1
2007

This is to certify that the
thesis entitled

Random Forests and Gene Selection to Classify Arabidopsis Thaliana Ecotypes

presented by

Hsueh-han Yeh

has been accepted towards fulfillment
of the requirements for the

M.S. degree in Statistics and Probability

Marianne Huebner
Major Professor's Signature

8-20-07
Date

Random Forests and Gene Selection to Classify Arabidopsis Thaliana Ecotypes

By
Hsueh-han Yeh

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Statistics and Probability

2007

ABSTRACT

Random Forests and Gene Selection to Classify Arabidopsis Thaliana Ecotypes

By

Hsueh-han Yeh

This thesis discusses the classification and gene selection of ecotype data for Arabidopsis thaliana. Gene expressions from Oligonucleotide gene expression arrays were used to classify Arabidopsis thaliana ecotypes using statistical methods. The hierarchical cluster method was used to group ecotypes according to latitude and altitude to distinguish ecotypes. Limma was used to select differentially expressed genes. The Random Forest algorithm provides a ranking of genes to indicate how well they can discriminate between ecotypes.

We focus on the Random Forest algorithm. It is an efficient approach and can deal with a large number of predictor variables in a classification process. Parameters are optimal to achieve a small classification error rate.

The final selection of genes may play an important role in adaptation to stress conditions. They were further examined for gene function and evidence regarding stress resistance.

Keywords: Arabidopsis thaliana, Microarray Data, Hierarchical Cluster, Limma, Random Forest, Classification.

ACKNOWLEDGEMENTS

I wish to thank many people who made this thesis possible. First of all, it is hard to overstate my gratitude to my advisor, Dr. Marianne Huebner, Department of Statistics, Michigan State University. With her enthusiasm, her patience, and her encouragement, she helped to make statistics and biology fun for me. Throughout my thesis-writing period, she provided many suggestions and lots of good ideas. Dr. Huebner also helped me revising my English. I am very glad and enjoyable to work with her. I wish thank to Dr. Andreas Weber for his support and grant. Dr. Weber also gave me suggestions to examine gene functions which makes this thesis complete. I wish to thank my parents. They raised me, supported me, taught me, and loved me. To them I dedicate this thesis. I wish to thank my best friend Hsiu-ching Chang, for helping me get through the difficult times, and for all the emotional support. My special gratitude is due to my brother, for his loving support. I also wish to thank William Robert Swindell for giving many helpful suggestions of biology section.

Finally, I have to say 'Thank You' to all my friends and family, wherever they are and where they go.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction of Microarray and Arabidopsis Ecotypes Data

1.1. Microarray Data

Regulatory regions of plant genes is likely to be more concise than those of animal genes, but the transcription factors encoded in plant genomes is larger than those of animals. Thus, plants can contribute to research regarding the influence of transcriptional factors in multicellular development. Here, we study the reference plant, *Arabidopsis thaliana*, for our study, and the dataset is AtGenExpress Ecotypes Expression estimated by gcRMA. The data is part of the public AtGenExpress expression atlas, which was created by Affymetrix ATH1 array platform. Microarray, obtained by Oligonucleotide Chips or spotted arrays, is a technology to study the expression of thousands of genes. Microarray technology requires statistical methods to analyze the dataset which are high dimensional data sets.

Statistical approaches can be used for multiple comparisons of genes to define the differentially expressed genes between arrays. Data mining is used widely for Microarray data since it can use a subgroup of genes to predict the observations (e.g. Ecotypes) that would help to reduce the dimension of Microarray data. In this study, we use classification approach and data mining technique, Random Forest, to classify the Arabidopsis thaliana Expression Ecotypes Data.

1.2. Arabidopsis Data

Arabidopsis thaliana

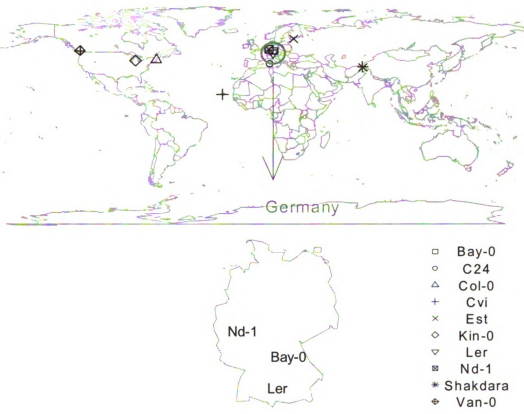
The Arabidopsis ATH1 Genome Array, built in TIGR (The Institute for Genomic Research), contains more than 22,500 probe sets displaying approximately 24,000 gene sequences on a single array. (http://www.affymetrix.com)

Arabidopsis thaliana is a flowering plant, an inconspicuous weed. It has been used as a model plant organism for many years and has been chosen for used in molecular genetic analysis. *Laibach (1943)* first specify that some significant characteristics of *Arabidopsis thaliana* make them are suitably used for model plant organism. It has a short life cycle; it only needs several weeks to mature. Due to its size, it can grow in a limited area. Furthermore, it has small genome size and nearly non-repetitive DNA (*S Barth, A E Melchinger, TH Lübberstedt, 2002*). These features make *Arabidopsis thaliana* plants much conveniently for genetic analysis. Due to these features in *Arabidopsis thaliana*, international effort has been devoted to build the methods to research its genome.

Arabidopsis thaliana at an early stage of flowering. [Drawing by K. Sutliff]

Arabidopsis thaliana Ecotype Data

Figure 1 First 10 Ecotypes Distribution Map



An ecotype is a population of a plant that survives as a distinct group through ecological environment. AtGenExpress Ecotypes Data used in this paper come from *weigelworld (www.weigelworld.org),* including 34 ecotypes. Each ecotype is composed by one or several arrays of 22810 genes each. Arabidopsis thaliana is widely distributed *(Meinke et al, 1998)*, and the 34 ecotypes in the Arabidopsis thaliana Ecotype data used in this study represent locations in Europe, North America and Africa. The location, longitude, latitude, and altitude of each ecotype were listed in the Table1. The latitudes of these ecotypes range from 16N to 59N. The longitudes range from 0.53E to 73E, and from 0.22W to 123W. The highest altitude is 3400m. Overview the distribution of the

ecotypes, 27 ecotypes distributed throughout Europe and 12 ecotypes among these 27 ecotypes in Germany. The other ecotypes are distributed in North America and Africa. We want to examine if we can use these gene expressions to classify Arabidopsis thaliana ecotypes by statistical methods. First of all, the problem we confront is the large size of genes in each ecotype. Dimension reduction can help deal with large variables efficiently and select the most important variables. We use Random Forest to decrease the size of dataset and classify ecotypes. Random Forest Algorithm will be discussed in the Chapter 2.

Table 1 Ecotypes Geography Information

Ecotype	Location	Altitude	Latitude	Longitude	Temperature (℃)
Bay-0	Bayreuth, Germany	350	49N	11 E	-2 – 18
C24	Coimbra, Portugal	179	40N	8 E	7.2 – 27
Col-0	Columbia University (U.S.)	49	39N	93 W	-3.3 – 28.9
Cvi	Cape Verde Islands	43	16N	24 W	24 – 29
Est	Estonia	15	59N	26 E	-5.2 – 17
Kin-0	Kinneville, MI	273	43N	85 W	-12.2 – 32.2
Ler	Landsberg, Germany	628	53N	16 E	-1.7 – 19.4
Nd-1	Niederzunzheim, Germany	250	50N	8 E	5.5 – 9.5
Shakdara	Pamiro-Alay, Tadjikistan	3400	37N	71 E	0 – 30
Van-0	UBC (Vancouver)	50	50N	123 W	0 – 26

Table 2 Resources of Arabidopsis thaliana GenomeL

Resources	Contact Person	Information of website
Arabidopsis database (AtDB)	M. Cherry	http://genome-www.stanford.edu/Arabidopsis/
ABRC* Stock Center (USA)	R. Scholl	http://aims.cps.msu.edu/aims
NASC† Stock Centre (UK)	M. Anderson	http://nasc.nott.ac.uk
TIGR‡ (USA)	S. Rounsley	http://www.tigr.org/tdb/at/at.html
SPP§ Consortium (USA)	R. Davis	http://sequence-www.stanford.edu/ara/SPP.html
CSHL\ Consortium (USA)	R. McCombie	http://nucleus.cshl.org/protarab/
ESSAConsortium (Europe)	M. Bevan	http://muntjac.mips.biochem.mpg.de/arabi/index.html
Genoscope (France)	F. Quetier	http://www.genoscope.cns.fr/externe/arabidopsis/Arabidopsis.html
Kazusa Institute (Japan)	S. Tabata	http://www.kazusa.or.jp/arabi/

David W. Meinke, J. Michael Cherry,* Caroline Dean, Steven D. Rounsley, Maarten Koornneef. Arabidopsis thaliana: A Model Plant for Genome Analysis (1998)

1.3. Gene Selection Process

1 Grouping 10 ecotypes (3 replications each) by latitude and altitude of first 10 ecotypes of Arabidopsis thaliana Ecotypes Data using Hierarchical Cluster. Four groups are as follows:

La4 (La-A, La-B, La-C, La-D) **Al4** (Al-A, Al-B, Al-C, Al-D)

2 For each of these two groupings (Al4 and La4) with *Limma* function of ***R*** software. A-B, A-C, A-D, B-C, B-D, and C-D in each of the grouping La4 and Al4, respectively.

