理學博士學位論文

생물정보학적 분석에 의한 SPS1 유전자 기능 및 마이크로RNA 전사 조절 인자의 예측에 관한 연구:

Studies on prediction of SPS1 gene function and microRNA transcriptional regulatory element by bioinformatical analysis

2012年 8月

서울大學校 大學院

協同科程 生物情報學 專攻

李 光 熙

생물정보학적 분석에 의한 SPS1 유전자 기능 및 microRNA 전사 조절 인자의 예측에 관한 연구

指導敎授 李 柄 宰

이 論文을 理學博士學位論文으로 提出함
2012年 5月

서울大學校 大學院
協同科程 生物情報學 專攻
李 光 熙

李 光 熙의 理學博士 學位論文을 認准함
2012年 6月

委 員 長 ______________________ (印)

副委員長 ______________________ (印)

委 員 ______________________ (印)

委 員 ______________________ (印)

委 員 ______________________ (印)

Studies on prediction of SPS1 gene function and microRNA transcriptional regulatory element by bioinformatical analysis

by

Kwang Hee Lee

Advisor

Professor Byeong Jae Lee, Ph.D.

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

August, 2012

Interdisciplinary Program in Bioinformatics
Seoul National University

Abstract

Studies on prediction of gene function and transcriptional regulatory elements by analyzing microarray data: Examples using SPS1 and microRNA genes

Kwang Hee Lee
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

Bioinformatics is an important area to analyze the massive biological data and predict the biological meanings using computational and statistical methods. Since the massive and highly qualified data have been accumulated by development of microarray technology, researches for finding biological meanings through predicting gene functions and transcriptional regulatory elements by using bioinformatical approaches are actively progressed. In these studies, we show the predicted and confirmed results for gene function and transcriptional regulatory element through two examples, selenophosphate synthetase 1 (SPS1) and microRNA genes.

For example predicting gene function, we used SPS1 which functions are unknown yet. There are two selenophosphate synthetases (SPSs) in higher eukaryotes, SPS1 and SPS2. Of these two isotypes, only SPS2 catalyzes selenophosphate synthesis. Although SPS1 does not contain selenophosphate synthesis activity, it was found to be essential for cell growth and embryogenesis. The function of SPS1, however, has not been elucidated. Using microarray data from obtained *SPS1* knockdown, differentially expressed genes were identified using two-way analysis of variance methods and clustered according to their temporal expression pattern. Gene ontology analysis was performed against differentially expressed genes and gene ontology terms related to vitamin B_6 biosynthesis were found to be significantly affected at the early stage (day 3). Interestingly, genes related to defense and amino acid metabolism were affected at the later stage (day 5) following knockdown. Levels of pyridoxal phosphate, an active form of vitamin B_6, were decreased by *SPS1* knockdown. Treatment of SL2 cells with an inhibitor of pyridoxal phosphate synthesis resulted in both a similar pattern of expression as that found by *SPS1* knockdown and the formation of megamitochondria, which is the major phenotypic change observed by *SPS1* knockdown. These results indicate that SPS1 regulates vitamin B_6 synthesis, which in turn impacts various cellular systems such as amino acid metabolism, defense and other important metabolic activities.

For example for predicting transcriptional regulatory elements, we selected miRNA genes. miRNAs are important post-transcriptional regulators of various biological processes. Although our knowledge of miRNA expression and regulation has increased considerably in recent years, the regulatory elements for miRNA gene expression, especially for intergenic miRNAs, are not fully understood. We identified the differentially methylated regions (DMRs) occurring 1000 bp upstream from all miRNAs in human neuroglioma cells using microarrays and discovered a unique sequence motif $C[N]_6CT$. This motif was preferentially located within 400 bp or from 800–1000 bp upstream of the intergenic miRNA start, corresponding to the highly methylated region. Interestingly, treatment of cells with a methyl transferase inhibitor (5-aza-2-deoxycytidine, DAC) significantly increased expression of miRNA genes with a high frequency of the C[N]$_6$CT motif in DMRs. Statistical analysis showed that the frequency of the C[N]$_6$CT motif in DMRs is highly correlated with intergenic miRNA gene expression, suggesting that C[N]$_6$CT motifs associated with DNA methylation regions play a role as regulatory elements for intergenic miRNA gene expression.

Keywords: selenophosphate synthetase 1, vitamin B$_6$, microRNA, promoter
Student Number: 2007-30782

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1.

LITERATURE REVIEW

1. BIOINFORMATICAL APPROCHES FOR EXPRESSION DATA

1.1. DNA microarray technology

DNA microarrays can simultaneously measure the expression level of thousans of genes within a specific sample. DNA microarray technology which is evolved from Southen blot is invented by Stephen P.A. Foder and colleagues in 1989 (Fodor *et al.*, 1993; Pease *et al.*, 1994). Microarrays are composed of short DNA probes which are implemented on a glass slide (Maskos *et al.*, 1992). The sequences of probes in array lattice are complementary to those of the interested genes, called target (Schena *et al.*, 1995). Targets prepared from samples are labeled with fluorescent dye and annealed with the probes on a glass slide (Figure 1.1).

Microarrays can be classified into two types according to the number of samples that can be applied on single array (Brown *et al.*, 1999); two-channel microarray and one-channel microarray (Figure 1.2). Two-channel microarrays are hybridized with cDNAs prepared from two samples to be compared. cDNAs are labeled with two different fluorescent dyes, Cyanin 3 (Cy3) and Cyanin 5 (Cy5), respectively. Cy3 emits a wavelength of 570 nm which is corresponding to the green part of the light spectrum, and Cy5 emits wavelength of 670 nm which is corresponding to the red part of the light spectrum. The two Cy-labeled cDNAs are mixed and hybridized to a single array chip, and then that is scanned in a platform-specific scanner to detect the fluorescences (Shalon *et al.*, 1996). Relative intensities of each fluorescent dye are used in ratio-based analysis to identify up-regulated and down-regulated genes (Tang *et al.*, 2007). In one-channel microarrays or single-channel microarray, cDNAs prepared from each sample are labeled with only single

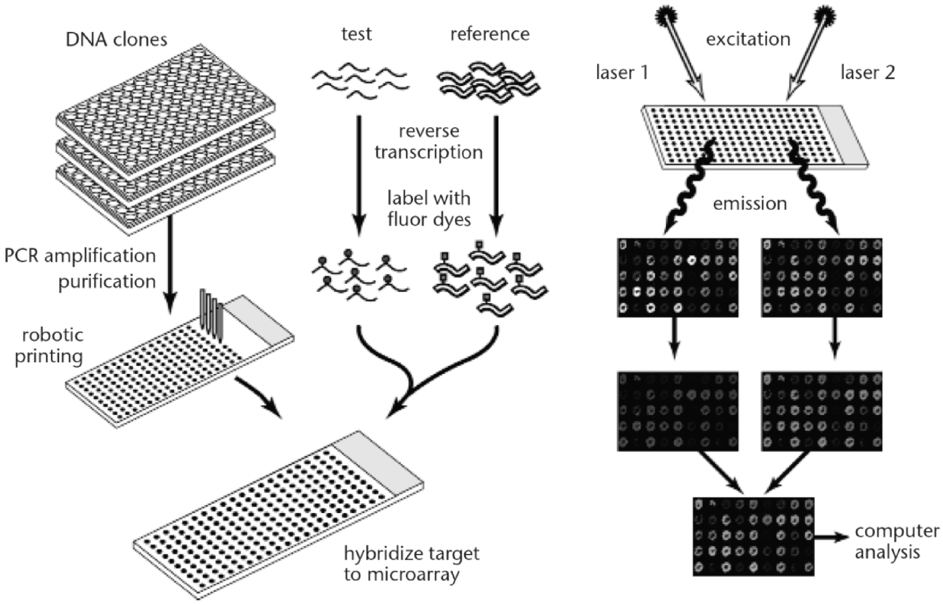


Figure 1.1. General procedure of microarray (from Duggan *et al*., 1999)

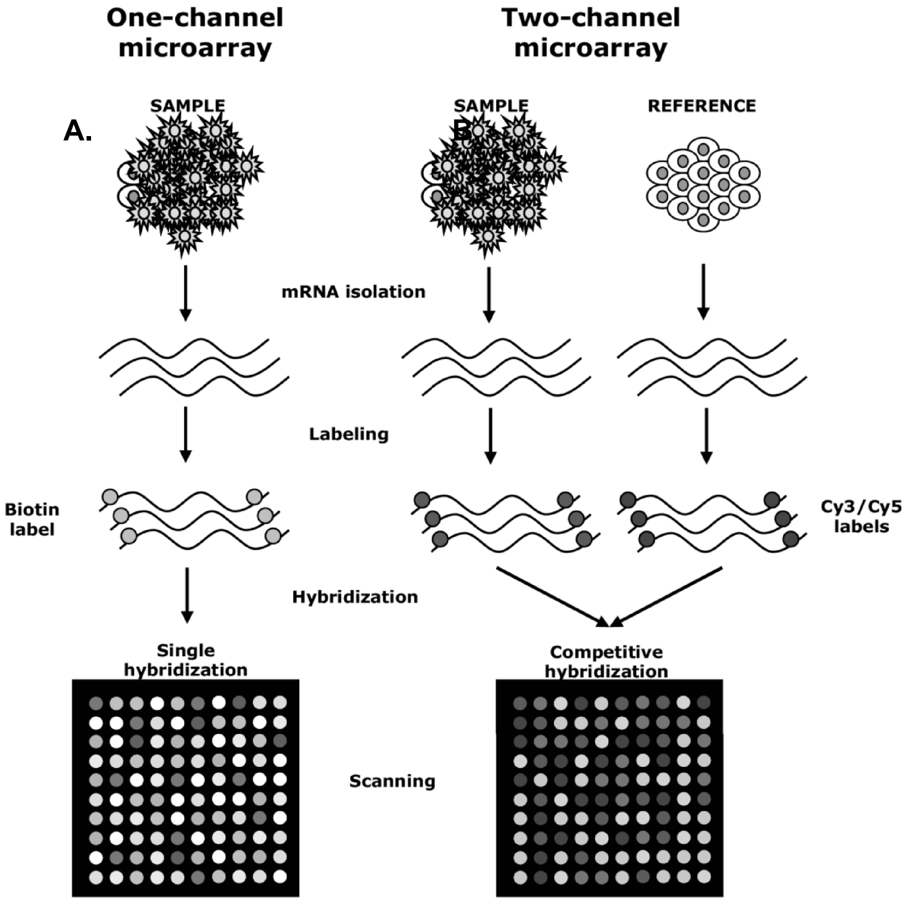


Figure 1.2. Schematics of experimental process using one-channel (A) and two-channel microarrays (B) (from Patterson *et al*, 2006)

dye and the labeled cDNAs are hybridized to each uniformative microarray chip, respectively (Patterson *et al*, 2006). For these reasons, the single-channel microarray system has the advantages that an aberrant sample cannot affect the raw data derived from other samples and that array data from different experiments are more easily compared to each other. However, one-channel microarray system requires twice more array chips compared to two-channel microarray system (Jaluria *et al*., 2007).

1.2. Applications of microarray

Microarrays are widely used to various biological researches, including gene expression profiling, chromatin immunoprecipitation on chip (ChIP-on-chip), detection of single nucleotide polymorphisms (SNPs), alternative spliced transcripts and fusion genes, and tiling array (Stoughton, 2005). The most well known use of microarrays is for profiling gene, especially messenger RNA (mRNA), expression levels (Shiu *et al*., 2008). To identify the differential expressed genes (DEGs) affecting in diseases or developmental stages of model animals, tens of thousands of DNA probes complementary to target DNAs are implemented on a array chip and the intensities of probes are measured simultaneously. Microarrays techonology have also been used to detect binding sites of DNA-protein interactions, called ChIP-on-chip. ChIP-on-chip technique which combines chromatin immunoprecipitation with microarray technology is generally used for identifying the binding sites of transcription factors or histone, and the methylated sequences of the genome (Zheng *et al*., 2007). To identify SNP among alleles within or between populations (McCarroll *et al*., 2008), potential splice sites of predicted exons for a gene (Milani *et al*., 2006), and fusion transcripts from cancer species (Jones *et al*., 2008), microarrays technology can be applied. Genomics tilling arrays consist of overlapping probes designed to densely represent a genomics region of interest or as large as entire region of human chromosome (Ishkanian *et al*., 2004). The purpose is to empirically detect expression of trasciprits or alternatively splice forms which may not have been previously known or predicted (Mockler *et al*., 2005).

1.3. Microarray data analysis

1.3.1. Standardization of microarray data

Since the data obtained from microarray experiment are very complicated and multi-dimensional structure, it is difficult to use the data among researchers. Therefore, biologists defined the guideline for microarrays, called the minimum information about a microarray experiment (MIAME), in 1999 to easily transfer and share the microarray data. The MIAME guidelines include a description of the following six sections (Brazma *et al*., 2001).

1. Experimental design: the set of hybridization experiments as a whole
2. Array design: each array used and each element (spot, feature) on the array
3. Samples: samples used, extract preparation and labeling
4. Hybridizations: procedures and parameters
5. Measurements: images, quantification and specifications
6. Normalization controls: types, values and specifications

For example, the information required to describe the measurement for a particular gene in a particular sample can be divided conceptually into three parts including gene annotation, sample annotation and a gene expression matrix (Figure 1.3).

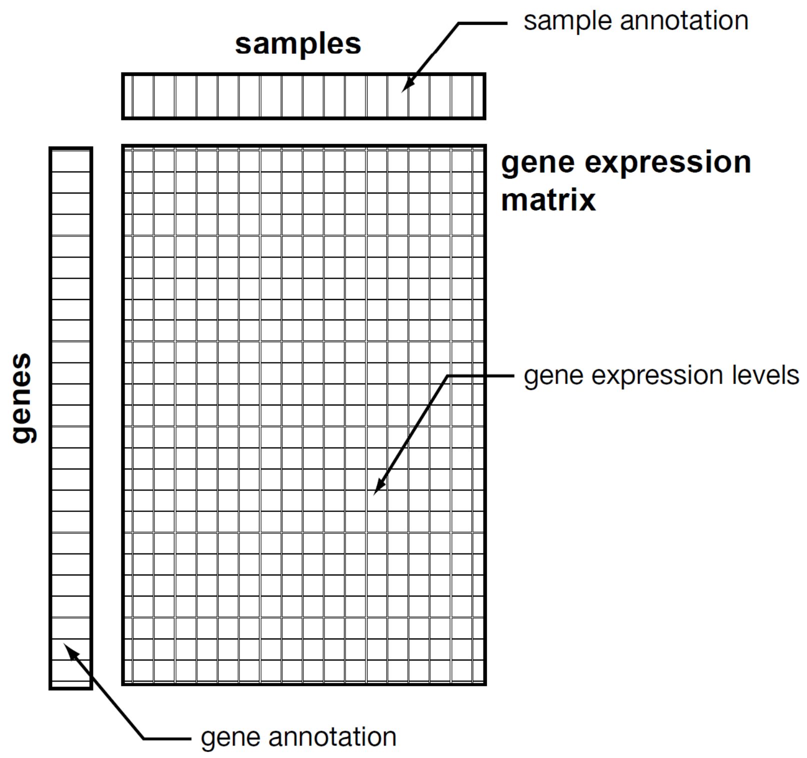


Figure 1.3. Conceptual view of gene expression data. The model has three parts: (i) gene annotation, which may be given as links to gene sequence databases, (ii) sample annotation, for which there currently are no public external databases (except the species taxonomy) and (iii) the gene expression matrix, in which each position contains information characterizing the expression of a particular gene in a particular sample (from Brazma *et al*., 2001).

1.3.2. Pre-processing of microarray data

Before microarray data analysis, several pre-processing steps are required. First, images are scanned to determine the signal intensity of each spot. This is usually done by image scanning software for a specific microarray platform such as GenePix (Fielden *et al.*, 2002), ImaGene (Medigue *et al.*, 1999), GeneSpring (Agilent Technologies, Palo Alto, CA), and many more (Bengtsson *et al.*, 2006). Second, normalization has to be performed to remove dye-related bias between two channels and various slide-specific artifacts that can exist between different microarrays. After pre-processing, the normalized microarray data can be further analyzed using various computational analysis methods such as clustering techniques and gene ontology analysis approaches.

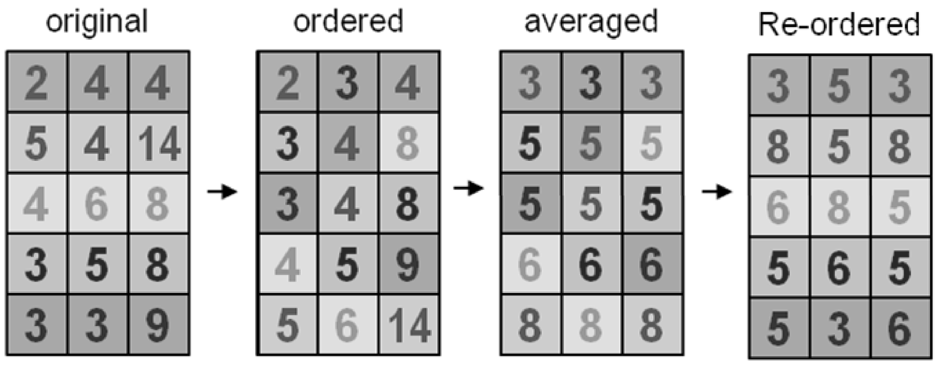
1.3.3. Normalization

When performing microarray experiments with multiple slides, non-biological variations are always exist between arrays such as dye biases, probes printing variations, and volume of initial RNA. To correct these variabilities, a series of normalization steps has to be performed with the scanned data before analysis. The main assumption behind normalization process is that most of the genes do not change their expression levels, and numbers of up- and down-regulated genes on each array are nearly equal. Overall average expression level of genes remains the same among different arrays. Thus, most normalization methods try to adjust expression levels of the genes by removing saturated signals from microarray or correcting the signal with background and probe spots.

Generally, when the signal intensities of genes is computed as the ratio between two hybridization signal, those of the over-expressed genes is greater than 1, and those of repressed genes is between 0 to 1. For these reasons, the signal intensities are transformed into the 2-based-logarithm to generate symmetrical distribution of the data and provide more interpretable comparisions between genes (Quackenbush *et al.*, 2002).

For one-channel microarrays such as Affymetrix arrays, robust multi-chip average (RMA) is used for background correction by calculating the average value of perfect match (PM) intensities (Irizarry *et al.*, 2003), and then quantile normalization is performed to adjust all probes intensities. Quantile normalization is based on the idea the the distribution of two array data is the same if the quantile-quantile associated plot is a straight diagonal line that we can represent by the unit vector (Balstad *et al.*, 2003). Quantile normalization process normalizes data by aligning ranked columns, computing their mean, and then replacing the the original data with the average quantiles (Figure 1.4).

In two-channel microarrays, many different methods have been developed to correct for dye-effects and other systematic errors between arrays. For example, global normalization (Quackenbush *et al.*, 2002) is the simplest methodology to adjust all log2-transformed measurements equal to zero. There are other methods that use similar global normalization approach, including log centering, rank invariant methods (Tseng *et al.*, 2001), and many other variations. However, these methods are inadequate in situations where systematic noise depends on overall intensity or spatial location within the array (Yang *et al.*, 2001). To account for these problems depends on the intensity and spatial location, lowess normalization methods based on robust



original

2	4	4
5	4	14
4	6	8
3	5	8
3	3	9

ordered

2	3	4
3	4	8
3	4	8
4	5	9
5	6	14

averaged

3	3	3
5	5	5
5	5	5
6	6	6
8	8	8

Re-ordered

3	5	3
8	5	8
6	8	5
5	6	5
5	3	6

Figure 1.4. Concept of quantile normalization

local weighted regression (Cleveland *et al*., 1988) have been proposed. This method adjusts the ratios through locally linear fits that depend on the intensity and has been proven to be a powerful normalization method for different types of two color microarray experiments (Berger *et al*., 2004). Although many new methods (Baird, *et al*., 2004; Chua *et al*., 2006; Wang *et al*., 2004; Yoon *et al*., 2004) and modifications (Wang *et al*., 2005) have been proposed, the comparison results are inconsistent and new methods outperform lowess normalization only in special cases (Fujita *et al*., 2006).

1.3.4. Statistical methods

Fold change (FC) was one of first methods used to identify differentially expressed genes (DEGs) between two different experimental conditions, such as samples or time points because of its simplicity. However, the method does not provide several questions; confidence intervals for results, and potential high false discovery rate (FDR) of results (Budhraja *et al*., 2003; Hsiao *et al*., 2004). For these reasons, more stringent statistical approaches have been suggested for microarray analysis. Among these, Student's t-test, analysis of variance (ANOVA), Mann-Whitney-Wilcoxon rank test, and significance analysis of microarray (SAM; Tusher *et al*., 2001) are most popular methods.

Student's t-test is used to test the hypothesis that a gene's expression levels differ between two sets of samples by using the T statistic and determining the significance level of the difference from *t* distribution (Gossett, 1908). This test is based on two assumption between two samples to be compared; the assumptions of normality and equal variance (Gossett, 1908).

ANOVA uses Fisher's F-distribution as part of the test of statistical significance and compares group variations to the overall variation observed (Kerr *et al*., 2000). One-way ANOVA method was most widely used for microarray data analysis. Recently, since more complicated experimental designs have been introduced with more than two independent variables, such as comparison with two samples according to several time points, two-way ANOVA method is also widely used (Churchill *et al*., 2004; Pavlidis *et al*., 2003).

Two-way ANOVA is an appropriate analysis method for a study with a quantitative outcome and two or more categorical variables. It requires three main assumptions of normality, equal variance, and independent samples (Zar, 1999). The two possible models for two-way ANOVA are the additive model and the interaction model. The additive model assumes that the effects on the outcome of a particular level change for one variable does not depend on the level of the other variable. On the other hand an interaction model assumes that the effects of a particular level change for one variable depend on the level of the other variable. For examples, suppose that microarray experiment was performed with two different groups over three time points to indentify the t_1 gene function, and t_1 gene can affect the expression level of t_2 gene. First, if the expression level and pattern of t_2 gene is significantly differ between groups over time, then those of t_2 gene is affected by the two variables, group and time, which means that the groups are changing over time (Time-by-group interactions; see in Figure 1.5A). Second, if the expression level of t_2 gene is differ between groups, but the expression pattern of t_2 gene in both groups is parallel without slope, then t_2 gene is affected by only variable group (Group effect; see in Figure 1.5B). Inversly, if the expression level of t_2 gene is superimposed and the expression pattern of t_2 gene is sloped, then t_2 gene is affected by only variable time (Time effect; see in Figure 1.5C). Finally, if the expression level of t_2 gene is differ between groups, and the expression pattern of t_2 gene in both groups is parallel and slope, then t_2 gene is affected by group and time (Group and time effect; see in Figure 1.5D) Overally, as shown in Figure 1.5, a plot drawn with parallel lines suggests an additive model, while non-parallel line suggests an interaction model.

SAM test is a statistical method for determining whether gene expression are significantly changed or not and uses slightly different statistic that is based on t-statistic. This analysis performs the correction that reduces the relative differences for low expressed genes and genes with similar expression levels (Tusher *et al.*, 2001). The input to SAM is gene expression measurements from a set of microarray experiments and a response variable, such as untreated and treated. SAM computes a statistic d_i for each gene i, measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine whether the expression of any genes is significantly related to the response. The cutoff for significance is determined by a parameter delta, chosen by the user based on the false positive rate (Tusher *et al.*, 2001; http://www-stat.stanford.edu/). SAM test is a non-parametric statistics and more useful for analyzing non-normal distributed data.

Mann-Whitney-Wilcoxon rank test is also a non-parametric statistical test compares for each gene the difference between measurements in two groups. However, it does not require assumptions about the form of the distributions of the measurements, so it is more reliable when used on microarray data with large number of outliers or high noise. The method's statistical power strictly depends on sample

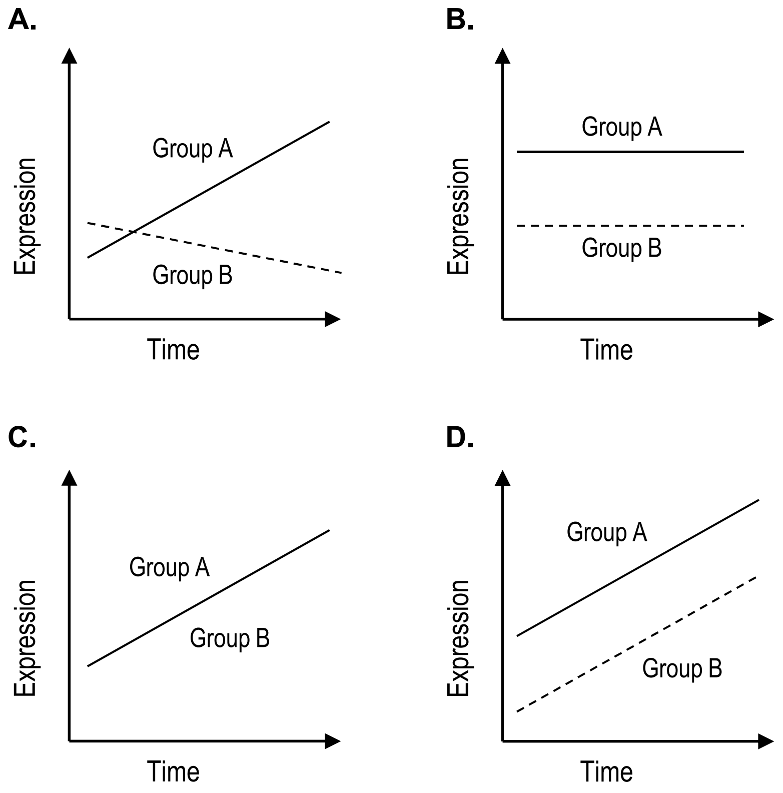


Figure 1.5. Models of two-way ANOVA for identifying DEG. Panel A shows that each element included group A and B has different expression patterns over time (Time-by-group interaction). Panel B represents only groups differences between two groups (group effect). Panel C represents superimposed patterns between groups (Time effect). Panel D shows that the elements of two groups are changed of their expression over time (Time and group effect)

sizes, and provides poor significance levels for groups with fewer than 6 samples (Ahmad, 1996).

1.4. Cluster analysis

Cells have adapted in different environmental conditions by responding a set of proteins required to use or oppose these conditions. To optimize the process, genes whose products function together are usually undergoing same regulatory mechanisms so they are coordinately expressed in response to stimuli. This property is used by many clustering methods that group genes together based on their expression profiles. In this case microarray data for analysis with clustering methods can be represented by a matrix with measurements of genes (rows) for multiple conditions (columns), where conditions can be of various kinds of samples, such as, different treatments, time points, and patients. Clustering algorithms can be divided into two categories: hierarchical clustering (Eisen *et al.*, 1998) and non-hierarchical clustering such as *k*-means clustering (Tibshirani *et al.*, 1999) and self-organizing maps (SOM; Tamayo *et al.*, 1999).

1.4.1. Hierarchycal cluster

Hierarchical clustering is one of the first clustering algorithms applied to microarray data (Eisen *et al.*, 1998; Alon *et al.*, 1999; Cho *et al.*, 1998). This method produces a classification by agglomeration in which small clusters of very similar molecules are agglomerated within larger cluster of less similar molecules. Using a distance metric, the method builds a hierarchical binary tree, called a dendrogram, starting from the individual gene expression profiles as leaves by progressively merging clusters, where each internal node represents the average of its two children (Eisen *et al*., 1998). There are several hierarchical clustering methods according to measure distances between internal nodes, including single linkage, complete linkage, centroid linkage, median linkage and average linkage (Figure 1.6). The constructed tree can be cut at some point according to a threshold value to receive clusters of required characteristics. Due to its simplicity and clear representation, hierarchical clustering has been used in many reported microarray experiments, but a number of drawbacks should be considered. First, hierarchical clustering is a greedy search algorithm, meaning that merging decisions on early steps are based only on the distance between nodes and cannot be undone, but not necessarily the best ones in global scale and can lead to mistakes in the overall clustering. Second, dendrograms and corresponding heatmaps, which used extensively in visualizations of the analysis results, suffer from inversion problems that complicate interpretation of the hierarchy (Morgan *et al*., 1995). In addition, complexity of dendrograms for larger data sets makes them difficult to understand, and the choice of location for tree cut to receive final clusters is unclear. And finally, analysis of yeast cell-cycle dataset with hierarchical clustering performed by Cherepinsky *et al*. showed that the method has very low accuracy of gene assignments to clusters (Cherepinsky *et al*., 2003).

1.4.2.Non-hierarchycal cluster

A non-hierarchical clustering generates a classification by partitioning a dataset, giving a set of non-overlapping groups having no hierarchical relationships between them (Downs and Barnard, 1995). These methods are to make a unique

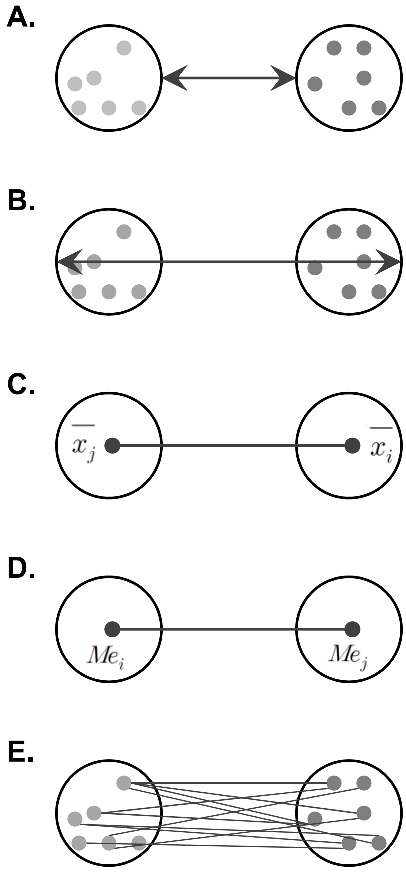


Figure 1.6. Graphical examples for inter-clusters distances Single linkage (A), complete linkage (B), centroid linkage (C), median linkage (D), and average linkage (E) (modified from *http://www.multid.se/genex/hs515.htm*)

partition of *k* groups, and optimal *k* value is determined by heuristic or by repeated clustering result. The most well known approaches of non-hierarchical clustering are *k*-means algorithm (Tavazoie *et al*., 1999) and self-organizing maps algorithm (Kohonen *et al*., 1997).

1.4.2.1. k-means clustering

The *k*-means clustering starts from randomly dividing genes into *k* groups and calculating cluster centers for each of these groups. New groups are formed by reassigning each gene to the closest centroid. Then the centroids are recalculated for the new clusters and the process repeats (Tavazoie *et al*., 1999). While simplicity and speed of the method are the main advantages of the method, the most important disadvantage is that results are not unique across different runs and depend on starting positions of centroids and the initial number of *k*. Another obstacle for analysis of microarray data is that this method finds sub-optimal clusters rather than optimal clusters from all data because of the limitation of *k*. Moreover, as it does not inform about the hierarchical structures among the data in each cluster, the similarity among the clusters are unknown (Heyer *et al*., 1999).

1.4.2.2. Self-organizing maps (SOM) clustering

SOM which is proposed by Kohonen in 1997 is a type of artificial neural netork that use unsupervised learning to produce a lower-dimensional, usually two-dimensional, representation of the input space of the traning data set samples (Kohonen *et al*., 1997). It performs the following procedure. After choosing an initial grid of nodes the nodes are mapped randomly into *k*-dimensional space. At each step of the algorithm a random data point is chosen and nodes are moved in the direction of it. Nodes are moved depending on distance between a node and that data point, so the closest node is moved the most compared to more distant nodes (Figure 1.7).

Although a number of successful application of SOMs have been reported (Tamayo *et al*., 1999; Baker *et al*., 2001; Toronen *et al*., 1999), several disadvantages exist for the method. Sensitivity of SOM to incomplete data is a problem that is very important in microarray data analysis, due to abundance of missing data points resulting from data flagging during preprocessing (Tamayo *et al*., 1999). Also, a SOM can yield different decompositions of the data depending on the choice of initial conditions. Another issue is that initial node grid is fixed and may not be changed during the analysis, and that can lead to inappropriate mapping of the data space. In addition, when two data points are mapped from high dimensional space to nearby locations on the two dimensional grid, it is possible that those points are actually far apart in the higher dimensional space (Baker *et al*., 2001; Toronen *et al*., 1999). However, this method is widely used to analysis of microarray data which are obtained from a series of time point experiments because of the availability of their multi-dimensional clustering. For examples, if microarray experiments are performed with two samples in three time points, the grid configuration for SOM clustering can be set by 3x3 or 3x2 because the expression of gene can be measured to be increased, decreased or unchanged in early and late time point, respectively.

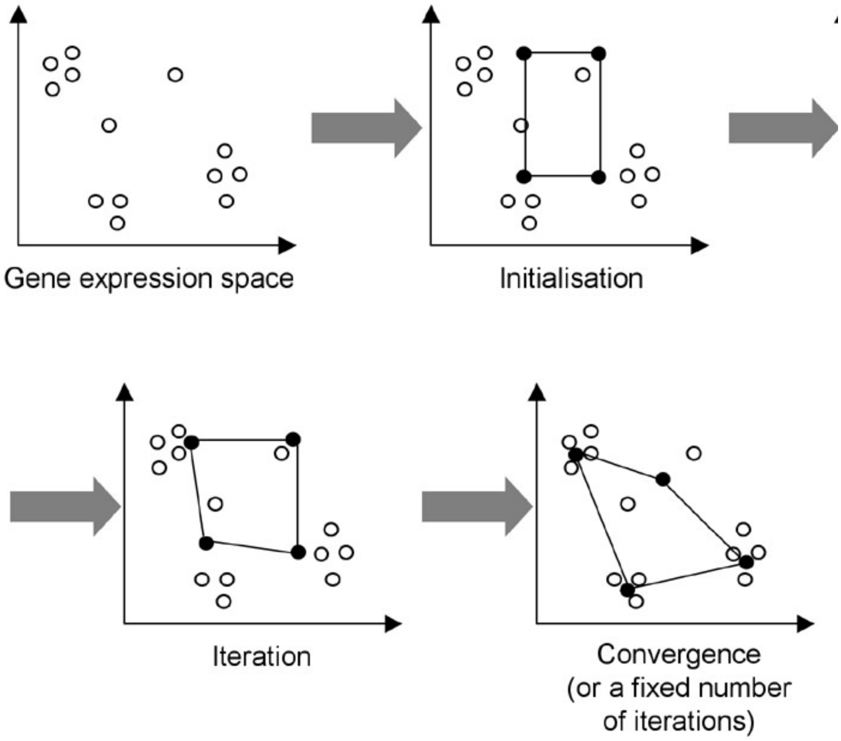


Figure 1.7. The principle behind SOMs. During the initialization step, a grid of nodes is projected onto the expression space and each gene is assigned its closest node. Following this step, one gene is chosen at random and the assigned node is 'moved' towards it. The other nodes are moved towards this gene depending on how close they are to the selected gene. This step is performed iteratively until convergence or is performed for a fixed number of iterations to get a final map of nodes (from http://www.mrc-lmb.cam.ac.uk/genomes/madanm/ microarray/)

1.5. Gene Ontology(GO) analysis

1.5.1. Foundation of GO

Biologists waste a lot of time and effort in searching for all of the available information about each small area of research because of the wide variations in terminology that may be common usage. For these reasons, the GO Consortium was founded to unify the biological terminology and the GO project was launched Ashburner *et al*., 2000). The GO project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The project began in 1998 as collaboration between three model organism databases; FlyBase (*Drosophila*; McQuilton *et al*., 2012), the *Saccharomyces* Genome Database (SGD; Cherry *et al*., 1998) and the Mouse Genome Database (MGD; Blake *et al*., 2003). Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes.

1.5.2. Three categories of GO

The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

Biological process refers to series of events to which the gene contributes. A process is accomplished via one or more ordered assemblies of molecular functions. They often involve a physical or chemical transformation, meaning that something goes into a process and some other thing comes out of it. ‘Cell growth and maintenance’ or ‘signal transduction’ are some examples of high-level or broad biological process terms; whereas ‘translation’ or ‘pyrimidine metabolism’ are a couple of low-level or more specific biological process terms.

Cellular component refers to the place in the cell where a gene product is active. Its ontology describes locations, at the levels of sub cellular structures and macromolecular complexes. Cellular component includes terms, such as ‘ribosome’ or ‘proteasome’ specifying where multiple gene products would be found. It also includes terms such as ‘nuclear membrane’ or ‘Golgi apparatus’.

Molecular function refers to the biochemical activity of a gene product and also the capability that a gene product carries as a potential. These may include transporting somethings, binding to somethings, holding with somethings and changing one into another. Examples of broad functional terms are ‘enzyme’, ‘transporter’ or ‘ligand’. Examples of narrower functional terms are ‘adenylate cyclase’ or ‘toll receptor ligand’.

The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are arcs between the nodes (Figure 1.8). The relationships used in GO are directed and the graph is acyclic, meaning that cycles are not allowed in the graph. The ontologies resemble a hierarchy, as child terms are more specialized and parent terms are less specialized, but unlike a hierarchy, a term may have more than one parent term.

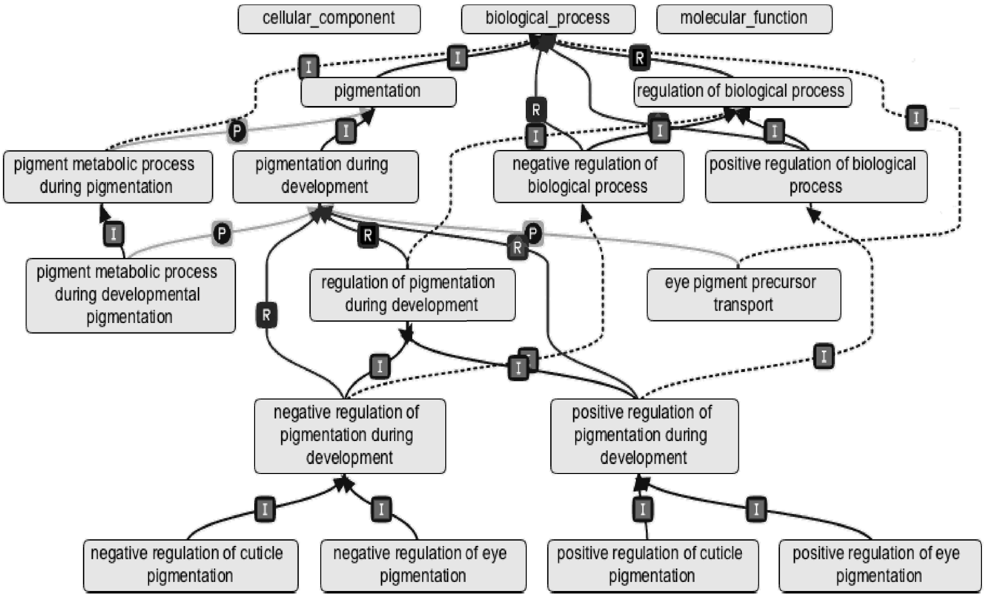

cellular_component
biological_process
molecular_function
pigmentation
regulation of biological process
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
negative regulation of pigmentation during development
positive regulation of pigmentation during development
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation
I
R
P

Figure 1.8. Example of the structure and the relationships between the GO nodes

(from http://www.*geneontology.org/GO.ontology.structure.shtml*).

1.5.3. Application of GO for functional genomics

Recently, GO analysis is an important step for identifying the function of the specific genes or gene products because massive data for gene expression obtained from microarray or high throughput sequencing experiment are accumulated. An approach making use of the GO for analyzing microarray experiments is that of Pavlidis *et al* in 2002. They suggested the method for scoring differential gene expression in groups of functionally related genes. The scoring method they proposed reflects the statistical significance of the expression pattern of each gene. Firstly, they calculate p-value for each gene by applying the analysis of variance (ANOVA) method on the gene specific expression values over the samples. And then, they calculate the experiment score by adding the negative logarithms of the ANOVA-p-values of genes in specific GO group (Pavlidis *et al.*, 2002). A second approach using the GO to analyze large-scale gene expression is to search for over-representation of particular GO categories from a list of genes. The over-represented GO category, or the most frequent GO category, is assigned as a major biological property in a gene cluster (Al-Shahrour *et al.*, 2004). The hypergeometric test is generally implemeted as a statistical model for calculating over-represented GO categories in a gene cluster. The hypergeometric testing is a discrete probability distribution that describes the probability of *i* successes in *n* draws from a finite population of size *N* without replacement (Chvatal, 1979). In analyzing gene expression data, *i* successes indicate the number of genes showing that significantly different pattern of their expression. Then the hypergeometric probability is:

$$p(X=i)=\frac{\binom{N-k}{n-i}\binom{k}{i}}{\binom{N}{n}}$$

Notation is as follow: *N* is the number of total genes in a species, *n* is the number of genes identified as DEGs, *k* is the number of total genes included in a particular GO category in a species, and *i* is the number of DEGs included in a particular GO category

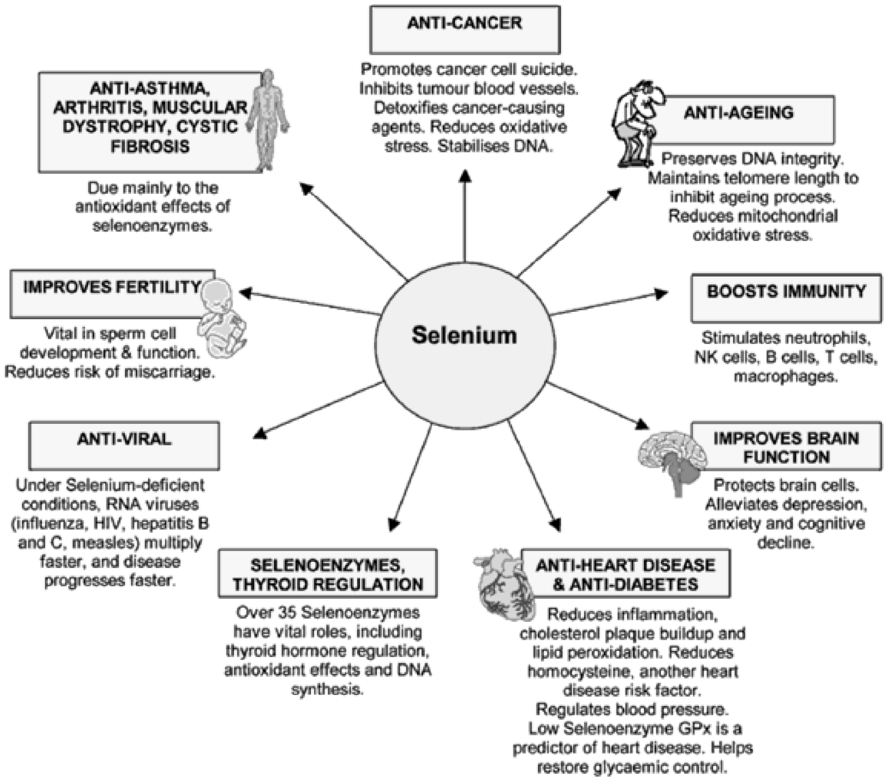
2. SELENIUM BIOLOGY

2.1. Selenium and human health

Selenium is an essential trance element found in soil, and vegetables are the major dietary sources of selenium. The content of selenium in food depends on the selenium content of the soil where plants are grown or animals are raised. Selenium also can be found in some meats and seafood. Animals that eat grains or plants that were grown in selenium-rich soil have higher levels of selenium in their muscle.

Selenium plays important roles for human health (Lee *et al*., 1996). Moderate selenium insufficiency may contribute to the pathogenicity of viral infection, the progression of AIDS in HIV positive patients, male infertility, and impaired immunity (Hatfield, Berry, and Gladyshev, 2006 and references therein; see Figure 1.9). Moreover, its deficiency has been implicated as a factor in Keshan disease, a cardiomyopathy that affects young women and children in certain regions of China that have selenium-poor soil. Since dietary selenium reduces the risk of certain types of cancers (Clark *et al*., 1996), selenium has also sparked a lot of interest as an anticancer nutrient.

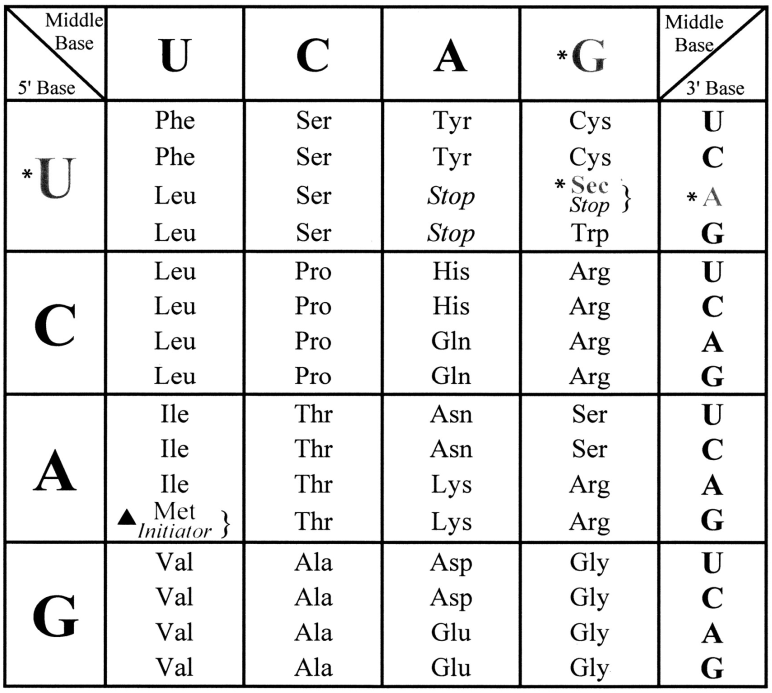
Selenium has also attracted the attention of molecular biologists because it is co-translationally incorporated into protein as the amino acid selenocysteine (Sec), the 21st amino acid (Lee *et al*., 1989; Longtin, 2004). Sec is encoded by a UGA codon in selenoprotein mRNA (Hatfield and Gladyshev, 2002; Birringer *et al*., 2002; Driscoll and Copeland, 2003; see Figure 1.10). Sec is widely used in all major domains of life and is responsible for the majority of biological effects of selenium. There are 25 known selenoproteins in humans and 24 in rodents (Kryukov *et al*., 2003). Therefore,



Selenium
ANTI-CANCER
Promotes cancer cell suicide. Inhibits tumour blood vessels. Detoxifies cancer-causing agents. Reduces oxidative stress. Stabilises DNA.
ANTI-ASTHMA, ARTHRITIS, MUSCULAR DYSTROPHY, CYSTIC FIBROSIS
Due mainly to the antioxidant effects of selenoenzymes.
ANTI-AGEING
Preserves DNA integrity. Maintains telomere length to inhibit ageing process. Reduces mitochondrial oxidative stress.
IMPROVES FERTILITY
Vital in sperm cell development & function. Reduces risk of miscarriage.
BOOSTS IMMUNITY
Stimulates neutrophils, NK cells, B cells, T cells, macrophages.
ANTI-VIRAL
Under Selenium-deficient conditions, RNA viruses (influenza, HIV, hepatitis B and C, measles) multiply faster, and disease progresses faster.
IMPROVES BRAIN FUNCTION
Protects brain cells. Alleviates depression, anxiety and cognitive decline.
SELENOENZYMES, THYROID REGULATION
Over 35 Selenoenzymes have vital roles, including thyroid hormone regulation, antioxidant effects and DNA synthesis.
ANTI-HEART DISEASE & ANTI-DIABETES
Reduces inflammation, cholesterol plaque buildup and lipid peroxidation. Reduces homocysteine, another heart disease risk factor. Regulates blood pressure. Low Selenoenzyme GPx is a predictor of heart disease. Helps restore glycaemic control.

