as well as their response to therapy and subsequent risk of relapse [Ross et al., 2003]. Current risk assignment includes clinical characteristics (e.g. age, sex), basic laboratory studies (presenting white blood cell count and presence or absence of leukaemia in cerebral spinal fluid), as well as characteristics of the leukaemic blasts (immunophenotype, cytogenetics, molecular diagnostics for the presence of translocation-encoded fusion transcripts, and response to therapy) [Pui and Evans, 1998; Pui et al, 2001].The early risk features are identified from epidemiologic studies correlating clinical characteristics with outcome data. For example, infants are found to

have poorer overall survival rate than kids between the ages of 2 and 10 years old [Ross et al., 2003]. Immunophenotypic characterization of leukemic blasts subsequently revealed that patients with T-cell lineage ALL (T-ALL) had a higher risk for relapse than B-precursor ALL (B-ALL) [Ross et al., 2003]. These include t(9;22)*(BCRABL)*, t(1;19)*(E2A-PBX1)*, t(12;21)*(TEL-AML1)*, rearrangement in the *MLL* gene on chromosome 11q23, and hyperdiploid karyotype with more than 50 chromosomes [Ross et al., 2003].

Contemporary treatment of childhood acute lymphoblastic leukaemia requires the assignment of patients to specific risk groups [Ross et al., 2003; Ross et al., 2004]. To find the molecular markers that naturally discriminate childhood acute lymphoblastic leukaemia subtypes, the Mahalanobis distance in K-means clustering is used to organise microarray gene expression profiling of bone marrow samples from pediatric patients with acute lymphoblastic leukaemia. The data used here was available from Gene Expression Omnibus under platform accession no GSE2605, and this microarray experiment was performed previously by De Pittà et al [2007].

We consider fifteen group arrangements corresponding to the measured fluorescence intensities of bone marrow samples from pediatric patients with acute lymphoblastic leukemia (Figure 5.1). At the finest level, we find that there is a cluster of 70 features that *naturally discriminate pediatric patients with B-ALL, T-ALL and B-ALL through MLL/AF4 chromosomal rearrangement in agreement with* clinical classification (Figure 5.2).
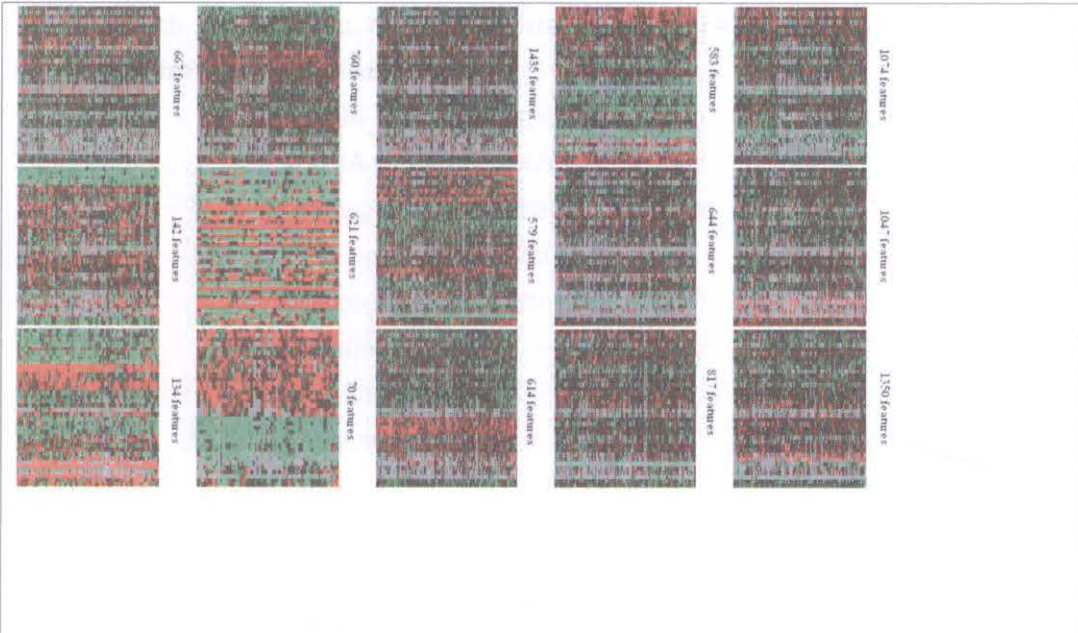
Figure 5.1 K-means clustering results of microarray gene expression profiling of bone marrow samples from pediatric patients with B-ALL, T-ALL and B-ALL through MLL/AF4 chromosomal rearrangement (K=15).
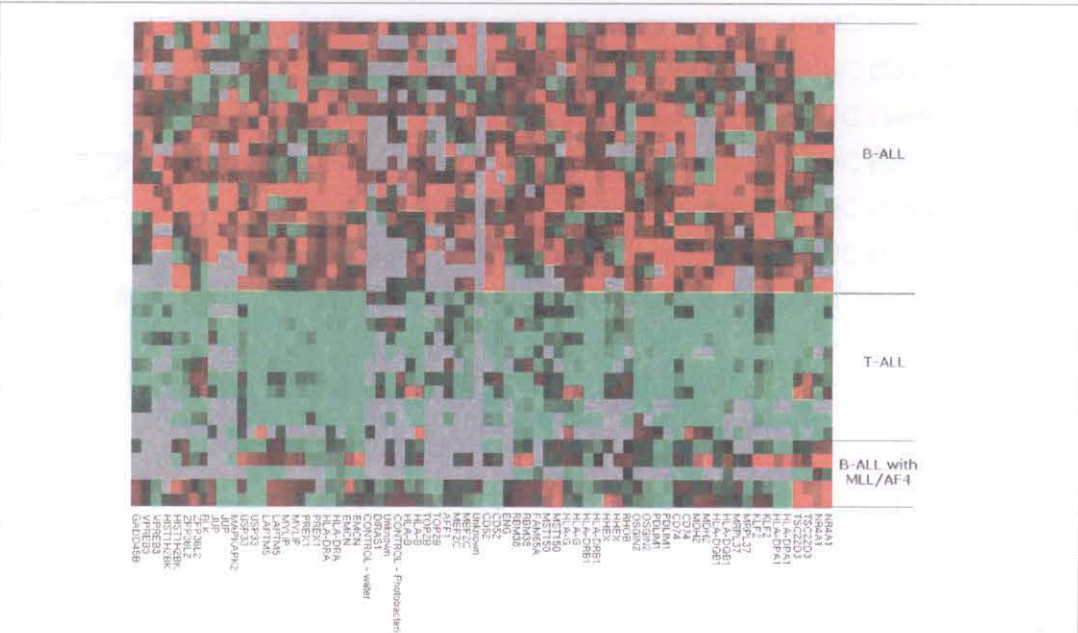
Figure 5.2 A cluster with 70 features that are up-regulated in bone marrow samples from pediatric patients with B-ALL, moderately expressed for those with B-ALL through MLL-Af4 chromosomal rearrangement, and down-regulated for those with T-ALL.

Table 5.1 Diagnostic discriminating genes in bone marrow samples from pediatric

patients with B-ALL, T-ALL and B-ALL through MLL/AF4 rearrangement.

Gene symbol	Gene ontology
RBM38	DNA damage response
TOP2B	DNA topological change
REX1	oxygen free radical metabolism
MDH2	carbohydrate metabolism
USP33	protein deubiquitinylation
LAPTM5	transport
FAM65A	unknown
MST150	unknown
MRPL37	translation
HHEX	mRNA export out of nucleus
NR4A1	transcription
KLF2	transcription
TSC22D3	transcription
AFF1	transcription
MEF2C	transcription
PDLIM1	transcription
HLA-DQB1	antigen processing and presentation (MHC class II)
HLA-DPA1	antigen processing and presentation (MHC class II)
CD74	antigen processing and presentation (MHC class II)
HLA-DRB1	antigen processing and presentation (MHC class II)
HLA-G	antigen processing and presentation (MHC class II)
HLA-B	antigen processing and presentation (MHC class II)
HLA-DRA	antigen processing and presentation (MHC class II)
EMCN	cell adhesion
ENG	cell adhesion
MYLIP	cell motility
OSGIN2	meiosis
DIRAS1	small GTPases mediated signal transduction
RHOB	Rho GTPases mediated signal transduction

To characterise their biological processes we use Entrez Gene to identify gene-specific information on all features within this cluster [Maglott et al., 2007]. This enables us to identify 31 distinct genes that are up-regulated in bone marrow samples from pediatric patients with B-ALL, moderately expressed for those with B-ALL through MLL/AF4 chromosomal rearrangement, and down-regulated for those

with T-ALL (Table 5.1). In particular, we find the expressions of the major histocompatibility complex (MHC) class II molecules (HLA-DQB1, HLA-DPA1, CD74 , HLA-DRB1, HLA-G, HLA-B, and HLA-DRA) are up-regulate din bone marrow samples from pediatric patients with B-ALL, moderately expressed for those with B-ALL through MLL/AF4 chromosomal rearrangement, and down-regulated for those with T-ALL.

Therefore, a unique gene set in classifying childhood ALL takes into account the expressions of MHC class II molecules. In addition, since it is well known that pediatric patients with T-ALL have a higher risk of relapse than those with B-ALL, the application of DNA dementhylating agent to induce MHC class II expressions will be beneficial in the prevention of pediatric ALL relapse [Holling et al., 2004].

5.2.2 Waldenström's Macroglobulinemia

Waldenström's macroglobulinemia is a pathological distinct B-cell malignancy characterised by intramedullary monoclonal expansion of predominantly small B lymphocytes with variable plasmacytoid differentiation in the bone marrow, associated with serum IgM paraprotein [Ghng et al., 2005]. Histologically, this represents bone marrow involvement by lymphoplasmacytic lymphoma [Owen et al., 2003]. Though the clinical disease spectrum has been well established, little is known about its biology [Chng et al., 2006]. Unlike other B cell malignancies, anslocations involving the IgH locus on chromosome 14q32 are rare in WM [Kriangkum et al., 2006]. The most common genetic abnormality is deletion of the long arm of chromosome 6 [Schop et al., 2002; Ackroyd et al., 2005]. Nonetheless, this genetic abnormality is not unique to Waldenström's macroglobulinemia and its biological and clinical significance in Waldenström's macroglobulinemia has yet to be elucidated [Chng et al., 2006].

At present, there is no cure for Waldenström's macroglobulinemia. The average survival rate for patients diagnosed with Waldenström's macroglobulinemia is approximately five to seven years. To find a unique gene set to Waldenström's macroglobulinemia, the Mahalanobis distance in the K-means clustering algorithm is used to

arrange gene expression profiling of bone marrow samples from patients with Waldenström's macroglobulinemia corresponding to those of other malignant B cells, including chronic lymphocytic leukaemia and multiple myeloma. The data used here was available from Gene Expression Omnibus under platform accession no GSE5591, and this microarray experiment was perform previously by Gutiérrez et al [2006].

We consider fourteen group arrangements corresponding to the measured fluorescence intensity of bone marrow samples from patients with Waldenström's macroglobulinemia, chronic lymphoblastic leukaemia and multiple myeloma (Figure 5.3). At the finest level, we have found that there is a cluster of 72 features that naturally discriminate patients with B-cell malignancies into Waldenström's macroglobulinemia, chronic lymphoblastic leukaemia and multiple myeloma (Figure 5.4), whereas there is a cluster of 92 features that reveal considerable fold increase in B lymphocytes from patients with Waldenström's macroglobulinemia and chronic lymphoblastic leukaemia (Figure 5.5).

Figure 5.3 K-means clustering results of microarray gene expression profiling of bone marrow samples from patients with Waldenström's macroglobulinemia, chronic lymphoblastic leukaemia and multiple myeloma (K=14).
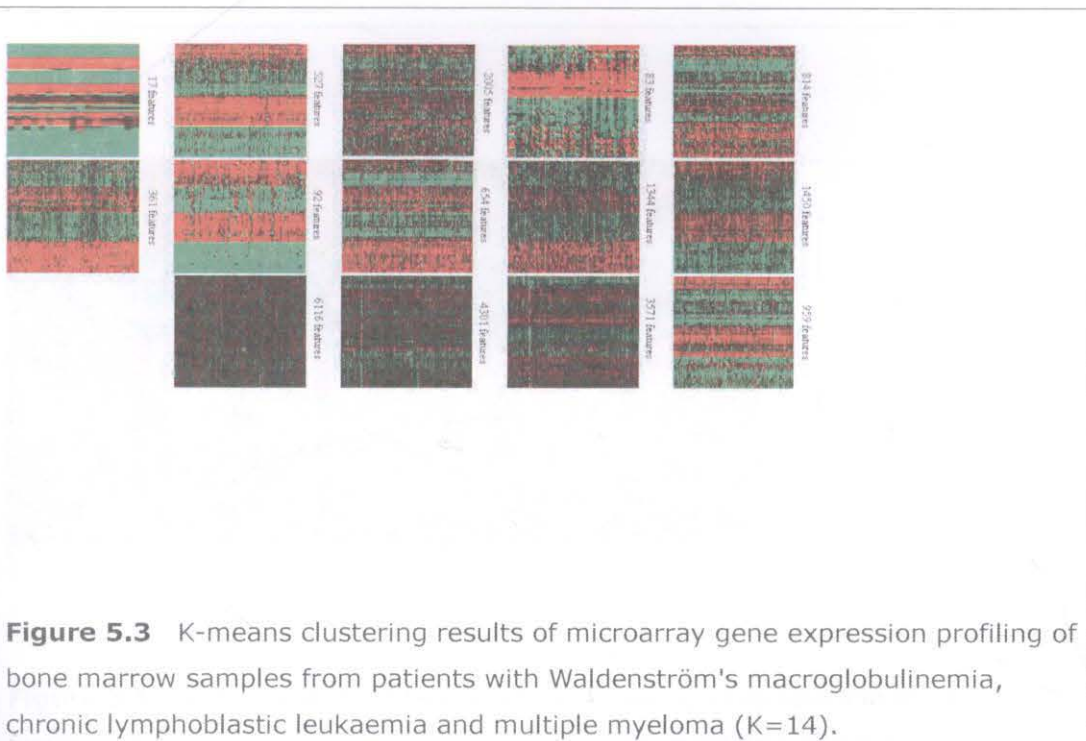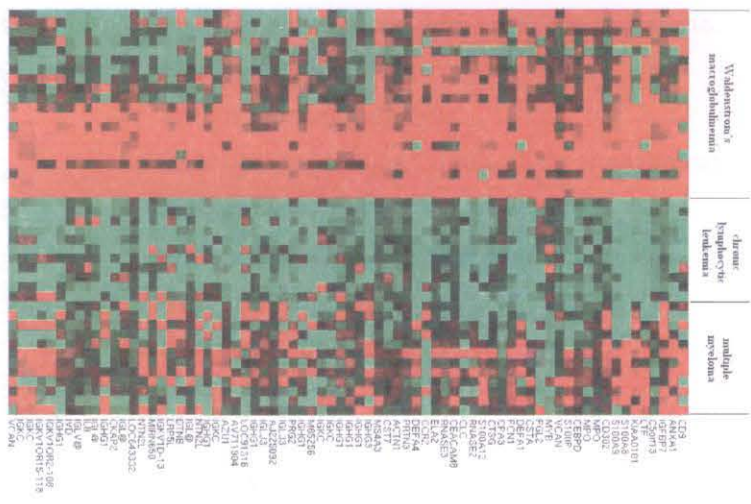
Figure 5.4   A cluster with 72 features that are up-regulated in plasma cells from patients with Waldenström's macroglobulinemia, down-regulated in B lymphocytes for those with chronic lymphoblastic leukaemia, and moderately expressed in plasma cell for those with multiple myeloma.
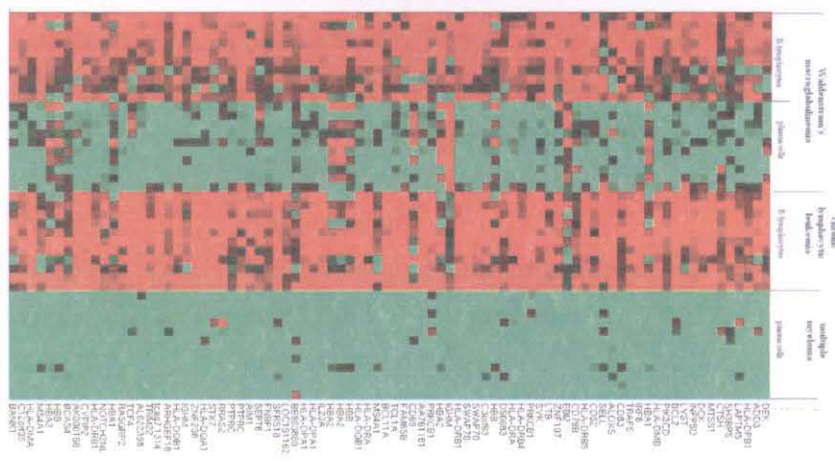
Figure 5.5   A cluster with 92 features that reveal considerable fold increase in B lymphocytes from patients with Waldenström's macroglobulinemia and chronic lymphoblastic leukaemia.

Table 5.2 Diagnostic discriminating genes whose expressions are up-regulated in plasma cells from patients with Waldenström's macroglobulinemia, down-regulated in B lymphocytes from those with chronic lymphoblastic leukaemia, and moderately expressed in plasma cells from those with multiple myeloma.

Gene symbol	Gene ontology
IVD	amino acid metabolism
MPO	energy metabolism; amino acid metabolism
PRTN3	collagen catabolism
PRG2	phosphatidic acid phosphatase activity
LTF	transport (iron ion)
FCN1	transport (phosphate)
CD302	sugar binding
FGL2	signal transduction
CEBPD	transcription
MYB	transcription
KIAA0101	unknown
LOC643332	unknown
LRP5L	unknown
X79782	unknown
MIRN650	microRNA
RNASE2	RNA catabolic process
RNASE3	RNA catabolic process
C5orf13	open reading frame
C6orf142	open reading frame
CKAP2	*apoptosis; cell cycle*
DTNB	*calcium binding; protein binding; zinc ion binding*
CD	cell adhesion; platelet activation
VCAN	cell adhesion
ANXA1	cell motion; peptide cross-linking
IGFBP7	regulation of cell growth
IL8	immune response
IGHG3	immune response
IGHG1	immune response
IGLJ3	immune response
IGL@	immune response
IGLV@	immune response
IGKC	immune response
IGKV1D-13	immune response

IGKV1OR2-108	immune response
IGKV1OR15-118	immune response
LOC91316	immune response
CST7	immune response
DEFA1	immune response
DEFA4	immune response
CEACAM8	immune response
CCR2	immune response; Jak-STATE cascade
CLC	B cell differentiation; JAK-STAT cascade
AZU1	inflammatory response
S100A8	inflammatory response
S100A9	inflammatory response
S100A12	inflammatory response
S100P	endothelial cell migration
ELA2	leukocyte migration
CPA3	renin-angio tensin system
CTSG	renin-angio tensin system
NTN2L	axon guidance
ACTN1	actin filament bundle formation
MS4A3	signal transduction

Table 5.3 Highly differentially expressed genes in B lymphocytes from patients with Waldenström's macroglobulinemia and chronic lymphoblastic leukaemia.

Gene symbol	Gene ontology
CHST6	carbohydrate metabolism
DCK	nucleotide metabolism
HBA1	oxygen transport
HBB	oxygen transport
HBA2	oxygen transport
LAPTM5	transport
LYST	transport (protein)
STX7	interscellular protein transport
SH3BP5	interscellular signaling cascade
PRKCB1	interscellular signaling cascade
CTSH	proteolysis
IRF8	myeloid cell differentiation
BCL2	B cell proliferation
MS4A1	B cell activation

BANK1	B cell activation
PIK3CD	B cell activation; B cell homeostasis
SYK	B cell receptor signaling pathway
CD24	B cell receptor translocation into membrane raft; T cell costimulation
MS4A1	B cell activation
IL23A	immune and inflammatory response
INPP5D	B cell receptor signaling pathway
CD83	immune response
CD79B	immune response
EBI2	immune response
TRIM22	immune response
TRAF5	immune response
FAIM3	immune and defense response
LTB	immune response
ALOX5	inflammatory response
IGHM	immune response
HLA-DPB1	antigen processing and presentation (MHC class II)
HLA-DMA	antigen processing and presentation (MHC class II)
HLA-DMB	antigen processing and presentation (MHC class II)
HLA-DRB5	antigen processing and presentation (MHC class II)
HLA-DRB4	antigen processing and presentation (MHC class II)
HLA-DRA	antigen processing and presentation (MHC class II)
HLA-DRB1	antigen processing and presentation (MHC class II)
HLA-DQB1	antigen processing and presentation (MHC class II)
HLA-DPA1	antigen processing and presentation (MHC class II)
HLA-DQA1	antigen processing and presentation (MHC class II)
FNBP1	endocytosis
CD52	cytosolic calcium ion concentration elevation
ADD3	structural constituent of cytoskeleton
MTSS1	actin cytoskeleton organization and biogenesis
ARHGEF18	Rho protein signaling transduction
RASGRP2	Ras protein signaling transduction
RRAS2	Ras protein signaling transduction
CD69	transmembrane receptor activity
DEK	transcription
ZNF107	transcription
ZNF238	transcription

TCL1A	transcription
TCF4	transcription
BCL11A	transcription
LBH	transcription
C3orf63	open reading frame
C12orf35	open reading frame
SFRS18	unknown
BCAS4	unknown
LOC151162	hypothetical protein
SWAP70	somatic cell DNA recombination
AIM1	cell cycle; cell division
FAM65B	cell differentiation
NOTCH2NL	cell differentiation
SELL	cell adhesion
CYFIP2	cell adhesion

To characterise their biological processes, we use Entrez Gene to search for gene-specific information [Maglott et al., 2007]. This enables us to identify 53 distinct genes that are up-regulated in plasma cells from patients with Waldenström's macroglobulinemia, down-regulated in B lymphocytes from patients with chronic lymphoblastic leukaemia, and moderately expressed in plasma cells from patients with multiple myeloma (Table 5.2). At the same time, we find that there are 78 distinct genes that reveal considerable fold increase in B lymphocytes from patients with Waldenström's macroglobulinemia and chronic lymphoblastic leukaemia (Table 5.3). In particular, the expressions of IgH (IGHG1 and IGHG3), IgK (IGKC, IGKV1D-13, IGKV1OR2-108 and IGKV1OR15-118) and IgL (IGLJ3, IGL@ and IGLV@) genes are up-regulated in plasma cells from patients with Waldenström's macroglobulinemia, and moderately expressed in plasma cells from patients with multiple myeloma. High expressions of bone marrow rennin-angiotensin system are observed in plasma cells from patients with Waldenström's macroglobulinemia. Several genes involved in potential interactions with Netrin-2 mediated axon guidance are up-regulate in Waldenström's macroglobulinemia [Dickson, 2002], suggesting Netrin-2 mediated axon guidance is linked to the susceptibility to Waldenström's macroglobulinemia. In addition, it can

be observed that the expressions of MHC class II molecules are up-regulated in B lymphocytes from patients with Waldenström's macroglobulinemia and chronic lymphoblastic leukaemia, but down-regulated in plasma cells from patients with multiple myeloma. As a result, the expressions of MHC class II molecules in Waldenström's macroglobulinemia are very similar to chronic lymphoblastic leukaemia but distinct from multiple myeloma.

In summary, a unique gene set to Waldenström's macroglobulinemia takes into account the high expressions of IgH, IgK and IgL genes in plasma cells plus high expressions of MHC class II molecules in B lymphocytes.

5.3.3 Bladder cancer

Bladder cancer represents a global public health problem. It ranks ninth in world cancer incidence [Murta-Nascimento et al., 2007]. Around 80 percent of the patients present originally with superficial disease, which comprise Ta tumours located in the mucosa only, submucosa invasive T1 tumours, and carcinoma in situ (CIS) lesions [Dyrskjøt et al., 2004]. The patients presenting Ta and T_1 tumours experience frequent tumour recurrences and to a lesser extent disease progression to a muscle invasive stage, whereas the patients presenting with isolated or concomitant CIS lesions have a high risk of disease progression to a muscle invasive stage [Dyrskjøt et al., 2004]. At present, there are no clinically useful markers available for identifying bladder cancer patients with a high risk of disease progression to a muscle invasive stage [Dyrskjøt et al., 2004]. If we would predict which tumours are likely to progress, we can gain a great impact on the clinical management of patients with superficial disease, since it would be possible to treat high-risk patients more aggressively.

In order to find clinical useful markers that naturally discriminate distinct classes of bladder cancer, the Mahalanobis distance in Ward's linkage hierarchical clustering algorithm is used to group microarray gene expression profiling of bladder biopsies from 15 superficial transitional cell carcinomas without CIS lesion, 13 superficial

transitional cell carcinomas with surrounding CIS lesions, and 12 muscle invasive transitional carcinomas (Figure 5.6). The data used here was available from Gene Expression Omnibus under platform accession no GSE3167 and this microarray experiment was implemented by [Dyrskjøt et al., 2004]. At the finest level, we find that there is a cluster of 239 features that naturally discriminate bladder cancer patients with three classes, those with superficial transitional cell carcinomas with or without CIS lesions, and muscle invasive carcinomas (Figure 5.7).

Figure 5.6 Clustering results of microarray gene expression profiles in bladder biopsies from patients with superficial transitional cell carcinoma with or without CIS lesions, and in muscle invasive carcinomas.
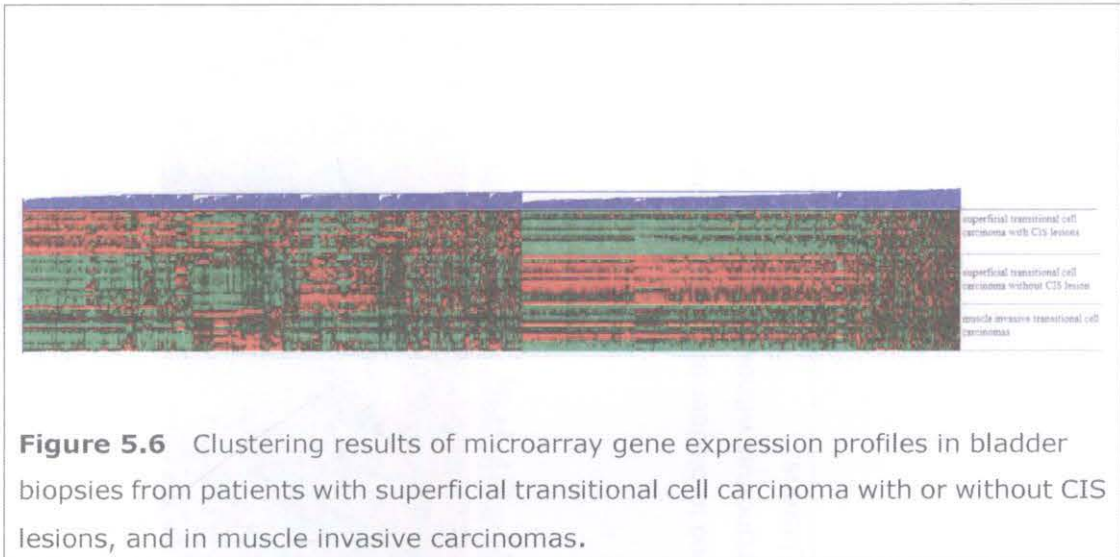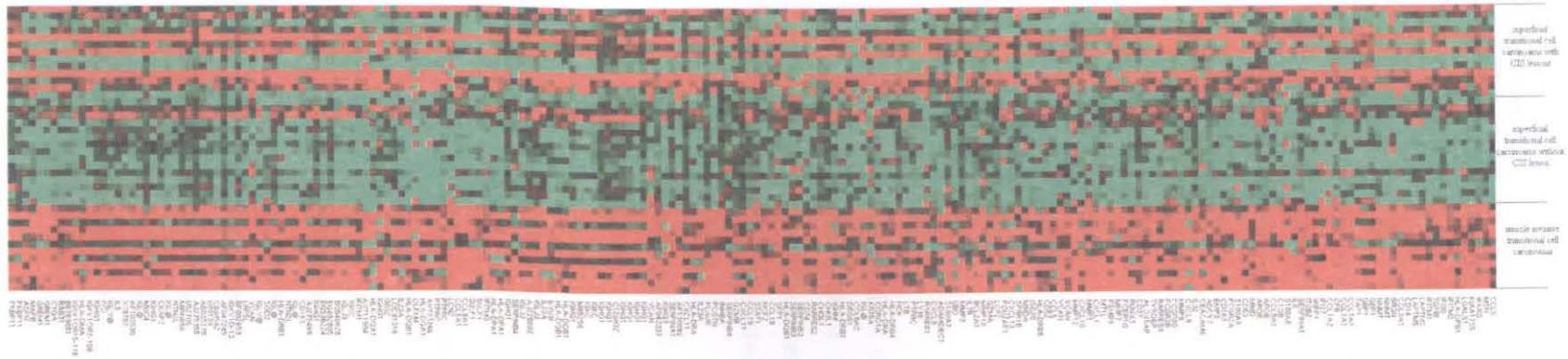
Figure 5.7 A cluster with 225 features that are up-regulated in bladder biopsies from patients with invasive muscle carcinomas, down-regulated for those with superficial transitional cell carcinomas without CIS lesion, and moderately expressed for those with superficial transitional cell carcinomas with surrounding CIS lesions.

Table 5.4 Diagnostic discriminating genes in bladder biopsies from patients with superficial transitional cell carcinoma with or without CIS lesions, and muscle invasive carcinomas.

Gene symbol	Gene ontology
CD74	antigen processing and presentation (MHC class II)
HLA-DPB1	antigen processing and presentation (MHC class II)
HLA-DMA	antigen processing and presentation (MHC class II)
HLA-DRA	antigen processing and presentation (MHC class II)
HLA-DRB4	antigen processing and presentation (MHC class II)
HLA-DRB1	antigen processing and presentation (MHC class II)
HLA-DRB5	antigen processing and presentation (MHC class II)
HLA-DPA1	antigen processing and presentation (MHC class II)
HLA-DQA1	antigen processing and presentation (MHC class II)
IFI30	antigen processing and presentation (MHC class II)
FCER1G	IgE binding
FCGR2A	IgE binding
FCGR3B	IgE binding
UBD	ubiqintin-dependent protein catabolism
PSMB9	ubiqintin-dependent protein catabolism
TYROBP	natural killer mediated cytotoxicity
GZMB	natural killer mediated cytotoxicity
NTN2L	axon guidance
CD52	cytosolic calcium ion concentration elevation
RAB31	small GTPases mediated signaling transduction
IFI27	biological process
STAT1	NIK-I-kappaB/NF-kappaB cascade
LGALS1	regulation of apoptosis; upregulation of I-kappaB kinase/NF-kappaB cascade
LCP2	mast cell activation; cytokine secretion; natural killer mediated cytotoxicity
SRGN	mast cell secretory granule organization and biogenesis; T lymphocyte secretory granule organization and biogenesis
IFITM3	interferon induced transmembrane protein 3 (1-8U)
IFITM2	immune response
IFITM1	cell surface receptor mediated signal transduction; B cell receptor signaling pathway
IL7R	cell surface receptor mediated signal transduction

LY96	immune response
GZMA	immune response
IGLJ3	immune response
IGLV@	immune response
IGHG1	immune response
IGHG3	immune response
IGHM	immune response
IGKC	immune response
IGKV1D-13	immune response
IGKV1OR2-108	immune response
IGKV1OR15-118	immune response
GBP1	immune response
LTB	immune response
CORO1A	immune response
IL8	immune response
SPP1	cytokine activity
IL23A	cell surface receptor mediated signal transduction; immune response
IL32	immune response
CXCL1	immune and inflammatory response
CCL5	immune and inflammatory response
CXCL6	immune and inflammatory response
CXCL9	immune and inflammatory response
CXCL10	immune and inflammatory response
CXCL11	immune and inflammatory response
CXCL13	immune and inflammatory response
CCL18	immune and inflammatory response
CCL19	immune and inflammatory response
NCF2	immune response
S100A7	inflammatory response
S100A8	inflammatory response
S100A9	inflammatory response
ALOX5AP	inflammatory response
CD163	inflammatory response
CD14	inflammatory response
CD53	signal transduction
CEACAM6	signal transduction
COL1A1	immune response

COL1A2	TGF-beta receptor signaling pathway
COL3A1	integrin-mediated signaling pathway; TGF-beta receptor signaling pathway
INHBA	TGF-beta receptor signaling pathway
GREM1	regulation of TGF-beta receptor signaling pathway
ADAMDEC1	integrin-mediated signaling pathway
MSN	leukocyte adhesion
COL5A1	cell adhesion
MUC4	cell adhesion
OLFM4	cell adhesion
POSTN	cell adhesion
VCAN	cell adhesion
ITGB2	activated T cell proliferation
PTPRC	B cell receptor signaling pathway; T cell receptor signaling pathway
ASPN	protein binding
FKBP11	protein folding
ISG20	cell proliferation
RUNX3	cell proliferation; apoptosis
RARRES3	regulation of cell proliferation
EMP3	cell growth
BCL2A1	regulation of apoptosis
G0S2	cell cycle
CKAP2	cell cycle
TGFBI	cell proliferation
SPRR1B	epidermis development; peptide cross-linking
KRT6A	ectoderm development
COL6A2	ectoderm development
HCK	mesoderm development
WIPF1	actin filament based movement
MMD	cytolysis
C1QA	complement component, C1 complex
C1QB	complement activation, classical pathway
CFB	complement activation, alternative pathway
PLAUR	blood coagulation
SERPINA1	blood coagulation
SERPINB3	blood coagulation
SERPINB4	blood coagulation

MAFB	transcription
MNDA	transcription
POU2AF1	transcription
MIRN650	microRNA
C6orf142	open reading frame
C12orf32	open reading frame
EVI2B	unknown
MT1L	unknown
KIAA1755	unknown
CHI3L1	carbohydrate metabolism
LOC91316	carbohydrate metabolism
DEGS1	lipid metabolism
NNMT	metabolisms of cofactors and vitamins
IVD	amino acid metabolism
WARS	amino acid metabolism
SULF1	heparan sulfate proteoglycan metabolism
CD37	protein amino acid N-glycan biosynthesis
ADCY7	nucleotide metabolism
RARRES2	retinoid metabolism
MMP1	metabolism
MMP3	metabolism
MMP7	metabolism
MMP9	metabolism
MMP10	metabolism
MMP12	metabolism
SOD2	response to oxidative stress
APOE	Cdc42 protein signal transduction; lipid metabolism
LAPTM5	transport
LCN2	transport
TAP1	transport (oligopeptide)

To characterise their biological processes, we search for gene-specific information on all features with the aid of Entrez Gene [Maglott et al., 2007]. This enables us to identify 132 distinct genes that are up-regulated in bladder biopsies from patients with muscle invasive carcinomas, moderately expressed for those with superficial transitional cell carcinomas with CIS lesions, and down-regulated for those with superficial transitional cell carcinomas without CIS lesion

(Table 5.4). At a result, these 132 genes naturally distinguish the patients into three distinct classes of bladder cancer, those with muscle invasive carcinomas, superficial transitional cell carcinomas with CIS lesions, and superficial transitional cell carcinomas without CIS lesion. Particularly, we observe the MHC class II expressions (CD74, HLA-DPB1, HLA-DMA, HLA-DRA, HLA-DRB4, HLA-DRB1, HLA-DRB5, HLA-DPA1, HLA-DQA1 and IFI30) are up-regulated in bladder biopsies from patients with muscle invasive carcinomas, moderately expressed for those with superficial transitional cell carcinomas with CIS lesions, and down-regulated for those with superficial transitional cell carcinomas without CIS lesion. The expressions of interferon –induced transmembrane proteins (IFITM1, IFITM2 and IFITM3), IgH (IGHG1, IGHG3 and IGHM),IgL genes (IGLJ3 and IGLV@) and IgK genes (IGKC, IGKV1D-13, IGKV1OR2-108, and IGKV1OR15-118) are also up-regulated in bladder biopsies from patients with muscle invasive carcinomas, moderately expressed for those with superficial transitional cell carcinomas with CIS lesions, and down-regulated for those with superficial transitional cell carcinomas without CIS lesion [Qhu and Burgess, 2000].

It can be observed that the progression of a muscle invasive stage of bladder cancer leads to the binding of IgE to the high-affinity Fc episilon RI and cytokine-cytokine receptor interactions promoting natural killer mediated cytotoxicity [Perussia , 2000; Vivier, 2004]. The progression of a muscle invasive stage of bladder cancer also leads to Netrin-2 mediated axon guidance [Dickson, 2002]. In addition, several genes involved in the regulation of TGF-beta receptor signaling pathway is up-regulated in bladder biopsies from patients with muscle invasive carcinomas, moderately expressed for those with superficial transitional cell carcinomas with CIS lesions, and down-regulated for those with superficial transitional cell carcinomas without CIS lesions, suggesting the activation of TGF-beta receptor signaling pathway is involved in the progression of a muscle invasive stage of bladder cancer.

In summary, the biological mechanisms of bladder cancer progress to a muscle invasive stage include: (1) the MHC class II expressions,

interferon-induced transmembrane proteins, IgL, IgH and IgK molecules are up-regulated in bladder biopsies from patients with muscle invasive carcinomas, moderately expressed for those with superficial transitional cell carcinomas with CIS lesions, and down-regulated for those with superficial transitional cell carcinomas without CIS lesion, (2) Netrin-2 like mediated axon guidance has a strong association with the progression of a muscle invasive stage of bladder cancer, (3) the progression of a muscle invasive stage of bladder cancer leads to the binding of IgE to the high-affinity Fc episilon RI and cytokine-cytokine receptor interactions promoting natural killer mediated cytotoxicity, and (4) TGF-beta receptor signaling pathway is involved in the progression of a muscle invasive stage of bladder cancer.

5.3.4 Prostate Cancer

Prostate cancer is the second most widespread cause of cancer death in males, with around 3,000 men dying from the disease annually in Australia. The prognosis and choice of therapy for prostate cancer is based primarily on three parameters obtained at the time of diagnosis: clinical stage, serum prostate-specific antigen (PSA) and the Gleason score of the cancer [Partin et al., 1997]. The Gleason grade system, which is based on the progressive deterioration of the microscopic tumour architecture, consists of five histological patterns that annotate cancers into five categories exhibiting well differentiated (pattern) to poorly differentiated (pattern 5) features [Gleason and Mellinger, 1974]. A number from 1 to 5 is assigned to the most dominant patterns [True et al., 2006]. A second number, also from 1 to 5, is assigned to the second most dominant patterns [True et al., 2006]. The Gleason grade, which is the sum of these two numbers, has a value between 2 and 10 [True et al., 2006]. In current clinical practice, the vast majority of prostate tumours have a Gleason score of ≥ 6 [Epstein, 2000]. Therefore, tumours composed of Gleason patterns 3, 4, and/or 5 are considered clinically significant [True et al., 2006]. The reporting of individual Gleason patterns is not a trivial distinction, since the amount of pattern 4 and the presence of any pattern 5 has been highly correlated with the probability of cancer dissemination, with respect to therapy, disease outcome,

patient-management decisions, and clinical-trial enrolment [Chan et al., 2000; True et al., 2006]. Several studies have suggested a direct correlation between the Gleason score and clinical measurements of disease outcome, including death due to tumour within 15 years and the likelihood of remaining free of biochemical evidence of the recurrence of the disease after either definitive, potentially curative radical prostatectomy or radiation therapy [True et al., 2006]. Although different Gleason patterns are histologically distinctive, the molecular features underlying these tumour phenotypes are still relatively unknown [True et al., 2006].