3 The number of significant genes for each contrast in each grouping is counted.

4 After counting the number of significant genes, we found that *Cvi* (La-D) has the largest number of significant genes differentially expressed in comparison with other 3 latitude groups. *Shakdara* (Al-A) has the largest number of significant genes differentially expressed in comparison with other 3 altitude groups.

5 *Cvi* (smallest latitude) and *Shakdara* (highest altitude) are compared to the other ecotypes to identify genes that differentiate these.

6 Contrasts to be considered :

- $Cvi - \frac{1}{8}(Bayo + C24 + Colo + Est + Kino + Ler + Nd1 + Vano)$
- $Sha - \frac{1}{8}(Bayo + C24 + Colo + Est + Kino + Ler + Nd1 + Vano)$

The top 500 differently expressed genes are selected from each of these two contrasts. Corresponding gene sets are *Cvi500* and *Sha500.*

7 Optimal parameters, *ntree* and *mtry,* in Random Forest are chosen for *Cvi500* and *Sha500.*

8 Highly ranked genes (variable importance) are selected from *Cvi500* and *Sha500.*

There are 43 genes chosen from Cvi500 and 84 genes chosen from Sha500.

9 Compare OOB error rate for the selected genes.

10 Discuss misclassification arrays in Random Forest.

11 Gene functions of the selected genes are considered.

Chapter 2 Statistical Methodology

In this chapter, clustering (2.1), linear models for Microarray Data (2.2), and Random Forest (2.3) will be discussed.

(2.1) Clustering is the first step in our gene selection process. In this section, we use Hierarchical Clustering method to group the 10 ecotypes into subsets and those subsets will be contrasted with linear models.

(2.2) Limma is the second step. In this step, we choose smaller subgroups of genes which are differentially expressed from Limma method by contrasting subsets of ecotypes obtained in clustering result. We explain the differentially expressed genes.

(2.3) Random Forest is a method to rank genes by their importance in classifying ecotypes. In this section, we will explain the Random Forest algorithm and the selection of important predictor variables (genes) from the gene sets chosen with the linear models.

2.1. Clustering

Grouping a collection of observations into subgroups (clusters) is called Clustering. Observations within the each cluster have smaller distance to each other than to observations assigned to other different clusters.

In Hierarchical Clustering (Jinwook Seo, Ben Shneiderman 2002), the observations are not separated into subgroups in only one step. Instead, observations are separated by a serious of partitions. Clustering may start from a single cluster containing all observations to subgroups of observations, called *Divisive method.* On the other hand (Figure 2), it may start from n clusters (if you have n observations) and each cluster contains one observation, then finding the closest distance pair of clusters and combining them into a single cluster. In the end, all clusters will be combined into one cluster, called *Agglomerative method. The Agglomerative method* is used here to identify latitude

and altitude groups (Table 3).

Table 3 Geography of Ecotypes.

	Ecotype	Location	Altitude	Group(Al)	Latitude	Group(La)
1	Bay-0	Bayreuth, Germany	350	C	49.56	B
2	C24	Coimbra, Portugal	179	C	40.2	C
3	Col-0	Columbia University (U.S.)	49	D	43.0125	C
4	Cvi	Cape Verde Islands	43	D	16	D
5	Est	Estonia	15	D	59	A
6	Kin-0	Kinneville, MI	273	C	42.466	C
7	Ler	Landsberg, Germany	628	B	48.2	B
8	Nd-1	Niederzunzheim, Germany	250	C	50.778	B
9	Shakdara	Pamiro-Alay, Tadjikistan	3400	A	37.183	C
10	Van-0	UBC (Vancouver)	50	D	49.85	B

The process of *Agglomerative Method* as follows:

Given a set of *n* observations (ecotypes) to be grouped, and a *nxn* distance matrix (Euclidean distance measure used) illustrates each pair of two observation distance.

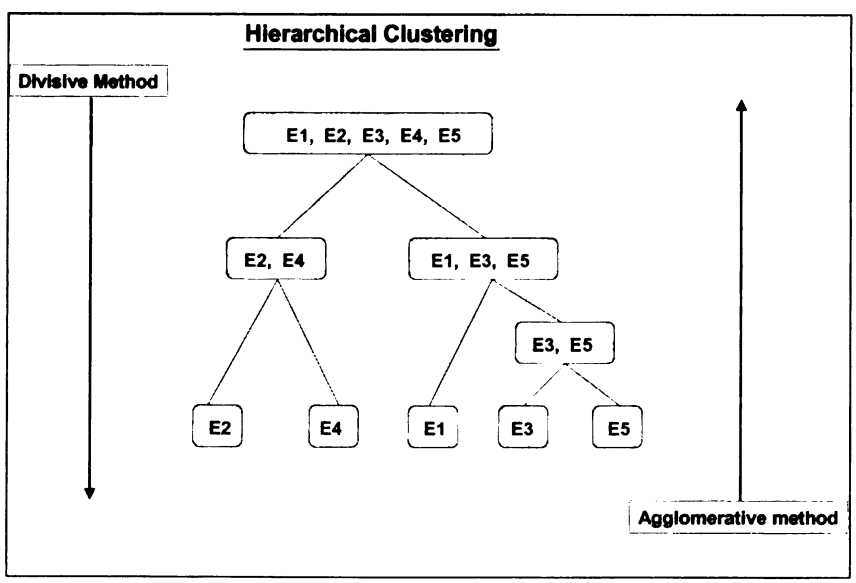
Step1. Start with *n* clusters, and each cluster contains a single observation.

Step2. Select the closest pair of clusters to merge into one new cluster.

Step3. Calculate the distance of the new cluster and other old single observation cluster.

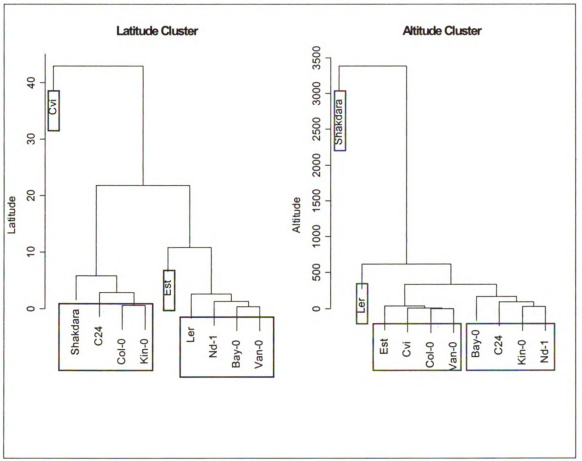
Step4. Repeat Step2 and Step3 until all observations merge into one cluster.

Figure 2 Hierarchical Clustering Process



Hierarchical cluster used to cluster 10 ecotypes into subgroups according to their altitude and latitude (Figure3). From Figure3, we can see that Cvi and Shakdara differ the most from the remaining ecotypes.

Figure 3 Ecotype Cluster



2.2. Limma – Linear Models for Microarray Data

Before Random Forest is applied to gene sets, we use *Limma*, Linear Models for Microarray Data (Smyth, G. K. 2004), to choose smaller subgroups of genes between ecotypes. The grouping will be discussed in the following paragraph. Differentially expressed genes will be used in Random Forest to classify ecotypes and to assign ranks to the genes.