Figure 1.9. The health benefits of selenium

(from http://1phil4everyill.wordpress.com/ 2008/11/20/dr-russell-blaylock-nutrition-and-behavior/)



Middle Base 5' Base	U	C	A	*G	Middle Base 3' Base
*U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	*Stop*	* Sec *Stop* }	*A
	Leu	Ser	*Stop*	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	▲Met *Initiator* }	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure. 1.10. The genetic code showing that Sec is the 21st amino acid and that Sec is coded by UGA. *, UGA and Sec; ▲, AUG, the other codon in the genetic code that serves a dual function (from Hatfield and Gladyshev, 2002)

it is very important to understanding how selenium is incorporated into protein and the biological functions of the selenoproteins.

2.2. Selenium metabolism

Selenium metabolism from selenium supplements indicates differences in the absorption and use of selenium between inorganic and organic forms in humans (Butler *et al.*, 1999; Brown *et al.*, 2000) and rats (Finley *et al.*, 2001). The absorption pathways have not yet been fully characterized; however, selenium as selenate or selenite appears to be very well absorbed but less well retained in the body than organic forms of selenium, such as selenomethionine and Sec (Schrauzer *et al.*, 2000). The proposed metabolic pathways for different forms of selenium are shown in Figure 1.11 (Fairweather-Tait *et al.*, 2010). Selenomethionine, Sec, selenate, and selenite enter the selenide pool and the selenium is either used for selenoprotein synthesis or excreted in the urine as a selenosugar. Selenomethionine can be incorporated directly into proteins through the replacement of methionine; however, optimal and suboptimal levels in diet, selenium mainly transformed into Sec in animals, and then Sec is incorporated into selenoproteins. Therefore, deficiency in dietary selenium results in decreased levels of selenoproteins and biological processes that are maintained by selenoproteins are compromised.

2.3. Selenoproteins biosynthesis

Most of the characterized selenoproteins are enzymes involved in oxidation reduction reactions and contain Sec in their active site (Stadtman, 2000). Sec is

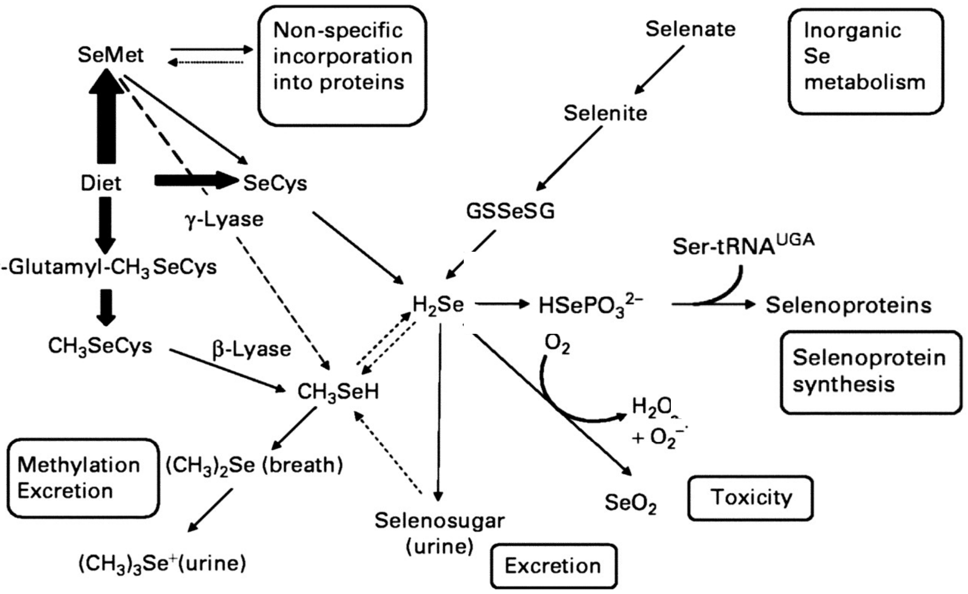
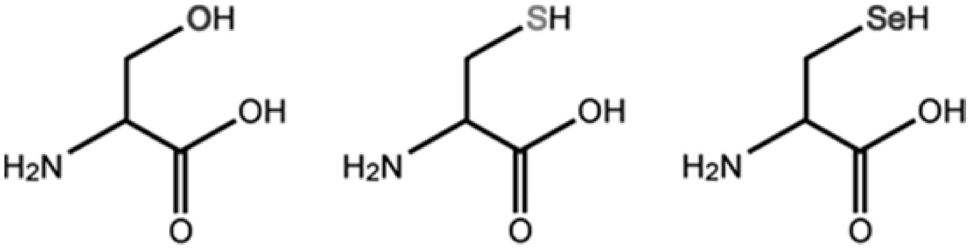


Figure 1.11. Metabolic pathway of dietary selenium in humans. Se, selenium; SeMet, selenomethionine; SeCys, selenocysteine; GSSeSG, selenodiglutathione; γ-glutamyl-CH_3SeCys, γ-glutamyl-Se-methylseleno cysteine; H_2Se, hydrogen selenide; $HSePO_3^{2-}$, selenophosphate; CH_3SeCys, Se-methylselenocysteine; CH_3SeH, methylselenol; $(CH_3)_2Se$, dimethyl selenide; SeO_2, selenium dioxide; $(CH_3)_3Se^+$, trimethyl selenonium ion (from Fairweather-Tait *et al.*, 2010).

structurally very similar to serine (Ser) and cytosine (Cys), except that it contains selenium instead of oxygen and sulfur, respectively (Figure 1.12). Since the selenol group is more fully ionized than the thiol group at physiological pH (Stadtman, 1996), the catalytic activity of a selenoprotein is strongly decreased after Sec is replaced with Cys (Axley *et al*., 1991). Although there are no known selenoproteins in higher plants, or in yeast, the majority of organisms have selenoproteins.

The selenoproteins play crucial roles in a variety of biological processes, and several of them are involved in antioxidant defense. Glutathione peroxidases (GPx) protect cells against peroxidative damage by reducing hydrogen peroxide and free fatty acid hydroperoxides (Flohe and Brigelius-Flohe, 2001). Another GPx family member, phospholipid hydroperoxide glutathione peroxidase (PHGPx), reduces phospholipid, cholesterol, and cholesteryl ester hydroperoxides, thereby protecting cells against membrane lipid peroxidation (Flohe and Brigelius-Flohe, 2001). PHGPx also plays a structural role in the mitochondrial capsule of mature spermatozoa where the protein becomes oxidatively cross-linked and inactive (Ursini *et al*., 1999). This noncatalytic function of PHGPx may also be involved in the male infertility similar to selenium deficiency (Flohe, 2001). In mammals, three distinct mammalian thioredoxin reductases function in cellular redox homeostasis by reducing thioredoxin and other substrates (Holmgren, 2001). Other oxido-reductases that contain Sec include the family of deiodinases, which are involved in thyroid hormone metabolism (Germain, 2001), and selenophosphate synthetase 2 (SPS2), which synthesizes the selenium donor for Sec biosynthesis. This enzyme is unique in that it is the only selenoprotein expressed in both prokaryotes and eukaryotes (Guimaraes *et al*., 1996).

There are several selenoproteins, including SelW, which is expressed in



Serine (Ser) Cysteine (Cys) Selenocysteine (Sec)

Figure 1.12. Comparison of selenocysteine (Sec) to the structurally similar amino acids serine (Ser) and cysteine (Cys).
(from http://www.riken.go.jp/engn/r-world/info/release/press/2010/100813 /index .html)

cardiac and skeletal muscle (Whanger *et al*., 2002), and Sep15, which is implicated in preventing prostate cancer (Gladyshev, Diamond and Hatfield, 2001). Novel selenoprotein genes have been identified in the human genome using bioinformatic approaches, but the functions of the encoded proteins are largely unknown. One of these selenoproteins, SelR, was shown to be a methionine sulfoxide reductase (Kryukov *et al*., 2002). An antioxidant function has also been proposed for the plasma protein, selenoprotein P (SelP). The SelP mRNA encodes 10 to 17 UGA codons, depending on the species (Burk and Hill, 2005). Understanding how multiple UGA codons are decoded as Sec in SelP would provide important insight into the mechanism of selenoprotein synthesis.

2.3.1. Mechanism of selenocysteine biosynthesis

Selenocysteine is known as the 21st amino acid in protein synthesis (Lee *et al*., 1989; Longtin, 2004) and its specific incorporation is directed by the UGA codon. Because UGA codon can recognized as a Sec codon as well as a stop codon, the cellular mechanisms that distinguish between these two functions exist. Unique tRNA, called Sec-tRNA$^{[Ser]Sec}$, that have complementary UGA anticodons, and several factors have been identified for being required in the recognition of UGA as Sec codon.

The mechanisms of Sec biosynthesis are slightly different between prokaryotes and eukaryotes. In prokaryotes, Sec is directly synthesized by replacing the hydroxyl moiety of serine into selenium moiety on serine charged tRNA$^{[Ser]Sec}$ (Forchhammer and Böck, 1991). In eukaryotes, one more enzyme is involved in Sec synthesis. tRNA$^{[Ser]Sec}$ is first aminoacylated with serine by Seryl-tRNA$^{[Ser]Sec}$ synthetase (Figure 1.13). The seryl moiety of Seryl-tRNA$^{[Ser]Sec}$ is then phosphorylated

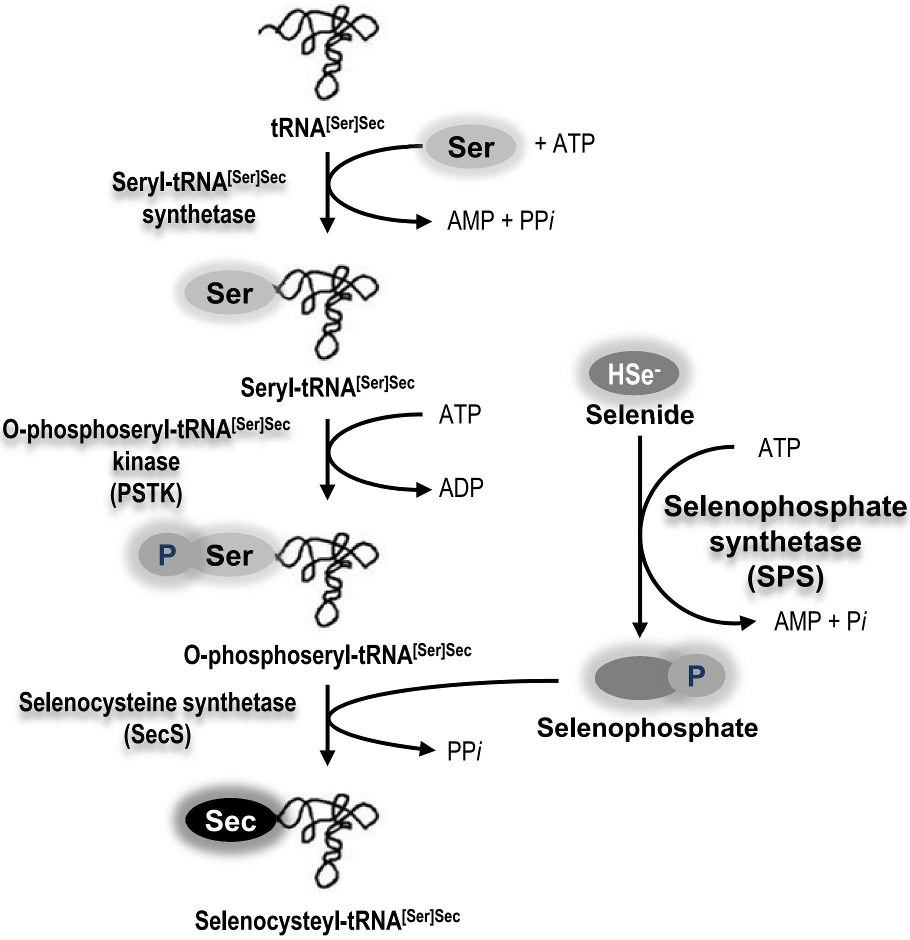

tRNA[Ser]Sec
Ser
+ ATP
Seryl-tRNA[Ser]Sec synthetase
AMP + PPi
Ser
Seryl-tRNA[Ser]Sec
HSe-
Selenide
O-phosphoseryl-tRNA[Ser]Sec kinase (PSTK)
ATP
ADP
ATP
Selenophosphate synthetase (SPS)
P
Ser
O-phosphoseryl-tRNA[Ser]Sec
AMP + Pi
P
Selenophosphate
Selenocysteine synthetase (SecS)
PPi
Sec
Selenocysteyl-tRNA[Ser]Sec

Figure 1.13. Biosynthesis of Sec in eukaryotes.

by O-phosphoseryl-tRNA$^{[Ser]Sec}$ kinase (PSTK; Carlson *et al*., 2004) to yield Phosphoseryl-tRNA$^{[Ser]Sec}$, which is converted to Sec-tRNA$^{[Ser]Sec}$ by selenocyteine synthase (SecS) (Yuan *et al*., 2006; Xu *et al*., 2007). Sec-tRNA$^{[Ser]Sec}$ is used on the ribosome to insert Sec into a specific site in a nascent polypeptide of selenoproteins. The active donor of selenium in Sec biosynthesis is monoselenophosphate (Glass *et al*., 1993), which is synthesized from selenite and ATP by an enzyme designated as selenophosphate synthetase (SPS; Ehrenreich *et al*., 1992).

2.3.2. Incorporation of selenocysteine into protein

One novel feature of selenoprotein mRNAs is the occurrence of a *cis*-stem-loop structure known as the SEC Insertion Sequence (SECIS) element or elements in the 3'-untranslated region (3'-UTR) of selenoprotein mRNAs (Low and Berry, 1996). SECIS elements are responsible for recoding the UGA codon as Sec and bypassing stop. In addition to these two *cis*-acting factors, there are several *trans*-acting factors involved in the insertion of this amino acid into protein: selenocystenyl-tRNA$^{[Ser]Sec}$-specific elongation factor, EFsec (Tujebajva *et al*., 2000; Fagegaltier *et al*., 2000); SECIS-binding protein, SBP2 (Copland *et al*., 2000); and L30 ribosomal protein, rpL30 (Chavatte *et al*., 2005).

Mechanism of selenocysteine incorporation into protein is summarized in Figure 1.14 (Hatfield and Gladyshev, 2002). After seryl-tRNA$^{[Ser]Sec}$ residue is converted to a selenocysteine-residue by the PLP-dependent enzyme selenocysteine synthase (SecS), selenocysteyl-tRNA$^{[Ser]Sec}$ is incorporated into ribosomal A-site mediated by a specific translational elongation factor, eEFsec, and SECIS binding protein 2 (SBP2). Once the selenocysteyl-tRNA$^{[Ser]Sec}$ is incorporated to the A-site,

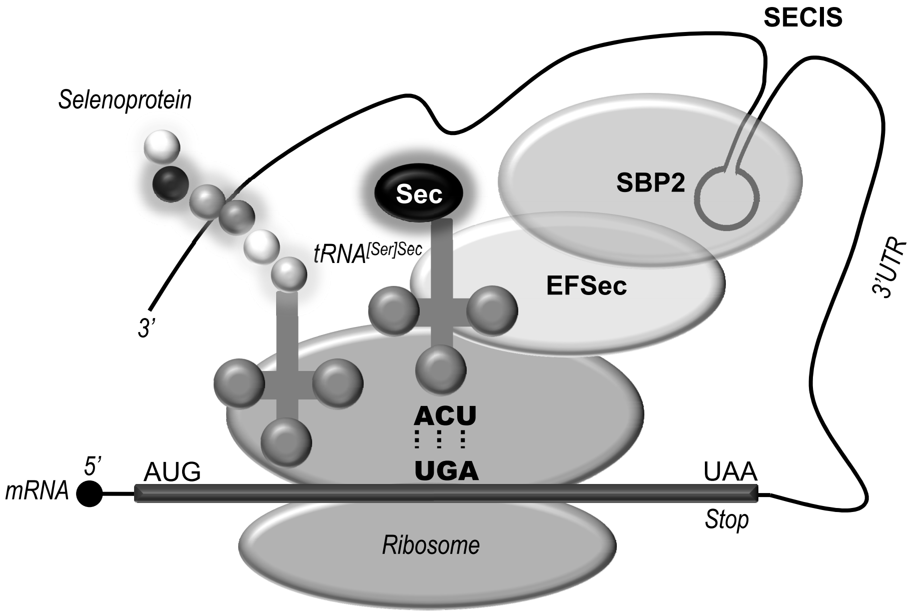


Figure 1.14. Mechanism of Sec incorporation at UGA codon. Co-translational incorporation of the 21st proteinogenic amino acid Sec into proteins occurs at the UGA codon, which recruits Sec-loaded tRNA$^{[Ser]Sec}$ (SelC) to the ribosome via an interaction of the Sec-specific translation factor EFSec with the SECIS binding protein 2 (SBP2). SBP2 recognizes the 3′-UTR hairpin loop SECIS mRNA structure found in all mRNAs encoding Sec-containing proteins (modified from Kohrle *et al*., 2005).

selenocysteyl-tRNA$^{[Ser]Sec}$ is transferred to the peptidyl site and Sec is incorporated into the nascent selenopeptide.

2.3.3. Components of selenocysteine biosynthesis

The SECIS element is a stem-loop and *cis*-acting RNA structure around 60 nucleotides in length on the 3' UTR of selenoprotein mRNA transcripts. This structural motif serves as the signaling factors for translating UGA stop codon as Sec (Low and Berry, 1996). Thus, SECIS elements are a fundamental aspect of mRNA encoding selenoproteins. Eukaryotic SECIS elements are composed of two helices separated by an internal loop, a SECIS core structure consist of a Quartet located at the base of helix 2, and an apical loop or bulge (Figure 1.15). The Quartet is specific base pairs composed of four non-Watson-Crick types. It appears to be the main functional site of the stem-loop structure when binding to SBP2 and the L30 protein. There is a spatial requirement regarding the distance between in-frame Sec UGA codon and 3'UTR SECIS elements. The minimal distance was measured to be between 51 and 111 nucleotides (Low and Berry, 1996), suggesting that SECIS elements are both necessary and sufficient for Sec insertion. In bioinformatics, several computer programs have been created that search for SECIS elements within a genome sequence, based on the sequence and secondary structure characteristics of SECIS elements (Chapple *et al*., 2009).

Elongation factor for selenoprotein translation (eEFsec), which is also called eSelB, is necessary translation factor for the incorporation of Sec into nascent polypeptides in response to UGA codon by recruiting selenocystenyl-tRNA and acting with SBP2 (Tujebajeva *et al*., 2000; Fagegaltier *et al*., 2000). eEFsec is specific for

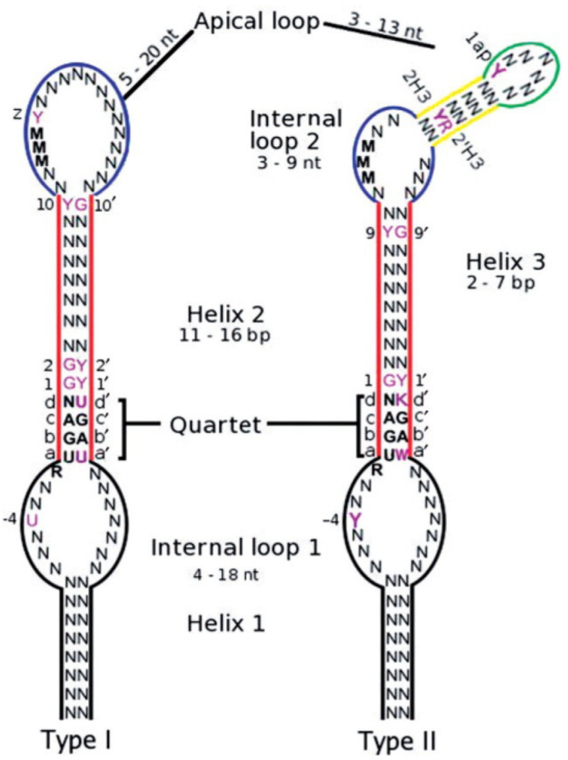


Figure 1.15. Eukaryotic SECIS elements. Novel conserved residues are shown in magenta. Where a specific nucleotide is shown, it was observed in that position in 50% or more of the aligned sequences. Where a class of nucleotides is shown, that class was observed in that position in 70% or more of the aligned sequences. Y=U or C, K=G or U, N=any nucleotide, W=A or U, R=A or G, M=A or C. Quartet: four consecutive non-Watson–Crick base pairs. Base pairs forming the quartet were called abcd/a′b′c′d′ for the sake of clarity in the text. Position ‘z’ is the first nucleotide after the run of Ms, positions 2H3/2′H3 are the second base pair of Helix 3 and 1ap the first nucleotide of the apical loop. The range of possible lengths for helix 1 is hard to determine because it depends on the local 2D structure of the mRNA 3′UTR. There are two types of SECIS elements in eukaryotes with type II being the most common (from Chapple *et al*., 2009).

Sec and does not bind seryl-tRNA$^{[Ser]Sec}$ (Tujebajeva *et al*., 2000) or phosphoseryl-tRNA$^{[Ser]Sec}$ (Carlson *et al*., 2004). eEFsec forms a complex with SBP2 and this complex formation is stimulated by the presence of selenocystenyl-tRNA$^{[Ser]Sec}$.

SBP2, is factor appears to have essential functions that include binding to the SECIS core and the ribosome, and the insertion of Sec into selenoprotein. SBP2 binds the Quartet in SECIS element to form SBP2-SECIS complex (Driscoll *et al*., 2003; Copeland, 2003). SBP2 also binds to the ribosome at the 28S RNA by binding with one or more kink turn structures (Kinzy *et al*., 2005). rpL30 also enhances the incorporation of Sec into protein (Chavatte *et al*., 2005). A model to accommodate the participation of both components in binding to the SECIS element has been proposed wherein SBP2 and rpL30 carry out different functions in the recoding of UGA and the SECIS element acts as a molecular switch upon protein binding (Chavatte *et al*., 2005). SBP2 is continually bound to the ribosome except at the time Sec is delivered to the A-site for decoding (Caban *et al*., 2006; Kinzy *et al*., 2005). Thus, SBP2 can serve to select a subset of ribosome and program them for Sec insertion competence by interacting with the SECIS element at the moment of Sec insertion. The role of rpL30 would then be to compete SBP2 off the SECIS element and back onto the ribosome, rendering the ribosome competent for another round of recoding (Caban *et al*., 2006)

2.4. Selenophosphate synthetase (SPS)

Selenophosphate synthetase (SPS) is the enzyme that catalyzes the formation of monoselenophosphate, which is the active selenium donor, from selenide and ATP (Glass *et al*., 1993). In lower eukaryotes and bacteria, there is only one type of SPS encoded by *SelD* gene; however, there are two isoforms of SPS in higher eukaryotes, SPS1 and SPS2 (Leinfelder *et al*., 1990; Guimaraes *et al*., 1996).

2.4.1. Structural characteristic of eukaryotic SPS

Although there are two types of SPS, or SPS1 and SPS2, in higher eukaryotes, the DNA sequences homology between SPS1 and SPS2 is different each other. Whereas the amino acid sequence homology has high similarity between them. For examples, the amino acid sequence homology between *Drosophila* SPS1 and SPS2 is approximately 45%, and that between human SPS1 and SPS2 is 72%. There are two well conserved motifs among two isoforms (Figure 1.16). As shown in Figure 1.16, motif A is similar to the consensus sequence for ATP binding site, and it contains well conserved Lys 20 residues. Biochemical analysis with E.coli SelD mutants shows that selenide-dependent formation fo AMP from ATP was dramatically reduced by replacement Cys 17 and Lys 20 to others (Kim *et al*., 1992; 1993). These results indicate that Cys 17 and Lys 20 positions are essential active sites for the formation of selenophosphate. However, Cys 17 residue is substituted to threonine (Thr) and arginine (Arg) in human (also mouse) and *Drosophila* SPS1, respectively. In contrast, SPS2 is a selenoprotein with Sec (U) substituted at the same site (Guimaraes *et al*., 1996). Motif B is a conserved glycine rich region similar to the conservd ATP/GTP binding consensus sequences, GXXXXGK(S/T) or GXGXXG found in many ATP/GTP binding proteins or protein kinases, respectively (Low *et al*., 1995; Guimaraes *et al*., 1996).

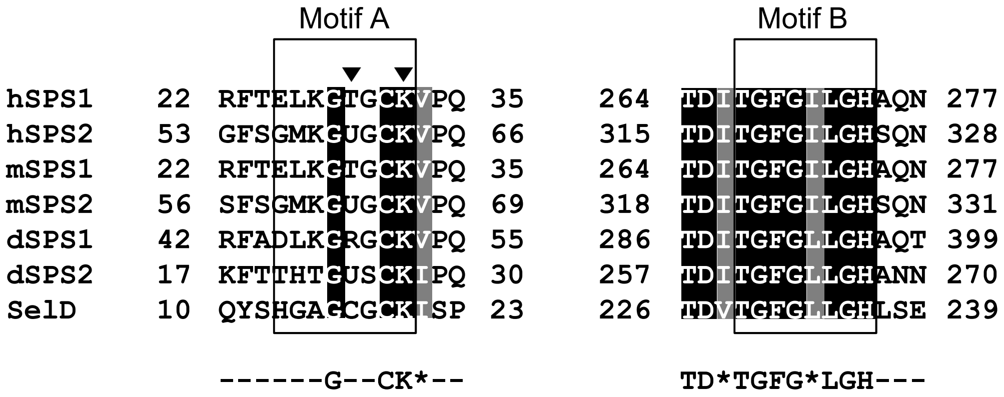

Motif A
Motif B
hSPS1 22 RFTELKGTGCKVPQ 35 264 TDITGFGILGHAQN 277
hSPS2 53 GFSGMKGUGCKVPQ 66 315 TDITGFGILGHSQN 328
mSPS1 22 RFTELKGTGCKVPQ 35 264 TDITGFGILGHAQN 277
mSPS2 56 SFSGMKGUGCKVPQ 69 318 TDITGFGILGHSQN 331
dSPS1 42 RFADLKGRGCKVPQ 55 286 TDITGFGLLGHAQT 399
dSPS2 17 KFTTHTGUSCKIPQ 30 257 TDITGFGLLGHANN 270
SelD 10 QYSHGAGCGCKISP 23 226 TDVTGFGLLGHLSE 239
------G--CK*--
TD*TGFG*LGH---

Figure 1.16. Two conserved motifs of SPS from human, mouse, *Drosophila* and *E.coli.* Strictly conserved residues are marked in black background and less conserved residues in grey. Consensus sequences are the bottom of the Figure. Asterisk marks represents the hydrophobic amino acid group: Ile (I), Leu (L), and Val (V). Positions corresponding to Cys 17 and Lys 20 of E.coli SelD are represented by an arrowhead. U denotes the Sec residue. Catalytic center (motif A) and the ATP/GTP binding site (motif B) are boxed. h, human; m, mouse; d, *Drosophila*.

2.4.2. Functional characteristics of eukaryotic SPS

Both *SPS1* and *SPS2*, homologous genes of *SelD*, were initially proposed to contain a catalytic activity in selenophosphate synthesis (Low *et al*., 1995; Guimaraes *et al*., 1996). In the earlier time, Tamura and colleagues had proposed that both SPS1 and SPS2 were able to synthesize selenophosphate by different pathways. When cDNA of *SPS1* from human lung adenocarcinoma was cloned and transformed into a *SelD* deficient mutant of *E.coli*, the *SelD* mutation was not complemented in selenite containing medium, but complemented when the cells were cultured in the medium supplemented with L-selenocysteine (L-Sec; Tamura *et al*., 2004). From these finding, it was proposed that SPS1 might be able to synthesize selenophosphate by recycling intracellular L-Sec through salvage pathway (Figure 1.17). However the mechanism of Sec recycling and how SPS1 regulates Sec recycling have not yet been determined.

On the other hand, there are some evidences that SPS1 does not participated in selenophosphate synthesis. *Drosophila* SPS1 purified from the overexpression in *E.coli* did not catalyze the selenide-dependent ATP hydrolysis reaction and its gene did not complement a *SelD* lesion in *E.coli*. Recently, it was subsequently shown that only SPS2 catalyzes selenophosphate synthesis. *In vitro* experiments, SPS2 synthesized selenophosphate from selenide and ATP, but SPS1 did not have this activity (Xu *et al*., 2007a). Knockdown of SPS2 in NIH3T3 cells led to the loss of selenoprotein biosynthesis, whereas the inhibition of SPS1 expression did not affect the biosynthesis of selenoproteins, and only SPS2 was capable of restoring selenoprotein synthesis (Xu *et al*., 2007b). In some insects such as the red beetle and silkworm, which have lost the selenoprotein synthesizing machinery, including *SPS2*, *SPS1* still encoded in the genome (Lobanov *et al*., 2008).

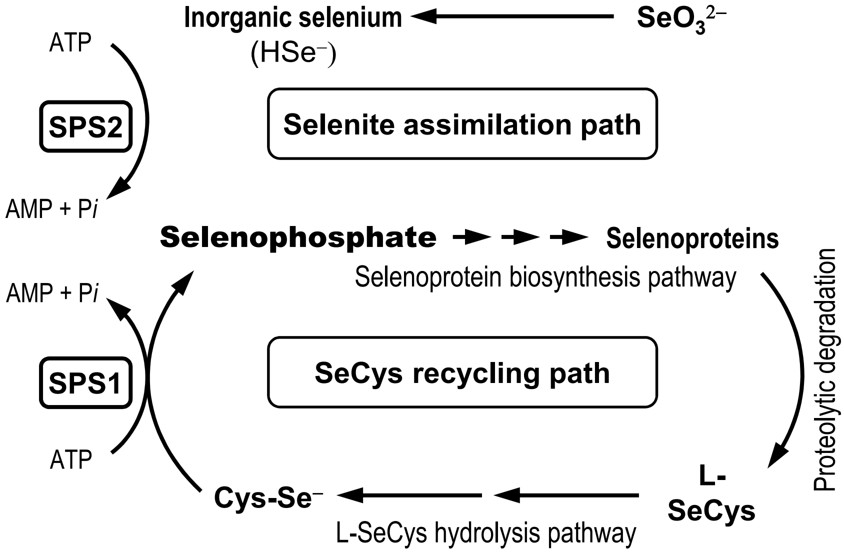


Figure 1.17. Hypothetical selenium assimilation reoutes in the lung adeno carcinoma cell. Up-regulation of SPS2, capable of using selenide derived from selenite, provides a bypass route, which directly converts selenide into monoselenophosphate, leading to an increased cellular selenium pool. The SeCys-60 residue in *Sps2* is proposed to provide a selenide binding site for enzyme-substrate complex formation. SPS1 that lacks a SeCys or Cys residue in the corresponding glycine-rich sequence would require a selenium-delivery system in which activated selenium is supplied as a perselenide (-S-SeH) derivative. (modified from Tamura *et al*., 2004)

Although SPS1 does not participate in the synthesis of selenophosphate, it has an essential function in *Drosophila* as the knock-out of the gene encoding SPS1, designed *patufet*, led to aberrant imaginal disc morphology and embryonic lethality (Alsina *et al*., 1998). The null mutation of *patufet* caused an accumulation of ROS and suggesting SPS1 plays a role in reducing the intracellular ROS level (Morey *et al*., 2003). Haplo-insufficiency of *patufet* dominantly suppressed the phenotypes caused by hyperactivation of the Ras/mitogen-activated protein kinase (MAPK) cassette, and the activation of the *Drosophila* epidermal growth factor receptor and Sevenless receptor tyrosine kinases (Morey *et al*., 2001). These evidences suggest that SPS1 may have novel function unrelated to selenophosphate synthesis, but related with important signaling pathway as well as cell growth.

3. VITAMIN B_6

Vitamin B_6 is a water-soluble compound and is part of the vitamin B complex group (Combs, 2008). There are several forms of the vitamin, such as pyridoxine (PN), pyridoxine 5'-phosphate (PNP), pyridoxal (PL), pyridoxal 5'-phosphate (PLP), Pyridoxamine (PM), and pyridoxamine 5'-phosphate (PMP). The structures of vitamin B_6 are summarized in Figure 1.18.

3.1. Physiological roles of vitamin B_6

PLP, the metabolically active form of vitamin B_6, is involved in the variety of macronutrient metabolism, including neurotransmitter, histamine and hemoglobin synthesis. Moreover, PLP generally served as a cofactor for many biological processes and can help facilitate decarboxylation, transamination, racemization, and replacement and β–group interconversion reactions (Grogan, 1988; Mihara *et al.*, 1997).

PLP is a cofactor in transaminases in amino acid metabolism (Dolphin *et al.*, 1986; Dakshinamurti, 1990). It is an essential component of two enzymes involved in cysteine and selenocysteine metabolism, or cystathionine β–synthase (CBS) and cystathionine β–lyase (CBL). In cysteine metabolism, CBS catalyzes to produce L-cystathionine, which is a precursor compound to be L-cysteine, from L-homocysteine and L-serine as substrates (Banerjee, 2005). CBL converts L-cystathionine to L-homocysteine in cysteine catabolism, and also L-selenocystathionine to L-selenohomocysteine in selenocysteine metabolism (Mihara *et al.*, 1997; Anderson *et al.*, 1979; Flavin and Slaughter 1964). Selenohomocysteine is then further transformed into hydrogen selenide. Low vitamin B_6 status will result in decreased activity of these

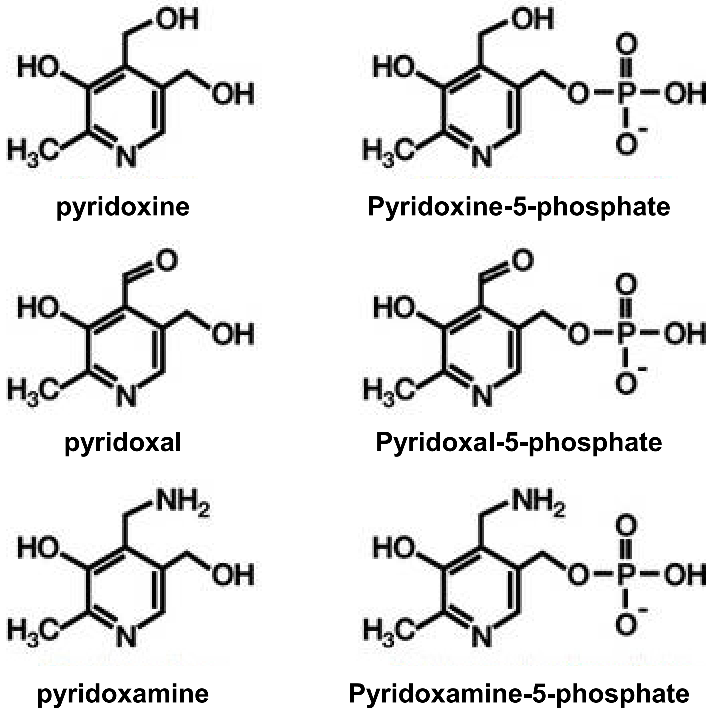

pyridoxine
Pyridoxine-5-phosphate
pyridoxal
Pyridoxal-5-phosphate
pyridoxamine
Pyridoxamine-5-phosphate

Figure 1.18. The structures of vitamin B$_6$

(modified from http://www.nutrition.tum.de/index.php?id=114) enzymes. PLP is also required for the conversion of tryptophan to niacin (Leklem *et al*., 1975) and used to produce physiologically active amines includes: histamine from histidin, serotonin from tryptophan, γ-aminobutyric acid (GABA) from glutamate and dopamine from dihydroxyphenylalanine (Lee *et al*., 1988; Schaeffer *et al*., 1998 and references therein).

PLP is also participated in gluconeogenesis and lipid metabolism. In gluconeogenesis, PLP is a required cofactor of glycogen phosphorylase, which is necessary for starting gluconeogenesis (Helmreich, 1992). In lipid metabolism, PLP is a crucial role of the biosynthesis of sphingolipids, particularly the synthesis of ceramide. In this reaction, serine is decarboxylated and combined with palmitoyl-CoA to form sphinganine which is combined with a fatty acyl-CoA to form dihydroceramide, which is then further desaturated to form ceramide.

3.2. Vitamin B_6 Biosynthesis

The biochemistry of *de novo* PLP biosynthesis has been studied in the gram-negative model organism *Escherichia coli* (Hill and Spenser, 1996). Also, molecular cloning and characterization of genes coding for enzymes involved in PLP biosynthesis was performed using this organism, mainly by Malcolm E. Winkler and coworkers: The PLP precursor pyridoxine 5'-phosphate (PNP) is synthesized by the PdxA and PdxJ enzymes using 4-phosphohydroxy-L-threonine (4PHT; synonym: 3-hydroxyhomoserine) and 1-deoxy-D-xylulose 5-phosphate (DXP) as substrates (Cane *et al*., 1998; Laber *et al*, 1999). PNP is oxidized to the active coenzyme PLP by the action of PdxH oxidase, a flavoprotein (Lam and Winkler, 1992; Notheis *et al*., 1995). The two substrates DXP and 4PHT are supplied by two independent pathways (Figure

1.19), which are both linked to carbohydrate metabolism, or glycolysis and the pentose phosphate cyle. DXP is also a precursor in isoprenoid and thiamine (vitamin B1) biosynthesis (Begley *et al.*, 1999), and synthesized by the transketolase-like enzyme DXP-synthase (DXS) (Lois *et al.*, 1998) using pyruvate and D-glyceraldehyde-3-phosphate as substrates. 4PHT (Zhao and Winkler, 1996) is formed in a series of reactions involving two oxidation steps and one transamination step in a pathway similar to serine biosynthesis starting from erythrose-4-phosphate (E4P), a central metabolite of the pentose phosphate pathway. E4P is also a precursor of aromatic amino acids (L-tryptophan, L-phenylalanine and L-tyrosine) and aromatic vitamins (p-aminobenzoate, p-hydroxybenzoate, 2,3-dihydroxybezoate) and it is produced directly by the action of transketolases TktA and TktB (Zhao and Winkler, 1994). These enzymes use D-glycerinaldehyde-3-phosphate and D-fructose-6-phosphate as substrates to produce Dxylulose-5-phosphate and E4P. In the first oxidation step, E4P is converted to 4-phosphoerythronate (4PE) by the action of dehydrogenases GapA or Epd (GapB) (Yang *et al.*, 1998). 4PE is further oxidized by the PdxB dehydrogenase to 3-hydroxy-4-phosphohydroxy-a-ketobutyrate. By a transamination reaction using glutamate as donor, this compound is finally transformed into 4PHT by the action of the PdxF (SerC) transaminase, a pyridoxal 5'-phosphate containing enzyme (Drewke *et al.*, 1996).

PLP can be synthesized by a salvage pathway (Figure 1.19) that uses B_6-vitamers pyridoxal (PL), pyridoxine (PN) and pyridoxamine (PM) present in the growth medium (Hill and Spenser, 1996; Yang *et al.*, 1996). In this pathway, the substrates PL, PN and PM are phosphorylated by kinases to form PLP, PNP and pyridoxamine 5'-phosphate (PMP). Two different kinases exhibiting a different

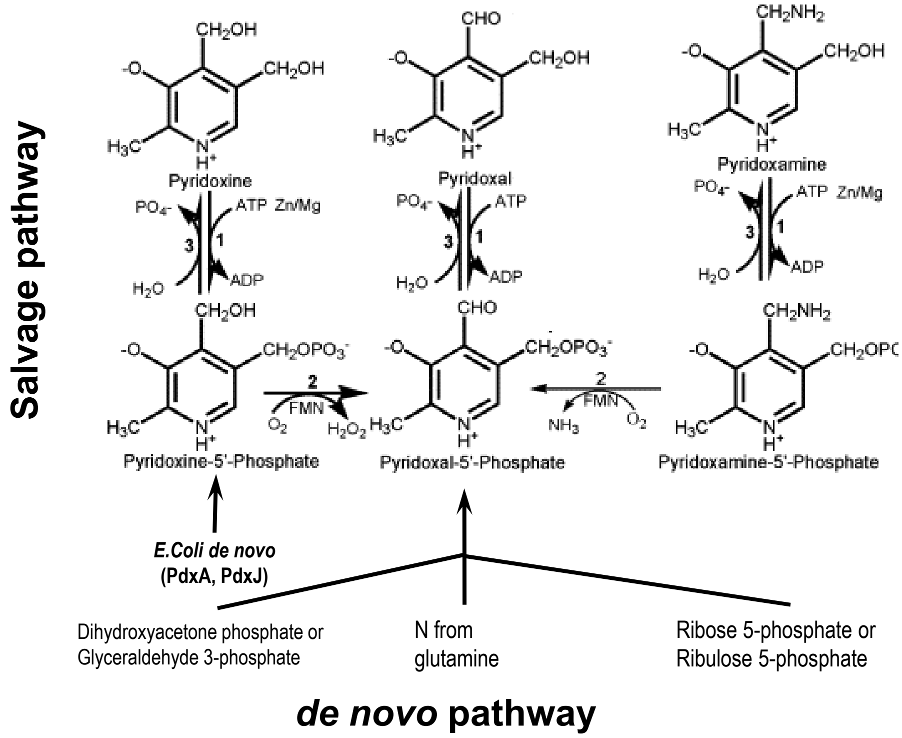


Figure 1.19. Biosynthetic pathways of vitamin B_6 biosynthesis. Six different vitamin B_6 vitamers, or pyridoxine (PN), pyridoxal (PL), pyridoxamine (PM) and their phosphorylated vitamers, pyridoxine 5'-phosphate (PNP), pyridoxal 5'-phosphate (PLP) and pyridoxamine 5'-phosphate (PMP), are interconverted between each other in the salvage pathway. PLP, the active vitamer, is synthesized directly in many organisms by a *de novo* biosynthetic pathway. Salvage pathway of B_6 vitaminers catalyzed by: (1) pyridoxal (PL) kinase, (2) pyridoxal phosphate (PLP) oxidase, (3) B6 vitamin kinase conversion can be reversed by phosphatases, (4) any unbound pyridoxal is oxidised by aldehyde oxidase/aldehyde dehydrogenase 2 (ALDH2) to form pyridoxic acid which is released into the plasma and excreted in the urine (modified from Depeint *et al.*, 2006).

substrate specifity have been identified in E. coli: The PN/PL/PM kinase PdxK (Yang *et al*., 1996) and the PL kinase PdxY (Yang *et al*., 1998a) PNP and PMP are oxidized to PLP by the PdxH oxidase which functions in both pathways. Alternatively, PMP can be converted to PLP by the action of transaminases. Although PdxA and PdxJ have not been found in animals (Ehrenshaft *et al*. 1999), similar salvage pathways involving oxidases and kinases exist in mammalian cells (McCormick and Chen, 1999). Therefore, species that synthesize PLP have one of the two PLP biosynthetic pathways. PLP Homoeostasis is maintaned by relatively unspecific PLP phosphatases.

4. MICRORNA BIOLOGY

MicroRNAs (miRNA) are small, evolutionarily conserved, non-coding ribonucleic acid (RNA) molecules involved in regulation of gene expression in essentially all eukaryotic organisms. They are on average 22 nucleotides in length, ranging from 18-25 nucleotides (Bartel *et al*., 2004). MicroRNAs are first described in 1993 with the identification of *lin-4*, a small RNA that repressed expression of Lin-14 protein in the nematode *C. elegans* (Lee *et al*., 1993). Presently, there are 1,049 human miRNA sequences registered in the miRBase (ver 16.0) miRNA database (Griffiths-Jones *et al*., 2008). It is estimated that up to 30% of human genes are regulated by miRNA expression (Rajewsky *et al*., 2004). MicroRNAs are involved in control of crucial cellular functions, including proliferation, apoptosis, development, differentiation and metabolism (Garzon *et al*., 2006; Alvarez-Garcia *et al*., 2005; Croce *et al*., 2005). They are tightly regulated and have been observed to show tissue-specific expression patterns during embryogenesis (Dalmay, 2008), though they are expressed in all tissues and at all stages of development (Hudder *et al*., 2008)

4.1. Biogenesis of microRNA

There are several steps involved in biogenesis of miRNA (Figure 1.20). The primary transcript, called pri-miRNA, is typically 3 to 4 kilobases in length with a 5'-7-methylguanosine (m7G) cap and 3'-polyadenylated [poly(A)] tail, similar to mRNA (Lee *et al*., 2004; Cai *et al*., 2004; Bracht *et al*., 2004). Following transcription, a stable hairpin structure of at least 30 bp is necessary to serve as the initiation signal for the processing steps (Lau *et al*., 2009). The pri-miRNAs are cleaved in the nucleus by

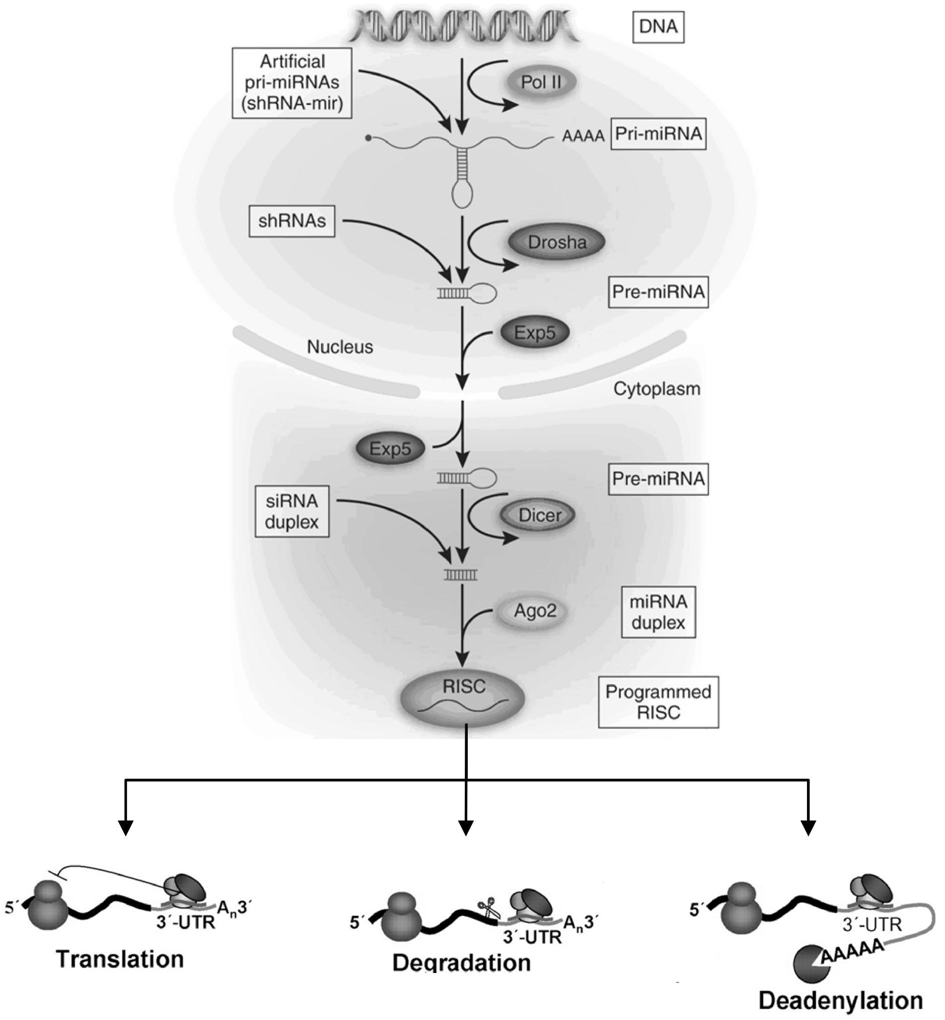


Figure 1.20. Biogenesis of miRNA. In the nucleus, the RNase III–type enzyme Drosha processes the long primary transcripts (pri-miRNA), yielding a hairpin precursors (pre-miRNA) consisting of approximately 70 nt. The pre-miRNA hairpins are exported to the cytoplasm where they are further processed into unstable, 19-25 nt miRNA duplex structures by the RNase III protein Dicer. The less stable of the two strands in the duplex is incorporated into a multiple-protein nuclease complex, the RNA-induced silencing complex (RISC), which regulates protein expression (modified from Cullen *et al*., 2005).

a multiprotein complex called Microprocessor, composed of the RNase III enzyme Drosha and double-stranded RNAbinding domain (dsRBD) protein DGCR8/Pasha, to produce one or more precursor miRNAs (pre-miRNA) (Lee *et al*., 2003; Hutvagner *et al*., 2001; Ketting *et al*., 2001). DGCR8/Pasha recognizes the junction of single and double-stranded RNA at the base of the pri-miRNA hairpin, binding Microprocessor to it, allowing Drosha to cleave it (Lau *et al*., 2009). Pri-miRNAs often contain several pre-miRNAs, known as clusters. Microprocessor activity can be inhibited through direct competitive binding of RNA-binding nuclear proteins (Lau *et al*., 2009; Ivan *et al*., 2008), structural alterations of pri-miRNA (Guil *et al*., 2007) or direct protein interaction with Microprocessor (Lau *et al*., 2009).