To know molecular alterations underlying prostate cancer grade, the Mahalanobis distance in Ward's linkage hierarchical clustering algorithm is used to group microarray gene expression profiles in prostate epithelial cells of Gleason grades 3, 4 and 5 (Figure 5.8). The data used here was available from Gene Expression Omnibus under platform accession no GSE5132, and this microarray experiment was implemented by True et al. [2006]. At the finest level, we find that there is a cluster of 55 features whose expressions are up-regulated in high-grade prostate tumours of Gleason patterns 4 and 5, and moderately expressed in low-grade prostate tumours of Gleason pattern 3 (Figure 5.9).
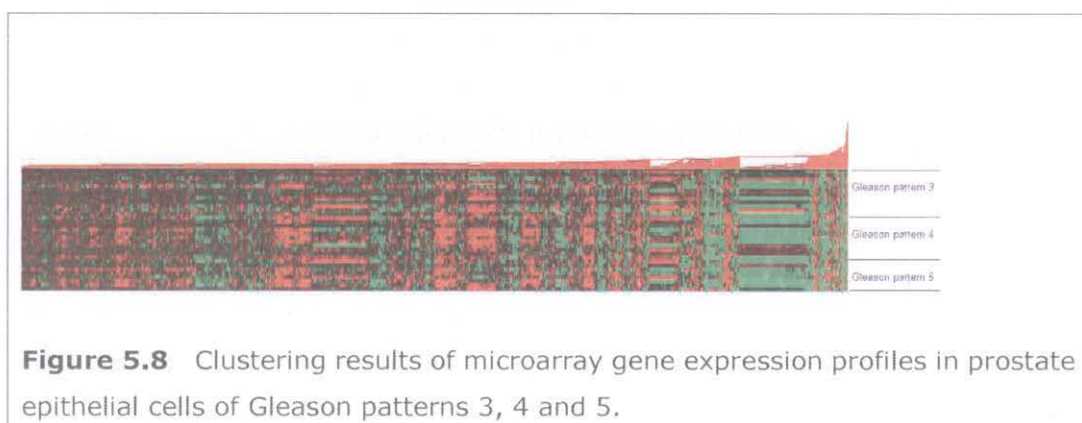
Figure 5.8 Clustering results of microarray gene expression profiles in prostate epithelial cells of Gleason patterns 3, 4 and 5.

Figure 5.9 A cluster of 55 features that are up-regulated in prostate epithelial cells of Gleason patterns 4 and 5, and moderately expressed from those with Gleason pattern 3.
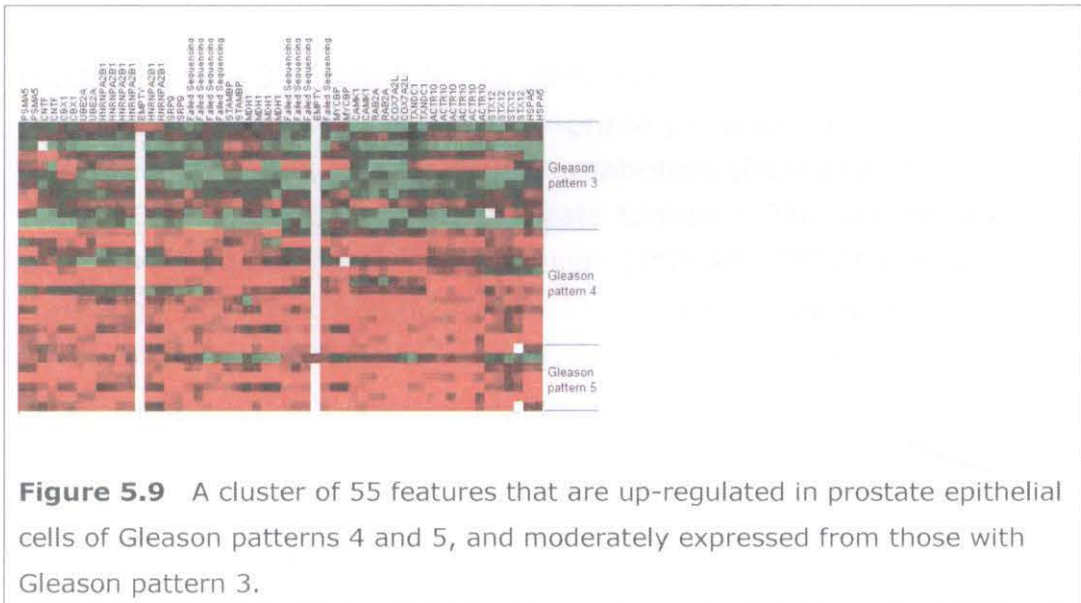
Table 5.5 Diagnostic discriminating genes whose expressions are up-regulated in prostate epithelial cells of Gleason patterns 4 and 5, and moderately expressed from those with Gleason pattern 3.

Gene symbol	Gene ontology
MYCBP	spermatogenesis
MDH1	carbohydrate metabolism
STX12	cholesterol efflux
HSPA5	endoplasmic reticulum overload response
COX7A2L	electron transport chain
TXNDC1	electron transport chain
PSMA5	ubiquitin-dependent protein catabolism
UBE2A	ubiquitin-dependent protein catabolism
STAMBP	ubiquitin-dependent protein catabolism
CBX1	chromatin remodeling
HNRNPA2B1	nuclear mRNA splicing, via spliceosome
RAB2A	protein transport
SRP9	protein export
CNTF	cell differentiation
CAMK1	cell differentiation

To characterise their biological processes, gene ontology within this cluster is perform on all features with the help of Entrez Gene [Maglott et al., 2007]. This enables us to identify 15 distinct genes that are up-regulated in high-grade prostate tumours of Gleason grades 4 and

5, and moderately expressed in low-grade prostate tumours of Gleason grade 3 (Table 5.5). In particular, we observe changes in spermatogenesis (MYCMP) in high-grade prostate tumours. We also observe changes in carbohydrate metabolism (MDH1) and cholesterol efflux (STX12) in high-grade prostate tumours. The occurrence of ubiquitin-mediated protein catabolism (PSMA5, UBE2A), electron transport chain (COX7A2L, TXNDC1), and endoplasmic reticulum overload response (HSPA5) are also observed in high-grade prostate tumours. In addition, several genes involved in potential interactions with chromatin remodeling and nuclear mRNA splicing, via spliceosome assembly, are up-regulated in high-grade prostate tumours.

In summary, the molecular alterations underlying high-grade prostate tumours of Gleason patterns 4 and 5 include: (1) changes in MYCMP-mediated spermatogenesis, (2) changes in carbohydrate metabolism and cholesterol efflux, (3) the occurrence of ubiquitin-mediated degradation and endoplasmic reticulum overload response, and (4) ubiqintin-mediated RNA splicing and nuclear mRNA splicing, via the spliceosome. Specifically, RNA splicing is a process that removes introns and joins exons in the primary transcript; whereas nuclear mRNA splicing, via spliceosome, is a process that the joining together of exons from one or more primary transcripts of mRNA and the excision of intron sequences, via a spliceosomal mechanism, so that mRNA consisting only of the joining exons is produced [Patel & Steitz, 2003].

5.3.5 Serous Ovarian Cancer

Ovarian cancer is the sixth most widespread cancer in women worldwide and a highly aggressive gynaecological malignancy causing around 125,000 deaths per year [Parkin et al., 2005]. Despite advances in detection and cytotoxic therapies, only 30 percent of patients with advanced stage ovarian cancer survive 5 years after initial diagnosis [Parkin et al., 2005]. As a result, an understanding of molecular features involved in ovarian cancer pathogenesis and progression has the potential to have a significant impact on the outcomes for this devastating disease [Wang et al., 2006].

Most ovarian neoplasms are adenocarcinomas that occur as four major histological subtypes, serous, mucinous, endometrioid and clear cell, with serous being the most widespread [Wang et al., 2006]. Current epidemiological data suggests that each of these histological subtypes is associated with distinct morphological and molecular genetic alterations [Bell, 2005]. Further investigations of biological mechanisms promoting ovarian cancer are necessary to determine how each of the subtypes emerges. Relatively little is known about the molecular events that lead to the progress of ovarian cancer.

In this section, the Mahalanobis distance in Ward's linkage hierarchical clustering algorithm is used to group microarray gene expression profiling of ovarian cancer cell lines from patients with serous adenocarcinomas corresponding to those with clear adenocarcinomas (Figure 5.10). The data used here was available from Gene Expression Omnibus under platform accession no GSE3001, and this microarray experiment was implemented by Komatsu et al [2006]. At the finest level, we find that there is a cluster of 25 features that exhibit substantial fold increase in serous adenocarcinomas (Figure 5.11). Thus, genes within this cluster are highly differentially expressed in serous ovarian adenocarcinomas.

Figure 5.10 Clustering results of microarray gene expression profiling of ovarian cancer cell lines from patients with serous adenocarcinomas corresponding to those with clear adenocarcinomas.

75
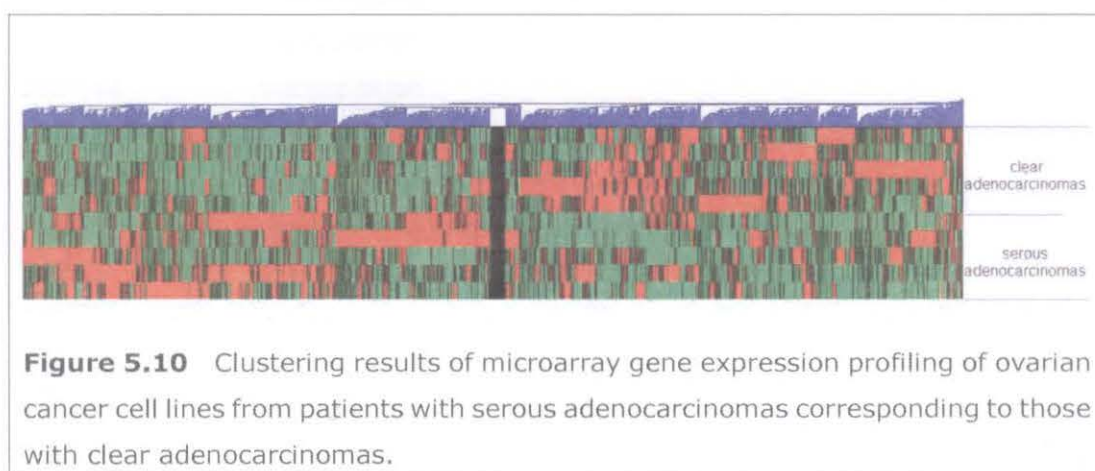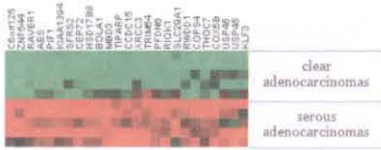
Figure 5.11 A cluster of 24 features that exhibit substantial fold increase in ovarian cancer cell lines from patients with serous adenocarcinomas versus those with clear adenocarcinomas.

Table 5.6 Highly differentially expressed genes in serous ovarian adenocarcinomas.

Gene Symbol	Gene Ontology
HSD17B8	estrogen biosynthesis
AES	Wnt receptor signaling pathway
XRCC3	DNA recombination; cell cycle
C6orf125	open reading frame
CCDC15	unknown
THOC7	unknown
BMD3	transcription
KLF3	transcription
ZNF544	transcription
KIAA1394	ATP binding
RAVER1	ribonucleoprotein, RNA/nucleotide/nucleic acid binding
PIF1	regulation of telomere maintenance
SFRS2	RNA splicing; nuclear mRNA splicing via spliceosome
RIOK1	kinase activity (prematurely truncated due to insertion of intronic sequence)
TRIM54	cell cycle
TFDP1	cell proliferation
COX5B	respiratory gaseous exchange
USP45	ubiquitin-dependent proteolysis
CEP72	identical protein binding
PFDN6	protein folding
COPS4	protein folding

BOLA1   protein folding

RWDD1   protein folding

SLC29A1   transport (nucleoside)

To characterise their biological processes, gene ontology within this cluster is performed on all features using Entrez Gene [Maglott et al., 2007]. This enables us to identify 24 distinct genes that are up-regulated in serous ovarian adenocarcinomas (Table 5.6). Specifically, we observe the occurrence of estrogen biosynthesis (HSD17B8) in serous ovarian adenocarcinomas. AES, a gene played a role in WNT receptor signaling pathway is up-regulated in serous ovarian adenocarcinomas as well. We also observe the occurrence of respiratory gaseous exchange (COX5B), ubiquitin-dependent proteolysis (USP45), cell cycle (TRIM54) and cell proliferation (TFDP1) respiratory gaseous exchange (COX5B), ubiquitin-dependent proteolysis (USP45), cell cycle (TRIM54) and cell proliferation (TFDP1). The expression of ribonucleoprotein (RAVER1) is also up-regulated in serous ovarian adenocarcinomas. In addition, it can be observed that several genes involved in potential interactions with ubiquitin-mediated RNA and nuclear mRNA splicing, via spliceosome, are up-regulated in serous ovarian adenocarcinomas.

To sum up, the molecular alterations underlying serous ovarian adenocarcinomas include: (1) the presence of estrogen biosynthesis, (2) the involvement of WNT receptor signaling pathway, and (3) ubiqintin-mediated RNA splicing and nuclear mRNA splicing, via the spliceosome. s

5.3.6 Breast Cancer

Breast cancer is the most frequently diagnosed cancer and the top cause of cancer death among women worldwide [Parkin et al., 2005]. At present, breast tumours are well known to be composed of phenotypically diverse groups of cells [Abraham et al., 2005]. However, it is unclear which of these cell types contribute to tumour development [Abraham et al., 2005]. In contrast to the hypothesis that all cell populations have the capacity to become tumorigenic through accumulation of mutations, another hypothesis limits this

ability to an elite group of cells that share classic features of stem cells such as the ability of self-renew and to differentiate [Reya et al., 2001].

To know molecular alterations underlying solid tumours heterogeneity in breast cancer, the Mahalanobis distance in Ward's linkage hierarchical clustering algorithm is used to group microarray gene expression profiling of tumorigenic and non-tumorigenic breast cancer cell lines corresponding to normal breast epithelial cells (Figure 5.12). The data used here was available from Gene Expression Omnibus under platform accession no GSE2618, and this microarray experiment was performed previously by Liu et al [2007]. At the finest level, we find that there is a cluster of 97 features that are up-regulated in tumorigenic breast cancer, moderately expressed in non-tumorigenic breast cancer, and down-regulated in normal breast epithelium (Figure 5.13). Thus, genes within this cluster naturally discriminate samples into normal breast epithelium, tumorigenic and non-tumorigenic breast cancer cells.

Figure 5.12 Clustering results of microarray gene expression profies in normal breast epithelium, tumorigenic and non-tumorigenic breast cancer cells.

Figure 5.13 A cluster of 97 features that are up-regulated in tumorigenic breast cancer, moderately expressed in non-tumorigenic breast cancer, and down-regulated in normal brest epithelium.
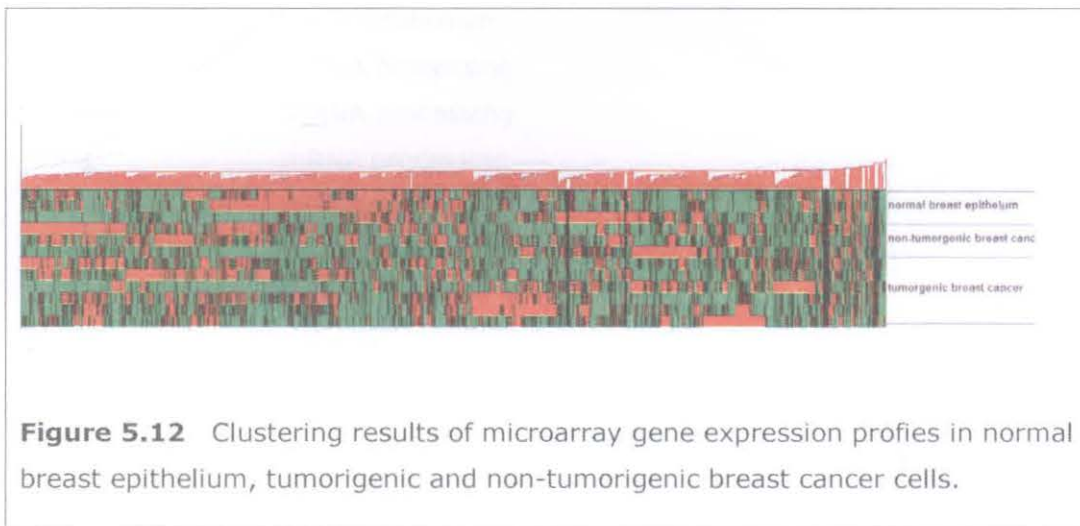
Table 5.7 Diagnostic discriminating genes whose expressions are up-regulated in tumorigenic breast cancer, moderately expressed in non-tumorigenic breast cancer, and down-regulated in normal brest epithelium.

Gene Symbol	Gene Ontology
LYCAT	lipid biosynthetic process
AGPS	lipid biosynthetic process
RP5-1022P6.2	lipid metabolism
INTS2	snRNA processing
INTS10	snRNA processing
ROD1	mRNA processing
THOC2	mRNA processing
TRMT6	tRNA processing; translation
EIF1AD	translational initiation
GFM1	translation elongation
GTF2H2	RNA elongation from Pol II promoter
NMD3	nonsense-mediated mRNA decay
MCM10	DNA replication
FAM33A	cell cycle
SERPINE2	cell differentiation
SASS6	cell cycle
FAM29A	hypothetical protein
FAM36A	hypothetical protein
FAM105B	hypothetical protein
FAM133B	hypothetical protein
FSD2	unknown
PCNX	unknown

QSER1	unknown
TMEM69	unknown
CCDC43	unknown
CTDSPL2	unknown
NARG2	unknown
SPATS2	unknown
TMEM182	unknown
VTA1	unknown
ZBTB37	transcription
ZNF280C	transcription
C15org23	open reading frame
C13orf3	open reading frame
C1orf104	open reading frame
C20orf7	open reading frame
C20orf20	open reading frame
CBX3	chromatin remodeling
KIAA1524	protein binding
WAC	protein binding
TRIM59	protein binding
S100A16	calcium ion binding protein
TMOD3	actin binding; tropomysin binding
CASC5	acrosome formation
CEP78	centrosomal protein
CENPL	centrosomal protein
CENPN	centrosomal protein
BIRC5	establishment of chromosome localization
PIF1	regulation of telomere maintenance
PMCHL2	biological process
KRIT1	small GTPases mediated signaling transduction
RAB39B	small GTPases mediated signaling transduction
GPR93	G protein coupled receptor signaling pathway
OR51E2	G protein coupled receptor signaling pathway
GOLT1B	vehicle-mediated transport
SEC22C	vehicle-mediated transport
ERGIC1	vehicle-mediated transport
SYT13	vehicle-mediated transport
XPO4	protein transport
SLC25A40	transport

SLC39A10	transport (ion, metal ion, zinc ion)
CSNK1G1	WNT receptor signaling pathway
TIMM23	protein targeting to mitochondrion
ASPM	maintenance of centrosome localization
PPIL5	ubiquitin-mediated protein catabolism
DTL	ubiquitin-mediated protein catabolism
DNHD2	ubiquitin-mediated protein catabolism
ASB5	ubiquitin-mediated protein catabolism
RAD18	ubiquitin-mediated protein catabolism
CDC2	ubiquitin-mediated protein catabolism
CDC23	ubiquitin-mediated protein catabolism
TBL1XR1	ubiquitin-mediated protein catabolism
UBE2R2	ubiquitin-mediated protein catabolism
XPNPEP1	proteolysis

To characterise their biological processes, gene ontology within this cluster is perform on all features with the help of Entrez Gene [Maglott et al., 2007]. This enables us to identify 74 distinct genes that are up-regulated in tumorigenic breast tumours, moderately expressed in non-tumorigenic breast tumours, and down-regulated in normal breast cancer epithelium (Table 5.7). In particular, we observe changes in lipid biosynthetic and metabolic processes (LYCAT, AGPS, and RP5-1022P6.2) in tumorigenic breast tumours. We also observe changes in (LYCAT, AGPS, and RP5-1022P6.2) in tumorigenic breast tumours. Changes in DNA replication (MCM10), cell differentiation (SERPINE2), cell cycle (FMA33A and SASS6) and WNT receptor signaling pathway (CSNK1G1) are also found in tumorigenic breast tumours.

It can be observed that several genes involved in ubiquitin-mediated protein catabolism (PPIL5, DTL, DNHD2, ASB5, RAD18, CDC2, CDC23, TBL1XR1 and UBE2R2) and proteolysis (XPNPEP1) are up-regulated in tumorigenic breast tumours, moderately expressed in non-tumorigenic breast tumours, and down-regulated in normal breast epithelium, suggesting the potential role of ubiquitin-mediated degradation machinery in the development of tumorigenic status. The occurrence of acrosome formation (CASC5), telomere maintenance (PIF1), centrosome (CEP78, CENPL and CENPN), chromatin

remodeling (CBX3), and establishment of chromosome localization (BIRC5) are also observed in tumorigenic breast cancer cells, suggesting the potential role of centrosome amplification and chromosome instability underlying the development of tumorigenic status in breast tumours.

In conclusion, gene expression profiling identified the potential roles of ubiquitin-mediated degradation machinery, centrosome amplification and chromosome instability underlying the development of tumorigenic status in breast tumours, and subsequently substantiated the hypothesis that all cell populations have the capacity to become tumorigenic through accumulation of mutations.

5.3 Conclusion

Microarray-based genomic surveys and other high-throughput approaches (ranging from genomics to combinational chemistry) are becoming increasingly important in clinical studies [Eisen et al., 1998]. As a result, we need to enable ourselves to "see" the information in the massive tables of quantitative measures that these techniques generate [Eisen et al., 1998]. Our approach to this problem can be generalised as follows: First, we use the Mahalanobis distance in the K-means or hierarchical clustering algorithm to classify cancer-related microarray gene expression data, which take the biological complexities and interdependencies into account. Next, the results as presented gain insight into toxic mechanisms underlying the progression of tumour cells, which lead to the development of diagnostic array platforms and promising new therapeutic strategies to treat or cure cancer. The success of this computational approach has enabled the diagnosis of toxic mechanisms of cancer, thus offering genome-based expression profiling as a single standardised microarray platform for diagnosis and therapeutic interventions in cancer.

Chapter 6

DNA Microarrays for the Study of Other Diseases

6-1 Introduction

The race to sequence and elucidate the human genome has culminated in an unparalleled gift to clinical investigators [Wade, 2001]. With the genome sequence in hand, clinical investigators can expect to understand all human diseases at the genetic level and, in time, to develop treatments or cures on the basis of this understanding [Wade, 2001]. At least in principal the prospect of conquering almost all diseases can be entertained [Wade, 2001].

In this chapter, the bioinformatics techniques described in Chapters 2 and 3 are used to analyse and make sense of microarray gene expression data in a choice of complex diseases, such as interstitial cystitis (IC), as asthma, type 2 diabetes, HIV/AIDS, organ transplant rejection, ovarian endometriosis and inflammatory papules. The analytical results lead to the classification of pathological samples of unknown, equivocal and ambiguous history, thus offering the basis of minimally invasive diagnostic tests and therapeutic interventions in the study of disease.

6.2 Microarray Data Analysis

In Section 6.2.1, the conditional logistic regression model is used to study the aetiology of interstitial cystitis (IC) in a matched case-control study. From Section 6.2.2 to Section 6.2.11, the Mahalanobis distance in the K-means or agglomerative hierarchical

clustering algorithms is sued to study the etiology of various complex diseases, such as asthma, diabetes, HIV/AIDS, organ transplant rejection, endometriosis and so on. Heat maps are used to represent the measured fluorescence intensities on the 2-dimensional grids. Cells with intensities of zero are coloured black, increasing positive intensities with reds of increasing intensity, and increasing negative intensities with greens of increasing intensities. This visual representation preserves all the quantitative information, but transmits to our brain by ways of a much higher-bandwidth channel than the "number-reading" channel [Eisen et al., 1998].

6.2.1 Interstitial Cystitis

Research into interstitial cystitis (IC) has a very long history. It was first documented by Mercier in 1937, and characterised by Skene in 1887, as the presence of an inflammation that destroys the urinary bladder mucous membrane, partly or wholly, and extends to muscular parietes [Rover and Esdridge, 2005]. In 1914, Hunner popularized IC with the description of a characteristic bladder wall ulcer [Rover and Esdridge, 2005]. In 1947, Hand made the first epidemiological description of IC as the widespread, small, submcosal, bladder hemorrhages and substantial shrinkage in the bladder capacity [Rover and Esdridge, 2005]. In 1987, Messeing and Stamey documented glomerulation as an indicator of IC [Rover and Esdridge, 2005]. To date, IC is still relatively unknown and has been acknowledged by the American Urological Association, as the most challenging disease.

One of the most difficult aspects in the treatment of IC is its diagnosis. Presently, the requirements for diagnosis involve excluding all other causes, such as urinary tract infection, bladder cancer, kidney diseases and sexually transmitted diseases. Also, an invasive cystoscopy must be performed in which the bladder is distended and hemorrhages are identified. Although Keay et al [2003] have isolated several potential candidate IC markers from the urine of patients with IC, their attempts remained undetermined for two reasons:

1. We all know that IC is a complex disease. An individual molecular marker for superior intelligence probably does not it does for IC. Instead, genes and gene products work together in a complex and interdependent way to cause symptoms of IC.

2. Keay et al [2003] studied gene expression based RNA/spotted DNA/cDNA array on bladder epithelial cells from human IC subjects corresponding to their age- and gender- matched controls. However, they used ANOVA to identify IC markers, which failed to take the matching into consideration. As a result, the results they presented suffered from excessive type I error: A "false-positive" conclusion that the outcome are in a causal relationship with confounding factors. Although the primary objective of matching is the elimination of biased comparisons between cases and controls, their objective can only be accomplished in general if matching is accompanied by an analysis that corresponds to the matched design [Schlesselman, 1982]. Unless the statistical analysis accounts for the matching used in the selection phases of a case-control, the estimated statistic associated with the onset of IC can be biased as a consequence of matching. Therefore, matching is only the first of a two-step process that can be used to control for confounding (i) matching design, followed by (ii) matched analysis.

We perform an analysis of gene expression of 4,132 genes using spotted DNA microarrays, and compared microarray gene expression profiling of bladder epithelial cells from patients with IC corresponding to their age- and gender- matched controls. To take the matching into consideration, the conditional logistic regression model is used to analyse gene expression profiling of bladder epithelial cells from patients with IC corresponding to their age- and gender- matched controls. To consider the tradeoff between the biological meaning and statistical significance, significantly differential expressed genes with IC are prioritised relating to combinations of the following two criteria:

$\hat{\beta}_i > 100$ and $\dfrac{\hat{\beta}_i - \mu(\hat{\beta}_i)}{\mu(\hat{\beta}_i)} > 1$. This enables us to identify 139

significant differentially expressed genes in response to IC stimulus. The data used here was collected from Gene Expression Omnibus

under platform accession no GSE621, and this microarray experiment was performed previously by Keay et al [2003].

The results were very interesting. First, complex metabolic reactions, including carbohydrate, lipid, cofactors, vitamins, xenobiotics, nucleotide and amino acid metabolisms, are found to have a strong relationship with bladder smooth muscle contraction through IC status. Next, based on the mapping of the resultant differentially expressed genes onto the biochemical pathways of cytoskeleton dynamics, including the regulation of actin cytoskeleton, transforming growth factor-beta (TGF-beta) signaling transduction, mitogen-activated protein kinase (MAPK) signaling transduction, peroxisome proliferators-activated receptor (PPAR) signaling transduction and adipocytokine signaling transduction, we are able to find the transcriptional regulations of IC-induced bladder smooth muscle contraction status, including the level of contractile force, tissue homeostasis, energy homeostasis and the development of nervous system. Thus, the mapping of a gene onto these biochemical pathways mimics its actual transcription. A more insightful explanation is given as follows.

Table 6.1 Significant differentially expressed genes in response to IC stimulus using fixed effect logistic regression model.

Gene Symbol	Gene Ontology
ALDH4A1	amino acid metabolism
IVD	amino acid metabolism
PSMC4	amino acid metabolism
PSMB6	amino acid metabolism
AMPD1	nucleotide metabolism
ATP5D	nucleotide metabolism
TSTA3	nucleotide metabolism
INDO	metabolisms of cofactors and vitamins
VNN1	metabolisms of cofactors and vitamins
GNLY	xenobiotics metabolism
INPP5D	carbohydrate metabolism
INPP1	carbohydrate metabolism
INSR	carbohydrate metabolism
LDHA	carbohydrate metabolism

HYAL2	carbohydrate metabolism
MAN2A1	carbohydrate metabolism
PGM1	carbohydrate metabolism
PIP5K1C	carbohydrate metabolism
ACSM3	lipid metabolism
ASAH1	lipid metabolism
PRKAG1	lipid metabolism
PLCG2	lipid metabolism
BTN2A1	lipid metabolism
MTPA	lipid metabolism
HADHA	lipid metabolism
ECHS1	lipid metabolism
HSD11B2	lipid metabolism
HSD17B3	lipid metabolism
NQO2	oxidation reduction
PRDX4	oxidation reduction
NDRG1	response to metal ion
VTI1B	vehicle mediated transport
ASNA1	transport (anion)
TCN2	*transport (cobalamin)*
STX4A	transport (neurotransmitter)
SLC7A2	transport (amino acid)
SLC22A18	transport (tetracycline)
SLC25A46	transport (phosphate)
CYP1A1	transport (electron)
DLD	transport (electron)
PCM1	transport (electron)
QP-C	transport (electron)
ARF4	transport (protein)
KCNS1	transport (potassium ion)
KCNN3	transport (potassium ion)
COL7A1	transport (phosphate)
SLC17A1	transport (phosphate)
VDBG	transport (vitamin D)
IL3	immune response; promote eosinophil production and activation
TNF	inflammatory response; stimulates cytokines and chemokines; activates endothelium
THY1	mast-cell activation

CXCL7	immune response
CCL3	immune response
DARC	immune response
IGJ	immune response
IL6	immune response
IL15	immune response
IFRD2	immune response
IFITM1	immune response
KIT	immune response
PSMB9	immune response
IGJ	immune response
TAPBP	immune response
TRB	immune response
XBP1	immune response
CLU	blood coagulation and complement; immune response
TFPI	blood coagulation and complement; immune response
TFPI2	blood coagulation and complement; immune response
SERPINA5	blood coagulation and complement; immune response
SERPINH1	blood coagulation and complement; immune response
CHGA	blood pressure regulation
GPS1	JNK cascade
PPP2R1B	signaling by MAPK
FGF1	signaling by MAPK
MADD	activation of MAPK
SHC	activation of MAPK
MAP2K2	Ras protein signal transduction
MAP2K3	signal transduction
RHOA	Rho protein signal transduction
RREB1	Ras protein signal transduction
GNAI1	signal transduction
ITPKB	signal transduction
RGS18	GTPases activator activity
MYL3	regulation of ATPase activity
XRCC1	regulation of GTPase activity
PTHLH	G protein signaling
PIM2	ATP binding
SNTB2	actin binding
CDC42BPA	actin cytoskeleton organization and biogenesis

CAPZA2	actin cytoskeleton organization and biogenesis
PTPN13	amino acid dephosphorylation
ICAM1	cell adhesion
CD47	cell adhesion
ITGB5	cell adhesion
PCDHB16	cell adhesion
FLOT2	cell adhesion
CASP10	apoptosis
DAD1	apoptosis
PDCD10	apoptosis
FAS	apoptosis
IER3	apoptosis
MCL1	apoptosis
BCAP31	apoptosis
SIAH2	apoptosis
KIT	TGF-beta receptor signaling pathway
ACVR2	TGF-beta receptor signaling pathway
BMP2	TGF-beta receptor signaling pathway
BMP6	TGF-beta receptor signaling pathway
AGER	TGF-beta receptor signaling pathway
FN1	TGF-beta receptor signaling pathway
PTN	TGF-beta receptor signaling pathway
PTPRG	TGF-beta receptor signaling pathway
PRKCZ	amino acid phosphorylation
PSKH1	amino acid phosphorylation
PTPRM	amino acid phosphorylation
TESK1	amino acid phosphorylation
SFRS5	mRNA processing
SF3B4	mRNA processing
SFRS10	mRNA processing
HNRPL	mRNA processing
HNRPU	mRNA processing
BRPF1	transcription
ETS2	transcription
GATA6	transcription
GTF2B	transcription
HMGA1	transcription
HOXA4	transcription

ID2	transcription
LDB2	transcription
NR3C1	transcription
PELP1	transcription
SMCX	transcription
ZNF3	transcription
ZNF161	transcription
ZNF345	transcription
C19orf56	open reading frame
C1orf2	open reading frame
METAP2	translation
EIF1AY	translation

Regulation of the level of contraction force. The contraction of bladder smooth muscle is controlled by the nervous system. According to the mapping representing the regulation of actin cytoskeleton (Figure 6.1), Ras protein of small GTPases domains are strong candidates to transmit axonal guidance signals in the growth cone in response to stress. Both Ras and ATP proteins are identified as responsible for the generation of mechanical forces upon the hydrolysis, consisting with their structural similarity for the catalytic regions of GTPases and ATPases domains [Kosztin et al., 2002]. The force transmission mechanism is based on an irreversible structural change, produced by the hydrolysis, which trigger a thermal switch between force-generating substates through change in the configurational space of the proteins [Kosztin et al., 2002].
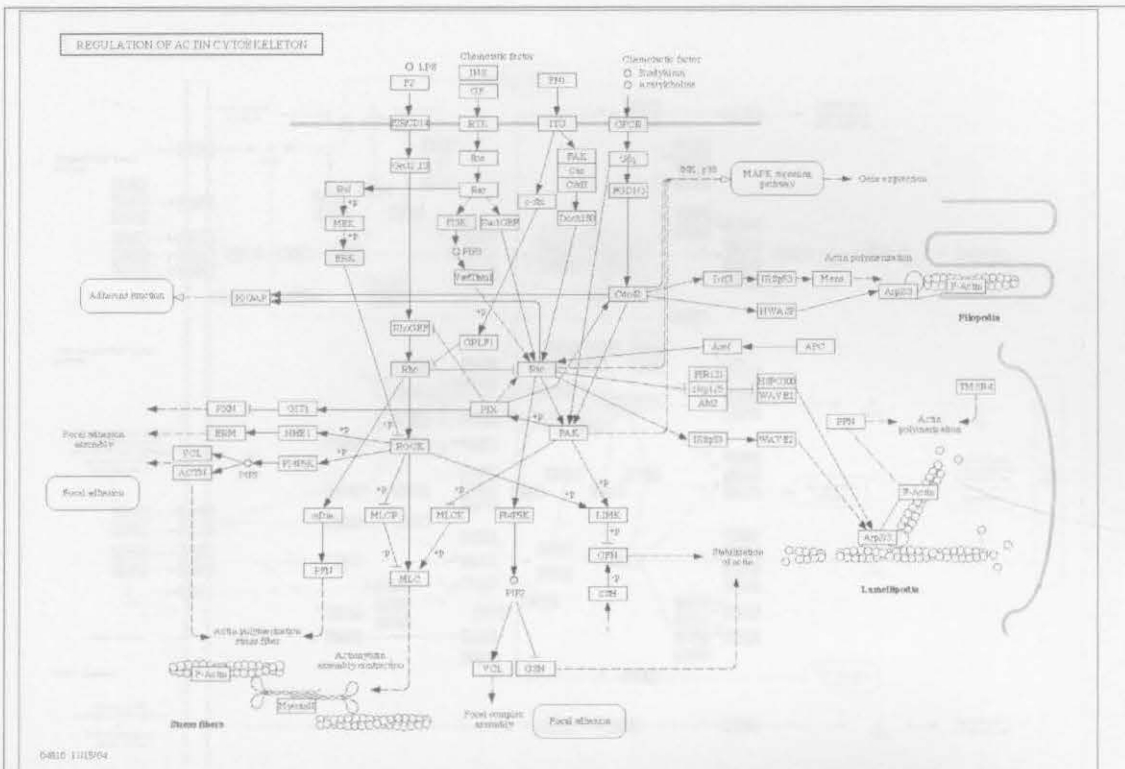
Figure 6.1 Reference KEGG pathway representing the regulation of actin cytoskeleton [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Regulation of nervous system development. Examining the mapping of the MAPK signaling transduction pathway (Figure 6.2) indicates the stress-induced materials, such as those downstream mitochondrial DNA damage signals, such as those downstream mitochondrial DNA damage signals, pro-caspase, FAS and TNFR, have activated neutrophil, JNK and P38 MAPK signaling pathways, respectively. The whole process involved the fate of neural precursor axon and dendrite growth patterns, the expression and activity of functionally important proteins through IC-induced bladder smooth muscle contraction status [Harper and LoGrasso, 2001].
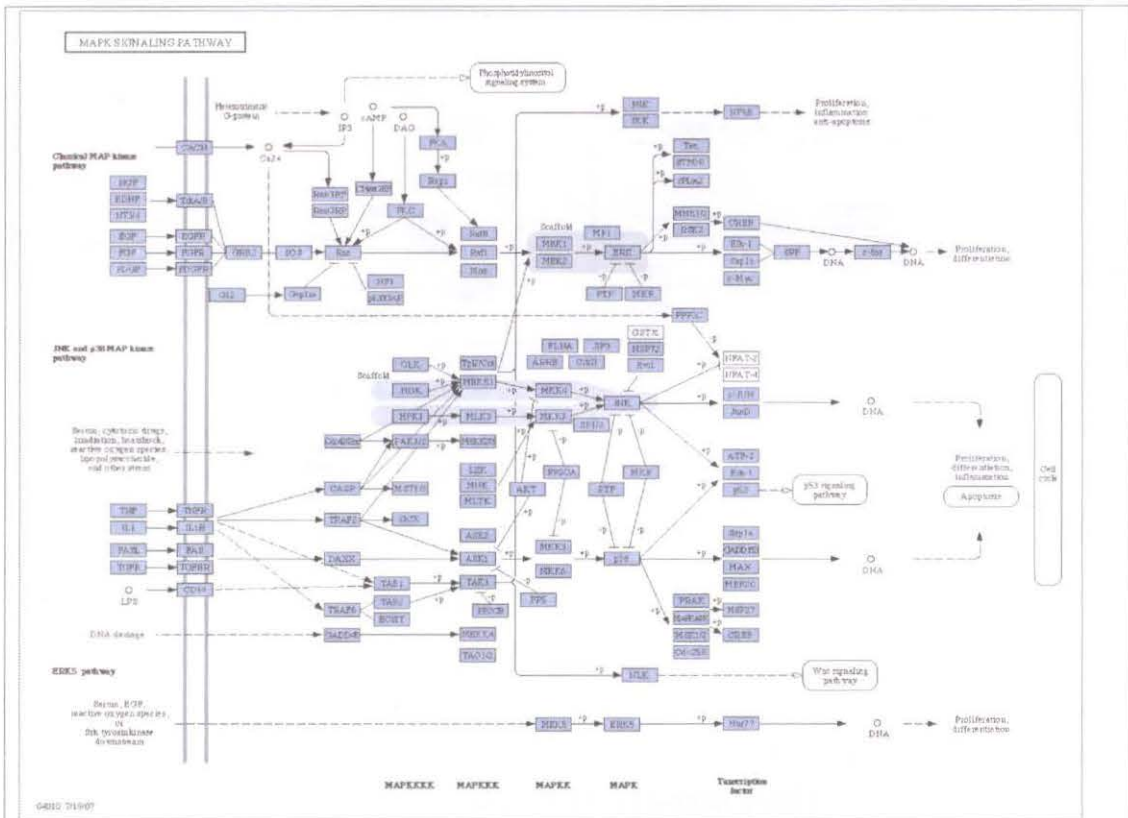
Figure 6.2 Reference KEGG pathway representing the regulation of MAPK signaling transduction [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Regulation of tissue homeostasis. Searching for the mechanisms of maintaining tissue homeostasis through IC status devoted in the discovery of TGF-beta signaling transduction system (Figure 6.3). The bone marrow proteins and activins are found to be activated through plasma membrane serine-threonine kinase receptors and cytoplasmic effectors in response to IC stimulus. The SMAD and non-SMAD proteins engage in specifying the vertral mesoderm, induction of mesoderm and endoderm, regulation of cell proliferation, cell differentiation, cell cycle and apoptosis through IC status [Deryncl and Zhang, 2003].
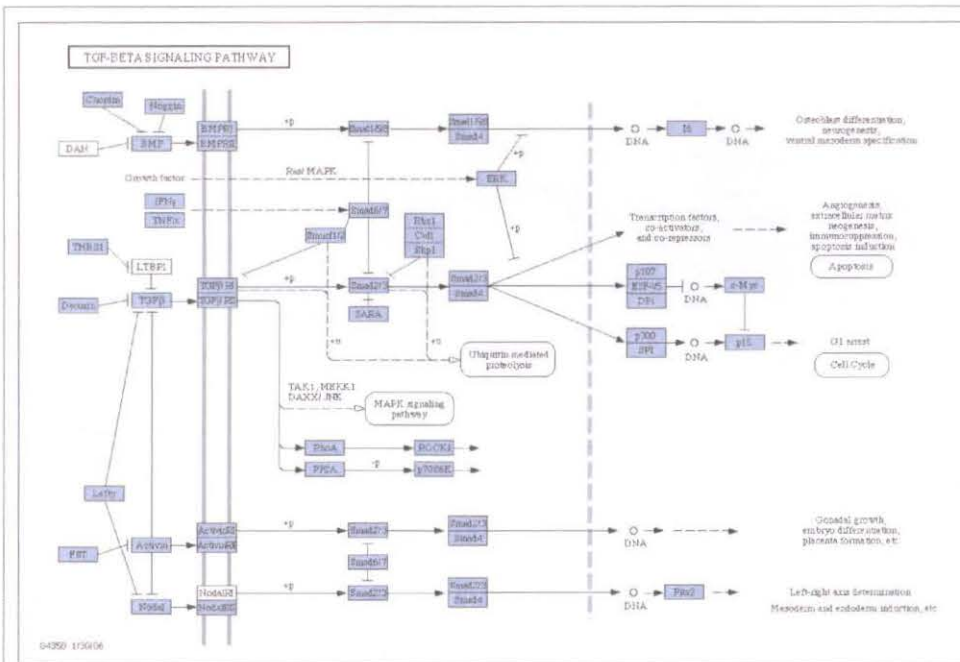
Figure 6.3 Reference KEGG pathway representing the regulation of TGF-beta signaling transduction [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Regulation of energy homeostasis. PPAR is a member of the nuclear receptor superfamily activated by various hydrophobic compounds [Takada and Kato, 2005]. According to the mapping representing the regulations of PPAR signaling transduction (Figure 6.4), PPARalpha was found to be activated in response to IC stimulus. Through PPAR/RXR heterodimetric complex, PPARalpha promotes the transport of long-chain fatty acids and retinoids X acids into mitochondrial that in turn stimulates the beta-oxidative degradation of fatty acids by the carnitine O-palmitoyltransferase I (CPTI) system [Takahashi et al., 2005]. While the mitochondria devote organelles in the cells and the sites of metabolic enzyme-specific reactions in the respiratory chains, it is also the site where oxygen acts as the final electron receptor in electron transport chain. In the presence of IC, mitochondrial oxidative stress could cause mitochondrial dysfunction, reduce production and triggering the apoptosis.