Limma is used to identify genes whose expression pattern differs from others. *Limma* is a software package in Bioconductor in R environment (http://www.r-project.org) for the analysis of gene expression microarray data. Linear models are constructed for each gene to determinate weather they are differentially expressed in subgroups of ecotypes defined by latitude an altitude clusters. In the *topTable* function of *Limma*, *M-value, t-statistic, B-statistic and P.Value* of each gene can provide overall ranking of genes in order of differential expression. *M-value* is $\log_2$-fold change between two groups.

$$M = log_2(\frac{expression\ value\ of\ gene\ in\ group\ A}{expression\ value\ of\ gene\ in\ group\ B})$$

The *t-statistic* is a well-known hypothesis to test the mean of two groups. The B-statistic is the log odds that the gene is differentially expressed. For example, if the B-statistic is 3.5, the probability that the gene is differentially expressed is $\frac{e^{3.5}}{1+e^{3.5}} = 97\%$. A larger *B-statistic* indicates higher probability that the gene is differentially expressed. The *P.Value* is adjusted for multiple hypothesis testing using *Benjamini- Hochberg's* method (BH). B-statistics and *P.Value* provide the same ranking when no data is missing. Besides, differentially expressed genes are ranked in *topTable* by their *P.Values*. *Benjamini- Hochberg's* method controls the *false discovery rate* (FDR) when testing thousands of hypotheses, such as in microarray data. We identify genes differentially

expressed in subgroups from Hierarchical Cluster (Figure3) and assign the letters of *A*, *B*, *C*, *D* to those four groups (Table3).

2.3. Random Forest

The Random Forest algorithm by *Leo Breiman* (L. Breiman 2001) is a classification procedure consisting of a collection of tree-structured classifiers. Each tree is independent, identically distributed random vectors. Each tree gives a unit vote for the class of input vectors (arrays). Random Forest can analyze high dimensional data efficiently. Two processes of randomization occur in Random Forest: trees and nodes. Trees were built by bootstrap samples, and each node was split by randomly selected predictor variables (genes).

In the ecotype data, there are ten ecotypes and each of them has 3 arrays , so there are 30 arrays in the ecotype data. Moreover, each array has 22810 genes. In the Random Forest, the 30 arrays are "input vectors" (class observations) and 22810 genes are as "predictor variables". Randomly select N arrays from those 30 arrays with replacement for the training set (in-bag). The arrays which are not included in the training set are called out-of-bag (OOB). The training set data are used to grow the tree. The OOB data are used to estimate the classification error rate and get a variable importance measure.

Figure 4 Random Forest Construction

Each tree is grown as follows

Step 1. The training set consists of N observations (arrays) selected at random. Take N observations (arrays) at random and with replacement from the original data set called "in-bag". The observations not selected are called "out-of-bag". On average, there will be two third observations "in-bag", and one third "out-of-bag".

Step 2. The observations selected from the training set are used to construct a decision tree. The number of variables is M. A fixed number m_{try} (m_{try} <<M) of variables are chosen randomly from M variables, and the number of m_{try} is held constant during forest growing. These m_{try} variables are candidate for splitting the node. The best split on these randomly chosen m-variables is used to split the node which visualizes the tree and examine diagnostic statistics of each tree. For example, if we have M=5 variables, we can choose m_{try}=3 variables to split the node. There are C_3^5=10 candidates and each candidate has 3 variables (m_{try}). Randomly choose one of these 10 candidates and apply the best predictor variable (genes) of these 3 variables to split the node of each tree. Each tree is grown as large as possible and without pruning.

Step 3. Repeat *Step1* and *Step2* to construct 500 trees, ie. n_{tree}=500 (default number in R). Thus, the algorithm is called "Random Forest."

Step 4. Each tree give a classification for 10 ecotypes, we say each array "votes" for that class (ecotype). For example, if *AtGE_111A* was predicted for Bay0 at the terminal node, we say *AtGE_111A* "votes" Bay0, similarly to other arrays. As

the tree is built, each array will be assigned to a class (ecotype) in the terminal node (vote). For each of the N bootstrap samples, a tree is built. The majority vote for an array in this forest will be the predicted class (ecotype).

Notations

M : 22810 Genes.

N : 30 Arrays.

m_{try} *: The number of variables (genes) used to split each tree node.*

n_{tree} *: The number of trees (bootstraps) in the forest.*

In the original paper (Leo Breiman 2001) of Random Forests, it was shown that the error rate in Random Forest depends on two properties: the pairwise correlation between trees and the strength of each individual tree. The correlation is the extent to which arrays in a tree are similar from one to another. The strength is the overall average prediction quality. Higher correlation between trees will increase the error rate, and larger strength of each individual tree will reduce the error rate. Increasing the number of variables, m_{try}, will increase both of correlation between trees and strength of each individual tree. Decreasing m_{try} decreases both of them. Therefore, we can use the error rate to estimate optimal m_{try}. The parameter m_{try} is the only modifiable parameter which is sensitive in random forest. The predicted class (ecotype) of overall trees establishes the classification of Random Forest by choosing the most votes of the class in overall trees.

Features of Random Forest

- It runs efficiently on thousands of observations.
- It can handle large number of predictor variables (genes).
- It can rank predictor variables (genes) importance in the classification.

Parameters of *ntree* and *mtry*

In the Random Forest, the most important and sensitive parameters are the number of trees (n_{tree}) and the number of variables (m_{try}) which are selected at random from all variables. Each ecotype represents a class in Random Forest. We want to find the optimal n_{tree} and m_{try} to lower the OOB error rate, since the OOB error rate means that the ecotypes can be classified well or not. The optimal values for n_{tree} and m_{try} are not unique.

OOB error estimate

There are about one-third of observations (arrays) not included in the training set. Building trees based on the observations in the test set (OOB). If the class *j* has the most of the votes every time as observation *n* is in OOB data, class *j* will be as the predicted class. The proportion of the number of times that *j* is not equal to the true class *i* over all observations *N* is the OOB error rate estimate.

$$\textbf{OOB error rate :} \quad \frac{\sum_{n=1}^{N} I(C_{nj} \mid C_{ni})}{N}$$ (There are 500 boostrap samples here)

For observation n :

C_{nj} : Class j gets the most votes (as every time observation n is in OOB data)

C_{ni} : The true class for observation n is i

N : There are N observations

$$I(C_{nj} \mid C_{ni}) = \begin{cases} 0 & if \ j = i \\ 1 & if \ j \neq i \end{cases}$$

Example:

There are 10 classes in Ecotypes and each class has 3 arrays, so there are 30 observations in the data. The confusion matrix is computed as follows. For example, observations of *AtGE_111_A*, *AtGE_111_B*, and *AtGE_111_C* belong to class of *Bay0, but* in random forest procedure, class *Est* gets the most votes for *AtGE_111_A* which imply that $I(C_{AtGE\_111\_A\,j} \mid C_{AtGE\_111\_A\,i}) = 1$, and class *Bay0* gets the most votes for *AtGE_111_B* and *AtGE_111_C* which implies that $I(C_{AtGE\_111\_B\,j} \mid C_{AtGE\_111\_B\,i}) = 0$ and $I(C_{AtGE\_111\_C\,j} \mid C_{AtGE\_111\_C\,i}) = 0$ (j: Est, i: Bay0). In our example, there are 21 observations with $I(C_{nj} \mid C_{ni}) = 1$, so OOB error rate is $\frac{\sum_{n=1}^{N} P(C_{nj} \mid C_{ni})}{N} = \frac{21}{30} = 70\%$.

OOB estimate of error rate: 70%

Confusion matrix:

	Bay0	C24	Col0	Cvi	Est	Kin0	Ler	Ndl	Shakdara	Van0	class.error
Bay0	2	0	0	0	**1**	0	0	0	0	0	0.3333333
C24	0	1	0	0	0	0	0	0	**2**	0	0.6666667
Col0	0	0	1	0	0	0	**1**	0	0	**1**	0.6666667
Cvi	0	0	0	1	0	**1**	0	0	0	**1**	0.6666667
Est	**2**	0	0	0	0	0	0	**1**	0	0	1.0000000
Kin0	0	0	**1**	**1**	0	0	0	0	0	**1**	1.0000000
Ler	0	0	**1**	**1**	**1**	0	0	0	0	0	1.0000000
Ndl	0	0	0	0	0	**1**	**1**	1	0	0	0.6666667
Shakdara	0	0	0	0	0	0	0	0	3	0	0.0000000
Van0	0	0	**1**	**2**	0	0	0	0	0	0	1.0000000

Variable importance

Much interest in bioinformatics is given to Variable Importance measures. In this study, we rank the genes and thus reduce the number of variables. A variable importance measure is obtained as the trees are built based on the OOB data set. The most important

predictor variables (genes) are identified by calculating an important score for each predictor variable (gene). For a predictor variable (gene) X, the gene expression values of the gene X are permuted in each OOB data set to build the tree. The *raw importance scores* are calculated by subtracting the number of votes for each correct class with permutation from the number of votes for the correct class without permutation. The average of the *raw value* over all trees is the raw importance score. The raw importance score is normalized by dividing by standard error. There are fewer correct votes when predictor variables (genes) are permuted. Thus, a higher importance score for a gene identifies this gene with more discriminatory power.

$$Raw\text{-}Score(X) = \left(\sum_{tree\_i} (N^{without-permutation} - N^{permutation})_{tree\_i} \right) \Big/ ntree$$

$$Z\text{-}Score(X) = \frac{Raw - Score(X)}{Square[Variance(N^{without-permutation} - N^{permutation})]}$$