Pre-miRNAs are 65-100 nucleotides long with a hairpin structure containing a doublestranded RNA stem (Lau *et al*., 2009). Exportin 5 (Exp5) recognizes the 3' overhang, which is characteristic of pre-miRNA, and a portion of the RNA duplex structure (Thomson *et al*., 2006; Karube *et al*., 2005) and transports the pre-miRNA from the nucleus to the cytoplasm. Once in the cytoplasm, the pre-miRNA is bound by a protein complex called RISC-loading complex (RLC), which consists of another RNase III, called Dicer, along with Argonaut 2 and TRBP proteins (Gregory *et al*., 2005). Dicer recognizes the stem of the hairpin structure as double-stranded RNA and cleaves it on the loop side, leaving an 18-25 base pair miRNA duplex (miRNA: miRNA*) (Dalmay, 2008; Garzon *et al*., 2006).

The strand of the duplex with its 5' end on the less thermodynamically stable end of the duplex, termed the guide strand, is retained and becomes the mature miRNA (Chen *et al*., 2008; Mishra *et al*., 2008). The other strand, denoted as miRNA*, is removed and degraded (Dalmay, 2008; Garzon *et al*., 2006). This is facilitated by

Dicer, which also may help stabilize the miRNA and play a role in mRNA target identification (Lee *et al*., 2004). The mature miRNA is then incorporated into a protein complex formed with an Argonaut family (Ago) protein, called RNA Induced Silencing Complex (RISC) (Gregory *et al*., 2005). There are 4 Argonaut proteins expressed in humans, of which only Ago 2 possesses endonucleolytic activity (Hudder *et al*., 2008).

4.2. Functions and regulation of microRNA

4.2.1. MicroRNA-mediated gene regulation

The mature miRNA forms a complex with a member of the Argonaut (Ago) protein family, termed the RNA-induced silencing complex (RISC), and guides it specifically to the target messenger RNA (mRNA) through base pairing interactions generally at the 3' UTR of the target. Nucleotides 2-7 in the 5' region of the miRNA, called the seed region, bind the target mRNA through near-perfect base-pairing (Lewis *et al*., 2005). The remainder of the miRNA binds the target mRNA with varying degrees of complementarity (Lewis *et al*., 2005). If the miRNA is a perfect or nearperfect complement, cleavage and degradation of the mRNA is induced through de-capping of the 5' m7G cap or de-adenylation of the poly(A) tail. If there is a partial complement, RISC inhibits translation through competitive m7G cap binding by Ago 2 with the translational initiating factor eIF4E (Kiriakidou *et al*., 2007). These translationally-silenced mRNA:RISC complexes remain in the cytoplasm and accumulate, forming processing bodies (P-bodies) (Bhattacharyya *et al*., 2006). P-bodies contain decapping proteins and exoribonuclease, and therefore are capable of degrading the mRNAs. However, there is newly emerging evidence that miRNA translational silencing may be reversible, allowing mRNAs to leave P-bodies and migrate to ribosomes for translation (Bhattacharyya *et al*., 2006). Since base pairing with the target does not have to be a perfect compliment, a single miRNA can potentially affect mRNA and protein levels of 200 or more genes (Rouhi *et al*., 2008).

4.2.2. Epigenetic regulation of microRNA

MicroRNA expression can be regulated epigenetically, either through DNA methylation (Rouhi *et al*., 2008) or histone modification (Scott *et al*., 2006). Approximately 10% of miRNAs are regulated by DNA methylation (Han *et al*., 2007) and are more frequently methylated than protein-coding genes (Weber *et al*., 2007). Although it is currently unknown exactly why this is, three general reasons have been suggested (Weber *et al*., 2007); first, the increased frequency of miRNA methylation could be due to the specific nucleotide sequences surrounding the miRNA-associated CpG islands; second, miRNA could be embedded in specific chromosomal structures predisposing them to methylation; third, the predilection for methylation could be related to the tight regulation of miRNA expression.

There are three general mechanisms by which miRNA expression can be controlled through methylation: first, most commonly, miRNA can be embedded within or near a CpG island, which functions as its promoter (Rouhi *et al*., 2008); second, miRNAs can be located within an imprinted region (Royo *et al*., 2008; Seitz *et al*, 2004), thus preventing transcription; third, intronic miRNAs can be regulated by CpG island methylation of the promoter of the host gene (Grady *et al*., 2008; Lin *et al*., 2006). About forty-seven percent of miRNAs in the miRNA registry database (miRBase), and all miRNAs currently linked to epigenetic regulation, are associated with CpG islands (Weber *et al*., 2007). Some miRNA promoters are unmethylated in normal tissue, while others are normally methylated (Rouhi *et al*., 2008; Brueckner *et al*., 2007). Promoter methylation of miRNA results in reduced expression or transcriptional silencing.

4.3. Promoters and miRNA

4.3.1. MicroRNA transcription

As previous mentioned, microRNA expression is regulated by transcription factors and transcribed by RNA Polymerase II (Pol II), similar to protein-coding genes, although the precise mechanisms of transcriptional control of miRNAs are not entirely understood. While most miRNAs reside within intergenic non-coding regions (Lujambio *et al*., 2007), they can also be located in introns, exons or untransclated regions (UTRs) of coding genes (Rouhi *et al*., 2008; see Figure 1.21). Many miRNAs are embedded close to other miRNAs in the genome, giving rise to miRNA clusters (Lujambio *et al*., 2007). Single and clustered miRNAs can be transcribed from their own promoters, generally located within 500 base pairs of the 5' end of the miRNA, as single or polycistron transcriptional units, respectively (Lujambio *et al*., 2007; Zhou *et al*., 2007).

4.3.2. Promoter regions

Transcription of protein coding genes as well as miRNAs, is carried out by

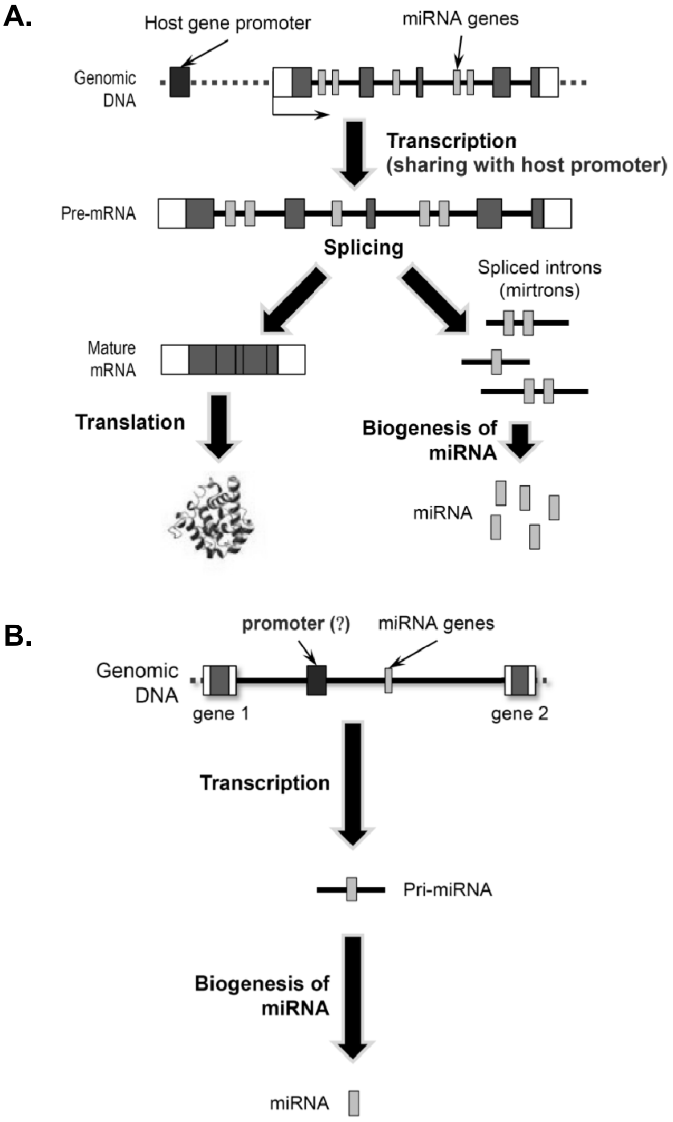


Figure 1.21. Transcription of miRNA. Transcription of intronic miRNAs (mirtron) (A) and intergenic miRNAs (B). Intronic miRNAs are located in exons, introns or 3', 5'-UTRs of annotated transcripts. For this reason, they can be transcribed by sharing host transcription unit. However, intergenic miRNAs are located in flank region of the genome, thus they are thought to be transcribed by using their unique transcription units.

Pol II. While Pol II binds to the DNA at the transcription initiation point, it is not capable of directly recognizing its target (Nikolov *et al*., 1997). A complex of proteins in a region known as the core promoter binds to the DNA whereupon they recruit Pol II to the transcription start site (TSS). Other proteins, called transcription factors (TFs), then bind to the proximal promoter or enhancer regions to either activators or repressors the activation of Pol II.

The core promoter region typically consists of the couple hundred base pairs surrounding the TSS of a gene. This region was once thought to contain a handful of known features able to be bound by elements of the Pol II protein complex; though it is now known that there is a wide diversity of properties that can be identified. It was initially believed that the core promoter regions consisted of a TATA box (~30bp upstream of the TSS) and an initiator sequence (Inr; overlaps the TSS). Recent studies have estimated the prevalence of these two sequences in only about 16% of human promoters (Yang *et al*., 2007), typically for tissue specific genes (Carninci *et al*., 2006). It has been identified that approximately half of the human core promoters are located around CpG islands. CpG islands are regions of a high concentration of CG dinucleotides, which are very underrepresented across the genome (Antequera, 2003).

For non-CpG island related promoters it was discovered that addition sequences such as the down stream promoter element (DPE; ~28bp downstream of the TSS) and the TFIIB recognition element (BRE; ~35bp upstream of the TSS) were also targets for proteins involved in the recruitment of RNA Pol II to the TSS besides the TATA box and the Inr. While some hypothesized that at least 2 of these 5 elements were necessary for the transcriptional initiation complex to bind (Gershenzon *et al*., 2005), other researchers have identified gene specific elements capable of binding this complex such as the downstream core element (DCE) in the human β-*globin* promoter (Smale and Kadonaga, 2003) and the multiple start site downstream element (MED-1) in the *pgp1* promoter (Ince *et al*., 1995). This suggests that the core promoter structure may be more complicated than originally hypothesized.

CpG islands are found at a low frequency in the genome because methylated CG dinucleotides can easily be mutated to TG dinucleotides, a process that is not corrected by DNA repair mechanisms (Smale and Kadonaga, 2003). Genes that have CpG islands in their promoters tend to be ubiquitously expressed across most tissues and throughout development. DNA methylation is a mechanism for which the cell can block the binding of transcription factors to promoters in order to prevent transcription of certain genes; because genes with CpG islands tend to be continuously expressed, they have not had as many opportunities to be methylated relative to the remainder of the genome (Antequera, 2003). Another feature that distinguishes genes with CpG islands is that they are more likely to have multiple TSSs that can span over 100bp whereas TATA-Inr containing promoters tend to have just one TSS (Carninci *et al*., 2006). The transcription factor *Sp1* is capable of binding to CpG islands and recruiting Pol II (Lee *et al*., 2005).

4.3.3. Transcription factors

The regulatory regions outside of the core promoter are the proximal promoter region and the enhancer regions. These regions, which can be located upstream of the gene, downstream of the gene, or even in the gene’s introns, are bound by TFs that activate or repress the functionality of Pol II (Ng *et al*., 1989). TFs will typically consist of a DNA binding domain and in the case of activators, an activation domain. Each TF usually binds to a specific set of sequence motifs 6-15bp in length (Stormo, 2000). The over 2,000 human TFs can be broken down into families based upon their structural properties that will typically correspond to their preferred binding motif (Mahony *et al*., 2007; Sandelin *et al*., 2004). TFs can function individually, in tandem, or in competition with each other (Umetani *et al*., 2001).

The identification of transcription factor binding sites (TFBSs) in the genome is an important and highly researched subject. The advancement of large-scale chromatin immunoprecipitation technology (ChIP-chip) has provided biologists with the tools necessary to identify many binding sites for a specific factor (Buck *et al*., 2004). The limitations of these studies are that the results only provide information on one specific cell type, one cellular condition, and only a single TF. In addition to laboratory procedures that can identify TFBSs, computational biologists have taken up the task of locating their motifs given the complete sequences of genomes and some external information.

A variety of computational methods have been developed over the years for the identification of TFBSs. The short and often degenerate nature of the sequences makes them a challenge to identify over large genomic regions (Corcoran *et al*., 2005). Algorithms have been developed that search for specific strings of sequences that match known TFBSs, called library based methods. Other methods use an IUPAC alphabet in the form of a consensus sequence to match the variability in TFBSs (Quandt *et al*., 1995). The most common method of representing the binding motifs of a TF is a position-specific scoring matrix (PSSM). A PSSM provides a mathematical model that represents all of the known binding sites for a given TF (Stromo *et al*., 2000). The PSSM can be combined with other features to help in the efficiency of identifying TFBSs such as sequence conservation across species or looking for common TFBSs in the promoters of genes that appear to be co-regulated given the similarity in their expression profiles (Corcoran *et al*., 2005; Bussemaker *et al*., 2001; Lenhard *et al*., 2003; Loots *et al*., 2004; Roth *et al*., 1998).

The identfication of TFBSs is an essential step in the understanding of gene regulatory networks. The further analysis of miRNA promoters and the transcription factors that bind them will be an essential component to the understanding of the regulatory networks in hich they participate.

5. EPIGENETICS

Epigenetics is defined as heritable but reversible changes in gene expression due to genetic modifications without a change in the DNA sequence (Esteller *et al*., 2008; Jiang *et al*., 2004; Liard., 1996). These changes include DNA methylation and histone modifications, and play crucial roles in various biological processes such as the regulation of gene expression, embryonic development and genomic stability

5.1. DNA methylation

DNA methylation occurs at dinucleotides in which cytosine is upstream and adjacent to guanine, called CpGs. This takes place when S-adenosylmethionine (SAM) donates a methyl (CH3) group that is covalently attached to the 5-carbon of a cytosine pyrimidine ring in a reaction catalyzed by the enzyme DNA methyltransferase (DNMT) (Zangi *et al*., 2010). Approximately 50-70% of CpG dinucleotides are methylated in normal human tissue, termed global methylation (Ehrlich *et al*., 1982). However, most of the human genome is depleted of CpG dinucleotides due to the relative instability of m5C, which can result in spontaneous hydrolytic deamination of the cytosine base to thymine (Gronbaek *et al*., 2007). CpGs are not randomly distributed throughout the genome but rather are concentrated in CpG enriched regions (Esteller *et al*., 2008; Gronbaek *et al*., 2007; Shi *et al*., 2007) referred to as CpG islands. Specifically, CpG islands are defined as a sequence greater than 0.5 kb with a G+C content greater than or equal to 55% and an observed vs. expected CpG ratio greater than 60% (Takai *et al*., 2002). They typically range from 0.5 to 5 kb in length and occur on average every 100 kb in the genome (Das *et al*., 2004). These

CpG islands are often disproportionately concentrated in the 5' promoter regions of genes. Approximately 50% of all human genes have CpG islands in the promoter region. While CpG dinucleotides are frequently methylated in normal tissue, promoter-associated CpG islands are generally not methylated, although methylation of subgroups of CpG islands may occur (Gronbaek *et al.*, 2007; Shi *et al.*, 2007).

During DNA replication, the methylation pattern of the parent strand is transferred onto the new strand by DNMT1 (Zangi *et al.*, 2010). DNMT1 has an affinity for hemimethylated DNA (Bestor *et al.*, 1992). However, during early embryonic development or carcinogenesis, previously unmethylated DNA may be methylated in a process mediated by DNMT3a or DNMT3b, which is termed *de novo* methylation (Gronbaek *et al.*, 2007; Das *et al.*, 2004). DNMT3b has a propensity to target pericentromeric satellite regions for methylation, which are prone to loss of stability as a result of hypomethylation, leading to chromosomal breakage (Okano *et al.*, 1999).

There are several mechanisms guarding against aberrant promoter methylation (hypermethylation), including active transcription, active demethylation, timing of replication and prevention of access to DNMT by local chromatin structure (Das *et al.*, 2004). Enzymes that actively demethylate DNA are called demethylases. These may include the glycosylases thymine-DNA glycosylase (TDG) and methyl-CpG-binding domain protein 4 (MBD4), which remove the methylated cytosine (5-meC) leaving the deoxyribose intact to be replaced with a new cytosine via DNA repair (Gehring *et al.*, 2009; Zhu *et al.*, 2009); methyl-CpG-binding protein 2 (MBD2), which is believed to demethylate by hydrolyzing 5-meC to cytosine and methanol (Das *et al.*, 2004); or thymine removal by glycosylases through the G/T mismatch base excision repair pathway following 5-meC deamination to thymine by DNMT3a or DNMT3b (Zhu *et al*., 2009).

Generally, methylation of CpG islands in the promoter region is associated with transcriptional silencing of the gene, whereas methylation of downstream gene sequences has no influence on expression (Gronbaek *et al*., 2007; Figure 1.22). DNA methylation is capable of repressing gene expression in three general ways. One mechanism is through direct interference with transcription factors. Several transcription factors, including AP-2, c-Myc, CREB, E2F, and NFkB, recognize and bind promoter regions containing CpG islands and are inhibited by methylation. A second mechanism involves inhibition of transcription through the direct binding of transcriptional repressors to 5-meC in the promoter region, including MBD1, MBD2, MBD4, Kaiso, MeCP1, and MeCP2. Finally, CpG methylation can guide the deacetylation of histones and subsequently alter chromosome structure to prevent transcription. Methylated cytosines of silenced promoters can bind methyl-CpG-binding-domain proteins (MBD), forming a complex involving histone deacetylase enzymes (HDAC) (Gronbaek *et al*., 2007; Das *et al*., 2004).

5.2. Histone modification

Promoter methylation is not the sole epigenetic mechanism capable of silencing gene expression. Modification of histone proteins can result in the alteration of chromatin structure, directly affecting gene transcription, DNA repair, DNA replication and chromosomal organization (Esteller *et al*., 2008; Carley *et al*., 2007; Gronbaek *et al*., 2007). Histones are protein octamers, containing 2 of each of H2A, H2B, H3, and H4, around which approximately 146 bp of DNA is wound, forming a

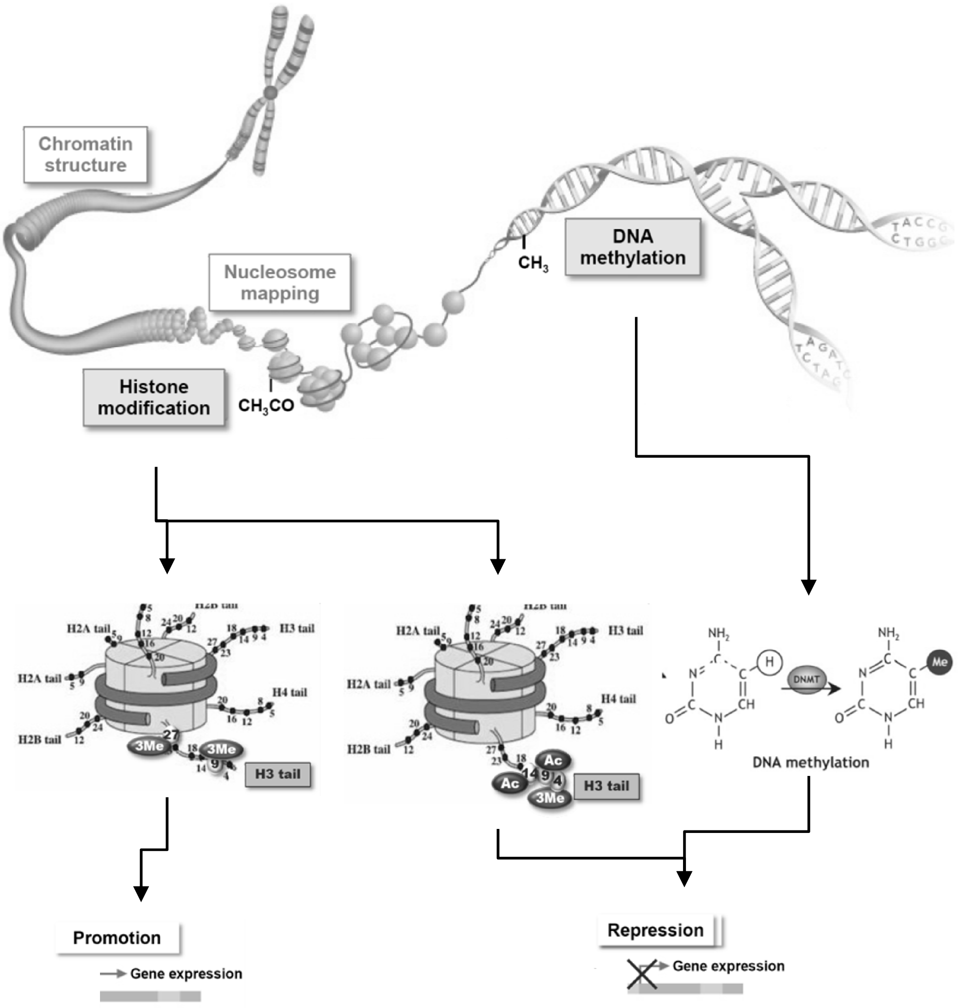


Figure 1.22. Transcription regulation by epigenetic signatures. Some histone modifications, such as H3K4Me3, H3K9Ac or H3K14Ac are generally active markers for gene expressions, whereas DNA methylation and other histone modification, such as H3K9Me3 and H3K27Me3, are repression markers for gene expressions.

nucleosome. The nucleosome is a recurring structure of eukaryotic DNA that comprises the chromosomes, condensing the DNA so that the entire genome can fit into the nucleus. Most chromatin exists as tightly compacted nucleosomes, called heterochromatin, which is transcriptionally incompetent. This is represented by the dark staining portion of the nucleus on light microscopy. Euchromatin has less compact nucleosomes, forming an open chromatin structure that can be readily transcribed. This appears as the lightly staining portion of the nucleus on light microscopy (Gronbaek *et al*., 2007).

Histone modification occurs in different histone proteins, histone variants and histone residues such as lysine, arginine and serine. Modifications typically involve addition or removal of acetyl or methyl groups to the histone proteins at the N-terminal tails protruding from the nucleosomes (Zangi *et al*., 2010; Esteller *et al*., 2008; Gronbaek *et al*., 2007).

Histone acetylation is associated with transcriptional activation (Figure 1.22). In transcriptionally active promoters with unmethylated cytosines, histones are acetylated by histone acetyl transferases (HAT). These form a complex with transcription activator and coactivator proteins to initiate transcription. Conversely, histone deacetylases (HDAC) form complexes with methyl-CpG-binding-proteins (MBD) and methylated cytosines in the promoter, allowing them to remove acetyl groups from the N-terminal tails of the histones, causing condensation of the nucleosome, resulting in transcriptional inactivation (Esteller *et al*., 2008; Gronbaek *et al*., 2007).

Histone methylation can result in either transcriptional activation or repression, depending upon the protein and amino acid type methylated and its position in the histone tail (Esteller *et al*., 2008; Gronbaek *et al*., 2007). It can also have different degrees, including mono-, di- and trimethylation. Histone methylation is catalyzed by a class of enzymes called histone methyltransferases, while histone demethylases are responsible for demethylation (Bartova *et al*., 2008). Trimethylation of lysine at position 9, 27, or 36 of the N-terminal tail of H3 (H3K9, H3K27, or H3K36) or lysine at position 20 on H4 (H4K20) results in chromosomal structure alterations (heterochromatin) leading to transcriptional silencing (Bracken *et al*., 2006; Kahlil *et al*., 2009). Trimethylation of lysine at position 4, 36, or 79 on H3 (H3K4, H3K36 or H3K79) is associated with a euchromatin conformation and active transcription (Okitsu *et al*., 2010; Bartova *et al*., 2008; Lachner *et al*., 2001; Santos-Rosa *et al*., 2002). Several other covalent methyl histone modifications have been identified, but their precise effects on transcription are presently unknown.

CHAPTER 2.

DROSOPHILA SELENOPHOSPHATE SYNTHETASE 1 REGULATES VITAMIN B_6 METABOLISM: PREDICTION AND CONFIRMATION

1. INTRODUCTION

Selenium has been reported to provide many health benefits in animals, including humans, when obtained from the diet in adequate amounts. For example, selenium has been known to play roles in cancer prevention, aging retardation, immune augmentation, prevention of heart diseases, muscle development and development (Boosalis *et al*., 2008; Flohé *et al*., 2007; Hatfield *et al*., 2006; Tamura *et al*., 2004 and references therein; see Figure 1.9). Many of the health benefits of selenium are mediated by selenoproteins, which contain selenocysteine (Sec) as a selenium containing amino acid (Hatfield *et al*., 2006).

Selenophosphate synthetase (SPS) synthesizes selenophosphate (SeP), the active selenium donor in Sec biosynthesis, using selenide and ATP as substrates (Ehrenreich *et al*., 1992; see Figure 1.13). SeP serves as a selenium donor during Sec biosynthesis (Glass *et al*., 1993). Sec is contained in all selenoproteins (Aitken *et al*., 2005). SPS was first isolated from *Escherichia coli* as one of the enzymes involved in selenoprotein synthesis and was designated SelD (Leinfelder *et al*., 1990). Only one type of SPS, SelD, exists in lower eukaryotes and eubacteria, however, there are two isoforms of SPS, SPS1 and SPS2, occur in higher eukaryotes (Guimaraes *et al*., 1996). One of the major differences in the sequences between SPS1 and SPS2 is that SPS1 has an arginine at the position corresponding to Sec in SPS2 (Low *et al*., 1995).

Although it is not clear why there are two SPSs in higher eukaryotes, recent studies have shown that SPS2 synthesizes SeP from selenide and ATP *in vitro*, while SPS1 does not have this activity (Xu *et al*., 2007). Loss of function in NIH3T3 cells using RNA interference technology showed that SPS2 is required for selenoprotein biosynthesis, while SPS1 does not affect the biosynthesis of this protein class (Xu *et al.*, 2007). While some insects such as the red beetle and silkworm have lost the selenoprotein synthesizing machinery including SPS2, SPS1 is still encoded in the genome of these insects, suggesting SPS1 is required for a function other than SeP synthesis (Lobanov *et al.*, 2008).

Although SPS1 does not catalyze SeP biosynthesis, it plays essential roles in the cell. When the gene encoding SPS1 (*SPS1*, also designated *patufet*) was deleted in *Drosophila*, the embryo showed lethality during development (Alsina *et al.*, 1998), and reactive oxygen species (ROS) levels increased (Morey *et al.*, 2003). The haploinsufficiency of genes involved in the *Ras*-regulated signaling pathway was also suppressed by *SPS1* knockout in *Drosophila* (Morey *et al.*, 2001). From the finding that the SelD (*E. coli* SPS) mutant of *E. coli* can be complemented by human SPS1 only when L-Sec is supplemented in the medium, it was suggested that SPS1 is involved in the recycling of Sec (Tamura *et al.*, 2004). However, the means by which SPS1 may be involved in Sec recycling has not been determined. Recently, it was found that the targeted depletion of *SPS1* by RNA interference in *Drosophila* SL2 cells causes growth inhibition, ROS induction and megamitochondrial formation by increasing intracellular glutamine levels (Shim *et al.*, 2009). Interestingly, human SPS1 was found to interact with the soluble liver antigen, which was recently identified as eukaryotic Sec synthase (SecS), and the binding reaction was enhanced by Sec tRNA methylase designated SECp43 (Small-Howard *et al.*, 2006; Xu *et al.*, 2007). It should be noted that SecS is a pyridoxal phosphate (PLP)-dependent enzyme and, therefore, the uptake and/or activation of vitamin B_6 may be related to selenium metabolism (Forchhammer *et al.*, 1991; Ganichkin *et al.*, 2008).

Vitamin B_6 is a water-soluble compound that contains a pyridine ring and is part of the vitamin B complex group (Combs, 2008). Vitamin B_6 is present in nature as several different forms such as pyridoxal (PL), pyridoxine (PN), pyridoxamine (PM) and their 5'-phosphorylated forms, including pyridoxine 5'-phosphate (PNP), pyridoxal 5'-phosphate (PLP), and pyridoxamine 5'-phosphate (PMP) (Fitzpatrick *et al*., 2007; see Figure 1.18). Before use, these vitamers are converted to PLP, which is the metabolically active form. PLP is used as a cofactor for PLP-dependent enzymes, where the pyridine ring acts as an electron sink during enzymatic reactions. PLP is involved in the variety of macronutrient metabolism, including neurotransmitter, histamine and hemoglobin synthesis and can help facilitate decarboxylation, transamination, racemization, and replacement and β–group interconversion reactions (Grogan, 1988; Mihara *et al*., 1997). Since animals, including humans, cannot synthesize vitamin B_6, they must obtain it from their diet (Mooney *et al*., 2009). PLP can be synthesized through several different pathways, and two types of enzymes, kinases and oxidases, participate in these pathways. For PM to be converted to PLP, it is first phosphorylated by a kinase (PL/PM/PN kinase) to form pyridoxamine phosphate (PMP), and then the PMP is oxidized to form PLP using an oxidase (PMP/PNP oxidase). PN can also be converted to PLP using the same kinase and oxidase used for PM. In this case, the phosphorylated intermediate is pyridoxine phosphate (PNP). However, PL can be directly converted to PLP by phosphorylation using a kinase (Gonza´lez *et al*., 2007). Therefore, kinases and oxidases are important enzymes for PLP synthesis.

There are more than 100 PLP-dependent enzymes in a cell that perform essential roles in various metabolic pathways including amino acid metabolism (such as amino acid synthesis and degradation), fatty acid metabolism (such as the synthesis of polyunsaturated fatty acids), and carbohydrate metabolism (such as the breakdown of glycogen) (Mooney *et al*., 2009 and references therein). The PLP-dependent enzymes that participate in amino acid metabolism can be classified into 4 categories: transaminase, racemase, decarboxylase and α, β-eliminase (Yoshimura *et al*., 2008). Interestingly, the biosynthesis of Sec can be mediated by cystathionine β-synthase (CBS) using serine as a precursor and it can also be synthesized by cystathionine ɣ-lyase (CGL) from selenocystathionine (Esaki *et al*., 1981; Meier *et al*., 2001). Both CBS and CGL are PLP-dependent enzymes (Aitken *et al*., 2005). In addition, enzymes that are involved in the degradation of Sec, such as selenocysteine lyase (SCL), D-selenocystine α, and β-lyase, use PLP as a cofactor (Soda *et al*., 1998). Recently, it was found that SCL can interact with SPS1 (Tobe *et al*., 2009). Therefore, it seems that vitamin B_6 participates in the metabolism of Sec, i.e., in the biosynthesis and/or decomposition of Sec.

In the present study, we found that the knockdown of *SPS1* led to the down regulation of genes involved in PLP biosynthesis, which, in turn, induced the formation of megamitochondria and the expression of genes responsible for innate immunity. Our findings suggest that SPS1 primarily regulates PLP biosynthesis, and the intracellular PLP level affects various biological processes such as amino acid metabolism, megamitochondrial formation and innate immune response.

2. MATERIALS AND METHODS

2.1. Reagents and other materials

Materials were purchased from the following sources: *Drosophila* Schneider cell line 2 (SL2) was purchased from Invitrogen, HyQ SFX-Insect medium from Hyclone, T3 Megascript kit from Ambion, RNeasy mini kit from Qiagen, GeneChip *Drosophila* genome 2.0 array from Affymetrix, SYBR Green mix from Applied Biosystems, TRIzol reagent from Invitrogen, Moloney murine leukemia virus reverse transcriptase from Super-Bio, 4-deoxypyridoxine hydrochloride from TCI, 5',6,6'-tetrachloro-1,1',3,3'-tetraethylbenzimidazolyl-carbocyanine iodide (JC-1) from Molecular Probes, and oligonucleotides from Cosmo Genetech. The sequences of oligos used for RT-PCR are listed in Table 2.1.

2.2. SL2 cell culture and RNA interference

SL2 cell culture and preparation of double-stranded RNAs were carried out as described (Shim *et al.*, 2009). Briefly, for RNA interference, 0.25×10^6 cells were plated on a 24-well plate containing 0.5 ml of HyQ SFX-Insect medium. Four micrograms of dsRNAs were added directly to the medium and incubated for 48 hr and cells were split into appropriate culture dishes for further incubation and other experiments.

2.3. Microarray experiment

Table 2.1. Oligonucleotides sequences used as primers for RT-PCR or real-time PCR

Gene	**Forward primer (5'à3')**	**Reverse primer (5'à3')**
SPS1	tactaggccacgctcaaa	gccggttacaactgaatg
CG31472	gggattcaccttcttcacga	gtctgaccctgccagaactc
CG11899	tcccttcgatgtctccaagt	gccataccagcgactcctc
Egr	gcacataccggcaccacgcc	ttgcgatcgttgtgaatgtc
AttB	acaatctggatgccaaggtc	tacatctataccagggtaatat
AttD	agtttatggagcggtcaacg	aggtgatgattggcacttcc
CG8745	gcccagaacgaatttcttga	Aaattaggcggatcaacgtg
Cec B	aatccgatcgtaagccaaca	Agagaaatgagcgggtcgag
Drs	gtacttgttcgccctcttcg	acaggtctcgttgtcccaga
Dro	cataccgcggagaagtcatc	ttaggggacaaacccattca
Dpt B	atcctgatccccgagagatt	tgaagtgccctaaaacctgaa
PGRP-SD	atgacttggatcggtttgct	cccgttcttcgaagttacca
Mtk	ccaccgagctaagatgcaa	tgttaacgacatcagcagtgtg
W	gagaacgacaaaaggcgaag	actttctgctttcgctcctg
Oat	actcgtctatgccagggaga	gatgatctttgcctggttgg
GS1	gatcgcgttttggacaaagt	gacgtccgtccacgtctaat
Arg	cccaaggatcagctggttta	gagtgcctccacgatgct
Cyp6a8	ttccagagtcccgctgca	ccatgtctcttgtcaacc
PGRP-LF	ttacccgaccctattggttg	ttcattccccttgctttcag
Toll-7	gtctgcagcacagagacctg	gagcggcgaattctatgaac
RP49	cagtcggatcgatatgcta	aatctccttgcgcttctt

Microarray experiments were performed using the GeneChip *Drosophila* genome 2.0 array. After the addition of double stranded RNAs targeting SPS1 to the culture medium, total RNA was extracted from SL2 cells treated with or without SPS1 dsRNA on day 1, 3 and 5 after treatment using the RNeasy mini kit according to the manufacturer's instructions. The cells that were not treated with any dsRNA were used as controls. The RNA quality was checked using Experion (Applied Biosystems) according to the manufacturer's instructions. Five micrograms of total RNAs were reverse transcribed with oligo-dT primer containing a T7 RNA polymerase promoter (TAATACGACTCACTATAGGG). Biotin-labeled cRNAs were generated from the cDNA sample by *in vitro* transcription with T7 RNA polymerase. The labeled cRNAs were fragmented to an average size of 35-200 bases by mild alkaline treatment at 94°C for 40 min. Fragmented cRNAs were hybridized with probes that are on GeneChip *Drosophila* genome 2.0 array, and the chips were washed and stained in the Affymetrix Fluidics Station 450 by following the procedures established by Affymetrix (Affymetrix GeneChip R Expression Analysis Technical Manual). The signals were scanned using the GeneChip Scanner 3000 7G (Affymetrix). Overall experimental scheme of microarray is shown in Figure 2.1.

2.4. Microarray data processing

The raw data were imported into Acuity 4.0 software (Molecular Devices, Inc.), and a background adjustment and normalization were performed using robust multichip average (RMA) and quantile methods, respectively, implemented in Acuity 4.0 software (Bolstad *et al*., 2003; Irizarry *et al*., 2003). To identify differentially expressed genes (DEGs), a two-way analysis of variance (ANOVA) model was used

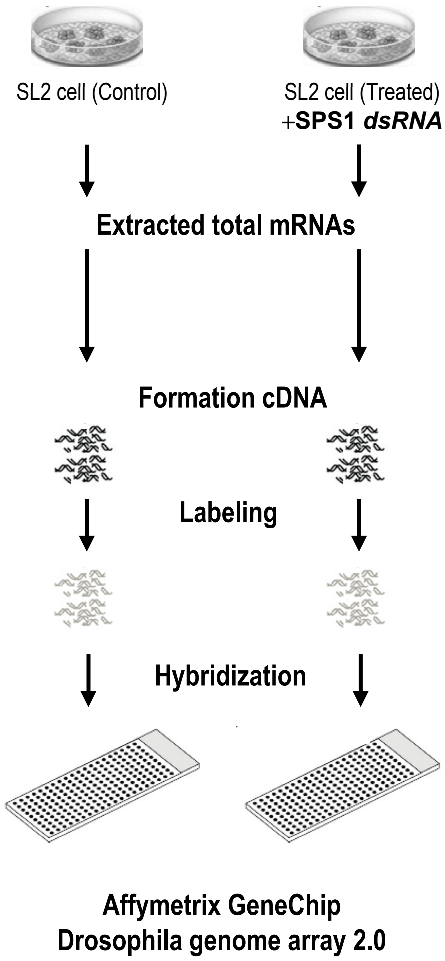


Figure 2.1. Procedure of Affymetrix GeneChip experiment. After the addition of double stranded RNAs targeting SPS1 to the culture medium, total RNA was extracted from SL2 cells treated with or without SPS1 dsRNA on day 1, 3 and 5. Extracted RNA samples were reverse transcribed with T7 RNA polymerase and labeled with biotin.

and fitted using the R software (www.r-project. org), as described by Park *et al* (Park *et al*., 2003). Two models were considered to identify DEGs. Model 1 contains group and time effects as well as their interactions. Model 1 allows the expression level of genes to change over time (days 1, 3 and 5) and these change patterns to differ between groups (control and knockdown). Model 2 includes only group and time effects assuming that the expression level of genes changes over time but these change patterns are the same between groups. From Model 1, DEGs were identified by the genes with significant interaction effects, while from Model 2 DEGs were identified by the genes with significant group effects. The p-values were adjusted by Westfall and Young's method (Westfall *et al*., 1993). The genes with adjusted p-values less than 0.1 were identified.

2.5. Temporal clustering

To classify DEGs according to their temporal expression pattern, DEGs were clustered using a self-organizing map (SOM) algorithm implemented in Acuity 4.0 (Tamayo *et al*., 1999). The ratios of normalized $\log_2$ values of DEGs between *SPS1* knockdown cells and control cells were used as input data and the SOM map size was set to 3 x 2. The ranges of expression ratios of DEGs within each cluster at each sampling date were displayed by box plot using R software. The interquartile ranges (IQRs) of each cluster were compared to select cluster(s) whose IQRs were significantly deviated. The criterion for determining clusters within which gene expressions were changed significantly was set to 0.75, i.e., when the interquartile range (IQR) of a cluster was larger than +0.75 or smaller than −0.75, the cluster was selected as significantly changed. This is because 0.75 is the threshold value to isolate clusters on day 3 (see Results for more details). The genes composing a cluster selected at the early stage (day 3) were defined as an early responding gene-set and those composing a cluster selected at the late stage (day 5) were defined as a late responding gene-set.

2.6. Gene ontology analysis

GO analysis was performed by BiNGO version 2.3 (Maere *et al.*, 2005), which is plugged in Cytoscape (Shannon *et al.*, 2003). Gene symbols of each gene-set were used as input data. The parameters were set as follows: assessment was set to overrepresentation, statistical test to binomial test, multiple testing correction to FWER correction, significance level to 0.05. Among GO evidence codes, inferred from electronically annotated (IEA) were discarded. The most significant pathway was predicted by considering the selected GO terms and visualized output.

2.7. RT-PCR and quantitative real time RT-PCR

RT- PCR and real time PCR were carried out as described (Shim *et al.*, 2009). Briefly, total RNA was isolated from the cells using the TRIzol reagent. cDNAs were synthesized from total RNAs with Moloney murine leukemia virus reverse transcriptase and oligo (dT) primers according to the manufacturer's protocols. RT-PCR was performed with 0.1 μg of template total RNA and specific primers (Table 2.1). RT-PCR products were electrophoresed on a 2 % agarose gel and visualized by ethidium bromide. For the measurement of relative mRNA levels of each gene, real time PCR was carried out using an ABI 7300 real time PCR system (Applied

Biosystems) as follows. cDNAs were amplified using SYBR Green mix and specific primers for 40 cycles [initial incubation at 50°C for 2 min and then at 95°C for 10 min, and 40 cycles (95°C for 15 sec, 55°C for 1 min and 72°C for 1 min)]. Output data were obtained as *Ct* values using Sequence Detection Software (SDS) version 1.3 (7300 System, Applied Biosystems) and the differential mRNA expression of each gene between control and knockdown cell was calculated using the comparative *Ct* method (Schmittgen *et al*., 2008). RP49 mRNA, an internal control, was amplified along with the target genes, and the *Ct* value of RP49 used to normalize the expression of target genes.

2.8. Measurement of intracellular PLP concentration

Cellular PLP levels were determined using the method previously described (Perry *et al*., 2007) with minor modifications. At day 5 after treatment with dsRNA or 4-DPN, cells were washed with phosphate buffered saline and harvested. Cells (6×10^7) were lysed by resuspension in 600 µl of distilled water. Cell extracts were induced to produce the semicarbazon derivative of PLP as follows: 40 µl of 250 mg/ml of both semicarbazide and glycine were added into 500 µl of cell extracts or PLP standard. The mixture was vortexed and incubated at room temperature in the dark for 30 min. Proteins were then precipitated by adding 50 µl of 60% $HClO_4$ into the mixture, and the solution was thoroughly mixed for 1 min. The solution was clarified by centrifugation for 10 min at 15,000×g, and 30-50 µl of a 25% NaOH solution was added to the supernatant to achieve a pH between 3.0 and 5.0. HPLC was performed using a ZORBAX SB-C18 column (4.6 mm × 25 cm, PN 880975902) and an isocratic mobile phase consisting of 60 mM sodium phosphate (pH 6.5), 400 mg/l EDTA and

9.5% methanol at a flow-rate of 1 ml/min, and the derivatized PLP was quantified using a Waters^{TM} 474 scanning fluorescence detector by setting excitation and emission wavelengths to 380 and 450 nm, respectively.

2.9. Mitochondrial staining and confocal microscopy

Mitochondrial staining and confocal microscopy were carried out as described (Shim *et al*., 2009). Briefly, SL2 cells (0.5×10^6) were plated onto a chambered coverglass one day before staining. Cells were incubated with 1 μg/ml JC-1 for 30 min at 25 °C, washed three times with HyQ-SFX-Insect medium and observed with a LSM510 confocal microscope (Carl Zeiss) at 512×512 pixel resolution through an X63 C-Apochromat objective. Excitation wavelengths for JC-1 aggregate and JC-1 monomer were 543 and 488 nm, respectively.

3. RESULTS

3.1. Identification of differentially expressed genes

To elucidate the molecular function of SPS1, DEGs were identified by microarray analysis in SL2 cells where *Drosophila* SPS1 was knocked-down. After the addition of double stranded RNAs targeting SPS1 to the culture medium, total RNAs were isolated on days 1, 3 and 5 after treatment and subjected to microarray analysis using Affymetrix microchips (GEO accession number: GSE 17685). Because megamitochondrial formation begins 3 days after knockdown (Shim *et al.*, 2009), transcriptomes were analyzed before and after megamitochondrial formation to find the primary target of SPS1. The knockdown efficiency was approximately 90% which was similar with that obtained in the previous work (Shim *et al.*, 2009). The $\log_2$ values of signal intensity of 18,952 transcripts on each chip were obtained after normalization. By performing two-way ANOVA analysis (adjusted P-value < 0.1) against the $\log_2$ values of signal intensities of transcripts, a total of 238 genes were found to be different in their expression between knockdown and control cells. Twenty-three genes were selected by model 1 and 227 genes by model 2, with 12 genes being common between the two models (Figure 2.2). Among these genes, only one gene (SPS1) showed more than two-fold change in expression on day 1. However, 55 and 201 genes among DEGs were found to have more than two fold expression changes on day 3 and 5, respectively. These results suggest that the effect of SPS1 knockdown on the expression of target genes were manifested slowly. List of DEGs was shown in Table 2.2.

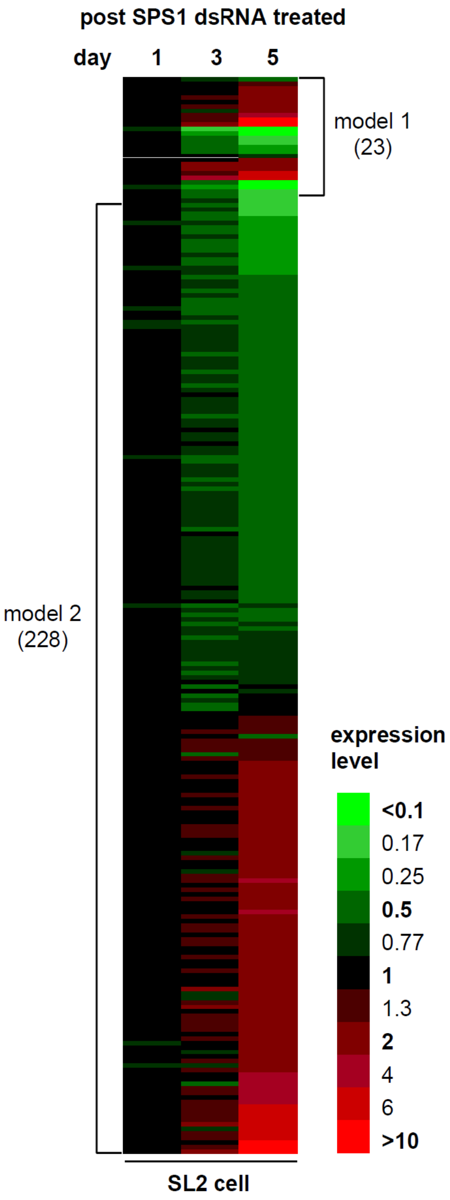


Figure 2.2. Identification of DEGs by microarray analysis. DEGs expression profiles ordered by fold changes on day 5 in each model. The number in bracket represents the gene numbers including each model. Genes selected by both model (model 1 and 2) overlapped between model 1 and model 2.

