CHAPTER IV

METHOD TO IDENTIFY RECTAL CANCER GENES WITH SIGNIFICANT DIFFERENTIAL EXPRESSION BY MICROARRAY ANALYSIS

4.1 Introduction

Microarray is a powerful technique suitable for providing analysis of many thousands of genes. Recently, the researchers are interested in the gene expression scales for comparing the gene expressions of two samples. This study concentrated on the measurement and analysis of genes from two different samples, normal and rectal cancer tissue samples in order to identify the genes which are differentially expressed by cDNA microarray. In this chapter the methodology of collecting raw microarray data and the procedure of analyzing microarray data are discussed.

4.2 Tissue sample

The procedure of collection of microarray data involved the use of reference designs in which normal tissues were assigned as reference and tumor tissues as interesting samples. The samples were collected from the biopsy specimens from 6 stage T3 rectal cancer patients at the Division of Gastroenterology, Department of Internal Medicine, Maharaj Nakorn Chiang Mai Hospital. In this study, the specimens used were the residual of tissue biopsy for diagnostic purposes without patient identification. Normal and tumor tissues, were obtained from the same patients. The patients staging were classified according to TNM staging of the

American joint committee on cancer (AJCC). After dissection, all specimens were immediately immersed in RNAlater™ to stabilize the RNA and later stored at –80 °C until required. The information of six stage III rectal cancer patients is shown in Table 4.1.

Table 4.1 Six stage III rectal cancer patients.

Code	Age	Tumor site	Pathologic	Stage
1	55	Rectum	Adeno CA, WD	T3N0M0
2	62	Rectum	Adeno CA, PD	T3N1M0
3	76	Rectum	Adeno CA, WD	T3N1M1
4	50	Rectum	Adeno CA, WD	T3N1M0
5	66	Rectum	Adeno CA, PD	T3N0M0
6	63	Rectum	Adeno CA, WD	T3N1M1

Ethical approval was granted form the Ethics Committee: Research Ethics Committee 2, Faculty of Medicine, Chiang Mai University and study code is NONMED-11-03-22A-12. The microarray experimental design of this research was a reference design as mentioned in which the control was normal tissue and the interesting sample was tumor tissue of the same rectal cancer patient. The control and samples were labeled with red and green fluorescent dyes, respectively. The Figure 4.1 shows the idea of labeling with control and sample tissues.

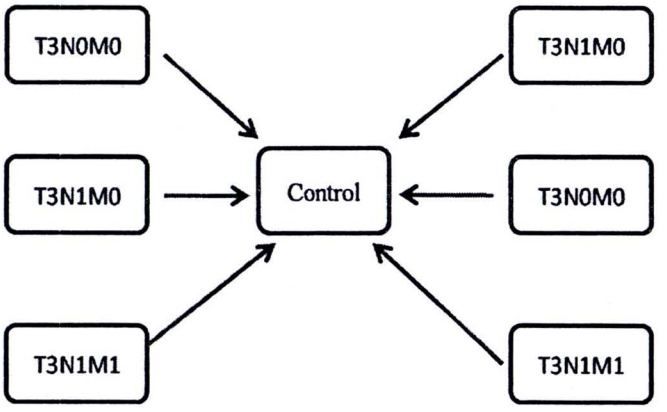

Figure 4.1 The reference design of six stage III rectal cancer patients for use in the microarray experiment.

4.3 Gene expression study by cDNA microarray

RNA was extracted from both specimens by dissolving them in a mixture of various organic solvents. RNAs were isolation by using the RNA Mini Spin kit (Illustra Tissue Mini Spin kit, GE Healthcare, Buckinghamshire, UK) followed the protocol provided by the kit company. The RNA quality control was checked by Gel electrophoresis method and the quantity of RNA was measured by UV spectrophotometer. The next process of microarray allowed mRNA to convert to cDNA and the labeled nucleotide was incorporated into the new cDNA molecules. The two samples were labeled with two different fluorescent dyes. The RNA from normal tissues used a labeling mix containing a red fluorescent dye (Alexa Fluor® 555) while the RNA from tumor used green fluorescent dye (Alexa Fluor® 647). With equal amount of the two different labelled cDNA, hybridization to the probe on the microarray glass slide was allowed to perform for 16 hours. The glass microarray slides were from the Human Genome opArray™ (Operon). After that, the microarray

glass slides were washed and scanned for gene expression level. The microarray scanning process used ScanArray Express (PerkinElmer Inc, USA) and its software for analysis of the microarray image. The quantitation method of ScanArray software is adaptive circle, and subsequently lowess method is used for the normalization procedure. The output of the quantitation method is shown in Table 4.2 (short of output Table).

Table 4.2 Some section of the output file of ScanArray Express Software.