$N^{without-permutation}$: *the number of votes for correct class after permutation*

$N^{permutation}$: *the number of votes for correct class without permutation*

Chapter 3 Results of *Limma* and *Random Forest* to Ecotypes

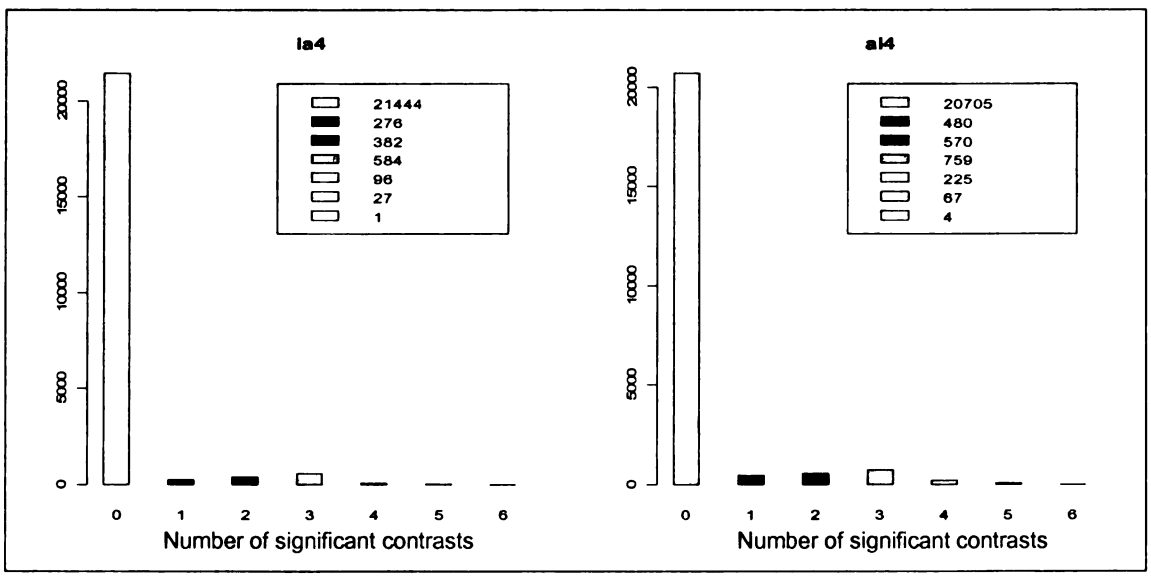
3.1. Gene Selection using Limma

There are thirty-four ecotypes input vectors in the original Ecotype data. Here, we just pick up first ten ecotypes that have been replicated. The locations of these ecotypes are located across every continent in the world.

Let's examine *latitude* clusters first, we divide those 22810 genes into seven sets (Figure 5). The seven sets are:

- 6: genes are significant in all six contrast combinations
- 5: genes are significant in any five of six contrast combinations
- 4: genes are significant in any four of six contrast combinations
- 3: genes are significant in any three of six contrast combinations
- 2: genes are significant in any two of six contrast combinations
- 1: genes are significant in any one of six contrast combinations
- 0: genes are not significant in any of the six contrast combinations.

Figure 5 Number of significant contrasts

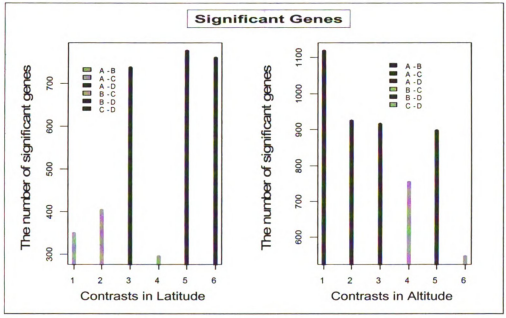


Similarly *altitude* clusters, genes are also divided into seven sets (Figure 5). As expected most genes are not statistically significant. Moreover, we are also interested in the number of significant genes per contrast (Table 4 and Figure 6)

Table 4 The number of significant genes for per contrast

Latitude		Altitude	
Contrast	**Number of Significant Genes**	**Contrast**	**Number of Significant Genes**
A-B	349	A-B	1118
A-C	403	A-C	924
A-D	736	A-D	916
B-C	295	B-C	754
B-D	775	B-D	897
C-D	759	C-D	547

Figure 6 The number of significant genes for per contrast



From Table4, we can see that the contrasts of *A-D, B-D, C-D* have the larger number of significant genes at significance level 0.05 in *Latitude* grouping, and the contrasts of *A-B, A-C, A-D* have the larger number of significant genes at significance level 0.05 in *Altitude* grouping. Therefore, Group *D* in *Latitude* and group *A* in *Altitude* are significant group within other groups. This corresponds to Cvi (group D in Latitude) and Shakdara

(group A in Altitude). Therefore, we will discuss these two ecotypes (*Cvi* and *Shakdara*) in more detail in the following chapter. 423 genes are significant for all three contrasts *A-D, B-D, C-D* among latitude, and 8 genes are significant for all three contrasts *A-B, A-C, A-D* among altitude (Figure5). Only one gene (247999_at) appears in both, in the 423-Latitude genes and the 8-Altitude genes. As expected, gene expression differ the most in the Shakdara and Cvi ecotypes compared to the others (Figure3).

Figure 7 Significant genes for Latitude and Altitude.

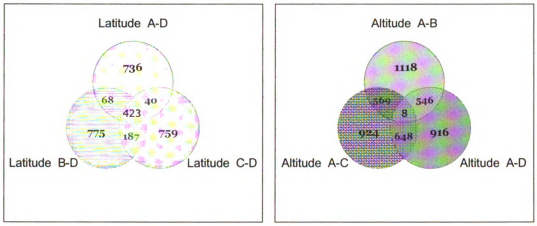
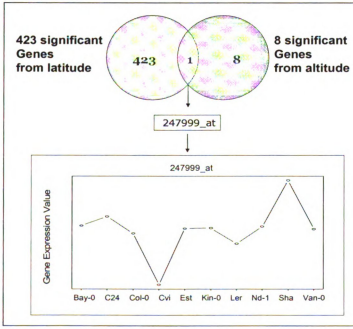


Figure 8 Gene expression of 247999_at



ID
247999_at (AT5G56150)
Annotation
ubiquitin-conjugating enzyme, putative, strong similarity to ubiquitin-conjugating enzyme UBC2 (Mesembryanthemum crystallinum) GI:5762457, UBC4 (Pisum sativum) GI:456568; contains Pfam profile PF00179: Ubiquitin-conjugating enzyme.

3.2. Ecotypes of Cvi and Shakdara

We are interested in how Cvi and Shakdara differ from the other 8 ecotypes. Thus, we examine the contrast between Cvi and the average of other 8 ecotypes, and the contrast between Shakdara and other 8 ecotypes.

- $Cvi - \frac{1}{8}(Bayo + C24 + Colo + Est + Kino + Ler + Nd1 + Vano)$
- $Sha - \frac{1}{8}(Bayo + C24 + Colo + Est + Kino + Ler + Nd1 + Vano)$

In each of these two contrasts, we perform multiple comparisons and select the top 500 differently expressed genes ranked by P.Values. Therefore, we have two sets of genes and each set has 500 genes.

We use Random Forest to reduce the number of genes and decide which of these highly significant genes mostly affect the classification performance of these 10 ecotypes.

3.3. Gene Selection from Cvi contrasts with other 8 ecotypes

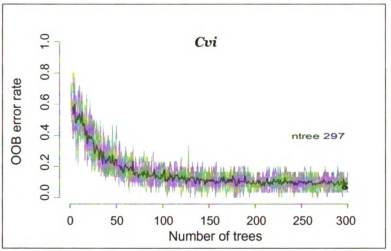
Genes were selected from *topTable* of *Limma* for the Contrast:

- $Cvi - \frac{1}{8}(Bayo + C24 + Colo + Est + Kino + Ler + Nd1 + Vano)$

500 top differently expressed genes were selected from *Limma* with this contrast, and called ***Cvi500***. Then we would like to use Random Forest to find the optimal number of *Cvi500* genes to improve classification. Before selecting top ranked genes from Random Forest, we need to find the optimal *ntree* and *mtry* first to reduce the OOB error rate. The procedure for finding the optimal *ntree* is as follows:

1. Run Random Forest with different number of trees but select *mtry* is the default (The default *mtry* is $\sqrt{the\ number\ of\ variables} \approx 151$).
2. Repeat 1. ten times and average the OOB error rate of these ten times for each of the number of trees.
3. To see which number of trees has the lowest average OOB error rate and this number is our optimal number of trees, ie. *ntree*. We found the optimal number of trees is 297 from *Cvi500*. (Figure 8)

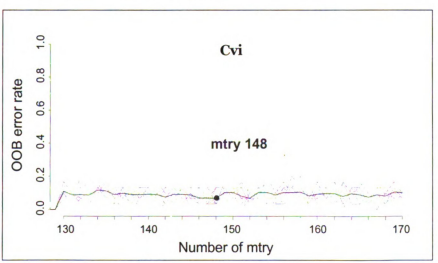
Figure 9 Optimal value of ntree for Cvi



After finding the optimal *ntree*, we would like to find optimal *mtry* as follows.