Table 2.2. List of differenctially expressed genes

	CG Symbol	Symbol	ANOVA	Fold change			Gene Ontology	
			p-value	Day 1	Day 3	Day 5	Molecular Function	Biological Process
*	CG1311	CG1311	2.64E-02	0.97	0.68	0.32	_	_
*	CG30059	CG30059	7.18E-03	0.96	1.14	1.43	N-acetylglucosamine-6-sulfatase activity	N-acetylglucosamine metabolic process
*	CG30046	CG30046	3.17E-02	0.95	1.18	2.04	_	_
*	CG14196	CG14196	1.13E-03	0.98	1.06	2.15	secondary active monocarboxylate transmembrane transporter activity	transmembrane transport
*	CG4926	Ror	1.21E-02	1.02	1.34	2.48	transmembrane receptor protein tyrosine kinase activity	central nervous system development
*	CG9042	Gpdh	3.87E-03	0.97	1.24	2.70	glycerol-3-phosphate dehydrogenase (NAD+) activity	triglyceride metabolic process
*	CG3413	wdp	6.97E-02	0.97	1.59	3.04	_	_
*	CG14253	CG14253	6.36E-02	0.96	0.64	3.22	_	_
*	CG6231	CG6231	7.63E-02	0.94	1.36	5.53	secondary active organic cation transmembrane transporter activity	transmembrane transport
*	CG7496	PGRP-SD	3.57E-03	0.96	1.73	11.73	peptidoglycan binding	defense response to Gram-positive bacterium
*	CG32185	CG32185	1.23E-02	1.06	2.33	12.82	_	_
$	CG31472	CG31472	2.85E-04	0.68	0.15	0.07	pyridoxamine-phosphate oxidase activity	pyridoxine biosynthetic process
$	CG1572	CG1572	8.40E-03	0.78	0.22	0.10	_	_
$	CG3625	CG3625	3.45E-02	0.93	0.42	0.11	_	_
$	CG31658	Nnf1b	2.26E-02	0.93	0.45	0.18	_	mitotic metaphase plate congression
$	CG14872	CG14872	4.20E-02	0.96	0.45	0.23	binding	_
$	CG32280	CG32280	1.58E-03	0.95	0.35	0.25	_	_
$	CG5272	gnu	3.87E-03	0.92	0.88	0.67	_	regulation of cell cycle
$	CG6639	CG6639	5.55E-02	1.10	1.18	1.95	serine-type endopeptidase activity	proteolysis
$	CG12919	egr	3.40E-03	0.98	2.03	2.85	protein binding	immune response
$	CG30456	CG30456	1.32E-02	1.06	2.19	3.75	Rho guanyl-nucleotide exchange factor activity	regulation of Rho protein signal transduction
$	CG9505	CG9505	1.62E-03	1.08	1.82	6.59	metalloendopeptidase activity	proteolysis
$	CG15526	CG15526	7.53E-03	1.01	4.38	9.68	_	_
	CG32985	CG32985	3.42E-02	0.85	0.26	0.07	catalytic activity	metabolic process
	CG12014	CG12014	7.46E-02	0.77	0.23	0.10	iduronate-2-sulfatase activity	metabolic process
	CG34008	CG34008	2.00E-02	1.00	0.49	0.13	_	_
	CG15078	Mctp	9.38E-02	0.92	0.45	0.13	_	_
	CG4398	CG4398	2.05E-04	0.97	0.65	0.14	_	_
	CG4210	CG4210	1.74E-02	0.84	0.32	0.15	N-acetyltransferase activity	metabolic process
	CG5535	CG5535	9.44E-02	0.93	0.54	0.16	amino acid transmembrane transporter activity	amino acid transport; amino acid transport
	CG5165	Pgm	6.70E-02	1.00	0.41	0.17	phosphoglycerate mutase activity	glycogen biosynthetic process
	CG31658	Nnf1b	1.93E-02	0.93	0.45	0.18	_	mitotic metaphase plate congression
	CG17566	gammaTub37C	7.58E-03	0.84	0.30	0.19	GTP binding	microtubule-based process
	CG4531	argos	8.05E-02	0.59	0.69	0.19	receptor antagonist activity	wing disc morphogenesis

CG Symbol	Symbol	ANOVA	Fold change			Gene Ontology	
		p-value	Day 1	Day 3	Day 5	Molecular Function	Biological Process
CG32446	Atox1	1.51E-02	0.86	0.43	0.21	metal ion binding	metal ion transporter
CG33462	CG33462	1.09E-02	0.83	0.34	0.21	serine-type endopeptidase activity	proteolysis
CG12390	dare	3.33E-02	1.05	0.53	0.21	NADPH-adrenodoxin reductase activity	steroid biosynthetic process
CG15399	CG15399	2.23E-02	0.91	0.42	0.23	_	_
CG6659	CG6659	9.78E-03	0.80	0.34	0.23	_	_
CG1318	Hexo1	7.44E-02	0.92	0.53	0.24	beta-N-acetylglucosaminidase activity	carbohydrate metabolic process
CG10063	CG10063	2.61E-02	0.90	0.30	0.25	_	_
CG12883	CG12883	4.82E-02	0.98	0.56	0.25	_	_
CG6965	mthl5	3.22E-02	0.89	0.44	0.25	G-protein coupled receptor activity	G-protein coupled receptor signaling pathway
CG12880	CG12880	3.80E-02	0.69	0.72	0.25	_	_
CG17610	grk	1.87E-03	0.78	0.45	0.27	epidermal growth factor receptor binding	maternal determination of dorsal/ventral axis
CG31142	CG31142	5.62E-02	1.02	0.74	0.27	_	_
CG31975	CG31975	9.67E-02	0.94	0.59	0.28	_	_
CG17181	CG17181	6.32E-02	0.94	0.50	0.29	zinc ion binding	_
CG1753	CG1753	1.15E-02	0.85	0.48	0.29	cystathionine beta-synthase activity	cysteine biosynthetic process from serine
CG1787	Hexo2	8.37E-02	0.94	0.65	0.29	beta-N-acetylglucosaminidase activity	negative regulation of growth of symbiont in to fungus
CG5008	GNBP3	7.93E-02	0.88	0.42	0.29	pattern recognition receptor activity	
CG32521	CG32521	2.63E-03	0.76	0.45	0.29	_	_
CG2715	Syx4	2.21E-02	0.92	0.41	0.30	SNAP receptor activity	inter-male aggressive behavior
CG6854	CTPsyn	7.95E-02	0.89	0.74	0.32	CTP synthase activity	CTP biosynthetic process
CG8588	pst	4.33E-03	0.59	0.27	0.32	_	olfactory learning
CG3920	l(2)k16918	1.54E-03	0.88	0.70	0.33	_	_
CG3960	CG3960	3.39E-02	0.77	0.59	0.33	actin binding	mesoderm development
CG13795	CG13795	1.10E-03	0.88	0.77	0.33	neurotransmitter transporter activity	neurotransmitter transport
CG9641	CG9641	7.85E-02	0.90	0.61	0.34	_	_
CG10824	CG10824	2.44E-03	0.78	0.36	0.35	_	_
CG18528	CG18528	6.30E-03	0.92	0.56	0.35	GTPase activity	tRNA modification
CG4057	tamo	8.73E-03	0.84	0.56	0.35	Ran GTPase binding	protein transport
CG11899	CG11899	1.44E-02	0.90	0.68	0.36	O-phospho-L-serine:2-oxoglutarate aminotransferase activity	pyridoxine biosynthetic process
CG10160	ImpL3	4.33E-02	1.00	0.65	0.36	L-lactate dehydrogenase activity	glycolysis
CG10424	CG10424	4.55E-02	0.95	0.64	0.37	_	_
CG12840	Tsp42El	6.81E-02	0.82	0.32	0.37	_	_
CG12643	CG12643	5.69E-02	1.01	0.94	0.38	_	_
CG8994	exu	1.33E-02	0.89	0.47	0.38	_	embryonic development
CG3779	numb	7.79E-03	0.85	0.66	0.38	protein binding	regulation of developmental process
CG2669	CG2669	4.23E-02	1.01	0.66	0.38	_	cell proliferation
CG17905	ChLD3	6.70E-02	0.92	0.55	0.38	hydrolase activity	chitin metabolic process
CG2177	CG2177	4.04E-02	0.88	0.55	0.39	metal ion transmembrane transporter activity	transmembrane transport
CG3950	CG3950	3.39E-02	0.88	0.68	0.39	actin binding	mesoderm development

CG Symbol	Symbol	ANOVA	Fold change			Gene Ontology	
		p-value	Day 1	Day 3	Day 5	Molecular Function	Biological Process
CG2200	CG2200	6.53E-03	0.85	0.55	0.39	dipeptidase activity	proteolysis
CG17129	CG17129	2.73E-02	0.93	0.57	0.39	_	_
CG33250	AlkB	1.28E-02	0.90	0.52	0.40	oxidoreductase activity	
CG4502	CG4502	1.89E-02	0.85	0.50	0.40	acid-amino acid ligase activity	post-translational protein modification
CG1218	CG1218	1.30E-02	0.87	0.65	0.40	_	_
CG17721	CG17721	2.74E-02	0.91	0.63	0.41	zinc ion binding	_
CG3770	CG3770	2.26E-02	0.87	0.78	0.41	_	establishment and or maintenance of cell polarity
CG18412	ph-p	6.96E-02	0.86	0.73	0.41	chaperone binding	central nervous system neuron development
CG11513	armi	6.98E-02	0.98	0.71	0.42	DNA helicase activity	nuclear-transcribed mRNA catabolic process
CG3961	CG3961	4.71E-02	0.92	0.87	0.42	long-chain-fatty-acid-CoA ligase activity	metabolic process
CG4330	CG4330	6.19E-02	0.93	0.62	0.42	high affinity inorganic phosphate:sodium symporter activity	transmembrane transport
CG8936	Arpc3B	3.73E-02	0.75	0.41	0.42	actin binding	actin filament organization
CG6137	aub	5.02E-03	0.83	0.44	0.43	piRNA binding	regulation of metabolic process
CG12367	Hen 1	1.15E-02	0.95	0.71	0.43	O-methyltransferase activity	posttranscriptional gene silencing by RNA
CG13117	CG13117	5.40E-02	0.78	0.44	0.43	_	_
CG14079	CG14079	3.82E-03	1.00	0.56	0.43	_	_
CG2100	CG2100	3.91E-02	0.90	0.67	0.43	polynucleotide adenylyltransferase activity	RNA processing
CG5126	CG5126	7.21E-02	0.96	0.75	0.43	_	_
CG6204	CG6204	3.84E-03	0.97	0.68	0.43	_	_
CG32706	CG32706	2.68E-02	0.87	0.49	0.43	nucleotide binding	_
CG3570	CG3570	2.01E-02	0.90	0.71	0.43	_	_
CG6383	crb	2.65E-02	0.99	0.68	0.43	protein kinase C binding	system development; biological regulation
CG34033	CG34033	4.15E-02	0.88	0.73	0.43	_	_
CG7737	CG7737	9.74E-02	0.88	0.71	0.43	_	_
CG33134	debcl	4.49E-02	0.94	0.65	0.44	_	negative regulation of neuron apoptosis
CG12608	CG12608	2.08E-02	0.92	0.63	0.44	_	_
CG15083	CG15083	1.52E-02	0.93	0.81	0.44	_	_
CG32043	CG32043	3.04E-02	0.81	0.71	0.44	_	_
CG4947	Tgt	1.44E-02	0.97	0.70	0.44	queuine tRNA-ribosyltransferase activity	queuosine biosynthetic process
CG13602	CG13602	8.91E-02	1.02	0.40	0.44	_	_
CG12283	kek1	3.01E-02	0.81	0.56	0.45	epidermal growth factor binding	negative regulation of EGFR signaling pathway
CG4711	squ	4.94E-02	0.96	0.65	0.45	_	gene silencing by RNA
CG1962	CG1962	8.39E-02	0.97	0.71	0.46		
CG9471	CG9471	9.22E-02	0.97	0.76	0.46	NADPH dehydrogenase activity	metabolic process
CG1344	CG1344	1.36E-03	1.02	0.66	0.46	protein kinase activity	_
CG16790	CG16790	2.91E-02	0.97	0.75	0.47	protein binding	RNA metabolic process
CG7381	CG7381	4.91E-02	0.99	0.67	0.48	_	_
CG14346	CG14346	1.68E-02	0.89	0.77	0.48	_	_

CG Symbol	Symbol	ANOVA	Fold change			Gene Ontology	
		p-value	Day 1	Day 3	Day 5	Molecular Function	Biological Process
CG14229	CG14229	3.99E-02	0.94	0.66	0.49	_	_
CG7504	CG7504	5.76E-02	0.93	0.72	0.49	ATP-dependent RNA helicase activity	_
CG18410	Ude	3.50E-02	0.89	0.58	0.49	DNA binding	pupation
CG2604	CG2604	9.83E-02	0.95	0.78	0.49	catalytic activity	metabolic process
CG7995	CG7995	1.19E-02	1.00	0.80	0.50	glycerol kinase activity	glycerol kinase activity
CG12018	CG12018	6.84E-02	0.96	0.71	0.50	DNA-directed DNA polymerase activity	cellular response to DNA damage stimulus
CG9008	CG9008	1.44E-02	0.74	0.47	0.51	isomerase activity	carbohydrate metabolic process
CG15922	CG15922	7.22E-02	0.95	0.74	0.51	_	_
CG31274	CG31274	4.07E-02	0.94	0.52	0.51	_	_
CG4953	CG4953	6.01E-02	0.94	0.74	0.52	_	_
CG15893	CG15893	7.77E-02	0.96	0.31	0.54	_	_
CG8157	CG8157	7.29E-02	0.83	0.59	0.54	_	_
CG4338	CG4338	7.02E-02	0.97	0.75	0.55	_	_
CG31431	CG31431	2.79E-02	0.89	0.49	0.59	fibroblast growth factor receptor activity	fibroblast growth factor receptor signaling pathway
CG9460	Spn42De	1.58E-02	0.85	0.68	0.60	serine-type endopeptidase inhibitor activity	_
CG10999	CG10999	6.98E-02	0.92	0.66	0.60	_	_
CG31151	wge	9.66E-02	0.93	0.66	0.60	DNA binding	_
CG40293	Stlk	1.93E-02	0.91	0.77	0.62	protein serine/threonine kinase activity	protein amino acid phosphorylation
CG5144	CG5144	2.32E-02	1.03	0.36	0.67	arginine kinase activity	_
CG17559	dnt	1.55E-02	0.90	0.74	0.67	transmembrane receptor protein tyrosine kinase activity	signal transduction
CG1628	CG1628	7.68E-03	0.90	0.66	0.69	amino acid transmembrane transporter activity	mitochondrial ornithine transport
CG9338	CG9338	6.45E-02	0.91	0.49	0.77	_	_
CG4615	CG4615	6.13E-02	0.90	0.78	0.79	_	phagocytosis
CG17259	CG17259	1.27E-02	0.95	0.84	0.80	serine-tRNA ligase activity	seryl-tRNA aminoacylation
CG2065	CG2065	6.96E-02	0.87	0.37	0.81	affinity inorganic phosphate:sodium symporter activity	transmembrane transport
CG1598	CG1598	7.61E-02	0.95	0.93	0.83	arsenite-transporting ATPase activity	cellular metal ion homeostasis
CG4786	Rcd2	1.68E-02	0.90	0.50	0.89	_	centriole replication
CG2259	Gclc	6.49E-02	0.94	0.72	0.92	glutamate-cysteine ligase activity	glutathione biosynthetic process
CG9681	PGRP-SB1	1.99E-02	0.93	0.38	0.94	peptidoglycan binding	immune response
CG17834	CG17834	5.39E-02	0.90	0.39	0.98	_	_
CG5793	CG5793	9.78E-02	1.01	1.26	1.26	catalytic activity	metabolic process
CG2984	Pp2C1	5.39E-02	1.15	1.06	1.38	protein serine/threonine phosphatase activity	protein amino acid dephosphorylation
CG3792	CG3792	2.26E-02	1.06	1.28	1.40	_	_
CG7903	CG7903	4.91E-02	1.04	1.27	1.45	mRNA binding	_
CG9154	CG9154	9.39E-02	1.03	1.66	1.46	methyltransferase activity	methylation
CG12316	CG12316	1.96E-02	1.21	1.19	1.57	_	_
CG2794	CG2794	4.23E-02	1.03	1.43	1.59	phosphotransferase activity	_

CG Symbol	Symbol	ANOVA	Fold change			Gene Ontology	
		p-value	Day 1	Day 3	Day 5	Molecular Function	Biological Process
CG8031	CG8031	1.19E-02	1.05	1.39	1.69	_	_
CG17186	CG17186	1.88E-02	1.07	1.54	1.73	zinc ion binding	_
CG9733	CG9733	3.13E-02	0.94	0.38	1.77	serine-type endopeptidase activity	proteolysis
CG7523	CG7523	5.76E-02	1.02	1.42	1.82	_	_
CG9739	fz2	1.38E-02	0.93	0.96	2.01	Wnt receptor activity	receptor-mediated endocytosis
CG17124	CG17124	1.27E-02	0.97	0.78	2.03	phosphoprotein phosphatase inhibitor activity	regulation of phosphorylation
CG8595	Toll-7	1.43E-02	0.92	1.31	2.03	transmembrane receptor activity	signal tranduction
CR8687	Cyp6a14	2.73E-02	0.97	1.04	2.03	electron carrier activity	oxidation reduction
CG9453	Spn4	8.56E-02	1.08	1.09	2.03	serine-type endopeptidase inhibitor activity	negative regulation of peptide hormone processing
CG4559	Idgf3	5.11E-02	1.01	1.27	2.04	imaginal disc growth factor activity	imaginal disc development
CG3074	CG3074	3.39E-02	0.99	0.89	2.05	cysteine-type endopeptidase activity	proteolysis
CG6199	CG6199	3.04E-02	1.01	1.32	2.05	procollagen-lysine 5-dioxygenase activity	oxidation reduction
CG6043	CG6043	4.59E-02	1.04	1.00	2.06	_	_
CG7123	LanB1	5.58E-02	1.02	1.16	2.08	_	organ development; cell migration
CG9968	Anxb11	6.05E-02	1.03	1.45	2.09	actin binding	regulation of cell shape
CG6042	Cyp12a4	1.43E-02	1.04	0.80	2.09	electron carrier activity	response to insecticide
CG32714	CG32714	5.55E-02	1.09	1.16	2.11	_	_
CG9623	if	3.33E-02	1.10	1.19	2.14	receptor activity	biological regulation
CG33521	CG33521	7.02E-02	0.99	1.43	2.14	zinc ion binding	_
CG14680	Cyp12e1	5.34E-02	1.06	1.47	2.18	electron carrier activity	oxidation reduction
CG8051	CG8051	5.61E-02	1.08	1.53	2.18	secondary active monocarboxylate transmembrane transporter activity	transmembrane transport
CG33503	Cyp12d1-d	4.89E-02	0.98	1.25	2.21	electron carrier activity	oxidation reduction
CG30489	Cyp12d1-p	5.01E-02	0.99	1.18	2.23	electron carrier activity	oxidation reduction
CG9102	bab2	5.62E-02	0.92	1.21	2.24	protein binding	imaginal disc-derived leg morphogenesis
CG9331	CG9331	1.58E-02	0.99	0.72	2.24	NAD or NADH binding	metabolic process
CG5123	W	8.44E-02	0.88	1.45	2.25	_	biological regulation
CG4026	IP3K1	9.64E-02	1.05	0.93	2.25	inositol trisphosphate 3-kinase activity	response to oxidative stress
CG10249	CG10249	9.36E-02	1.21	1.06	2.27	_	_
CG33468	CG33468	5.14E-02	1.05	0.66	2.30	_	_
CG2086	drpr	5.60E-03	1.10	1.31	2.31	protein binding	phagocytosis
CG40498	CG40498	5.72E-02	1.08	1.41	2.31	_	_
CG2003	CG2003	4.13E-02	0.95	1.25	2.32	_	_
CG6126	CG6126	5.18E-02	0.98	1.40	2.32	organic cation transmembrane transporter	transmembrane transport
CG10810	Drs	5.68E-03	0.97	1.35	2.33	_	defense response to fungus
CG31075	CG31075	4.07E-02	1.03	1.27	2.33	aldehyde dehydrogenase (NAD) activity	pyruvate metabolic process
CG3424	path	5.39E-02	0.95	1.21	2.35	acid transmembrane transporter activity	growth
CG13654	CG13654	6.94E-02	1.00	1.11	2.37	_	_
CG3884	CG3884	3.39E-02	0.97	1.19	2.38	_	_

CG Symbol	Symbol	ANOVA	Fold change			Gene Ontology	
		p-value	Day 1	Day 3	Day 5	Molecular Function	Biological Process
CG6127	Ser	6.00E-02	1.16	1.42	2.40	protein binding	biological regulation
CG8846	Thor	5.68E-02	1.09	1.01	2.41	eukaryotic initiation factor 4E binding	immune response
CG11822	nAcRbeta-21C	3.95E-03	0.87	1.61	2.46	nicotinic acetylcholine-activated cation-selective channel activity	ion transport
CG5630	CG5630	4.73E-02	1.09	1.43	2.52	_	_
CG4475	Idgf2	6.58E-02	1.00	1.24	2.54	imaginal disc growth factor activity	imaginal disc development
CG6871	Cat	9.54E-02	1.10	1.45	2.55	catalase activity	response to hydrogen peroxide
CG5210	CG5210	3.56E-02	1.10	1.82	2.56	chitin binding	cuticle chitin catabolic process (Chit)
CG4501	bgm	9.39E-02	0.95	0.90	2.58	long-chain-fatty-acid-CoA ligase activity	long-chain fatty acid metabolic process
CG6612	Adk3	4.73E-02	1.11	1.15	2.60	nucleoside triphosphate adenylate kinase activity	ADP biosynthetic process
CG2913	yin	6.44E-02	0.92	1.61	2.66	proton-dependent oligopeptide secondary active transmembrane transporter activity	oligopeptide transport
CG32680	spri	4.13E-02	0.94	1.25	2.70	RasGTPase binding	border folicle cell migration
CG4500	CG4500	1.55E-02	1.04	0.86	2.79	long-chain fatty acid-CoA ligase activity	mesoderm development
CG8147	CG8147	7.29E-02	0.95	1.35	2.80	alkaline phosphatase activity	metabolic process
CG2718	Gs1	3.54E-03	1.06	1.27	2.84	glutamate-ammonia ligase activity	glutamine metabolic process
CG18550	yellow-f	9.66E-02	1.01	1.02	2.85	dopachrome isomerase activity	indole derivative biosynthetic process
CG4472	Idgf1	4.12E-02	0.89	1.23	2.90	imaginal disc growth factor activity	imaginal disc development
CG13907	CG13907	9.56E-02	1.07	2.15	2.92	secondary active monocarboxylate transmembrane transporter activity	transmembrane transport
CG10248	Cyp6a8	1.55E-02	0.88	0.76	2.94	alkane 1-monooxygenase activity	insecticide metabolic process
CG3348	CG3348	7.79E-02	0.96	0.70	2.97	chitin binding	chitin metabolic process
CG11299	Sesn	6.49E-02	1.13	1.78	2.99	_	negative regulation of cell growth
CG3085	CG3085	1.26E-02	1.15	2.00	3.05	_	microtubule cytoskeleton organization
CG8782	Oat	4.55E-02	0.89	0.91	3.12	ornithine-oxo-acid transaminase activity	ornithine metabolic process
CG5958	CG5958	1.23E-02	1.07	1.61	3.14	retinal binding	transport
CG16888	CG16888	1.88E-02	1.07	1.77	3.15	_	_
CG6863	tok	2.91E-02	0.91	1.72	3.19	metalloendopeptidase activity	motor axon guidance
CG6018	CG6018	4.13E-02	1.06	1.82	3.29	carboxylesterase activity	_
CG6006	CG6006	9.78E-02	0.82	1.15	3.34	transporter activity	transmembrane transport
CG18104	arg	5.95E-03	1.05	1.64	3.42	arginase activity	arginine catabolic process to ornithine
CG30484	CG30484	5.36E-02	0.70	0.82	3.43	_	_
CG32170	CG32170	2.96E-02	0.97	1.09	3.55	transition metal ion binding	oxidation reduction
CG15678	pirk	7.56E-02	0.95	0.58	3.71	receptor binding	negative regulation of innate immune response
CG10564	Ac78C	4.93E-02	0.95	1.23	3.77	adenylate cyclase activity	response to sucrose stimulus
CG10816	Dro	6.49E-02	1.06	1.61	3.81	_	defense response to Gram-positive bacterium
CG4437	PGRP-LF	3.56E-02	0.76	0.75	3.82	peptidoglycan binding	innate immune response
CG5304	l(2)01810	3.79E-02	1.12	1.98	3.91	high affinity inorganic phosphate:sodium symporter activity	transmembrane transporter
CG1878	CecB	1.55E-02	0.98	1.15	4.79	_	defense response to Gram-negative bacterium

CG Symbol	Symbol	ANOVA	Fold change			Gene Ontology	
		p-value	Day 1	Day 3	Day 5	Molecular Function	Biological Process
CG17725	Pepck	5.08E-02	1.11	1.23	4.88	phosphoenolpyruvate carboxykinase (GTP) activity	gluconeogenesis
CG13315	CG13315	2.29E-02	0.95	0.37	4.95	_	_
CG32625	CG32625	1.19E-02	1.19	1.41	5.08	_	_
CG12002	Pxn	3.69E-02	0.98	1.74	5.17	peroxidase activity	response to oxidative stress
CG3132	Ect3	7.21E-02	0.97	1.25	5.25	beta-galactosidase activity	autophagic cell death
CG42280	ome	3.76E-02	0.93	1.42	5.45	dipeptidyl-peptidase activity	proteolysis
CG14629	CG14629	9.20E-02	1.01	1.82	6.08	_	_
CG13077	CG13077	9.84E-02	0.90	1.71	6.10	_	_
CG5322	CG5322	9.80E-02	1.11	1.41	6.18	alpha-mannosidase activity	mannose metabolic process
CG4250	CG4250	4.63E-03	1.09	1.55	6.66	_	_
CG8745	CG8745	1.28E-02	1.17	2.30	7.04	ornithine-oxo-acid transaminase activity	arginine catabolic process to glutamate
CG4259	CG4259	2.00E-02	0.99	0.65	7.87	serine-type endopeptidase activity	proteolysis
CG18372	AttB	8.94E-03	1.02	1.74	8.84	_	defense response to bacterium
CG10794	DptB	4.69E-02	1.02	1.98	8.90	_	antibacterial humoral response
CG6124	eater	1.93E-02	0.96	1.17	10.24	bacterial cell surface binding	phagocytosis
CG7629	AttD	3.66E-02	1.00	1.60	10.27	_	antibacterial humoral response
CG8175	Mtk	9.97E-03	1.08	3.38	25.02	_	defense response to fungus

3.2. Functional distribution of differentially expressed genes

Total 238 identified DEGs were classified according to their GO terms, especially biological process and molecular function terms. One hundred and forty-eight genes (61.9%) could be annotated with their GO terms by direct searching GO database. As shown in Figure 2.3, the annotated functions of DEGs were classified into 9 categories. Genes participating in primary metabolic process, including carbohydrate, amino acid and protein metabolic process, took the largest portion among the genes whose functions were annotated (31.16%). The portion of genes participating in developmental process, transport and defense response were 17.6%, 14.9% and 12.8%, respectively. Therefore, the number of genes fallen into those four GO categories were more than 70% among the annotated genes suggesting SPS1 participates majorly in metabolic process, development, cell transport and defense response. Minor portion of DEGs was taken by the genes participating in signaling (5.4%), oxidation/reduction (5.4%), cellular component organization (4.7%), response to stimulus (4.1%) and gene expression (4.1%). It should be noted that the major effect of SPS1 knockdown was megamitochondrial formation mediated by the accumulation of glutamine and growth inhibition (Shim and Kim *et al.*, 2009). Therefore, it is interesting that genes related with the metabolic process were the most abundant among DEGs. The DEGs related with development can also participate in cell growth. It is, however, uncertain that defense response is related with megamitochondrial formation and cell growth. Cell growth inhibition and activation of cell defense system by SPS1 knockdown seems to be resulted by the secondary effect after knockdown, because those effects were shown at relatively late stage (approximately 4 days after knockdown) and were very wide. Therefore, it is necessary to identify the

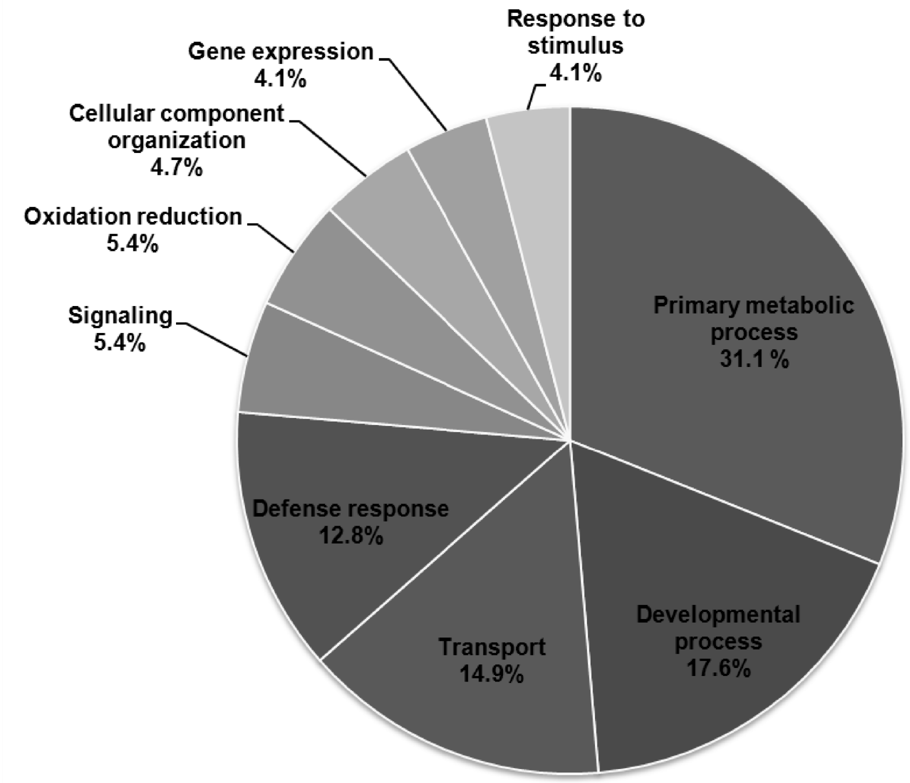


Figure 2.3. Pie chart representing the functional annotation of DEGs. Functional categories of differentially expressed genes in Table 1 were determined by analyzing the annotated functions in gene ontology database. The percentage of the category in each sector is marked.

primary effect by analyzing the DEGs detected at early stage. To identify the primary effect of SPS1 knockdown, temporal clustering and GO analysis were performed.

3.3. Construction of gene sets by temporal clustering

3.3.1. Clustering DEGs by SOM algorithm

To analyze the expression pattern of DEGs generated by *SPS1* knockdown, clustering of DEGs was performed according to their temporal expression using by self-organizing map (SOM) algorithms (Tamayo *et al.*, 1999). As a result, the DEGs were classified into 6 clusters (Figure 2.4). Genes belonging to cluster 1 (33 genes) showed continuous increase in their expression by *SPS1* knockdown, and most of them showed more than 4-fold increase on day 5. The expression patterns of genes in cluster 2 (77 genes) were similar to those of cluster 1, but the average expression level was lower than that of cluster 1. Genes in cluster 3 (9 genes) showed down-up patterns of expression. The expression of cluster 4 genes (12 genes) was decreased until day 3, and the expression level was maintained afterward. The expression pattern of genes in cluster 5 (27 genes) was a down-down type. Genes in cluster 6 (80 genes) showed an expression pattern similar to that of cluster 5 genes. However, the average level of expression of cluster 5 genes was much lower than that of cluster 6 genes.

3.3.2. Functional distribution of six clusters

We further examined how genes within each functional category distributed into the clusters. As shown in Figure 2.5, genes participating in metabolism seem to be

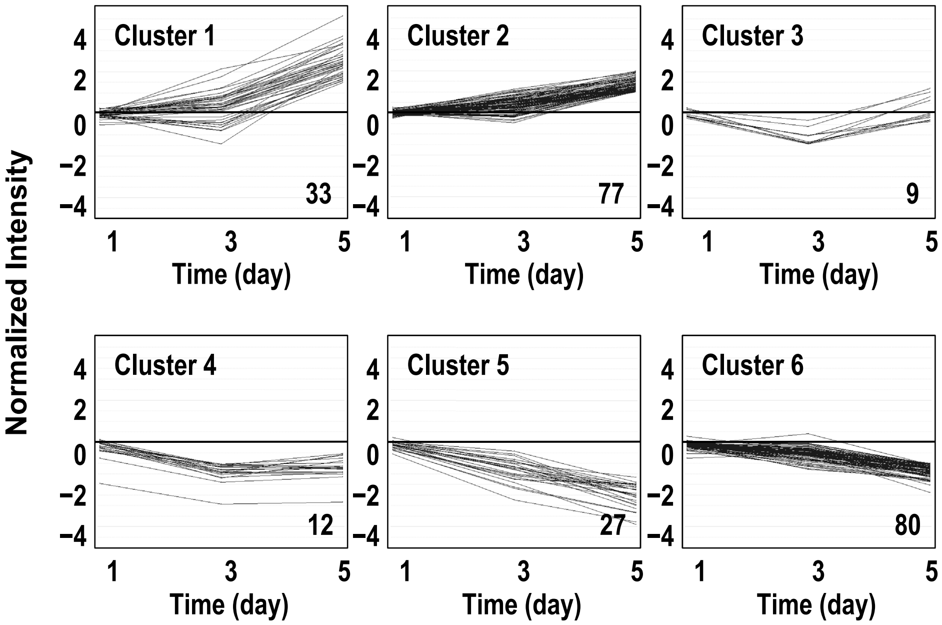


Figure 2.4. Result of SOM clustering. DEGs were classified into six clusters according to their temporal expression patterns using SOM clustering methods. The number in each panel represents the number of genes in each cluster. Normalized intensities are $\log_2$-values of signal intensity.

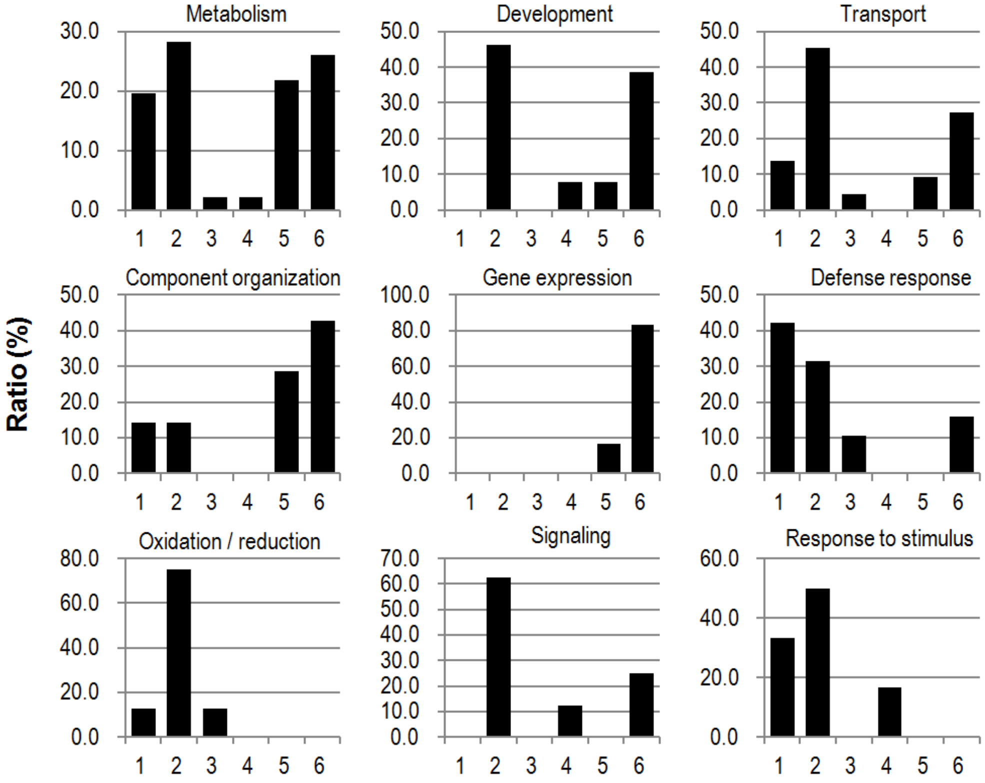


Figure 2.5. Comparison of the ratio of the number of genes in each cluster according to nine functional categoris. Distribution of genes consisting of each functional category. The y- and x-axis in each graph denote percentile of genes in each cluster and cluster number, respectively.

evenly distributed in all clusters, although the relative percentages are slightly different suggesting various genes were regulated in various ways. However, most of genes involved in development were distributed in cluster 2 and 6 suggesting these genes were either slightly up-regulated or slightly down-regulated. Majority of genes responsible for transport, oxidation/reduction, signaling, and response to stimulus are found in cluster 2 suggesting they were slightly up-regulated. Most of genes responsible for cellular component organization and gene expression are fallen into cluster 6 suggesting these gene were slightly down-regulated. Genes responsible for defense response are equally distributed in both cluster 1 and 2 suggesting these genes were up-regulated.

3.3.3. Construction of three gene sets for GO analysis

Using six clusters resulted from SOM clustering, the expression ratios of DEGs composing a cluster were drawn as a box plot according to their sampling date (days 1, 3, and 5). As shown in Figure 2.6, the median values (Q2s) of all clusters were close to zero on day 1. However, Q2s of clusters 3, 4 and 5 on day 3 were significantly decreased. On day 5, Q2s of clusters 1 and 2 were significantly increased, while those of clusters 4, 5 and 6 decreased. The interquartile ranges (IQRs) of each cluster were compared to select cluster(s) whose IQRs were significantly deviated. Clusters 3, 4 and 5 revealed significant down regulation compared to the other clusters on day 3. The IQRs of those clusters on day 3 were lower than −0.75. Therefore, the threshold to select clusters whose expression was significantly changed at a specific sampling date was set to the absolute value of 0.75 (see the dashed lines in Figure 2.6). A gene pool composing the selected clusters that showed the same expression pattern at the same sampling date was used as a gene-set for gene ontology analysis. As shown in Figure 2.6, there is no cluster showing that their IQRs were located at the outside of the threshold range (−0.75~ +0.75) on day 1; thus, no gene was selected for GO analysis from day 1 samples. However, on day 3, the IQRs of clusters 3, 4 and 5 were lower than the lower threshold (−0.75), and the genes in these clusters were defined as the early/down gene-set because their expressions were decreased. Clusters 1 and 2 showed a significant increase in their expression on day 5, and the genes in those clusters were defined as the late/up gene-set. On the other hand, genes in clusters 4, 5 and 6 showed significant down-regulation in their expression, and they were defined as the late/down gene-set (the dotted boxes in Figure 2.6; Table 2.3. for the list of genes in these gene-sets).

3.4. Identification of statistically overrepresented biological processes

To predict overrepresented metabolic pathway or biological process that is significantly affected by *SPS1* knockdown, gene ontology (GO) analysis (Ashburner *et al*., 2000) was performed with 3 gene-sets (early/down, late/up and late/down) previously defined using BinGO software (Maere *et al*., 2005). The parameters for statistical test and multiple testing corrections were used to binomical test and Bonferroni family-wise error rate (FWER) correction (Benjamini and Hochberg., 1995), respectively. As a result, total 29 GO biological process terms and 23 genes, which are included in each GO term, were selected. (Table 2.4). The terms related to vitamin B_6 biosynthesis were selected as significant GO terms from the early/down gene-set (p-value=2.48e-02). Changing the parameters for statistical tests and multiple testing corrections to Benjamini-Hochberg false discovery rate (FDR) (Benjamini and

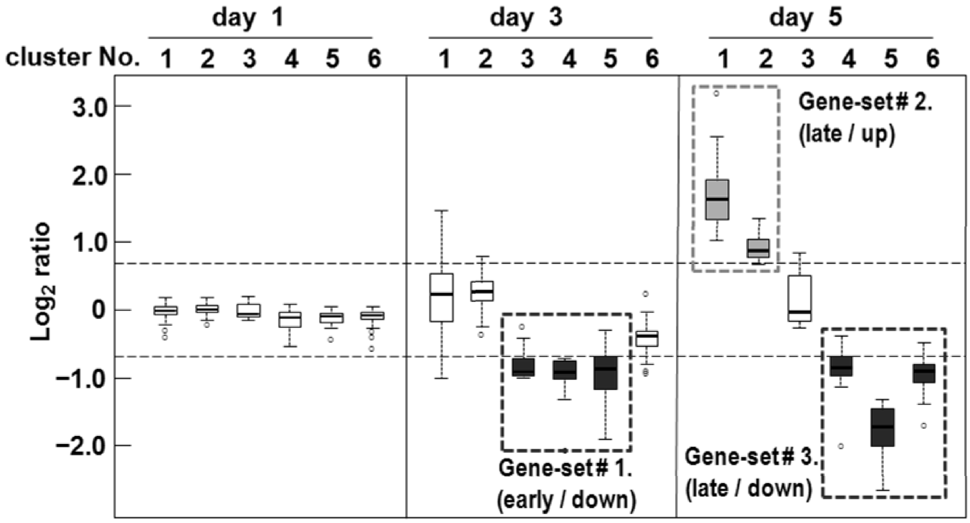


Figure 2.6. Construction of three temporally responded gene sets for GO analysis. The range of expression ratios of DEGs in each cluster was drawn with a box plot. The line in each box designates the median quartile (Q2). Dashed lines designate the threshold values (log$_2$ ratio of +0.75 and -0.75) for determining clusters of genes whose expressions were changed significantly. The dotted boxes represent the clusters showing their inter-quartile ranges (IQRs) and are the outliers of the threshold, and the genes in those clusters were selected as gene sets for GO analysis.

Table 2.3. Clusters of DEGs and gene-sets used for gene ontology analysis

				Early / down gene-set				
	Late / up gene-set				Late / down gene-set			
Cluster No.	1	2		3	4	5	6	
Gene No.	33 genes	77 genes		9 genes	12 genes	27 genes	80 genes	
Gene Symbol	AttB	Ac78C	CG5210	CG1962	argos	Atox1	AlkB	CG3950
	AttD	Adk3	CG7903	CG2065	aub	CG10424	armi	CG3960
	CecB	Anxb11	CG8031	CG31274	CG10824	CG11899	Arpc3B	CG3961
	CG13077	arg	CG8051	CG31975	CG13602	CG12014	CG10063	CG4330
	CG15526	bab2	CG8147	CG33468	CG15893	CG1218	CG10999	CG4338
	CG30484	bgm	CG9154	CG9733	CG17834	CG12643	CG12018	CG4502
	CG3085	Cat	CG9331	Gclc	CG31431	CG1572	CG12608	CG4615
	CG32170	CG10249	Cyp12a4	PGRP-SB1	CG5144	CG1753	CG12880	CG4953
	CG32185	CG12316	Cyp12d1-d	Pp2C1	CG9008	CG31472	CG12883	CG5126
	CG32625	CG13315	Cyp12d1-p		CG9338	CG32280	CG1311	CG6204
	CG32714	CG13654	Cyp12e1		pst	CG32985	CG13117	CG6659
	CG3348	CG13907	Cyp6a14		Tsp42El	CG33462	CG1344	CG7381
	CG4250	CG14196	Dro			CG34008	CG13795	CG7504
	CG4259	CG14253	drpr			CG3625	CG14079	CG7737
	CG5322	CG14629	Drs			CG4210	CG14229	CG7995
	CG6006	CG16888	egr			CG4398	CG14346	CG8157
	CG6018	CG17124	fz2			CG5535	CG14872	CG9471
	CG6231	CG17186	Gpdh			dare	CG15083	CG9641
	CG8745	CG2003	Gs1			gammaTub37(	CG15399	ChLD3
	CG9505	CG2794	Idgf1			Hen 1	CG15922	crb
	Cyp6a8	CG30046	Idgf2			Hexo1	CG1598	CTPsyn
	DptB	CG30059	Idgf3			l(2)k16918	CG1628	debcl
	eater	CG30456	if			Mctp	CG16790	dnt
	Ect3	CG3074	IP3K1			mthl5	CG17129	exu
	Mtk	CG31075	l(2)01810			Nnf1b	CG17181	GNBP3
	Oat	CG33521	LanB1			numb	CG17259	gnu
	ome	CG3792	nAcRbeta-21C			Pgm	CG17721	grk
	Pepck	CG3884	path				CG18528	Hexo2
	PGRP-LF	CG40498	Ror				CG2100	ImpL3
	PGRP-SD	CG4500	Ser				CG2177	kek1
	pirk	CG5630	Spn4				CG2200	ph-p
	Pxn	CG5793	Thor				CG2604	Rcd2
	wdp	CG5958	tok				CG2669	Spn42De
		CG6043	Toll-7				CG31142	squ
		CG6126	W				CG32043	Stlk
		CG6199	yellow-f				CG32521	Syx4
		CG6639	yin				CG32706	tamo
		CG7523					CG34033	Tgt
		Sesn					CG3570	Ude
		Spri					CG3770	wge

Table 2.4. List of biological process terms selected by GO analysis with three gene-sets

Discarded evidence codes : IEA
Overrepresentation
Selected statistical test : Binomial test
Selected correction : Bonferroni Family-Wise Error Rate (FWER) correction
Selected significance level : 0.05
Testing option : Use whole annotation as reference set

Gene-set	GO ID	p-value	corr. p-value	selected gene No.	genes No. in GO term	Annotated gene No. in gene-set	reference gene No.	GO term (Biological process)	Genes in test set
early/down	42816	5.29E-05	2.48E-02	2	3	26	7410	vitamin B6 biosynthetic process	*CG11899, CG31472*
	42819	5.29E-05	2.48E-02	2	3	26	7410	vitamin B6 metabolic process	*CG11899, CG31472*
	8615	5.29E-05	2.48E-02	2	3	26	7410	pyridoxine biosynthetic process	*CG11899, CG31472*
	8614	5.29E-05	2.48E-02	2	3	26	7410	pyridoxine metabolic process	*CG11899, CG31472*
late/up	9617	8.83E-11	6.22E-08	10	71	53	7410	response to bacterium	*CecB, PGRP-SD, Mtk, egr, pirk, Dro, AttB, DptB, AttD, Drs*
	42742	7.41E-10	5.22E-07	9	63	53	7410	defense response to bacterium	*CecB, PGRP-SD, Mtk, egr, Dro, AttB, DptB, AttD, Drs*
	6952	9.85E-10	6.95E-07	12	157	53	7410	defense response	*CecB, PGRP-LF, W, PGRP-SD, Mtk, egr, Dro, AttB, DptB, AttD, Drs, Toll-7*
	19731	1.14E-09	8.06E-07	7	27	53	7410	antibacterial humoral response	*CecB, Mtk, Dro, AttB, DptB, AttD, Drs*
	6955	6.69E-09	4.72E-06	11	147	53	7410	immune response	*CecB, PGRP-LF, W, PGRP-SD, Mtk, egr, Dro, AttB, DptB, AttD, Drs*
	9308	8.80E-09	6.21E-06	9	151	53	7410	amine metabolic process	*yellow-f, Gs1, PGRP-LF, W, PGRP-SD, arg, Oat, CG8745, Drs*
	2376	1.09E-08	7.68E-06	11	195	53	7410	immune system process	*CecB, PGRP-LF, W, PGRP-SD, Mtk, egr, Dro, AttB, DptB, AttD, Drs*
	51707	2.73E-08	1.93E-05	10	130	53	7410	response to other organism	*CecB, PGRP-SD, Mtk, egr, pirk, Dro, AttB, DptB, AttD, Drs*
	9607	4.15E-08	2.93E-05	10	136	53	7410	response to biotic stimulus	*CecB, PGRP-SD, Mtk, egr, pirk, Dro, AttB, DptB, AttD, Drs*
	19752	1.16E-07	8.17E-05	8	152	53	7410	carboxylic acid metabolic process	*yellow-f, Gs1, Cyp6a8, arg, Pepck, Oat, CG8745, Drs*
	43436	1.16E-07	8.17E-05	8	152	53	7410	oxoacid metabolic process	*yellow-f, Gs1, Cyp6a8, arg, Pepck, Oat, CG8745, Drs*
	6082	1.16E-07	8.17E-05	8	152	53	7410	organic acid metabolic process	*yellow-f, Gs1, Cyp6a8, arg, Pepck, Oat, CG8745, Drs*
	42180	2.74E-07	1.93E-04	8	167	53	7410	cellular ketone metabolic process	*yellow-f, Gs1, Cyp6a8, arg, Pepck, Oat, CG8745, Drs*
	51704	3.52E-07	2.48E-04	11	269	53	7410	multi-organism process	*CecB, W, PGRP-SD, Mtk, egr, pirk, Dro, AttB, DptB, AttD, Drs*
	19730	9.40E-07	6.63E-04	7	73	53	7410	antimicrobial humoral response	*CecB, Mtk, Dro, AttB, DptB, AttD, Drs*
	44106	9.46E-07	8.67E-04	7	107	53	7410	cellular amine metabolic process	*yellow-f, Gs1, W, arg, Oat, CG8745, Drs*
	50829	2.67E-06	1.88E-03	5	30	53	7410	defense response to Gram-negative bacterium	*CedB, Mtk, egr, Dro, Drs*
	6959	3.20E-06	2.26E-03	7	88	53	7410	humoral immune response	*CecB, Mtk, Dro, AttB, DptB, AttD, Drs*
	6519	3.54E-06	3.50E-03	6	128	53	7410	cellular amino acid derivative metabolic process	*yellow-f, Gs1, W, arg, Oat, CG8745*
	6950	5.60E-06	7.95E-03	12	485	53	7410	response to stress	*PGRP-LF, CecB, egr, AttB, AttD, W, PGRP-SD, Mtk, Dro, DptB, Toll-7, Drs*
	6520	3.54E-05	2.49E-02	6	86	53	7410	cellular amino acid metabolic process	*yellow-f, Gs1, arg, Oat, CG8745, Drs*
	50830	3.89E-05	2.74E-02	4	26	53	7410	defense response to Gram-positive bacterium	*CecB, PGRP-SD, Mtk, Dro*
	50832	5.17E-05	3.65E-02	4	28	53	7410	defense response to fungus	*CecB, Mtk, Dro, Drs*
	50896	1.72E-05	3.91E-02	15	1138	53	7410	response to stimulus	*CecB, PGRP-LF, egr, pirk, AttB, AttD, W, PGRP-SD, Cyp6a8, Mtk, Dro, DptB, Cyp12a4, Toll-7, Drs*
	9620	6.74E-05	4.75E-02	4	30	53	7410	response to fungus	*CecB, Mtk, Dro, Drs*
late/down	No significant biological processes are selected.								