Figure 6.4 Reference KEGG pathway representing the regulation of PPAR signaling transduction [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Complement and coagulation cascades. According to the mapping on the pathway representing the regulation of complement and coagulation cascades (Figure 6.5), the classical and lectin pathways, respectively, are triggered to mediate complement activation through IC status. Complement activation products exert a number of effects on macrophage via specific surface receptors causing internalization of offending agents, microbes and immune complexes, promoting of intracellular killing, controlling migration behavior, inducing release of potent biological substances [Norris, 2003; Ruf et al., 2003; Campbell, 2003]. Thus, complement is also implemented as a pathogeneses of IC. In addition, blood coagulations are found to arise through the tissue factor (FC) pathway. The TF pathway triggers the production of Thrombin in response to IC stimulus.

Figure 6.5 Reference KEGG pathway representing the regulation of complement and coagulation cascades [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

Allergic inflammatory response. Finally, there is an indication of the mast-cell activation mediated by the high-affinity IgE receptor of Fc episilon RI on the basis of the mapping onto the pathway of Fc episilon RI signaling transduction (Figure 6.6). To date, it is well known that Fc episilon RI is considered as a key event in the allergy inflammatory response [Siraganian, 2003; Gilfillan and Tkaczyk, 2006]. Therefore, the activation of Fc episilon RI signaling transduction pathway has triggered allergic inflammation through IC status.

Figure 6.6 Reference KEGG pathway representing the regulation of Fc episilon RI signaling pathway [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

In conclusion, the pathogenic mechanisms of human IC can be summarised as follows: (1) complex metabolic reactions, including carbohydrate, lipid, cofactors, vitamins, xenobiotics, nucleotide and amino acid metabolisms, are found to have a strong association with bladder smooth muscle contraction through IC status, (2) we have found the transcriptional regulations of IC-induced bladder smooth muscle contraction status, including the level of contractile force, tissue homeostasis, energy homeostasis and the development of nervous system, and (3) there is an indication of the mast cell activation mediated by the high-affinity IgE receptor of Fc episilon RI triggering allergic inflammation through IC status. Such genetic alterations, jointly termed "bladder remodeling" can constitute an important long-term consequence of IC.

6.2.2 Asthma

Patients with asthma suffer from repeated episodes of airway narrowing that can cause shortness of breath along with wheezing and/or coughing. Airway passages clench up and contract as if they have been clamped in a vice [Stephenson, 2006]. Watery mucus spills into the upper respiratory system [Stephenson, 2006]. The lung

struggle and have to force the next breath [Stephenson, 2006]. It is warning to both the sufferer and to the parent with a child who is choking for air because of it.

Asthma is triggered by a respiratory track infection or allergic reaction-the body's immune response to a molecule or agent it seems an unwelcome or "foreign" [Stephenson, 2006]. Several molecules or agents can trigger an asthma attack. That which triggers an asthma attack in one patient, however, may not do so in anther. Dust mites, cockroach feces, mouse feces, cat or dog dander, pollen, wood smoke, smog, ozone, paint, exhaust, exercise, cold and heat can, in the susceptible person, precipitate an allergic reaction [Stephenson, 2006]. Inhalation of an allergen such as a cat dander of fecal material from dust mites triggers the release into the bloodstream of a special kind of antibody called IgE, which subsequently attach to inflammatory cells of the immune system [Stephenson, 2006]. That is, in asthma and other allergic disorders, the binding of IgE to the high-affinity Fc episilon receptor (Fc episilon RI) on mast cells activate signaling transduction pathway resulting in the release of histamines and other allergic inflammation mediators. Several pathological abnormalities, including airway inflammation and airway remodeling (thickening of the smooth muscle layer, increased epithelial cell mucus content, and subepithelial fibrosis), are generally seen in asthma [Fahy et al., 2000]. Though the clinical disease spectrum has been well established, little is known about its molecular mechanisms.

The advance of microarray technology and the completion of the human genome project offer a new opportunity to gain insight into global gene expression profiles in asthma, leading to the identification of mast cell products mediated via the high-affinity Fc episilon RI. Microarray analyses of the effects of targeted interventions, such as transgenic high expression of cytokines of gene targeting, could help to elucidate the contributions of specific molecules and pathways [Erle and Yang, 2003]. While Sayama et al [2002] reported that IL11 was the mast cell product mediated via the Fc episilon RI, their attempts remained undetermined for three reasons:

1. The results they presented were disconnected with the

acknowledged Fc episilon RI signaling pathway at the KEGG pathway [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

2. Sayama et al [2002] deposited their microarray data on the Gene Expression Omnibus under platform accession no GSE3271. We scrutinised the results they represented and did not find any clue providing evidence on the activation of IL11.

3. We all know that asthma and other allergic disorders are complex diseases. An individual gene (e.g. IL11) for superior intelligence probably does not exist as it does for asthma and other allergic disorders. Instead, genes and gene products work together in complex and dependent ways to cause allergic inflammation and anaphylactic reactions.

To actually identify new mast cell products, the Mahalanobis distance in the K-means clustering algorithm is used to group microarray gene expression profiling of the human umbilical cord blood-derived mast cells two hours post stimulation by the high-affinity IgE receptor of Fc episilon RI (Figure 6.7). The data used here was available from the Gene Expression Omnibus under platform accession no GSE3281, and this microarray experiment was performed by Sayama et al [2002]. At the finest level, we find that there is a cluster of 48 features that reveal considerable fold increase in expressions over baseline "o" levels in two human umbilical cord blood-derived mast cell populations (Figure 6.8). Thus, genes within this cluster are highly differentially expressed in the human umbilical cord blood-derived mast cells mediated via the high-affinity Fc episilon RI.
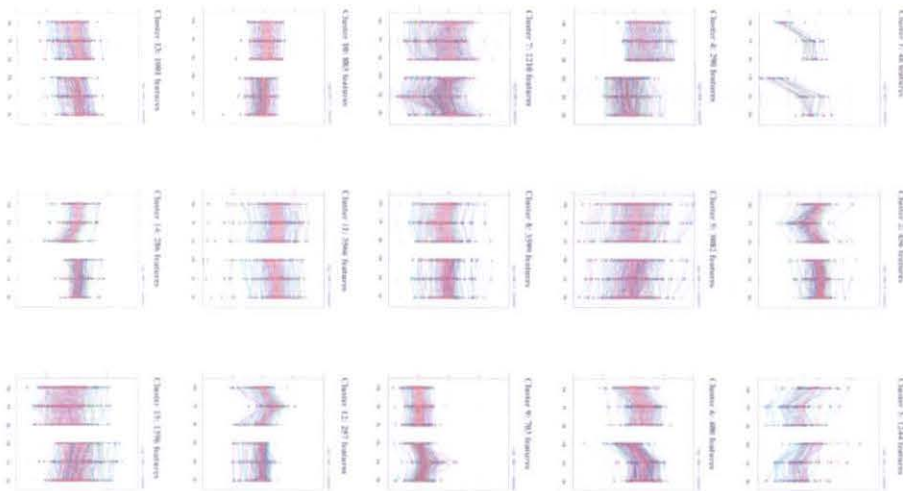
Figure 6.7 K-means clustering results of microarray gene expression profiles in the human umbilical cord blood-derived mast cells mediated by the high-affinity IgE receptor of Fc episilon RI (K=15).
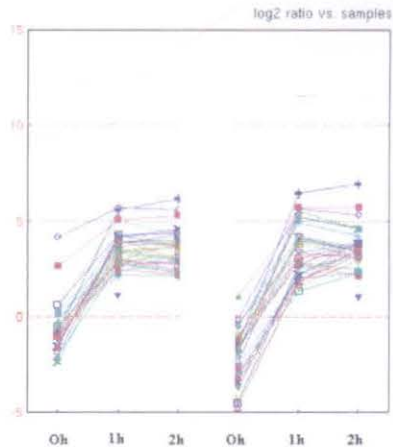
Figure 6.8 A cluster of 48 features that reveal considerable fold increase in expressions over baseline "0" levels in two human umbilical cord blood-derived mast cell populations.

Table 6.2 Highly differentially expressed genes in the human umbilical cord blood-derived mast cells stimulated via the high-affinity IgE receptor of Fc episilon

RI.

Gene Symbol	Gene Ontology
IL3	immune response; promote eosinophil production and activation
CSF2	immune response; promote eosinophil production and activation
TNF	inflammatory response; stimulates cytokines and chemokines; activates endothelium
ITGA2	cell adhesion
NR4A3	transcription
NR4A2	transcription
BTG2	transcription
SMARCA2	transcription
ZNF264	transcription
MIRN155	transcription
KPNA5	protein import into nucleus
CD69	activate T lymphocyte
CCL4	immune and inflammatory response
CXCL2	immune and inflammatory response
CCL3	immune and inflammatory response
CCL3L1	immune and inflammatory response
H62864	immune and inflammatory response
IL1A	immune and inflammatory response
IL8	immune and inflammatory response
LIF	immune response
JMJD3	inflammatory response
NFKBID	inflammatory response
GEM	immune response
MMP25	inflammatory response

To characterise their biological processes, we perform gene ontology on all features within this cluster with the help of Entrez Gene [Maglott et al., 2007]. This enables us to identify 24 distinct genes that are up-regulated in the human umbilical cord blood-derived mast cells mediated by the high-affinity IgE receptor of Fc episilon RI (Table 6.2). The nature of five cytokines (IL3, CSF2, TNF, IL1A and LIF) and six chemokines (CCL4, CXCL2, CCL3, CCL3L1, H62864 and IL8) have been altered through these processes. Several genes involved in

transcriptional activities are also up-regulated (NR4A3, NR4A2, BTG2, SMARCA2, ZNF 264 and MIRN155). Anti-IgE activated human mast cells also change the expression of a cell adhesion molecule (ITGA2) and additional genes involved in inflammatory responses (JMJD3, NFKBID, GEM and MMP25).

According to the mapping onto the Fc episilon RI signaling transduction pathway at the KEGG pathway [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008] (Figure 6.9), there is an indication that mast cell activation mediated via the Fc episilon RI can secret TNF, IL3 and CSF. In addition, the production of TNF promotes inflammation, activates endothelium as well as stimulates cytokines and chemokines; whereas the production of IL3 and CSF2 activates eosinophils [Turner and Kinet, 1999; Nadler et al., 2000; Nadler and Kinet., 2002; Gu et al., 2001; Kawakami et al., 2002; Siraganian, 2003; Gilfillan and Tkaczyk, 2006]. Such changes, jointly termed "airway remodeling," can constitute an important long-term consequence of asthma.
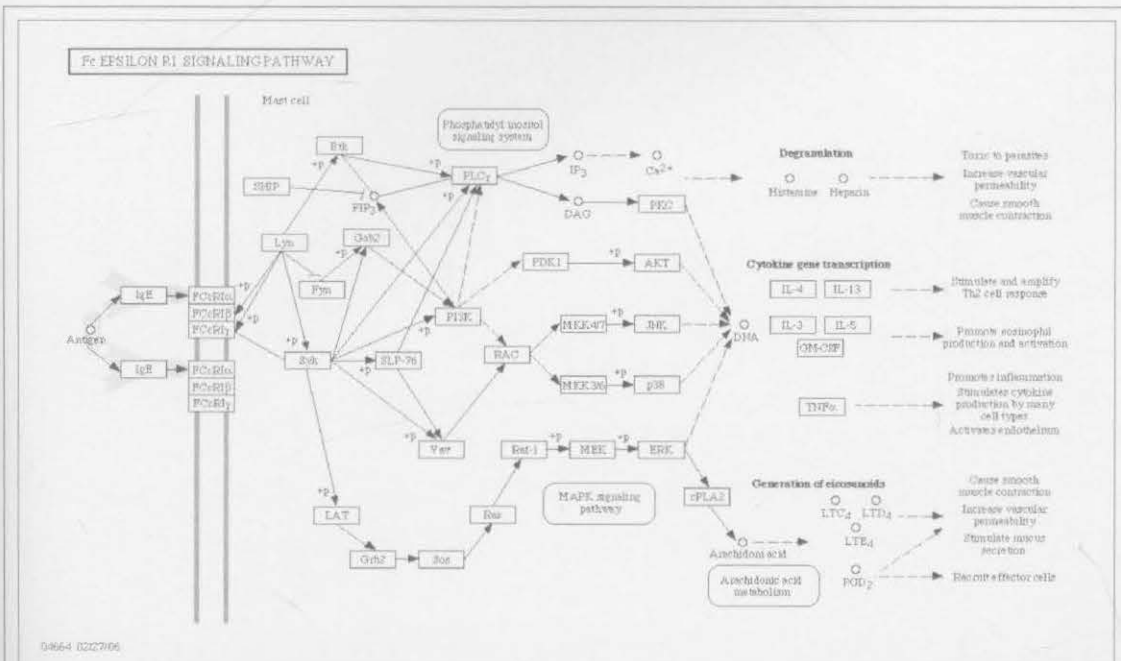
Figure 6.9 Reference KEGG pathway representing the regulation of Fc episilon RI signaling transduction [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

6.2.3 Obesity and Type 2 Diabetes

Suppose you bite off a piece of cake. As the cake is digested and its carbohydrate converted into the sugar glucose, the pancreas, a banana-shaped gland next to your stomach, squirts out a shot of insulin, helping to guide that sugar down the pathway that ends in its storage as fat or in its use as the gasoline for the engine that drives muscle movement [Stephenson, 2006].

As part of its function, insulin regulates glucose production and controls the movement of glucose into muscle cells through both direct and indirect effects on the liver [Edgerton et al, 2006]. If insulin fails to do its business, either because it is not made or because the body is unable to respond to it, glucose accumulates [Stephenson, 2006]. Lifted blood sugar (glucose), a condition called hyperglycemia, can lead to the onset of diabetes.

Generally speaking, diabetes comes in two versions, type 1 and type 2. While they are both the same disease, they start off by different ways. Type 1 diabetes is genetic and affects the patient from birth, whereas type 2 diabetes afflicts those who are obese.

Although obesity is a strong risk factor for the onset of type 2 diabetes, not all obese persons are diabetic. Actually, only around 10 percent of obese persons develop diabetes [Beck-Nielsen and Hother-Niesen, 1996]. Up to the present time, it is still much a mystery on what decide which obese non-diabetic individuals will transfer to diabetes.

To know what decide which obese non-diabetic individuals will transfer to diabetes, the Mahalanobis distance in Ward's linkage hierarchical clustering algorithm is used to group microarray gene expression profiling of adipose tissue from lean, obese and obese diabetic (hyperglycemia and severely diabetic) mice (Figure 6.10). The data used here was available from Gene Expression Omnibus under platform accession no GSE2952, and this microarray experiment was implemented by Nadler et al [2000]. At the finest level, we find that there is a cluster of 108 features that exhibit

substantial fold increase in the adipose tissue from obese diabetic mice (Figure 6.11). Thus, genes within this cluster are highly differentially expressed in the adipose tissue from obese diabetic mice.
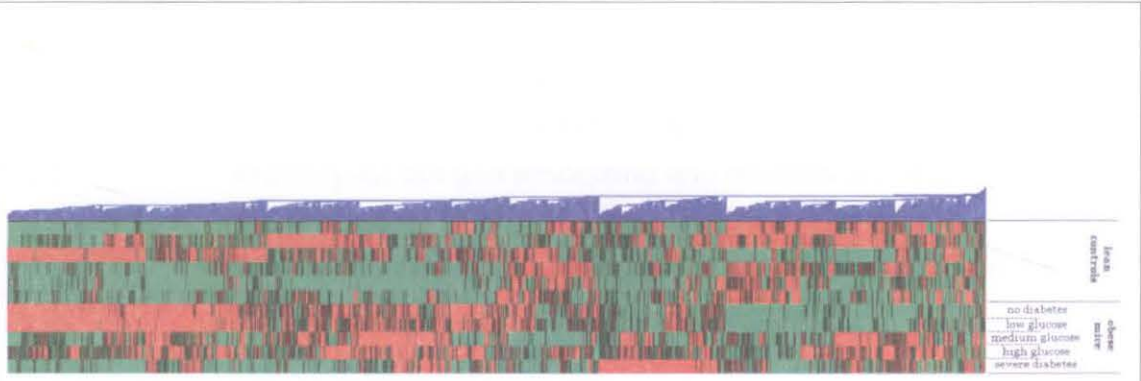
Figure 6.10 Clustering results of microarray gene expression profiling of adipose tissue from lean, obese and obese-diabetic (increasing hyperglycemia and severely diabetic) mice.
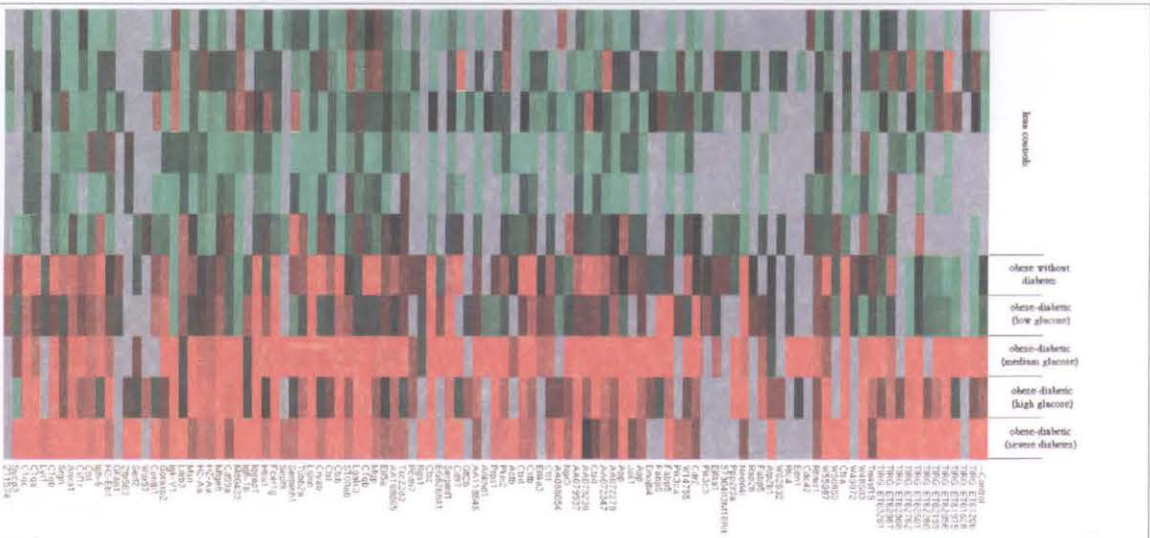
Figure 6.11 A cluster of 108 features that exhibit considerable fold increase in adipose tissue from obese diabetic (increasing hyperglycemia and severely diabetic) mice.

Table 6.3 Highly differentially expressed genes associated with the development of obese diabetes.

Gene Symbol	Gene ontology
Fabp5	glucose metabolism; lipid metabolism
Pkm2	glycolysis
H2-Aa	antigen processing and presentation (MHC class II)
H2-Eb1	antigen processing and presentation (MHC class II)
Ctsl	Proteolysis; antigen processing and representation (MHC class II)
Lgals3	IgE binding
Fcer1g	Fc epision RI pathway; regulation of mast cell cytokine production; immunoglobulin mediated immune response
Igk-V1	IgE mediated immune response
Igh-1b	IgE mediated immune response
Pik3c3	phosphoinositide-mediated signaling
Pik3ca	phosphoinositide-mediated signaling
Jak1	cytokine and chemokine mediated signaling pathway
C1qb	complement activation; classical pathway
Serpinf1	serine protease inhibitor activity
Serpinh1	serine protease inhibitor activity
Lilrb3	B cell mediated immunity
Mfge8	phagocytosis
Edn1	diuresis; natriuresis
Atp2b1	ion transport
Rfc4	DNA replication
Rbms1	DNA replication
Tsc22d3	transcription
Zfp503	transcription
Gtf2b	transcription; translational initiation
Eif4a1	translation
Eif4a3	RNA splicing, mRNA processing; rRNA processing
Eif5a	translation
Wdr57	RNA splicing, mRNA processing
Ankhd1	RNA binding; protein binding
Prps1	nucleotide biosynthesis
Tm4sf19	unknown
Serf2	unknown
EG626841	predicted gene

5730403M16Rik	hypothetical protein
Cdr2	biological process
Gorasp2	biological process
Dnajb4	protein folding
Nedd4l	ubiquitin ligase activity
Ppp2r3a	mitochondrial protein transporting ATP synthase complex
Ctsa	proteolysis
Cstb	proteolysis
Ctsd	proteolysis
Ctsz	proteolysis
Lgmn	proteolysis
Ctsa	proteolysis
Pcdh7	cell adhesion
S100a6	cell cycle; cell proliferation
App	Axonogenesis
Actb	F-actin
Msn	moesin; cytoskeleton protein binding
Mgp	cartilage development
Npr3	peptide receptor activity, G protein coupled
Cd97	G protein coupled receptor signaling pathway
Rgs1	G protein coupled receptor signaling pathway
Cdc42	Rho protein singal transduction; small GTPases mediated signaling transduction
Iqgap1	small GTPases mediated signaling transduction
Rab28	small GTPases mediated signaling transduction
Gkap1	G protein signaling, coupled to cGMP nucleotide second messenger
Tubb2a	GTPase activity; microtubule based movement
Centb1	regulation of ARF GTPase activity
Cryab	transmembrane receptor protein tyrosine kinase signaling pathway
Htra1	regulation of transmembrane receptor protein tyrosine kinase signaling pathway

To characterise their biological processes, we perform gene ontology on all features within this cluster with the help of Entrez Gene [Maglott et al., 2007]. This enables us to identify 62 distinct genes that are associated with the development of hyperglycemia (Table 6.3). Specifically, we observe changes in lipid and glucose metabolism

(Fabp5). These metabolisms require the release of energy, which is made available from mRNA encoding small GTPases domains, including Ras, CDC42 and Rho isoforms and ATP pmotor rotein [Kosztin et al., 2002]. The expressions of MHC class II molecules are up-regulated through these processes, suggesting that MHC class II expressions have a strong association with the progress of obese diabetes. We also observe the expression of Endothelin-1 (Edn1) is up-regulated through these processes. According to Hoffman et al [2000], the diuresis and natriuresis are caused the differential effect of Edn1.

It can be observed that obese diabetes initiate classical pathway of complement cascades promoting muscle contraction, phagocyte recruitment and B cell mediated immune responses [Kirschfink, 1997; Bhole D and Stahl, 2003]. Several genes involved in potential interactions with the binding of IgE to the high-affinity Fc episilon RI and the activation of mast cells are up-regulated through these processes, suggesting that obese diabetes trigger the binding of IgE to the high-affinity Fc episilon RI and the activation of mast cells causing allergic reactions [Gilfillan, A.M. and Tkaczyk, 2006; Zhang et al, 2007]. Changes in ubiquitin proteasome system are also observed. Furthermore, we observe changes in the regulation of Wnt receptor (Csnk2a2) and TGF-beta receptor signaling pathway (Cryab and Htra1), indicating the potential roles of Wnt receptor and Tgf-beta receptor signaling pathways are involved in the development of obese diabetes.

In conclusion, the pathogenic mechanisms underlying the progress of obese diabetes include: (1) MHC class II expressions are strongly linked to the susceptibility of the obese diabetes, (2) the diuresis and natriuresis are caused by the differential effect of Edn1, (3) obese diabetes initiate classical pathway of complement cascades promoting muscle contraction, phagocyte recruitment and B cell mediated immune responses, (4) changes in ubiquitin proteasome system are observed, (5) obese diabetes promote the binding of IgE to the high-affinity Fc episilon RI and the activation of mast cells triggering allergic reactions, and (6) Wnt receptor and Tgf-beta receptor signaling pathway are involved in the onset of obese diabetes.

6.2.4 HIV/AIDS

The human immunodeficiency virus (HIV) infection originated from a natural infection of chimpanzees [Barnes, 2005]. The virus evolved into a silent infection causing no harm within its natural host [Barnes, 2005]. Chimpanzees that were hunted and slaughtered for food by rural people gave the virus the opportunity to jump species and infected humans during the butchering process [Barnes, 2005]. It would only take a kin scratch or nick in contract with infected blood or other body fluids of the animal to permit infection [Barnes, 2005]. Most likely the virus occasionally infected rural people in isolated area of the Conge with deadly results for many before its escape from total central Africa during the latter half of the twentieth century [Barnes, 2005]. The virus quickly adapted to human beings to turn into a human disease transmitted through intravenous drug abuse, blood transfusions and sexual contact.

But the world did not know what was taking place in Africa. People outside Africa mostly ignored reports of the mysterious of HIV infection and the disease it caused, acquired immune deficiency syndrome (AIDS), despite the fact that tens of thousands of people were killed in this part of the world [Barnes, 2005]. Many Africans blamed witchcraft for the carnage, and African government denied the seriousness of epidemic [Garrett, 1994]. At the same time, the global blood market and illicit heroin traffic greatly facilitated movement of this virus to different regions of the globe.

Until the summer of 1981, the recognition of AIDS put the world on notice of the impending HIV pandemic [Quagliarell, 1982; Jaffe et al, 1983; Mansell et al, 1984]. The symptom was first described as Kaposi's sarcoma, an unusual form of skin cancer, for which they were a symptom [Mansell et al, 1984; Longo et al, 1983]. Physicians also began to take note of an unusual clustering of cases of rare Pneumocystis carinii pneumonia [Pass et al, 1983; Wollschlager et al, 1984]. Both Kaposi's sarcoma and the Pneumocystitis form of pneumonia were opportunistic diseases; they were usually seen only with patients whose immune systems have been compromised

[Stephenson, 2006]. The patients were all found to be suffering from a substantial reducing in the number of circulating T cells of a type known as CD4+ and their immune defense had collapsed [Stephenson 2006].

HIV coming from central Africa has been labeled HIV-1 to distinguish it from another HIV type from west Africa, HIV-1, first identified in 1985 [Marlink, 2001]. The virus from west Africa bears a strong relationship with Mangabey monkeys, macaques, green monkeys and mandrills [Strauss and Strauss, 2002]. This indicates that humans picked up HIV-2 from monkeys in West Africa separately from transmission of HIV-1 in central Africa from chimpanzees to humans. Like HIV-1 viruses, HIV-2 viruses probably jumped species from monkeys to humans in isolated rural areas more than once over many years [Barnes, 2005]. But HIV-2 most remains confined to west Africa, and it follows a far less virulent path in humans than HIV-1.

HIV-1's target is a white blood cell called a T cell-one of the very defenders the immune system enlists to engage viruses when they appear [Stephenson, 2006]. On the surface of a T cell are receptors used to receive communication from other cells and to snag other proteins or molecules, very specifically, for transport across the cell membrane [Stephenson, 2006]. HIV-1 uses CD4 receptors and CCR5 receptors to tightly dock with its victim T cell [Liu et al, 1996; Abbas et al, 1997]. With attachment, the outer coat of HIV fuses with the outer membrane of its host and the virus spills its contents into the cell [Liu et al, 1996; Abbas et al, 1997].

On average, it takes eight years from the time of initial HIV-1 infection to the onset of full-brown AIDS [Stephenson, 2006]. Without the protection of T cells, all the other microbes, constantly at the gate, finally breach what is left of the body's defenses [Stephenson, 2006]. The AIDS patient finally succumbs to any number of other diseases such as pneumonia, tuberculosis or cancer against which the body can no longer mount resistance [Stephenson, 2006].

Although it is well known that changes in T cell function are a hall mark of HIV-1 infection, the pathogenic mechanisms leading to these

changes are still relatively unknown [Hyrcza et al, 2007]. To identify the pathogenic mechanisms leading to these changes, the Mahalanobis distance in Ward's linkage hierarchical clustering algorithm is used to group microarray gene expression profiling of ex vivo human CD4+ T cells from untreated HIV-1 patients at different disease stages and rates of disease progression (Figure 6.12). The data used here was available from Gene Expression Omnibus under platform accession no GSE6740, and this microarray experiment was performed previously by Hyrcza et al [2007]. At the finest level, we find there is a cluster of 328 features that exhibit considerable fold increase in the CD4+ T cells from early ($\leq$6 months) untreated HIV-1 patients (Figure 6.13). Thus, genes within this cluster are highly differentially expressed in the CD4+ T cells from early untreated HIV-1 patients.

early HIV-1 infection

chronic HIV-1 infection

non-progressive HIV-1 infection

uninfected
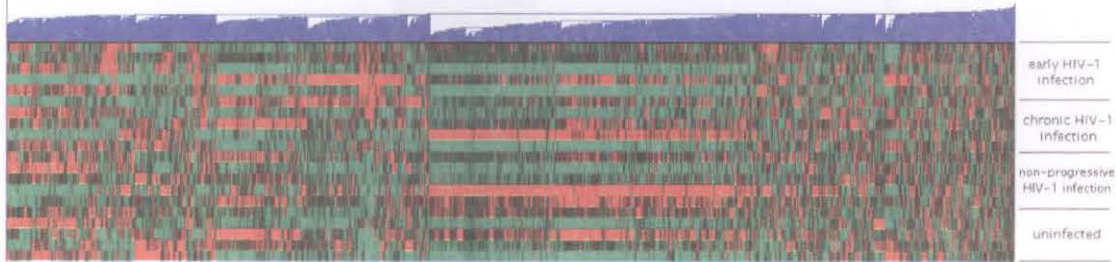
Figure 6.12 Clustered display of transcriptional responses in ex vivo human CD4+ cell from untreated HIV-1 patients at different disease stages and rates of disease progression.
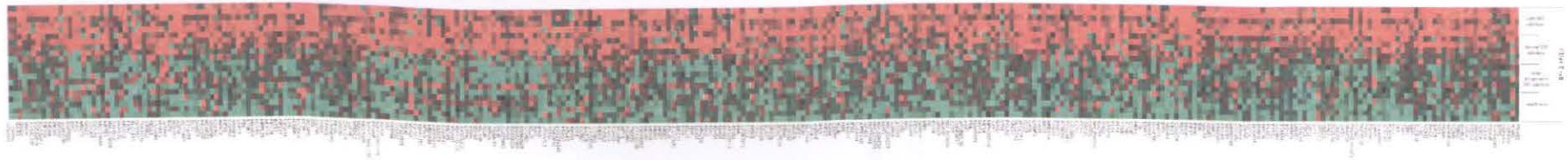
Figure 6.13 A cluster of 293 features that exhibited considerable fold increases in ex vivo human CD4+ cell from patients infected with HIV-1 within 6 months of study.

Table 6.4 Highly differentially expressed genes in ex vivo human CD4+T cell from patients infected with HIV-1 within 6 months of study.

Gene symbol	Gene ontology
TOP2A	retroviral genome replication
ISG15	response to virus
MX1	response to virus
PLSCR1	response to virus
RSAD2	defense response to virus
AURKA	mitosis
KIF2C	mitosis
NCAPH	mitosis
PBK	mitosis
TPX2	mitosis
MRE11A	meiosis
RAD51	meiosis
MSH5	meiosis
ESPL1	spindle organization and biogenesis during meiosis
TTK	spindle organization and biogenesis during meiosis
ASPM	cell cycle
AURKB	cell cycle
BUB1	cell cycle
BIRC5	cell cycle
DCLRE1A	cell cycle
FANCI	cell cycle
HELLS	cell cycle
SMC4	cell cycle
SPC25	cell cycle
CCNA2	cell cycle
CCNB1	cell cycle
CCNB2	cell cycle
CCNF	cell cycle
CHEK1	cell cycle
CENPE	cell cycle
CKAP2	cell cycle
CHAF1B	cell cycle
CDC2	cell cycle
CDC6	cell cycle
CDC20	cell cycle

CDC25A	cell cycle
CDC45L	cell cycle
CDCA3	cell cycle
CDCA8	cell cycle
CDKN3	cell cycle
CDT1	cell cycle
CEP55	cell cycle
DLGAP5	cell cycle
E2F8	cell cycle
KIF11	cell cycle
KIF23	cell cycle
KIFC1	cell cycle
MCM2	cell cycle
MCM4	cell cycle
MCM5	cell cycle
MCM6	cell cycle
MCM10	cell cycle
NCAPG2	cell cycle
NUSAP1	cell cycle
FBXO5	cell cycle
GMNN	cell cycle
MKI67	cell cycle
NEK2	cell cycle
NCAPG	cell cycle
NDC80	cell cycle
NUSAP1	cell cycle
PTTG1	cell cycle
SMC2	cell cycle
UBE2C	cell cycle
ZWINT	cell cycle
ZWILCH	cell cycle
MPHOSPH9	M phase of mitotic cell cycle
GTSE1	G2 phase of mitotic cell cycle
CENPF	G2 phase of mitotic cell cycle
BARD1	cell cycle arrest
BUB1B	cell cycle; cell cycle checkpoint
BRCA1	cell cycle checkpoint
RBBP8	cell cycle checkpoint

PSMA4	cell cycle checkpoint; signaling by WNT
PML	DNA damage response, signal transduction by P53 class mediator resulting in cell cycle arrest
FANCL	DNA repair
DTL	DNA replication
GINS1	DNA replication
GINS2	DNA replication
ORC1L	DNA replication
POLE2	DNA replication
RRM2	DNA replication
TREX1	DNA replication
HIST1H1E	nucleosome assembly
ANXA4	anti-apoptosis; signal transduction
XAF1	apoptosis
LAMP3	cell proliferation
MYF6	cell proliferation
LMNB1	cell communication
OIP5	cell communication
GEMIN6	metabolism of non-coding RNA
CHI3L2	carbohydrate metabolism
GK	carbohydrate metabolism
GALK1	carbohydrate metabolism
SUCLA2	carbohydrate metabolism
EPRS	amino acid metabolism
HIBCH	amino acid metabolism
PCBD1	amino acid metabolism
PSMB2	amino acid metabolism
PSMA5	amino acid metabolism
WARS	amino acid metabolism
TK1	nucleotide metabolism
OAS1	nucleotide metabolism
TYMS	nucleotide metabolism
TYMP	nucleotide metabolism
OAS1	nucleotide metabolism
OAS2	nucleotide metabolism
DHFR	metabolisms of vitamins and cofactors
CD38	metabolisms of vitamins and cofactors
ACOT7	lipid metabolism

APOL1	lipid metabolism
ETNK1	lipid metabolism
FAR2	lipid metabolism
FDXR	lipid metabolism
SLC27A2	lipid metabolism
MANEA	glycoprotein glucosylmannohydrolase activity
IVD	metabolism
KIAA0859	metabolism
MT1E	biological process
SPTLC2	biosynthetic process
DHRS7B	oxidation reduction
PRDX3	oxidative reduction
ATP6V0E1	transport (ATP hydrolysis coupled proton)
ATP6V1D	transport (ion, proton)
PCTP	transport (lipid)
TAP1	transport (oligopeptide)
SLC1A4	transport (glutamate/neutral amino acid)
SLCO4A1	transport (ion)
SLC6A8	transport (neurotransmitter)
SLC13A1	transport (sodium ion, sulfate)
HBB	transport (oxygen)
AP3D1	intracellular protein transport
HSPA1A	response to stress; ATP binding
DDX60	ATP binding
NVL	ATP binding
GIMAP4	GTP bindig
ESPN	actin binding
NUDT1	response to oxidative stress
TRAFD1	metal ion binding; zinc binding; protein binding
NTN2L	axon guidance
KIF4A	anterograde axon cargo transport
TUBG1	microtubule cytoskeleton organization and biogenesis
KIF14	microtubule based movement
KIF15	microtubule based movement
KIF18A	microtubule based movement
KIF20A	microtubule based movement
MYL6B	muscle filament sliding
AHNAK	nervous system development

SPAG6	cell projection organization
CELSR3	cell adhesion
PXN	cell adhesion
RARRES1	cell proliferation
GPR19	G-protein coupled receptor signaling pathway
GPR107	G-protein coupled receptor signaling pathway
GPR171	G-protein coupled receptor signaling pathway
GPSM2	G-protein coupled receptor signaling pathway
ARL4C	small GTPases mediated signal transduction
HMMR	ECM-receptor interaction
LY6E	cell surface receptor linked signal transduction
CCR1	immune and inflammatory response
CCR5	immune and inflammatory response
MR1	immune response
AIM2	immune response
GBP1	immune response
IFI6	immune response; anti-apoptosis
IFI16	DNA damage response, signal transduction by P53 class mediator resulting in cell cycle arrest
IFI35	immune response
IFI44	immune response
IFI44L	immune response
IFIT3	immune response
EDN3	regulation of MAP kinase
TNFSF10	immune response;NF-kappaB and MAPK8/JNK activation
TNFRSF17	immune response;NF-kappaB and MAPK8/JNK activation
ECT2	regulation of NF-kappaB cascade
STAT1	regulation of NF-keppa B cascade
LGALS9	regulation of NF-keppa B cascade
IGL@	immune response
IGLV@	immune response
IGLJ3	immune response
IGHM	immune response
IGHG1	immune response
IGHG3	immune response
IGKC	immune response
IGKV1D-13	immune response
IGKV1OR15-118	immune response

IGKV1OR2-108	immune response
OASL	immune response
OAS3	immune response
CST5	thiol protease inhibitor
MELK	protein amino acid phosphorylation
PLK4	protein amino acid phosphorylation
CPM	Proteolysis
LXN	perception of pain, sensory transduction of temperature stimulus
FBXO38	ubiqintin mediated protein catabolism
HERC6	ubiqintin mediated protein catabolism
USP18	ubiqintin mediated protein catabolism
CENPA	nucleosome assembly
HJURP	DNA binding
TIMM17A	protein targeting to mitochondria
PAQR4	progestin and adipoQ receptor
APOBEC3B	biological process
IFIT1	biological process
IFI27	biological process
C1orf112	open reading frame
C3orf14	open reading frame
C5orf42	open reading frame
C6orf142	open reading frame
C11orf75	open reading frame
C12orf48	open reading frame
ATXN7L1	unknown
BTN2A2	unknown
CENPM	unknown
CLDND1	unknown
LRP5L	unknown
NT5DC2	unknown
PCNAP	unknown
TMEM156	unknown
KIAA0101	hypothetical protein
KIAA1751	hypothetical protein
LOC284244	hypothetical protein
MGC5370	hypothetical protein
MGC29506	hypothetical protein

MIRN650	microRNA
EXOSC4	RNA processing
INTS7	snRNA processing
ASF1B	transcription
BAZ1B	transcription
CARHSP1	transcription
MED18	transcription
EZH2	transcription
FHL2	transcription
EYA4	transcription
FOXM1	transcription
HOXB7	transcription
HIVEP3	transcription
MED8	transcription
MLF1IP	transcription
MYBL2	transcription
PKNOX1	transcription
POU2AF1	transcription
SATB2	transcription
WDHD1	transcription
WHSC1	transcription
ZBTB32	transcription
PALM2-AKAP2	co-transcription
MRPL35	translation
TSFM	translation

To characterise their biological processes, gene ontology within this cluster was performed on all features using Entrez Gene [Maglott et al., 2007]. This enabled us to identify 247 distinct genes that are highly differentially expressed in the CD4+ cells through early HIV-1 infectious status (Table 6.4). Particularly, changes in carbohydrate, lipid, cofactors, vitamins, nucleotide, amino acid and non-coding RNA metabolisms are observed. These metabolisms require the release of energy, which is made available for mRNA encoding small GTPases and ATP proteins [Kosztin et al., 2002]. The expressions of interferon-induced proteins (IFI6, IFI16, IFI35, IFI44, IFI44L and IFIT3) are up-regulated in the CD4+ T cells through early HIV-1 infectious status. The expressions of IgH (IGHM, IGHG1 and IGHG3),

IgK (IGKC, IGKV1D-13, IGKV1OR15-118 and IGKV1OR2-108) and IgL (IGL@ and IGLV@) molecules are up-regulated in the CD4+ T cells through early HIV-1 infectious status.