4. Run *randomforest* with ntree=297 and different number of *mtry* which is near $\sqrt{\textit{the number of variables}} \approx 151$. Here taking the range of *mtry* from 130 to 170.
5. Repeat 4. ten times and average the OOB error rate of these ten times for each of the number of *mtry*.
6. To see which number of *mtry* has the lowest average OOB error rate and this number is our optimal mtry. We found the optimal *mtry* is 148 from *Cvi500*. (Figure10)

Figure 10 Optimal value of mtry for Cvi



3.4. Gene Selection from Shakdara contrasts with other 8 ecotypes

Genes were selected from *topTable* of *Limma* for the Contrast:

- $Sha - \frac{1}{8}(Bayo + C24 + Colo + Est + Kino + Ler + Nd1 + Vano)$

500 top differently expressed genes was selected from *Limma* with this contrast, and called ***Sha500***. We follow the same procedure of finding the optimal *ntree* and *mtry*, and choose optimal *ntree* = 291 and there are 4 optimal numbers of *mtry* which can make Random Forest OOB error rate smallest, 133, 148, 161, 163. (Figure11) (Figure12)

Figure 11 Optimal value of ntree for Shakdara

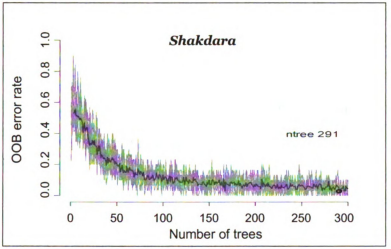
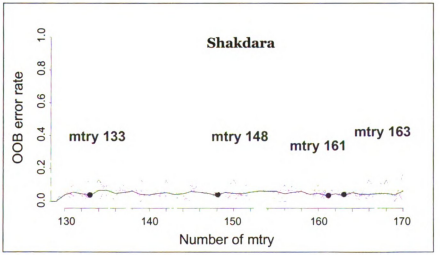


Figure 12 Optimal value of mtry for Shakdara



3.5. Gene Selection from Cvi500 and Sha500 by Random Forest

In order to reduce the number of *Cvi500* and *Sha500*, we select important genes from Random Forest, but the question is how many genes are needed for the best performance of classification. Beside *ntree* and *mtry*, the number of genes which has smallest OOB error rate is which we are interested in. From above procedure of finding optimal *mtry* and *ntree* (Figure 9, 10, 11, 12), the value of *ntree* greater than 200 can get stable smaller OOB error rate, but the value of *mtry* is not significant association with the OOB error rate. Thus, we select the number of most important genes from Random Forest with *ntree*=200, but keep *mtry* be default in Cvi500 and Sha500 respectively. To rank the genes the measure MeanDecreaseAccuracy was used to measure reliable importance.

In *Cvi500*, 43 genes is the smallest number for optimal classification. In *Sha500*, 84 genes is the smallest number for optimal classification. Then we compare those two sets of selected genes, there are 43 genes from the intersection of Cvi500 and Sha500, and there are 4 genes from the intersection of Cvi43 and Sha84. (Figure 15)

Figure 13 Optimal value of the number of genes for Cvi

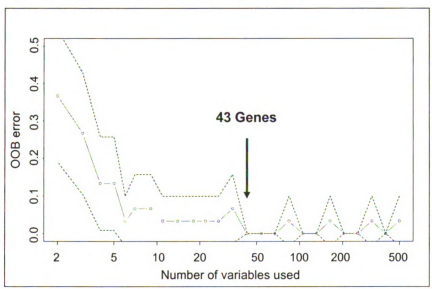


Figure 14 Optimal value of the number of genes for Shakdara

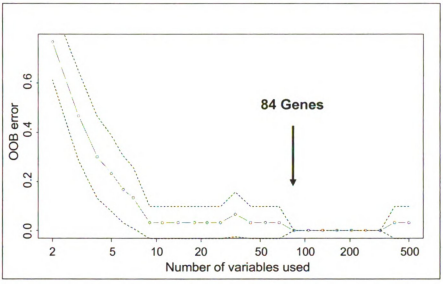
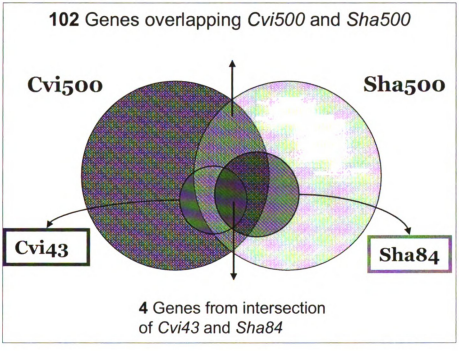


Figure 15 Overlapping genes from Cvi500 and Sha500



3.6. Compare the OOB error rate of Random Forest

Several sets of genes were selected with *Limma* and Random Forest. We have two sets of 500 genes selected from *topTable* of *Limma*; they are *Cvi500* and *Sha500*. Moreover, we have a set of 43 genes from *Cvi500*, and a set of 84 genes from *Sha500*. The following table will show the OOB error rate for *Cvi500* and *Sha500* and compare the status of using the optimal *ntree* and *mtry* with the status of without optimal *ntree* and *mtry*. Besides, Table5 also shows that the OOB error rate for the selected 84 genes and selected 50 genes without adjusting parameters

Table 5 Comparison of OOB error rate

Genes	Status	Number of Genes	OOB error rate
Cvi500	Without optimal value of *ntree* and *mtry*	500	16.67%
	With optimal values of *ntree* and *mtry* and the smallest number of genes	43	6.67%
Sha500	Without optimal value of *ntree* and *mtry*	500	10%
	With optimal values of *ntree* and *mtry* and the smallest number of genes	84	3.33%

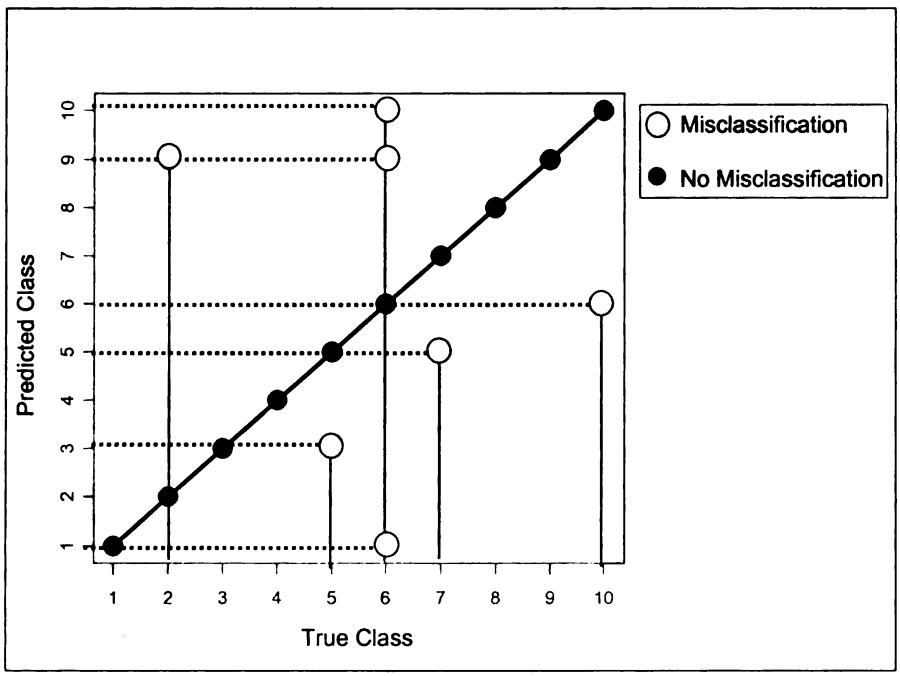
3.7. Misclassifications of Ecotypes

When running the Random Forest, there are some arrays which are misclassified. Each run gives us different misclassified ecotypes. Table6 shows misclassified arrays from all Random Forest runs. The most frequent misclassifications are Van0 and Kin0. The array ATGE_116_B.CEL (Kino) is often misclassified.