Hochberg., 1995) and hypergeometric test did not change the results (Figure 2.7), suggesting vitamin B_6 biosynthesis is the only significant biological process affected by *SPS1* knockdown at the early stage. GO terms selected from the late/up gene-set could be categorized into two distinct biological processes: defense (immune) response and carboxylic acid (amino acid) metabolism (Figure 2.8). Both defense response (p=6.22e-08) and carboxylic acid processes (p=8.17e-05) were selected with significantly high probabilities. Interestingly, 15 genes among 21 genes (72%) selected from the late/up gene-set are known to participate in defense response (Table 2.5). In addition, 7 of 15 defense response genes encode antimicrobial peptide (AMP). No GO term was selected from the late/down gene-set. These results strongly suggest that SPS1 affects vitamin B_6 biosynthesis at the early stage and then defense response and amino acid metabolism through vitamin B_6 activity.

3.5. Validation of expression by quantitative PCR

Since the cells that were not transfected with double stranded RNA (dsRNA) as control for microarray analysis, it was necessary to confirm that the selected DEGs have the same expression pattern with the cells transfected with control dsRNA. We used GFP dsRNA as a control RNA and quantitative PCR (qPCR) was carried out to measure the expression levels. Of 23 DEGs from the selected GO terms (see Table 2.5 for gene list of selected GO terms), 15 genes were arbitrarily chosen, and their expressions were compared between *SPS1* knockdown and GFP dsRNA treated control cells. The addition of dsRNA of GFP did not show any significant differences in the expression of genes when compared to the negative controls to which no dsRNA was added. As shown in Figure 2.9, all tested genes showed the same pattern of

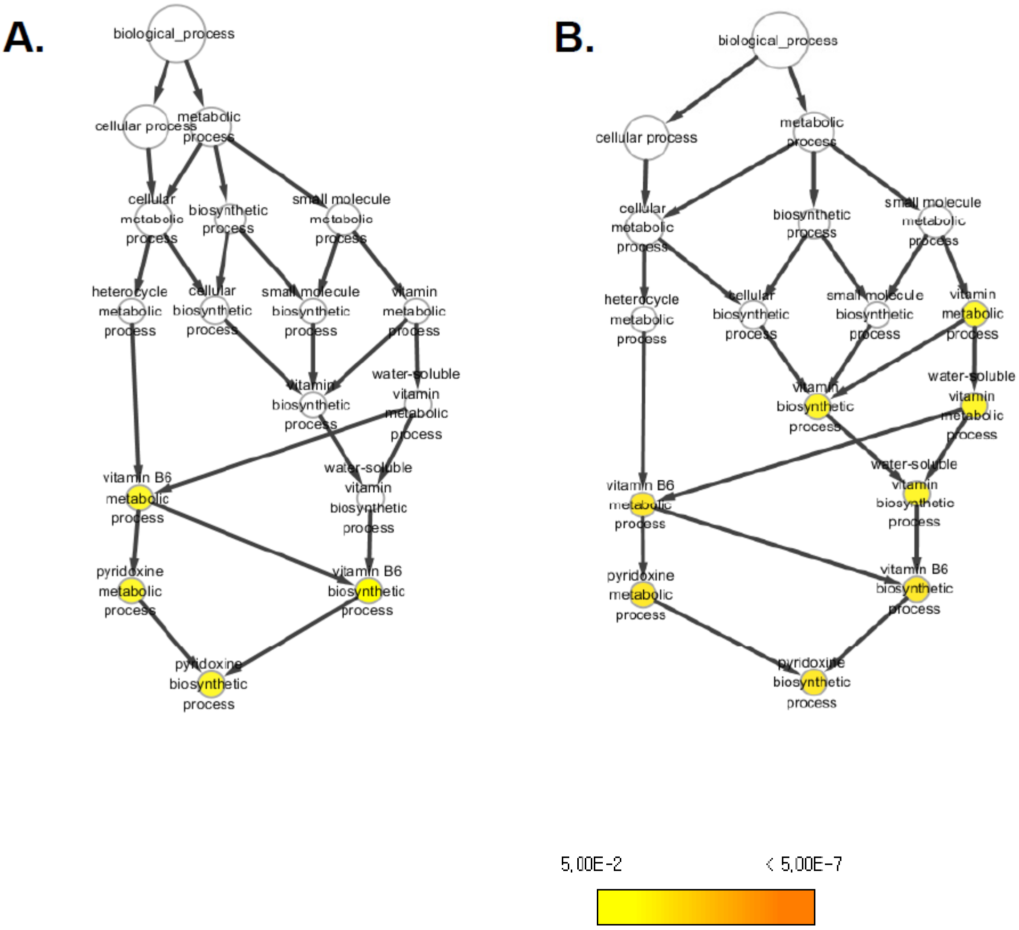

A.
biological_process
cellular process
metabolic process
cellular metabolic process
biosynthetic process
small molecule metabolic process
heterocycle metabolic process
cellular biosynthetic process
small molecule biosynthetic process
vitamin metabolic process
water-soluble vitamin metabolic process
vitamin biosynthetic process
vitamin B6 metabolic process
water-soluble vitamin biosynthetic process
pyridoxine metabolic process
vitamin B6 biosynthetic process
pyridoxine biosynthetic process
B.
biological_process
cellular process
metabolic process
cellular metabolic process
biosynthetic process
small molecule metabolic process
heterocycle metabolic process
cellular biosynthetic process
small molecule biosynthetic process
vitamin metabolic process
vitamin biosynthetic process
water-soluble vitamin metabolic process
vitamin B6 metabolic process
water-soluble vitamin biosynthetic process
pyridoxine metabolic process
vitamin B6 biosynthetic process
pyridoxine biosynthetic process
5,00E-2
< 5,00E-7

Figure 2.7. Hierarchical structures of GO terms obtained by performing early/down gene-set with different parameters. These Figures show examples of hierarchical structures of GO terms obtained by analyzing early/down gene set with BinGO software. Panels A were obtained by hypergeometric test and FDR for multiple corrections, and panel B were obtained by binomial test and FWER for multiple corrections. The results using the different parameters showed similar patterns.

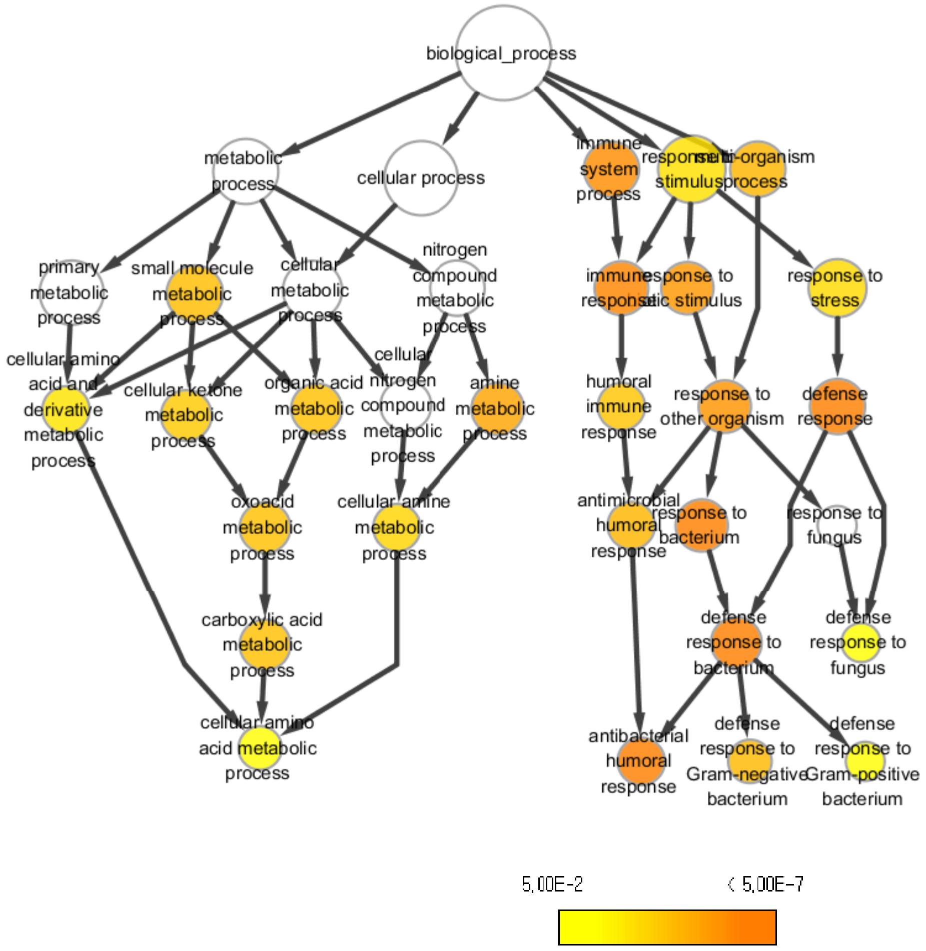

biological_process
metabolic process
cellular process
immune system process
response to stimulus
multi-organism process
primary metabolic process
small molecule metabolic process
cellular metabolic process
nitrogen compound metabolic process
immune response
response to biotic stimulus
response to stress
cellular amino acid and derivative metabolic process
cellular ketone metabolic process
organic acid metabolic process
cellular nitrogen compound metabolic process
amine metabolic process
humoral immune response
response to other organism
defense response
oxoacid metabolic process
cellular amine metabolic process
antimicrobial humoral response
response to bacterium
response to fungus
carboxylic acid metabolic process
defense response to bacterium
defense response to fungus
cellular amino acid metabolic process
antibacterial humoral response
defense response to Gram-negative bacterium
defense response to Gram-positive bacterium
5,00E-2
< 5,00E-7

Figure 2.8. Hierarchical structures of GO terms obtained by performing late/up gene-set. This figure shows the hierarchical structure of GO terms obtained by analyzing late/up gene set with BinGO software. This structure was obtained by hypergeometric test and FDR for multiple corrections.

Table 2.5. Representative biological processes selected according to the common hierarchical ancestor

Gene-set	Represented biological process	Max. corrected p-value	Selected genes
Early/down	Vitamin B6 biosynthesis	2.48E-02	*CG11899, CG31472*
Late/up	Defense response	6.22E-08	*AttB*, AttD*, CecB*, DptB*, Dro*, Drs*, Mtk*, egr, pirk, PGRP-LF, PGRP-SD, W, Cyp6a8, Cyp12a4, Toll-7*
	Carboxylic acid metabolism (Amino acid metabolism)	8.17E-05	*arg*$^{+}$, *CG8745*$^{+}$, *Gs1*$^{+}$, *Oat*$^{+}$, *Pepck, yellow-f*

Antimicrobial peptides (AMPs) are marked as ∗

Genes responsible for amino acid metabolism are marked as +

expression as that obtained from microarrays. It should be addressed that all the genes involved in vitamin B_6 synthesis and encoding AMP were tested and showed the same expression patterns. Interestingly, the level of expression between qPCR and microarray was significantly different in some AMP encoding genes such as AttB, CecB, Drs, DptB and Mtk, and in a gene encoding Pepck that is involved in carboxylic acid metabolism, although their expression patterns are similar between microarray and qPCR. The relative mRNA amounts measured by qPCR were more than three times higher than those obtained from microarray analysis. We assume that the probes on microarray against those genes were saturated by the transcripts and accordingly the values did not increase proportionally to the amount of RNAs. However, the saturation problem can be avoided by qPCR suggesting the values obtained by qPCR are more accurate. Therefore, the expression of those genes encoding AMPs was increased several hundred folds compared to that in control cell.

3.6. Intracellular pyridoxal phosphate level was decreased by SPS1 knockdown

Because GO analysis predicted that vitamin B_6 biosynthesis was the only pathway affected at early stage by *SPS1* knockdown and the expression patterns of genes involved in vitamin B_6 synthesis were confirmed, it can be speculated that levels of PLP will decrease by *SPS1* knockdown. To test this hypothesis, intracellular PLP levels were measured after *SPS1* knockdown. As shown in Figure 2.10, PLP levels in the cells where *SPS1* was knocked down decreased by approximately twofold compared to the control cells. The PLP concentration in *SPS1* knockdown cells was 37.23±0.66 pmol/mg protein. On the other hand, the PLP levels in the non-treated

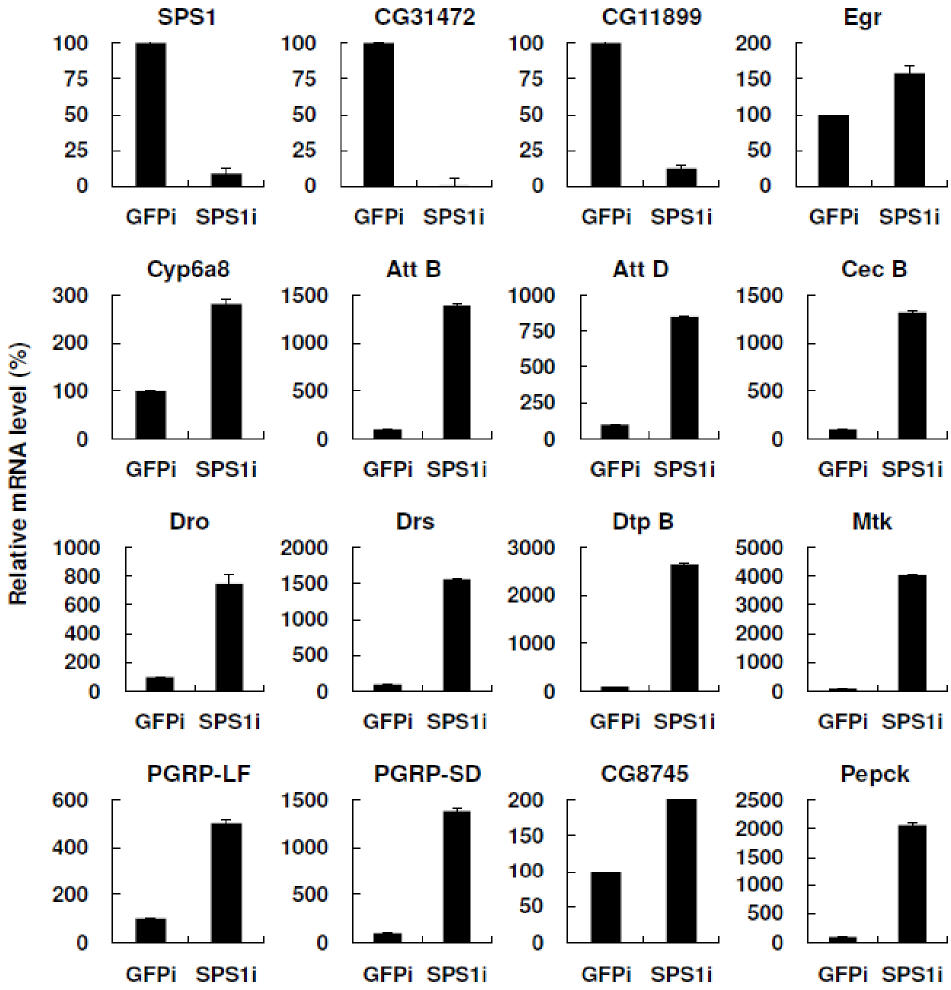


Figure 2.9. Validation of the selected genes by quantitative PCR. Five days after adding dsRNA, the mRNA levels of selected genes were measured by real time RT-PCR using rp49 for normalization. The y axis represents the relative mRNA level of each gene in the cells treated SPS1 dsRNA (SPS1*i*) to that treated GFP dsRNA (GFP*i*). The mRNA level of GFP*i* was set to 100%. The gene symbol is marked above each graph.

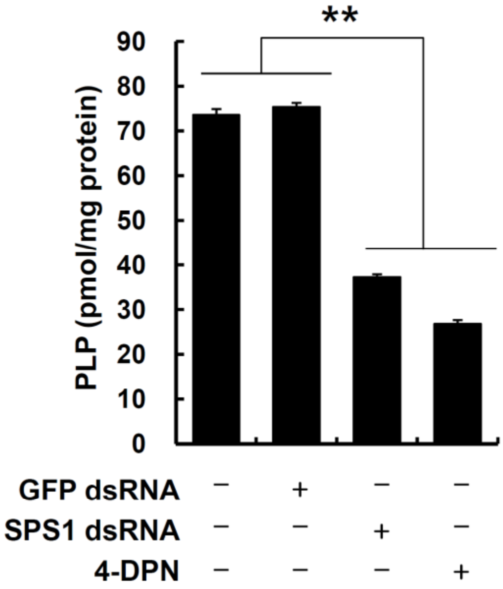


Figure 2.10. ***SPS1*** **knockdown causes a decrease in intracellular PLP levels.** Five days after SPS1 dsRNA or 4-DPN was added to the medium, intracellular PLP levels were measured as described in Methods. dsRNAs and 4-DPN used are shown on the x-axis. Experiments were performed in triplicate and error bars denote the standard deviation from the mean of three independent experiments. Statistical significance was tested by one-way ANOVA followed by Tukey's multiple comparison tests. ** indicates significance at $p< 0.01$.

control and in GFP dsRNA treated cells (negative control cells) were 73.59±1.31 and 75.37±0.89 pmol/mg protein, respectively. PLP levels in *SPS1* knockdown cells were similar to those observed in 4-deoxypyridoxine (4-DPN), which is an inhibitor of PLP biosynthesis, treated cells (positive control cells). These results indicate that the function of SPS1 is to regulate the biosynthesis of PLP in the cells, and also suggest that SPS1 can affect PLP requiring reactions such as amino acid metabolisms. It should be noted that some DEGs involved in amino acid metabolism were increased in their expression after SPS1 knockdown (see Table 2.5) suggesting the lack of PLP in the cell provides a signal for compensatory induction of some genes responsible for amino acid metabolism.

3.7. Inhibition of PLP biosynthesis and SPS1 knockdown showed similar expression patterns

Because intracellular PLP levels were significantly reduced after *SPS1* knockdown, it can be assumed that PLP biosynthesis is the primary target of SPS1, and the inhibition of PLP synthesis by treating cells with inhibitors will cause similar gene expression patterns as those resulting from *SPS1* knockdown. To test this hypothesis, *Drosophila* cells were treated with 4-DPN for 5 days, and the expression level of genes selected by GO analysis was measured with RT-PCR. As shown in Figure 2.11, the level of expression of the early/down genes (*CG31472* and *CG11899*) was not changed by 4-DPN treatment. Because 4-DPN inhibits only the function of proteins that participate in PLP synthesis and does not affect the expression of genes encoding those proteins, it is reasonable that 4-DPN does not affect the expression of *CG31472* and *CG11899*. However, the treatment of 4-DPN affected the expression of

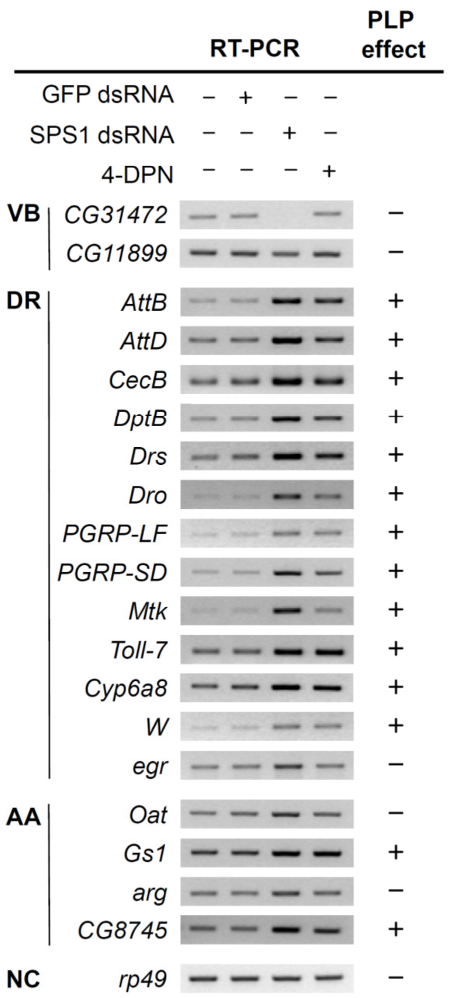


Figure 2.11. RT-PCR patterns of selected genes. Five days after dsRNAs and 4-DPN treatment, expression patterns of genes selected by GO analysis were measured by RT-PCR. Tested genes and GO terms of the gene are shown on the left of each panel. VB, vitamin B_6 biosynthesis; DR, defense response; AA, amino acid metabolism; Con, internal control. The effect of 4-DPN on gene expression is represented as the PLP effect. The + and – symbol designate consistency and inconsistency of expression pattern of each gene between *SPS1* knockdown and 4-DPN treated cells, respectively. rp49 was used as an internal control.

genes comprising the late/up and late/down gene-sets. Of the 17 genes tested, 14 genes (82 %) showed expression patterns similar to those observed by microarray analysis. It should be noted that the late gene-sets include genes responsible for defense response and amino acid metabolism. These results strongly suggest that PLP synthesis is the primary target of SPS1 and that intracellular PLP levels regulate other important biological processes such as defense system and amino acid metabolism.

3.8. The reduction of intracellular PLP level inhibits cell growth and induces megamitochondrial formation

In our previous study, we discovered that *SPS1* knockdown leads to cell growth inhibition and induction of megamitochondrial formation (Shim *et al*., 2009). As shown in Figure 2.12A, cell growth was significantly inhibited after the cells were treated with 4-DPN suggesting that the cell growth retardation induced by SPS1 knockdown was due to vitamin B_6 starvation. Another prominent phenotypic change induced by SPS1 knockdown is megamitochondrial formation. *Drosophila* SL2 cells were treated with 4-DPN for 3 days and examined under a confocal microscope after the mitochondria were stained with JC-1. As shown in Figure 2.12B, the cells treated with 4-DPN formed megamitochondria that were similar to those observed in the *SPS1* knockdown cells in terms of their size and number. Interestingly, the number of polar mitochondria (red dots in Figure 2.12B) in 4-DPN treated cells was similar to that in the control cells, and this mitochondrial polarity pattern was also similar to that observed in the *SPS1* knockdown cells. Since megamitochondria formation can arise from several different pathways, we examined whether megamitochondrial formation occurred by the activation of *Gs1* and *l(2)01810*. As shown in Figure 2.12C, both the level of *Gs1* and *l(2)01810* expression was increased. These results strongly suggest that the formation of megamitochondria, which is the most prominent phenotype from *SPS1* knockdown, is induced by the lack of intracellular PLP.

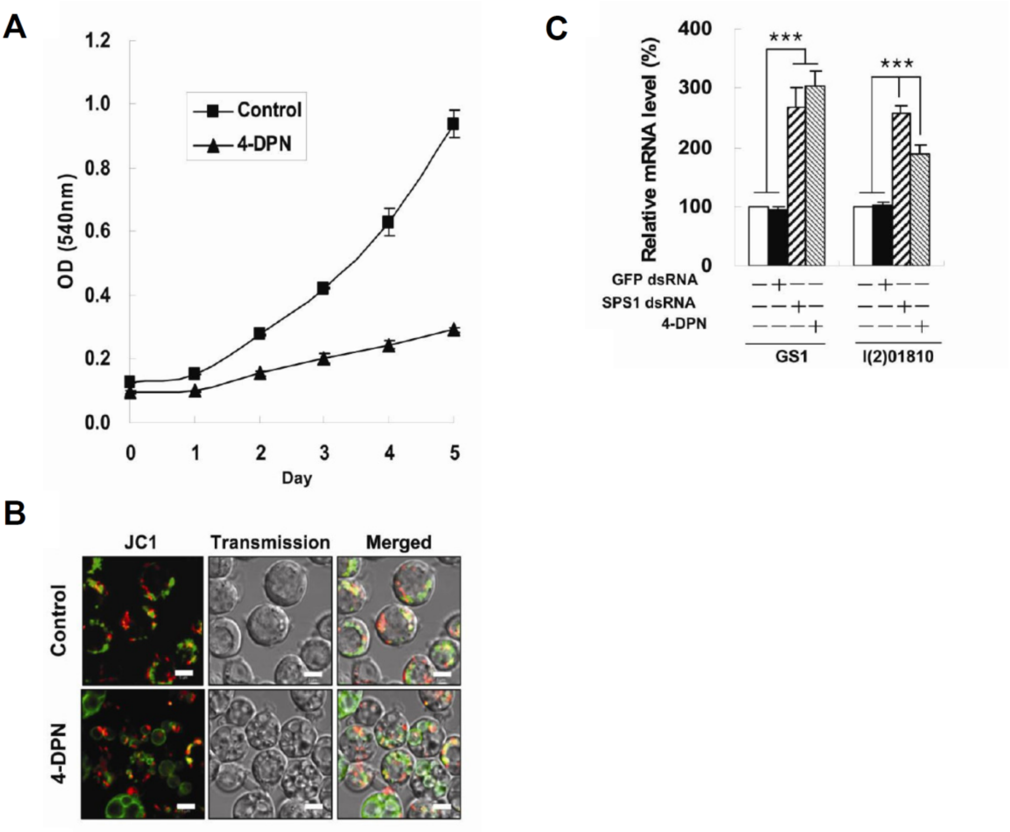


Figure 2.12. The effect of PLP synthesis inhibition on cell growth and megamitochondrial formation. **(A)** The growth rate of SL2 cells was examined by the MTT assay in 4-DPN-treated cells. Control cells were not treated with 4-DPN. Experiments were performed in triplicate, and error bars denote the standard deviation of three independent experiments. **(B)** Three days after 4-DPN treatment, cells were stained with JC-1 and then observed under a confocal microscope. Control cells were grown in the absence of 4-DPN. Scale bars represent 5 μm. **(C)** Five days after treatment of cells with dsRNAs and 4-DPN, mRNA levels of GS1 and l(2)01810 were measured by real-time RT-PCR. dsRNAs and 4-DPN treated are shown on the x axis. Statistical significance was tested by one-way ANOVA followed by Tukey's multiple comparison test. *** indicates significance at $p<0.001$.

4. DISCUSSION

We assumed that the genes whose expression was changed at the early stage after knockdown are involved in the primary target process regulated by SPS1. To identify the primary target, DEGs were isolated after microarray analysis and classified according to their temporal expression pattern; GO terms of early changed DEGs were analyzed using BinGO software. It is interesting that only PLP biosynthesis was predicted from the early/down gene set, even though the parameters were changed.

As shown in Table 2.5, the DEGs in the early/down gene set that are involved in vitamin B_6 synthesis are *CG31472* and *CG11899*. CG31472 is an ortholog of mammalian pyridoxine phosphate oxidase (PNPO), which catalyzes PLP production from PMP and PNP and PL production from PN or PM by oxidizing the substrates (Musayev *et al*., 2003). The function of CG11899 was not determined experimentally. However, it has high homology with mammalian phosphoserine aminotransferase and PdxC of *E. coli,* which are responsible for producing 4-phospho-hydroxy threonine, a precursor of the pyridoxine ring (Drewke *et al*., 1996). Therefore, it seems that CG11899 plays a role in producing precursors of vitamin B_6. Interestingly, intracellular PLP levels were decreased even though only two genes among four genes that are involved in the PLP biosynthesis pathway in *Drosophila* cell were down-regulated (Figure 2.13). This result suggests that these two genes are involved in an essential step of PLP biosynthesis, or SPS1 may also regulate the other proteins involved in PLP biosynthesis post-transcriptionally.

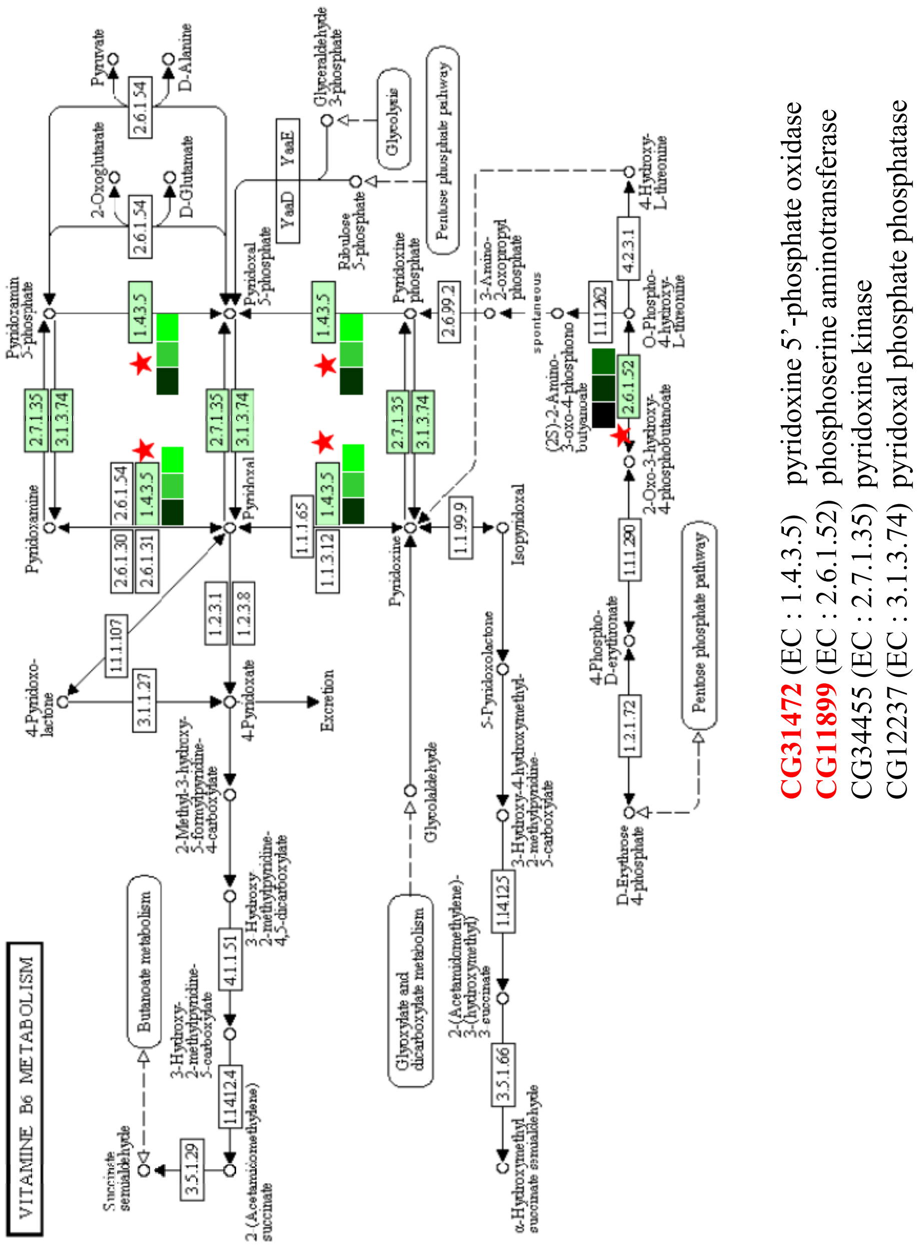

VITAMINE B6 METABOLISM
Pyridoxamine 5'-phosphate
Pyridoxamine
Pyridoxal 5'-phosphate
Pyridoxal
Pyridoxine phosphate
Pyridoxine
2-Oxoglutarate
D-Glutamate
Pyruvate
D-Alanine
Glyceraldehyde 3-phosphate
Glycolysis
Ribulose 5-phosphate
Pentose phosphate pathway
YaaD
YaaE
3-Amino-2-oxopropyl phosphate
spontaneous
(2S)-2-Amino-3-oxo-4-phosphonooxybutanoate
O-Phospho-4-hydroxy-L-threonine
4-Hydroxy-L-threonine
2-Oxo-3-hydroxy-4-phosphobutanoate
4-Phospho-D-erythronate
D-Erythrose 4-phosphate
Isopyridoxal
5-Pyridoxolactone
4-Pyridoxolactone
4-Pyridoxate
Excretion
Glycolaldehyde
Glyoxylate and dicarboxylate metabolism
Butanoate metabolism
2-Methyl-3-hydroxy-5-formylpyridine-4-carboxylate
3-Hydroxy-2-methylpyridine-4,5-dicarboxylate
3-Hydroxy-2-methylpyridine-5-carboxylate
2-(Acetamidomethylene)succinate
Succinate semialdehyde
3-Hydroxy-4-hydroxymethyl-2-methylpyridine-5-carboxylate
2-(Acetamidomethylene)-3-(hydroxymethyl)-3-succinate
α-Hydroxymethyl succinate semialdehyde
2.6.1.54
1.4.3.5
2.7.1.35
3.1.3.74
2.6.1.30
2.6.1.31
1.1.1.65
1.1.3.12
1.1.99.9
2.6.99.2
1.1.1.262
2.6.1.52
4.2.3.1
1.1.1.290
1.2.1.72
1.1.1.107
3.1.1.27
1.2.3.1
1.2.3.8
4.1.1.51
1.14.12.4
3.5.1.29
1.14.12.5
3.5.1.66
CG31472 (EC : 1.4.3.5) pyridoxine 5'-phosphate oxidase
CG11899 (EC : 2.6.1.52) phosphoserine aminotransferase
CG34455 (EC : 2.7.1.35) pyridoxine kinase
CG12237 (EC : 3.1.3.74) pyridoxal phosphate phosphatase

Figure 2.13. Schematic diagram of vitamin B_6 metabolic pathway. The original vitamin B_6 metabolic pathway diagram (collected from KEGG database) was modified by indicating DEGs and showing their expression levels after SPS1 was knocked down.

Because PLP is used as a cofactor for various enzymes that are important for many metabolic pathways, including amino acid metabolism, the inhibition of PLP biosynthesis will lead to the inhibition of cell growth. The inhibition of cell growth induced by *SPS1* knockdown seems to be mediated by a decrease in intracellular PLP levels. Specific inhibition of PLP synthesis by 4-DPN treatment led to growth inhibition (data not shown), suggesting the growth inhibition by *SPS1* knockdown is caused by down-regulation of PLP synthesis.

As described in the Results, down-regulation of genes responsible for PLP synthesis stimulated the expression of DEGs that participate in the defense response. In addition, most of the late gene-sets showed the same pattern of expression as that seen when cells were treated with 4-DPN (Figure 2.11). The relationship between vitamin B_6 and cellular defense, however, has not been demonstrated before this study. Previously, it was reported that the knockdown of *SPS1* induced diphthericin expression in *Drosophila* SL2 cell when a genome-wide knockdown was performed (Foley *et al*., 2004). The inhibition of PLP synthesis also induced the expression of various AMPs, including dipththericin. Therefore, SPS1 plays a key role in innate immune responses, including AMP production, by regulating PLP level in the cell. The mechanism by which vitamin B_6 regulates the innate immune system remains to be elucidated.

The fact that the treatment of 4-DPN, like *SPS1* knockdown, induced megamitochondrial formation indicates that intracellular glutamine levels increased with the inhibition of PLP synthesis. Because PLP is used as a cofactor for enzymes that have transaminase activity, it is reasonable to assume that low levels of PLP will lead to the inhibition of synthesis of amino acids such as glutamate or glutamine.

However, the inhibition of PLP biosynthesis induced the expression of *Gs1* and *l(2)01810* (Figure 2.12C). These two genes are involved in the increase of intracellular glutamine levels (Shim *et al*., 2009). Interestingly, *SPS1* knockdown also induced down-regulation of CG1753, which encodes cystathionine β-synthase (Table 2.2). Cystathionine β-synthase catalyzes both L-cystathionine and L-selenocysteine synthesis (Tamura *et al*., 2004). In cysteine metabolism, Cystathionine β-synthase catalyzes to produce L-cystathionine, which is a precursor compound to be L-cysteine, from L-homocysteine and L-serine as substrates (Banerjee, 2005), and cystathionine β–lyase converts L-cystathionine to L-homocysteine in cysteine catabolism, and also L-selenocystathionine to L-selenohomocysteine in selenocysteine metabolism (Mihara *et al*., 1997; Anderson *et al*., 1979; Flavin and Slaughter 1964). Selenohomocysteine is then further transformed into hydrogen selenide. PLP is an essential component of these two enzymes (Figure 2.14). Therefore, it seems that SPS1 regulates the synthesis of Sec indirectly by regulating the expression of Sec synthesizing enzymes. These results suggest that the lack of PLP in the cell provides a signal for compensatory induction of some genes responsible for amino acid metabolism. PLP regulation of the expression of *Gs1* and *l(2)01810* has not been elucidated.

A model for the molecular pathways regulated by SPS1 is summarized in Figure 2.14. SPS1 regulates the intracellular level of PLP by regulating the expression of genes responsible for PLP biosynthesis. Optimal levels of PLP do not induce defense response signaling and glutamine synthesis. However, low levels of PLP induce both defense signaling and glutamine synthesis. Once defense signaling is stimulated, genes responsible for the innate immune system, including AMPs, are activated. The activation of genes responsible for glutamine synthesis leads to

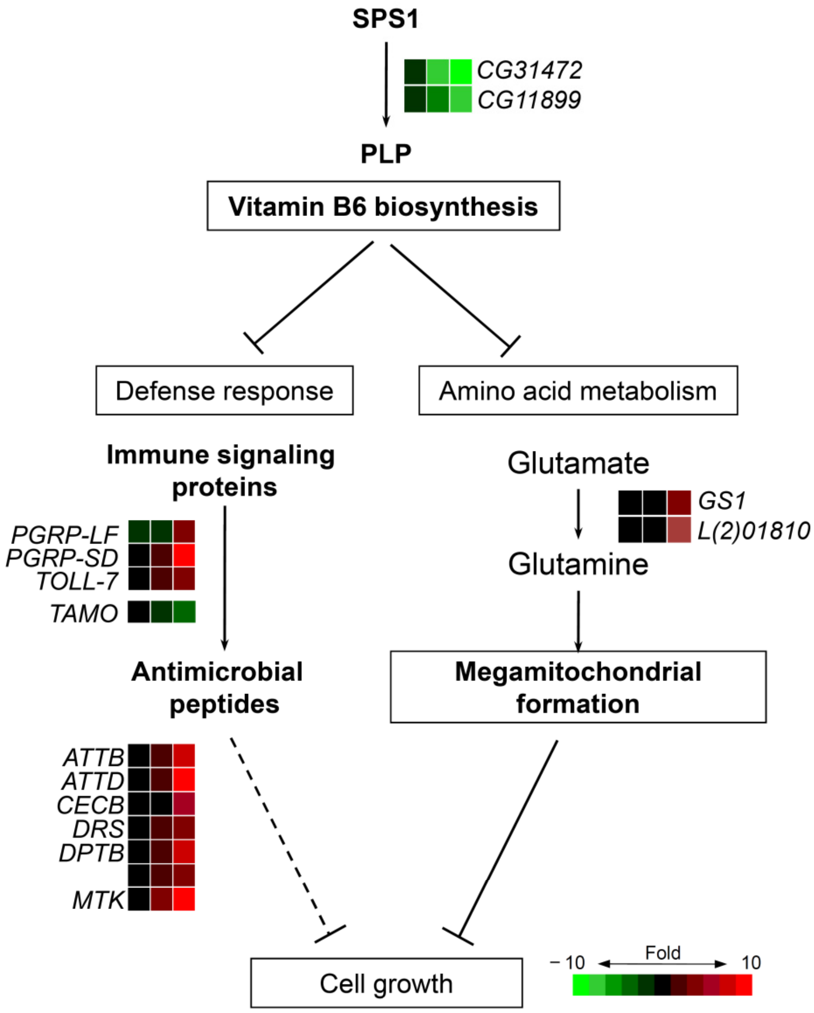


Figure 2.14. A hypothetical model for molecular pathways regulated by SPS1. A detailed explanation is provided in the Discussion. Molecular or cellular processes are marked with boxes. Proteins and molecules are in boldface letters. The expression levels of genes are marked with colors. The arrow and blocked line (┤) represent positive and negative regulation, respectively. The dashed line indicates that the effect was not proved experimentally.

megamitochondrial formation. The low level of intracellular PLP also leads to growth inhibition, presumably through induction of megamitochondrial formation and/or other biological processes. This hypothesis is supported by the observation of cell growth inhibition after the treatment of cells with 4-DPN (data not shown). However, it is not clear whether the growth inhibition is caused by the induction of both glutamine and AMP synthesis or one of these. In our previous study, it was found that conditions inducing megamitochondrial formation, such as the over expression of *GS1* and *l(2)01810*, also resulted in cell growth inhibition (Shim *et al*., 2009). But there is no report showing that the condition for the induction of defense system inhibits cell growth. Therefore, the inhibition of cell growth by AMP induction is represented as a dotted line in Figure 2.14.

Although SPS1 was found to regulate the biosynthesis of vitamin B_6, the mechanism or signal pathway to which SPS1 is related has not been determined. Because SPS1 is localized to both plasma and nuclear membranes (Kim *et al*., 2010), it can be speculated that SPS1 regulates signal transduction by transducing signals on the plasma membrane or by transporting messengers or transcription factors through the nuclear membrane. The treatment of cell with 4-DPN or *SPS1* knockdown induced the expression of *PGRP-SD* and *Toll-7*, which are involved in the Toll signaling pathway, and *PGRP-LF*, which is an activator of the IMD pathway (Figure 2.14). In addition, Tamo, which is a negative regulator for nuclear import of Dorsal, was found to be one of the down-regulated DEGs. These results strongly suggest that PLP, which is regulated by SPS1, participates in both the Toll and the IMD pathways.

CHAPTER 3.

IDENTIFICATION OF METHYLATION-DEPENDENT REGULATORY ELEMENTS FOR INTERGENIC MIRNAS IN HUMAN H4 CELLS

1. INTRODUCTION

MicroRNAs (miRNAs) are small, non-coding RNA molecules that act as post-transcriptional regulators of gene expression by inhibiting translation or degrading mRNA genes through partial or complete base pairing with complementary sequences of target genes (Bartel *et al.*, 2004). In addition, some miRNAs participate in the remodeling of chromatin structures (Yoo *et al.*, 2009). miRNAs are initially transcribed as large precursor RNAs, or primary miRNAs (pri-miRNA), and sequentially processed by Drosha and Dicer to produce ~22-nucleotide-long active mature miRNAs (Lee *et al.*, 2003; Hutvagner *et al.*, 2001; Ketting *et al.*, 2001). miRNAs are highly conserved in multiple organisms and play crucial roles in development, cell differentiation, determination of cell fate, and cancer (Alvarez-Garcia *et al.*, 2005; Croce *et al.*, 2005).

miRNA genes can be classified into two categories according to their genomic contexts: intronic and intergenic miRNAs (see Figure 1.21). Intronic miRNAs are embedded within other genes. Therefore, they are thought to be transcribed by sharing promoters with host genes (Rodriguez *et al.*, 2004). On the other hand, intergenic miRNAs are believed to have independent transcription units because they are positioned within flanking regions or in antisense orientation to annotated genes (Lagos-Quintana *et al.*, 2001). Intronic miRNAs are generally believed to be transcribed by RNA polymerase II (pol II); however, it remains unclear what type of RNA polymerase is responsible for intergenic miRNA transcription, although pol II and RNA polymerase III (pol III) are obvious candidates. For example, pri-miR-23a~27a~24-2 and pri-miR-21 are transcribed by pol II and have a 5’-7-

methylguanosine cap structure and a 3’-polyadenylated [poly(A)] tail similar to the structure of mRNAs (Lee *et al*., 2004; Cai *et al*., 2004), while miR-517a and miR517c, which are interspersed among Alu repeats in the human chromosome 19, are transcribed by pol III (Borchert *et al*., 2006).