It can be observed that several genes involved in potential interactions with NF-kappaB activation are up-regulated in the human CD4+ T cells through early HIV-1 infectious status. We also observe changes in potential interactions with cell cycle progression through early HIV-1 infectious status, suggesting that early HIV-1 infection changes cell cycle progression in vivo promoting virus production. It can be observed that HIV-1 genome replication is regulated by topoisomerase II, alpha isozyme (TOP2A) as well. In addition, several genes involved in the regulation of Netrin-2 like (NTN2L) mediated axon guidance are up-regulated in the CD4+ T cell through early HIV-1 infectious status [Dickson, 2002], indicating that Netrin-2 like mediated axon guidance promote T cell turnover through early HIV-1 infectious status.

In summary, gene expression profiling identified the pathogenic mechanisms underlying changes in T cell function through early HIV-1 infectious status, which include: (1) carbohydrate, lipid, cofactors, vitamins, nucleotide, amino acid and non-coding RNA metabolisms have a direct association with changes in T cell function, (2) the expressions of interferon-induced protein, IgH, IgK and IgL genes are up-regulated in the human CD4+ T cells, (3) genes involved in NF-kappaB activation are up-regulated in the human CD4+ T cells (4) early HIV-1 infection changes cell cycle progression in vivo promoting virus production, (5) Netrin-2 like mediated axon guidance promote T cell turnover, and (6) the transcriptional evidence that CD4+ T cells in early untreated HIV-1 patients are activity responsible for viral control, whose genome replication are regulated by DNA topoisomerase II, alpha isozyme (TOP2A).

6.2.5 *Chlamydia Pneumoniae* Infection

C. pneumoniae, the obligate intracellular gram-negative bacterium, causes acute respiratory infections in humans and mice [Yang et al., 1993]. The normal route of entry of this bacterium is oral and nasal

mucosa [Rodríguez et al., 2007]. Infection with *C. pneumoniae* has been related to asthma, chronic obstructive pulmonary disease and also to non-respiratory disease like atherosclerosis [Clementsen et al., 2002; Johnston, 2002; Belland et al., 2004; Grayston, 2005]. It is widely distributed in the population, and up to 50 percent of the people of the developed world are sero-positive by the age of 20 years [Rodríguez et al., 2007]. Nonetheless, the host immune defense mechanisms triggered by *C. pneumoniae* are still relatively unknown.

Although Rodríguez et al. [2007] reported that myeloid differentiation factor-88 (MYD88) and toll-like receptors play a critical role in inducing immunity against *C. pneumoniae*, their attempts remained undetermined for two reasons:

1. Rodríguez et al. [2007] deposited their microarray data on the Gene Expression Omnibus under platform accession no GSE6690. We inspected the resulted they represented and did not find any clue providing evidence on the activation of toll-like receptors and toll like receptor signaling transduction pathway.
2. We all know that *C. pneumoniae* infection is a complex disease. An individual gene (e.g. MYD(9) for superior intelligence probably does not exist as it does for *C. pneumoniae* infection. Instead, genes and gene products work together in complex and dependent ways to cause pathogeneses underlying *C. pneumoniae* infection.

To actually find the host immune defense mechanisms underlying *C. pneumoniae* infection, the Mahalanobis distance in K-means clustering is used to organise microarray gene expression profiling of the macrophages of C57BL/6 mice after infection with *C. pneumoniae*. The data used here was available from Gene Expression Omnibus under platform accession no GSE6690, and this microarray experiment was performed by Rodríguez et al. [2007]. We consider thirteen group arrangements corresponding to the measured fluorescence intensity of the macrophages after infection with *C. pneumoniae* versus uninfected controls (Figure 6.14). At the finest level, we find that there is a cluster of 60 features that reveal considerable fold increase in the macrophages after infection with *C. pneumoniae* (Figure 6.15). Thus, genes within this cluster are highly

differentially expressed in the macrophages after infection with *C. pneumoniae*.
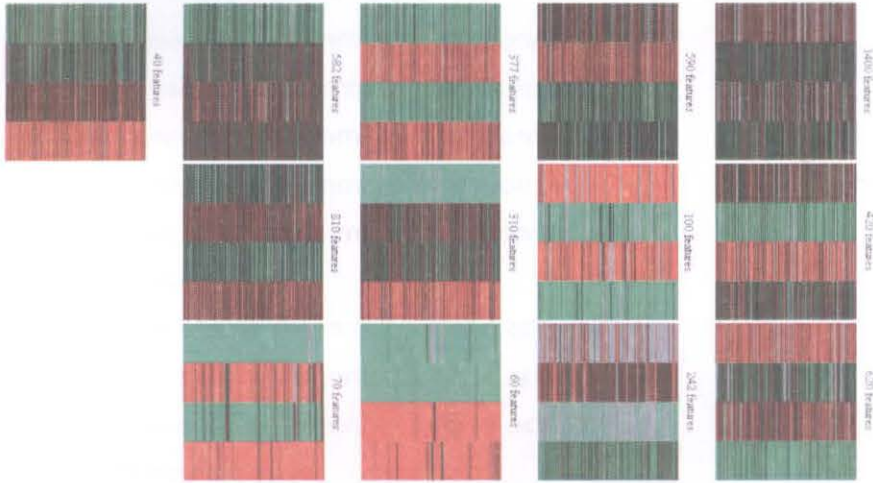
Figure 6.14 K-means clustering results of transcriptional responses in the macrophage of 57BL/6 mice after infection with *C. pneumoniae* versus uninfected controls (K=13).
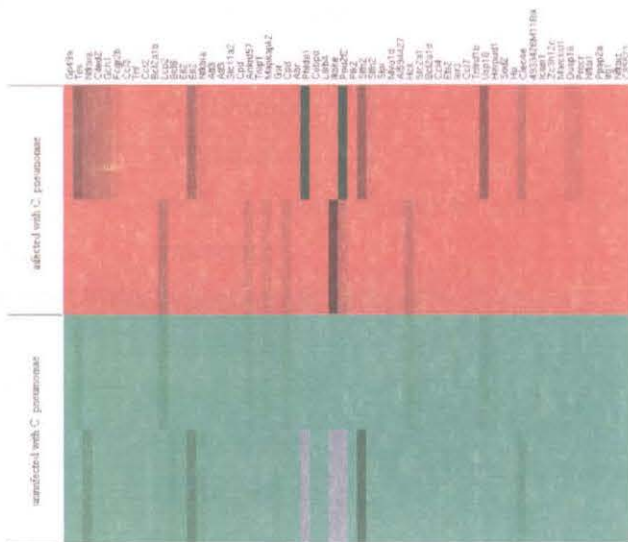
Figure 6.15 A cluster of 60 features that reveal substantial fold increase in the macrophages of 57BL/6 mice after infection with *C. pneumoniae*.

Table 6.5 Highly differentially genes in the macrophages of 57BL/6 mice after

infection with *C. pneumoniae.*

Gene symbol	Gene ontology
Fcgr2b	IgE mediated immune response; IgE binding
Pou2f2	IgE secretion during immune response
Lcp2	mast cell activation; cytokine secretion
Gp49a	mast cell surface glycoprotein precursor
Ccl2	immune and inflammatory response
Ccl3	immune and inflammatory response
Ccl4	immune and inflammatory response
Ccl7	immune and inflammatory response
Bcl6	immune and inflammatory response
Tnf	immune and inflammatory response
Dusp16	apoptosis
Mapkapk2	protein serine-threonine kinase activity
Plk2	protein serine-threonine kinase activity
Slpi	protein serine-threonine kinase activity
Ikbke	immune response
Lilrb4	immune response
Clec4e	immune response
Gch1	regulation of lung blood pressure
Cited2	blood vessel development
Icam1	leukocyte adhesion
Marcksl1	cell proliferation
Slfn2	cell proliferation
Myo1d	ATP binding
Hck	ATP binding
Abr	small GTPases mediated signaling transduction
Ankrd57	unknown
4933426M11Rik	hypothetical protein
Tnip1	translation
Atf3	transcription
Cebpd	transcription
Ell2	transcription
Ets2	transcription
Tes	transcription
Zc3h12c	protein binding
Cpd	proteolysis
Gsr	amino acid metabolism

Slc2a1	transport (facilitated glucose)
Slc11a2	transport (cobalt ion, iron ion)
Hp	response to reactive oxygen species
Sod2	mitochondrial organization and biogenesis
	respiratory electron transport chain
Nfkbia	ubiquitin-mediated proteolysis;
Usp18	ubiqintin-dependent protein catabolism
Herpud1	protein modification process
Bcl2a1b	apoptosis
Bcl2a1d	apoptosis
Ier3	apoptosis
Phlda1	apoptosis
Tnfrsf1b	apoptosis

To characterise their biological processes, we perform gene ontology on all features within this cluster using Entrez Gene [Maglott et al., 2007]. This enables us to identify 49 distinct genes that are highly differentially expressed in the macrophages after infection with *C. pneumoniae* (Table 6.5). In particular, we observe changes in a cytokine (Tnf) and four chemokines (Ccl2, Ccl3, Ccl4, and Ccl7) through these processes. We also observe changes in ubiquitin proteasome. Changes in electron transport chain, response to reactive oxygen species (ROS) and apoptosis are also observed, suggesting that *C. pneumoniae* infection lead to the deleterious effects of oxygen from the metabolic reduction of the highly active and toxic ROS and subsequently promote endothelial damage and atherosclerosis [MJ Nadler and JP Kinet, 2002].

C. pneumoniae infection also promotes IgE secretion (Pou2f2), the binding of IgE to the high-affinity Fc episilon RI (Fcgr2b), and the activation of mast cell (Lcp2), suggesting that the binding of IgE to the high-affinity Fc episilon RI and the activation of mast cell promote allergic inflammation in the macrophages of C57BL/6 mice triggered *C. pneumoniae*. The activation of mast cell mediated via the high-affinity Fc episilon RI trigger the secretion of Tnf. The production of Tnf promotes allergic inflammation, activates endothelium as well as stimulates cytokines and chemokines in response to *C. pneumoniae* infection [Vivier et al., 2004; Bonnema and Leibson, 1996].

Furthermore, according to the mapping onto the KEGG reference pathway representing the regulation of natural killer cell mediated cytotoxicity [Perussia, 2000], it can be observed that natural killer cells are activated triggering cytotoxicity, secreting cytokines and chemokines in the macrophages of C57BL/6 mice triggered *C. pneumoniae*. As a result, *C. pneumoniae* infection initiates signaling transduction pathway regulation NK cell activation –induced host immunity against this microorganism.
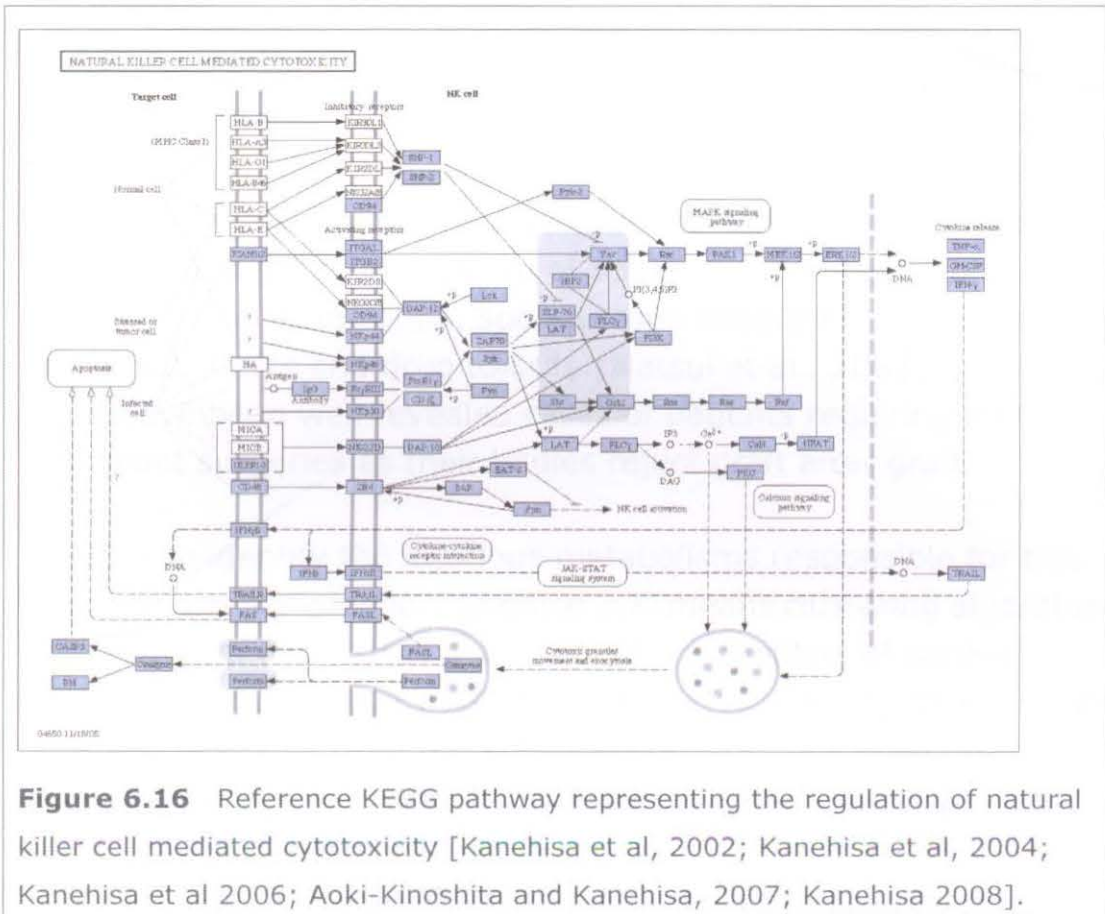
Figure 6.16 Reference KEGG pathway representing the regulation of natural killer cell mediated cytotoxicity [Kanehisa et al, 2002; Kanehisa et al, 2004; Kanehisa et al 2006; Aoki-Kinoshita and Kanehisa, 2007; Kanehisa 2008].

To sum up, the host immune defense mechanisms in the macrophages of C57BL/6 mice after infection with *C. pneumoniae* can be summarised as follows: (1) *C. pneumoniae* infection lead to the deleterious effects of oxygen from the metabolic reduction of the highly active and toxic reactions oxygen species, which promote endothelial damage and atherosclerosis, (2) changes in ubiquitin proteasome are observed, (3) *C. pneumoniae* infection initiate natural killer cell mediated cytotoxicity, (4) *C. pneumoniae* infection cause the binding of IgE to the high-affinity Fc episilon RI and the activation of

mast cell promoting Tnf-mediated allergic inflammation; such changes termed "airway remodeling" can constitute an important consequence of *C. pneumoniae* –induced asthma.

6.2.6 Cardiac Allograft Rejection

Heart and lung transplants, introduced into clinical practice in 1981, have been performed in over 20,000 patients worldwide for whom other effective medical therapy was not available [Masters et al., 1999]. Though the administration of immunosuppressive therapy has enabled the early survival of cardiac transplant recipients, patients still have a 30 percent risk of rejecting the grafts during the first two years following transplantation [Masters et al., 1999; Matsui et al., 2003]. The administration of immunosuppressive therapy also suppresses the immune system of patients and entails many problems such as infection, spontaneous neoplasm, undesirable metabolic effects and drug toxicity [Matsui et al., 2003]. In addition, there have been well revealed cases of patients requiring multiple transplant surgeries as their bodies reject graft after graft.

In order to identify the injurious metabolisms responsible for this rejection, the Mahalanobis distance in K-means clustering algorithm is used to group microarray gene expression profiling of cardiac allografts from BALB/c to C57BL/6 mice (Figure 6.17). The data used here was available from Gene Expression Omnibus under platform accession no GSE437, and this microarray experiment was implemented previously by Matsui et al [2003]. We consider thirteen group arrangements given by K-means clustering corresponding to the measured fluorescence intensity of cardiac allografts. At the finest level, we find that there is a cluster of 194 features that exhibit substantial fold increase in cardiac rejection allografts (Figure 6.18).
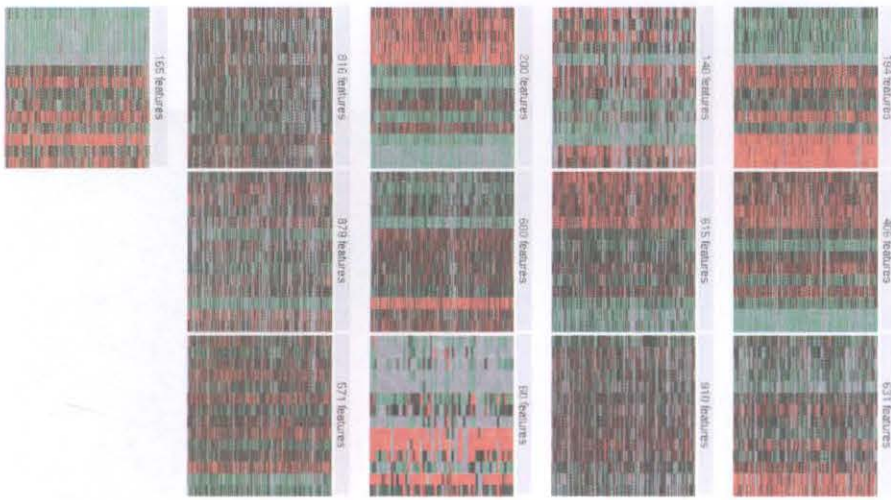
Figure 6.17 K-means clustered displayed of transcriptional response in cardiac rejecting allografts (K=13).

Figure 6.18 A cluster of 194 features that exhibit substantial fold increase in cardiac rejection allografts.

Table 6.6 Highly differentially expressed gene in cardiac rejecting allografts.

Gene Symbol	Gene Ontology
Anxa1	arachidonic acid secretion

Atp5c1	ion transport
P2rx4	transport (ion)
Adfp	transport (long-chain fatty acid)
Slc3a2	transport (amino acid)
Slc11a1	transport (iron ion)
Slc31a1	copper ion transport
Laptm5	transport
Msr1	transport (phosphate, lipoprotein)
Txn1	transport
Yipf5	transport
Sec61a1	protein transport
Sec61b	protein transport
Gltp	glycolipid transport
Npc2	intercellular cholesterol transport
Nasp	transport; histain exchange
Cyba	oxidation reduction; transpport (leukocyte transendothelial migration)
Cd24a	chemokine transport
Ncam1	cell adhesion
Fermt3	cell adhesion
Selplg	cell adhesion
Tgfbi	cell adhesion
Tnc	cell adhesion
Thbs1	cell adhesion; inflammatory response
Pscdbp	regulation of cell adhesion
Emp3	cell growth
Uck2	metabolism
Sdcbp	metabolism
Ltb4dh	metabolism
Gpd2	carbohydrate metabolism
Man2a1	carbohydrate metabolism
Pgam1	carbohydrate metabolism
Ugdh	carbohydrate metabolism
Pdk3	carbohydrate metabolism; glucose metabolism
Ugt1a6a	carbohydrate; lipid; xenobiotics
Car2	energy metabolism (nitrogen metabolism)
Angptl4	lipid metabolism
Apoc1	lipid metabolism

Grn	lipid catabolism
Pla2g4a	lipid metabolism
Pla2g7	lipid metabolism
Ptgs1	lipid metabolism
Sptlc2	lipid metabolism
Scd1	lipid metabolism
Ugcg	lipid metabolism
Arg1	amino acid metabolism
P4ha1	amino acid metabolism
Gclm	amino acid metabolism
Sat1	amino acid metabolism
Dck	nucleotide metabolism
Tyms	nucleotide metabolism
Rrm2	nucleotide metabolism; oxidation reduction
Pnp1	nucleotide metabolism; metabolism of cofactors and vitamins
Bst1	metabolisms of cofactors and vitamins
Hmox1	metabolisms of cofactors and vitamins
Ncf2	superoxide metabolism; NADP catabolism (leukocyte endothelial migration)
Mmp13	collageniscatabolism; proteoly
Anxa2	collagen fibrinolysis; angiogenesis; fibrinolysis
Lyz2	cell wall catabolism; cytolysis
Hist1h2ab	nucleosome assembly
Ctsz	proteolysis
Ctsc	proteolysis
Ctsh	proteolysis
Lgmn	proteolysis
Hck	amino acid phosphorylation
Lgals3	IgE binding
Srgn	mast cell secretory granule organization and biogenesis
Klrc1	antigen processing and representation
H13	antigen processing and presentation
Procr	antigen processing and presentation
Cd1d1	antigen processing and presentation (via MHC classs Ib)
Fcgr1	antigen processing and presentation (via MHC classs I)
Fcgr2b	antigen processing and presentation (via MHC classs II)

Tapbp	antigen processing and presentation (MHC classs I)
Hmha1	antigen processing and presentation (MHC classs I); signal transduction; intercellular signal cascade
Iqgap1	small GTPases mediated signal transduction
Birc5	apoptosis
Shisa5	apoptosis
Adam9	proteolysis; integrin-mediated signaling pathway
Mcm5	cell cycle; transcription
Atf3	transcription
Bach1	transcription
Cebpb	transcription
Nfil3	transcription
Fos	transcription
Litaf	transcription; apoptosis
Lrrfip1	transcription
Gtf2e2	transcription
Ostf1	transcription
Sfpi1	transcription
Tgif1	transcription
Basp1	transcription
Nfkbiz	transcription; inflammatory response
Ppp1r15b	translation; endoplasmic reticulum overload response
Glrx	transport; cell redox homeostasis
Rnu35a	RNA
Pabpc1	mRNA processing; mRNA splicing
H2afz	chromosome organization and biogenesis; nucleosome assembly
Ssr2	cotranslational protein targeting to membrane
Arl6ip1	cotranslational protein targeting to membrane
Tpd52	protein heterodimerization activity; protein homodimerization activity
Gp49a	unknown
Id2	transcription
Mkrn1	biological process
Ctla2b	biological process
Ctla2a	biological process
Pira3	receptor activity
Cstb	adult locomotory behavior

Apbb1ip	signal transduction
Rassf5	signal transduction (apoptosis; cell cycle) (leukocyte transendothelial migration)
Ptk2b	single complex assembly
Wsb1	intercellular signaling cascade
Cxcr4	axon guidance
Gnai3	axon guidance
Sema4d	axon guidance
Btg1	cell migration
Capg	cell projection biogenesis
Arhgdib	Gtpase activator activity
Rasd1	small GTPases mediated signal transduction
Actn1	cortical cytoskeleton organization and biogenesis
Csrp1	actin cytoskeleton organization and biogenesis
Tmsb10	actin cytoskeleton organization and biogenesis
Kif22	microtubule-based movement
Tubb6	microtubule-based movement
Tubb5	microtubule-based process
Wipf1	actin filament based movement
Tpm1	actin binding; structural constituent of cytoskeleton
Ccl9	immune response
Mobkl1b	metal ion binding
Rnf145	metal ion binding
Rnf149	metal ion binding
S100a4	calcium ion binding
S100a6	calcium ion binding
S100a8	calcium ion binding
S100a9	actin cytoskeleton organization; leukocyte chemotaxis;
S100a10	calcium ion binding
S100a11	calcium ion binding
Diap1	actin cytoskeleton organization and biogenesis
Rgs16	GPCR signaling pathway
Timp1	red blood cell maturation
Ptpn1	insulin receptor signaling pathway; amino acid dephosphorylation
Igf2bp2	regulation of cytokine biosynthesis
G6pdx	cytokine production
C1qc	complement activation; immune response

Hp	blood coagulation
Serpine1	blood coagulation
Serpina1b	blood coagulation
Igk-V20	immune response
Ccr5	immune response
Il4ra	immune response
Lilrb4	immune response
Olr1	immune response
Saa3	immune response
Clec4a2	immune response
Clec4d	immune response
Cd14	immune rand inflammatory response
Tnfrsf1b	inflammatory response
Lilrb3	B cell mediated immunity
Skap2	B cell activation
Il7r	T cell differentiation
Cd68	macrophage activation
Cdkn1a	cell cycle
Cd1d1	cell cycle
Prr13	unknown
Tmem49	unknown
D17H6S56E-5	unknown
Bxdc1	unknown
Slfn4	unknown
Slfn2	regulation of cell proliferation
Evi2a	cell proliferation
Mki67	cell proliferation

To characterise their biological processes, we perform gene ontology on all features within this cluster with the help of Entrez Gene [Maglott et al., 2007]. This enables us to identify 127 distinct genes that are up-regulated in rejecting allografts of cardiac transplants (Table 6.6). Thus, these 127 genes are highly differentially expressed in rejecting allografts of cardiac transplants. Specifically, changes in some chemokines (Cxcr4, Ccl9, Ccr5), cytokine-related molecules (Il7r, Tnfrsf1b, Igf2bp2 and G6pdx), leukocyte immunoglobulin-like receptors ((Lilrb3 and Lilrb4), and cell adhesion-related molecules (Ncam1, Fermt3, Pscdbp, Selplg, Tgfbi, Tnc and Thbs1) are observed;

these molecules play essential roles in leukocyte migration. Changes in complex metabolisms are also observed, including carbohydrate, energy, lipid, cofactors, vitamins, nucleotide and amino acid metabolisms. Several genes involved in transport, transcription and translation are also up-regulated in cardiac rejection allografts.

It can be observed that rejecting allografts initiate complement and coagulation cascades promoting muscle contraction, phagocyte recruitment, apoptosis, T cell and B cell mediated immune responses. Rejecting allografts also change the MHC class I and II molecules (Klrc1, H13, Procr, Cd1d1, Fcgr1, Fcgr2b, and Tapbp), in which the host T cell detects "foreign" antigens in rejecting allografts are derived from MHC class I and II molecules. Rejecting allografts also lead to gE production (Il4ra), IgE binding (Lgals3), the high-affinity Fc episilon receptors (Fcgr1 and Fcgr2b) and the mast cell secretory granule organisation and biogenesis (Srgn), suggesting that the binding of IgE to the Fc episilon RI promotes allergic reaction in this rejection. Several genes involved in Cxcr4 and Semaphorin 4D mediated axon guidance are up-regulated, indicating the potential roles of Cxcr4 and Semaphorin 4D mediated axon guidance promote immunological rejection of cardiac transplants.

In conclusion, the injurious mechanisms underlying allograft rejection of cardiac transplant can be summarised as follows: (1) complex metabolisms are involved in, including carbohydrate, energy, lipid, cofactors, vitamins, nucleotide and amino acid metabolisms, (2) the presence of leukocyte migration is found in this rejection, (3) rejection allografts initiate complement and coagulation cascades promoting muscle contraction, phagocyte recruitment, apoptosis, T cell and B cell mediated immune responses, (4) the host T cell which detected "foreign" antigen in rejection allografts is derived from MHC class I and II molecules, and (5) rejection allografts lead to the binding of IgE to the high-affinity Fc episilon RI promoting allergic reaction, and (5) Cxcr4 and Sema6D mediated axon guidance promote immunological rejection of cardiac transplants. As a result, the blockade of individual pathways generally does not prevent allograft rejection and long-term survival is achieved only after the simultaneous blockade of all of them.

6.2.7 Kidney Allograft Rejection

Kidney transplants, since the first successful transplant in 1954, have now extended and improved the quality of life for the majority of patients with end stage renal diseases [Andrew, 2002; Flechner et al., 2004]. While the administration of immunosuppressive therapy has enabled the survival of kidney transplant recipients, these drugs suppress the immune system of patients and entail many problems such as infection, spontaneous neoplasm, undesirable metabolic effects and drug toxicity [Matsui et al., 2003].

To know the injurious mechanisms underlying kidney allograft rejection, the Mahalanobis distance in K-means clustering is used to arrange microarray gene expression profiling of kidney biopsies and peripheral blood lymphocytes in transplant patients including normal donor kidneys, well-functioning transplant without rejection, kidneys undergoing acute rejection and transplants with renal dysfunction without rejection (Figure 6.19). The data used here was available from Gene Expression Omnibus under platform accession no GSE1563, and this microarray experiment was implemented previously by Flecher et al [2005]. We consider fifteen group arrangements given by K-means clustering corresponding to the measured fluorescence intensity of kidney biopsies and peripheral blood lymphocytes. At the finest level, we find that there is a cluster of 158 features that exhibit substantial fold increase in kidney biopsies undergoing acute rejection (Figure 6.20).
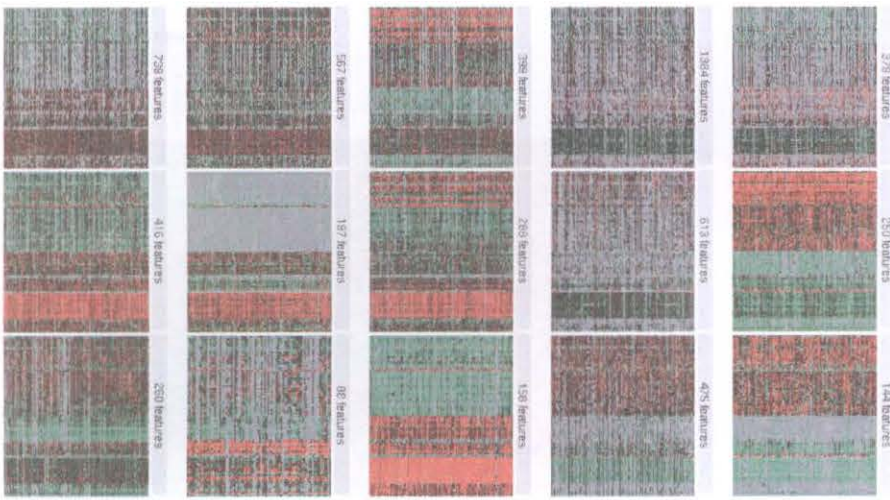
Figure 6.19 K-means clustering results of microarray gene expression profiling in kidney biopsies and peripheral blood lymphocytes from transplant patients including normal donor kidneys, well-functioning transplants without rejection, kidneys undergoing acute rejection, transplants with renal dysfunction without rejection.

Figure 6.20 A cluster of 158 features that exhibit substantial fold increase in kidney biopsies from transplant patients undergoing acute rejection.

Table 6.7 Highly differentially expressed genes in kidney biopsies from transplant patients undergoing acute rejection.

Gene symbol	Gene ontology
AKR7A2	carbohydrate metabolism
FBP1	carbohydrate metabolism
FUCA1	carbohydrate metabolism
IMPA2	carbohydrate metabolism
IDH1	carbohydrate metabolism
IDH2	carbohydrate metabolism
MDH1	carbohydrate metabolism
SDHC	carbohydrate metabolism
UGDH	carbohydrate metabolism
ALDH2	carbohydrate metabolism; amino acid metabolism
ALDH9A1	amino acid metabolism
GOT1	amino acid metabolism
QDPR	amino acid metabolism
GLUD2	amino acid metabolism
GLS	amino acid metabolism
GOT2	amino acid metabolism
OAT	amino acid metabolism
SEPHS2	amino acid metabolism
IARS	amino acid metabolism
CYB5A	metabolisms of cofactors and vitamins
AKR1A1	lipid metabolism
ACAA1	lipid metabolism
ACAA2	lipid metabolism
ACADSB	lipid metabolism; amino acid metabolism
ADH5	lipid metabolism
ALDH3A2	lipid metabolism
CLU	lipid metabolism
GPX4	lipid metabolism
HADH	lipid metabolism
HADHB	lipid metabolism
ANXA7	calcium dependent phospholipid binding
PEX3	peroxisome organization and biogenesis
ABCD3	peroxisome organization and biogenesis
ATP5H	nucleotide metabolism
ATP5J	nucleotide metabolism

ATP5O	nucleotide metabolism
ATP5A1	nucleotide metabolism
ATP5C1	nucleotide metabolism
ATP5J2	nucleotide metabolism
TXN	nucleotide metabolism
PON2	xenobiotics degradation
PAM	peptide metabolism
NDUFA2	mitochondrial electron transport, NADH to ubiquinone
NDUFA5	mitochondrial electron transport, NADH to ubiquinone
NDUFB5	mitochondrial electron transport, NADH to ubiquinone
NDUFS4	mitochondrial electron transport, NADH to ubiquinone
NDUFAF1	mitochondrial electron transport, NADH to ubiquinone
PRDX1	hydrogen peroxide catabolism
PRDX2	response to oxidative stress
PRDX3	hydrogen peroxide catabolism
PRDX4	oxidation reduction
HSP90AB1	regulation of nitric oxide biosynthesis process
SURF1	aerobic respiratory
BNIP3	oxygen and reactive oxygen species metabolism
NDUFS1	oxygen and reactive oxygen species metabolism
HSPD1	protein transport into mitochondrial matrix
SKP1	ubiquitin mediated proteolysis
UBA1	ubiquitin mediated proteolysis
UQCRQ	ubiquitin mediated proteolysis
CTSL1	proteolysis
CTSH	proteolysis
COX5A	electron transport chain
COX6A1	electron transport chain
COX6C	electron transport chain
COX7B	electron transport chain
ATP5G3	ATP synthesis coupled proton transport
ATP6V1A	ATP biosynthesis process
VTI1B	vehicle mediated transport
GM2A	transport (lipid)
SLC25A4	transport (ATP)
ATP6V0E1	transport (ion, protein)
PKD2	transport (calcium ion)
LAPTM4A	transport

LAPTM4B	transport
DDB1	DNA repair
MSH3	base-excision repair; mismatch repair
MRPS18B	translation
EARS2	translation
CANX	protein folding
CTAGE5	protein folding
FKBP2	protein folding
HSPE1	protein folding
ST13	protein folding
TNPO1	protein import into nucleus
SARS	tRNA processing
SNRPN	RNA splicing
ID1	transcription
ZNF117	transcription
ZBTB20	transcription
MORF4L2	transcription
HOXD4	transcription
ESD	biological process
REEP5	biological process
CFDP1	biological process
TRIM2	biological process
MT1B	biological process
MT1E	biological process
MT1F	biological process
MT1H	biological process
FAM171A1	unknown
MXRA7	unknown
TMEM59	unknown
C5orf13	open reading frame
C6orf108	open reading frame
PGRMC1	axon guidance
DSTN	actin filament severing
PFN2	actin cytoskeleton organization and biosynthesis
DYNLL1	actin cytoskeleton organization and biosynthesis
PRKCI	actin cytoskeleton organization and biosynthesis; signaling by NGF
RTN4	regulation of axon extension; signaling by NGF

RTN4	Regulation of axon guidance; signaling by NGF
DUSP3	signaling by NGF
PALLD	cytoskeleton organization and biosynthesis
CAPRIN1	cell projection
CTNNA1	establishment of cell polarity
GDI2	regulation of GTPases activity
KRIT1	small GTPases mediated signaling transduction
RIN2	small GTPases mediated signaling transduction
FGF9	signal transduction
ANXA4	signal transduction
PRKAR1A	signal transduction
FCGRT	IgE binding; antigen processing and presentation
SOD1	hemostasis
CD9	hemostasis
CD63	hemostasis
AQP3	regulation of immune system process
CD46	complement activation, classical pathway
SERPING1	complement activation, classical pathway
TNFSF10	immune response; apoptosis
AIFM1	apoptosis
ACTA1	muscle contraction
SERPINE2	serine proteinase inhibitor activity
APP	serine proteinase inhibitor activity
APLP2	serine proteinase inhibitor activity
ITIH5L	serine proteinase inhibitor activity
PEBP1	serine proteinase inhibitor activity

To characterise their biological processes, we perform gene ontology on all features within this cluster with the help of Entrez Gene [Maglott et al., 2007]. This enables us to identify 135 distinct genes that are up-regulated in kidney biopsies from transplant patients undergoing acute rejection (Table 6.7). Thus, those 135 genes are highly differentially expressed in kidneys from transplant patients undergoing acute rejection. In particular, changes in complex metabolisms are observed, including carbohydrate, lipid, cofactors, vitamins, xenobiotics, nucleotide and amino acid metabolisms. These metabolisms require the release of energy, which is made available from mRNA encoding small GTPases and ATP protein [Kosztin et al.,

2002]. Changes in oxygen and reactive oxygen species metabolisms are also observed. Several genes involved in potential interactions with electron transport chain and ubiquitin proteasome are up-regulated, suggesting rejecting allografts cause the deleterious effects of oxygen from the metabolic reduction of the highly active and toxic reactive oxygen species promoting endothelial damage and atherosclerosis. Several genes involved in transport, transcription and translation are up-regulated in rejection allografts as well.

It can be observed that rejecting allografts initiate classical pathways of complement cascades promoting muscle contraction, hemostasis, immune and inflammatory responses [Bhold and Stahl, 2003; Turnberg and Motto, 2003]. Rejecting allografts also promote the binding of IgE to the high-affintiy Fc episilon RI promoting allergic reactions in this rejection [Draft and Novak, 2006]. Furthermore, rejecting allografts lead to progesterone receptor membrane protein (PGRMC1), a putative steroid membrane receptor expressed in the liver and kidney, mediated axon guidance promoting immunological rejection of kidney transplants.

In conclusion, the injurious metabolisms underlying allograft rejection of kidney transplants can be summarised as follows: (1) complex carbohydrate, lipid, cofactors, vitamins, xenobiotics, nucleotide and amino acid metabolisms are involved in, (2) rejection allografts cause the deleterious effect of oxygen from the metabolic reduction of the reactive oxygen species promoting endothelial damage and atherosclerosis, (3) rejecting allografts initiate classical pathway of complement cascades promoting muscle contraction, hemostasis, immune and inflammatory responses, (4) rejecting allografts lead to the binding of IgE to the high-affinity Fc episilon RI promoting allergic reactions, and (5) rejecting allografts lead to progesterone receptor membrane protein (PGRMC1) mediated axon guidance promoting immunological rejection of kidney transplants.