Table 6 Misclassification List

Array	Actual Ecotype	Predicted Ecotype
ATGE_112_A.CEL	C24	Shakdara
ATGE_115_D.CEL	Est	Colo
ATGE_116_A.CEL	Kino	Vano
ATGE_116_B.CEL	Kino	Vano, Shakdara, Bayo
ATGE_116_C.CEL	Kino	Vano
ATGE_117_D.CEL	Ler	Est
ATGE_120_A.CEL	Vano	Kino
ATGE_120_C.CEL	Vano	Kino

Figure 16 Misclassification figure



Chapter 4 Gene Ontology

4.1. Gene Ontology with *Classification Superviewer*

We have identified genes that may be important in adaptation. We selected two groups of genes, Cvi43 and Sha84 based on Random Forest. Cvi is close to the equator off the coast of Africa with higher temperature than other ecotypes, and Shakdara is a mountainous (around Himalayas) landlocked country in Central Asia and thus exposed to climate (eg. Temperature). The adaptation of these two ecotypes has likely been driven by these stress conditions. We would like to argue that these selected genes are important for stress resistance.

In order to validate the genes we selected from Random Forest, we classify the gene function on a group of genes based on the website: "The Bio-Array Resource *for Arabidopsis thaliana* Functional Genomics" http://bar.utoronto.ca/. The web-based tool of *Classification SuperViewer* creates an overview of gene functional classification of a group of AGI genes based on the MIPS database (Munich Information Center for Protein Sequences). Currently, there are 25450 genes for MIPS classifications in the MAtDB (MIPS Arabidopsis Thaliana Database). Here we do not focus on single genes. Instead, we want to find gene functions overrepresented in the selected sets of genes that can provide important information on stress response. Gene function classification is an approach for grouping genes based on functional similarity. However, Functional Classification Pie Chart often used in Bioinformatics provides the absolute numbers and percentage of gene function. Absolute numbers of genes on functional classification might be misleading in a different treatment and situation, but normalizing the group of genes can avoid this misdirection. In this way, the differences of gene function are more easily detected. *Classification SuperViewer* includes normalization, bootstrap sampling,

and provides a confidence estimate for the accuracy of results. The standard deviation may make results spurious and unreliable. Moreover, if the confidence intervals include one, the genes of this functional classification may be due to a small number of genes, and thus the class score is unreliable. We only consider a class score greater than one and confidence intervals not including one to check if these categories of functions are associated with stress response.

A class score for normalization was calculated based on the following equation: (*N* is gene number)

$$Score_{class} = \frac{N_{class(inputset)} / N_{classified(inputset)}}{N_{class(25K)} / N_{classified(25K)}}$$

(inputset : Cvi43 and Sha84)

One hundred Bootstrap samples were chosen from the input set. After sampling, classifying each set and generating them to get class score as above equation. Furthermore, the standard deviation of each class was shown along with the class score. If the class scores are greater than one and confidence intervals not including one, the gene ontology categories are overrepresented within a group of genes. In the following section is applied to gene groups Cvi43 and Sha84 in *Classification Superviewer* and discuss how their overrepresented gene functions affect the stress response. After that, we simplify the broad and wide spectrum of known protein functions based on *FunCat* annotation which includes 7 main gene categories (Table7).

Table 7 Main Function categories of FunCat

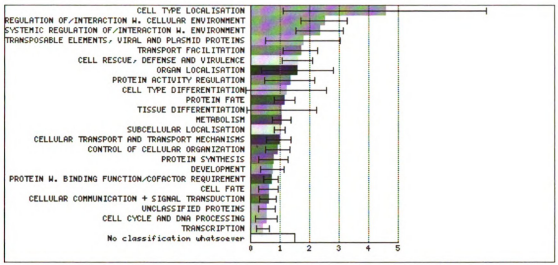
Main Function categories of FunCat	
Metabolism	
01	Metabolism
02	Energy
04	Storage protein
Information pathways	
10	Cell cycle and DNA processing
11	Transcription
12	Protein synthesis
14	Protein fate (folding, modification and destination)
16	Protein with binding function or cofactor requirement (structural or catalytic)
18	Protein activity regulation
Transport	
20	Cellular transport, transport facilitation and transport routes
Perception and response to stimuli	
30	Cellular communication/signal transduction mechanism
32	Cell rescue, defense and virulence
34	Interaction with the cellular environment
36	Interaction with the environment (systemic)
38	Transposable elements, viral and plasmid proteins
Developmental processes	
40	Cell fate
41	Development (systemic)
42	Biogenesis of cellular components
43	Cell type differentiation
45	Tissue differentiation
47	Organ differentiation
Localization	
70	Subcellular localization
73	Cell type localization
75	Tissue localization
77	Organ localization
78	Ubiquitous expression
Experimentally uncharacterized proteins	
98	Classification not yet clear-cut
99	Unclassified proteins

With the exception of categories 78, 98 and 99, all main categories are the origin of hierarchical, tree-like structures. To make the introduction of new main categories possible, the numbering of the categories is not strictly sequential.

The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, Nucleic Acids Research, 2004, Vol.32, No.18: 5539-5545.

4.2 Gene Ontology of Cvi43 and sha84

Figure 17 Cvi43 – Classification Superviewer



As we can see, there are five terms whose class scores greater than one and confidence intervals not including one. The number of genes, Cvi43, associated with terms (1)-(5) below is greater than expected on the basis of chance. In other words, terms (1)-(5) are overrepresented in the gene set of Cvi43.

(1) CELL TYPE LOCALISATION

(2) REGULATION OF/INTERACTION W. CELLULAR ENVIRONMENT

(3) SYSTEMIC REGULATION OF/INTERACTION W. ENVIRONMENT

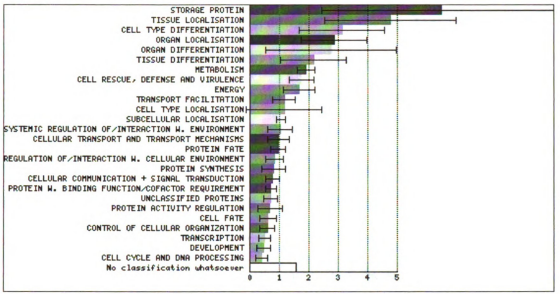
(4) TRANSPORT FACILITATION

(5) CELL RESCUE, DEFENSE AND VIRULENCE

Refer to Table7 , (2) (3) (5) are in category of *Perception and response to stimuli*. Plant perception indicates the change in the environment. The stimuli which plants perceive can respond to the environmental effects of chemicals, gravity, light, moisture, infections, temperature, oxygen, and carbon dioxide. Plants detect stimuli in different methods and a variety of reaction response, but generally plant perception occurs at the cellular level.

Thus, the selected genes are related to climatic conditions for Cvi.

Figure 18 Sha84 – Classification Superviewer



In Figure18, there are eight terms whose class scores greater than one and confidence intervals not including one. Thus, terms (1)-(8) below are overrepresented in the gene set of Sha84.

(1) STORAGE PROTEIN

(2) TISSUE LOCALISATION

(3) CELL TYPE DIFFERENTIATION

(4) ORGAN LOCALISATION

(5) TISSUE DIFFERENTIATION

(6) METABOLISM

(7) CELL RESCUE, DEFENSE AND VIRULENCE

(8) ENERGY

Terms of (1) (6) (8) covered all sub-functions of the metabolism. The definition for metabolism is: "Chemical process occurring within a living cell or organism, including anabolism and catabolism. Metabolism is a chemical process that typically transforms

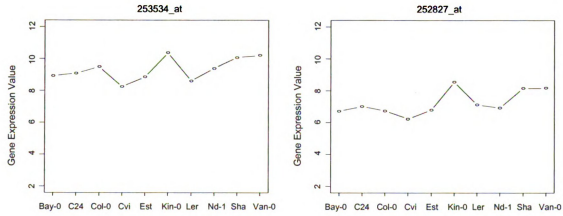
small molecules, but also includes macromolecular process and protein synthesis and degradation." Metabolism is associated with energy in some ways. Under stress, in metabolism some compounds are broken down to yield energy. Then this energy is directed at repairing the damage made by stress. Thus, metabolism would be an important factor under many different types of stressors. Under stress, plants may undergo a change of metabolism which would direct energy away from growth and reproduction and focus on cellular defense and maintenance. Instead, this helps plants survive in tough environments. Thus, the selected genes Sha84 may be important for adapting to the climatic conditions in high altitude.