Index	Array Row	Array Column	Spot Row	Spot Column	Name	ID	X	Y	Ch1 Median	Ch1 Mean	Ch1 SD	Ch2 Median	Ch 2 Mean	Ch2 SD
1	1	1	1	1	Dye Marker	97: D-01 Dye Marker	2382	8217	4792	12641	15739.24	1949	5496	6234.68
2	1	1	1	2	ENSG00000109846	01: D-01 H200000498	2527	8209	140	206	135.68	119	158	91.61
3	1	1	1	3	Buffer	96: D-01 Buffer	2689	8209	128	179	128.3	125	180	147.48
4	1	1	1	4	ENSG00000059377	01: D-13 H200000511	2851	8209	166	226	145.14	120	156	81.36
5	1	1	1	5	ENSG00000056558	01: H-01 H200000542	3013	8209	159	206	124.98	120	171	115.7
6	1	1	1	6	ENSG00000180644	01: H-13 H200000557	3175	8209	183	243	144.74	118	161	96.75
7	1	1	1	7	ENSG00000160654	01: L-01 H200000577	3337	8209	151	208	126.66	118	157	81.14
8	1	1	1	8	ENSG00000163032	01: L-13 H200000583	3499	8209	218	260	170.98	119	163	94.87
9	1	1	1	9	ENSG00000075213	01: P-01 H200000613	3661	8209	178	247	152.02	118	167	103.18
10	1	1	1	10	ENSG00000080608	01: P-13 H200000623	3823	8209	172	248	187.95	116	148	74.25

4.4 Microarray data analysis

The quantitation results of six arrays were analyzed using R programming language which is a program for statistical and graphical computing. R programming is provided for free download from http://www.R-project.org. In addition, the bioinformatics library BioConductor provides many packages for microarray data analysis. In this research, we chose the linear models for microarray data from Bioconductor version release 2.8. The linear models for microarray data are known as the limma package proposed by Gordon K *et al.*, (2010) (30). The microarray data were analyzed based on analysis steps in limma user's guides. The procedure for analysis of two-color microarray data consists of five main steps: importing data, quality assessment, pre-processing and normalization, data analysis with specific design and statistics for differential expression. Before the analysis procedure, we changed the working directory of the data so that all data files were assumed to be located in the same folder in this directory. The importing steps were designing the targets frame for reading raw two-color data. The layout of the target file is shown in Table 4.3. The targets file was read by ScanArray Express with the other default estimates used for foreground and background intensities provided by the limma package. We obtained the result of RG object, red-green list that used to store the raw intensity microarray data. Human OpArray.GAL file provided the information of genes in list of around 30000 genes. The information included position, gene ID, and gene name. Table 4.4 shows example information of ten genes in the top of the Human OpArray.GAL file.

Table 4.3 The layout of targets file for reading six arrays into R software.

Slide Numbers	Names	Filenames
1	C14	14.txt
2	C15	15.txt
3	C20	20.txt
4	C22	22.txt
5	C25	25.txt
6	C28	28.txt

Table 4.4 The information of ten genes in the top of Human OpArray.GAL file.

Block	Column	Row	ID	Name
1	1	1	97: D-01 Dye Marker	Dye Marker
1	2	1	01: D-01 H200000498	ENSG00000109846
1	3	1	96: D-01 Buffer	Buffer
1	4	1	01: D-13 H200000511	ENSG00000059377
1	5	1	01: H-01 H200000542	ENSG00000056558
1	6	1	01: H-13 H200000557	ENSG00000180644
1	7	1	01: L-01 H200000577	ENSG00000160654
1	8	1	01: L-13 H200000583	ENSG00000163032
1	9	1	01: P-01 H200000613	ENSG00000075213
1	10	1	01: P-13 H200000623	ENSG00000080608

The next processes were to print tip and the spot types file. The process was the arrangment of the spot from the gene list and identification of the different types of all spots. The spot types were used to set the control status of each spot. The usefulness of this process is that it is usually used to distiguish the control spots from those corresponding exactly to genes of interested in the difference of gene expression level in two different cell types. The spot types for reading into the program are specifically designed files for the spot types which are shown in the gal file. For this research eight different spot types were defined. The information of the spot types and the file format for assigning R programimg for detecting and defining the spots is shown in Table 4.5.

Table 4.5 Eight spot types of microarray data.