6.2.8 Human Endometrial Receptivity throughout the Menstrual Cycle

The human endometrium, the anatomic prerequisite for establishing

and sustaining pregnancy, undergoes remarkable histological changes throughout the menstrual cycle, in preparation for embryonic implantation and subsequent shedding and regeneration in non-conception cycle [Talbi et al., 2006]. The differential histological appearance of this tissue was first described by Hitschmann and Adler [1908], and since the publication of the paper by Noyes et al. [1950], it has been understood that endometrial histology correlates with the changes in the circulating of estradiol and progesterone. Based on an ideal 28-day cycle, Noyes et al. [1950] described distinct histological phases, which are known as proliferative, early-secretory, mid-secretory and late-secretory phases of the menstrual cycle.

While histological criteria described by Noyes et al. [1950] has remained the gold standard for more than 50 years in clinical diagnosis and management of women with endometrial disorders, substantial evidence has claimed the accuracy of such criteria [Talbi et al., 2006]. Recently, the usefulness of histological dating of the endometrium for couples with infertility has been addressed since histological delay in endometrial maturation fails to discriminate between fertile and infertile couples [Giudice and Feerenczy, 1996; Talbi et al., 2006]. In another recent study, histological features fail to reliably distinguish between specific menstrual cycle days and narrow intervals of days, leading to the conclusion that histological dating has neither the accuracy nor the precision to the useful in clinical practice [Talbi et al., 2006].

To search for biochemical signatures for the endometrium that may prove to be more useful, the Mahalanobis distance in Ward's linkage hierarchical clustering algorithm is used to group microarray gene expression profiling of the human endometrium at early-secretory, mid-secretory and late-secretory phases during the menstrual cycle (Figure 6.21). The data used here was available at Gene Expression Omnibus under platform accession no GSE4888 and this microarray experiment was implemented previously by Talbi et al [2006]. At the finest level, we find that there is a cluster of 209 features that are up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretory phase, and down-regulated for those at early-secretory phase (Figure 6.22).
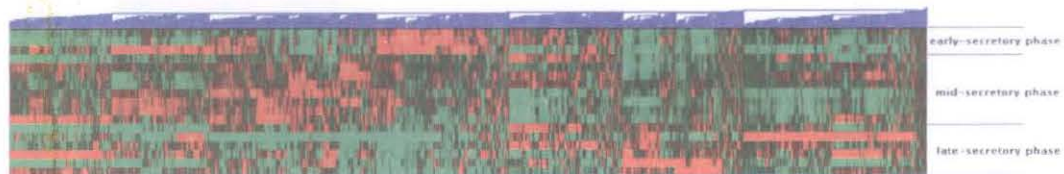
Figure 6.21 Clustering results of microarray gene expression profiles in the human endometriosis at early-secretory, mid-secretory, and late-secretory phases throughout the menstrual cycle.

Figure 6.22 A cluster of 209 features that are up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretory phase, and down-regulated for those at early-secretory phase.

TABLE 6.8 Diagnostic discriminating genes whose expressions are up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretory phase, and down-regulated for those at early-secretory phase.

Gene symbol	Gene ontology
ANG	ovarian follicle development
ANXA4	apoptosis
IER3	apoptosis
SGK1	apoptosis
AMIGO2	cell adhesion
MUC16	cell adhesion
CLDN4	cell adhesion
COMP	cell adhesion
HABP2	cell adhesion
LAMB3	cell adhesion
SPP1	cell adhesion
CATSPERB	cell differentiation
RARRES1	cell differentiation
GADD45A	cell cycle
LEFTY1	cell growth
EFNA1	axon guidance
MET	axon guidance
DNER	neuron migration
EDNRB	neural crest cell migration
S100P	endothelial cell migration
SYNE2	actin binding
KRT7	cytoskeleton reorganization and biogenesis
CRYAB	muscle contraction
GABARAPL1	autophagy
GABARAPL3	autophagy
DPP4	proteolysis
PTPRR	amino acid phosphorylation
TSPAN8	amino acid glycosylation
MUC20	protein homooligomer assembly
C3	immune response; blood coagulation
C4BPA	immune response; blood coagulation
SERPING1	immune response; blood coagulation
THBD	immune response; blood coagulation
CD55	immune response; complement activation

CFD	immune response; complement activation
GBP2	immune response
CXCL14	immune response
IL15	immune response
CRISP3	immune response
DEFB1	immune response
CLEC4E	immune response
BCL6	immune response
PAPLN	serine-type endopeptidase inhibitor activity
SLPI	serine-type endopeptidase inhibitor activity
GDF15	signal transduction
GPR110	signal transduction
RHPN2	signal transduction
DEPDC6	intercellular signaling cascade
MAP3K5	MAPKKK cascade
DKK1	down regulation of frizzled signaling pathway
ARID5B	transcription
ELL2	transcription
FOXO1	transcription
CITED2	transcription
CEBPD	transcription
KLF6	transcription
NFIL3	transcription
PAX8	transcription
FOSL2	transcription
NFKBIZ	transcription
IRX3	transcription
MYOCD	transcription
FAM3C	biological process
SCGB2A2	biological process
HRASLS3	biological process
GRAMD1C	unknown
MUM1L1	unknown
PHYHIPL	unknown
SESTD1	unknown
RAD21L1	unknown
C10orf10	open reading frame
C21org63	open reading frame

RNASE4	mRNA cleavage
FAM84B	hypothetical protein
FAM134B	hypothetical protein
FAM149A	hypothetical protein
TC2N	calcium binding
STC1	calcium ion homeostasis
NDRG1	response to metal ion
LIMS3	metal ion binding
MT2A	metal ion binding
MT1E	metal ion binding
MT1F	metal ion binding
MT1M	metal ion binding
MT1G	metal ion binding
MT1H	metal ion binding
MT1X	metal ion binding
GCNT3	carbohydrate metabolism
IRS2	glucose metabolism
ALDH1A3	lipid metabolism
APOD	lipid metabolism
SOD2	lipid metabolism
SRD5A3	lipid metabolism
CP	cofactors and vitamins metabolism
INDO	cofactors and vitamins metabolism
NNMT	cofactors and vitamins metabolism
VNN1	cofactors and vitamins metabolism
AOX1	amino acid metabolism
ARG2	amino acid metabolism
GPX3	amino acid metabolism
MGST1	amino acid metabolism
MAOA	amino acid metabolism
HAL	amino acid metabolism
GNLY	xenobiotics metabolism
RDH10	oxidation reduction
INDOL1	oxidation reduction
ABCC3	transport
PAEP	transport
MFSD4	transport
RBP4	transport

TRPM8	transport (ion)
SCARA5	transport (phosphate)
SLC15A1	transport (oligopeptide)
SLC15A4	transport (oligopeptide)
SLC16A3	transport (monocarboxylic acid)
SLC18A2	transport (monoamine)
SLC44A4	transport (choline)
SLC1A1	transport (L-glutamate, dicarboxylic acid)
SLC3A1	transport (amino acid)
SLC4A1	transport (anion)
SLC7A2	transport (L-amino acid)
SLC30A2	transport (cation, zinc ion)
SLC38A1	transport (neutral amino acid)
TMEM37	transport (calcium ion)
GABRE	transport (ion)
TCN1	transport (cobalt ion, cobalamin)

To identify their biological processes, gene ontology within this cluster is perform on all features with the help of Entrez Gene [Maglott et al., 2007]. This enables us to identify 127 distinct genes that are up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretory phase, and down-regulated for those at early-secretory phase. Thus, those 127 genes naturally classify tissues into three distinct phases, those at early-secretory, mid-secretory and late-secretory phases (Table 6.8). In particular, the nature of genes involved in ovarian follicle development (ANG), apoptosis (ANXA4, IER3 and SGK1), cell adhesion (AMIGO2, CLDN4, COMP, HABP2, LAMB3, MUC16 and SPP1), cell differentiation and growth (CATSPERB, RARRES1 and GADD45A) are altered throughout the menstrual cycle. Changes involved in complex metabolisms are observed, including carbohydrate, glucose, lipid, cofactors, vitamins, xenobiotics and amino acid metabolisms, at mid-secretory phase, suggesting the presence of extracellular remodeling at mid-secretory phases. These metabolisms require the release of energy, which was made available for mRNA encoding small GTPases and ATP protein [Kosztin et al., 2002]. Two genes involved in vehicle explanation and completion intended for autophagy (GABARAPL1, GABARAPL3) are up-regulated in the endometrium at mid-secretory phase, moderately

expressed for those at late-secretory phase and down-regulated for those at early-secretory phase [Klionsky and Emr, 2000; Bechet et al., 2005], suggesting that autophagic degradation is the mechanism responsible for new blood vessel formation in the human endometrium throughout the menstrual cycle.

It can be observed that several genes involved in potential interactions with complement and coagulation cascades are up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretory phase and down-regulated for those at early-secretory phase, which subsequently promoted muscle contraction, chemotaxis, phagocyte recruitment and peritoneal inflammation [Bhold and Stahl, 2003; Turnberg and Botto, 2003]. Several genes involved in potential interactions with Ephrin A mediated axon guidance are up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretory phase and down-regulated for those at late-secretory phase, indicating that Ephrin A mediated axon reflexes promoting retrograde menstruation at mid-secretory phase. In addition, DKK1, a gene involved in the regulation of the WNT receptor signaling pathway, is up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretory phase and down-regulated for those at late-secretory phase, suggesting that WNT receptor signaling pathway may play a role in the progress of endometrial disorders.

In conclusion, the biological mechanisms implicated in the human endometrium at mid-secretory phase can be summarised as follows: (1) complex metabolisms, including carbohydrate, glucose, lipid, cofactors, vitamins, xenobiotics and amino acid metabolisms, are involved in, (2) the initiation of complement and coagulation cascades promote muscle contraction, chemotaxis, phagocyte recruitment and peritoneal inflammation, (3) the occurrence of Ephrin A mediated axon guidance promote retrograde menstruation, (4) autophagy is the mechanism responsible for the new blood vessel formation, and (5) WNT receptor signaling pathway may contribute to the development of endometrial disorders.

6.2.9 Ovarian Endometriosis

Endometriosis, adenomysis and uterine fibroids represent the most widespread benign gynecological diseases in reproductive age and are collectively responsible for significant morbidity [Hever et al., 2007]. Endometriosis is associated with pain and other discomfort and plays a role in infertility [Berkeley et al., 2005]. Despite its high prevalence, the pathogenic mechanisms underlying endometriosis is relatively unknown. Diagnosis of endometriosis is also difficult, relying on symptoms, and laparoscopy examination being the only definitive confirmation [Hever et al., 2007]. Since the latter one is a surgical procedure, the diagnosis and subsequent treatment of endometriosis are usually delayed [Brosens et al., 2003]. Treatment options for endometriosis include surgery and/or suppression of ovarian steroids [Hever et al., 2007]. However, these strategies offer only temporary relief since the disease usually recurs [Hever et al., 2007].

To recognise the pathogenic mechanisms of ovarian endometriosis, the Mahalanobis disease in Ward's linkage hierarchical clustering is used to organise microarray gene expression profiling of ovarian endometriosis lesions corresponding to normal endometrial tissues obtained from the same patient at the same time (Figure 6.23). The data used here was available from Gene Expression Omnibus under platform accession no GSE7305, and this microarray experiment was performed by Hever et al [2007]. At the finest level, we find that there is a cluster of 293 features that reveal substantial fold increase in endometriosis lesions (Figure 6.24). Thus, genes within this cluster are highly differentially expressed in endometriosis lesions.

Figure 6.23 Clustering results of microarray gene expression profiles in ovarian endometriosis lesions corresponding to normal endometrial tissues obtained from

the same patient at the same time.
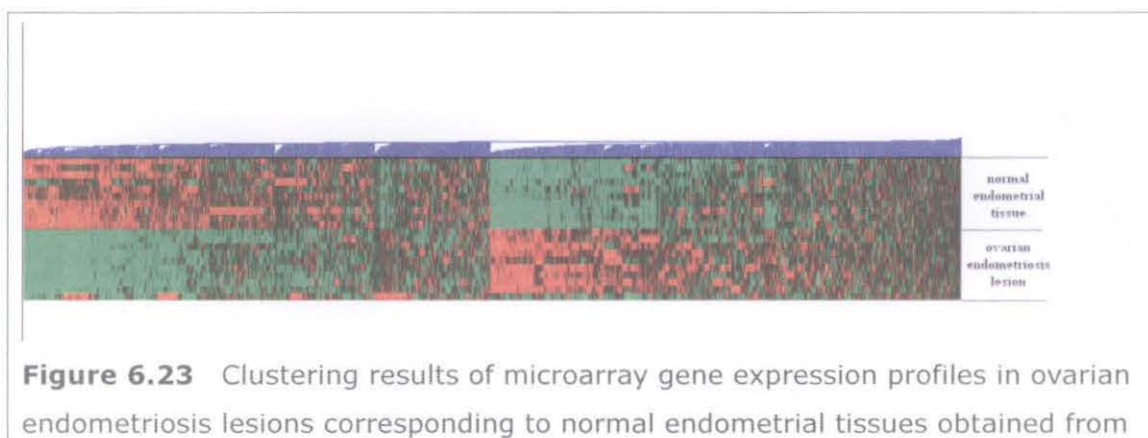
Figure 6.24 A cluster of 293 features that reveal substantial fold increase in ovarian endometriosis lesions corresponding to normal endometrial tissues obtained from the same patient at the same time.

Table 6.9 Highly differentially expressed genes associated with ovarian endometriosis lesions.

Gene symbol	Gene ontology
FOXC1	ovarian follicle development
OSAP	ovary-specific protein
GNG4	hormone mediated signaling
GPC3	embryogenesis and morphogenesis
LXN	perception of pain
FRZB	Wnt receptor signaling pathway
FZD7	Wnt receptor signaling pathway
WISP2	Wnt receptor signaling pathway
SFRP2	Wnt receptor signaling pathway
TGFBR3	TGF-beta receptor signaling pathway
FST	TGF-beta receptor signaling pathway
ABLIM1	axon guidance
ARX	axon guidance
CHL1	axon guidance
NRP2	axon guidance
NTN2L	axon guidance
SEMA6D	axon guidance
CLDN11	axon ensheathment
SYNPO	cell projection
NEXN	regulation of cell migration; regulation of cytoskeleton reorganization and biogenesis
PROK1	activation of MAPK
SCG2	MAPKKK cascade
NTRK2	signaling by NGF
ITPR1	signaling by NGF
LMOD1	tropomyosin binding
CALD1	cell motility
MYH11	muscle contraction
DMD	muscle development
TAGLN	muscle development
PRELP	skeletal development
MEG3	apoptosis
PEG10	apoptosis
PEG3	apoptosis

PRKAR2B	apoptosis
ITPR1	apoptosis
CDON	cell adhesion
ITGBL1	cell adhesion
NCAM1	cell adhesion
NEGR1	cell adhesion
VCAM1	cell adhesion
CDH3	cell adhesion
CPXM2	cell adhesion
NCAM1	cell adhesion
NFASC	cell adhesion
PPFIBP1	cell adhesion
SIGLEC11	cell adhesion
THBS2	cell adhesion
FLRT3	cell adhesion
ECM2	cell-matrix adhesion
SGCE	cell-matrix adhesion
ANGPT1	cell differentiation
TMEM176B	cell differentiation
PDGFD	cell proliferation
RARRES1	cell proliferation
GPNMB	cell proliferation
TIMP1	cell proliferation
GAS1	cell cycle
RGS2	cell cycle
NDE1	cell cycle
AEBP1	transcription
BNC2	transcription
ZFPM2	transcription
NR4A3	transcription
HOXC6	transcription
LHX9	transcription
MYOCD	transcription
MKX	transcription
FHL2	transcription
KLF2	transcription
PBX3	transcription
TCF21	transcription

GATA4	transcription
GATA6	transcription
PDLIM3	transcription
NRK	transcription
ZNF521	transcription
RNASE4	mRNA cleavage
RBMS3	RNA binding; nucleotide binding
FAM129A	biological process
MATN2	biological process
SH3D19	biological process
M85256	hypothetical protein
C10orf116	open reading frame
C10orf10	open reading frame
ITM2A	unknown
GPM6A	unknown
FAM70A	unknown
KLHDC8A	unknown
CCDC3	unknown
CCDC80	unknown
ABI3BP	unknown
MEG3	unknown
CLIP4	unknown
FIBIN	unknown
PLCXD3	unknown
ST7L	unknown
ANG	multicellular organismal development
DLK1	multicellular organismal development
PLSCR4	blood coagulation
PROS1	blood coagulation
SERPINE2	blood coagulation; complement activation (classical pathway)
SERPING1	blood coagulation; complement activation (classical pathway)
C1R	complement activation; immune response
C3	complement activation; immune response
C4A	complement activation; immune response
C4B	complement activation; immune response
C7	complement activation; immune response
CFH	complement activation; immune response
CFHR1	complement activation

CLU	complement activation; classical pathway
LY96	immune response
IGL@	immune response
IGLV1-44	immune response
IGLV@	immune response
IGHG3	immune response
IGKC	immune response
IGLJ3	immune response
IGHG1	immune response
IGKC	immune response
IGHM	immune response; activation of MAPK; B-cell receptor signaling pathway
CCL2	immune and inflammatory response
FAM19A2	immune and inflammatory response
CD163	inflammatory response
TNFSF13B	B cell costimulation T cell costimulation
BST2	immune response; B-cell activation
SERPINA3	inflammatory response
PGM5	carbohydrate metabolism; glucose metabolism
PDK4	carbohydrate metabolism; glucose metabolism
CHI3L1	carbohydrate metabolism
FABP4	lipid metabolism; cholesterol homeostasis
PLA2G2A	lipid metabolism
PLA2G2F	lipid metabolism
STAR	lipid metabolism
HSD11B1	lipid metabolism
PTGIS	lipid metabolism
ENPP6	lipid metabolism
ADH1B	lipid metabolism; amino acid metabolism; ethanol oxidation; oxidative reduction
GPX3	amino acid metabolism
AOC3	amino acid metabolism
AOX1	amino acid metabolism; metabolisms of cofactors and vitamins
GATM	amino acid metabolism
INMT	amino acid metabolism
ST6GALNAC5	amino acid glycosylation
NNMT	metabooisms of cofactors and vitamins
DPYSL3	nucleotide metabolism

DHRS2	xenobiotics metabolism
CSGALNACT1	glycan biosynthesis and metabolism
HS3ST1	glycan biosynthesis and metabolism
DSE	glycan biosynsis
HS6ST2	transferace activity
PTPRZ1	one-carbon compound metabolism
FMO2	reactive oxidative species metabolism; oxidative reduction
FMO1	oxidation reduction
MSRB3	oxidation reduction
TSPAN8	amino acid glycosylation
LRRC2	protein binding
SDPR	protein binding
CPVL	proteolysis
PRSS35	proteolysis
MRC1	receptor-mediated endocytosis
LTBP2	protein secretion; protein targeting
DIRAS3	genetic imprinting; regulation of cyclin-dependent protein kinase activity
CYBRD1	transport; electron transport chain
ABCA8	transport
ABCA6	transport (ABC)
COLEC11	transport (phosphate)
COL10A1	transport (phosphate)
SCN7A	transport (sodium ion)
SLC16A4	transport (monocarboxylic acid)
ABCA9	transport
ACTG2	ATP binding; protein binding
PDLIM3	metal ion binding; zinc binding; protein binding
EFEMP1	calcium ion binding; protein binding
PLN	cellular calcium ion homeostasis
RCAN2	calcium-mediated signaling transduction
RERGL	small GTPases mediated signaling transduction
SYTL2	Rab GTPase binding
AKAP12	signal transduction
ANGPTL1	signal transduction
RERG	signal transduction
FGF7	signal transduction
PNOC	signal transduction

TGFBR3	signal transduction
MS4A4A	signal transduction
MS4A7	signal transduction
CHN2	intercellular signaling cascade

To characterise their biological processes, gene ontology within this cluster is perform on all features using Entrez Gene [Maglott et al., 2007]. This enables us to identify 185 distinct genes that are highly differentially expressed in ovarian endometriosis lesions (Table 6.9). In particular, the nature of genes involved in ovarian follicle development (FOXC1), apoptosis (MEG3, PEG3, PEG10, PRKAR2B and ITPR1), cell adhesion (CDON, ITGBL1, NCAM1, NEGR1, VCAM1, CDH3, CPXM2, NCAM1, NFASC, PPFIBP1, SIGLEC11, THBS2 and FLRT3), cell-matrix adhesion (ECM2 and ECM2), cell differentiation (ANGPT1 and TMEM176B), cell proliferation (GPNMB , PDGFD, RARRES1 and TIMP1) and cell cycle (GAS1 ,NDE1 and RGS2) have been altered through these processes. The expressions of IgH (IGHG3, IGHG1 and IGHM), IgK (IGKC and IGKC) and IgL (IGL@, IGLV1-44, IGLV@ and IGLJ3) molecules are also up-regulated in ovarian endometriosis lesions. Changes in complex metabolisms are also observed, including carbohydrate, glucose, lipid, cofactors, vitamins, xenobiotics and amino acid metabolisms. These metabolisms require the release of energy, which is made available from mRNA encoding small GTPases and ATP motor proteins [Kosztin et al., 2002].

It can be observed that several genes involved in potential interactions with complement and coagulation cascades are activated in ovarian endometriosis lesions promoting muscle contraction, phagocyte recruitment, T cell and B cell mediated immune responses [Holers, 2008; Klaska and Nowak, 2007]. Several genes involved in WNT receptor signaling pathway (FRZB, FZD7, WISP2 and SFRP2) and TGF-beta receptor signaling pathway (FST and TGFBR3) are up-regulated in ovarian endometriosis lesions, suggesting the potential roles of WNT and TGF-beta receptor signaling pathways involved in the development of endometriosis. In addition, several genes involved in potential interactions with Netrin2 like (NTN2L) and Semaphorin 6D (SEMA6D) mediated axon guidance are up-regulated [Dickson, 2002], suggesting that Netrin2 like and Semaphorin 6D

mediated axon guidance promoting retrograde menstruation in endometriosis.

In summary, the pathogenic mechanisms of ovarian endometriosis can be summarised as follows: (1) complex metabolic reactions, including carbohydrate, glucose, lipid, cofactors, vitamins, xenobiotics, nucleotide and amino acid metabolisms are involved, (2) complement and coagulation cascades are activated in ovarian endometriosis lesions promoting muscle contraction, phagocyte recruitment, T cell and B cell mediated immune responses, (3) WNT and TGF-bet receptor signaling pathways are participated in the development of ovarian endometriosis, and (4) Netrin-2 like and Semaphorin 6D mediated axon guidance promoting retrograde menstruation in ovarian endometriosis.

6.2.10 Inflammatory Acne

Acne is the most widespread skin condition affecting millions of people worldwide [Trivedi et al., 2006]. The pathogenesis of acne vulgaris is complex and incompletely understood [Trivedi et al., 2006]. Inflammation is a key component of the pathogenesis of acne [Trivedi et al., 2006]. An immunological reaction to the Gram-positive microbe *Propionibacterium acnes* may play a major role in the initiation of the inflammatory reaction [De Young et al., 1984; Jappe, et al, 2002]. Furthermore, viable *P.acnes* and not heat-killed organisms can stimulate the release of cytokines [Schroder, 2004]. While the initiating events causing acne still remain a mystery, there exists a debate as to whether hyperkeratinisation of the follicular duct precedes the influx of inflammatory cells or *vice versa* [Trivedi et al., 2006].

To know the pathogenic mechanisms underlying inflammatory acnes, the Mahalanobis distance in the K-means clustering algorithm is used to group microarray gene expression profiling of inflammatory papules corresponding to normal skin biopsies. The data used here was available from Gene Expression Omnibus under platform accession no GSE6475, and this microarray experiment was implemented by Trivedi et al [2006]. We consider nine group

arrangements given by K-means clustering corresponding to the measured fluorescence intensity of inflammatory acnes versus normal skin biopsies (Figure 6.25). At the finest level, we find that there is a cluster of 60 features that reveal considerable fold increase in inflammatory papules (Figure 6.26).
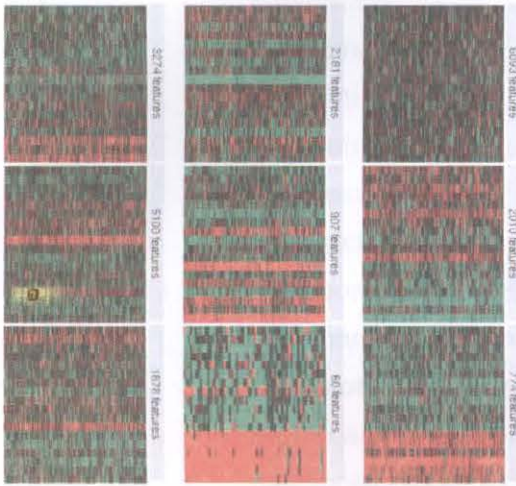

Figure 6.25 K-means clustered display of microarray gene expression profiles in inflammatory papules corresponding to normal skin biopsies (K=9).
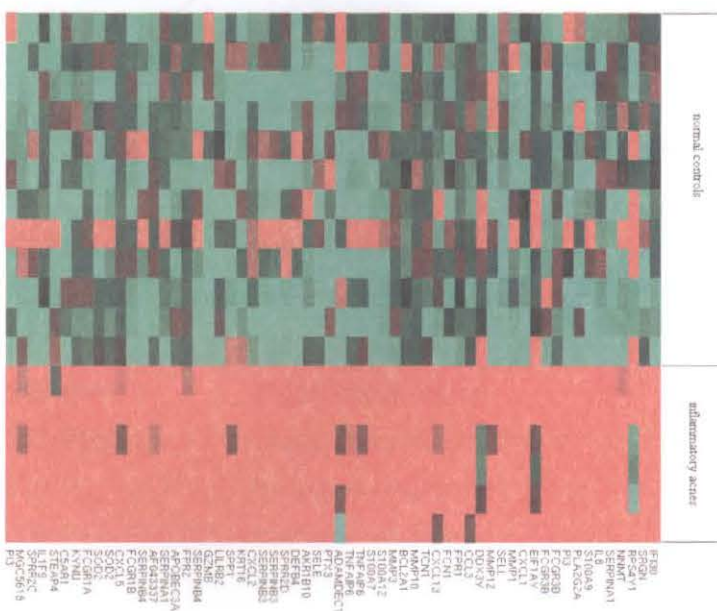

Figure 6.26 A cluster of 60 features that reveal considerable fold increase in inflammatory papules from acne patients corresponding to normal skin biopsies.

TABLE 6.10 Highly differentially expressed genes in inflammatory acne.

Gene symbol	Gene Ontology
IFI30	antigen process and representation (MHC class II)
SRGN	T-cell secretory granule organization and biogenesis
FCGR3B	IgE binding;
PLA2G2A	Fc episilon RI signaling transduction pathway
PI3	copulation
DDX3Y	DNA binding
EIF1AY	translation
RPS4Y1	translation
BCL2A1	apoptosis
GZMB	apoptosis
FCN1	transport (phosphate)
TCN1	transport (colbalamin ; cobalt ion)
APOBEC3A	xenobiotics biodegradation
AKR1B10	amino acid metabolism
NNMT	metabolisms of cofactors and vitamins
MMP3	metabolism
MMP1	metabolism
MMP10	metabolism
MMP12	metabolism
SPRR2D	epidermis development
KRT16	epidermis development
ADAMDEC1	integrin-mediated mediated pathway
SELL	cell adhesion
SPP1	cell adhesion
SELE	cell adhesion
SERPINA1	blood coagulation
SERPINB3	blood coagulation
SERPINB4	blood coagulation
IL8	immune response
CXCL1	immune response
CCL3	immune response
CXCL13	immune response
CXCL2	immune response
LILRB2	immune response
S100A7	immune response
S100A9	inflammatory response

S100A12	inflammatory response
FPR2	inflammatory response
PTX3	inflammatory response
TNFAIP6	inflammatory response

To characterise their biological processes, we perform gene ontology on all features within this cluster using Entrez Gene [Maglott et al., 2007]. This enables us to identify 41 distinct genes that are highly differentially expressed in inflammatory papules (Table 6.10). In particular, we observe changes in five chemokines (IL8, CXCL1, CCL3, CXCL13 and CXCL2), a leukocyte immunoglobulin-like receptor (LILRB2), three cell adhesion molecules (SELL, SPP1 and SELE) and a gene involved in integrin-mediated signaling pathway (ADAMDEC1) through these processes; these molecules play essential roles in leukocyte migration. We also observe two genes involved in epidermis development (SPRR2D, KRT16) are up-regulated. Changes in metabolisms are also observed, including cofactors, vitamins, xenobiotics and amino acid metabolisms, in inflammatory papules.

It can be observed that inflammatory acne activates intrinsic pathway of blood coagulations promoting hemostasis or the stoppage of bleeding [Kaufmann and Schaible, 2005]. A MHC class II molecule (IFI30) and a gene involved in T cell ecretory granule organisation and biogenesis (SRGN) are up-regulated, suggesting that inflammatory acne activates MHC class II expression triggering T cell mediated immune response in response to bacteria infection. Furthermore, there is an indication of the binding of IgE to the high-affinity Fc episilon RI promoting allergic inflammation in inflammatory papules.

In conclusion, the pathogenic mechanisms underlying inflammatory acne can be summarised as follows: (1) the presence of leukocyte migration, (2) the activation of intrinsic pathway of blood coagulations promoting hemostasis or the stoppage of bleeding, (3) the activation of MHC class II expression triggering T cell mediated immune response in response to bacteria infection, and (4) the binding of IgE to the high-affinity Fc episilon RI promoting allergic inflammation.

6.3 Conclusion

A natural way of viewing a complex data set is first to scan and survey the large-scale features and then to address the details of interest [Eisen et al., 1998]. In this chapter, we first find variant genes associated with the disease under investigation where the correlation between genes is taken into account. Next, we identify gene-specific information using Entrez Gene or map variant genes onto KEGG pathways to obtain a linkage between key molecules and biochemical pathways associated with the disease under investigation in a cause-effect format. The approaches are generalised ones, with no inherent specificity to the particular method used to acquire microarray gene expression data. Therefore, these approaches may be used to develop genome-based diagnostic tests and therapeutic targets in many other types of diseases.

Finally, we have seen the possibility that medicine of the future could include a scan of a patient's entire genome, looking for all variant genes that predispose toward disease, and this work has made possible a useful data analysis tool in the outlook of modern medicine.

Chapter 7

Conclusion

7.1 Summary of Research Contributions

In this thesis, a choice of bioinformatics techniques used in the analysis of disease-related microarray gene expression data is developed. They examined variant genes associated with the study disease and subsequently derived insight into the molecular aetiology of disease using: (1) the linear logistic regression model, (2) the conditional or fixed effect logistic regression model, and (3) the Mahalanobis distance in the K-means or hierarchical clustering algorithms. For an unmatched study, the variant genes can be assessed using the linear logistic regression model. For a matched study, the variant genes can be assessed using the conditional logistic regression model. Alternatively, one can employ the Mahalanobis distance in K-means or hierarchical clustering algorithms for data mining and pattern recognition in gene expression analysis. In these analytical procedures, strong emphasis is placed on the interpretation of disease-related microarray gene expression data in terms of biological processes that contribute to the disease under investigation. This requires the consideration of biological interdependencies and potential confounding, as well as the identification of gene-specific information using Entrez Gene [Maglott et al., 2007]. In other words, fundamental to this thesis is that the mass of numbers produced by gene chips are primarily driven by their genetic roots in the human genome.

A central step in implementing K-means or hierarchical clustering algorithms is to select a distance measure, which will calculate the similarity between genes on the basis of their expression profiles. Given that all clustering procedures are based on some form of

disease measure, the method of calculating distance matrix can have an important effect on the clustering results [Mimmack et al., 2001]. This will influence the shape of the clusters, as some genes may be close to one another according to one distance measure, and farther away according to another.

We can demonstrate the important influence on the choice of distance measure by the example with which we dealt in Section 6.2.10. In this example, the applicability of the Mahalanobis distance and of the frequently used Pearson correlation coefficient are examined in the context of clustering gene expression profiles in inflammatory papules corresponding to normal skin biopsies using K-means clustering algorithm.
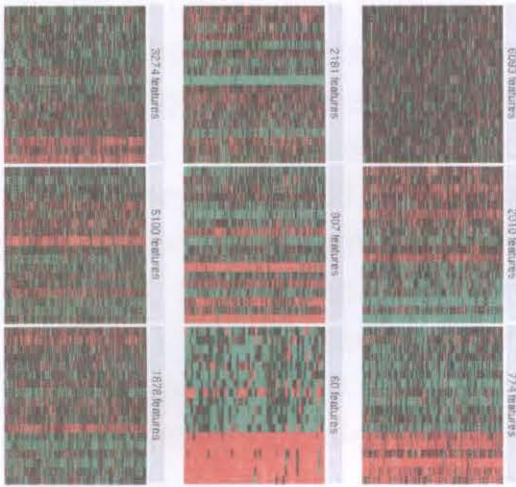

Figure 7.1 The Pearson correlation coefficient in K-means clustered display of microarray gene expression profiles in inflammatory papules corresponding to normal skin biopsies (K=9).
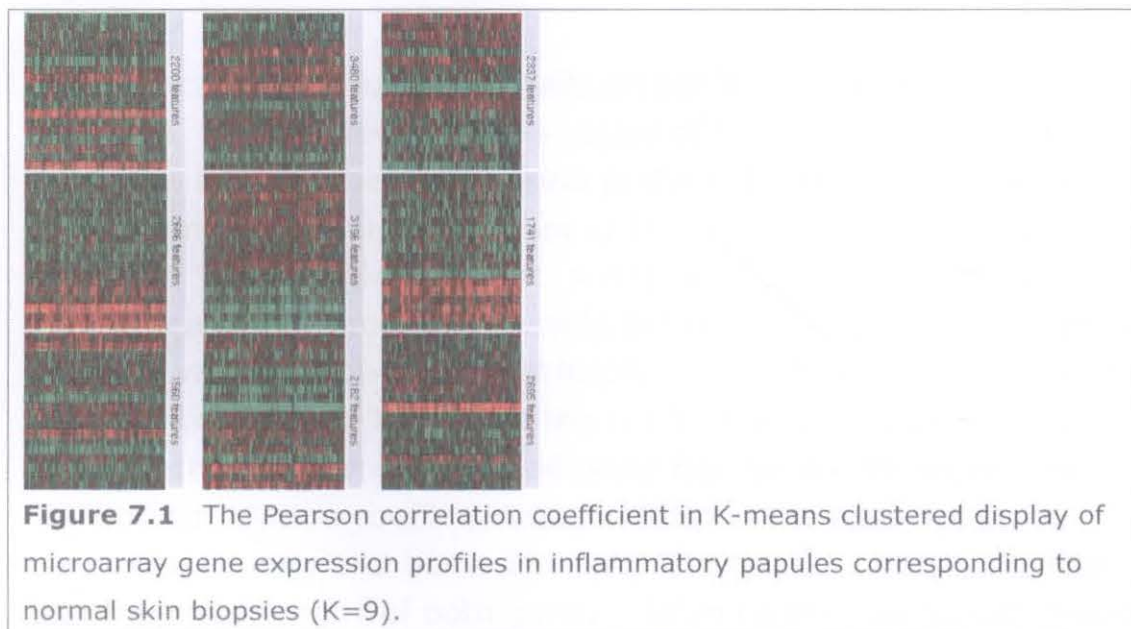
Figure 7.2 The Mahalanobis distance in K-means clustered display of microarray gene expression profiles in inflammatory papules corresponding to normal skin biopsies (K=9).

Geometrically, the Pearson correlation coefficient between two genes X and Y can be viewed as the cosine of the angle between X and Y on the basis of their expression profiles. Therefore, the Pearson correlation coefficient varies from -1 to +1 wheres the cosin varis from 0 to 1 in a single quadrant. A positive value for the coefficient implies a positive linear relationship between two genes, whereas a negative value for the coefficient implies a negative linear relationship between two genes. The closer the coefficient is to -1 or +1, the stronger the linear correlation between two genes. Moreover, the Pearson correlation coefficient depends upon the assumption of normality, lienarly and homoscedasity. These assumptions mandate that the distributions of both genes related by the coefficient should be normal and that the scatter-plots should be linear and homoscedatic. In genomics, on the other hand, the distributions of two genes may not satisfy the assumption of linearly, normality and homoscedasity. Therefore, the clustering result with the Pearson correltion coefficient as a distance measure in K-means clustering algorithm, as shown in Figure 7.1, are disturbed as a result of omitting the assumption of linearly, normality and homoscedasity.

A better distance measure that incorporates the differences in variance as well as the covariance between genes is the Mahalanobis distnce [Tseng et al., 2009]. The Mahalanobis distance gives less

weight to genes with high variance and to high correlated genes so that all characteristics are treated as equally important. Apart from accounting for covariances between genes and for differences in vairances, the Mahalanobis distance has other attractive properties, such as being related to the log-likeliood of multivariate normal distributions, and to multidimensional scaling [Greenacre and Underhill, 1982; Stephenson and Doblas-Reyes, 2000; Mimmack et al., 2001]. The clustering result with the Mahalanobis distance as a distance measure in K-means clsutering algorithm, as shown in Figure 7.2, efficiently revealed a cluster of 60 features that exhibit substantial fold increase in inflammatory papules corresponding to normal controls. The variant gene clusters can be revealed for a further explanation of molecular biological mechanisms [as presented in Section 6.2.10].

7.2 Perspectives

In discussing the role of gene chips in clinical practice, the bioinformatics techniques in the analysis of disease-related microarray gene expression data presented here have contributed to the exploitation of genome-based expression profiling as a single standardised microarray platform for minimally invasive diagnostic tests and therapeutic interventions in the study disease.

Presently, most microarray-based diagnostic tests are still in the development stage since many challenges remain in adopting DNA microarrays as a commonplace platform for diagnostic testing [Aitman 2001; Pollack, 2007]. For instance, the existing procedure used to classify prognostic subtypes of childhood ALL using DNA microarrays is the selection of a number of discriminating genes for the various ALL subtypes using Chi-square statistic, and there are marked differences in the number of diagnostic discriminating genes for the various ALL subtypes [Ross et al., 2003; Ross et al., 2004]. This simple technique, however, does not support the use of genome-based expression profiling as a single standardised microarray platform for diagnosis of the known prognostic subgroups of pediatric ALL in clinical practice, since it can be extremely laborious, time-consuming and expensive. Furthermore, such procedures do not

address the full potential of genome-based experiments to alter our understanding of cellular biology by offering, through an inclusive analysis of the entire repertoire of transcripts, a continuing comprehensive window into the biology underlying the clinical difference among these leukaemia subtypes.

In order to support the use of genome-based expression profiling as a single standardised microarray platform for minimally invasive diagnostic tests and therapeutic targets in childhood ALL, the Mahalanobis distance in the K-means clustering algorithm has been has been used to analyse and make sense of microarray gene expression profiling of bone marrow samples from pediatric patients with various prognostic subtypes of childhood ALL (as presented in Section 5.1). The results show that the expressions of seven MHC class II molecules (HLA-DQB1, HLA-DPA1, CD74, HLA-DRB1, HLA-G, HLA-B and HLA-DRA) are a unique gene set in classifying childhood acute lymphoblastic leukaemia, and the application of DNA dementhylating agent to induce MHC class II expressions will be beneficial in the prevention of pediatric acute lymphoblastic leukaemia relapse [Holling et al., 2004]. As a result, a commonplace diagnostic array platform consisting of these seven MHC class II molecules can be developed for classifying childhood ALL in clinical practice. According to the digitized intensity output patterns from DNA microarrays, clinicians can provide promise of a more accurate diagnostic tests and tailored treatments to each pediatric patient with ALL in clinical practice in a straightforward manner.