Moreover, cytochrome P450 genes and glutathione-S-transferase genes may play an important role in oxidative stress resistance since oxidative stress is generated by all forms of stress in some ways. Several papers mention that Cytochrome P450 genes is important for plants. Oxidative detoxification of some herbicides in plant tissues is obtained by a Cytochrome P450-dependent monooxygenase system (Donaldson and Luster 1991, Hatzios 1991, and Sandermann 1992). Cytochrome P450s play important roles in biosynthesis of a variety of endogenous lipophilic compounds (Donaldson and Luster 1991 and Bolwell et al. 1994). Cytochrome P450 monooxygenases are a group of haem-containing proteins which catalyze various oxidative reactions (Schuler 1996 and Chapple 1998). In addition, some papers support that Glutathione-S-transferase plays an important role in plants. Glutathione S-transferases (GSTs) appear to be ubiquitous in plants and have defined roles in herbicide detoxification (Lamoureux and Rusness 1993). The fundamental function of GSTs is the detoxification of both endogenous and xenobiotic compounds (Marrs 1996). GSTs play a fundamental role in protection against endogenous or exogenous toxic chemicals (Sheehan et al. 2001). Furthermore,

cytochrome P450 genes and glutathione-S-transferase are phase I and phase II detoxification enzyme, respectively. Therefore, finding such genes associated with any form of stress may be biologically meaningful.

Besides, a gene (At5g10140) in Cvi43 is FLC (FLOWERING LOCUS C) gene which is a main determinant of flowering time. *Arabidopsis thaliana* locates in the Northern Hemisphere with long day time light hours which may affect flowering time. The transition to flowering is an important event in the plant life cycle and is adapted by several environmental factors of photoperiod, light quality, vernalization, and growth temperature, as well as biotic and abiotic stresses. Thus, FLC can respond to stresses and environmental effects. The following 5 genes were identified in both Cvi43 and Sha84 corresponding to these 3 specific genes and the graph also shows the expressions of these 5 genes.

Figure 19 Expression graph for 5 specific genes.



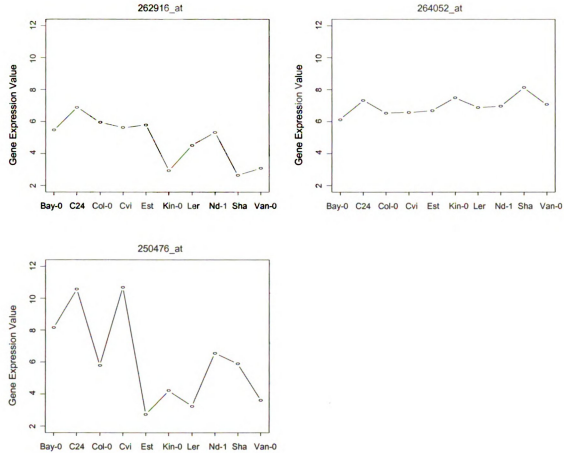


Table 8 FLC, Cytochrome P450 and Glutathione-S-transferase genes

AGI ID	Affy ID	Annotation
At4g31500	253534_at	CYP83B1_ATR4_RED1_RNT1_SUR2_CYP83B1 (CYTOCHROME P450 MONOOXYGENASE 83B1); oxygen binding
At4g39950	252827_at	CYP79B2_CYP79B2 (cytochrome P450, family 79, subfamily B, polypeptide 2); oxygen binding
At1g59700	262916_at	ATGSTU16_ATGSTU16 (Arabidopsis thaliana Glutathione S-transferase (class tau) 16); glutathione transferase
At2g22330	264052_at	CYP79B3_CYP79B3 (cytochrome P450, family 79, subfamily B, polypeptide 3); oxygen binding
At5g10140	250476_at	FLC_AGL25_FLF__FLC (FLOWERING LOCUS C)

Annotation from "TAIR, affy_ATH1_array_elements-2006-07-14.txt"

APPENDICES

APPENDIX A

Selected groups of Genes – *Cvi43 & Sha84*

Cvi43

Affy ID	AGI ID	Annotation
246173_s_at	At3g61520 At5g28370 At5g28460	pentatricopeptide (PPR) repeat-containing protein
246671_at	At5g30450	
246862_at	At5g25760	UBC21_PEX4__PEX4 (PEROXIN4); ubiquitin-protein ligase
247760_at	At5g59130	subtilase family protein
247791_at	At5g58710	ROC7__ROC7 (rotamase CyP 7); peptidyl-prolyl cis-trans isomerase
248460_at	At5g50915	basic helix-loop-helix (bHLH) family protein
249752_at	At5g24660	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G24655.1); similar to unknown protein [Brassica rapa subsp. pekinensis] (GB:AAQ92331.1)
249780_at	At5g24240	phosphatidylinositol 3- and 4-kinase family protein / ubiquitin family protein
250476_at	At5g10140	FLC_AGL25_FLF__FLC (FLOWERING LOCUS C)
251241_s_at	At3g62460 At3g62530	similar to PBS lyase HEAT-like repeat-containing protein [Arabidopsis thaliana] (TAIR:AT3G62530.1); similar to 80C09_3 [Brassica rapa subsp. pekinensis] (GB:AAZ41814.1); similar to Os07g0637200 [Oryza sativa (japonica cultivar-group)] (GB:NP_001060400.1); contains InterPro domain Protein of unknown function DUF537; (InterPro:IPR007491)
251962_at	At3g53420	PIP2A_PIP2__PIP2A (plasma membrane intrinsic protein 2;1)
252168_at	At3g50440	hydrolase
252231_at	At3g49720	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G65810.1); similar to Os01g0144000 [Oryza sativa (japonica cultivar-group)] (GB:NP_001042001.1); similar to conserved hypothetical protein [Medicago truncatula] (GB:ABE78370.1); contains domain S-adenosyl-L-methionine-dependent methyltransferases (SSF53335)
252459_s_at	At3g47220 At3g47290	phosphoinositide-specific phospholipase C family protein
252529_at	At3g46490	oxidoreductase, 2OG-Fe(II) oxygenase family protein
252723_at	At3g43520	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G26240.1); similar to Os04g0653100 [Oryza sativa (japonica cultivar-group)] (GB:NP_001054104.1); similar to transmembrane protein 14C [Argas monolakensis] (GB:ABI52790.1); similar to Os03g0568500 [Oryza sativa (japonica cultivar-group)] (GB:NP_001050510.1); contains InterPro domain Protein of unknown function UPF0136, Transmembrane; (InterPro:IPR005349)
253532_at	At4g31570	similar to myosin-related [Arabidopsis thaliana] (TAIR:AT1G24460.1); similar to hypothetical protein, conserved [Leishmania major] (GB:CAJ07774.1); contains InterPro domain Prefoldin; (InterPro:IPR009053); contains InterPro domain t-snare; (InterPro:IPR010989)

253534_at	At4g31500	CYP83B1_ATR4_RED1_RNT1_SUR2__CYP83B1 (CYTOCHROME P450 MONOOXYGENASE 83B1); oxygen binding
254351_at	At4g22300	carboxylic ester hydrolase
254361_at	At4g22212	Encodes a defensin-like (DEFL) family protein.
254928_at	At4g11410	short-chain dehydrogenase/reductase (SDR) family protein
255257_at	At4g05050	UBQ11__UBQ11 (UBIQUITIN 11); protein binding
255307_at	At4g04900	RIC10__RIC10 (ROP-INTERACTIVE CRIB MOTIF-CONTAINING PROTEIN 10)
255578_at	At4g01450	nodulin MtN21 family protein
256497_at	At1g31580	ECS1_CXC750__ECS1
256863_at	At3g24070	zinc knuckle (CCHC-type) family protein
257071_at	At3g28180	ATCSLC04_ATCSLC4_CSLC04__ATCSLC04 (Cellulose synthase-like C4); transferase, transferring glycosyl groups
257205_at	At3g16520	UDP-glucoronosyl/UDP-glucosyl transferase family protein
259067_at	At3g07550	F-box family protein (FBL12)
259591_at	At1g28150	similar to Os04g0528100 [Oryza sativa (japonica cultivar-group)] (GB:NP_001053373.1)
259733_at	At1g77480	nucellin protein, putative
260232_at	At1g74640	similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:BAD28539.1); contains domain no description (G3D.3.40.50.1820); contains domain alpha/beta-Hydrolases (SSF53474)
260244_at	At1g74320	choline kinase, putative
260252_at	At1g74240	mitochondrial substrate carrier family protein
263034_at	At1g24020	Bet v I allergen family protein
263777_at	At2g46450	ATCNGC12_CNGC12__ATCNGC12 (cyclic nucleotide gated channel 12); cyclic nucleotide binding / ion channel
265142_at	At1g51360	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G31670.1); similar to Hypothetical protein [Oryza sativa] (GB:AAK55783.1); contains InterPro domain Stress responsive alpha-beta barrel; (InterPro:IPR013097); contains InterPro domain Dimeric alpha-beta barrel; (InterPro:IPR011008)
265162_at	At1g30910	molybdenum cofactor sulfurase family protein
265486_at		
265699_at	At2g03550	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G48690.1); similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G48700.1); similar to Esterase/lipase/thioesterase [Medicago truncatula] (GB:ABE83378.1); contains InterPro domain Esterase/lipase/thioesterase; (InterPro:IPR000379); contains InterPro domain Alpha/beta hydrolase fold-3; (InterPro:IPR013094)
265768_at	At2g48020	sugar transporter, putative
266643_s_at	At2g29710 At2g29730	UDP-glucoronosyl/UDP-glucosyl transferase family protein
267093_at	At2g38170	CAX1_RCI4__CAX1 (CATION EXCHANGER 1); calcium:hydrogen antiporter