SpotType	ID	Name	Colour
Blank	*	-*	black
ENSG	*	ENSG*	pink
Dye Marker	*	*Dye*	yellow
Buffer	*	*Buffer	red
HIX	*	HIX*	brown
EMPTY	*	*EMPTY	blue
null	*	*null	orange
MTTC	*	MTTC*	green

Next, the quality assessment is the method concentrated in the quality for each array. The quality assessment considers the MA-plot. The usefulness of the MA-plot is to enhance and to highlight the diversity of control spots on an array while the box plot in the quality assessment can show the laser quality for each array in which the array is divergent from the others. After that, pre-processing and normalization of two-color data were performed including background correction, within-array normalization and between-array normalization. The background correction method endeavors to adjust the data for the surrounding intensity of each spot feature to get correct foreground intensities corresponding with background intensities. Different models are available to use with background correction such as none, substract , edwards normexp. In this study, we used the *normexp* method. The further details of normexp method described in chapter II. We estimated true signal (S) by using the condition expectation $E(S|X = x)$ and calculated three parameters, α, μ, σ^2 by using the maximum likelihood estimation (MLE) or saddle point estimation. The true signal results from the background correction method were denoted as R and G of each spot. The results data of background allowed to us to normalize by using loess method for within-array normalization while Aquantile method was used for between-array normalization.

Loess normalization method for analysis of gene expression data was as followed:

1. Each spot i on an array contained red and green intensities that were defined as R_i and G_i. First procedure of loess normalization calculated M and A values for each spot i: $M_i = log_2 R_i - log_2 G_i$ and $A_i = \frac{1}{2}(log_2 R_i + log_2 G_i)$ the M_i and A_i are observed value of each spot i.
2. Control loess normalization: all probes were used to calculate as

$$M_{new} = M_i - f_{loess}(A_i)$$

where $f_{loess}(A_i)$ is loess curve through the points of data.

3. Composite loess normalization: The normalization loess curve was weighted for each spot by using loess smoothing function. We selected first order polynomial to fit M_j and R_j with distance weights and spot j is the neighborhood of spot i

$$d_{ij} = d(A_i - A_j)$$

where d_{ij} is a decreasing function and close to zero if all j is the neighborhood of A_i.

4. Iteration for identification of the best points of loess curve by using a robust weights:

Let the value of polynomial at $i = j$ becomes $f(A_i)$

$$r_i = r(|M_i - f(A_i)|)$$

where $r()$ is decreasing function for giving large weight to small residuals and small weight to a large residuals. Fitting robust weight by using lest square of polynomial to obtained $f(A_i)$ and weight is change into $d_{ij}r_j$

Finally, $f(A_i)$ are subtracted from M_i normalization.

5. The spot of loess curve was defined as less proportion comparing genes as spot j with spot i .

6. Loess normalization method was checked with M-A scatter plot, and subsequently we printed out the graphic of MA-plot, M-values and A-values (28).

Following by Aquantile normalization method of between-array normalization, this method was used to reduce variability between arrays. The technique to perform

quintile-transformation of the A-values was used, so that the empirical distribution of the A-values (average intensities) was the same across all arrays and across channels.

Then, the data analysis in the limma package provided an approach for the linear model for analysis of the microarray experiment in different designs. The model specifically depends on the matrix design which corresponds to the microarray experiment objective. The Table 4.6 shows the targets frame for reading data into R programming.

Table 4.6 The targets frame of factorial design.

Filename	Lymph Node	Metastasis
CRC14	N0	M0
CRC15	N1	M0
CRC20	N1	M1
CRC22	N1	M0
CRC25	N0	M0
CRC28	N1	M1

The last process was the statistics method for detection of gene differential expressions. The limma package provides the function that can be summarized the results of linear model as follows: t-statistics, moderated t-statistics, p-value, adjust p-value and B-value. For this study, we used p-value of global t-test for identification of differential gene expressions between normal and tumor tissues. Gene regulation was measured by the values of fold change. The critical threshold of p-value of t-test and fold change was defined as 0.005 and 2-fold, respectively. The genes with

p-value less than the threshold level genes would be differ on two different samples while fold change value less than 2 indicated down-regulated between normal and tumor samples and in the same way, if fold change values were greater than 2 the genes were up-regulated between two samples. In addition, we obtained the gene description of each spot by merging the information file prepared by Human Genome OpArrayTM (Operon) (Table 4.7). Finally, we identified significantly different gene expressions correlated with major clinical characteristics of colorectal cancer by matching gene data previously published by the researches mainly study under colorectal cancer using microarray technique from literature review (47-53). Furthermore, we identified the genetic function for each gene by merging with gene ontology data from http://www.geneontology.org/GO.database.shtml.

Some information of the GO database is shown in Table 4.8.

Table 4.7 Some sections of the information provided by Human Genome OpArrayTM (Operon).