7.3 Conclusion Remarks

The main theme of this thesis is to provide better fundamental bioinformatics techniques in the analysis of disease-related microarray gene expression data. This work resolved the challenges in adopting DNA microarrays as a commonplace platform for diagnostic testing and therapeutic targets in clinical practice. In the next three to five years, the development of diagnostic microarray platform is likely to offer promise of more accurate diagnostic tests and tailored treatments to each patient in clinical practice.

Furthermore, the identification of biological mechanisms, new indicators of disease prognosis, and new therapeutic targets are likely to be understood in almost any kind of diseases, and will start to enter clinical practice with new treatment strategies within this time frame. The success of these computational approaches has given us confidence to "see" the information in disease-related gene expression data based on biological necessities.

Bibliography

Abbas AK, Lichtman AH & Pober JS (1997), *Cellular and molecular immunology*, Philadelphia: W.B. Saunders.

Ackroyd S, O'Connor SJ & Owen RG (2005), Rarity of IgH translocations in Waldenstrom macroglobulinemia, *Cancer Genetics and Cytogenetics* **163**, pp.77-80.

Armstrong NJ & van de Wiel MA (2004), Microarray data analysis: from hypotheses to conclusions using gene expression data, *Cellular Oncology* **26**, pp. 279-290.

Aitman TJ (2001), Science, medicine and the future: DNA microarrays in medical practice, *British Medical Journal* **323**, pp. 611-615.

Alexander Fleming (1944), The discovery of Penicillin, *British Medical Bulletin* **2**, pp. 4-5.

Alperin M & Moalli PA (2006), Remodeling of vaginal connective tissue in patients with prolapse, *Current Opinion in Obstetrics & Gynecology* **18**, pp.544-550.

Alwine JC, Kemp DJ & Stark GR (1977), Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymetyl-paper and hybridization with DNA probes, *Proceedings of the National Academy of Sciences of the United States of America* **74**, pp. 5350-5354.

Aoki-Kinoshita & Kanehisa (2007), Gene annotation and pathway mapping in KEGG, *Methods in Molecular Biology* **396** , pp. 71-91.

Baldwin KM (2000), Research in the exercise science: Where do we go from there?, *Journal of Applied Physiology* **88**, pp. 332-336.

Barnes (2005), *Diseases and human evolution*, University of New Mexico Press.

Bechet D, Tassa A & Taillandier D et al. (2005), Lysosomal proteolysis in skeletal muscle, *The International Journal of Biochemistry & Cell Biology* **37**, pp. 2098-2114.

Beck-Nielsen H & Hother-Niesen O (1996), *Diabetes Mellitus: A Fundamental and Clinical Text*, Philadelphia: Lippincott-Raven, pp. 475–484.

Bell DA (2005), Origins and molecular pathology of ovarian cancer, *Modern Pathology* **18 (Suppl2)**, pp. S19-S32.

Belland RY, Ouellette SP, Gieffers J & Gyrne GI (2004), Chlamydia pneumoniae and atherosclerosis, *Cellular Microbiology* **6**, pp.117-127.

Bhole D & Stahl GL (2003), Therapeutic potential of targeting the complement cascade in critical care medicine, *Critical Care Medicine* **31 (Suppl 1)**, pp. S97-S104.

Bonnet D & Dick JE (1997), Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell, *Nature Medicine* **3**, pp. 730-737.

Brown PO & Botstein D (1999), Exploring the new world of the genome with DNA microarrays, *Nature Genetics* **21(Suppl)**, pp. 33-37.

Bouchard C, Leon AS & Rao DC et al (1995), The HERITAGE family study: aims, design, and measurement protocol, *Medicine and Science in Sports and Exercise* **27**, pp. 721-729.

Burney RO, Talbi S & Hamilton AD (2007), Gene expression analysis of endometrium reveals progesterone resistance and candidate susceptibility genes in women with endometriosis, *Endocrinology* **148**, pp. 3814-3827.

Campbell DJ (2003), The kallikrein-kinin system in humans, *Clinical and Experimental Pharmacology & Physiology* **28**, pp.1060-1065.

Chan TY, Partin AW, Walsh PC & Epstein JI (2000), Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy, *Urology* **56**, pp. 823-827.

Chap T Le (1997), *Applied survival analysis*, Wiley.

Chen B, Wen Y & Zhang Z et al. (2006), Microarray analysis of differentially expressed genes in vaginal tissues from women with stress urinary incontinence compared with asymptomatic women, *Human Reproduction* **21**, pp. 22-29.

Chen I, Homma H, Jin C & Yan H (2007), Cluster analysis of elite swimmers' race patterns, *International Journal of Sport Science and Coaching* **2**, pp. 203-303

Chen I, Homma H, Jin C & Yan H (2007), Clustering and display of elite swimmers' race patterns across various comparable criteria at the same time, *International Journal of Sport Science and Engineering* **1**, pp. 129-136.

Chen I, Tseng L, Homma H et al (2007), Monitoring athletes' physiological responses to endurance training with genome-wide expression data, *International Journal of Sport Science and Engineering* **1**, pp. 147-156.

Chen I, Chen M, Jin C & Yan H (2007), Large-scale cluster analysis of elite male and female swimmers' race patterns, *International Journal of Sport Science and Engineering* **2**, pp.123-128.

Chng WJ, Schop RF & Price-Troska T et al (2006), Gene expression profiling of Waldenstrom macroglobulinemia reveals a phenotype more similar to chronic lymphocytic leukemia than multiple myeloma, *Blood* **108**, pp. 2755-2763.

Clementsen P, Permin H & Norn S (2002), Chlamydia pneumoniae infection and its role in asthma and chronic obstructive pulmonary disease, *Journal of Investigational Allergology & Clinical Immunology* **12,** pp.73-79.

Colwin C (2002), B*reakthrough Swimming: Physiological Research and Applications (Chapter 9)*, Human Kinetics.

Collet D (1991), *Modeling Binary Data*, CRC Press.

De Pittà C, Tombolan L & Campo DM et al., (2005), A leukemia-enriched cDNA microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia, *Haematologica* **90**, pp. 890-898.

De Young LM , Young JM & Ballaron SJ et al. (1984), Intradermal injection of Propionibacterium acnes: a model of inflammation relevant to acne, *The Journal of Investigative Dermatology* **83**, pp. 394-398.

Dent P, Yacoub A & Fisher PB et al. (2003), Pathways in radiation responses, *Oncogene* **22**, pp.5885-5896

Derynck R & Zhang YE (2003), Smad-dependent and Smad-independent pathways in TGF-beta family signaling, *Nature* **425**, pp. 577-584.

Drewes PG, Yanagisawa H & Starcher B et al. (2007), Pelvic organ prolapse in fibulin-5 knockout mice: pregnancy-induced changes in elastic fiber homeostasis in mouse vagina, *The American Journal of Pathology* **170**, pp. 578-589.

Dyrskjøt L (2003), Classification of bladder cancer by microarray expression profiling: towards a general clinical use of microarrays in cancer diagnostics, *Expert Review of Molecular Diagnostics* **3**, pp. 635-647

Edgerton DS, Lautz M & Scott M et al. (2006), Insulin's direct effects on the liver dominate the control of hepatic glucose production, *The Journal of Clinical Investigation* **116**, pp. 302-304.

Eisen MB, Spellmen PT, Brown PO & Botstein D (1998), Cluster Analysis and Display of Genome-Wide Expression Pattern, *Proceedings of the National Academy of Sciences of the United States of America* **95**, pp. 14863-14868.

Epstein JI (2000), Gleason score 2-5 adenocarcinoma of the prostate on needle biopsy: a diagnosis that should not be made, *The American Journal of Surgical Pathology* **24**, pp. 1443-1444.

Eyre H, Kahn R & Robertson RM (2004), Preventing cancer, cardiovascular disease and diabetes: a common agenda for the American Cancer Society, the American Diabetes Association, and the American Heart Association, *Diabetes Care* **26**, pp.1812-1824.

Fahy JV, Corry DB & Houshey HA (2000), Airway inflammation and remodeling in asthma, *Current Opinion in Pulmonary Medicine* **6**, pp. 15-20.

Fearon ER & Vogelstein B (1990), A genetic model for colorectal tumorigenesis, *Cell* **61**, pp. 759-767.

Fielden MR & Zacharewski TR (2001), Challenge and limitations of gene expression profiling in mechanistic and predictive toxicology, *Toxicological Sciences* **60**, pp. 6-10.

Fluck M, Dapp C & Schmutz S (2005) , Transcriptional profiling of tissue plasticity: role of shifts in genes expression and technical limitations, *Journal of Applied Physiology* **99**, pp.397-413.

Gailani D & Renne T (2007), Intrinsic pathway of coagulation and arterial thrombosis, *Arteriosclerosis, Thrombosis, and Vascular Biology* **27**, pp. 2507-2513.

Garrett L (1994), *The coming plague*, New York: Farrar, Straus & Giroux.

Greenacre MJ & Underhill LG (1982), Scaling a data matrix in a low-dimensional Euclidean space. *Topics in Applied Multivariate analysis*, Cambridge University Press, pp. 183-266;

Gilfillan AM & Tkaczyk C (2006), Integrated signaling pathways for mast-cell activation, *Nature Reviews Immunology* **6**, pp.218-630.

Gleason DF & Mellinger GT (1974), Prediction of prognosis for prostatic adenocarcinomas by combined histological grading and clinical staging, *Journal of Urology* **111**, pp. 58-64.

Golub TR, Slonim DK & Tamayo P et al. (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **15**, pp.531-537.

Grayston JT (2005), Chlamydia pneumoniae and atherosclerosis, *Clinical Infectious Diseases* **40**, pp.1131-1132.

Gu H, Saito K & Klaman LD et al. (2001), Essential role for Gab2 in the allergic response, *Nature* **412**, pp. 186-190.

Gunther S, Alexander RW & Atkinson WJ et al. (1982), Functional angiotensin II receptors in cultured vascular smooth muscle cells, *The Journal of Cell Biology* **92**, pp. 289-298.

Giudice LC & Ferenczy A (1996), *The endometrial cycle: morphologic and biochemical events. In: Adashi EY, Rosenwaks Z, eds. Reproductive endocrinology, surgery, and technology*, New York: Raven Press pp. 271-300.

Gutiérrez NC, Ocio EM & de Las Rivas J (2007), Gene expression profiling of B lymphocytes and plasma cells from Waldenström's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic

leukemia, multiple myeloma and normal individuals, *Leukemia* **21**, pp. 541-549.

Hanai T, Hamada H & Okamoto M (2006), Application of bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields, *Journal of Bioscience and Bioengineering* **101**, pp. 377-384.

Harper SJ & LoGrasso P (2001), Signaling for survival and death in neurons: the role of stress-activated kinases, JNK and p38, *Cell Signals* **13**, pp.299-310.

Heck H, Mader A & Hess G et al. (1985), Justification of the 4-mmol/1 lactate threshold, *International Journal of Sports Medicine* **6**, pp.117-130.

Henry T & Fonseca R (2007), Genomics and proteomics in multiple myeloma and Waldenström macroglobulinemia, *Current Opinion in Hematology* **14**, pp. 369-374.

Hitschmann F & Adler L (1908), er Bau der Uterus-schleimhaut des geschlectsreifen Weibes mit besonderer Berucksightigung der Menstruation, Monatschr, *Geburtsch. Gynakol* **27**, pp. 1–23.

Hoffman A, Abassi ZA & Brodsky S et al. (2000), Mechanisms of big endothelin-1-induced diuresis and natriuresis: role of ET(B) receptors. *Hypertension* **35**, pp.732-739.

Holling TM, SchootenE & Langerak AW et al (2004), Regulation of MHC class II expression in human T-cell malignancies, *Blood* **103,** pp.1438-1444.

Horcajadas JA, Pellicer A & Simón C (2007), Wide genome analysis of human endometrial receptivity: new times, new opportunities, *Human Reproduction Update* **13**, pp. 77-86.

Hollman W, Rost R & Meirleir K De et al. (1985), Cardiovascular effects of extreme physical training, *Acta Medica Scandinavica Scandivavica: Supplementum* **11**, pp. 193-203.

Hotamisligil GS (2007), Endoplasmic reticulum stress and inflammation in obesity and type 2 diabetes, *Novartis Foundation Symposium* **286**, pp. 86-94.

Hyrcza MD, Kovacs C & Loutfy M (2007), Distinct transcriptional profiles in ex vivo CD4+ and CD8+ T cells are established early in human immunodeficiency virus type

1 infection and are characterized by a chronic interferon response as well as extensive transcriptional changes in CD8+ T cells, *Journal of Virology* **81**, pp. 3477-3486.

James J Schlesselman (1982), *Case-control studies: design, conduct, analysis*, Oxford University Press.

Jaffe HW, Bregman DJ & Selik RM (1983), Acquired immune deficiency syndrome in the United States: the first 1,000 cases, *The Journal of Infection* **148**, pp. 339-345.

Jappe U,Ingham E, Henwood J & Holland KT (2002), Propionibacterium acnes and inflammation in acne; P. acnes has T-cell mitogenic activity, *The British Journal of Dermatology* **146**, pp. 202–209.

Johnston SL (2002), Influence of viral and bacterial respiratory infections on exacerbations and symptom severity in childhood asthma, *Pediatric Pulmonology Supplement* **16**, pp.88-89.

Klaska I and Nowak JZ (2007), The role of complement in physiology and pathology, *Postępy Higieny i Medycyny Doświadczalnej* **61**, pp. 167-177.

Kaneto H, Matsuoka TA & Nakatani Y et al. (2005), Oxidative stress and the JNK pathway in diabetes, *Current Diabetes Reviews* **1**, pp.65-72.

Kanehisa M, Goto S & Hattori M et al. (2006), From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Research* **34**, D354.-D357,

Kanehisa M, Goto S, Kawashima S & Nakaya A (2002), The KEGG database at GenomeNet, *Nucleic Acids Research* **30**, pp. 42-46.

Kanehisa M, Goto S & Kawashima S (2004) , The KEGG resource for deciphering the genome, *Nucleic Acids Research* **32**, pp. D288-D280.

Kanehisa M, Goto S & Hattori M et al. (2006), From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Research* **34**, pp. D354-D357.

Kanehisa M, Araki M, Goto S et al. (2008) , KEGG for linking genome to life and the environment, *Nucleic Acids Research* **35**, pp. D480-484.

Kaufmann SH & Schaible UE (2005), Antigen presentation and recognition in bacterial infections, *Current Opinion in Immunology* **17**, pp. 79-87.

Kawakami T & Galli SJ (2002), Regulation of mast-cell and basophil function and survival by IgE, *Nature Reviews Immunology* **2**, pp. 773-786.

Keay S, Seillier-Moiseiwitsch F & Zhang CO et al. (2003), Changes in human bladder epithelial cell gene expression associated with interstitial IC or anttiproliferative factor treatment, *Physiological Genomics* **14**, pp. 1186-1189.

Kampuraj D, Saito H & Kaneko A et al. (1999), Characterization of mast cell-committed progenitors present in human umbilical cord blood, *Blood* **93**, p. 3338-3346.

Khatri P & Drăghici S (2005), Ontological analysis of gene expression data: current tools, limitations and open problems, *Bioinformatics* **21**, pp. 3587-3595.

Kim SJ, Gershov D & Ma X et al. (2003), Opsonization of apoptotic cells and its effect on macrophage and T cell immune responses, *Annals of the New York Academy of Sciences* **987** , pp.68-78.

Kirschfink M (1997), Controlling the complement system in inflammation, *Immunopharmacology* **38**, pp. 51-62.

Klionsky DJ & Emr SD (2000), Autophagy as a regulated pathway of cellular degradation, *Science* **290**, pp. 1717-1721.

Kosztin, R. Bruinsma R, O'Lague P & Schulten K (2002), Mechanical force generation by G proteins, *Proceedings of the National Academy of Sciences of the United States of America* **99**, pp. 3575-3580.

Kriangkum J, Taylor BJ & Strachan E et al (2006), Impaired class switch recombination (CSR) in Waldenstrom macroglobulinemia (WM) despite apparently normal CSR machinery, *Blood* **107**, pp. 2920-2927.

Le Moine A, Goldman M & Abramowicz D (2002), Multiple pathways to allograft rejection, *Transplantation* **73**, pp.1373-1381.

Liu R, Paxton WA & Choe S et al. (1996), Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection, *Cell*

86, pp.367-377.

Liu R, Wang X & Chen GY et al. (2007), The prognostic role of a gene signature from tumorigenic breast-cancer cells, *The New England Journal of Medicine* **18**, pp. 217-226.

Longo DL, Steis RG & Lane HC et al. (1984), Malignancies in the AIDS patient: natural history, treatment strategies and preliminary results, *Annals of the New York Academy of Sciences* **437**, pp. 421-430.

Mader A & Heck H (1986), A theory of the metabolic origin of anaerobic threshold, *International Journal of Sports Medicine Supplement* **7 (Suppl)**, pp. S45-S65.

Maglischo EW (2002), *Swimming Fastest*, Human Kinetics Press.

Maglott D, Ostell J, Pruitt KD & Tatusova T (2007), Entrez Gene: gene-centered information at NCBI, *Nucleic Acids Research* **35**, pp.D26-D31.

Mansell PW (1983), Acquired immune deficiency syndrome, leading to opportunistic infections, *Kaposi's Aarcoma and Other Malignancies* **20**, pp. 191-204

Marlink R (2001), *The biology and epidemiology of HIV-2. In: AIDS Clinical Review 2000/2001*, New York: Marcel Dekker, pp.47-65

Maass M & Uarig U (1995), Evaluation of culture conditions used for isolation of Chlamydia pneumoniae, *American Journal of Clinical Pathology* **103**, pp.141-148.

Masters RG, Davies RA & Veinot JP et al. (1999), Discoordinate modulation of natriuretic peptides during acute cardiac allograft rejection in humans, *Circulation* **100**, pp. 287–291.

Matsui Y, Saiura A & Sugawara Y et al. (2003), Identification of gene expression profile in tolerizing murine cardiac allograft by costimulatory blockade, *Physiological Genomics* **14**, pp. 199-208

Matsushita M, Thie S & Jensenius JC et al. (2000), Proteolytic activities of two types of mannose-binding lectin-associated serine protease, *Journal of Immunology* **165**, pp. 2637-2642.

Mimmack GM, Mason SJ & Galpin JS (2001), Choice of distance matrices in cluster analysis of defining regions, *Journal of Climate* **14**, pp. 2790-2797.

Abraham BK, Fritz P & McClellan M et al. (2005), Prevalence of CD44^{+}/CD24$^{-/low}$ cells in breast cancer may not be associated with clinical outcome but may favor distinct metastasis, *Clinical Cancer Research* **11**, pp.1154-1159.

Nadler ST, Stoehr JP & Schueler KL (2000), The expression of adipogenic genes is decreased in obesity and diabetes mellitus, *Proceedings of the National Academy of Sciences of the United States of America* **97**, pp. 11371-11376.

Nadler, M. J. et al. 2000. Signal transduction by the high-affinity immunoglobulin E receptor Fc epsilon RI: coupling form to function, *Advances in Immunology* **76**, pp.325-355.

Nadler, M. J. and Kinet, J. P. 2002. Uncovering new complexities in mast cell signaling, *Nature Immunology* **3**, pp.707-708.

Nagy I, Pivarcsi A & Koreck A et al. (2005), Distinct strains of Propionibacterium acnes induce selective human beta-defensin-2 and interleukin-8 expression in human keratinocytes through toll-like receptors, *The Journal of Investigative Dermatology* **124**, pp. 931-938.

Nanni S, Priolo C & Grasselli et al. (2006), Epithelial-restricted gene profile of primary cultures from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer, *Molecular Cancer Research* **4**, pp.79-92.

Norris LA (2003), Blood coagulation, *Best Practice & Research: Clinical Obstetrics & Gynaecology* **17**, pp.369-383

Novembre FJ, Saucier M & Anderson DC et al. (1997), Development of AIDS in a chimpanzee infected with human immunodeficiency virus type 1, *Journal of Virology* **71**, pp. 4086-4091.

Noyes RW, Hertig AT & Orck J (1950), Dating the endometrial biopsy, *Fertility and Sterility* **1**, pp. 3-17.

Ogata H, Goto S & Sato K et al.(1999), KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Research* **27**, pp. 29-34.

Olbrecht K, Madsen O & Mader A et al. (1985), Relationship between swimming velocity and lactic concentration during continuous and intermittent training exercises, *International Journal of Sports Medicine* Supplement **7**, pp. 74-77.

Owen RG, Treon SP & Al-Katib A (2003), Clinicopathological definition of Waldenström's macroglobulinemia: consensus panel recommendations from the Second International Workshop on Waldenström's macroglobulinemia, *Seminars in Oncology* **30**, pp. 110-115.

Patel S, Parmar D, Dupta YK & Singh MP (2005), Contribution of genomics, proteomics, and single-nucleotide polymorphism in toxicology research and Indian scenario, *Indian Journal of Human Genetics* **11**, pp. 61-75.

Pearsons CL & Pearsons JK (1996), *Interstitial cystitis. In: Raz S, ed. Female urology 2d ed.*, Philadelphia: Saunders

Pollheimer J & Knöfler M (2005), Signaling pathways regulating the invasive differentiation of human trophoblasts: a review, *Placenta* **26(Suppl A)**, pp.S21-S30.

Parsons CL (2007), The role of the urinary epithelium in the pathogeneses of interstitial cystitis/prostates/urethritis, *Urology* **69(Suppl 4A)**, pp. 9-16.

Partin AW, Kattan MW & Subong EN et al. (1997), Combination of prostate-specific antigen, clinical stage, and Gleason score to predice pathological stage of localized prostate cancer: a multi-institutional update, *JAMA* **277**, pp. 1445-1451.

Pass HI, Potter DA & Macher AM et al. (1983), Thoracic manifestations of the acquired immune deficiency syndrome, *The Journal of Thoracic and Vardiovascular Surgery* **88 (5 pt 1)**, pp. 654-658.

Patel AA & Steitz JA (2003), Splicing double: insights from the second spliceosome, *Nature Reviews Molecular Cell Biology* **4**, pp. 960-970.

Pisani P, Parkin DM, Bray F & Ferlay J (1999), Estimates of the worldwide mortality from 25 cancers in 1990. *International Journal of Cancer* **83,** pp. 18–29.

Pollack JR (2007), A perspective on DNA microarrays in pathology research and

practice, *The American Journal of Pathology* **171**, pp. 375-385.

Pui CH & Evans WE (1998), Acute lymphoblastic leukemia, *New England Journal of Medicine* **339**, pp.605-615.

Pui CH, Campana D & Evans WE (200), Childhood acute lymphoblastic leukaemia—current status and future perspectives, *Lancet Oncology* **2**, pp. 597-607.

Quagliarello V (1982), The acquired immunodeficiency syndrome: current status, *The Yale Journal of Biology and Medicine* **55**, pp. 443-452.

Qhu HJ & Burgess AQ (2001), Regulation of transforming growth factor-betasignaling, *Molecular Cell Biology Research Communications* **4**, pp. 321-330.

Ramaswamy S, Osteen RT & Shulman LN (2001), *Clinical Oncology*, American Cancer Society pp. 711–719.

Parkin DM, Bray F, Ferlay J & Pisani P (2005), Global cancer statistics, 2002, *CA: a Cancer Journal for Clinicians* **55**, pp. 74-108.

Reya T, Morrison SJ & Clarke MF (2001), Stem cells, cancer and cancer stem cells, *Nature* **414**, pp. 105-111.

Rodríguez N, Mages J & Dietrich H et al. (2007), MyD88-dependent changes in the pulmonary transcriptome after infection with Chlamydia pneumoniae, *Physiological Genomics* **18**, pp.134-145.

Rover E & Esdridge M (2005), Interstitial cystitis. eMedicine from Web MD.

Ross ME, Zhou X & Song G et al. (2003), Classification of pediatric acute lymphoblastic leukemia by gene expression profiling, *Blood* **102**, pp. 2951-2959.

Ross ME, Mahfouz R & Onciu M et al. (2004), Gene expression profiling of pediatric acute myelogenous leukemia, *Blood* **104**, pp. 3679-3698.

Roukos DH (2000), Current status and future perspectives in gastric cancer management, *Cancer Treatment Reviews* **26**, pp. 243–552

Ruf W, Dorfleutner A & Riewald M (2003), Specificity of coagulation factor signaling, *Journal of Thrombosis and Haemostasis* **1**, pp.1495-1503.

Sayama K, Diehn M & Matsuda K et al. (2002), Transcriptional response of human mast cells stimulated via the Fc(episilon)RI and identification of mast cells as a course of IL-11, *BMC Immunology* **3**:5.

Schlesselman JJ (1982). *Case-control studies: Design, Conduct, Analysis*, Oxford University Press

Schmitz-Dräger CO, Zeegers MP & Steineck G et al. (2007), Epidemiology of urinary bladder cancer: from tumor development to patient's death, *World Journal of Urology* **25**, pp. 285-295.

Schroder JM (2004), Epithelial antimicrobial peptides: local innate defense effector molecules, *Annales de Dermatologie et de Vénéréologie* **131**, pp. 411-416.

Schop RF, Kuehl WM & Van Wier SA et al. (2002), Waldenström macroglobulinemia neoplastic cells lack immunoglobulin heavy chain locus translocations but have frequent 6q deletions, *Blood* **100**, pp. 2996-3001.

Simoneau JA, Lortie G & Boulay MR et al. (1985), Skeletal muscle histochemical and biochemical characteristics in sedentary male and female subjects, *Canadian Journal of Physiology and Pharmacology* **63**, pp.30–35.

Siraganian RP (2003), Mast cell signal transduction from the high-affinity IgE receptor, *Current Opinion in Immunology* **15,** pp.639-646.

Smith PG, Pike MC & Hill AP et al. (1981), Algorithm AS 162: multivariate conditional logistic analysis of stratum matched case-control studies, *Applied Statistics* **30**, pp. 190-197.

Stekel D (2003), *Microarray bioinformatics*, Cambridge University Press.

Stephenson DB & Doblas-Reyes FJ (2000). Statistial methods for interprting Monte Carlo ensemble forecasts. *Tellus* **52A**, pp. 300–322.

Wade N, Life sceipt (2001), *the genome and the new medicine*, Free Press.

Stephenson FH (2006), *DNA: how the biotech revolution is changing the way we fight disease*, Prometheus Books.

Stern E, Steenblock ER, Reed MA & Fabmy TM (2008), Label-free electronic detection of the antigen-specific T-cell immune response, *Nano Letters* **8**, pp. 3310-3314.

Strauss JH & Strauss EG (2002), *Viruses and human disease*, Academic Press.

Takada I & Koto X (2005), PPARs target genes, *Nippon Rinsho* **63**, pp. 573-577.

Takahashi N, Goto T & Kusudo T et al. (2005), The structures and functions of peroxisome proliferator-activated receptors (PPARs), *Nippon Rinsho* **63**, pp. 557-564.

Talbi S, Hamilton E & Vo KC et al. (2006), Molecular phenotyping of human endometrium distinguishes menstrual cycle phases and underlying biological processes in normo-ovulatory women, *Endocrinology* **147**, pp. 1097-1121.

Tamura G (2002) Genetic and epigenetic alterations of tumor suppressor and tumor-related genes in gastric cancer, *Histology and Histopathology* **17**, 323–329

Tamura G, Yin J & Wang S et al. (2000), E-Cadherin gene promoter hypermethylation in primary human gastric carcinomas, *Journal of the National Cancer Institute* **92**, pp. 569–573.

Terashima M, Maesawa C & Oyama K et al. (2005), Gene expression profiles in human gastric cancer: expression of maspin correlates with lymph node metastasis, *British Journal of Cancer* **92**, pp. 1130-1136.

Tomatis L (1990), *Cancer: Causes, Occurrence and Control. WHO International Agency for Research on Cancer*, Scientific Publications No. 100 LYON: WHO.

Tomaszewski J E & LiVolsi V A (1999), Cancer 86, pp. 2198–2200.

Trivedi NR, Gilliland KL, Zhao W & Lin W (2006), Gene array expression profiling in acne lesions reveals marked upregulation of genes involved in inflammation and matrix remodeling, *The Journal of Investigative Dermatology* **126**, pp. 1071-1079.

True L, Coleman I & Hawley S (2006), A molecular correlate to the Gleason grading system for prostate adenocarcinomas, *Proceedings of the National Academy of Sciences of the United States of America* **103**, pp. 10991-10996.

Turner H & Kinet JP (1999), Signalling through the high-affinity IgE receptor Fc epsilonRI, *Nature* **402**, pp. B42-B30.

Turnberg D & Botto M (2003), The regulation of the complement system: insights from genetically-engineered mice, *Molecular Immunology* **40**, pp. 145-153.

Viemann D, Schmidt M & Tenbrock K et al. (2007), The contact allergen nickel triggers a unique inflammatory and proangiogenic gene expression pattern via activation of NF-keppaB and hypoxia-inducible factor-1 alpha, *Journal of Immunology* **178**, pp. 3198-3207.

Wade N (2001), *Life sceipt: the genome and the new medicine,* London: Free press.

Wang Y, Hewitt SM & Liu S et al (2006), Tissue microarray analysis of human FRAT1 expression and its correlation with the subcellular localization of beta-catenn in ovarian tumours, *British Journal of Cancer* **13**, pp. 686-691.

Waris W & Ahsah H (2006), Reactive oxygen species: role in the development of cancer and various chronic conditions, *Carcinogenesis* **5**, pp. 1186-1192.

Wen Y, Polan ML & Chen B (2006), Do extracellular matrix protein expressions change with cyclic reproductive hormones in pelvic connective tissue from women with stress urinary incontinence?, *Human Reproduction* **21**, pp. 1266-1273..

Wen Y, Zhao YY & Li S (2006) , Differences in mRNA and protein expression of small proteolycans in vaginal wall tissue from women with and without stress urinary incontinence, *Human Reproduction* **22**, pp. 1718-1724.

Werner T (2008), Bioinformatics applications for pathway analysis of microarray data, *Current Opinion in Biotechnology* **19**, pp. 50-54

Wheeler DL, Benson DA & Bryant S et al. (2005), Database resources of the National Center for Biotechnology Information: Update, *Nucleic Acid Research* **33**, pp. D-39-D45.

Williams HH, Ward CW, Spangenburg EE & Nelson RM (1998), Functional aspects of skeletal muscle contractile apparatus and sarcoplasmic reticulum after fatigue, *Journal of Applied Physiology* **85**, pp. 619-626.

Wollschlager CM, Khan FA, Chitkara RK & Shivaram U (1984), Pulmonary smanifestations of the acquired immunodeficiency syndrome (AIDS), *Chest* **85**, pp. 197-202.

Wu X & Dewey TG (2006), From microarray to biological networks: analysis of gene expression profiles, *Methods in Molecular Biology* **316**, pp. 35-48.

Yamaguchi O (2004), Response of bladder smooth muscle cells to obstruction: signal transduction and the role of mechanosensors, *Urology* **63 (Suppl1)**, pp.11-16.

Yang ZP, Cuo CC & Grayston JP (1993), A mouse model of Chlamydia pneumoniae strain TWAR pneumonitis, *Infection and Immunity* **61**, pp. 2037-2040.

Yasui W, Oue N & Kuniyasu H et al. (2001), Molecular diagnosis of gastric cancer: present and future, *Gastric Cancer* **4**, pp. 113-121.

Zhang M, Murphy RF & Agrawal DK (2007), Decoding IgE Fc receptors, *Immunologic Research* **37**, pp. 1-16.

Zhou YT, Shimabukuro M & Wang MY et al. (1998), Role of peroxisome proliferator-activated receptor alpha in disease of pancreatic beta cells, *Proceedings of the National Academy of Sciences of the United States of America* **95**, pp. 8898-8903.

Appendix A

Genome-based expression profiles as a single standardised microarray platform for the diagnosis of experimental interstitial cystitis: an array of 75 genes model

ORIGINAL ARTICLE

Genome-based expression profiles as a single standardized microarray platform for the diagnosis of experimental interstitial cystitis: an array of 75 genes model

Ling-Hong Tseng · Ilene Chen · Ming-Yang Chen ·
Chyi-Long Lee · Tsia-Shu Lo · L. Keith Lloyd

Abstract

Introduction and hypothesis To investigate the molecular signature underlying experimental interstitial cystitis (IC) using cDNA microarray.

Methods Microarray gene expression profiles are studied in bladder epithelium of C57BL/6 mice with ovalbumin or substance P-induced experimental IC versus *Escherichia coli* lipopolysaccharide-induced bacterial cystitis.

Results Main findings are summarized as follows: firstly, a "75-gene" model was discovered to contain high expressions of bladder epithelium which feature in experimental IC. Secondly, glucose, lipid, nucleotide, xenobiotics, and amino acid metabolisms are involved in. Thirdly, T-cell-mediated immune and inflammatory responses are observed. Fourthly, Wnt, Tgf-beta, Mapk, and insulin growth factor receptor signaling pathways are also involved in. In addition, experimental IC leads to Ephrin- and Semaphorin-mediated

L.-H. Tseng (✉) · C.-L. Lee · T.-S. Lo
Department of Obstetrics and Gynecology,
Chang Gung Memorial Hospital and University of Chang Gung
School of Medicine,
Tao-Yuan, Taiwan
e-mail: 3g7330@yahoo.com.tw

I. Chen
School of Electrical and Information Engineering,
University of Sydney,
Sydney, Australia

M.-Y. Chen
Department of Electrical Engineering, Stanford University,
Stanford, CA, USA

L. K. Lloyd
Division of Urology, University of Alabama School of Medicine,
Birmingham, AL, USA

axon guidance promoting parasympathetic inflammatory reflexes.

Conclusions Further characterization of human IC-induced gene expression profiles would enable the use of genome-based expression profiling for the therapeutic targets and diagnosis of IC.

Keywords Bladder pain syndrome/interstitial cystitis (BPS/IC) · C57BL/6 mice · Gene expression · cDNA microarray

Introduction

Research into bladder pain syndrome/interstitial cystitis (BPS/IC) has a very long history. It was first documented by Mercier in 1937, and characterized by Skene in 1887, as the presence of an inflammation that destroys the urinary bladder mucous membrane, partly or wholly, and extends to muscular parities [1]. In 1914, Hunner popularized IC with the description of a characteristic bladder wall ulcer [1]. In 1947, Hand made the first epidemiological description of IC as the widespread, small, submucosa bladder hemorrhage and substantial shrinkage in the bladder capacity [1]. In 1987, Messing and Stamey recognized glomerulation as an indicator of IC [1]. Up to the present time, the etiology of BPS/IC is still relatively unknown and has been acknowledged by the American Urological Association, as a most challenging disease.

The urothelium plays a pivotal role as a barrier between urine and its solutes and the underlying bladder. Bladder surface mucus is a critical component of this function. Currently, the most accepted hypothesis is glycosaminoglycans (GAG) deficiency [2]. The GAG layer is extremely hydrophilic and traps water which forms a barrier at the critical interface between urine and the bladder. The result

is a highly impermeable urothelium that serves as a key protective barrier for the bladder interstitium [2].

One of the most difficult aspects in the treatment of BPS/IC is its diagnosis. Presently, the requirements for diagnosis involve excluding all other causes, such as urinary tract infection, bladder cancer and sexually transmitted disease. Also, a cystoscopy is optional in which the bladder is distended and hemorrhage are identified, however, the cystoscopy with hydrodistention is only used to rule-out other disorders but not used to rule in the diagnosis of BPS/IC. Although several potential candidate BPS/IC markers were isolated from the urine of patients with BPS/IC, they were a few molecules surrounded by the networks of molecular interaction and biological coordination. We all know that BPS/IC is a complex disease [3]. An individual gene for superior intelligence probably nor does it exist for BPS/IC. Instead, genes and gene products work together in complex and mysterious ways to cause the symptoms of BPS/IC [4].

The completion of the human genome project and the advance of microarray technology, which amounts to hundreds of data points for thousands or tens of thousands of genes, have enabled the use of gene chips in the diagnosis of BPS/IC. DNA molecules can be attached to small squares of glass that by analogy with computer circuits are called microarray. Micro-array platforms can be designed so that they will detect altered genes associated with BPS/IC. These advances promise to revolutionize medicine as they are expected to reveal the mechanisms underlying BPS/IC as well as to pinpoint potential targets for drug discovery and diagnostics.

To support the use of genome-based expression profiling as a single standardized microarray platform for the diagnosis of BPS/IC, we wish to put forward a comprehensive analysis of bladder epithelium of C57BL/6 mice with ovalbumin (OVA) or substance P (SP)-induced experimental IC versus *Escherichia coli* lipopolysaccharide (LPS)-induced bacterial cystitis.

Materials and methods

Source of data

We studied the gene expression based DNA/cDNA array in bladder epithelium of C57BL/6 mice with OVA, SP-induced cystitis corresponding to LPS-induced bacterial cystitis. Bladder epitheliums were examined at 1 and 4 h after OVA, SP, or LPS challenge. The data used here was available from gene expression omnibus under platform accession non GSE597.

Experimental procedures

Bladder epithelium samples from C57BL/6 mice with OVA, SP, or LPS-induced cystitis were obtained as described in Saban et al. [5]. Three bladders from each group were homogenized together in Ultraspec RNA solution (Biotecx, Houston, TX, USA) for isolation and purification of total RNA [5]. RNA was DNase-treated according to manufacturer's instructions (Clontech Laboratories, Palo Alto, CA, USA), and 10 μg of RNA was evaluated by denaturing formaldehyde/agarose gel electrophoresis [5]. This procedure was repeated using an additional three bladders in each experimental group [5]. Therefore, two pools of RNA were generated per experimental group for a total of six mice and two separate hybridizations per group [4]. Then, cDNA probes prepared from DNase-treated RNAs obtained from each of the experimental groups were hybridized simultaneously to membranes containing Atlas Mouse 1.2 Arrays (Clontech, Cat. #7853–1) [5]. The radioactively labeled complex cDNA probes were hybridized overnight to mouse cDNA expression arrays (Clontech) using ExpressHyb hybridization solution with continuous agitation at 68°C [5]. After high- and low-stringency washes, the hybridized membranes were exposed overnight at room temperature to a ST Cyclone phosphor screen (Packard BioScience Company, Downers Grove, IL, USA) [5].

Cluster analysis

Cluster analysis is a generic term that attempts to determine whether or not a given dataset contains distinct groups, and, if so, to find the groups. In this analysis, a hierarchical clustering algorithm is used, which leads to a series of hierarchical groups, reported by tree-like dendrograms. This method starts from a similarity matrix calculation between the genes to be clustered on the basis of their expression profiles. The similarity metric used here is the Mahalanobis distance, defined as follows:

$$d = \sqrt{\sum_{k=1}^{n} \frac{\left(X_{ik} - X_{jk}\right)^2}{\sigma_i^2}}$$

where X_{ik} and X_{jk} are the measured fluorescence ratio of test samples i and j and σ_i is the standard deviation of the X_{ik} over the sample set. The similarity between two clusters is based on Ward's linkage, where the inter-cluster distance is computed as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster.

Results

We have used Ward's linkage hierarchical clustering to cluster the global gene expression profiles in OVA or SP-induced experimental IC versus LPS-induced bacterial cystitis (Fig. 1). Heap maps are used to represent the measured fluorescence ratio of bladder epithelium in the 2-
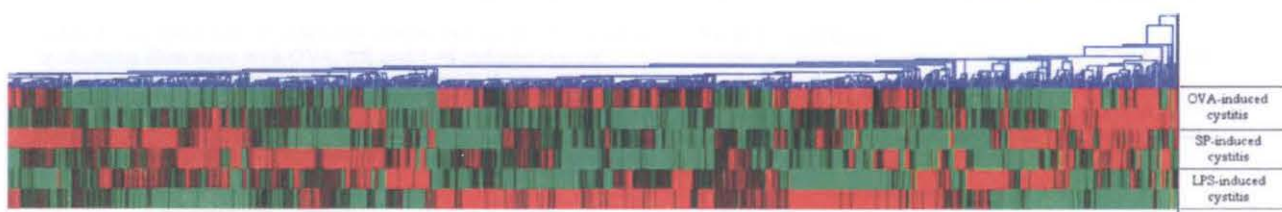
Fig. 1 Clustered display of gene expression profiles in bladder epithelium of C57BL/6 mice with experimental cystitis (OVA or SP-induced cystitis) corresponding to LPS-induced cystitis

dimensional grids and the rank-by-feature color scale scheme is used to represent the magnitude of the data on the basis of their expression patterns [6–8]. Medium data points are colored black, data points with increasing ranks are colored with reds of increasing intensity, and decreasing ranks are colored with greens of increasing intensity [6–8]. At the finest level, we have found that there is a cluster of 75 features that exhibit considerable fold increases in experimental IC (Fig. 2).