The transcriptional start site of intergenic miRNA genes usually occurs within 2 kb upstream from the start site of miRNAs (Saini *et al*., 2007). Using computational methods, several conserved sequence patterns for intergenic miRNA genes, including putative promoters, have been proposed from various species (Zhou *et al*., 2007; Heikkinen *et al*., 2008). Among these, CT repeats are most well known. They are highly conserved among four species, such as *Caenorhabditis elegans*, *Homo sapiens*, *Arabidopsis thaliana* and *Oryza sativa*, and are abundant within 1000 bp upstream sequences from miRNA hairpins (Zhou *et al*., 2007). Another sequence pattern, GANNNNGA, was identified within 1000 bp upstream of worm miRNAs (Heikkinen *et al*., 2008). However, there is no direct evidence that these conserved patterns play a role as promoter or regulatory elements. Currently, the transcriptional mechanisms of most intergenic miRNAs are largely unknown.

Epigenetic signatures such as DNA methylation and histone modification, and the regulation of expression of miRNA genes are tightly linked similarly to other genes (Heintzman *et al*., 2007; Ozsolak *et al*., 2008; Suzuki *et al*., 2011; Toyota *et al*., 2008, see Figure 1.22). For example, chromatin signatures such as trimethylation of histone H3 at lysine 4 (H3K4me3) and acetylation of histone H3 at lysines 9 and 14 (H3K9/14Ac) are established as markers for transcriptionally active promoters (Li *et al*., 2007; Okitsu *et al*., 2010), whereas trimethylation of histone H3 at lysine 27 (H3K27me3) is characterized as a marker for transcriptional repression (Kahlil *et al*., 2009). Recently it was reported that hypermethylation of the human miR-124 loci, which is the most abundant miRNA in the adult brain and plays a key role in neurogenesis, inhibits miR124a expression and results in brain tumors (Agirre *et al*., 2009; Cao *et al*., 2007; Silber *et al*., 2008). Interestingly, some miRNAs control the expression of epigenetic regulators, including DNA methyltransferases and histone deacethylases (Fabbri *et al*., 2007; Noonan *et al*., 2009). The fact that miRNA gene expression can be regulated by DNA methylation indicates the feasibility of using methylated sequences to predict miRNA gene promoters or regulatory elements.

In this study, to identify the putative transcriptional regulatory elements for concering with the intergenic miRNA expresseion, we tried to analysis with various bioinformatical tools, such as motif search, for the data retrieved from methyl-binding domain (MBD)-chip array experiments. Then we found a novel sequence motif, C[N]$_6$CT, for intergenic miRNA gene expression by predicting sequence patterns in the differentially methylated regions (DMRs), and by examining the relationship between the occurrence of this motif and methylation dependence of gene expression.

2. MATERIALS AND METHODS

2.1. Cell lines and culture

H4 cells, a human neuroglioma cell line, were purchased from the American Type Culture Collection and cultured in Dulbecco's modified Eagle's medium (DMEM; Invitrogen, Carlsbad, CA, USA) supplemented with 10% fetal bovine serum (FBS; Invitrogen, Carlsbad, CA, USA) and 1% antibiotics-antimycotics (Invitrogen, Carlsbad, CA, USA) at 37°C in a humidified incubator containing 5% CO_2

2.2. Identification of miRNAs from sequence and annotation data

The genomic coordinates of 1,049 human miRNAs were obtained from the miRBase (ver. 16.0) (Griffiths-Jones *et al*., 2006), and all sequences and the annotated data were from the UCSC genome browser (http://genome.ucsc.edu). A total of 1,049 miRNAs were classified into 621 intronic and 428 intergenic miRNAs according their genome contexts.

2.3. Probe design

Sequences up to 1,000 bp upstream from the start site of 428 intergenic miRNAs were retrieved and cleaved into 60-bp-long sequences overlapped by 40 bp of adjacent sequence (Figure 3.1). Chopped sequences were filtered based on sequence redundancy, low GC ratios (GC ratio < 0.6), and low melting temperatures (Tm < 85°C). A total of 7,646 sequences were selected as probes for printing on an

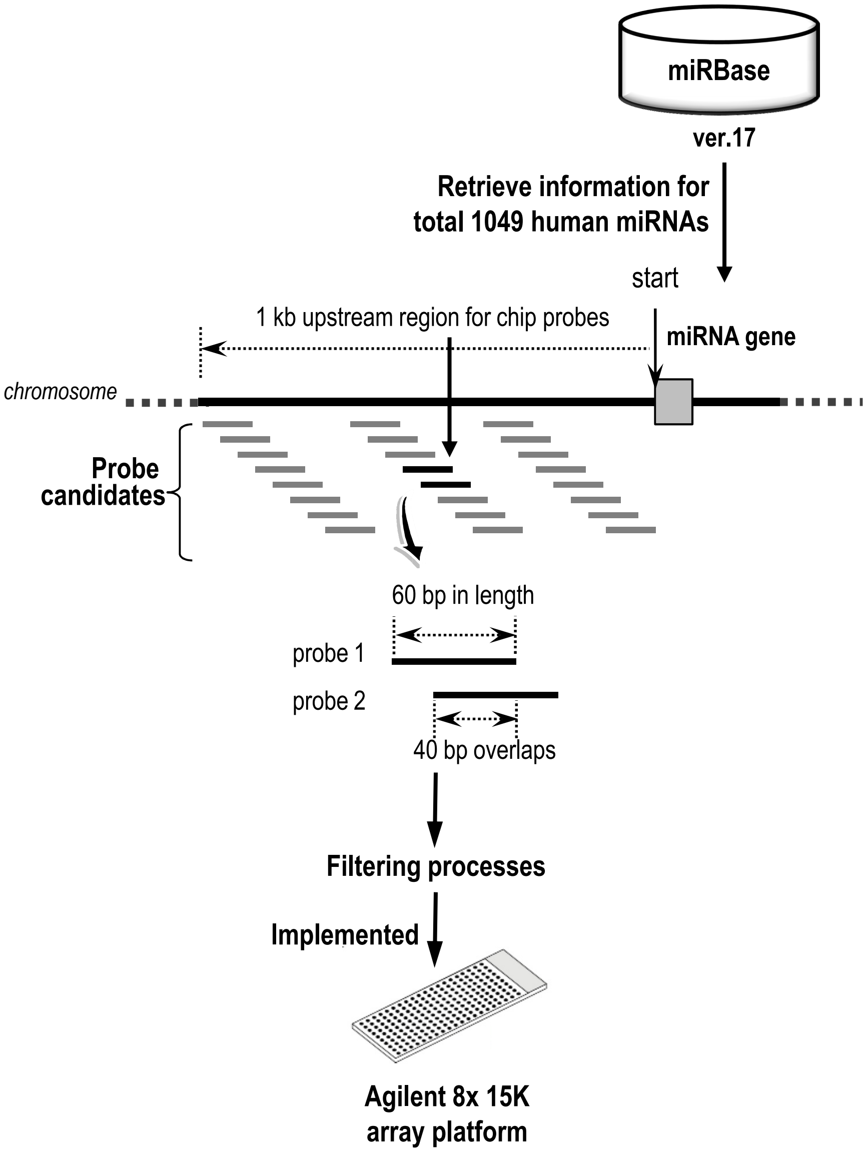


Figure 3.1. A schematic diagram of our custom designed probes. Probe sequences were retrieved from candidate regions and cleaved into 60-bp lengths with a 40-bp overlap with the adjacent probe. Probe candidates were selected by several filtering processes, and the final selected probes were implemented on Agilent 15K array platform.

Agilent 15K array platform to build a customized array chip (Chip No. 253347810001).

2.4. Microarray experiment

Genomic DNA was isolated from H4 cells cultured in the presence or absence of 5 µM 5-aza-2-deoxycytidine (DAC), an inhibitor of DNA methyltransferase. Briefly, after sonicating the genomic DNA (0.5 µg), the fragments were incubated with 2 µg recombinant methylation-specific binding protein (MBD2bt) at 4°C for 4 h on a rocking platform. The enriched methylated DNA was amplified using a Whole Genome Amplification Kit (GenomePlex®, Sigma-Aldrich, St. Louis, MO, USA) as recommended by the manufacturer's instructions. The amplified DNA from DAC-treated and untreated cells were labeled with cyanine 5 (Cy5) and cyanine 3 (Cy3), respectively. The labeled DNA samples were purified using a PCR Purification Kit (QIAquick, Qiagen, Valencia, CA, USA) and co-hybridized to the customized microarrays according to the manufacturer's protocol. The microarrays contained a total of 7,646 oligonucleotide probes, including control probes and those covering the sequences upstream of the miRNA genes (Figure 3.2).

2.5. Microarray data analysis

The hybridized images were analyzed using an Agilent DNA Microarray Scanner (Agilent Technology, Palo Alto, CA, USA) and data quantification was performed using Feature Extraction software version 10.7.3.1 (Agilent Technologies, Palo Alto, CA, USA). Preprocessing of raw data and normalization steps were

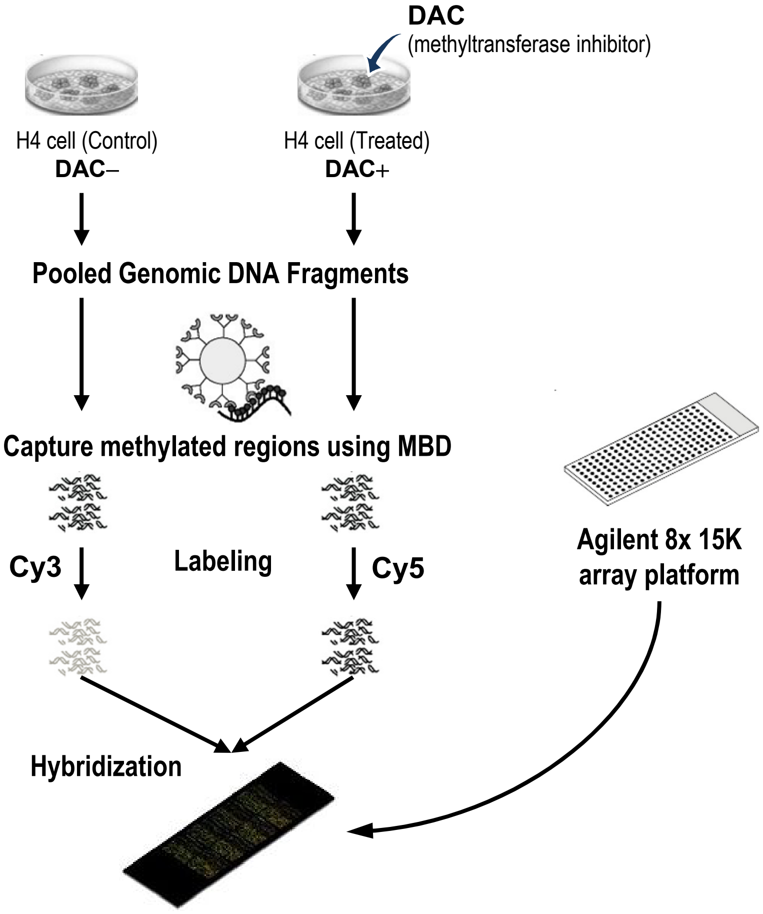


Figure 3.2. Design for identification of diffentially methylated probes using microarray experiment. After the addition of DAC, methyl-transferase inhibitor in H4 cells, fragmented genomic DNAs were captured by using methylation-specific binding (MBD) proteins. The enriched methylated DNAs were labeled with Cy3 (DAC-) and Cy5 (DAC+), and loaded on Agilent 15K array chips.

performed using R software (http://www.r-project.org). Background-corrected intensity data were normalized using the intensity-dependent LOWESS method to remove the dye bias within each array. The p-values for each probe were calculated using linear fit models implemented in the Limma package (http://bioconductor.org/packages/release/bioc/html/limma.html), and the probes within the threshold (p-value < 0.05) were selected as differentially methylated probes (DMPs). Since the probes were designed to 40-bp overlapped with each other, overlapped DMPs were constructed to be a contig, called differentially methylated regions (DMRs)

2.6. Distribution analysis

2.6.1.Predicted transcription factor binding sites (TFBSs)

Conserved transcription factor binding sites (TFBSs) were identified in the 1-kb region upstream of 1,049 miRNAs using PROMO (Messenguer *et al.*, 2002), a web-based program to predict putative TFBSs in DNA sequences. PROMO uses the TRANSFAC database to construct specific binding site weight matrices for TBFS prediction. Many TFs are provided as a list in the TRANSFAC database, enabling us to restrict the prediction to TF sets or to use all TFs. The general transcription factors TFIIB and TFIID, which recognize TATA box, and TFII-I, which binds to a pyrimidine-rich initiator (Inr) and a recognition site (E-box) for upstream stimulatory factor 1 (USF1), were used to predict the TFBSs.,

2.6.2. Predicted transcription start sites (TSSs)

TSSs were searched for in the same region as TFBSs using with the application implemented Eponine method (Down *et al*., 2002), a probabilistic method for detecting TSSs in mammalian genomic sequence. The default Eponine threshold of 0.990 was used because this threshold generally provides a useful number of predictions without sacrificing accuracy.

2.6.3. DMRs

Distribution curves were drawn with the number of DMRs, TFBSs, and TSSs calculated by scanning in a 40-bp window. The counted data were transformed to a standard normalized value because the number of probes for intragenic and intergenic miRNA, as well as predicted TFBSs and TSSs, were substantially different, making it difficult to compare them in a single plot. All analyses were performed using R software (http://www.r-project.org).

2.7. Motif analysis

2.7.1. Multiple alignments of DMRs and clustering

Alignment of DMR sequences was performed using ClustalX (ver. 2.0.12) software (Larkin *et al*., 2007) with the default parameters. A phylogeny tree was constructed based on the Neighbor-joining (NJ) model (Saitou *et al*., 1987) using MEGA5 (Tamura *et al*., 2011) with the following parameters: substitution model was set to maximum composite likelihood, missing data treatment was set to partial deletion, site coverage cutoff was from 100% to 95%, and the other parameters were set at default.

2.7.2. Identification of statistically significant motifs

MEME (Bailey *et al*., 2006) was used to search for top-ranking degenerate motifs within the probe sequences in each cluster, and its optional parameters were set to as follows: optimum motif width was set to 8-12 bp, occurrence of motif in the input sequences was set to any number of repetitions in the input sequence, and others parameters were left as default. The MEME algorithm calculates the significance of each identified motif against random sequences with the same nucleotide composition, and assigns an E-value based on the probability of obtaining the observed number of motifs.

2.8. Reverse transcription PCR reaction

Total RNA was isolated using the RecoverAll, Total Nucleic Acid Isolation Kit (Ambion, Austin, TX, USA), according to the manufacturer's protocol. RNA quantity and purity was determined using the NanoDrop 1000 spectrophotometer (Thermo Scientific, Rockford, IL, USA). Reverse transcription was performed using the TaqMan MicroRNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA). All reactions were performed as per the manufacturer's protocol. *Rnu6b* was used as a negative control.

2.9. Quantitative real-time PCR

Quantitative real-time PCR was performed to amplify miRNAs with specific primer sets against target miRNAs (Applied Biosystems, Foster City, CA, USA) using the ABI-7500 Real Time PCR system according to the manufacturer's protocol. Template (10 ng) was amplified in 20 µl reaction volumes. PCR conditions were as follows: 50°C for 2 min, 95°C for 10 min, 50 cycles of 95°C for 15 s, and 60°C for 1 min. Experiments were performed in triplicate.

2.10. miRNA targets prediction

Target genes for selected intergenic miRNAs were predicted by using miRanda open source software (http://www.microrna.org/microrna/) associated with the miRBase with a cutoff p-value less than 0.05. Common target genes of selected intergenic miRNAs were extracted and input for GO analysis.

2.11. Gene ontology analysis

GO analysis was performed by BiNGO version 2.3 (Maere *et al.*, 2005), which is plugged in Cytoscape (Shannon *et al.*, 2003). Gene symbols of predicted target genes were used as input data. The parameters were set as follows: assessment was set to overrepresentation, statistical test to binomial test, multiple testing correction to FDR correction, significance level to 0.05. The most significant pathway was predicted by considering the selected GO terms and visualized output.

2.12. Analysis of bisulfate sequencing data

The raw data of bisulfate sequencing were obtained from GEO database (Accession number: GSE15007). The number of three types of motifs within reads, C[N]$_6$CT, T[N]$_6$CT, and C[N]$_6$TT, were counted by a customized Perl scripts. Mapping of bisulfate sequencing reads was performed with chromosome 12 and 20 using BSMAP web tools. The position of C[N]$_6$CT motif in chromosome 12 and 20 were determined whether the reads corresponding to the bisulfate-converted strand or the reverse-complementary strand.

3. RESULTS

3.1. Identification of DMRs

To identify DMRs in the upstream of miRNA genes, microarray analysis was performed using our custom chips. A total of 7,646 probes against the 5'-flanking region of 428 intergenic miRNA genes were designed (Figure 3.1) and implemented on the Agilent 15K array chip platform. Genomic DNA was isolated from H4 cells cultured in the absence or presence of 5-aza-2-deoxycytidine (DAC). Methylated sequences were enriched using MBD2bt proteins. The methyl group-enriched DNAs were labeled with fluorescent dyes and then hybridized with the probes on chips (see 2. Materials and Methods in this chapter).

The signal intensities of the 7,646 probe spots on each chip were obtained after LOESS normalization. The Pearson's correlation of signal intensities between chips was 0.99 (Figure 3.3A). By performing a linear fitting and Bayes function analysis, a total of 161 probes (adjusted p-value < 0.05) were found to have different methylation levels between DAC-treated and untreated samples (Figure 3.3B) and these were defined as differentially methylated probes (DMPs). These DMPs were derived from 98 intergenic miRNAs. The sequences of the DMPs are shown in Table 3.1.

3.2. Distribution of DMRs, TFBSs and TSSs

To determine the distribution of the DMPs on the 5'-flanking region of each gene, we calculated the relative distance of DMPs from the 5'-end of each intergenic

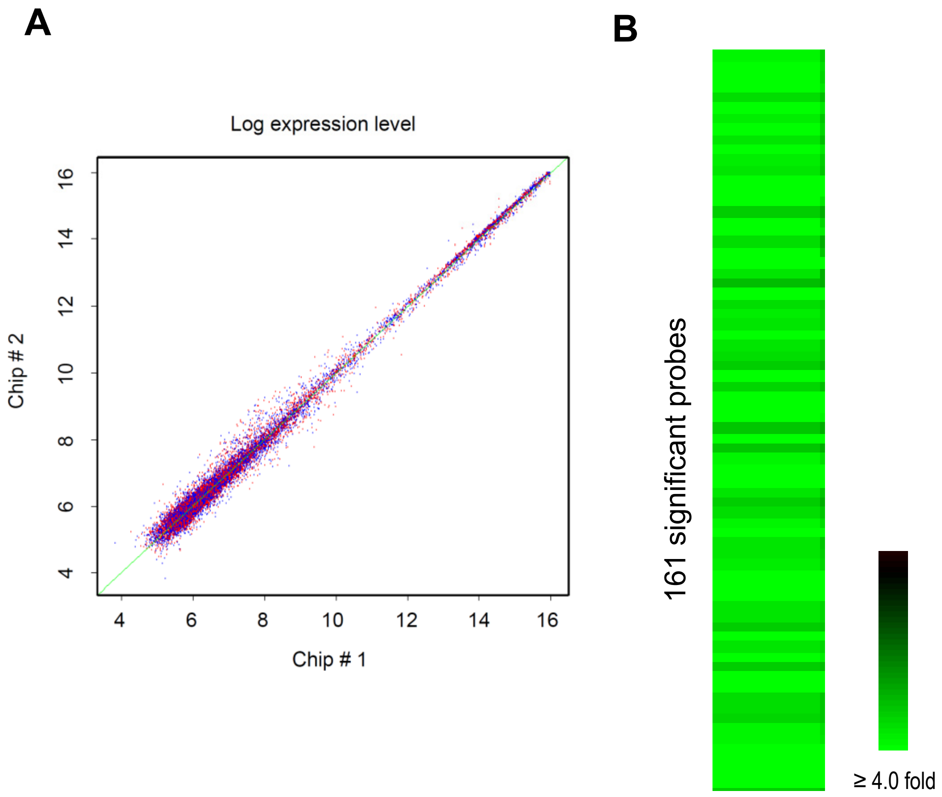


Figure 3.3. Correlation between replicated chips and selected differentially methylated probes. (A) Pearson correlations between replicated chips (r=0.99) (B) Profiles of 161 differentially methylated probes (DMPs) identified from the 5'-flanking regions of 98 intergenic miRNAs. Green color designates hypomethylated probes, while black color represents non-methylated.

Table 3.1. List of DMPs of intergenic miRNAs selected from microarray analysis

probe ID	start	end	probe sequences	adj.P-Value	miRNA name
NIH043068	52,195,979	52,196,038	atccctggtcctcctggtccctgtctgtctgtctgtcgggtctgtccacctgccgcgccc	1.65E-04	hsa-let-7e
NIH043069	52,195,999	52,196,058	ctgtctgtctgtctgtcgggtctgtccacctgccgcgccccccgggctgaggtaggaggt	4.45E-03	hsa-let-7e
NIH027954	62,996,846	62,996,905	gccggcggcagcctctgcgggatgcccgccggcgattatcggcgccgagagcgggccggc	1.91E-02	hsa-let-7i
NIH027975	62,997,266	62,997,325	tctcatccccgtagctctcggggcgcgccgggagtagccgagccgctgggtgctgggcgg	1.67E-03	hsa-let-7i
NIH032671	101,508,614	101,508,673	ccagagcacccagtgtggctacaggacttgctgctgggacctctaagccccactgctctc	3.87E-03	hsa-mir-1185-1
NIH032714	101,509,655	101,509,714	cctcttctccataccaagtcatacccagagggcttcctgcgtccagcactaacctcgcac	5.78E-03	hsa-mir-1185-2
NIH032236	101,496,289	101,496,348	cggtgggtggcgcctcttcagtgtgagatcttgcttggatggccttctggacacgctagc	3.97E-05	hsa-mir-1193
NIH031906	101,491,441	101,491,500	gcttcccagagagaggcaaagtcaccttcgtgggagacacactgatctggcttcagcgcc	4.96E-05	hsa-mir-1197
NIH019775	129,162,182	129,162,241	cccagcacttcctcttctccctccctgtgtcccactgccactgagtgaagccacactcgg	4.29E-02	hsa-mir-1208
NIH019776	129,162,202	129,162,261	ctccctgtgtcccactgccactgagtgaagccacactcggtccgactggcggatctttct	4.44E-03	hsa-mir-1208
NIH017983	9,761,863	9,761,922	ccacctccctcgctcccgccctccctgctgcattttgcatgagctatctaggtcattagc	9.38E-03	hsa-mir-124-1
NIH018000	9,761,464	9,761,523	gaggtaaagaggcggcggccgcgcgccctgccccggctgccgctgccgagtccccgctgc	6.58E-03	hsa-mir-124-1
NIH018002	9,761,424	9,761,483	cgctgccgagtccccgctgctctccttgtccttcgctctcttcttcctcctcctagtccc	1.29E-02	hsa-mir-124-1
NIH018012	9,761,224	9,761,283	gtaattaacacgggggaggcacccctccgtctcccacttccacccacacccccatccctc	2.91E-02	hsa-mir-124-1
NIH018020	9,761,064	9,761,123	gggggagctgcggcggggaggatgctgtggtcccttcctccggcgttccccacccccatc	7.49E-03	hsa-mir-124-1
NIH018022	9,761,024	9,761,083	cggcgttccccacccccatccctctccccgctgtcagtgcgcacgcacacgcgccgcttt	1.29E-02	hsa-mir-124-1
NIH027295	9,392,428	9,392,487	gggcagtgccacccgcagatgacacgcgctctccaccacccaacccaaaccatgagaatt	4.30E-02	hsa-mir-1244-3
NIH033865	102,027,240	102,027,299	gtggccccggctcccgccccaccgaggacgcgggccagcgaggaggcaactttgcgttcc	1.55E-02	hsa-mir-1247
NIH033884	102,026,860	102,026,919	gagctggggaaagcggcgccggccctgcgcccccagccgacccgggcgcggggcgccacg	4.68E-02	hsa-mir-1247
NIH033885	102,026,840	102,026,899	ggccctgcgcccccagccgacccgggcgcggggcgccacgggcgtggggcgcggggcagc	1.10E-02	hsa-mir-1247
NIH043095	52,195,967	52,196,026	gccctccccctcatccctggtcctcctggtccctgtctgtctgtctgtcgggtctgtcca	9.98E-05	hsa-mir-125a
NIH043096	52,195,987	52,196,046	tcctcctggtccctgtctgtctgtctgtcgggtctgtccacctgccgcgcccccgggct	1.25E-03	hsa-mir-125a
NIH043119	52,196,447	52,196,506	tgtgcctatctccatctctgacccccaccccagggtctaccgggccaccgcacaccatgt	1.71E-02	hsa-mir-125a
NIH031492	101,348,816	101,348,875	gaactggctcttgtccagggagcagtagacgttgtgatggcggaagcggaccaggacttg	5.10E-02	hsa-mir-127
NIH05054	70,480,378	70,480,437	cagtgattgtggtactgcattccagtctgggcacagagtgagactctgccccccaacctc	1.04E-02	hsa-mir-1285-2
NIH033199	101,520,564	101,520,623	caggatggtggttggcagccactccccttggagaagtggaaggggactccttgtctgtct	6.62E-04	hsa-mir-134
NIH07409	44,155,144	44,155,203	ttggaggtttggggaccctggacagcatacaaatggaggggtgaggggcagccgagggaa	4.39E-03	hsa-mir-138-1
NIH027244	7,072,940	7,072,999	tgtgtcagcaacatccatcgcctcaggtccccagcccttagctggctgcagccccctccc	1.59E-02	hsa-mir-141
NIH027256	7,072,960	7,073,019	cctcaggtccccagcccttagctggctgcagccccctccccacttcccacgcaccccgga	4.97E-02	hsa-mir-141
NIH027257	7,072,980	7,073,039	gctggctgcagccccctccccacttcccacgcaccccggaagcccctcgtcttgagctga	2.77E-02	hsa-mir-141
NIH027258	7,073,000	7,073,059	cacttcccacgcaccccggaagcccctcgtcttgagctgagagcgttgcacaaggggtgg	5.67E-03	hsa-mir-141
NIH013119	159,911,859	159,911,918	ttccaggccagagggatggcatatggaagggtcatgaggcaggaaaggccagctaccatg	7.11E-04	hsa-mir-146a

probe ID	start	end	probe sequences	adj.P-Value	miRNA name
NIH015537	25,990,427	25,990,486	ggaagatgggaaagcactttccagacctgttgcacgcgcgcaacagctgttcaggtgcgg	7.45E-03	hsa-mir-148a
NIH015549	25,990,187	25,990,246	gtggaacggaggggatgggacgacttcgacccgagttcccggggctgggcgcgggcggct	1.42E-02	hsa-mir-148a
NIH041934	13,984,513	13,984,572	gggctcaaaagatcctcccgcctcagcctcccaaaatactgaattacaggcctgagccgc	3.56E-02	hsa-mir-181c
NIH041964	13,985,113	13,985,172	gcatcaatgccctcggggccaggatcctccgttgccagacttcccgggaacacttggaat	2.33E-04	hsa-mir-181c
NIH041966	13,985,153	13,985,212	ttcccgggaacacttggaatgcctcctccacctcggactagcagtggcctttggcctcag	1.66E-03	hsa-mir-181c
NIH041968	13,985,193	13,985,252	gcagtggcctttggcctcagtctccccagttgacaaagggggtaatctgcacctccagg	3.96E-02	hsa-mir-181c
NIH041972	13,985,273	13,985,332	aggagcgggcttgaggccagcactcccctgcactgctacatctccatccccatagcaaag	3.53E-02	hsa-mir-181c
NIH041995	13,984,849	13,984,908	gctgcgcggactccctccggaagtgcccgagttcagatgctgtgtgaccgcgcacgcggg	8.03E-04	hsa-mir-181d
NIH017133	129,410,413	129,410,472	ctccacaccagggcgaccctgcaggaaggaccttgtcgcagttgcggggatgggcgcctc	1.68E-02	hsa-mir-182
NIH028732	100,583,022	100,583,081	tccccactcgacccaggaagtccagctggcttcacctcccaccaggaggccaaacaaaaa	4.29E-02	hsa-mir-1827
NIH028733	100,583,042	100,583,101	tccagctggcttcacctcccaccaggaggccaaacaaaaaggacagtgctgagaggaacc	1.47E-04	hsa-mir-1827
NIH038732	29,886,095	29,886,154	ccgccgtcagggtctctcccgccggtgccgcaggctggagcgggcgtcgcggccgggggc	2.57E-02	hsa-mir-193a
NIH038733	29,886,115	29,886,174	gccggtgccgcaggctggagcgggcgtcgcggccgggggcggtgccgcctaggggcgggc	5.29E-03	hsa-mir-193a
NIH038738	29,886,215	29,886,274	ggaggcggtgcgggcgggcgctggccgcgtccgcccgccgggagccgcggggcgcgcagc	1.59E-02	hsa-mir-193a
NIH038769	29,886,835	29,886,894	tcgtgtaacccttggagggctgggtttggagcccgcgacccgaggtcgggcggggcggtg	1.06E-02	hsa-mir-193a
NIH036289	14,396,844	14,396,903	ctccacgtgcgcccgagccccgctggcatgggcggcggcgcctgcgggggccgggccccg	3.13E-02	hsa-mir-193b
NIH062	1,101,624	1,101,683	tgagggctgaacctccctcgggtcctgagtgtgcctggagtagaagcctagggtctctgg	1.14E-04	hsa-mir-200b
NIH081	1,102,004	1,102,063	cgttctgtctcgagagcctcgcagacaccgggcctttgagaagagaaggggctgggcagg	3.84E-02	hsa-mir-200b
NIH027162	7,071,922	7,071,981	gcctgcccttcacaggcctggcggggccttctccctcccctttccctcaggggatcccag	1.67E-02	hsa-mir-200c
NIH027163	7,071,942	7,072,001	gcggggccttctccctcccctttccctcaggggatcccagcacaggctgggcactgcggg	8.60E-05	hsa-mir-200c
NIH027165	7,071,982	7,072,041	cacaggctgggcactgcggggcggcacagcccaactcctgcccagctgacccctcgctga	2.32E-04	hsa-mir-200c
NIH027166	7,072,002	7,072,061	gcggcacagcccaactcctgcccagctgacccctcgctgacctcagggatctctctggcc	6.58E-03	hsa-mir-200c
NIH027168	7,072,042	7,072,081	cctcagggatctctctggcctgcagctccgctgtgggcagggtctgaggccacagaggaa	1.87E-04	hsa-mir-200c
NIH027170	7,072,082	7,072,141	ggtctgaggccacagaggaatgggctagtcctgggggcagcatctgctgtggggaggga	2.46E-02	hsa-mir-200c
NIH027188	7,072,442	7,072,501	aatctggggccttaaagccccttcgtctccccagcacccactctctgggggcaggtgggc	3.55E-02	hsa-mir-200c
NIH033971	104,583,202	104,583,261	ccgggaggccaggtgcgcccaggccaggcgctggaggctggggcgccggatggggcgggg	1.14E-02	hsa-mir-203
NIH024836	568,499	568,558	ccccctcgcgcccactggctgcgttgcggtagggcgggcccggctaggccggccattggc	3.44E-02	hsa-mir-210
NIH024865	568,439	568,498	tgagggaccaggtcatttgcatacggggctgggcgtggagccgcggggcccggggccggt	1.24E-02	hsa-mir-210
NIH024873	568,279	568,338	gatcccaggttggcgggcggggcgcccctcagaggccgccctcccgcgggggctgcgggg	1.22E-02	hsa-mir-210
NIH024877	568,199	568,258	gcggggctcggacgcccaagttggaggggacgggggtggggtcaatccctccgcccgccc	4.62E-02	hsa-mir-210
NIH038945	41,521,354	41,521,413	ccccattgactggggtggggccctggtgtctactccctgataaagaccacgtatgcctgg	6.04E-03	hsa-mir-2117
NIH037911	1,954,315	1,954,374	acgggaggacgaggaggggacgggcaggcggcgggtggggcggcccctccggggtgggga	1.48E-02	hsa-mir-212
NIH037934	1,953,855	1,953,914	cgcccacggcttcccgcctctgcgagcggagctgtcctctcaggaccgggggggcggggc	2.52E-02	hsa-mir-212
NIH022345	131,155,654	131,155,713	ggcgttattcaaatccagtaccaattgcagccgagcccctttctccgcgccaggcccgta	3.35E-02	hsa-mir-219-2

probe ID	start	end	probe sequences	adj.P-Value	miRNA name
NIH022358	131,155,394	131,155,453	tcagccacaggaaagcggagagcgcccgagccggtgcgtgcgcggggctgggcgggggcg	2.04E-02	hsa-mir-219-2
NIH049825	45,606,495	45,606,554	ccccagaaggcaaaggatcacccagctgctggaaggtgtaggtaccctcaatggctcagt	3.31E-04	hsa-mir-221
NIH049919	45,606,511	45,606,570	cttccacagagcccctccccagaaggcaaaggatcacccagctgctggaaggtgtaggta	3.73E-04	hsa-mir-222
NIH010329	111,782,244	111,782,303	gccaacagtggccagagttggctcccatgctgtaaggttggaggaaggacacaccaggct	3.93E-04	hsa-mir-297
NIH019981	135,813,011	135,813,070	gcctgtagtgccccccactgccctggacacgcccccacttccaaaaaaacaaacacccaa	7.95E-03	hsa-mir-30b
NIH016054	57,471,831	57,471,890	acgcagggactacactgcgcatgttcaaagggggcgtgtcaggggtggggcgagacggag	2.64E-02	hsa-mir-3147
NIH036189	2,582,847	2,582,906	agggccagtctgggagggagtacccggacgctccggagctgtcccagactccaggggcgg	1.82E-03	hsa-mir-3178
NIH036232	2,581,987	2,582,046	ccagagttgcgcgtcagcctggactcagcgcctgcgctacgaggctgggcggggcgcggc	7.40E-03	hsa-mir-3178
NIH036418	14,994,885	14,994,944	tcattttccagtgggcctactgtgtgctggggtggcgagactcttcctgccctggaactt	2.65E-02	hsa-mir-3179-1
NIH036419	14,994,905	14,994,964	tgtgtgctggggtggcgagactcttcctgccctggaacttggtgaggtcaggagggaaga	2.94E-02	hsa-mir-3179-1
NIH036529	15,004,757	15,004,816	tctgtctctctcttagccaggaaacctggggtagggaggcttggagccagcgggtgcgtc	1.70E-02	hsa-mir-3180-1
NIH036841	16,403,416	16,403,475	tctctctcttagccaggaaacctggggtagggaggcttggagccagcgggtgcgtcggga	2.30E-02	hsa-mir-3180-2
NIH042174	18,392,487	18,392,546	ccgcccccgttcgccattgggtgcgggtgacgtcgctcatttgcatggaggggcggggcc	1.44E-02	hsa-mir-3188
NIH042179	18,392,587	18,392,646	gcgcggtaaacgccacaacagcgcgctgccttgtgggttgacgtcatgggggcggccgcg	2.53E-02	hsa-mir-3188
NIH042188	18,392,767	18,392,826	cccgaggccccgccccgtgtccgcgcgtggggccccggccgggcgggggcgggggcgggg	1.93E-02	hsa-mir-3188
NIH046096	30,194,329	30,194,388	cttcttgggttgaggggagggcctgggctgggggttgacccctgaccctcttcacagct	3.64E-02	hsa-mir-3193
NIH048080	42,538,924	42,538,983	aattagacagtccgcagagagctggctggggatagaaggggaggtgggggagaagggcag	3.00E-04	hsa-mir-3197
NIH031978	101,492,009	101,492,068	ccaggaggtgatatcagctttgcggaagagccactgtcctggtgtcagtacggctgctgc	6.25E-03	hsa-mir-323
NIH032126	101,493,217	101,493,276	ggccttctggtccagacctcagcttcagggaagggcgttactctcagctccagtccactg	1.24E-04	hsa-mir-329-2
NIH031122	100,773,576	100,773,635	agtggagtcgggccagatcgcagcgcctccgtcagcggcggccggacggggcgagaccgg	2.32E-03	hsa-mir-345
NIH031123	100,773,596	100,773,655	cagcgcctccgtcagcggcggccggacggggcgagaccggaggggctcggtaggggccgc	3.61E-02	hsa-mir-345
NIH026738	111,383,503	111,383,562	ctgcgaggccggcgggggtcccgctgggcccgggggtgtcctcgggggccgcttgcgccc	3.58E-02	hsa-mir-34b
NIH026765	111,384,004	111,384,063	ctgtatgctgtgattcactgtgtctatttgccatcgtctagtagagtattcaccaagcta	3.81E-02	hsa-mir-34c
NIH026776	111,384,024	111,384,083	tgtctatttgccatcgtctagtagagtattcaccaagctagcaactcagttgagctccaa	1.33E-02	hsa-mir-34c
NIH036488	15,001,454	15,001,513	gtgtggctgtgcctgtttggtcaccgtgtctgttctgattggtcggtgctcctgcatgtc	1.05E-02	hsa-mir-3670
NIH017802	1,748,911	1,748,970	gtccttcagccacccgcaggaattgccgccagagtaggacatgtcttctaccctcttgga	3.75E-04	hsa-mir-3674
NIH038245	8,090,173	8,090,232	ggctgtctgcagacgaggtggccgagtggttaaggcgatggactgctaatccattgtgct	1.47E-04	hsa-mir-3676
NIH036098	2,320,054	2,320,113	gaacagaggggccgccccccctgctgaagggccccagagtcggcgctaggggagggggca	2.32E-03	hsa-mir-3677
NIH036123	2,320,554	2,320,613	taaaactcaggagcctacccccaccctgcccttccaggagctcagccccagagccaggct	2.03E-03	hsa-mir-3677
NIH036131	2,320,714	2,320,773	ggcagtggccagagccctgcagtgctgggcatgggcttctcgtgggctctggccacggcc	2.04E-03	hsa-mir-3677
NIH039935	73,401,430	73,401,489	ggattgcgctccacgtcccaagcctcgccgccccgacgcgaaccccggggccggtcccgg	4.13E-03	hsa-mir-3678
NIH037123	29,611,407	29,611,466	gcagaacgcaccctgtcattacaaatgactcctggaggcagtccccgggggcctggcagg	5.09E-02	hsa-mir-3680
NIH019832	130,496,389	130,496,448	tcctctccctgtgctgtggcctggaaactgtctgtgcggtcagctggggcagtcagaggg	4.59E-03	hsa-mir-3686
NIH022423	137,741,551	137,741,610	ccgtgcttcctgggaggtgtgatcctgtgcttcctgggaggtgtgataccatgcttcctg	5.67E-03	hsa-mir-3689a

probe ID	start	end	probe sequences	adj.P-Value	miRNA name
NIH022476	137,742,219	137,742,278	gcttcctgggaggtgtgatcctgtgcttcctgggaggtgtgataccatgcttcctgggag	1.41E-02	hsa-mir-3689b
NIH045574	54,290,069	54,290,128	tgaaggtgtgtctgactaaggcaagctaggatcaaagggagcagggtggcttggggcgga	1.43E-03	hsa-mir-371
NIH045627	54,290,324	54,290,383	tcaggatctcactgtcgcccaggatgaagtgcacaggtaggatgatggcgccttgcagcc	1.49E-02	hsa-mir-372
NIH045628	54,290,344	54,290,403	aggatgaagtgcacaggtaggatgatggcgccttgcagcctcgacctcctgggactcacc	2.31E-02	hsa-mir-372
NIH06346	219,867,371	219,867,430	acactgcagctggactgagacctggccgctagctgccggccgctggaccgggtctggcca	3.71E-02	hsa-mir-375
NIH06354	219,866,851	219,866,910	cagcaccctcccctccgcccccgccaccaaggcctcggagaagctccggtctcagagccc	2.50E-02	hsa-mir-375
NIH032399	101,505,466	101,505,525	gagacattagggttacccccacccagacgagtgacagcagggcagaccccaaacttacgt	7.22E-04	hsa-mir-376a-2
NIH032437	101,506,226	101,506,285	atgtctgcggaccacccacccgccttgaggagacccctctcgcaagctgcccgctcggcc	9.49E-03	hsa-mir-376a-2
NIH032387	101,505,867	101,505,926	tagcaaggtctgtcctgggacatgcgtcctccgcagggcccatgtcactgccgctagccc	5.10E-02	hsa-mir-376c
NIH032794	101,511,937	101,511,996	attcctatagaaagtgagaagctgagctctgcgtgtgcccaggaaggcccgtggggtgga	4.22E-03	hsa-mir-381
NIH032795	101,511,957	101,512,016	gctgagctctgcgtgtgcccaggaaggcccgtggggtggatctccttttcaggaacaagt	3.84E-02	hsa-mir-381
NIH033167	101,520,563	101,520,622	acaggatggtggttggcagccactccccttggagaagtggaaggggactccttgtctgtc	6.39E-04	hsa-mir-382
NIH028149	69,979,404	69,979,463	ctggcgcaggccgcacgccccgcgcatggctcgcaccatgtgcggagggccggggagaca	3.15E-02	hsa-mir-3913-1
NIH014044	36,590,950	36,591,009	cagcatgtgactgtgccgcctctgacactgaccacccaggcctttctcttcactccttcc	2.03E-03	hsa-mir-3925
NIH033618	101,531,297	101,531,356	gtgcaggggtccctacagggtcacccccctctcagggtctgggaatgaaaagcgggtgcga	1.30E-03	hsa-mir-409
NIH031731	101,488,782	101,488,841	agctttggagggcttcgtggagccaatactaccttcaggggggacccaccagtccatcctt	1.30E-04	hsa-mir-411
NIH031756	101,489,282	101,489,341	tcccctaaccagctctgttagccatgtgccatgtcctccctcccatctacccactccttc	4.65E-02	hsa-mir-411
NIH033644	101,530,944	101,531,003	ggctgcccgctccaggaggcaccttctgggtgttctctgagtctggggaaggttgggttc	3.95E-02	hsa-mir-412
NIH05320	110,827,660	110,827,719	ctggatgaggttggcactagggctgccatctcagaacccagagtaggcccaggggtggta	2.80E-02	hsa-mir-4267
NIH06530	220,771,947	220,772,006	aagccagcccagagcccgagacccccccagcagaggcctgatttcacacacattttcccca	2.46E-02	hsa-mir-4268
NIH0179	1,103,785	1,103,844	agtggcctctctcacgtggtccgggctccgtggagggtctgctgggggcgcaggaaggac	3.18E-02	hsa-mir-429
NIH0199	1,104,185	1,104,244	tcccgggtacccccacacagccaggctgggctgggaccaggtgctgggatgggggctgtc	4.97E-02	hsa-mir-429
NIH033921	103,005,481	103,005,540	atgtgcccccccagctgtccaagcaggggtgtacagagactctggggtggggtgagggtga	5.61E-03	hsa-mir-4309
NIH033935	103,005,761	103,005,820	aggactaggttcagcccacagcaggtgggatgcacccgccccttggcgggccgggggtgg	4.44E-03	hsa-mir-4309
NIH031375	101,346,604	101,346,663	agaagaagtcagtggagcacccgaggtcagtgtcccagggtggccatcaccaggccgggt	3.06E-03	hsa-mir-431
NIH031399	101,347,084	101,347,143	gctcttctagccttgcctgctccctgggctggctctcccaggcggggatgggcagggccc	4.03E-02	hsa-mir-431
NIH031519	101,349,820	101,349,879	tgccgtcagctcccgaatccaccagggcctggaccgcgacgctgtggtaggggttcactc	3.23E-02	hsa-mir-432
NIH042456	42,638,586	42,638,645	ccctcaccaggcccagctctggtctgaggcactgagaatggagccagcatgggaggcgag	4.86E-03	hsa-mir-4323
NIH033316	101,521,596	101,521,655	gtaagtgcgcctcgggtgagcatgcacttaatgtgggtgtatgtcactcggctcggccca	1.09E-02	hsa-mir-485
NIH033088	101,518,163	101,518,222	gagtaagactcacatgctgtggcctccaggctctggaggcctcgggggcagagcagggtt	4.16E-02	hsa-mir-487a
NIH032849	101,512,552	101,512,611	tggtctgggtccctgcttcctgagcgggaagtcatcagcccgggccgatggctggggtgg	8.83E-03	hsa-mir-487b
NIH031248	101,335,217	101,335,276	cctctccctctcgtctttgggggggtccatctcatgcatgtctgttaccatggactcca	4.49E-02	hsa-mir-493
NIH053094	146,308,071	146,308,130	gtgattctcctgcctcagcagctgcggttacagttgcccaccaccacacccggctaattt	2.74E-03	hsa-mir-513a-2
NIH053007	146,280,866	146,280,925	agtggcctgatctcctctcactgaaacctgactcccagggtcaaggtgtactcctgcctc	6.35E-04	hsa-mir-513b

probe ID	start	end	probe sequences	adj.P-Value	miRNA name
NIH053008	146,280,846	146,280,905	ctgaaacctgactcccagggtcaaggtgtactcctgcctcagctcctgggtggctgggat	4.69E-04	hsa-mir-513b
NIH045480	54,263,727	54,263,786	ctcccacctaagcctgctgcattagtgtctacaggcatgcaccaccacacccggctcact	1.69E-03	hsa-mir-516a-2
NIH044437	54,215,841	54,215,900	tagcaggatctctgctcaccggaaactccacctctcgggttccagtgattctcccacctc	5.16E-03	hsa-mir-519d
NIH044248	54,210,347	54,210,406	tcggatgcccccatgaggactgtgcgctcctgtactggaactcaagcgaccacttggctc	9.53E-05	hsa-mir-520c
NIH044645	54,224,680	54,224,739	cttccggcgattctcccacctcagcctgccgaatagttgggaatagagatgcccgccatc	2.31E-02	hsa-mir-520g
NIH032864	101,512,678	101,512,737	ttggggtatgtgacccggtccactaaccctcagcatctaattcatccccaggaccgcgcc	3.12E-02	hsa-mir-539
NIH032967	101,514,035	101,514,094	agcaaaccttagggaccgatcattggggccagacccccttcccctgcccaagagtgtgac	8.63E-03	hsa-mir-544
NIH08461	114,035,657	114,035,716	gctgctaaggtgcccggagtccagaatgtccattaatcactcaggcacgagcctggcact	7.11E-04	hsa-mir-568
NIH09773	11,370,291	11,370,350	ttggcgcggaatccgaccgcatccgtgtttcgggggctgcccggcaggacgcatcgtgaa	2.59E-02	hsa-mir-572
NIH025877	65,211,169	65,211,228	tttgcaaccccactggccagagggaagggccagtcacttggctctctcactgccctgcgc	5.66E-03	hsa-mir-612
NIH025878	65,211,189	65,211,248	agggaagggccagtcacttggctctctcactgccctgcgccccagatggttctagggctg	4.10E-04	hsa-mir-612
NIH025879	65,211,209	65,211,268	gctctctcactgccctgcgccccagatggttctagggctgctgttttcccttggccctgc	4.96E-05	hsa-mir-612
NIH046541	49,202,243	49,202,302	atgctgcagctgatctaacaggagatggagctcaggtggtcatgctcagtcgctcgccac	1.43E-02	hsa-mir-645
NIH033061	101,515,767	101,515,826	gtgctttctttgcaggatgtgaacacctccctgccccaaccctgggattcagctcatccc	1.01E-04	hsa-mir-655
NIH031987	101,491,437	101,491,496	gagggcttcccagagagaggcaaagtcaccttcgtgggagacacactgatctggcttcag	4.96E-05	hsa-mir-758
NIH042684	46,521,650	46,521,709	gagaaggggtactacccctccatcccccacccacttgctgggtatggtgtgggggggtggg	1.51E-02	hsa-mir-769
NIH042685	46,521,670	46,521,729	catcccccacccacttgctgggtatggtgtgggggggtggggccagggtggcctccagccc	2.19E-02	hsa-mir-769
NIH042707	46,522,110	46,522,169	gtggtgggaaggaggtgtcttgcagcgtggttcactgcccaggaggacccccaggaccca	2.86E-03	hsa-mir-769
NIH032953	101,514,018	101,514,077	gcacttcttggacatgaagcaaaccttagggaccgatcattggggccagacccccttccc	5.49E-03	hsa-mir-889
NIH032954	101,514,038	101,514,097	aaaccttagggaccgatcattggggccagacccccttcccctgcccaagagtgtgactga	7.40E-03	hsa-mir-889
NIH02347	155,164,548	155,164,607	gctgggcggggtggggacatctgacgtcagcgcgcccgggagcgcgggccggggaggcgg	2.30E-02	hsa-mir-92b
NIH02353	155,164,668	155,164,727	tcaactcccggcattgccaagcaacagccattcagttcggttgctgggacacgcgtcacc	2.48E-02	hsa-mir-92b
NIH02370	155,164,688	155,164,747	gcaacagccattcagttcggttgctgggacacgcgtcaccatggcgacggctccgcgccg	4.60E-02	hsa-mir-92b
NIH02371	155,164,908	155,164,967	tccccagcgcgccccgggttcccgggtcgggccggcttcaggcggtggggagcgggatcc	3.96E-02	hsa-mir-92b
NIH035563	89,910,328	89,910,387	gggcctcgcccaagtggagcatagctgaggaacctctgagtgccaggtgttatgggtggg	1.54E-02	hsa-mir-9-3
NIH035568	89,910,428	89,910,487	aatccctggtctctgccgcgtgggctagatctactgcaagtgctgggcatggggaaagga	2.65E-04	hsa-mir-9-3
NIH035590	89,910,868	89,910,927	ccccgcggggcgattagcctgcgagaggagggccggcggtccagtgcgctgggggcggcc	2.94E-02	hsa-mir-9-3

miRNA and constructed a frequency graph for DMPs in 200-bp intervals. As shown in Figure 3.4, the majority of DMPs were located within 400 bp upstream from the start site of each miRNA, suggesting that the major DMR, which is defined as the contig of DMPs, overlaps with the region containing the transcriptional regulatory elements such as promoter and proximal sequence elements. Interestingly, there is an additional DMR (20.9% of total DMPs) spanning from -800 to -1,000 where enhancers are usually found.