384_number	384_position	oligo_id	oligo_sequence	Gene_id	gene_symbol	description
1	A01	-	-	-	-	-
1	A03	H2000 00001	TGGGGAGAAATCTCGTGCCCAAA CCTGGTGATGGATCCCTTACTATT TAGAATAAGGAACAAAATAAAC	ENSG000 00156006	NAT2	Arylamine N-acetyltransferase 2 (EC 2.3.1.5) (Arylamide acetylase 2) (Arylamine N-acetyltransferase, polymorphic) (PNAT) (N-acetyltransferase type 2) (NAT-2). [Source:Uniprot/SWISSPROT;Acc:P11245]
1	A05	H2000 00005	GAAGGCTCTGGGTTACAGAGGCC CAAGATCCTCAACGTTGGGGACAT TGGAGGCAATGAAACAGTGACA	ENSG000 00092295	TGM1	Protein-glutamine gamma-glutamyltransferase K (EC 2.3.2.13) (Transglutaminase K) (TGase K) (TGK) (TG(K)) (Transglutaminase 1) (Epidermal TGase). [Source:Uniprot/SWISSPROT;Acc:P22735]
1	A07	H2000 00006	ATGGGTTACAGAATGCTAGGGAG GCAATTTGGTTACCTGCAATGGCT GCTTTTGCCAGCGAGGCCACCA	ENSG000 00066926	FECH	Ferrochelatase, mitochondrial precursor (EC 4.99.1.1) (Protoheme ferro-lyase) (Heme synthetase). [Source:Uniprot/SWISSPROT;Acc:P22830]
1	A09	H2000 00007	TATGGAGATCAGCACCTGGTTTGT ACCTGCCCACCCATGGAAGTTTAT GAGTCTCCATTTTCTGAACAA	-	-	-
1	A11	H2000 00008	GTCATCTTCTCCATGAAGACCACT GAATGAACACCTTTTCATCCAGCC TTAATTTCTTGCTCCATAACT	ENSG000 00149534	MS4A2	High affinity immunoglobulin epsilon receptor beta-subunit (FcERI) (IgE Fc receptor, beta-subunit) (Fc epsilon receptor I beta-chain). [Source:Uniprot/SWISSPROT;Acc:Q01362]

Table 4.7 Some sections of the information provided by Human Genome OpArrayTM (Operon) (continued).

384_number	384_position	oligo_id	oligo_sequence	Gene_id	gene_symbol	description
1	A13	H2000 00010	CATGGAGGAGCTTGGGGGATGAC TAGAGGCAGGGAGGGGACTATTT ATGAAGGCAAAAAAATTAAATTA	ENSG000 00173503	LTA	Lymphotoxin-alpha precursor (LT-alpha) (TNF-beta) (Tumor necrosis factor ligand superfamily member 1). [Source:Uniprot/SWISSPROT;Acc:P01374]
1	A15	H2000 00011	GAACAGGACGCTTATGCTATTAAT TCTTATACCAGAAGTAAAGCAGCA TGGGAAGCTGGGAAATTTGGA	ENSG000 00075239	ACAT1	Acetyl-CoA acetyltransferase, mitochondrial precursor (EC 2.3.1.9) (Acetoacetyl-CoA thiolase) (T2). [Source:Uniprot/SWISSPROT;Acc:P24752]
1	A17	H2000 00014	GTGCTGTGGGTCTTTGCCCGCCTG TACCCTTGCAAGAAATTCAATGAG ATAAAGATCTTCATGGTGAAC	ENSG000 00169403	PTAFR	Platelet activating factor receptor (PAF-R). [Source:Uniprot/SWISSPROT;Acc:P25105]
1	A19	H2000 00016	TGCACTGGTCGGTATATGGAAACA CATTGCTCTACCCTGCTACTTAGTT GATTTTAAAGTGAATTTACA	ENSG000 00165195	PIGA	Phosphatidylinositol N-acetylglucosaminyltransferase subunit A (EC 2.4.1.198) (GlcNAc-PI synthesis protein) (Phosphatidylinositol- glycan biosynthesis, class A protein) (PIG-A). [Source:Uniprot/SWISSPROT;Acc:P37287]

Table 4.8 Some information of GO database.

Ensembl Gene ID	GO ID	GO description	GO evidence code
ENSG00000176269	GO:0004984	olfactory receptor activity	IEA
ENSG00000176269	GO:0007186	G-protein coupled receptor protein signaling pathway	IEA
ENSG00000176269	GO:0016021	integral to membrane	IEA
ENSG00000172748	GO:0003676	nucleic acid binding	IEA
ENSG00000172748	GO:0008270	zinc ion binding	IEA
ENSG00000172748	GO:0006355	regulation of transcription, DNA-dependent	IEA
ENSG00000172748	GO:0005634	nucleus	IEA
ENSG00000172748	GO:0003676	nucleic acid binding	IEA
ENSG00000172748	GO:0008270	zinc ion binding	IEA
ENSG00000172748	GO:0006355	regulation of transcription, DNA-dependent	IEA