To characterize their molecular functions and processes, we perform gene ontology on all features within this cluster. This enables us to identify 75 distinct genes that are up-regulated in bladder epithelium from C57BL/6 mice with experimental IC (Table 1). Therefore, these 75 genes naturally distinguish those mice with cystitis into two kinds of cystitis, those with experimental IC and LPS-induced bacterial cystitis. In particular, we observe changes in cytokine production, cytokine- and chemokine-mediated signaling pathway, hematopoietin/interferon-class (D200-domain) cytokine receptor activity, T cell activation as well as T cell antigen processing and presentation through these processes. Several genes involved in cell adhesion, transcription factors, DNA damage and repair, mRNA cleavage, apoptosis, cell cycle, and cell proliferation are up-regulated. Changes in complex metabolic reactions are also observed, including glucose, lipid, nucleotide, xenobiotics, and amino acid metabolisms.

It can be observed that several genes involved in the regulation of Wnt, Tgf-bet and Mapk, and insulin growth factor signaling pathways are up-regulated. Therefore, the Wnt, Tgf-beta, and insulin growth factor receptor signaling pathways participate the pathophysiological development process through experimental IC process. Changes in Ephrin- and Semaphorin-mediated axon guidance are also observed [9–11]. Therefore, experimental IC leads to Ephrin- and Semaphorin-mediated axon guidance promoting

parasympathetic inflammatory reflexes [12]. In addition, a gene involved in urogenital system development (Pax2) and a gene involved in ovulation from ovarian follicle (Nos3) are up-regulated. Therefore, experimental IC also activated genes involved in urogenital system development and ovulation from ovarian follicle.

Discussion

Microarray-based genomic surveys and other high-throughput approaches (ranging from genomics to combinational chemistry) are becoming increasingly important in clinical research [6]. Therefore, we need to enable ourselves to "see" the information in the massive tables of quantitative measures that these techniques produce [6]. Our approaches to this problem can be generalized as follows: firstly, we use Ward's linkage hierarchical clustering to arrange genes, based on the order inherent in the genes. Secondly, a "75-gene" model is developed as a single standardized microarray platform for the diagnosis of experimental IC. Thirdly, the results we represent gain insight into the mechanisms underlying experimental IC.

Considerable evidence suggests that bladder pain syndrome/ interstitial cystitis (BPS/IC) involves urogenital permeability defect that allows urinary potassium to penetrate tissue and promote symptoms [2]. Studies show that this defect is present in most patients with BPS/IC who have been tested for it, as well as in many patients with symptoms of BPS/IC who have received a variety of other diagnoses.

Consequently, if we can scan a BPS/IC patient's entire genome on urinary epithelium searching for all altered genes that predispose toward symptoms of BPS/IC, we can use genome-based minimally invasive test and therapeutic targets for BPS/IC. In future studies, we would like to

Fig. 2 A cluster with 75 features that were up-regulated in bladder epithelium of C57BL/6 mice with experimental IC (OVA or SP-induced cystitis) corresponding to LPS-induced cystitis

Table 1 Representation of diagnostic discriminating genes in bladder epitheliums from mice with OVA, SP, and LPS-induced cystitis

Gene symbol	Gene ontology
Pla1	Lipid catabolism
Ipf1	Glucose metabolism
LOC14433	Glucose metabolism
Pold1	Nucleotide metabolism
Gstm2	Amino acid metabolism; xenobiotics metabolism
Fzd4	Wnt receptor signaling pathway
Dkk1	Wnt receptor signaling pathway
Erbb2	TGF-beta receptor signaling pathway
Tgfbr1	TGF-beta receptor signaling pathway
Tgfb2	TGF-beta receptor signaling pathway
Ksr	TGF-beta receptor signaling pathway; regulation of MAPK
Map2k4	MAPKKK cascade
Igfbp1	Insulin-like growth factor binding; regulation of cell growth
Igfbp4	Insulin-like growth factor binding; regulation of cell growth
Igfbp5	Insulin-like growth factor binding; regulation of cell growth
Pax2	Urogenital system development
Nos3	Ovulation from ovarian follicle
Efna3	Axon guidance
Gnai2	Axon guidance
Sema3d	Axon guidance
Kif5c	Axon guidance
Kitl	Neural crest cell migration
Tbx1	Neural crest cell migration
Nkx6-2	Regulation of spinal cord motor neuron fate
Grin2a	Synaptic transmission
Phox2a	Somatic neuron differentiation; sympathetic nervous system development; parasympathetic nervous system development
Fmn1	Actin cytoskeleton organization and biogenesis
Cnr1	G protein coupled receptor signaling pathway
Mc4r	G protein coupled receptor signaling pathway
Rln1	Signal transduction
Cttn	Protein binding
Mdfi	Protein binding
Stap2	Protein binding
Apaf1	Apoptosis
Dad1	Apoptosis
Ei24	Apoptosis
Nbl1	Cell cycle
Cks2	Cell cycle
Ccnf	Cell cycle
Pml	DNA damage response; DNA repair
Brca1	DNA damage response; DNA repair
Dbp5	Transcription
Hoxc10	Transcription
Cand1	Transcription
Ebf2	Transcription
Pax2	Transcription
Pou6f1	Transcription
Nrli2	Transcription

Table 1 (continued)

Gene symbol	Gene ontology
Dbp	Transcription
Dlx3	Transcription
Tcf21	Transcription
Cutl1	Transcription
E2f3	Transcription
Per1	Transcription
Cited	Transcription
Bapx1	Transcription
Rara	Transcription
Dmtf1	Transcription
Zic3	Transcription
Eed	Transcription
Mafb	Transcription
Thra	Transcription (thyroid hormone receptor alpha)
Skil	Transcription
Pop4	mRNA cleavage
Hsp110	Response to stress
Icam1	T cell antigen processing and presentation
Csk	Regulation of T cell activation
Cntn1	Cell adhesion
Cd44	Cell adhesion
Col6a1	Cell adhesion
Ifnar2	Cell proliferation
Lta	Immune and inflammatory response
Rel	Cytokine production
Mpl	Hematopoietin/interferon-class (D200-domain) cytokine receptor activity
Il1r1	Cytokine and chemokine-mediated signaling pathway; immune response

define the programs of gene expression in human BPS/IC bladder epithelium. Such analysis would enable the classification of samples of unknown, equivocal or ambiguous history, thus supporting the use of genome-based expression profiling as a single standardized microarray platform for the diagnosis of BPS/IC in clinical practice.

Conflicts of interest None.

References

1. Eric S Rovner and Brett D Lebed (2005) Interstitial cystitis, eMedicine from WebMD, http://emedicine.medscape.com/article/441831-overview
2. Parsons CL (2007) the role of the urinary epithelium in the pathogenesis of interstitial cystitis/prostatitis/urethritis Urology 69 (4 Suppl):9–16. Review

3. Fall M, Oberpenning F, Peeker R (2008) Treatment of bladder pain syndrome/interstitial cystitis 2008: can we make evidence-based decisions? Eur Urol 54:65–75 Epub 2008. Review
4. Erickson DR, Schwarze SR, Dixon JK, Clark CJ, Hersh MA (2008) Differentiation associated changes in gene expression profiles of interstitial cystitis and control urothelial cells. J Urol 180:2681–2687 Epub 2008
5. Saban MR, Nguyen NB, Hammond TG, Saban R (2002) Gene expression profiling of mouse bladder inflammatory responses to LPS, substance P, and antigen-stimulation. Am J Pathol 160:2095–2110
6. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95:14863–14868
7. Chen I, Chen M, Jin C, Yan H (2008) Large-scale cluster analysis of elite male and female swimmers' race patterns. Int J Sport Sci Eng 2:123–129
8. Chen I, Homma H, Jin C, Yan H (2007) Clustering and display of elite swimmers' race patterns across various comparable criteria at the same time. Int J Sport Sci Eng 1:129–135
9. Dickson BJ (2002) Molecular mechanisms of axon guidance. Science 298:1564–1959
10. Huot J (2004) Ephrin signaling in axon guidance. Prog Neuropsychopharmacol Biol Psychiatry 28:813–818
11. Guan KL, Rao Y (2003) Signalling mechanisms mediating neuronal responses to guidance cues. Nat Rev Neurosci 4:941–956
12. Tracey KJ (2002) Inflammatory reflexes. Nature 420:853–859

Appendix B

Genome-based expression profiling study following spinal cord injury
in the rat: an array of 48-gene model

Genome-Based Expression Profiling Study Following Spinal Cord Injury in the Rat: An Array of 48-Gene Model

Ling-Hong Tseng, Ilene Chen, Yi-Hao Lin, Ching Chung Liang,* and L. Keith Lloyd
Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital and University of Chang Gung Memorial Hospital, Linkou Medical Center, Tao-Yuan, Taiwan

Aim: To explore the potential molecular mechanisms underlying experimental neurogenic bladder dysfunction. **Methods:** With the aid of Affymetrix GeneChip Rat Genome U34A arrays, we examined microarray gene expression profiles in bladder wall tissue from female Sprague–Dawley rats within the first 3 weeks following spinal cord injury. Gene transcripts expressed in rat bladder wall tissue at 3 days, 7 days, and 3 weeks following spinal cord injury were compared to normal rat bladder wall tissue. **Results:** The Mahalanobis distance in hierarchical cluster analysis revealed a 48-gene model, which contained high expressions in rat bladder wall tissue at 3 days, 7 days, and 3 weeks following spinal cord injury. According to gene ontology, plausible molecular alterations in rat bladder wall tissue following spinal cord injury include: (1) the release of nerve growth factor (NGF) and transforming growth factor beta 1 (Tgfb1) (2) the secretion of histamine from mast cells, (3) the occurrence of blood coagulation, (4) the occurrence of N-terminal protein myristoylation, and (5) Axon guidance mediated by Ena/Vasodilator-stimulated phosphoprotein (Ena/VASP) promotes reestablishment of the bladder reflex following spinal cord injury. Such changes, jointly termed "bladder remodeling," can constitute an important long-term consequence of neurogenic bladder dysfunction. **Conclusion:** The success of this innovation has supported the use of microarray-based expression profiling as a commonplace platform for the pathogenesis and therapeutic interventions of experimental neurogenic bladder dysfunction. *Neurourol. Urodynam.* © 2009 Wiley-Liss, Inc.

Key words: neurogenic bladder (NGB); spinal cord injury (SCI); DNA microarray; gene expression; cluster analysis

INTRODUCTION

The bladder is a unique organ which needs the central nervous system and peripheral nerves to synchronize the urine storage and emptying. Neurogenic bladder (NGB) refers to dysfunction of the bladder due to neurogenic dysfunction or insult emanating from internal or external trauma, disease, or injury. Symptoms of NGB range from neurogenic detrusor overactivity to detrusor sphincter dyssynergia, depending upon the site of neurogenic insult.[1] The urethra also may be affected, resulting in sphincter underactivity or overactivity and loss of coordination with bladder function.[1] In addition, NGB can cause urine leakage-a major quality of life issue- and can lead to severe complications, such as recurrent urinary tract infection, vesicoureteral reflux, autonomic dysreflexia, renal failure, and urosepsis.

One of the most difficult aspects in the treatment of NGB is its management. In addition to a complete medical and voiding history, the requirements for management rely primarily upon accurate diagnosis with a variety of clinical examination, including videourodynamics and selective radiographic imaging studies.[1] One difficulty in the management of NGB is the establishment of low-pressure urinary drainage without the use of an indwelling urethral or suprapubic catheter.[2] An indwelling urethral or suprapubic catheter presents its own problem, which include increased risk of urinary tract infection and formation of bladder stones. When bladder compliance falls to a dangerous level despite the catheterization, enterocystoplasty has been the gold-standard therapy.[3] Although this procedure increases bladder capacity and compliance, it has been plagued by complications, such as metabolic disorders, perforations, excessive

mucous productions, and stones.[3] We need to find some better fundamental methods for the management of NGB.

The completion of the human genome project and the advance of microarray technology offer a new opportunity to gain insight into global gene expression profiles in NGB. Several clinical researchers have started to examine transcriptional changes in the central nervous system or peripheral nerves following spinal cord injury (SCI).[4–9] Based on our prior works which have shown this concept is acceptable,[10,11] we focus on examining transcriptional changes in rat bladder wall tissue following SCI. These advances help to revolutionize medicine as they are expected to reveal the potential pathogenic mechanisms of NGB as well as to pinpoint potential targets for therapeutic interventions and diagnostics.

To explore the potential pathogenic mechanisms underlying NGB, we wish to put forward a better fundamental analysis for transcriptional responses in rat bladder wall tissue following experimental SCI. The results contain novel features which are of considerable biologist interest. Fundamental to this study is that the mass of numbers produced by

WILEY
InterScience®
DISCOVER SOMETHING GREAT

DNA microarrays relating to experimental NGB are primarily driven by their genetic roots in the genome. It enables the genome to become a unifying explanation for the pathogeneses of NGB.

MATERIALS AND METHODS

Source of Data

The data used in this analysis was collected from Gene Expression Omnibus under platform accession no. GSE14096. Female Sprague–Dawley rats (170 g) were anesthetized with halothane and a complete laminectomy at the T9–T10 spinal level was performed according to the established protocols.[12] Bladder were harvested from normal and spinal cord injury (SCI) rats (3 days, 7 days, or 3 weeks post-surgery) under halothane anesthesia, snap-frozen in liquid nitrogen, and stored until further processing.[7] Institutional Review Board approval was obtained for the study.

RNA Processing

Total RNA was isolated using Trizol reagent and following the manufacturer's instructions (Invitrogen, Carlsbad, CA). Total soluble proteins were collected by homogenizing the tissue samples in Tissue Protein Extraction Reagent (Pierce Biotechnology, Rockford, IL) supplemented with 0.5 mM PMSF and by subjecting the homogenates to three freeze-thaw cycles. The tissue lysates were centrifuged and protein concentrations of the supernatant samples were measured spectrophotometrically using the Bradford method.

Microarray Hybridization

Gene transcripts expressed by rat bladder wall tissue in normal, 3 days, 7 days, and 3 weeks following SCI (1 rat from each group) were evaluated using oligonucleotide arrays. Briefly, 20 µg t of biotin-labeled cRNA was fragmented following the manufacture's instruction (Affymetrix, Santa Clara, CA). All labeled samples were hybridization to Affymetrix GeneChip Rat Genome U34A Array Set RG-U34A (Affymetrix).

Cluster Analysis

Data derived from GeneChips have two potential sources of error: experimental and analytical. While experimental error can be limited by higher quality samples and standardized hybridization procedures, analytical error is difficult to control. The standard goal of minimizing both false positives and false negatives is compromised by the assumption-laden tools used for analysis.[5] Cluster analysis is a commonly used method in the analysis of microarray gene expression data. This method starts from a similarity matrix calculation between the genes to be clustered on the basis of their expression profiles. At present, cluster analysis with the Pearson correlation coefficient, Spearman correlation coefficient or standard correlation coefficient as a distance measure has been frequently used in the analysis of microarray gene expression data. Although the Pearson correlation coefficient, Spearman correlation coefficient or standard correlation coefficient are scale invariant, they do not take the covariance between genes into consideration. However, changes in expression levels of genes do not occur as individual events, but participate in a complex and interdependent manner due to gene co-expression and co-regulation, giving to the working of lively organisms and their parts. Unless the distance measure properly takes the covariance between genes into account, the organized profiling can be disturbed as a result of ignoring interdependencies.

To account for the covariance between genes, a proper distance measure can be used in the Mahalanobis distance defined as follows:

$$d_{ij}^2 = (x_i - x_j)' S^{-1} (x_i - x_j)$$

where x_i and x_j are two data points, S is the $p \times p$ covariance matrix of X, and X is assumed to be of full rank so that S^{-1} exists.[10,11,13] The Mahalanobis distance incorporates covariance between gene expression levels as well as differences in variances.[10,11,13] It gives less weight to genes with high variance and to high covariated genes, so that all characteristics are treated as equally important by considering the covariance between genes and the results can be derived insight into the molecular biology of the disease or condition under investigation.

RESULTS

With the aid of Affymetrix GeneChip Rat Genome U34A Arrays, we examined microarray gene expression profiling of 8,799 distinct features in bladder wall tissue from female Sprague–Dawley rats within the first 3 weeks following spinal cord injury (SCI). Gene transcripts expressed by rat bladder wall tissue in 3 days, 7 days, and 3 weeks following SCI were compared to normal rat bladder wall tissue. To account for the biological interdependencies,[10,11] the Mahalanobis distance in Ward's linkage hierarchical clustering was used to cluster gene expression profiles in the rat bladder wall tissues within the first 3 weeks following SCI (Fig. 1).

A cluster of 48 genes that revealed considerable fold increase in rat bladder wall tissues in 3 days, 7 days and 3 weeks following spinal cord injury.
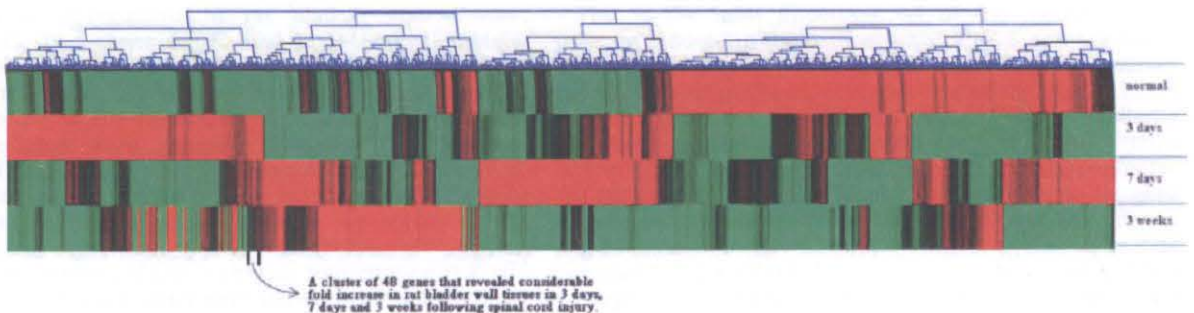
Fig. 1. Clustering results of gene transcripts expressed by rat bladder wall tissue in 3 days, 7 days, and 3 weeks following spinal cord injury were compared to normal rat bladder wall tissue.

Fig. 2. A cluster of 48 genes that revealed considerable fold increase in rat bladder wall tissues in 3 days, 7 days, and 3 weeks following spinal cord injury.

Here we used heat maps to represent the measured fluorescence intensities of the bladder wall tissues in the two-dimensional grids. Cells with intensities of zero are colored black, increasing positive intensities with reds of increasing intensity, and increasing negative intensities with greens of increasing intensity.[14] This visual representation preserves all the quantitative information, but transmits it to our brain by ways of a much higher-bandwidth channel than the "number-reading" channel.[14] At the finest level, we found that there was a cluster of 48 genes that revealed considerable fold increase in rat bladder wall tissue at 3 days, 7 days, and 3 weeks following SCI (Fig. 2). Thus, these 49 genes contained high expressions of rat bladder wall tissue in 3 days, 7 days, and 3 weeks following SCI.

To characterize their biological processes, we performed gene ontology on those 48 genes with the aid of Entrez Gene (Table I).[15–17] Specifically, changes in complex metabolisms were observed in rat bladder wall tissue following SCI, including carbohydrate, lipid, histamine, amino acid, xenobiotics, androgen, and estrogen metabolisms. We observed the gene expression for proprotein convertase subtilisin/kexin type 6 (Pcsk6) was up-regulated, which produced nerve growth factor (NGF) in rat bladder wall tissue following SCI. We also observed the gene expression for transforming growth factor beta 1 (Tgfb1) was up-regulated, which caused epithelial-mesenchymal transition in rat bladder wall tissue following SCI. The gene expression of tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta polypeptide (Ywhaz) was also up-regulated,[18] which caused histamine secretion from mast cells in rat bladder wall tissues flowing SCI.

Two genes contributed to blood coagulation (Thbd and Gzmg) were up-regulated in rat bladder wall tissues flowing SCI. Two genes contributed to N-terminal protein myristoylation (Ptpn18 and Tnk2) were up-regulated as well. Furthermore, several genes involved in potential interactions with Ena-vasodilator stimulated phosphoprotein (Evl) mediated axon guidance triggered by ubiquitin-mediated degradation were up-regulated in rat bladder wall tissues flowing SCI, suggesting the incidence and recurrence of bladder reflex following SCI. Lastly, we confirmed the changed gene expressions of four genes (Pcsk6, Tgfb1, Ywhaz, and Evl) by semiquantitative reverse transcription-PCR and observed quantitative agreement between the methods for all four genes. Semiquantitative reverse transcription-PCR revealed the elevated steady-state mRNA levels for these four genes.

DISCUSSION

Comprehensive assessment of changes in gene expression that contribute to experimental NGB has resulted in the identification of potential pathogenic mechanisms and novel therapeutic interventions. To that end, we have undertaken a comprehensive analysis to assess changes in gene expression in rat bladder wall tissue following spinal cord injury (SCI) with the aid of high-density oligonucleotide microarrays. In the present study, we sought to recognize pathogenic mechanisms underlying experimental neurogenic bladder (NGB). For this purpose, we investigated the time course of genomic up-regulations in rat bladder wall tissue following SCI. We selected genes that were up-regulated in rat bladder wall tissue following SCI and therefore, represent a relative universal bladder responses to SCI. The array analysis exhibited that of the over 8,000 distinct features examined, the 48 genes summarized in Table I were up-regulated. According to gene ontology,[11,13,14] potential molecular mechanisms in rat bladder wall tissue following SCI include: the release of nerve growth factor (NGF) and transforming growth factor beta 1 (Tgfb1), (2) the secretion of histamine from mast cells, (3) the occurrence of blood coagulation, (4) the occurrence of N-terminal protein myristoylation, and (5) Axon guidance mediated by Ena/Vasodilator-stimulated phosphoprotein (Ena/VASP) promotes re-establishment of the bladder reflex following spinal cord injury. Furthermore, several genes contributed to complex metabolisms were up-regulated in rat bladder wall tissue following SCI, including carbohydrate, lipid, histamine, amino acid, xenobiotics, androgen and estrogen metabolisms. Taken these results strongly suggest the existence of "bladder remodeling" promoting NGB formation.

This is the most important set of the transcriptional program induced in the pathogenesis of NGB, and identified 48 up-regulated genes. It is the first report in the microarray literature to employ a novel bioinformatics technique, cluster analysis with the Mahalanobis distance as a similarity measure, for identifying gene with over-expressed patterns in two populations. A major endeavor in functional genomics, in addition to revealing genome-wide changes in gene expression under defined conditions, is the "reverse engineering of signaling networks within cells" such that array data are used to inter potential upstream regulatory events.[19] This concept is based on the premise that co-expressed genes are likely to share common transcriptional regulators.[20] To address this problem, we used the Mahalanobis distance in agglomerative hierarchical clustering algorithm in order that

TABLE I. Representation of Highly Differential Expressed Genes in Bladder Wall Tissues From Female Sprague–Dawley Rats Within the First 2 Weeks Following Spinal Cord Injury

Gene symbol	Full name	Gene ontology
Pcsk6	Proprotein convertase subtilisin/kexin type 6	Nerve growth factor production
Tgfb1	Transforming growth factor, beta 1	Epithelial-mesenchymal transition
Ywhaz	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta polypeptide	Histamine secretion by mast cells
Evl	Ena-vasodilator stimulated phosphoprotein	Axon guidance
Cntf	Ciliary neurotrophic factor	Regulation of axon regeneration
Grn	Granulin	Regulation of epithelial cell differentiation
Arpcib	Actin related protein 2/3 complex, subunit 1B	Regulation of actin polymerization
Prph1	Peripherin	Intermediate filament cytoskeleton organization
Grid2	Glutamate receptor ionotropic delta 2	Synaptic transmission by the neurotransmitter glutamate
Tctex1	Dynein light chain Tctex-type 1	Microtubule-based process
Coro1b	Coronin, actin binding protein, 1B	Actin binding
Olr1191	Olfactory receptor 1191	G-Protein-coupled receptors signaling pathway
Gnb1	Guanine nucleotide binding protein beta 1	G-Protein-coupled receptors signaling pathway
Ralgds	Ral guanine nucleotide dissociation stimulator	Ras protein signaling transduction
Nras	Neuroblastoma ras oncogene	Small GTPases mediated signaling transduction
Rab2	RAB2A, member RAS oncogene family	Small GTPases mediated signaling transduction
Rab11A	RAB11A, member RAS oncogene family	Small GTPases mediated signaling transduction
Ube2a	Ubiquitin-conjugating enzyme E2A	Ubiquitin-mediated protein catabolism
Ripk3	Receptor-interacting serine-threonine kinase 3	NIK-I-kappaB/NF-kappaB cascade
Csnk2b	Casein kinase 2, beta polypeptide	NIK-I-kappaB/NF-kappaB cascade
Nup62	Nucleoporin 62 kDa	Upregulation of I-kappaB kinase/NF-kappaB cascade
Timp	Metallopeptidase inhibitor 1	Cartilage biogenesis
Col12a1	Collagen, type XII, alpha 1	Collagen fibril organization
Rgd1563250	RGD1563250	Transmembrane protein
Cstb	Cystatin B	Adult locomotory behavior
Meor	Mitochondrial trans-2-enoyl-CoA reductase	Fatty acid elongation in mitochondria
Fmod	Fibromodulin	Extracellular small interstitial proteoglycans
Hmgcs1	3-Hydroxy-3-methylglutaryl-Coenzyme A synthase 1	Carbohydrate metabolism; amino acid metabolism
Tpi1	Triosephosphate isomerase 1	Carbohydrate metabolism; lipid metabolism
Ugt1aa1	UDP glucuronosyltransferase 1 family polypeptide A1	Carbohydrate metabolism; xenobiotics metabolism; androgen and estrogen metabolism
Hdlbp	High density lipoprotein binding protein	Lipid metabolism
Pcyt2	Phosphate cytidylyltransferase 2 ethanolamine	Lipid metabolism
Sult2a1	Sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone-preferring, member 1	Lipid metabolism
Dlst	Dihydrolipoamide S-succinyltransferase	Amino acid metabolism
Hal	Histidine ammonia-lyase	Histidine metabolism
Sirt2	Sirtuin 2	Histone deacetylation
Ftl1	Ferritin light chain 1	Transport (iron ion)
Adfp	Adipose differentiation related protein	Transport (long-chain fatty acid)
Ppib	Peptidylprolyl isomerase B	Protein folding
Basp1	Brain abundant, membrane attached signal protein 1	Unknown
Lrrc8	Leucine rich repeat containing 8 family member A	Biological process
Pdlim1	PDZ and LIM domain 1	Regulation of transcription
Thbd	Thrombomodulin	Blood coagulation
Gzmg	Granzyme G	Blood coagulation
Ptpn18	Protein tyrosine phosphatase, non-receptor type 18	Protein amino acid dephosphorylation
Tnk2	Tyrosine kinase non-receptor 2	Protein amino acid dephosphorylation
Nmt1	N-Myristoyltransferase 1	N-Terminal protein myristoylation
Mgat2	Mannoside acetylglucosaminyltransferase 2	N-Terminal protein myristoylation

the correlation between genes and the differences in variances were taken into consideration.[21–24] We were able to identify 48 genes whose expression levels in rat bladder positively associated with the pathogenic mechanisms underlying experimental NGB. This technique helps us for data mining and pattern recognition in gene expression analysis. The upregulated gene clusters can be exhibited for further explanation of pathogenic mechanisms.

The present study differs from most previous approaches whose findings were obtained by gene isolating (i.e., observing the activities of one gene at any one time).[17–19,22–24] Seki et al.[22,23] checked NGF levels using enzyme-linked immunosorbent assay in chronic spinalized rats. Kikuno et al.[24] studied the combined effects of NGF and vascular endothelial growth factor on regeneration of the bladder acellular matrix graft in rats following SCI by assessing smooth muscle induction, collagen and nerve fiber regeneration. Instead of observing the activities of one gene at any one time, we performed an explanatory study, colloquially called a "fishing expedition," which is one in which multiple hypotheses were proposed for scientific investigation with the aid of microarray gene expression profiles in rat bladder wall tissue following SCI versus normal controls.[25] The primary theme of such study is to learn enough about possible causes of experimental NGB so that one or more specific hypotheses may be proposed and be sufficiently supported to justify a detailed investigation.[25]

Furthermore, our explanatory results consistent with analytical findings from other groups[21–24] in which experimental neurological bladder dysfunction resulted in the up-regulation of transcripts from Tgfb1 and genes with roles in production of NGF, metabolisms, protein synthesis and the axonal cytoskeleton. Our analysis presented here indicated that genes with roles in transports, histamine secretion by mast cells and blood coagulation were up-regulated, whereas changes in genes involved in oxidative stress, apoptosis, cell adhesion, cell growth/differentiation, and proteolysis were less common. As a result, the blockade of individual pathway generally des not fully eliminate the symptoms of NGB and optimal management is achieved only after the simultaneous blockade of multiple pathways contributing to the pathogenesis of NGB.

In summary, the present study constitutes the most definitive assessment of the functional genomics of bladder wall tissue following SCI to date. However, these results are very preliminary and further studies are needed to substantiate the clinical utility of these methods. Continued efforts of large-scale genomic data should contribute to our knowledge of NGB and hopefully will ultimately lead to improvements in treatment and recovery. In future studies, we would like to define the programs of gene expression in the human bladder in response to NGB formation. These advances promise to revolutionize medicine as they are expected to reveal the mechanisms underlying NGB as well as to pinpoint potential targets for drug discovery.

REFERENCES

1. Rackley R. Neurogenic bladder, eMedicine from web MD, http://emedicine. medscape.com/article/453539-overview.
2. Chartier-Kastler EJ, Mozer P, Denys P, et al. Neurogenic bladder management and cutaneous non-continent ileal conduit. Spinal Cord 2002;23:443–8.
3. Hodges SJ, Atala A. Regenerative medicine and the neurogenic bladder. Curr Bladder Dysfunct Rep 2008;3:67–74.
4. Gevaert T, Owsianik G, Hutchings G, et al. The loss and progressive recovery of voiding after spinal cord interruption in rats is associated with simultaneous changes in autonomous contractile bladder activity. Eur Urol 2008 [Epub ahead of print; PMID: 18599189].
5. Aimone JB, Leasure JL, Perreau VM, et al. The Christopher Reeve Paralysis Foundation Research Consortium. Spatial and temporal gene expression profiling of the contused rat spinal cord. Exp Neurol 2004;189:204–21.
6. Yang Y, Xie Y, Chai H, et al. Microarray analysis of gene expression patterns in adult spinal motoneurons after different types of axonal injuries. Brain Res 2006;23:1–12.
7. Yongzhi L, Benkang S, Jianping Z, et al. Expression of transforming growth factor beta1 gene, basic fibroblast growth factor gene and hydroxyproline in diabetes-induced bladder dysfunction in a rat model. Neurourol Urodyn 2008;27:254–9.
8. Ma Z, Liu T, Li X, et al. Identification of up-regulated genes after complete spinal cord transection in adult rats. Cell Mol Neurobiol 2006;26:277–88.
9. Wilson TS, Aziz KA, Vazques D, et al. Changes in detrusor smooth muscle myosin heavy chain mRNA expression following spinal cord injury in the mouse. Neurourol Urodyn 2005;24:89–95.
10. Tseng LH, Chen I, Chen MY, et al. Genome-based expressions profiling as a single standardized microarray platform for the diagnosis of endometrial disorder: An array of 126 genes model. Fertility and Sterility 2009 [Epub ahead of print].
11. Tseng LH, Chen I, Chen MY, et al. Genome-based expression profiling as a single standardized microarray platform for the diagnosis of experimental cystitis: An array of 75 genes model. Int J Urogynecol 2009;20:509–513.
12. Nagatomi J, DeMiguel F, Torimoto K, et al. Early molecular-level changes in rat bladder wall tissue following spinal cord injury. Biochem Biophys Res Commun 2005;9:1159–64.
13. Mimmack GM, Mason SJ, Galpin JS. Choice of distance matrices in cluster analysis of defining regions. J Climate 2001;14:2790–7.
14. Eisen MB, Spellmen PT, Brown BS, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.
15. Maglott D, Ostell J, Pruitt KD, et al. Entrez Gene: Gene-centered information at NCBI. Nucleic Acids Res 2007;35:D26–31.
16. The Gene Ontology Consortium. The gene ontology project in 2008. Nucleic Acids Res 2008;36:D440–4.
17. Carbon S, Ireland A, Mungall CJ, et al. Web Presence Working Group. AmiGO: Online access to ontology and annotation data. Bioinformatics 2008;25:288–9.
18. Fujii T, Ueeda T. Stimulation of 14-3-3 protein and its isoform on histamine secretion from permeabilized rat peritoneal mast cells. Bio Pharm Bull 2002;25:1524–7.
19. Adam RM, Eaton SH, Estrada C, et al. Mechanical stretch is a highly selective regulator of gene expression in human bladder smooth muscle cells. Physiol Genomics 2004;15:36–44.
20. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinform 2004;5:18.
21. Patel AA, Steitz JA. Splicing double: Insights from the second spliceosome. Nat Rev Mol Cell Biol 2003;4:960–70.
22. Seki S, Sasaki K, Fraser MO, et al. Immunoneutralization of nerve growth factor in lumbosacral spinal cord reduces bladder hyperreflexia in spinal cord injured rats. J Urol 2002;168:2269–74.
23. Seki S, Sasaki K, Igawa Y, et al. Suppression of detrusor-sphincter dyssynergia by immunoneutralization of nerve growth factor in lumbosacral spinal cord in spinal cord injured rats. J Urol 2004;171:478–82.
24. Kikuno N, Kawamoto K, Kirata H, et al. Nerve growth factor combined with vascular endothelial growth factor enhances regeneration of bladder acellular matrix graft in spinal cord injury-induced neurogenic rat bladder. BJU Int 2009 103;10:1424–8.
25. Schlesselman JJ. Case-control studies, design, conduct, analysis. Oxford University Press, New York, NY; 1982.

Appendix C

Genome-based expression profiling as a single standardised microarray platform for the diagnosis of endometrial disorder: an array of 126-gene model

Genome-based expression profiling as a single standardized microarray platform for the diagnosis of endometrial disorder: an array of 126-gene model

Ling-Hong Tseng, M.D.,[a,*] Ilen Chen, M.S.,[b,*] Ming-Yang Chen, M.S.,[c] Hong Yan, Ph.D.,[b]
Chao-Nin Wang, M.D.,[a] and Chyi-Long Lee, M.D.[a]

[a] Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital and University of Chang Gung School of Medicine,
Taiwan, Republic of China; [b] School of Electrical and Information Engineering, University of Sydney, Sydney, Australia; and
[c] Department of Electrical Engineering, Stanford University, Stanford, California

Objective: To assess the molecular signatures underlying endometrial disorder using cDNA microarray.
Design: Gene expression-based oligonucleotide array of the normal endometrium.
Setting: University hospital.
Patient(s): Humans.
Intervention(s): Endometrial tissues were obtained from 28 normal cycling women undergoing endometrial biopsy. RNA was extracted from each tissue and all labeled samples were hybridized to Affymetrix Human U133 plus 2.0 array.
Main Outcome Measure(s): Transcriptional response.
Result(s): Hierarchical cluster analysis with the Mahalanobis distance revealed a "126-gene" model, which are up-regulated at mid-secretory phase, moderately expressed at late-secretary phase, and down-regulated at late-secretory phase. Furthermore, the mechanisms underlying the receptivity of human endometrium at mid-secretary phase can be summarized: first, complex metabolic reactions are involved. Second, the activation of complement and coagulation cascades promotes muscle contraction, chemotaxis, phagocyte recruitment, and peritoneal inflammation. Third, Ephrin A-mediated axon guidance promotes retrograde menstruation. Fourth, autophagic degradation is suggested to be responsible for the new blood vessel formation. In addition, DKK1 is up-regulated, indicating that WNT signaling pathway may contribute to the development of endometrial disorders.
Conclusion(s): The success of this innovation has supported the use of microarray-based genome expression profiling as a single standardized platform for diagnosis of endometrial disorders. (Fertil Steril® 2009; ■:■–■.
©2009 by American Society for Reproductive Medicine.)

Key Words: Human endometrium, menstrual cycle, endometrial disorder, cDNA microarray

The human endometrium, the anatomic prerequisite for establishing and sustaining pregnancy, undergoes remarkable histologic changes throughout the menstrual cycle, in preparation for embryonic implantation and subsequent shedding and regeneration in nonconception cycle (1). The differential histologic appearance of this tissue was first described by Hitschmann and Adler (2), and since the publication of the article by Noyes et al. (3), it has been understood that endometrial histology correlates with the changes in the circulating of estradiol and progesterone. Based on an ideal 28-day cycle, Noyes et al. (3) described distinct histologic phases, which are known as proliferative (PE), early-secretory (ESE), mid-secretory (MSE), and late-secretory (LSE) phases of the menstrual cycle.

Although histologic criteria described by Noyes et al. (3) has remained the gold standard for more than 50 years in clinical diagnosis and management of women with endometrial disorder, substantial evidence has claimed the accuracy of such criteria (1). Recently, the usefulness of histologic dating of the endometrium for couples with infertility has been addressed, because histologic delay in endometrial maturation fails to discriminate between fertile and infertile couples (1, 4). In another recent study, histologic features fail to reliably distinguish between specific menstrual cycle days and narrow intervals of days, leading to the conclusion that histologic dating has neither the accuracy nor the precision to the useful in clinical practice (1).

The completion of the human genome project and the advance of microarray technology, which amounts to hundreds of data points for thousands or tens of thousands of genes, offer a new opportunity to gain insight into global gene expression profiles in the human endometrium throughout the menstrual cycle. DNA molecules can be attached to small squares of glass so called microarray by analogy with computer circuits. Microarray platforms can be designed so that they will detect altered genes within the human endometrium throughout the menstrual cycle and thus diagnose

a patient's vulnerability to endometrial disorder. Our study is to show how such methods can used to disclose the undiagnosed endometrial disorder based on all altered genes .These advances promise to revolutionize medicine, as they are expected to reveal the mechanisms responsible for endometrial disorder as well as to pinpoint potential targets for drug discovery and diagnostics.

To support the use of genome-based expression profiling as a single standardized microarray platform for the diagnosis of endometrial disorder, we wish to put forward a comprehensive analysis to study the gene expression model-based RNA/spotted DNA/cDNA array on the human endometrium at distinct phases of the menstrual cycle.

MATERIALS AND METHODS
Source of Data

The data used here were available at Gene Expression Omnibus under platform accession no. GDS4888. Endometrial samples (n = 28) were obtained from normally cycling undergoing endometrial biopsy. All subjects of the well-characterized samples were normoovulatory, with regular cycles (24–35 days), had not been on steroid hormone medications within 3 months of endometrial sampling, and were between the ages of 23 and 50 years old. Indications for the endometrial sampling (uterine leiomyoma, n = 14; endometrial biopsies, n = 8; uterine prolapse, n = 4; pelvic pain, n = 1; and ovarian cyst, n = 1). In this analysis, we study the gene expression-based oligonucleotide array of the normal endometrium at distinct phases of the menstrual cycle. The menstrual cycle was categorized into seven stages based on histopathologic criteria (3): early, mid, late-PE phases; ESE, MSE, LSE, and menstrual. Endometrial tissues were then snap frozen on dry ice immediately after collection and stored at −80°C until RNA extraction. Institutional review board approval was obtained for the study.