The sequence elements responsible for gene expression are closely regulated by methylation, therefore, we also screened for the distributions of TFBSs and TSSs upstream of the miRNAs. We searched 1 kb upstream of 1,049 human miRNAs to find TFBSs and TSSs using PROMO (ver 3.0) (Messequer *et al*., 2002) and Eponine (Down *et al*., 2002), respectively, as described in the Materials and Methods.

A total of 2,457 predicted TFBSs and 1,346 predicted TSSs prediction were identified. The majority of predicted TFBSs (n = 1406; 57.2% of the total) and TSSs (n = 905; 67.1% of the total) were upstream of intergenic miRNAs, which is consistent with a previous study showing that predicted TSSs are mainly distributed within 2 kb upstream of the intergenic miRNAs (Bracht *et al*., 2004). As shown in Figure 3.5, TFBSs were abundant at 170–230, 300–470, 680–780, and 850–960 bp upstream (blue bar in Figure 3.5), and TSSs were located at 50–250 bp upstream from the pre-miRNA starting site (blue-dotted bar in Figure 3.5). It is noteworthy that the regions upstream of intergenic miRNAs that were enriched for TFBSs and TSSs were generally overlapped with highly methylated regions, suggesting that these regions may be related to a putative epigenetic regulatory site for intergenic miRNA gene expression.

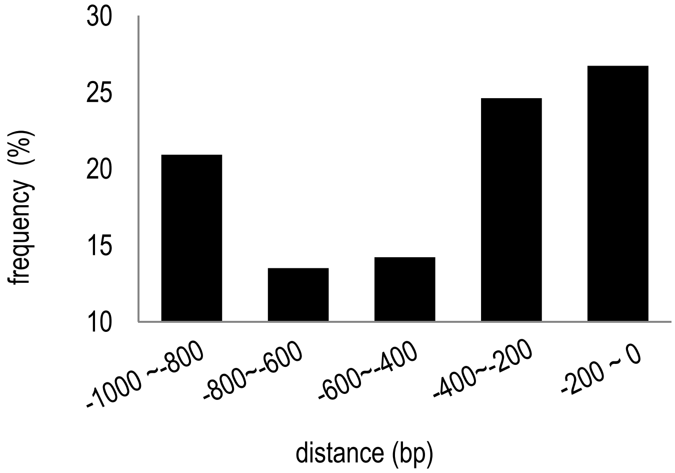


Figure 3.4. Distribution of DMPs in 200-bp intervals. The frequency is represented as a percentile of DMPs in each interval.

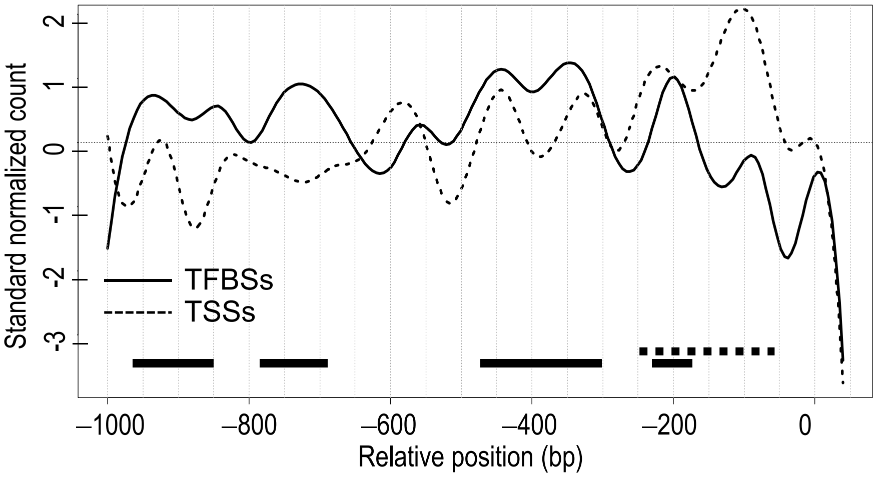


Figure 3.5. Distribution of predicted TSSs and TFBSs upstream from the starting site of the intergenic miRNAs. The black bar indicates the regions enriched for TFBSs, and the black-dotted bar indicates the regions enriched for TSSs. In both figures, the x-axis represents the relative distance from the staring site of the miRNAs, and the y-axis represents standard normalized values of the number of DMRs, TFBSs, and TSSs, shown as mean and standard deviation.

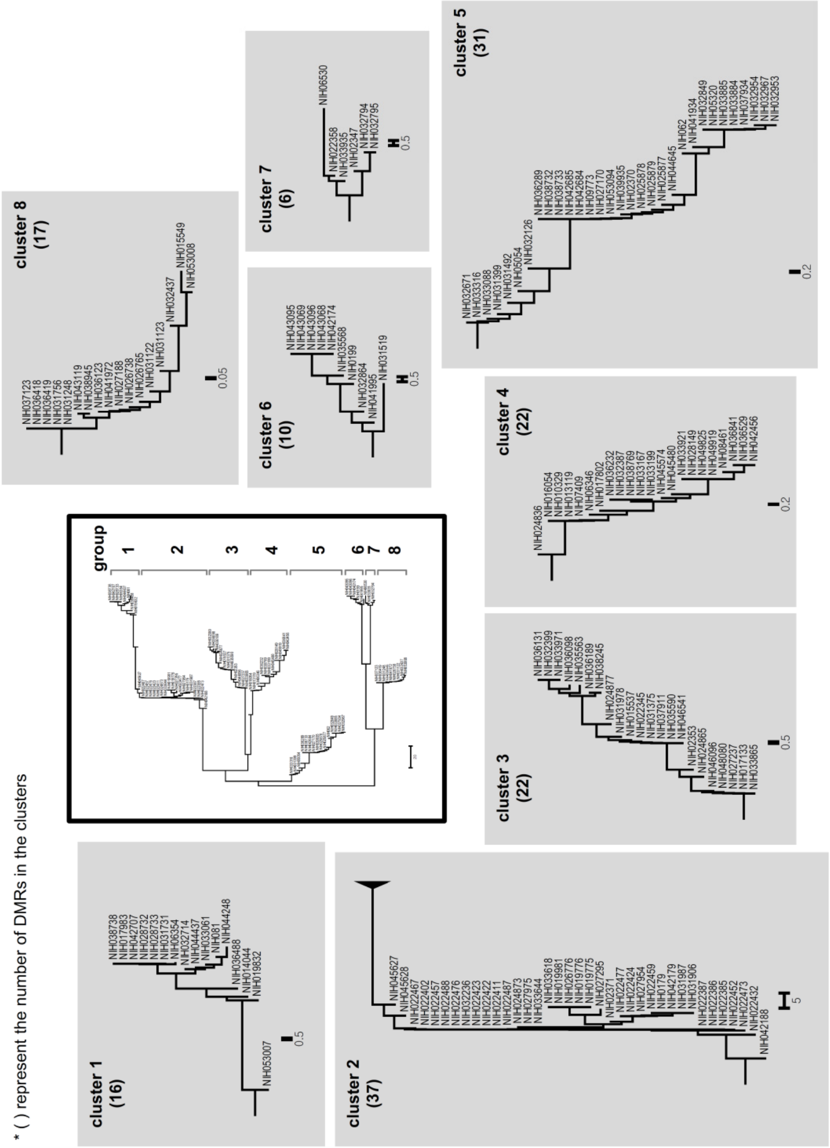
3.3. Clustering DMRs using phylogenetic method

We predicted that most or all of the DMR sequences could be clustered by their sequence similarities to identify specific miRNA regulatory motifs. To test this prediction, DMR sequences were aligned using clustalX (Larkin *et al*., 2007), and a phylogenetic tree was constructed based on the NJ model (Saitou *et al*., 1987) using MEGA5 (Tamura *et al*., 2011).

Analysis of the DMRs of the intergenic miRNA revealed that the DMRs were clustered into eight distinct clusters in the tree structure, which was constructed with site coverage cutoff at 99% (Figure 3.6). Clusters 2 and 5 were the largest, containing 37 and 31 DMRs, respectively. The number of the DMRs in clusters 1, 3, 4, 6, 7, and 8 were 16, 22, 22, 10, 6 and 17, respectively. When performing multiple sequence alignment with 161 DMPs using the ClustalX program (Larkin *et al*., 2007), only a single consensus sequence, CNNNNNNC (C[N]$_6$C), was found in all DMPs, which was also composed of conserved cytosine residues at position 1 and 8 and almost equal frequencies of different nucleotides in positions 2 to 7 (Figure 3.7).

3.4. Prediction of sequence motifs from DMPs

Because altering genome DNA methylation usually affects the efficiency of gene expression, it can be assumed that a specific sequence motif that regulates the transcription of its target miRNA gene is located within DMRs. Therefore, we analyzed the DMP sequences to predict a sequence motif using MEME software (Bailey *et al*., 2006), which is used to predict statistically overrepresented sequence motifs.



* ( ) represent the number of DMRs in the clusters

Figure 3.6. Clusters of DMRs upstream of intergenic miRNAs. A tree structure constructed by NJ method with site coverage cut off at 99% using DMRs upstream of intergenic miRNA. Overview structure of the tree is located on center box. Each box represents the DMRs belonging to each cluster for detail. The numbers in brackets represent the number of DMRs in the clusters

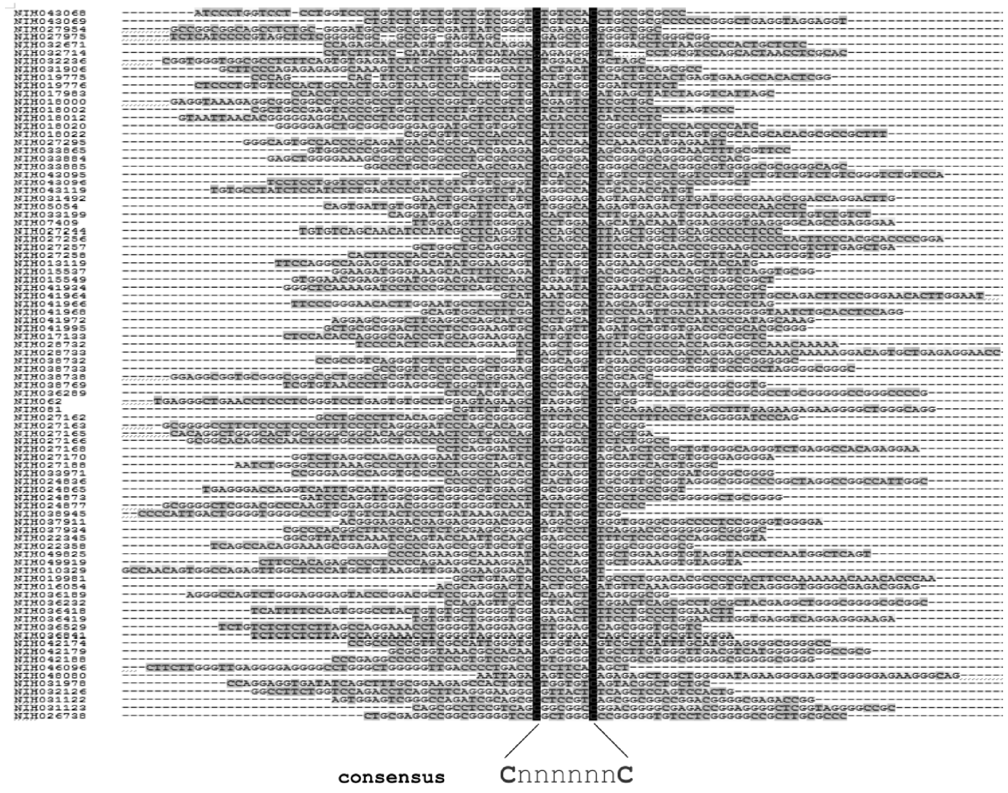


Figure 3.7. A single consensus sequence obtained by multiple alignment. The results of multiple sequence alignments by ClustalX using DMRs upstream of intergenic miRNA. A consensus sequence, C[N]$_6$C, was included in all DMRs in their clusters

After performing MEME with DMP sequences, we obtained six significant sequence patterns (p-value < 1.00e-05; Figure 3.8) which were 8–11 bp long. The most significantly overrepresented pattern was CNNNNNNCT ($C[N]_6CT$, p-value = 3.22e-18). N designates a non-conserved nucleotide. CTANCCTC, CTCTNCNC, TCTNNNTNT, GAGGTNTGATC, and CNNAGNGAC were also selected as significant patterns, the p-values of which were 2.15e-07, 5.26e-07, 2.34e-05, 2.80e-05, and 8.22e-05, respectively. It should be noted that there was a significant difference between the p-values for the $C[N]_6CT$ pattern and CTANCCTC (11 logarithmic order), suggesting that $C[N]_6CT$ is a major sequence motif in DMRs. Interestingly, CTANCCTC and CTCTNCNC patterns, as well as $C[N]_6CT$, contain two cytosine residues at positions 1 and 8, suggesting the cytosine residues at these positions are important for these patterns. It should be noted that this $C[N]_6CT$ motif is highly similar to the $C[N]_6C$ pattern performed by multiple alignment. These results suggest that the two conserved cytosine residues at positions 1 and 8 are likely embedded in the regulatory elements, and that the $C[N]_6CT$ pattern may be a potential regulatory motif for intergenic miRNA gene expression. Thus, we selected the $C[N]_6CT$ pattern as a candidate motif of regulatory element for intergenic miRNA expression.

We next examined the $C[N]_6CT$ pattern within 1000 bp upstream of 98 intergenic miRNAs that contain DMPs, and identified a total of 1,766 $C[N]_6CT$ motifs. Among the 1,766 $C[N]_6CT$ motifs, 189 (10.7%) were located in DMPs (Table 3.2). Overall, the number of $C[N]_6CT$ motifs found in DMPs is small compared to that found in non-DMPs. However, further analysis of this motif showed that each gene contains a different frequency of $C[N]_6CT$ motifs in its DMPs. For example, miR-

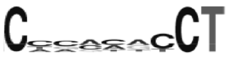
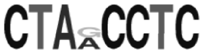
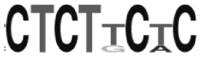
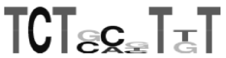
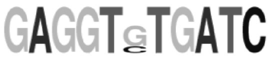
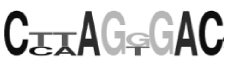
Consensus	motif logo	p-value [a]
CNNNNNNCT		3.22e-18
CTANCCTC		2.15e-07
CTCTNCNC		5.26e-07
TCTNNNTNT		2.34e-05
GAGGTNTGATC		2.80e-05
CNNAGNGAC		8.22e-05

Figure 3.8. Identification of significant sequence patterns. Six highly significant motifs were identified in the DMRs using MEME software (see Materials and Methods). Consensus designates the conserved sequence at each position. N represents non-conserved nucleotides. Motif logos are the graphically represented sequences showing homology at each position. The p-value designates significance of the sequences calculated against the random sequences.

Table 3.2. The number of DMPs and C[N]$_6$CT motifs within 1 kb upstream region of 98 intergenic miRNAs.

miRNA name	start	end	chr.	strand	# of DMPs	# of C[N]$_6$CT motifs		
						Total	within DMRs	ratio (%)
* hsa-mir-200c	7,072,862	7,072,929	12	+	7	22	16	72.73
hsa-mir-141	7,073,260	7,073,354	12	+	4	20	8	40.00
* hsa-mir-124-1	9,760,898	9,760,982	8	-	6	28	11	39.29
hsa-mir-181c	13,985,513	13,985,622	19	+	5	24	9	37.50
* hsa-mir-612	65,211,929	65,212,028	11	+	3	16	6	37.50
* hsa-mir-375	219,866,367	219,866,430	2	-	2	18	6	33.33
* hsa-mir-34c	111,384,164	111,384,240	11	+	2	14	4	28.57
* hsa-mir-210	568,089	568,198	11	-	4	7	2	28.57
hsa-mir-3677	2,320,714	2,320,773	16	+	3	18	5	27.78
hsa-mir-3179-1	14,995,365	14,995,448	16	+	2	19	5	26.32
hsa-mir-125a	52,196,507	52,196,592	19	+	3	25	6	24.00
hsa-mir-513b	146,280,562	146,280,645	X	-	2	18	4	22.22
hsa-mir-432	101,350,820	101,350,913	14	+	1	15	3	20.00
hsa-mir-411	101,489,662	101,489,757	14	+	2	16	3	18.75
hsa-let-7i	62,997,466	62,997,549	12	+	2	17	3	17.65
* hsa-mir-212	1,953,565	1,953,674	17	-	2	18	3	16.67
hsa-mir-148a	25,989,539	25,989,606	7	-	2	12	2	16.67
hsa-mir-3178	2,581,923	2,582,006	16	-	2	12	2	16.67
hsa-mir-3676	8,090,493	8,090,577	17	+	1	12	2	16.67
hsa-mir-376a-2	101,506,406	101,506,485	14	+	2	19	3	15.79
hsa-let-7e	52,196,039	52,196,117	19	+	2	21	3	14.29
hsa-mir-2117	41,522,174	41,522,253	17	+	1	21	3	14.29
hsa-mir-329-2	101,493,437	101,493,520	14	+	1	14	2	14.29
hsa-mir-1827	100,583,662	100,583,727	12	+	2	15	2	13.33
hsa-mir-487a	101,518,783	101,518,862	14	+	1	15	2	13.33
hsa-mir-516a-2	54,264,387	54,264,476	19	+	1	15	2	13.33
hsa-mir-655	101,515,887	101,515,983	14	+	1	15	2	13.33
hsa-mir-3925	36,590,213	36,590,289	6	-	1	23	3	13.04
hsa-mir-1208	129,162,362	129,162,434	8	+	2	16	2	12.50
hsa-mir-30b	135,812,763	135,812,850	8	-	1	8	1	12.50
hsa-mir-3686	130,496,303	130,496,388	8	-	1	24	3	12.50
hsa-mir-487b	101,512,792	101,512,875	14	+	1	16	2	12.50
hsa-mir-1185-1	101,509,314	101,509,399	14	+	1	17	2	11.76
hsa-mir-513a-2	146,307,344	146,307,470	X	-	1	17	2	11.76

miRNA name	start	end	chr.	strand	# of DMPs	# of C[N]$_6$CT motifs		
						Total	within DMRs	ratio (%)
hsa-mir-219-2	131,154,897	131,154,993	9	-	2	18	2	11.11
hsa-mir-519d	54,216,601	54,216,688	19	+	1	18	2	11.11
hsa-mir-3680	29,610,500	29,610,586	16	-	1	19	2	10.53
hsa-mir-520c	54,210,707	54,210,793	19	+	1	19	2	10.53
hsa-mir-4323	42,637,597	42,637,665	19	-	1	10	1	10.00
hsa-mir-193a	29,887,015	29,887,102	17	+	4	21	2	9.52
hsa-mir-1285-2	70,480,050	70,480,137	2	-	1	11	1	9.09
hsa-mir-758	101,492,357	101,492,444	14	+	1	11	1	9.09
hsa-mir-769	46,522,190	46,522,307	19	+	3	23	2	8.70
hsa-mir-9-3	89,911,248	89,911,337	15	+	3	23	2	8.70
hsa-mir-372	54,291,144	54,291,210	19	+	2	23	2	8.70
hsa-mir-200b	1,102,484	1,102,578	1	+	2	25	2	8.00
hsa-mir-381	101,512,257	101,512,331	14	+	2	25	2	8.00
hsa-mir-4309	103,005,981	103,006,063	14	+	2	26	2	7.69
hsa-mir-221	45,605,585	45,605,694	X	-	1	13	1	7.69
hsa-mir-1185-2	101,510,535	101,510,620	14	+	1	27	2	7.41
hsa-mir-429	1,104,385	1,104,467	1	+	2	14	1	7.14
hsa-mir-544	101,514,995	101,515,085	14	+	1	14	1	7.14
hsa-mir-493	101,335,397	101,335,485	14	+	1	29	2	6.90
hsa-mir-203	104,583,742	104,583,851	14	+	1	15	1	6.67
hsa-mir-222	45,606,421	45,606,530	X	-	1	15	1	6.67
hsa-mir-297	111,781,738	111,781,803	4	-	1	15	1	6.67
hsa-mir-376c	101,506,027	101,506,092	14	+	1	15	1	6.67
hsa-mir-3913-1	69,978,502	69,978,603	12	-	1	15	1	6.67
hsa-mir-3678	73,402,150	73,402,243	17	+	1	16	1	6.25
hsa-mir-520g	54,225,420	54,225,509	19	+	1	16	1	6.25
hsa-mir-568	114,035,322	114,035,416	3	-	1	16	1	6.25
hsa-mir-431	101,347,344	101,347,457	14	+	2	17	1	5.88
hsa-mir-3193	30,194,989	30,195,043	20	+	1	17	1	5.88
hsa-mir-3674	1,749,291	1,749,358	8	+	1	17	1	5.88
hsa-mir-382	101,520,643	101,520,718	14	+	1	17	1	5.88
hsa-mir-1197	101,491,901	101,491,988	14	+	1	18	1	5.56
hsa-mir-323	101,492,069	101,492,154	14	+	1	18	1	5.56
hsa-mir-539	101,513,658	101,513,735	14	+	1	18	1	5.56
hsa-mir-889	101,514,238	101,514,316	14	+	2	20	1	5.00
hsa-mir-3197	42,539,484	42,539,556	21	+	1	20	1	5.00
hsa-mir-4268	220,771,223	220,771,286	2	-	1	20	1	5.00

miRNA name	start	end	chr.	strand	# of DMPs	# of C[N]$_6$CT motifs		
						Total	within DMRs	ratio (%)
* hsa-mir-3188	18,392,887	18,392,971	19	+	3	21	1	4.76
hsa-mir-134	101,521,024	101,521,096	14	+	1	21	1	4.76
hsa-mir-485	101,521,756	101,521,828	14	+	1	21	1	4.76
hsa-mir-4267	110,827,538	110,827,619	2	-	1	23	1	4.35
hsa-mir-412	101,531,784	101,531,874	14	+	1	24	1	4.17
* hsa-mir-92b	155,164,968	155,165,063	1	+	4	19	0	0.00
hsa-mir-1247	102,026,624	102,026,759	14	-	3	26	0	0.00
hsa-mir-345	100,774,196	100,774,293	14	+	2	18	0	0.00
hsa-mir-1193	101,496,389	101,496,466	14	+	1	14	0	0.00
hsa-mir-1244-3	9,392,063	9,392,147	12	-	1	16	0	0.00
hsa-mir-127	101,349,316	101,349,412	14	+	1	15	0	0.00
hsa-mir-138-1	44,155,704	44,155,802	3	+	1	15	0	0.00
hsa-mir-146a	159,912,359	159,912,457	5	+	1	10	0	0.00
hsa-mir-181d	13,985,689	13,985,825	19	+	1	24	0	0.00
hsa-mir-182	129,410,223	129,410,332	7	-	1	18	0	0.00
hsa-mir-193b	14,397,824	14,397,906	16	+	1	17	0	0.00
hsa-mir-3147	57,472,731	57,472,796	7	+	1	19	0	0.00
hsa-mir-3180-1	15,005,077	15,005,170	16	+	1	29	0	0.00
hsa-mir-3180-2	16,403,736	16,403,823	16	+	1	29	0	0.00
hsa-mir-34b	111,383,663	111,383,746	11	+	1	17	0	0.00
hsa-mir-3670	15,001,574	15,001,638	16	+	1	19	0	0.00
hsa-mir-3689a	137,741,333	137,741,410	9	-	1	2	0	0.00
hsa-mir-3689b	137,741,971	137,742,118	9	-	1	7	0	0.00
hsa-mir-371	54,290,929	54,290,995	19	+	1	19	0	0.00
hsa-mir-409	101,531,637	101,531,715	14	+	1	25	0	0.00
hsa-mir-572	11,370,451	11,370,545	4	+	1	17	0	0.00
hsa-mir-645	49,202,323	49,202,416	20	+	1	20	0	0.00
Total number					161	1766	189	10.70

Intergenic miRNAs which were tested for their expression after DAC treatment are marked as asterisk (*)

200c has three DMRs, and 16 out of 22 C[N]$_6$CT motifs (72.7%) are concentrated in the DMRs (top line in Figure 3.9). In miR-124-1, there are four largely methylated regions that contain 11 out of 28 C[N]$_6$CT motifs (39.3%) (middle line in Figure 3.9). On the other hand, miR-92b contains three methylated regions, but there is no C[N]$_6$CT motif in these regions (bottom line in Figure 3.9).

3.5. Effect of demethylation on intergenic miRNA expression

It is important to determine whether the expression level of intergenic miRNA changes depending on the frequency of C[N]$_6$CT motifs in DMRs. Therefore, we first treated H4 cells with DAC to demethylate DNA. Then we isolated total RNA and measured the expression levels using quantitative PCR analysis against eight selected intergenic miRNA genes: one miRNA which had the highest frequency of C[N]$_6$CT motifs in DMRs (miR-200c), two with a 30–40% frequency (miR-124-1, miR-375), two with a 20–30% frequency (miR-34c, miR-210), one with a 10–20% frequency (miR-212), and two with less than 10% frequency (miR-3188 and miR-92b), all of which had more than three DMRs in their upstream regions but had no or only one C[N]$_6$CT motif in the DMRs (Table 3.2). *Rnu6b* was used as a negative control. As shown in Figure 3.10, the expression levels of six of eight intergenic miRNAs (miR-200c, miR-124-1, miR-375, miR-34c, miR-210, and miR-212) increased significantly in DAC-treated cells compared to untreated normal cells. The expression level of miR-200c, which shows the highest frequency of C[N]$_6$CT motifs in DMRs (72.7%), was increased by 22.3-fold after DAC treatment. The expression levels of miR-124-1, miR-375, miR-34c, miR-210, and miR-212 were also increased by 11.3, 9.4, 8.4, 3.5, and 13.2-fold after DAC treatment, respectively. On the other hand, DAC

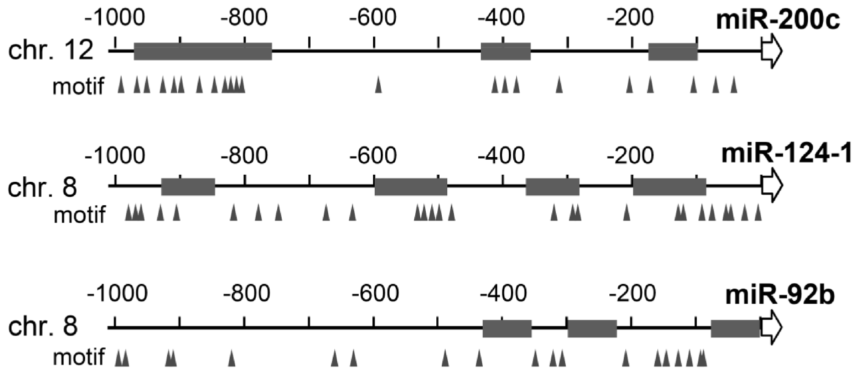


Figure 3.9. The distribution of C[N]$_6$CT motifs on the 5'-flank of intergenic miRNA genes. Schematic representation of the genomic region encompassing miR-200c, miR-124-1, and miR-92b showing the distribution of C[N]$_6$CT motifs. The gray box represents DMR, which is the contig of DMPs. C[N]$_6$CT motifs are marked with triangles. All features were drawn based on the distance from the 5'-end of each miRNA.

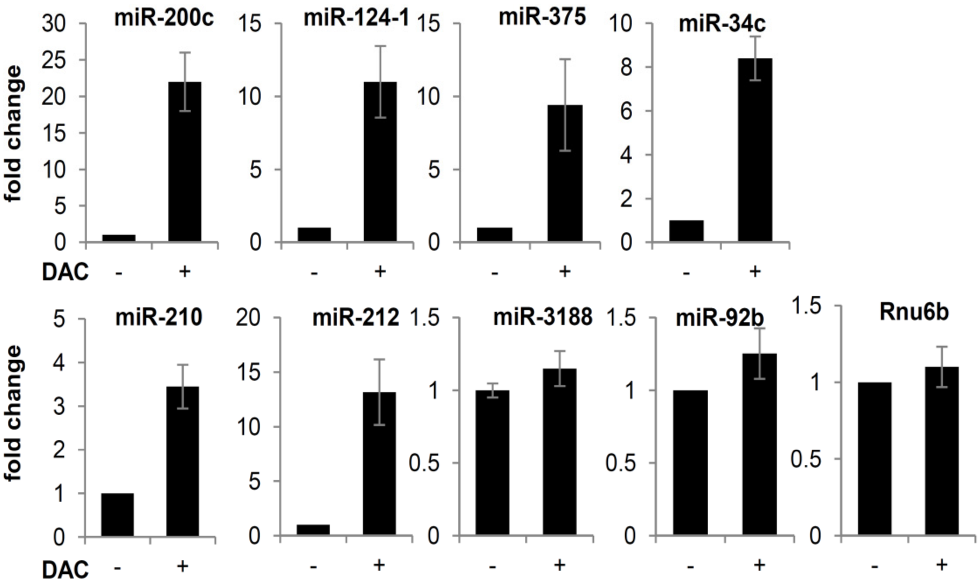


Figure 3.10. Measurement of intergenic miRNA gene expression by qPCR. The intergenic miRNA levels were measured by qPCR. *Rnu6b* was used as a negative control. The x-axis shows the experimental condition treated with (+) and without (–) DAC, respectively. The y-axis represents the relative fold change of expression level of each intergenic miRNA after treating the cells with DAC. The fold change of non-treated cells was set to 1. The gene symbol is marked above each panel.

treatment did not significantly change the expression levels of miR-92b or miR-3188. It should be noted that the frequencies of the C[N]$_6$CT motif in the DMR of miR-92b and miR-3188 are less than 10%, although they have more than three DMRs. These results strongly suggest that the frequency of C[N]$_6$CT motifs in DMRs is related to intergenic miRNA expression.

3.6. Correlations between the C[N]$_6$CT motif in DMRs and intergenic miRNA expression

Because the expression of intergenic miRNA genes were changed after DAC treatment in a motif-frequency dependent manner, we analyzed the relationship between the frequency of the C[N]$_6$CT motif in DMRs and the expression of intergenic miRNAs induced by demethylation. Correlation analysis between the fold changes of gene expression and the frequency of C[N]$_6$CT motifs in DMRs was performed with the eight miRNA genes described in the previous section. After performing Pearson's correlation test using those miRNAs, we obtained a high correlation value of 0.87 (p-value = 4.3e-03) between the frequency of C[N]$_6$CT motifs in DMRs and the fold changes in miRNA expression after demethylation (Figure 3.11). These results indicate that the frequency of C[N]$_6$CT motifs in DMRs plays a role in the expression efficiency of intergenic miRNAs in conjunction with the methylation status, and strongly suggests that the C[N]$_6$CT sequence pattern in DMRs is a methylation-dependent regulatory motif for intergenic miRNA expression.

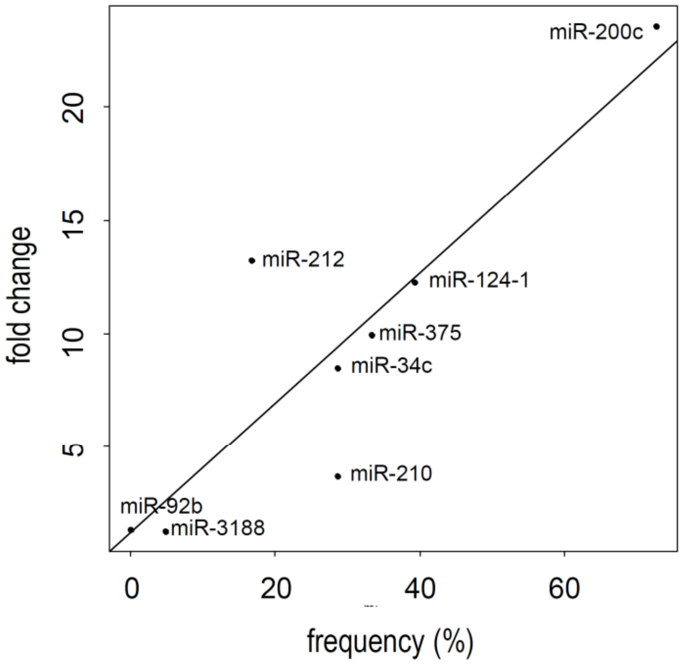


Figure 3.11. Pearson's correlations between the expression levels and the frequency of the C$[N]_6$CT motif in DMRs of intergenic miRNAs. The y-axis represents fold change measured by RT-qPCR and the x-axis represents the frequency of the motifs within DMRs. The calculated Pearson's correlation, *r,* was 0.86.

4. DISCUSSION

After identifying 161 DMPs within 1000 bp upstream of human intergenic miRNAs using microarray analysis, we searched for motifs within the DMRs and found a sequence motif, C[N]$_6$CT, which is conserved in the DMRs. Previous studies have reported that CT-repeat microsatellites are abundant within 1000 bp upstream of most intergenic miRNA (Zhou *et al.*, 2007 and references therein). Some motifs containing CT-repeats, including (CCT)$_n$, (CCTT)$_n$, (CGCT)$_n$, and (CCTCT)$_n$, have previously been identified in plants (Fujimori *et al.*, 2003; Molina *et al.*, 2005). Among these, the (CCT)$_n$ and (CCTCT)$_n$ motifs are very similar to the C[N]$_6$CT motif, which have two cytosine residues at positions 1 and 8. In other words, when n equals 3 in (CCT)$_n$, the sequence of the motif will be CCTCCTCCT, which can be represented as C[N]$_6$CT. Similarly, (CCTCT)$_n$ can also be represented as CC[N]$_6$CT when n equals 2. Therefore, the C[N]$_6$CT motif is highly similar to (CCT)$_n$ and (CCTCT)$_n$.

Generally, promoters contain mehylated CpG islands is well understood in long-term silencing of genes (Lin *et al.*, 2007; Jones *et al.*, 1999). Despite the fact that 45% of all human gene promoters, particularly those controlling the expression of tissue-specific genes, do not lie within CpG islands (Takai *et al.*, 2002), is almost unknown about their regulation and the potential role of methylation as a transcriptional control mechanism. A number of genes with non-CpG island promoters have been reported to be methylated in normal tissues, displaying a tissue-specific methylation pattern, suggesting DNA methylation may play a role in the establishment and maintenance of tissue-specific expression patterns (Eckhardt *et al.*, 2006). Some studies have shown that there is an inverse correlation between DNA methylation and gene expression, as it has been demonstrated for CpG island promoters (Eckhardt *et al*., 2006; Gal-Yam *et al*., 2008; Han *et al*., 2011), where other studies have reported that CpG-poor promoters could be still expressed when they are methylated (Weber M. *et al*., 2007). Moreover, Barres *et al* have reported that the non-CpG methylation level was highly enriched on the promoters of PGC-1α and TFAM compared to global levels (Barres *et al*., 2009). They showed that 7% of cytosines within the sequence CCAGG or CCTGG are methylated in human skeletal muscle suggests that further attention should be given to non-CpG methylation when using current DNA methylation analysis techniques.

The $C[N]_6CT$ motif must be located in DMRs to play a role as a regulatory element because the expression was increased in the genes with high frequency of the $C[N]_6CT$ motif in DMRs after DAC treatment. For examples, the *mir*-200c/141 cluster and *mir*-124-1 genes are functionally involved in carcinogenesis regulated by DNA methylation-based silencing (Neves *et al*., 2010; Hashimoto *et al*., 2010; Wilting *et al*., 2010). The majority of DMRs in these genes were found within 400 bp and 800–1000 bp upstream from the start sites of each miRNA, and the $C[N]_6CT$ motif was also predominantly found in these regions. It is possible that the cytosine residues at positions 1 and 8 of the $C[N]_6CT$ motif in DMRs are methylated because they are found in methylated regions. The fact that this motif is associated with the expression of intergenic miRNA genes in a motif-frequency dependent manner suggests that the $C[N]_6CT$ motif regulates gene expression by methylation/demethylation of cytosine.

Our findings indicate that the C[N]$_6$CT motif is closely associated with DNA methylation regions and might be a regulatory factor binding site or recognition sequence for miRNA processing. Combining DNA methylation signature with the C[N]$_6$CT motif may be useful for computationally predicting intergenic miRNAs.

CHAPTER 4.

CONCLUSIONS AND FURTHER REMARKS

1. CONCLUSIONS

Microarray technology is widely used in various biological researches such as finding gene functions by measuring expression profiles, detecting genetic variation among populations, and epigenetic analysis. Especially, microarrays are very useful tool because the expression levels of thousands of genes can be monitored simultaneously. However, the prediction accuracy that can be obtained from microarrays is dependent on the statistical methods and the several bioinformatical approaches including gene clustering, GO analysis.

For example predicting gene function, we used SPS1 which functions are unknown yet, and predicted that vitamin B_6 biosynthesis is the primary target of SPS1 by employing two-way ANOVA, SOM clustering and GO analysis. We confirmed the prediction experimentally by showing that PLP levels were decreased by *SPS1* knockdown and that the inhibition of PLP biosynthesis caused the same phenotypes as *SPS1* knockdown.

For example for predicting regulatory elements, we selected intergenic miRNAs because the transcriptional mechanism of intergenic miRNAs is not well understood. Microarray data is also used to identify the differentially methylated regions of intergenics miRNAs by statistical methods. We found that the regulation of gene expression by $C[N]_6CT$ motif is closely associated with DNA methylation status and the frequency of $C[N]_6CT$ motif occurrence in DMRs of intergenic miRNA gene. This motif may be a regulatory factor binding site for transcription factors or demethylase.

We showed that analysis of microarray data using bioinformatical approaches is very useful to predict gene functions and regulatory elements. Though microarray techonolgy can give high-throughput, high-density information, it is still difficult for researchers to plan and decide the various parameters and options associated with microarray data analysis. However, the combinatorial analysis of various bioinformatical tools and statistical models for analyzing microarray data will give us more precisely predicted information for biological problems. Nowdays, new high throughput sequencing, or next generation sequencing (NGS), technology, is developed and widely used for lots of research areas, such as genomics, transcriptomics, physiomics and so on. More powerful and obvious results for biological interpretation can be served by the bioinformatical prediction and analysis using both high-throughput technologies in the future.

2. FURTHER REMARKS

In prediction of transcriptional regulatory elements using intergenic miRNAs, we found that the C[N]$_6$CT motif in DMRs can act as a transcriptional regulatory motif for intergenic miRNA expression. However, there are no direct evidences whether conserved cytosine residues positioned at 1 and 8 are methylated or not. Generally, it is though that DNA methylation is occurred within CpG islands in human, although it has been reported that some DNA methylation is found on non-CpG regions recently. For these reasons, bisulfide sequencing have to be performed to confirm the methylation status of the cytosine residues of C[N]$_6$CT motif.

REFERENCES

Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M: Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* 2010; 11:383.

Agirre X, Vilas-Zornoza A, Jimenez-Velasco A, Martin-Subero JI, Cordeu L, Garate L, San Jose-Eneriz E, Abizanda G, Rodriguez-Otero P, Fortes P, Rifon J, Bandres E, *et al.*: Epigenetic silencing of the tumor suppressor microRNA Hsa-miR-124a regulates CDK6 expression and confers a poor prognosis in acute lymphoblastic leukemia. *Cancer Res* 2009; 69:4443–4453.

Ahmad IA: A class of Mann-Whitney-Wilcoxon type statistics. *The American Statistician* 1996; 50(4):324-327

Aitken SM, Kirsch JF: The enzymology of cystathionine biosynthesis: strategies for the control of substrate and reaction specificity. *Arch Biochem Biophys* 2005; 433(1):166–175.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* 1999; 96(12): 6745–6750.

Al-Shahrour F, Diaz-Uriarte R, Dopazo J: FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 2004; 20:578–580

Alsina B, Serras F, Baguna J, Corominas M: Patufet, the gene encoding the Drosophila melanogaster homologue of selenophosphate synthetase, is involved in imaginal disc morphogenesis. *Mol Gen Genet* 1998; 257(2):113–123.

Alvarez-Garcia I, Miska EA: Micro RNA functions in animal development and human disease. *Development* 2005; 132:4653–4662

Anderson NW, Thompson JF: Cystine lyase: beta-cystathionate from turnip roots. *Phytochemistry* 1979; 18(12):1953–1958

Antequera F: Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* 2003; 60(8):1647–1658.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardsom E, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. *Nat Genet* 2000; 25(1):25–29.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, *et al.*: Gene ontology: tool for the unification of biology. *Nature Genet* 2000; 25: 25–29.

Axley MJ, Bock A, Stadtman TC: Catalytic properties of an Escherichia coli formate dehydrogenase mutant in which sulfur replaces selenium. *Proc Natl Acad Sci U S A* 1991; 88:8450–8454

Bailey TL, Williams N, Misleh C, Li WW: MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acid Res* 2006; 34: W369–W373

Baird D, Johnstone P, Wilson T: Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics* 2004; 20(17): 3196–205.

Balstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; 19(2): 185–193

Banerjee R: Redox regulation and reaction mechanism of human cystathionine-beta-synthease: a PLP-dependent hemesensor protein. *Arch Biochem Biophys* 2005; 433(1):144–156

Barres R, Osler ME, Yan J, Rune A, Fritz T, Caidahl K, Krook A, Zierath JR: Non-CpG methylation of the PGC-1α promoter through DNMT3B controls mitochondrial density. *Cell Metabol* 2009; 10:189–198

Bartel DP: MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* 2004; 116: 281–297

Bartova E, Krejci J, Harnicarova A, Galiova G, Kozubek S: Histone modifications and nuclear architecture: a review. *J Histochem Cytochem* 2008; 56(8):711–721.

Begley TP, Downs DM, Ealick SE, McLafferty FW, van Loon AP *et al*: Thiamin biosynthesis in prokaryotes. *Arch Microbiol* 1999; 171:293–300.

Bengtsson A, Bengtsson H: Microarray image analysis: background estimation using quantile and morphological filters. *BMC Bioinformatics* 2006; 7: 96.

Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc* B 1995; 57:289–300

Berger JA, Hautaniemi S, Järvinen AK, Edgren H, Mitra SK, Astola J: Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics* 2004; 5:194.

Bestor TH: Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *EMBO J* 1992; 11(7):2611–2617.

Bhattacharyya SN, Habermacher R, Martine U, Closs EI, Filipowicz W: Stress induced reversal of microRNA repression and mRNA P-body localization in human cells. Cold Spring Harb Symp Quant Biol 2006; 71:513–521.

Bidaut G, Manion FJ, Garcia C, Ochs MF: WaveRead: automatic measurement of relative gene expression levels from microarrays using wavelet analysis. *J Biomed Inform* 2006;

39(4): 379–388.

Birringer M, Pilawa S, Flohe L: Trends in selenium biochemistry. *Nat Prod Rep* 2002; 19:693–718

Blake JA, Richardson JE, Bult CJ, Kadin JA *et al*.: MGD: the Mouse Genome Database. *Nucleic Acids Res* 2003; 31:193–195

Bolstad BM, Irizarry RA, Astrand M, Speed TP: A Comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* 2003; 19(2):185–193

Boosalis MG: The Role of Selenium in Chronic Disease. *Nutr Clin Pract* 2008; 23(2):152–160.

Borchert GM, Lanier W, Davidson BL: RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol* 2006; 13:1097–1101

Bracht J, Hunter S, Eachus R, Weeks P, Pasquinelli AE: Trans-splicing and polyadenylation of let-7 microRNA primary transcripts. RNA 2004; 10:1586–1594

Bradford JR, Hey Y, Yates T, Li Y, Pepper SD, Miller CJ: A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* 2010; 11:282

Brazma PH, Quackenbush J, Sherlock G, *et al*.: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001; 29:365–371.

Brown KM, Pickard K, Nicol F, Beckett GJ, Duthie GG, Arthur JR: Effects of organic and inorganic selenium supplementation on selenoenzyme activity in blood lymphocytes, granulocytes, platelets and erythrocytes. *Clin Sci* 2000; 98:593–599.

Brown PO, Botstein D: Exploring the new world of the genome with DNA microarrays. *Nat Genet* 1999; 21(1 Suppl):33–37.

Brueckner B, Stresemann C, Kuner R, *et al*: The human let-7a-3 locus contains an epigenetically regulated microRNA gene with oncogenic function. *Cancer Res* 2007; 67(4):1419–423.

Buck MJ, Lieb JD: ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* 2004; 83(3):349–360.

Budhraja V, Spitznagel E, Schaiff WT, Sadovsky Y: Incorporation of gene-specific variability improves expression analysis using high-density DNA microarrays. *BMC Biol* 2003; 1: 1.

Burk RF, Hill KE: Selenoprotein P: an extracellular protein with unique physical characteristics and a role in selenium homeostasis. *Annu Rev Nutr* 2005; 25:215–235

Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. *Nat Genet* 2001; 27(2):167–171.