RNA Extraction

Total RNA was extracted using Trizol reagent (Invitrogen, Tokyo, Japan) according to the manufacture's instruction. Curetting was homogenized (1 mL Trizol/100 mg tissue) and incubated at room temperature for 5 minutes. After the addition of chloroform (0.2 volume of Trizol), samples were nurtured for another 3 minutes at room temperature, centrifuged for 15 minutes at 12,000 g (4°C), and the aqueous phase RNA was precipitated with an equal volume of 100% ethanol. Total RNA was further purified using the RNeasy Mini Kit (Qiagen, Tokyo, Japan) according to the manufacture's instruction. Some samples with a lower RNA yield were concentrated by adding 10% sodium acetate, 1% glycogen, and 2.5 vol 100% cold ethanol. They were then incubated overnight at −20°C, centrifuged at 4°C for 30 minutes at 20,000 g, and washed once with cold 80% ethanol, with air drying of the RNA pellet. Samples were stored in RNase-free H_2O, and the purity was analyzed by both the 260/280 absorbance ratio as well as gel electrophoresis.

Microarray Hybridization

All microarray procedures followed Affymetrix protocols. Concisely, biotinylated cRNA was then prepared using the Enzo BioArray high-yield T7 transcript labeling kits (Enzo, Farmingdale, NY.) cRNA was subsequently purified using the clean-up module and fragmented using 5× fragmentation buffer (200 mM Tris [pH 8.1], 500 mM KOAc, 150 mM MgOAc). All labeled samples were hybridized to GeneChip Human U133 plus 2.0 array (Affymetrix, Emeryville, CA).

Cluster Analysis

Cluster analysis is a commonly used method in the analysis of microarray gene expression profiles. This method starts from a similarity matrix calculation between the genes to be clustered on the basis of their expression profiles. The similarity measure can be used is the Mahalanobis distance defined as follows:

$$d_{ij}^2 = \left(x_i - x_j\right)' S^{-1} \left(x_i - x_j\right)$$

where x_i and x_j are two data points, S is the $p \times p$ covariance matrix of X, and X is assumed to be of full rank so that S^{-1} exists (5). In this note, a hierarchical clustering algorithm is used, which leads to a series of hierarchical grouping, reported by tree-like dendrograms. The similarity between two clusters is based on the Ward's linkage, where the intercluster distance is computed as the increase in the "error sum of squares" after fusing two clusters into a single cluster.

RESULTS

The Ward's linkage hierarchical clustering is used to arrange the global gene expression profiles in the human endometrium at ESE, MSE, and LSE phases throughout the menstrual cycle (Fig. 1). Here we use heat maps to represent the measured fluorescence ratio of the endometrium on the two-dimensional grids. Cells with log ratios of zero are colored black, increasing positive log ratios with reds of increasing intensity, and increasing negative log ratios with greens of increasing intensity (6). This visual representation preserves all the quantitative information, but transmits it to our brain by means of a much higher bandwidth channel than the "number-reading" channel (6). At the finest level, we have found that there is a cluster of 209 features that were up-regulated in the endometrium at the MSE phase, moderately expressed for those at LSE phase, and down-regulated for those at LSE phase (Fig. 2).

To characterize their molecular functions, gene ontology within this cluster is performed on all features using Entrez Gene Database (7). This enables us to identify 126 distinct genes that were up-regulated in the endometrium at midsecretory phase, moderately expressed for those at LSE phase, and down-regulated for those at LSE phase. As a result, those 126 genes naturally classify tissues into three distinct phases, those at ESE, MSE, and LSE (Table 1). In particular, the nature of genes involved in ovarian follicle development

FIGURE 1

Cluster analysis displayed the global gene expression profiles in the human endometrium at ESE, mid-secretory, and LSE phases throughout the menstrual cycle.

Tseng. Endometrial disorder: a 126-gene model. Fertil Steril 2009.

(ANG), apoptosis (ANXA4, IER3, and SGK1), cell adhesion (AMIGO2, CLDN4, COMP, HABP2, LAMB3, MUC16, and SPP1), cell differentiation and growth (CATSPERB, RARRES1, and GADD45A) are altered throughout the cycle. Changes involved in complex metabolism are observed, including carbohydrate, glucose, lipid, cofactors, vitamins, xenobiotics, and amino acid, at mid-secretory phase, suggesting the presence of extracellular remodeling at the MSE phase. These metabolisms require the release of energy, which is made available for mRNA encoding small GTPases and ATP motor proteins.

Two genes involved in vehicle explanation and completion intended for autophagy (GABARAPL1, GABARAPL3) are up-regulated in the endometrium at the MSE phase, moderately expressed for those at the LSE phase, and down-regulated for those at LSE phase (8, 9), suggesting that autophagic degradation is the mechanism responsible for new blood vessel formation in the endometrium throughout the menstrual cycle. It can be observed that several genes involve in potential interactions with complement and coagulation cascades are up-regulated in the endometrium at the MSE phase, moderately expressed for those at the LSE phase, and down-regulated for those at the LSE phase, which subsequently promote muscle contraction, chemotaxis, phagocyte recruitment, and peritoneal inflammation (10, 11). Several genes involved in potential interaction with Ephrin A-mediated axon guidance and up-regulated in the endometrium at MSE phase, moderately expressed for those at the LSE phase and down-regulated for those at LSE phase, indicating that Ephrin A-mediated axon reflexes trigger retrograde menstruation at the MSE phase. In addition, DKK1, a gene involved

in the regulation of the WNT signaling pathway, is up-regulated in the endometrium at the MSE phase, moderately expressed for those at the LSE phase, and down-regulated for those at LSE phase, suggesting that Wnt signaling pathway may play a role in the progress of endometrial disorders.

In summary, the mechanisms implicated in the human endometrium at the MSE and LSH phases can be summarized as follows: first, complex metabolic reactions, including carbohydrate, glucose, lipid, cofactors, vitamins, xenobiotics, and amino acid, are involved. Next, the activation of complement and coagulation cascades promotes muscle contraction, chemotaxis, phagocyte recruitment, and peritoneal inflammation. Third, Ephrin A-mediated axon guidance promotes retrograde menstruation. Fourth, there is an indication of vehicle explanation and completion intended to autophagy, suggesting that autophagic degradation is the mechanism responsible for the new blood vessel formation in the human endometrium during the menstrual cycle. In addition, WNT signaling pathway may contribute to the development of endometrial disorders.

DISCUSSION

This is the most significant set of the transcriptional program on human endometrium to date, and identified a total of 126 genes that are up-regulated in the human endometrium at the MSE phase, moderately expressed for those at the LSE phase, and down-regulated for those at the LSE phase. It is the first report in the microarray literature to employ a novel bioinformatics technique, cluster analysis with the Mahalanobis distance as a similarity for identifying genes that naturally

FIGURE 2

A cluster of 209 features that was up-regulated in the endometrium at MSE phase, moderately expressed at late-secretory phase, and down-regulated at LSE phase.

Tseng. Endometrial disorder: a 126-gene model. Fertil Steril 2009.

TABLE 1

Representation of gene in the human endometrium, which were up-regulated in the endometrium at mid-secretory phase, moderately expressed for those at late-secretary phase, and down-regulated for those at late-secretory phase.

Gene symbol	Gene ontology
ANG	ovarian follicle development
ANXA4	apoptosis
IER3	apoptosis
SGK1	apoptosis
AMIGO2	cell adhesion
MUC16	cell adhesion
CLDN4	cell adhesion
COMP	cell adhesion
HABP2	cell adhesion
LAMB3	cell adhesion
SPP1	cell adhesion
CATSPERB	cell differentiation
RARRES1	cell differentiation
GADD45A	cell cycle
LEFTY1	cell growth
EFNA1	axon guidance
MET	axon guidance
DNER	neuron migration
EDNRB	neural crest cell migration
S100P	endothelial cell migration
SYNE2	actin binding
KRT7	cytoskeleton reorganization and biogenesis
CRYAB	muscle contraction
GABARAPL1 (ATP8)	autophagy
GABARAPL3	autophagy
DPP4	proteolysis
PTPRR	amino acid phosphorylation
TSPAN8	amino acid glycosylation
MUC20	protein homooligomer assembly
C3	immune response; blood coagulation
C4BPA	immune response; blood coagulation
SERPING1	immune response; blood coagulation
THBD	immune response; blood coagulation
CD55	immune response; complement activation
CFD	immune response; complement activation
GBP2	immune response

TABLE 1

Continued.

Gene symbol	Gene ontology
CXCL14	immune response
IL15	immune response
CRISP3	immune response
DEFB1	immune response
CLEC4E	immune response
BCL6	immune response
PAPLN	serine-type endopeptidase inhibitor activity
SLPI	serine-type endopeptidase inhibitor activity
GDF15	signal transduction
GPR110	signal transduction
RHPN2	signal transduction
DEPDC6	intercellular signaling cascade
MAP3K5	MAPKKK cascade
DKK1	down-regulation of frizzled signaling pathway
ARID5B	transcription
ELL2	transcription
FOXO1	transcription
CITED2	transcription
CEBPD	transcription
KLF6	transcription
NFIL3	transcription
PAX8	transcription
FOSL2	transcription
NFKBIZ	transcription
IRX3	transcription
MYOCD	transcription
FAM3C	biological process
SCGB2A2	biological process
HRASLS3	biological process
GRAMD1C	unknown
MUM1L1	unknown
PHYHIPL	unknown
SESTD1	unknown
RAD21L1	unknown
C10orf10	open reading frame
C21org63	open reading frame
RNASE4	mRNA cleavage
FAM84B	hypothetical protein
FAM134B	hypothetic protein
FAM149A	hypothetic protein
TC2N	calcium binding
STC1	calcium ion homeostasis
NDRG1	response to metal ion
LIMS3	metal ion binding
MT2A	metal ion binding
MT1E	metal ion binding

TABLE 1

Continued.

Gene symbol	Gene ontology
MT1F	metal ion binding
MT1M	metal ion binding
MT1G	metal ion binding
MT1H	metal ion binding
MT1X	metal ion binding
GCNT3	carbohydrate metabolism
IRS2	glucose metabolism
ALDH1A3	lipid metabolism
APOD	lipid metabolism
SOD2	lipid metabolism
SRD5A3	lipid metabolism
CP	cofactors and vitamins metabolism
INDO	cofactors and vitamins metabolism
NNMT	cofactors and vitamins metabolism
VNN1	cofactors and vitamins metabolism
AOX1	amino acid metabolism
ARG2	amino acid metabolism
GPX3	amino acid metabolism
MGST1	amino acid metabolism
MAOA	amino acid metabolism
HAL	amino acid metabolism
GNLY	xenobiotics metabolism
RDH10	oxidation reduction
INDOL1	oxidation reduction
ABCC3	transport
PAEP	transport
MFSD4	transport
RBP4	transport
TRPM8	transport (ion)
SCARA5	transport (phosphate)
SLC15A1	transport (oligopeptide)
SLC15A4	transport (oligopeptide)
SLC16A3	transport (monocarboxylic acid)
SLC18A2	transport (monoamine)
SLC44A4	transport (choline)
SLC1A1	transport (L-glutamate, dicarboxylic acid)
SLC3A1	transport (amino acid)
SLC4A1	transport (anion)
SLC7A2	transport (L-amino acid)
SLC30A2	transport (cation, zinc ion)
SLC38A1	transport (neutral amino acid)
TMEM37	transport (calcium ion)
GABRE	transport (ion)
TCN1	transport (cobalt ion, cobalamin)

Tseng. Endometrial disorder: a 126-gene model. Fertil Steril 2009.

discriminate endometrial tissues into ESE, MSE, and LSE phases.

In microarray-based studies of gene expression, a confounder or confounding gene can be defined as an extraneous gene that satisfies both of two conditions: [1] it is a differential expressed gene for the study disease or physiologic condition, and [2] it is associated with other differential expressed gene that is not a direct consequence of that differential expressed gene. In planning microarray data analysis of gene expression, one should always regard any known differential expressed gene for the study disease or condition as a potential confounder. The bioinformatics algorithm should be developed to either assess or eliminate the effects of confounder. As a rule, a differential expressed gene that is confounder in the sample may be dealt with by adjusting procedures that reply on linear logistic regression model (12) and cluster analysis with the Mahalanobis distance as a similarity measure (5), which gives straightforward ways to account for the confounding. The consequence of ignoring the confounding is that incorrect differential expressed genes underlying the study disease or physiologic condition can be obtained.

Our approach is a generalized one, compatible with all known longitudinal or matched case–control data analysis, and free of arbitrary effects because of the role of shifts in gene expression (13). Therefore, this approach may be used to develop genome-based diagnosis for many other kinds of diseases or conditions (14, 15). Our original idea is to see if there is a unique pattern for each individualized endometrial disorder and if we had tissue with an unknown entity we can compare with what we had known, then offer proper treatment. In this present study, we have used this approach to search for the roots of endometrial disorder based on all altered genes that predispose toward this. This example expresses a feature of gene expression that makes conditional regression modeling particularly worthwhile, namely, the tendency of gene expression data to derive insight into the pathogeneses of endometrial disorder. It is, of course, not very surprising that this approach as presented has resolved problems caused by the role of shifts in the expression of genes, and that is has made possible a solid philosophy of space, time, and gene expression data. The success of this computational approach has given us confidence to "see" the information in gene expression based on biologic necessities.

In conclusion, these advances promise to revolutionize medicine, as they are expected to reveal the mechanisms underlying endometrial disorder as well as to pinpoint potential targets for drug discovery and diagnostics. However, these results are very preliminary, and further studies are needed to substantiate the clinical utility of these methods.

Authors' Contribution

L.H. Tseng wrote the article; I. Chen and M.Y. Chen developed analytical tools: I. Chen analyzed data and wrote the article; H. Yan and C.N. Wang validated the results; C.L. Lee supervised the project.

REFERENCES

1. Talbi S, Hamilton E, Vo KC, Tulac S, Overgaard MT, Dosiou C, et al. Molecular phenotyping of human endometrium distinguishes menstrual cycle phases and underlying biological processes in normo-ovulatory women. Endocrinology 2006;147:1097–121.

2. Hitschmann F, Adler L. er Bau der Uterus-schleimhaut des geschlechtsreifen Weibes mit besonderer Berucksightigung der Menstruation. Monatschr Geburtsch Gynakol 1908;27:1–23.

3. Noyes RW, Hertig AT, Rock J. Dating the endometrial biopsy. Fertil Steril 1950;1:3–17.

4. Giudice C, Ferenczy A. The endometrial cycle: morphologic and biochemical events. In: Adashi EY, Rosenwaks Z, eds. Reproductive endocrinology, surgery, and technology. New York: Raven Press, 1996: 271–300.

5. Mimmack GM, Mason SJ, Galpin JS. Choice of distance matrices in cluster analysis of defining regions. J Climate 2001;14:2790–7.

6. Eisen MB, Spellman PT, Brown PS, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.

7. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 2007;35:D26–31.

8. Klionsky J, Emr SD. Autophagy as a regulated pathway of cellular degradation. Science 2000;290:1717–21.

9. Bechet D, Tassa A, Taillandier D, Combaret L, Attaixet D. Lysosomal proteolysis in skeletal muscle. Int J Biochem Cell Biol 2005;37:2098–114.

10. Bhold D, Stahl GL. Therapeutic potential of targeting the complement cascade in critical care medicine. Crit Care Med 2003;31:S97–104.

11. Turnberg D, Botto M. The regulation of the complement system: insights from genetically-engineered mice. Mol Immunol 2003;40:145–53.

12. Collet D. Modeling binary data. London: Chapman and Hall, 1991.

13. Chen I. Monitoring athletes' physiological responses to endurance training with genome-wide expression data. Int J Sport Sci Eng 2007;1:147–56.

14. Tseng LH, Chen I, Chen MY, Lee CL, Lo TS, Lloyd LK. Genome-based expression profiles as a single standardized microarray platform for the diagnosis of experimental interstitial cystitis: An array of 75 genes model. Int Urogynecol J Pelvic Floor Dysfunct. Published online January 20, 2009 [Epub ahead of print].

15. Tseng LH, Chen I, Chen MY, Lee CL, Lin YH, Lloyd LK. Genome-based expression profiles as a single standardized microarray platform for the diagnosis of bladder pain syndrome/interstitial cystitis: An array of 139 genes model. Int Urogynecol J Pelvic Floor Dysfunct. Published online February 13, 2009 [Epub ahead of print].

Appendix D

Genome-based expression profiles study for the pathogenesis of pelvic organ prolapse: an array of 33 genes model

ORIGINAL ARTICLE

Genome-based expression profiles study for the pathogenesis of pelvic organ prolapse: an array of 33 genes model

Ling-Hong Tseng · Ilene Chen · Yi-Hao Lin ·
Ming-Yang Chen · Tsia-Shu Lo · Chyi-Long Lee

Abstract

Introduction and hypothesis To explore the potential
molecular mechanisms contributing to the pathogenesis of
pelvic organ prolapse (POP) with the aid of high-density
oligonucleotide microarrays.

Methods We compared microarray gene expression profiles
in pelvic connective tissue from women with POP and
nonprolapse controls. The round ligament and uterosacral
ligament tissues were removed from each subject at the
time of laparoscopic hysterectomy. RNA was then
extracted, and all labeled samples were hybridized to ABI
Human Genome Survey Microarray version 2.0 (Applied
Biosystems, CA, USA).

Results The Mahalanobis distance in hierarchical cluster
analysis revealed a model of 33 genes, which contained
high expressions of round and uterosacral ligaments from
women with POP. According to gene ontology, the
expressions of mitochondrial genes encoding ribosomal
protein were upregulated. Genes involved in potential

Ling-Hong Tseng and Ilene Chen contributed equally to this paper.

L.-H. Tseng · Y.-H. Lin · T.-S. Lo · C.-L. Lee (✉)
Department of Obstetrics and Gynecology,
Chang Gung Memorial Hospital and University of Chang Gung
School of Medicine,
5 Fu-Hsing Street, Kwei-Shan,
Tao-Yuan 333, Taiwan
e-mail: 3g7330@yahoo.com.tw

I. Chen
School of Electrical and Information Engineering,
University of Sydney,
Sydney, Australia

M.-Y. Chen
Department of Electrical Engineering, Stanford University,
Stanford, California, USA

interactions with mitochondrial electron transport, nucleosome
assembly, cell cycle, and apoptosis were also upregulated.
As a result, defective mitochondrial translation caused by
ribosomal protein contributes to the potential molecular
etiology of POP. Such changes, jointly termed "remodeling
of pelvic connective tissue", can constitute an important
long-term consequence of POP.

Conclusions Our results support the use of genome-based
expression profiling as a commonplace platform for
diagnostic tests of POP.

Keywords Pelvic organ prolapse (POP) · Gene expression ·
DNA microarray · Cluster analysis

Introduction

Pelvic organ prolapse (POP) is a widespread problem that
affects quality of life of countless millions of women [1].
Although POP is rarely life-threatening and is not as
dramatic as cancer, it has been shown that self-perception
of the body is significantly affected in those with symptoms
[1]. Data from 1997 revealed approximately 350,000 POP
operations performed in the United States alone [2]. Also, it
is estimated that 11% of women over the age of 80 will
undergo surgery for such conditions, with an additional
30% who will require a repeat surgery [3]. Possible risk
factors for POP include genetic predisposition, parity
(particularly vaginal birth), connective tissue disorders,
and factors associated with elevated intra-abdominal pres-
sure (e.g., obesity, chronic constipation) [4–6]. However,
the underlying mechanisms remain undetermined and
irrespective of the inciting factor; the potential molecular
mechanisms contributing to the pathogenesis of POP are
still unclear.

One of the most difficult aspects in the treatment of POP is its assessment and diagnosis. Currently, POP is measured with the POP quantification (POPQ) system [7]. This exam measures nine points in and around the vaginal canal using centimeters to give an outline of the POP [7]. Though POPQ system has an excellent inter- and intraobserver correlation and allows comparison within and between patients [8], it is not easy to be used for two reasons: First, it has been shown that POPQ staging of the anterior and apical compartments suggested a greater stage than clinically suggested [9]. Next, during any assessment of pelvic organ support, the clinician should confirm the results of examinations of patients complaining of symptoms suggestive of prolapse [7].

The advance of microarray technology and the draft sequence of the human genome offer a new opportunity to gain insight into global gene expression profiles in pelvic connective tissue from women with POP. Prior studies [10–16] offered some clues for the molecular etiology of POP. In addition, our prior works [17, 18] have supported the use of genome-based expression profiling as a commonplace microarray platform for the study of clinical disease. Therefore, we wish to put forward a better fundamental analysis for transcriptional responses in pelvic connective tissue from women with POP corresponding to nonprolapse controls. The results might contain novel features which are of considerable biological interest. Fundamental to this study is that the mass of numbers produced by DNA microarrays relating to POP are primarily driven by their genetic roots in the human genome. It enables the genome to become a unifying explanation for the pathogenesis of POP.

Material and methods

Study subjects

The data used here was available at Gene Expression Omnibus under platform accession number GSE 12852. This is a case-control study of seven premenopausal women with POP corresponding to nine premenopausal nonprolapse controls. Women with a history of endometriosis, adenomyosis, uterine fibroids, connective tissue disorders, and pelvic inflammatory conditions were excluded. Two separate pelvic support tissues were collected from each subject. The round and uterosacral ligaments were obtained at the time of laparoscopic hysterectomy. Institutional Review Board approval was obtained for the study.

RNA isolation and microarray hybridization

RNA from each tissue was extracted with the Trizol reagent according to the manufacturer's instruction (Gibco BRL

Life Technologies, NY, USA). At least 30 μg total RNA was extracted from each tissue. 1 μg RNA was subjected to gel analysis to verify the integrity of the RNA.

Microarray hybridization

All microarray procedures followed Applied BioSystems protocols. Briefly, 20 μg of biotin-labeled cRNA was fragmented following the manufacture's instruction (Applied Biosystems, CA, USA). All labeled samples were then hybridized to ABI Human Genome Survey Microarray version 2 (Applied Biosystems, CA, USA)

Cluster analysis

Cluster analysis is a commonly used method in the analysis of microarray gene expression profiles. This method starts from a similarity matrix calculation between the genes to be clustered on the basis of their expression profiles. At present, cluster analysis with the correlation coefficient (e.g., Pearson correlation coefficient, Spearman correlation coefficient, and standard correlation coefficient) as a distance measure has been commonly used in the analysis of microarray gene expression data. Although the correlation coefficient is scale invariant, it does not take the covariance between genes into consideration. However, changes in expression levels of genes do not occur as individual events, but participate in a complex and interdependent manner due to gene coexpression and coregulation, giving rise to the working of lively organisms and their parts. Unless the distance measure properly takes the covariance between genes into account, the organized profiling can be disturbed as a result of ignoring interdependencies.

To account for the correlation between genes, a proper distance measure can be used is the Mahalanobis distance defined as follows:

$$d_{ij}^2 = \left(x_i - x_j\right)' S^{-1} \left(x_i - x_j\right)$$

where x_i and x_j are two data points, S is the $p \times p$ covariance matrix of X, and X is assumed to be of full rank so that S^{-1} exists [17–19]. The Mahalanobis distance is a distance measure introduced by P. C. Mahalanobis in 1936 [20]. It is based on correlations between variables by which different patterns can be identified and analyzed. The Mahalanobis distance incorporates covariance between gene expression levels as well as differences in variances [17–19]. It gives less weight to genes with high variance and to high correlated genes, so that all characteristics are treated as equally important [17–19]. Thus far, the Mahalanobis distance has not explicitly been commonly used in the organization of microarray gene expression data.
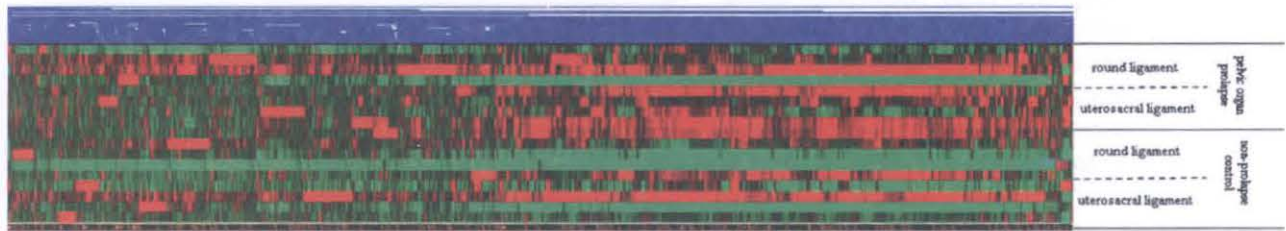
Fig. 1 Clustered display of gene expression profiles in round and uterosacral ligaments from women with pelvic organ prolapse corresponding to nonprolapse controls

Results

With the aid of high-density oligonucleotide microarrays, we examined microarray gene expression profiling of 32,878 distinct features in round and uterosacral ligaments from women with POP corresponding to nonprolapse controls. Gene transcripts expressed by round and uterosacral ligaments from women with POP were compared to those without history of POP. To take the interdependencies between genes into account [17, 18], the Mahalanobis distance in Ward's linkage hierarchical clustering was used to cluster the global gene expression profiles in round and uterosacral ligaments from women with POP corresponding to nonprolapse controls (Fig. 1). Here, we used heat maps to represent the measured intensity of the round and uterosacral ligaments in the two-dimensional grids. Cells with intensities of zero are colored black, increasing positive intensities with reds of increasing intensity and increasing negative intensities with greens of increasing intensity [21]. This visual representation preserves all the quantitative information, but transmits it to our brain by ways of a much higher bandwidth channel than the "number-reading" channel [22]. At the finest level, we found that there was a cluster of 33 genes that revealed

substantial fold increase in round and uterosacral ligaments from women with POP (Fig. 2).

To characterize their biological processes, we performed gene ontology on those 33 genes with the aid of Entrez Gene (Table 1) [23–25]. Specifically, we observed a gene contributing to O-glycan biosynthesis (C1GALT1C1) was upregulated. Several genes contributing to transport, transcription, and translation were also upregulated. The expressions of genes encoding ribosomal protein (RPL14, RPL14A, RPL15, RPL15A, RPL16, RPL22, RPL24, and RPSA) were upregulated in round ligament and uterosacral ligaments from women with POP. Genes involved in potential interactions with mitochondrial electron transport (PSMA3, NDUFB5, SDHB, COX7A2L, and C1GALT1C1), nucleosome assembly (H2BFS, HIST1H2BH, and HIST1H2BE), cell cycle (CKS1B and CCDC5), and apoptosis (PDCD10) were upregulated in round and uterosacral ligaments from women with POP as well. Lastly, we confirmed the changed gene expressions of 20 genes (C1GALT1C1, RPL14, RPL14A, RPL15, RPL15A, RPL16, RPL22, RPL24, RPSA, PSMA3, NDUFB5, SDHB, COX7A2L, C1GALT1C1, H2BFS, HIST1H2BH, HIST1H2BE, CKS1B, CCDC5, and PDCD1) by semiquan-

Fig. 2 A cluster with 33 genes those were upregulated in round and uterosacral ligaments from women with pelvic organ prolapse
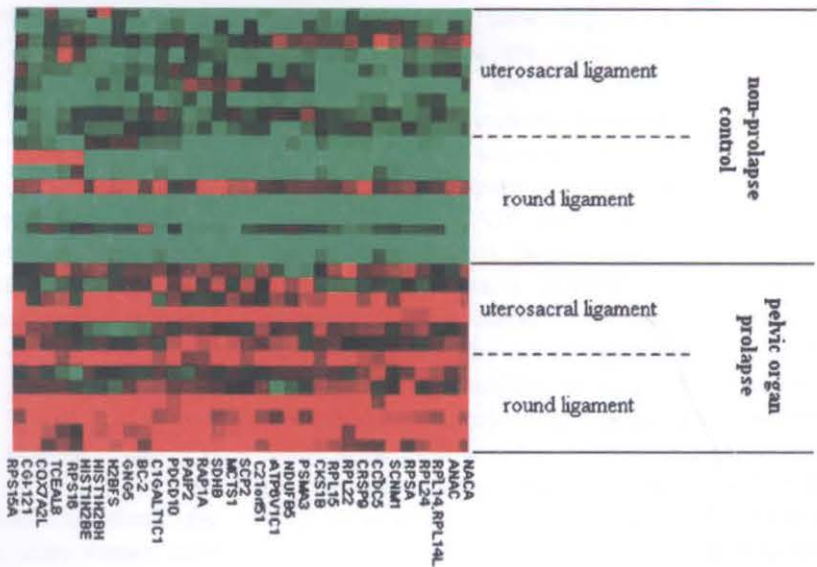
Table 1 Representation of highly differential expressed genes in round ligament and uterosacral ligaments from female patients with pelvic organ prolapse

Gene symbol	Gene ontology
CRSP9	Transcription
MCTS1	Transcription
TCEAL8	Transcription
C21orf51	Open reading frame
SCNM1	mRNA processing
PAIP2	Translation
RPL14	Translation
RPL14A	Translation
RPL15	Translation
RPL15A	Translation
RPL16	Translation
RPL24	Translation
RPL22	Translation
RPSA	Translation
NACA	Protein transport
ANAC	Protein transport
BC-2	Protein transport
ATP6V1C1	Proton transport
SCP2	Lipid transport
CKS1B	Cell cycle
CCDC5	Cell cycle
PDCD10	Apoptosis
PSMA3	Regulation of ubiquitin ligase activity during mitotic cell cycle
NDUFB5	Mitochondrial electron transport, NADH to ubiquinone
SDHB	Electron transport chain
COX7A2L	Electron transport chain
C1GALT1C1	*O*-glycan biosynthesis
CGI-121	Protein catabolism
RAP1A	Small GTPases-mediated signal transduction
GNG5	G protein-coupled receptor signaling pathway
H2BFS	Nucleosome assembly
HIST1H2BH	Nucleosome assembly
HIST1H2BE	Nucleosome assembly

titative reverse transcription-polymerase chain reaction (PCR) and observed quantitative agreement between the methods for all 20 genes. Semiquantitative reserve transcription-PCR revealed the elevated steady-state messenger RNA (mRNA) levels for these 20 genes.

Discussion

In the present study, we hypothesized that, in addition to the mechanical stresses on the pelvic floor, some women have genetic differences that predispose them to POP. To that end, we have undertaken a comprehensive analysis to assess upregulated patterns of gene expression in pelvic connective tissue from women with POP. We sought to recognize the potential molecular mechanisms contributing to the pathogenesis of POP. We selected genes that were upregulated in round and uterosacral ligaments from women with POP that represent relative universal transcriptional responses in POP. The array analysis revealed that of over 30,000 distinct features examined, the 33 genes summarized in Table 1 were upregulated. According to gene ontology, our analysis of transcriptional responses in pelvic connective tissue has suggested the potential molecular alterations underlying POP. First, the expressions of mitochondrial genes encoding ribosomal protein were upregulated. Second, several genes involved in potential interactions with mitochondrial electron transport, nucleosome assembly, cell cycle, and apoptosis were also upregulated. Such changes, jointly termed "remodeling of pelvic connective tissue", can constitute an important long-term consequence of POP. Taken these results entail defective mitochondrial translation caused by ribosomal protein contributes to the potential molecular etiology of POP. As a result, the therapies designed to prevent or inhibit defective mitochondrial translation caused by ribosomal protein would be beneficial in preventing or ameliorating the clinical manifestations of this disease.

A major endeavor in functional genomics, in addition to revealing genome-wide changes in gene expression under defined conditions, is the reverse engineering of signaling networks within cells such that array data are used to explain potential upstream regulatory events [26]. This concept is based on the premise that coexpressed genes are likely to share common transcriptional regulators [27]. To address this problem, we used the Mahalanobis distance in agglomerative hierarchical clustering algorithm in order that the correlation between genes and the differences in variances were taken into account [17, 18]. We were able to identify 33 genes whose expression levels in pelvic connective tissue positively associated with the molecular mechanisms contributing to the pathogenesis of POP. POP comes later as a consequence of the activation of those risk genes associated POP probably with some commonly known epigenetic risk factors.

The present study differed from most previous approaches whose analysis of microarray gene expression data were assessed by some statistical tests on an individual basis, essentially resulting in long lists of genes that are thought to have significantly changed transcriptional levels. Visco et al. [10] reported differential gene expression in pubococcygeus muscle from patients with POP, and these findings were subsequently confirmed by Hundley et al. [11]. In another study, Chen et al. [12] reported that elastin

remodeling is the molecular etiology of POP resulting in stress urinary incontinence, and they later claimed the differences in mRNA and protein expression in vaginal wall tissue from women with and without stress urinary incontinence [13, 14]. Recently, Hundley et al. [15] indicated differential messenger ribonucleic acid levels of actin- and myosin-related genes in women with POP. On the other hand, the results they presented remained uncertain for two reasons: First, they only investigated transcriptional responses of vaginal tissues or pelvic muscle from women with POP, without investigating their pelvic organ supportive tissue (e.g., round and uterosacral ligaments) which plays the most important role in uterine suspension [16]. Next, they hypothesized that elastin metabolism is the mechanism of POP and chose to examine genes involved in extracellular matrix metabolism only. Such analysis failed to address the full potential of genome-based experiments to alter our understanding of cellular biology and to reveal the potential molecular mechanisms contributing to the pathogenesis of POP. However, in biology, these changes do not occur as independent events as the lists suggest, but in a complex and interdependent manner due to gene coregulation and coexpression. To support the human genome becoming a unifying explanation for all of human biology and medicine, we need to take the interdependencies between genes into consideration. We performed an explanatory study, colloquially called a "fishing expedition", which was one in which multiple hypotheses were proposed for scientific investigation with the aid of microarray gene expression profiles in pelvic connective tissue from women with POP. The primary theme of such study was to learn enough about possible causes of POP so that one or more specific hypotheses may be proposed and be sufficiently supported to justify a detailed investigation in the future.

Although the round and uterosacral ligaments used in this study were removed from each subject at the time of laparoscopic hysterectomy, the biopsy can be done using less invasive laser capture microdissection [28]. Combining laser capture microdissection, RNA amplification protocols, and microarray technologies, it is possible to develop a commonplace microarray platform for diagnostic tests and therapeutic interventions of POP in clinical practice in a straightforward manner. For example, a 20-gene min-microarray platform for gene expression profiling of C1GALT1C1, RPL14, RPL14A, RPL15, RPL15A, RPL16, RPL22, RPL24, RPSA, PSMA3, NDUFB5, SDHB, COX7A2L, C1GALT1C1, H2BFS, HIST1H2BH, HIST1H2BE, CKS1B, CCDC5, and PDCD1 can be developed for the classification of samples of unknown, equivocal, or ambiguous history, thus, offering the basis for a genetic screening tool for women at risk of developing POP, as well as future treatments for prevention of POP.

In summary, the present study constitutes the most definitive assessment of the functional genomics of POP to date. However, these results are very preliminary, and further studies are needed to substantiate the clinical utility of these methods. Continued effects of large-scale genomic data should contribute to our knowledge of POP and hopefully, will lead to the improvement in treatment. Moreover, since POP contributes to causing stress urinary incontinence (SUI), we would like to define the programs of gene expression in SUI in the future. It will be of great interest to know the interrelationship between POP and SUI. Such analyses would give additional insights into what "works" are needed in sustaining a relationship between POP and SUI in present.

References

1. Patel PD, Amrute KV, Badlani GH (2007) Pelvic organ prolapse and stress urinary incontinence: a review of etiological factors. Indian J Urol 23:135–141
2. Jelovsek JE, Barber MD (2006) Women seeking treatment for advanced pelvic organ prolapse have decreased body imagine and quality of life. Am J Obstet Gynecol 194:1455–1461
3. Olen AL, Smith VJ, Bergstrom JO, Colling JC, Clark AL (1997) Epidemiology of surgically managed pelvic organ prolapse and urinary incontinence. Obstet Gynecol 89:501–506
4. Mant J, Painter R, Vessey M (1997) Epidemiology of genital prolapse: observations from the Oxford Family Planning Association Study. Br J Obstet Gynaecol 104:579–585
5. Maher C, Baessler K (2005) Surgical management of posterior vaginal wall prolapse: an evidence based literature review. Int Urogynecol J 17:84–88
6. Weber AM, Richter HE (2005) Pelvic organ prolapse. Obstet Gynecol 106:615–634
7. Pearce M, Swift S, Goodnight W (2008) Pelvic organ prolapse: is there a difference in POPQ exam results based on time of day, morning or afternoon? Am J Obstet Gynecol 192:200–201
8. Ghoniem M, Stanford E, Kenton K et al (2008) Evaluation and outcome measures in the treatment of female urinary stress incontinence: International Urogynecological Association (IUGA) guidelines for research and clinical practice. Int Urogynecol J Pelvic Floor Dysfunct 19:5–33
9. Harvey M, Eason E, Wells G (2003) Pelvic organ prolapse quantification (POPQ) system: should we challenge the anterior and apical staging? Neurourol Urodyn 22:274
10. Visco AG, Yuan L (2003) Differential gene expression in pubococcygeus muscle from patients with pelvic organ prolapse. Am J Obstet Gynecol 189:102–112
11. Hundley AF, Yuan L, Visco AG (2006) Skeletal muscle heavy-chain polypeptide 3 and myosin binding protein H in the pubococcygeus muscle in patients with and without pelvic organ prolapse. Am J Obstet Gynecol 194:1404–1410

12. Chen B, Wen Y, Zhang Z, Guo Y, Warrington JA, Polan ML (2006) Microarray analysis of differentially expressed genes in vaginal tissues from women with stress urinary incontinence compared with asymptomatic women. Hum Reprod 21:22–29

13. Wen Y, Polan ML, Chen B (2006) Do extracellular matrix protein expressions change with cyclic reproductive hormones in pelvic connective tissue from women with stress urinary incontinence? Hum Reprod 21:1266–1273

14. Wen Y, Zhao YY, Li S, Polan ML, Chen B (2006) Differences in mRNA and protein expression of small proteoglycans in vaginal wall tissue from women with and without stress urinary incontinence. Hum Reprod 22:1718–1724

15. Hundley AF, Yuan L, Visco AG (2008) Gene expression in the rectus abdominus muscle of patients with and without pelvic organ prolapse. Am J Obstet Gynecol 198:220.e1-7

16. DeLancey JO (1992) Anatomic aspects of vaginal eversion after hysterectomy. Am J Obstet Gynecol 166:1717–1724

17. Tseng LH, Chen I, Chen MY, Yan H, Wang CN, Lee CL (2009) Genome-based expressions profiling as a single standardized microarray platform for the diagnosis of endometrial disorder: an array of 126 genes model. Fertil Steril (Epub ahead of print)

18. Tseng LH, Chen I, Chen MY, Lee CL, Lo TS, Lloyd LK (2009) Genome-based expression profiling as a single standardized microarray platform for the diagnosis of experimental cystitis: an array of 75 genes model. Int Urogynecol J Pelvic Floor Dysfunct 20:509–513

19. Mimmack GM, Mason SJ, Galpin JS (2001) Choice of distance matrices in cluster analysis of defining regions. J Climate 14:2790–2797

20. Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2:49–55 http://ir.isical.ac.in/dspace/handle/1/1268

21. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95:14863–14868

22. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 35:D26–D31

23. The Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucleic Acids Res 36:D440–444

24. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub; Web Presence Working Group (2008) AmiGO: online access to ontology and annotation data. Bioinformatics 25:288–289

25. Adam RM, Eaton SH, Estrada C, Nimgaonkar A, Shih SC, Smith EH, Kohane IS, Bägli D, Freeman MR (2004) Mechanical stretch is a highly selective regulator of gene expression in human bladder smooth muscle cells. Physiol Genomics 20:36–44

26. Allocco DJ, Kohane IS, Butte AJ (2004) Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics 5:18

27. Simone NL, Bonner RF, Gillespie JW et al (1998) Laser-capture microdissection: opening the microscopic frontier to molecular analysis. Trends Genet 14:272–276

28. Espina V, Wulfkuhle JD, Calvert VS et al (2006) Laser-capture microdissection. Nat Protoc 1:586–603