Butler JA, Thomson CD, Whanger PD, Robinson MF: Selenium distribution in blood fractions of New Zealand women taking organic or inorganic selenium. *Am J Clin Nutr* 1991; 53:748–754.

Caban K, Copeland PR: Size matters: A view of selenocysteine incorporation from the ribosome. *Cell Mol Life Sc* 2006; 63:73–81

Cai X, Hagedorn CH, Cullen BR: Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 2004; 10:1957–1966

Cane, D.E., Hsiung, Y., Cornish, J.A., Robinson, K., and Spenser, I.D: Biosynthesis of vitamin B6: the oxidation of 4-(phosphohydroxy)-Lthreonine by PdxA. *J. Am. Chem. Soc* 1998; 120: 1936–1937.

Cao X, Pfaff SL, Gage FH: A functional study of miR-124 in the developing neural tube. *Genes Dev*. 2007; 21:531–536.

Carlson BA, Xu XM, Kryukov GV, Rao M, Berry MJ, Gladyshev VN, Hatfield DL: Identification and characterization of phosphoseryl-tTNA[Ser]Sec kinase. *Proc Natl Acad Sci U S A* 2004; 101:12848–12853

Carninci P, Sandelin A, Lenhard B, Katayama S, *et al*: Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006; 38(6):626–35.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M, Morris QD, Hughes TR: Conservation of core gene expression in vertebrate tissues. *J Biol* 2009; 8:33.

Chapple CE, Guigo R, Krol A: SECISaln, a web-based tool for the creation of structure-based alignments of eukaryotic SECIS elements. *Bioinformatics* 2009; 25(5):674–675.

Chavatte L, Brown BA, Driscoll DM: Ribosomal protein L30 is a component of the UGA-selenocysteine recoding machinery in eukaryotes. *Nat Struct Mol Biol* 2005; 12:408–416

Chen JX, Zheng Y, West M, Tang MS: Carcinogens preferentially bind at methylated CpG in the p53 mutational hot spots. *Cancer Res* 1998; 58(10):2070–2075.

Chen K, Song F, Calin GA, Wei Q, Hao X, Zhang W: Polymorphisms in microRNA targets: a gold mine for molecular epidemiology. *Carcinogenesis* 2008; 29(7):1306–1311.

Cherry JM, Adler C, Ball C, Chervitz SA *et al.*: SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 1998; 26:73–79

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg, TG, Gabrielian AE *et al.*: A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998; 2(1): 65–73.

Chua SW, Vijayakumar P, Nissom PM, Yam CY, Yang H: A novel normalization method for effective removal of systematic variation in microarray data. *Nucleic Acids Res* 2006; 34(5): e38.

Churchill GA: Using ANOVA to analyze microarray data. *Biotechniques* 2004; 37(2): 173–177.

Chvatal V: The tail of the hypergeometric distribution. *Discrete Mathematics* 1979; 25(3):285–287, 1979.

Clark LC, Combs GF Jr, Turnbull BW, Slate EH, Chalker DK *et al*: Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin. A randomized controlled trial. *JAMA* 1996; 276:1957–1963

Cleveland WS, Devlin SJ: Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J American Statistical Association*, 1988; 83: 596–610.

Combs GF: The Vitamins: Fundamental Aspects in Nutrition and Health. San Diego: Elsevier 2008

Copeland PR, Fletcher JE, Carlson BA, Hatfield DL, Driscoll DM: A novel RNA binding protein, SBP2, is required for the translation of mammalian selenoprotein mRNAs. *EMBO J* 2000; 19:306–314

Copeland PR: Regulation of gene expression by stop codon recoding: Selenocysteine. *Gene* 2003; 312:17–25

Corcoran DL, Feingold E, Dominick J, Wright M, Harnaha J, Trucco M, Giannoukakis N, Benos PV: Footer: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. *Genome Res* 2005; 15(6):840–847.

Croce CM, Calin GA: miRNAs, cancer, and stem cell division. *Cell* 2005; 122:6–7

Cullen BR: RNAi the natural way. *Nat Genet* 2005; 37:1163–1165

Dakshinamurti K: Vitamin B6. *Ann NY Acad Sci* 1990; 585

Dalmay T: MicroRNAs and cancer. *J Intern Med* 2008; 263(4):366–375

Das PM, Singal R: DNA methylation and cancer. *J Clin Oncol* 2004;22(22):4632–4642.

Down TA, Hubbard TJ: Computational detection and location of transcription start sites in mammalian genomeic DNA. *Genome Res* 2002; 12:458–461

Downs GM, Barnard JM: Hierarchical and non-hierarchical clustering. EUROMUG meeting, UK., 1995 (http://www.daylight.com/meetings/mug96/barnard/E-MUG95.html)

Drewke C, Klein M, Clade D, Arenz A, Müller R, Leistner E: 4-O-Phosphoryl-L-threonine, a substrate of the pdxC(serC) gene product involved in vitamin B6 biosynthesis. *FEBS*

Lett 1996; 390(2):179–182

Driscoll DM, Copeland PR: Mechanism and regulation of selenoprotein synthesis. *Annu Rev Nutr* 2003; 23:17–40

Duggan D, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. *Nat Genet* 1999; 21:10–14.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA *et al*: DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 2006; 38:1378–1385.

Ehrenreich A, Forchhammer K, Tormay P, Veprek B, Böck A: Selenoprotein synthesis in E. coli. Purification and characterization of the enzyme catalyzing selenium activation. *Eur J Biochem* 1992; 206(3):767–773.

Ehrlich M, Gama-Sosa MA, Huang LH, *et al*: Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* 1982; 10(8):2709–2721.

Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; 95(25): 14863–14868.

Esaki N, Nakamura T, Tanaka H, Suzuki T, Morino Y, Soda K: Enzymatic synthesis of selenocysteine in rat liver. *Biochemistry* 1981; 20(15):4492–4496.

Esteller M: Epigenetics in cancer. *N Engl J Med* 2008;358(11):1148–1159.

Fabbri M, Garzon R, Cimmino A, Liu Z, Zanesi N, Callegari E, Liu S, Alder H, *et al.*: MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. *Proc Natl Acad Sci U S A* 2007; 104:15805–10.

Fagegaltier D, Hubert N, Yamada K, Mizutani T, Carbon P, Krol A: Characterization of mSelB, a novel mammalian elongation factor for selenoprotein translation. *EMBO J* 2000; 19:4796–4805

Fairweather-Tait SJ, Collings R, Hurst R: Selenium bioavailabilty: current knowledge and future research requirements. *Am J Clin Nutr* 2010; 91(suppl):1484S–1491

Feng Z, Hu W, Hu Y, Tang MS: Acrolein is a major cigarette-related lung cancer agent: Preferential binding at p53 mutational hotspots and inhibition of DNA repair. *Proc Natl Acad Sci U S A* 2006; 103(42):15404–15409.

Fielden MR, Halgren RG, Dere E, Zacharewski TR: GP3: GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics* 2002; 18(5): 771–3.

Finley JW, Davis CD: Selenium (Se) from high-selenium broccoli is ulitized differently than selenite, selenate and selenomethionine, but is more effective in inhibiting colon carcinogenesis. *Biofactors* 2001; 14:191–196.

Fitzpatrick TB, Amrhein N, Kappes B, Macheroux P, Tews I, Raschle T: Two independent routes of de novo vitamin B6 biosynthesis: not that different after all. *Biochem J* 2007; 407(1):1–13.

Flavin M, Slaughter C: Cystathionine cleabase enzymes of neurospora. *J Biol Chem* 1964; 239:2212–2219

Flohé L, Brigelius-Flohé R, Maiorino M, Roveri A,Wissing J, Ursini F. Selenium and male reproduction. Springer 2001

Flohé L, Brigelius-Flohé R: Selenoproteins of the glutathione system. Springer 2001

Flohé L: Selenium in mammalian spermiogenesis *J Biol Chem* 2007; 388(10):987–995.

Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL: Multiplexed biochemical assays with biological chips. *Nature*, 1993; 364:555–6.

Foley E, O'Farrell PH: Functional dissection of an innate immune response by a genome-wide RNAi screen. *PLoS Biol* 2004; 2(8):e203

Forchhammer K, Böck A: Selenocysteine synthase from Escherichia coli. Analysis of the reaction sequence. *J Biol Chem* 1991; 266(10):6324–6328.

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, *et al.*: Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 2009; 10:161

Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, *et al.*: A novel feature of microsatellites in plants: A distribution gradient along the direction of transcription. *FEBS Lett* 2003; 554:17–22.

Fujita A, Sato JR, Rodrigues LO, Ferreira CE, Sogayar MC: Evaluating different methods of microarray data normalization. *BMC Bioinformatics* 2006; 7(1): 469.

Gal-Yam EN, Egger G, Iniguez L, Holster H, Einarsson S, Zhang X, Lin JC, Liang G, Jones PA, Tanay A: Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc Natl Acad Sci U S A* 2008; 105:12979–12984.

Ganichkin OM, Xu XM, Carlson BA, Mix H, Hatfield DL, Gladyshev VN, Wahl MC: Structure and catalytic mechanism of eukaryotic selenocysteine synthase. *J Biol Chem* 2008; 283(9):5849–5865.

Garzon R, Fabbri M, Cimmino A, Calin GA, Croce CM: MicroRNA expression and dunction in cancer. *Trends Mol Med* 2006; 12(12):580–587

Gehring M, Reik W, Henikoff S: DNA demethylation by DNA repair. *Trends Genet* 2009;

25(2):82–90.

Gershenzon NI, Ioshikhes IP: Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* 2005; 21(8):1295–1300.

Gladyshev VN, Diamond AM, Hatfield DL. The 15 kDa selenoprotein (Sep 15): functional studies and a role in cancer etiology. Springer 2001

Glass RS, Singh WP, Jung W, Veres Z, Scholz TD, Stadtman TC: Monoselenophosphate: synthesis, characterization, and identity with the prokaryotic biological selenium donor, compound SePX. *Biochemistry* 1993; 32(47):12555–12559.

Gonza´lez E, Danehower D, Daub ME: Vitamer levels, stress response, enzyme activity, and gene regulation of Arabidopsis lines mutant in the pyridoxine/pyridoxamine 5'-phosphate oxidase (PDX3) and the pyridoxal kinase (SOS4) genes involved in the vitamin B_6 salvage pathway. *Plant Physiol* 2007; 145(3):985–996.

Gossett WS: The probable error of a mean. *Biometrika*, 1908; 6:1–25.

Grady WM, Parkin RK, Mitchell PS, *et al*: Epigenetic silencing of the intronic microRNA hsa-miR-342 and its host gene EVL in colorectal cancer. *Oncogene* 2008; 27(27):3880–3888.

Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R: Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* 2005; 123:631–640

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2006; 34:D140–D144.

Grogan DW: Temperature-senetive murein synthesis in an Escherichia coli pdx mutant and the role of alanine racemase. *Arch Microbiol* 1988; 150:363–367

Gronbaek K, Hother C, Jones PA: Epigenetic changes in cancer. *Apmis* 2007; 115(10):1039–1059.

Groth D, Lehrach H, Hennig S: Goblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Res* 2004; 32:W313–317

Guil S, Caceres JF: The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat Struct Mol Biol* 2007; 14(7):591–596.

Guimaraes MJ, Peterson D, Vicari A, Cock, BG, Copeland NG, Gilbert DJ, Jenkins NA, Ferrick DA, Kastelein RA, Bazan JF, Zlotnik A: Identification of a novel selD homolog from eukaryotes, bacteria, and archaea: Is there an autoregulatory mechanism in selenocysteine metabolism? *Proc Natl Acad Sci U S A* 1996; 93(26):15086–15091.

Han H, Cortez CC, Yang X, Nichols PW, Jones PA, Liang G: DNA methylation directly silences gnes with non-CpG island promoters and extablishes a nucleosome occupied promoter. *Human Mol Genet* 2011; 20(22): 4299–4310

Han L, Witmer PD, Casey E, Valle D, Sukumar S: DNA methylation regulates MicroRNA expression. *Cancer Biol Ther* 2007; 6(8):1284–1288.

Hashimoto Y, Akiyama Y, Otsubo T, Shimada S, Yuasa Y: Involvement of epigenetically silenced microRNA-181c in gastric carcinogenesis. *Carcinogenesis* 2010; 31:777–784

Hatfield DL, Berry MJ, Gladyshev VN: Selenium: Its molecular biology and role in human health. 2nd edition. New York: Springer-Verlag Inc.; 2006.

Hatfield DL, Gladyshev VN: How selenium has altered our understanding of the genetic code. *Mol Cell Biol* 2002; 22:3565-3576.

Heikkinen L, Asikainen S, Wong G: Identification of phylogenetically conserved sequence motifs in microRNA 5' flanking sites from C. elegans and C. briggsae. *BMC Mol Biol* 2008; 9

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007; 39:311–318

Helmreich EJ: How pyridoxal 5'-phosphate could function in glycogen phosphorylase catalysis. *Biofactors* 1992; 3:159–172

Heyer LJ, Kruglyak S, Yooseph S: Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 1999; 9(11): 1106–1115.

Hill RE, Spenser ID: Biosynthesis of vitamin B6. In: Escherichia coli and Salmonella: cellular and molecular biology, 2nd ed., ASM Press, Washington, D.C. 1996; 695–703.

Holmgren A: Selenoproteins of the thioredoxin system. Springer 2001

Houbaviy HB, Dennis L, Jaenisch R, Sharp PA: Characterization of a highly variable eutherian microRNA gene. *RNA* 2005; 11:1245–1257

Hsiao A, Worrall DS, Olefsky JM, Subramaniam S: Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes. *Bioinformatics*, 2004; 20(17): 3108–3127.

Hudder A, Novak RF: miRNAs: effectors of environmental influences on gene expression and disease. *Toxicol Sci* 2008; 103(2):228–240

Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD: A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* 2001; 293:834–838

Ince TA, Scotto KW: A conserved downstream element defines a new class of RNA polymerase II promoters. *J Biol Chem* 1995; 270(51):30249–30252.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: Summaries of

Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003; 31(4):e15.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; 4(2): 249–264.

Ishkanian AS, Malloff CA, Watson SK *et al.*: A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet* 2004; 36:299–303

Ivan M, Harris AL, Martelli F, Kulshreshtha R: Hypoxia Response and microRNAs: No Longer Two Separate Worlds. *J Cell Mol Med* 2008.

Jaluria P, Konstantopoulos K, Betenbaugh M, Shiloach J: A perspective on microarrays: current applications, pitfalls, and potential uses. *Microbial Cell Factories* 2007; 6:4

Jiang Y, Bressler J, Beaud*et al*: Epigenetics and Human Disease, *Annu Rev Genom Human Genet* 2004; 5:479–510

Jones PA: The DNA methylation paradox. *Trends Genet* 1999; 15:34–37.

Jones TW, Kocialkowski S, Liu L, Pearson DM, Backlund LM, Ichimura K, Collins VP: Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of Pilocytic Astrocytomas. *Cancer Res* 2008; 68:8673

Kai T, Williams D, Spradling AC: The expression profile of purified Drosophila germline stem cells. *Dev Biol* 2005; 283:486–502.

Kerr MK, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. *J Comput Biol* 2000; 7(6): 819–837.

Ketting RF, Fischer SE, Bernstein E, Sijen T, Hannon GJ, Plasterk RH: Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev* 2001; 15:2654–2659

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales DR, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL: Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* 2009; 106:11667–11672

Kim IY, Veres Z, and Stadtman TC: Biochemical analysis of Escherichia coli selenophosphate synthetase mutants. Lysine 20 is essential for catalytic activity and cysteine 17/19 for 8-azido-ATP derivatization. *J Biol Chem* 1993; 268:27020–27025

Kim IY, Veres Z, and Stadtman TC: Escheichia coli mutant SELD enzymes. The cysteine 17 residue is essential for selenophosphate formation from ATP and selenide. *J Biol Chem* 1992; 267:19650–19654

Kim JY, Lee KH, Shim MS, Shim H, Xu XM, Carlson BA, Hatfield DL, Lee BJ: Human selenophosphate synthetase 1 has five splice variants with unique interactions, subcellular localizations and expression patterns. *Biochem Biophys Res Commun* 2010; 397(1):53–58

Kinzy SA, Caban K, Copeland PR: Characterization of the SECIS binding protein 2 complex required for the co-translational insertion of selenocysteine in mammals. *Nucleic Acids Res* 2005; 33:5172–5180

Kiriakidou M, Tan GS, Lamprinaki S, De Planell-Saguer M, Nelson PT, Mourelatos Z: An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* 2007; 129(6):1141–1151.

Kohonen T, Kaski S, Lappalainen H: Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation* 1997; 9: 1321–1344.

Kohrle J, Jakob F, Comtempre B, Dumont JE: Selenium, the tyroid, and the endocrine system. *Endocrine Rev* 2005; 26(7) 944–984.

Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN: Characterization of mammalian selenoproteomes. *Science* 2003; 300:1439–1443

Kryukov GV, Kumar RA, Koc A, Sun Z, Gladyshev VN. 2002. Selenoprotein R is a zinc-containing stereo-specific methionine sulfoxide reductase. *Proc Natl Acad Sci U S A* 2002; 99:4245–4250

Laber, B., Maurer, W., Scharf, S., Stepusin, K., and Schmidt, F.S: Vitamin B6 biosynthesis: formation of pyridoxine 5'-phosphate from 4-(phosphohydroxy)-L-threonine and 1-deoxy-D-xylulose-5-phosphate by PdxA and PdxJ protein. *FEBS Lett* 1999; 449: 45–48.

Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T: Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* 2001;410(6824):116–120.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: Identification of novel genes coding for small expressed RNAs. *Science* 2001; 294:853–538

Laird PW: The role of DNA methylation in cancer genetics and epigenetics, *Annu Rev Genet* 1996; 20:441–464

Lam HM, Winkler ME: Characterization of the complex pdxHtyrS operon of Escherichia coli K-12 and pleiotropic phenotypes caused by pdxH insertion mutations. *J Bacteriol* 1992; 174: 6033–6045.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, *et al.*: Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; 23: 2947–2948.

Lau PW, Macrae IJ: The molecular machines that mediate microRNA maturation. *J Cell Mol Med* 2009; 13(1):54–60.

Lee BJ, Worland PJ, Davis JN, Stadtman TC, Hatfield DL: Identification of a selenocysteyl-tRNA(Ser) in mammalian cells that recognizes the nonsense codon, UGA. *J*

Biol Chem 1989; 264:9724–9727

Lee BJ, Park SI, Park JM, Chittum HS, Hatfield DL: Molecular biology of selenium and its role in human health. *Mol Cells* 1996; 6:509–520

Lee KH, Shim MS, Kim JY, Jung HK, Lee E, Carlson BA, Xu XM, Park JM, Hatfield DL, Park TS, Lee BJ: Drosophila selenophosphate synthetase 1 regulates vitamin B6 metabolism: Prediction and confirmation. *BMC Genomics* 2011; 12:426

Lee KH, Kim HY, Lee BJ, Park K: Identification of methylation-dependent regulatory elements for intergenic miRNAs in human H4 cells. *Biophys Biochem Res Comm* 2012; 420:391–396

Lee NS, Muhs G, Wagner GC, Reynolds RD, Fisher H: Dietary pyridoxine interaction with tryptophan or histidine on brain serotonin and histamine metabolism. *Pharmacol Biochem Behav.* 1988; 29(3):559–564.

Lee RC, Feinbaum RL, Ambros V: The C.elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993; 75:843–857

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN: The nuclear RNase III Drosha initiates microRNA processing. *Nature* 2003; 425:415–419

Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, Kim VN: MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 2004; 23:4051–4060

Leinfelder W, Forchhammer K, Veprek B, Zehelein E, Böck A: In vitro synthesis of selenocysteinyl-tRNA (UCA) from seryl-tRNA (UCA): involvement and characterization of the selD gene product. *Proc Natl Acad Sci U S A* 1990; 87(2):543–547.

Leklem JE, Brown RR, Rose DP, Linkswiler H, Arend RA: Metabolism of tryptophan and niacin in oral contraceptives users receiving controlled intakes of vitamin B_6. *Am J Clin Nutr* 1975; 28(2):146–156.

Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2003; 2(2):13.

Lewis BP, Burge CB, Bartel DP: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005; 120(1):15–20.

Lewis SE: Gene Ontology: looking backwards and forwards. *Genome Biol* 2005; 6:103

Li B, Carey M, Workman JL: The role of chromatin during transcription. *Cell* 2007; 128:707–719

Lin JC, Jeong S, Liang G, Takai D, Fatemi M, Tsai YC, Egger G, Gal-Yam EN, Jones PA: Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island. *Cancer Cell* 2007; 12:432–444.

Lin SL, Miller JD, Ying SY: Intronic MicroRNA (miRNA). *J Biomed Biotechnol* 2006; 2006(4):26818.

Lobanov AV, Hatfield DL, Gladyshev VN: Selenoproteinless animals: selenophosphate synthetase SPS1 functions in a pathway unrelated to selenocysteine biosynthesis. *Protein Sci* 2008; 17(1):176–182.

Lois LM, Campos N, Putra SR, Danielsen K, Rohmer M, Boronat A: Cloning and characterization of a gene from Escherichia coli encoding a transketolase-like enzyme that catalyzes the synthesis of D-1-deoxyxylulose 5-phosphate, a common precursor for isoprenoid, thiamin, and pyridoxol biosynthesis. *Proc Natl Acad Sci U S A* 1998; 95: 2105–2110.

Longtin R: A Forgotten Debate: Is Selenocysteine the 21st Amono Acid? *J Natl Cancer Inst* 2004; 96:504–505

Loots GG, Ovcharenko I: rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res* 2004; 32(Web Server issue):W217–221.

Low SC, Berry MJ: Knowing when not to stop: Selenocysteine incorporation in eukaryotes. *Trends Biochem Sci* 1996; 21:203–208

Low SC, Harney JW, Berry MJ: Cloning and functional characterization of human selenophosphate synthetase, an essential component of selenoprotein synthesis. *J Biol Chem* 1995; 270(37):21659–21664.

Lujambio A, Esteller M: CpG island hypermethylation of tumor suppressor microRNAs in human cancer, *Cell Cycle* 2007; 6(12):1455–1459.

Maere S, Heymans K, Kuiper M: BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005; 21(16):3448–3449.

Mahony S, Auron PE, Benos PV: DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol* 2007; 3(3):e61.

Malone JH, Oliver B: Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology* 2011; 9:34

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008; 18:1509–1517

Maskos U, Southern EM: Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic Acids Res* 1992; 20 (7): 1679–1684.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, Maller JB *et al.*: Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 2008; 40:1166–1174

McCormick DB, Chen H: Update on interconversions of vitamin B_6 with its coenzyme. *J Nutr* 1996; 129:325–327.

McQuilton P, Susan E, and the FlyBase Consortium: FlyBase 101 – the basics of navigating FlyBase. *Nucleic Acids Res* 2012; 40(Database issue):D706–714.

Medigue C, Rechenmann F, Danchin A, Viari A: Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics* 1999; 15(1): 2–15.

Meier M, Janosik M, Kery V, Kraus JP, Burkhard P: Structure of human cystathionine β-synthase: a unique pyridoxal 5'-phophate-dependent heme protein. *EMBO J* 2001; 20(15):3910–3916.

Messeguer X, Escudero R, Farré D, Nuñez O, Martínez J, Albà MM: PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics* 2002; 18:333–334

Mihara, H., Kurihara, T., Yoshimura, T., Soda, K., and Esaki, N: Cysteine sulfinate desulfinase, a NIFS-like protein of Escherichia coli with selenocysteine lyase and cysteine desulfurase activities. Gene cloning, purification, and characterization of a novel pyridoxal enzyme. *J Biol Chem* 1997; 272: 22417–22424.

Milani L, Fredriksson M, Syvanen AC: Detection of alternatively spliced transcripts in Leukemia cell lines by minisequencing on microarrays. *Clinical Chem* 2006; 52(2):202–211

Mishra PJ, Humeniuk R, Mishra PJ, Longo-Sorbello GS, Banerjee D, Bertino JR: A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. *Proc Natl Acad Sci U S A* 2007; 104(33):13513–13518.

Mittenhuber G: Phylogenetic analyses and comparative genomics of vitamin B6 (pyridoxine) and pyridoxal phosphate biosynthesis pathway. *J Mol Microbiol Biotechnol* 2001; 3(1):1–20

Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR: Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 2005; 85(1): 1–15.

Molina C, Grotewold E: Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* 2005; 6:25.

Mooney S, Leuendorf JE, Hendrickson C, Hellmann H: Vitamin B6: A long known compound of surprising complexity. *Molecules* 2009; 14(1):329–351

Morey M, Corominas M, Serras F: DIAP1 suppresses ROS-induced apoptosis caused by impairment of the selD/sps1 homolog in Drosophila. *J Cell Sci* 2003; 116(22):4597–4604.

Morey M, Serras F, Baguñà J, Hafen E, Corominas M: Modulation of the Ras/MAPK signalling pathway by the redox function of selenoproteins in Drosophila melanogaster. *Dev Biol* 2001; 238(1):145–156.

Morgan BJ, Ray AP: Non-uniqueness and Inversions in Cluster Analysis. *Appl Stat* 1995; 44(1):117–134.

Musayev FN, Di Salvo ML, Ko TP, Schirch V, Safo MK: Structure and properties of recombinant human pyridoxine 5'-phosphate oxidase. *Protein Sci* 2003; 12(7):1455–1463

Neves R, Scheel C, Weinhold S, Honisch E, Iwaniuk KM, Trompeter HI, Niederacher D, Wernet P, Santourlidis S, Uhrberg M: Role of DNA methylation in mir-200c/141 cluster silencing in invasive breast cancer cells. *BMC Research Note* 2010; 3:219

Ng SY, Gunning P, Liu SH, Leavitt J, Kedes L: Regulation of the human beta-actin promoter by upstream and intron domains. *Nucleic Acids Res* 1989; 17(2):601–615.

Nikolov DB, Burley SK: RNA polymerase II transcription initiation: a structural view. *Proc Natl Acad Sci U S A* 1997; 94(1):15–22.

Noonan EJ, Place RF, Pookot D, Basak S, Whitson JM, Hirata H, Giardina C, Dahiya R: miR-449a targets HDAC-1 and induces growth arrest in prostate cancer. *Oncogene* 2009; 28:1714–1724.

Notheis C, Drewke C, Leistner E: Purification and characterization of the pyridoxol-5'-phosphate:oxygen oxidoreductase (deaminating) from Escherichia coli. *Biochim Biophys Acta* 1995; 1247: 265–271.

Ohler U, Uekta S, Lim LP, Bartel DP, Burge CB: Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* 2004; 10:1309–1322

Okano M, Bell DW, Haber DA, Li E: DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 1999; 99(3):247–257.

Okitsu CY, Hsieh JCF, Hsieh C-L: Transcriptional activity affacts the H3K4me3 Level and distribution in the coding region. *Mol Cel Biol* 2010; 30:2933–2946

Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, Zhang X, Song JS, Fisher DE: Chromatin structure analyses identify miRNA promoters. *Genes Dev* 2008; 22:3172–3183

Park T, Yi SG, Lee S, Lee S, Yoo DH, Ahn JI, Lee YS: Statistical tests for identifying differentially expressed genes in time course microarray experiments. *Bioinformatics*, 2003; 19(6):694–703

Patterson TA, Lobenhofe EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, Fang H *et al.*: Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat Biotechnol* 2006; 24(9):1140–1150.

Pavlidis P, Lewis DL, Noble WS: Exploring gene expression data with class scores.

Proceedings of the Pacific Symposium on Biocomputing 2002; 474–485

Pavlidis P: Using ANOVA for gene selection from microarray studies of the nervous system. *Methods* 2003; 31(4): 282–289.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Foder SP: Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A*, 1994; 91, 5022–5026

Perry C, Yu S, Chen J, Matharu KS, Stover PJ: Effect of vitamin B6 availability on serine hydroxymethyltransferase in MCF-7 cells. *Arch Biochem Biophys* 2007; 462(1):21–27

Quackenbush J: Microarray data normalization and transformation. *Nat Genet* 2002;. 32 Suppl: 496–501.

Quandt K, Frech K, Karas H, Wingender E, Werner T: MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* 1995; 23(23):4878–4884.

Rajewsky N, Socci ND: Computational identification of microRNA targets, *Dev Biol* 2004; 267(2): 529–535

Rederstorff M, Krol A, Lescure A: Understanding the importance of selenium and selenoproteins in muscle function. *Cell Mol Life Sc* 2006; 63:52–59

Rideout WM, Coetzee GA, Olumi AF, Jones PA: 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* 1990; 249(4974):1288–1290.

Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A: Identification of mammalian microRNA host genes and transcription units. *Genome Res* 2004; 14:1902–1910.

Roth FP, Hughes JD, Estep PW, Church GM: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 1998; 16(10):939–945.

Rouhi A, Mager DL, Humphries RK, Kuchenbauer F: MiRNAs, epigenetics, and cancer. *Mamm Genome* 2008.

Royo H, Cavaille J. Non-coding RNAs in imprinted gene clusters: *Biol Cell* 2008; 100(3):149–166.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J: TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 2003; 34(2): 374–378.

Saini HK, Griffiths-Jones S, Enright AJ: Genomics analysis of human microRNA transcripts. *Proc Natl Acad Sci U S A* 2007; 104:17719–17724

Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; 4: 406–425

Sandelin A, Wasserman WW: Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* 2004; 338(2):207–215.

Santos-Rosa H, Schneider R, Bannister AJ, *et al*: Active genes are tri-methylated at K4 of histone H3. *Nature* 2002;419(6905):407–411.

Schaeffer MC, Gretz D, Gietzen DW, Rogers QR: Dietary excess of vitamin B6 affects the concentrations of amino acids in the caudate nucleus and serum and the binding properties of serotonin receptors in the brain cortex of rats. *J Nutr.* 1998; 128(10):1829–1835.

Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; 270:467–470.

Schmittgen TD, Livak KJ: Analyzing real-time PCR data by the comparative CT method. *Nat Protoc* 2008; 3(6):1101–1108

Schrauzer GN. Selenomethionine: a review of its nutritional significance, metabolism and toxicity. *J Nutr* 2000; 130:1653–1656.

Scott GK, Mattie MD, Berger CE, Benz SC, Benz CC: Rapid alteration of microRNA levels by histone deacetylase inhibition. *Cancer Res* 2006; 66(3):1277–1281.

Seitz H, Royo H, Bortolin ML, Lin SP, Ferguson-Smith AC, Cavaille J: A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res* 2004; 14(9):1741–1748.

Shalon D, Smith SJ, Brown PO: A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996; 6(7):639–645

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; 13(11):2498–2504

Shi H, Wang MX, Caldwell CW: CpG islands: their potential as biomarkers for cancer. *Expert Rev Mol Diagn* 2007; 7(5):519–531.

Shim MS, Kim JY, Jung HK, Lee KH, Xu XM, Carlson BA, Kim KW, Kim IY, Hatfield DL, Lee BJ: Elevation of glutamine level by selenophosphate synthetase 1 knockdown induces megamitochondrial formation in Drosophila cells. *J Biol Chem* 2009; 284(47):32881–32894.

Shiu SH, Borevitz JO: The next generation of microarray research: applications in evolutionary and ecological genomics Heredity 2008; 100:141–149

Silber J, Lim DA, Petritsch C, Persson AI, Maunakea AK, Yu M, Vandenberg SR *et al*: miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. *BMC Med* 2008; 6:14.

Smale ST, Kadonaga JT: The RNA polymerase II core promoter. *Annu Rev Biochem* 2003; 72:449–479.

Small-Howard A, Morozova N, Stoytcheva Z, Forry EP, Mansell JB, Harney JW, Carlson BA, Xu XM, Hatfield DL, Berry MJ: Supramolecular complexes mediate selenocysteine incorporation in vivo. *Mol Cell Biol* 2006; 26(6):2337–2346.

Soda K, Oikawa T, Esaki N: Vitamin B6 enzymes participating in selenium amino acid metabolism. *BioFactors* 1999; 10(2):257– 262.

St. Germain DL: Selenium, deiodinases and endocrine function. Springer 2001

Stadtman TC: Selenium biochemistry. Mammalian selenoenzymes. *Ann NY cad Sci* 2000; 99:399–402

Stadtman TC: Selenocysteine. *Annu. Rev. Biochem*. 1996; 65:83–100

Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000; 16(1):16–23.

Stoughton RB: Applications of DNA microarrays in biology. *Ann Rev Biochem* 2005; 74:53–82

Suzuki H, Takatsuka S, Akashi H, Yamamoto E, Nojima M, Maruyama R, Kai M, Yamano H, Sasaki Y, Tokino T, Shinomura Y, Imai K, Toyota M: Genome-wide profiling of chromatin signatures reveals epigenetic regulation of microRNA genes in colorectal cancer. *Cancer Res* 2011; 71:5646–5658

Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 2002; 99(6):3740–3745.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 1999; 96(6):2907–2912.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* 2011; 28:2731–2739.

Tamura T, Yamamoto S, Takahata M, Sakaguchi H, Tanaka H, Stadtman TC, Inagaki K: Selenophosphate synthetase genes from lung adenocarcinoma cells: Sps1 for recycling L-selenocysteine and Sps2 for selenite assimilation. *Proc Natl Acad Sci U S A* 2004; 101(46):16162–16167.

Tang T, François N, Glatigny A, Agier N, Mucchielli MH, Aggerbeck L, Delacroix H: Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment. *Bioinformatics* 2007; 23(20):2686–2691.

The Gene Ontology Consortium: Creating the Gene Ontology Resource: design and implementation. *Genome Res* 2001; 11:1425–1433.

Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, Brouwn P: Clustering Methods for the Analysis of DNA Microarray Data. Stanford University, 1999.

Tobe R, Mihara H, Kurihara T, Esaki N: Identification of proteins interacting with selenocysteine lyase. *Biosci Biotechnol Biochem* 2009; 73(5):1230–1232.

Toyota M, Suzuki H, Sasaki Y, Maruyama R, Imai K, Shinomura Y: Epigenetic silencing of microRNA-34b/c and B-cell translocation gene 4 is associated with CpG island methylation in colorectal cancer. *Cancer Res* 2008; 68:123–132

Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* 2001; 29(12): 2549–2557.

Tujebajeva RM, Copeland PR, Xu XM, Carlson BA, Harney JW, Driscoll DM, Hatfield DL, Berry MJ: Decoding apparatus for eukaryotic selenocysteine insertion. *EMBO Rep* 2000; 1:158–163

Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001; 98(9):5116–5121.

Umetani M, Mataki C, Minegishi N, Yamamoto M, Hamakubo T, and Kodama T: Function of GATA transcription factors in induction of endothelial vascular cell adhesion molecule-1 by tumor necrosis factor-alpha. *Arterioscler Thromb Vasc Biol* 2001; 21(6):917–922.

Ursini F, Heim S, Kiess M, Maiorino M, Roveri A, *et al*: Dual function of the selenoprotein PHGPx during sperm maturation. *Science* 1999; 285:1393–1396

Verducci JS, Melfi VF, Lin S, Wang Z, Roy S, Sen CK: Microarray analysis of gene expression: considerations in data mining and statistical treatment. *Physiol Genomics* 2006; 25: 355–363

Wang D, Huang J, Xie H, Manzella L, Soares MB: A robust two-way semi-linear model for normalization of cDNA microarray data. *BMC Bioinformatics* 2005; 6:14.

Wang, J, Ma JZ, Li MD: Normalization of cDNA microarray data using wavelet regressions. Comb Chem High Throughput Screen, 2004; 7(8): 783–791.

Weber B, Stresemann C, Brueckner B, Lyko F: Methylation of human microRNA genes in normal and neoplastic cells. *Cell Cycle* 2007; 6(9):1001–1005.

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D: Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 2007; 39:457–466.

Westfall PH, Young SS: Resampling-based multiple testing: examples and methods for p-value adjustment. New York, Wiley 1993

Whanger PD: Selenoprotein W. *Methods Enzymol*. 2002; 347:179–187

Wilting SM, van Boerdonk RA, Henken FE, Meijer CJ, Diosdado B, Meijer GA, le Sage C, Agami R, Snijders PJF, Steenbergen RD: Methylation-mediated silencing and tumour suppressive function of hsa-miR-124 in cervical cancer. *Mol Cancer* 2010; 9:167

Xu XM, Carlson BA, Irons R, Mix H, Zhong N, Gladyshev VN, Hatfield DL: Selenophosphate synthetase 2 is essential for selenoprotein biosynthesis. *Biochem J* 2007a; 404(1):115–120.

Xu XM, Carlson BA, Mix H, Zhang Y, Saira K, Glass RS, Berry MJ, Gladyshev VN, Hatfield DL: Biosynthesis of selenocysteine on its tRNA in eukaryotes. *PLoS Biol* 2007b; 5(1):e4.

Xu XM, Carlson BA, Zhang Y, Mix H, Kryukov GV, Glass RS, Berry MJ, Gladyshev VN, Hatfield DL: New developments in selenium biochemistry: selenocysteine biosynthesis in eukaryotes and archaea. *Biol Trace Elem Res* 2007c; 119(3):234–241.

Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E: Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* 2007; 389(1) :52–65.

Yang Y, Zhao G, Man TK, Winkler ME: Involvement of the gapA- and epd (gapB)-encoded dehydrogenases in pyridoxal 5'-phosphate coenzyme biosynthesis in Escherichia coli K-12. *J Bacteriol* 1998; 180: 4294–4299.

Yang Y, Zhao G, Winkler ME: Identification of the pdxK gene that encodes pyridoxine (vitamin B6) kinase in Escherichia coli K-12. *FEMS Microbiol Lett* 1996; 141: 89–95.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002; 30(4): e15.

Yoo AS, Staahl BT, Chen L, Crabtree GR: MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. *Nature* 2009; 460: 642–6.

Yoon D, Yi SG, Kim JH, Park TS: Two-stage normalization using background intensities in cDNA microarray data. *BMC Bioinformatics* 2004; 5: 97.

Yoon JH, Smith LE, Feng Z, Tang M, Lee CS, Pfeifer GP: Methylated CpG dinucleotides are the preferential targets for G-to-T transversion mutations induced by benzo[a]pyrene diol epoxide in mammalian cells: similarities with the p53 mutation spectrum in smoking-associated lung cancers. *Cancer Res* 2001; 61(19):7110–7117.

Yoshimura T, Goto M: D-amino acids in the brain: structure and function of pyridoxal phosphate-dependent racemases. *FEBS J* 2008; 275(14):3527–3537.

Yuan J, Palioura S, Salazar JC, Su D, O'donoghue P, Hohn MJ, Cardoso AM, Whitman WB, Soll D: RNA-dependent conversion of phosphoserine forms selenocysteine in eukaryotes and archaea. *Proc Natl Acad Sci U S A* 2006; 103:18923–18927

Zangi R, Arrieta A, Cossio FP: Mechanism of DNA methylation: the double role of DNA as a substrate and as a cofactor, *J Mol Biol* 2010; 400(3):632–644

Zar JH: Biostatistical Analysis, 4th ed. Upper Saddle River, NJ. Prentice-Hall. 1999

Zhao G, Winkler ME: 4-Phospho-hydroxy-L-threonine is an obligatory intermediate in pyridoxal 5'-phosphate coenzyme biosynthesis in Escherichia coli K-12. *FEMS Microbiol Lett* 1996; 135: 275–280.

Zhao G, Winkler ME: An Escherichia coli K-12 tktA tktB mutant deficient in transketolase activity requires pyridoxine (vitamin B6) as well as the aromatic amino acids and vitamins for growth. *J Bacteriol* 1994; 176:6134–6138.

Zheng M, Barrera LO, Ren B, Wu YN: ChIP-chip: Data, Model, and Analysis. *Biometrics* 2007, 63(3):787–796

Zhou X, Ruan J, Wang G, Zhang W: Characterization and identification of microRNA core promoters in four model species. *PLoS Comp Biol* 2007; 3:e37

Zhu JK: Active DNA demethylation mediated by DNA glycosylases. *Annu Rev Genet* 2009; 43:143–166.

국문초록

생물정보학은 대량의 데이터를 컴퓨터 기술과 통계적 이론을 통해 생물학적 의미를 해석하고 예측하는 중요한 생물학의 한 분야이다. 마이크로어레이 기술의 발전으로 대량의 연구 결과가 축적됨에 따라 이를 생물정보학적 방법을 통해 분석하여 유전자의 기능 및 조절 서열을 예측하고자 하는 연구가 활발히 진행되고 있다. 이 논문에서는 생물정보학적 방법으로 마이크로어레이 자료를 분석하여 셀레늄 인산 합성효소 1 (SPS1)의 유전자의 기능과 마이크로RNA의 전사조절인자에 대한 예측과 실험적으로 검증한 연구 결과를 제시한다.

마이크로어레이 데이터를 이용하여 유전자의 기능을 예측한 연구는 *Sps1* 유전자를 모델로 이용하였다. 고등 진핵생물에는 두 가지 형태의 셀레늄 인산 합성효소 (SPS), 즉 SPS1과 SPS2가 존재하는데, 두 가지 형태의 효소 가운데 단지 SPS2만이 셀레늄 인산 합성반응을 촉매 하는 것으로 알려져 있으며, SPS1의 기능은 아직 명확하지 않다. 그러나, SPS1은 세포의 성장과 배아 발달 등 세포 활동에 필수적인 기능을 가진 것으로 알려져 있다. 따라서, SPS1의 세포 내 기능을 생물정보학적 방법을 이용하여 예측하고 실험적으로 확인하였다. 초파리 세포를 이용하여 *Sps1* 유전자의 발현을 억제시키고 각각 1, 3, 5일이 경과한 후에 메신저RNA을 분리하여 마이크로어레이 실험을 통해 정상과 차이를 보이는 유전자를 이원분산분석법을 통해 동정하였다. 선별된 유전자들을 시간에 따라 변화되는 발현량의 차이를 SOM 클러스터링 방법으로 3x2 그리드에서 클러스터링한 후, 초기와 후기에 발현이 변하는 유전자들을 이용하여 유전자셋을 만든 후, 이들이 가지는 생물학적으로 유의한 기능을 gene ontology (GO)를 이용하여 통계적으로 유추하였다. GO 분석 결과 비타민B_6

합성에 관여하는 유전자가 가장 초기에 (*Sps1* 유전자 발현 억제 후 3일 째) 유의한 변화를 보이는 것을 발견하였고, 후반부에는 (5일 째) 아미노산 대사 및 면역과 관련된 유전자들의 변화가 크게 나타나는 것을 관찰할 수 있었다. GO분석 SPS1의 일차 타깃이 비타민B_6의 합성과 관련되어 있다는 예측 결과를 실험적으로 검증하였다. 흥미로운 점은 SPS1의 발현을 억제시켰을 때, 이미 보고된 바와 같이 거대 미토콘드리아 표현형을 보인 연구 결과와 동일한 표현형을 보임과 동시에 비타민B_6의 활성 형태인 피리독살인산 (PLP)의 농도가 또한 현저히 감소되는 것을 확인하였다. 아울러 PLP 농도를 인위적으로 감소시켰을 때, SPS1 유전자의 발현을 억제한 경우와 매우 유사하게 아미노산 대사와 면역에 관련된 유전자들의 발현이 감소되는 현상을 보였다. 이러한 결과를 통해 SPS1이 아미노산대사와 면역과 같은 세포 내 중요한 대사 활성에 영향을 주는 비타민B_6의 합성을 조절하는 기능을 하고 있음을 증명하였다.

생물정보학적으로 유전자 발현을 조절하는 조절 서열을 예측은 마이크로RNA를 모델로 연구하였다. 마이크로RNA는 표적 유전자들이 단백질로 번역되는 과정을 조절하는 중요한 전사 후 조절 기능을 하는, 평균 22염기로 이루어진 짧은 RNA분자이다. 이들은 알려진 유전자의 내부 서열에 존재하는 인트로닉 마이크로RNA와 유전자가 존재하지 않는 지역에 독자적으로 존재하는 인터제닉 마이크로RNA로 크게 구분되는데, 인트로닉 마이크로RNA의 경우, 그들이 속해 있는 호스트 유전자의 발현 조절인자를 공유하여 발현되는데 반해, 인터제닉 마이크로RNA의 발현에 대한 기작은 아직 완전히 이해하지 못하고 있다. 따라서, 인터제닉 마이크로RNA에 대한 프로모터 또는 조절 서열을 예측하기 위하여 후생유전학 정보, 특히 DNA 메틸화 정보를 이용하였다. 신경교종세포의 마이크로RNA의 상부 1kb 부위에서 메틸화가 많이 되어 있는 부위를 마이크로어레이 기술을

이용하여 동정한 후, 이들 서열을 다량서열정렬방법과 모티프 검색방법을 적용하여 통계적으로 매우 유의한 결과를 보이는 C[N]$_6$CT 모티프 서열을 발견하였다. 인터제닉 miRNA들의 1 kb 상위 서열에서 C[N]$_6$CT 모티프 서열의 분포를 조사한 결과, 이 모티프가 메틸화가 많이 되어 있는 부위에 밀집되어 존재하는 miRNA에서부터 그렇지 않은 miRNA에 이르기까지 다양한 양상을 보였다. 이 모티프가 인터제닉 마이크로RNA의 발현에 관여한다는 사실을 증명하기 위해 실험적 검증을 진행하였다. 메틸화 서열에 C[N]$_6$CT 모티프 서열이 가장 많이 밀집되어 있는 has-mir-200c를 비롯하여 메틸화 서열에서 이 모티프를 하나도 포함하지 않는 has-mir-3188 등 8개의 인터제닉 miRNA들을 이용하여 메틸전이효소 억제제를 처리한 전후에 발현 정도를 qPCR를 통해 실험적으로 측정하여 본 결과, 메틸화가 된 부위에 C[N]$_6$CT 모티프 서열의 출현 빈도가 높은 인터제닉 마이크로RNA의 발현은 크게 증가하는 양상을 보인 반면, 빈도가 낮은 인터제닉 마이크로RNA의 경우 그렇지 못했다. 마이크로RNA의 발현과 메틸화된 부위에서의 C[N]$_6$CT 모티프 서열의 밀집 정도에 대해 피어슨 상관분석을 통해 분석한 결과, 상관계수가 0.87로 매우 밀접한 상관관계가 있음을 보여 주었다. 이 연구 결과는 DNA 메틸화와 밀접히 관련된 C[N]$_6$CT 모티프 서열을 발견함으로서 이 모티프가 인터제닉 마이크로RNA의 발현 조절과 밀접한 관련이 있음을 생물정보학적인 방법으로 예측하였으며, 또한 새로운 인터제닉 마이크로RNA의 발견을 위한 중요한 단서를 제공해 주는 결과이다.

마이크로어레이 자료를 이용하여 생물정보학적 방법을 통해 유전자의 기능과 전사 조절인자의 예측에 대한 두 가지 예측 연구를 진행하고 이를 확인한 결과, 예측 결과가 매우 정확성이 높음을 확인할 수 있었다. 이는 지금까지 개발된 다양한 생물정보학적 이론의 적절한 조합을

통해 보다 높은 확률로 여러 가지 연구 분야에 적용할 수 있음을 말해주는 연구 결과이다. 최근에는 차세대서열분석이라는 새로운 기술이 등장함에 따라, 마이크로어레이를 통한 생물학적 연구가 다소 사양되고 있으나, 두 가지 기술을 병행하여 다양한 생물정보학적 도구 및 모델을 적용시켜 분석한다면 보다 정확하고 생물학적으로 의미있는 해석을 위한 결과를 도출할 수 있을 것으로 기대한다.

주요어 : 셀레늄인산 합성효소 1, 비타민B_6, 마이크로RNA, 프로모터
학 번 : 2007-30782