Sha84

Affy ID	AGI ID	Annotation
245038_at	At2g26560	PLA2A_PLA IIA_PLP2_PLA IIA__PLP2 (PHOSPHOLIPASE A 2A); nutrient reservoir
245400_at	At4g17040	ATP-dependent Clp protease proteolytic subunit, putative
245456_at	At4g16950	RPP5__RPP5 (RECOGNITION OF PERONOSPORA PARASITICA 5)
245977_at	At5g13110	G6PD2__G6PD2 (GLUCOSE-6-PHOSPHATE DEHYDROGENASE 2); glucose-6-phosphate 1-dehydrogenase
246642_s_at	At5g34920 At5g59620	
246708_at	At5g28150	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G04860.1); similar to Os07g0572300 [Oryza sativa (japonica cultivar-group)] (GB:NP_001060057.1); similar to Os03g0806700 [Oryza sativa (japonica cultivar-group)] (GB:NP_001051637.1); similar to Protein of unknown function DUF868, plant [Medicago truncatula] (GB:ABE92686.1); contains InterPro domain Protein of unknown function DUF868, plant; (InterPro:IPR008586)
247210_at	At5g65020	ANNAT2__ANNAT2 (ANNEXIN ARABIDOPSIS 2); calcium ion binding / calcium-dependent phospholipid binding
247313_at	At5g63980	SAL1_FRY1_HOS2__SAL1 (FIERY1); 3'(2'),5'-bisphosphate nucleotidase/ inositol or phosphatidylinositol phosphatase
247404_at	At5g62890	permease, putative
247814_at	At5g58310	hydrolase, alpha/beta fold family protein
247999_at	At5g56150	UBC30__UBC30; ubiquitin-protein ligase
248079_at	At5g55790	unknown protein
248200_at	At5g54160	ATOMT1_OMT1__ATOMT1 (O-METHYLTRANSFERASE 1)
248427_at	At5g51750	subtilase family protein
248796_at	At5g47180	vesicle-associated membrane family protein / VAMP family protein
248800_at	At5g47320	RPS19__RPS19 (40S ribosomal protein S19); RNA binding
248961_at	At5g45650	subtilase family protein
249258_at	At5g41650	lactoylglutathione lyase family protein / glyoxalase I family protein
249567_at	At5g38020	S-adenosyl-L-methionine:carboxyl methyltransferase family protein
249610_at	At5g37360	similar to Os02g0815400 [Oryza sativa (japonica cultivar-group)] (GB:NP_001048502.1)
249645_at	At5g36910	THI2.2.2__THI2.2 (THIONIN 2.2); toxin receptor binding
249733_at	At5g24400	EMB2024__EMB2024 (EMBRYO DEFECTIVE 2024); catalytic
250072_at	At5g17210	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G61065.1); similar to unknown protein [Saussurea involucrata] (GB:ABC68264.1); similar to Os06g0114700 [Oryza sativa (japonica cultivar-group)] (GB:NP_001056606.1); similar to Os05g0434800 [Oryza sativa (japonica cultivar-group)] (GB:NP_001055640.1); contains InterPro domain Protein of unknown function DUF1218; (InterPro:IPR009606)
250633_at	At5g07460	PMSR2__PMSR2 (PEPTIDEMETHIONINE SULFOXIDE REDUCTASE 2); protein-methionine-S-oxide reductase
250751_at	At5g05890	UDP-glucoronosyl/UDP-glucosyl transferase family protein
251032_at	At5g02030	HB-6_LSN_BLH9_BLR_PNY_RPL_VAN__LSN (LARSON, VAAMANA); DNA binding / transcription factor
251903_at	At3g54120	reticulon family protein (RTNLB12)

252318_at	At3g48730	GSA2__GSA2 (GLUTAMATE-1-SEMIALDEHYDE 2,1-AMINOMUTASE 2); glutamate-1-semialdehyde 2,1-aminomutase
252462_at	At3g47250	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G47200.2); similar to hypothetical protein LOC_Os12g29620 [Oryza sativa (japonica cultivar-group)] (GB:ABA98257.1); similar to Os11g0543300 [Oryza sativa (japonica cultivar-group)] (GB:NP_001068043.1); similar to Os04g0505400 [Oryza sativa (japonica cultivar-group)] (GB:NP_001053253.1); contains InterPro domain Protein of unknown function DUF247, plant; (InterPro:IPR004158)
252478_at	At3g46540	epsin N-terminal homology (ENTH) domain-containing protein / clathrin assembly protein-related
252529_at	At3g46490	oxidoreductase, 2OG-Fe(II) oxygenase family protein
252659_at	At3g44430	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G41660.1)
252678_s_at	At3g44300 At3g44310	NIT2__NIT2 (NITRILASE 2)
252724_at	At3g43540	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G47860.1); similar to Os09g0436900 [Oryza sativa (japonica cultivar-group)] (GB:NP_001063263.1); similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:BAD36432.1); contains InterPro domain Protein of unknown function DUF1350; (InterPro:IPR010765)
252827_at	At4g39950	CYP79B2__CYP79B2 (cytochrome P450, family 79, subfamily B, polypeptide 2); oxygen binding
252863_at	At4g39800	MI-1-P SYNTHASE__MI-1-P SYNTHASE (Myo-inositol-1-phosphate synthase); inositol-3-phosphate synthase
253422_at	At4g32240	unknown protein
253666_at	At4g30270	MERI5B_BRU1_MERI-5__MERI5B (MERISTEM-5); hydrolase, acting on glycosyl bonds
254248_at	At4g23270	protein kinase family protein
254508_at	At4g20170	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G44670.1); similar to Os06g0328800 [Oryza sativa (japonica cultivar-group)] (GB:NP_001057533.1); similar to Os02g0712500 [Oryza sativa (japonica cultivar-group)] (GB:NP_001047907.1); similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:BAD72474.1); contains InterPro domain Protein of unknown function DUF23; (InterPro:IPR008166)
254553_at	At4g19530	disease resistance protein (TIR-NBS-LRR class), putative
255437_at	At4g03060	AOP2__AOP2 (ALKENYL HYDROXALKYL PRODUCING 2); oxidoreductase, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors
255859_at	At5g34930	arogenate dehydrogenase
256021_at	At1g58270	ZW9__ZW9
256096_at	At1g13650	similar to 18S pre-ribosomal assembly protein gar2-related [Arabidopsis thaliana] (TAIR:AT2G03810.3); similar to hypothetical protein [Trypanosoma cruzi strain CL Brener] (GB:XP_813437.1)
256221_at	At1g56300	DNAJ heat shock N-terminal domain-containing protein
256454_at	At1g75280	isoflavone reductase, putative
256458_at	At1g75220	integral membrane protein, putative
256489_at	At1g31550	carboxylic ester hydrolase/ lipase
256940_at	At3g30720	unknown protein

257205_at	At3g16520	UDP-glucoronosyl/UDP-glucosyl transferase family protein
257228_at	At3g27890	NQR__NQR (NADPH:QUINONE OXIDOREDUCTASE); FMN reductase
257580_at	At3g06210	binding
258124_at	At3g18215	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G24600.1); similar to hypothetical protein [Oryza sativa (japonica cultivar-group)] (GB:BAC55679.1); similar to Os02g0292800 [Oryza sativa (japonica cultivar-group)] (GB:NP_001046597.1); similar to Os08g0153900 [Oryza sativa (japonica cultivar-group)] (GB:NP_001061011.1); contains InterPro domain Protein of unknown function DUF599; (InterPro:IPR006747)
258322_at	At3g22740	HMT3__HMT3 (Homocysteine S-methyltransferase 3); homocysteine S-methyltransferase
259770_s_at	At1g07780 At1g29410 At5g05590	PAI1_TRP6__PAI1 (PHOSPHORIBOSYLANTHRANILATE ISOMERASE 1); phosphoribosylanthranilate isomerase
260546_at	At2g43520	ATTI2__ATTI2 (ARABIDOPSIS THALIANA TRYPSIN INHIBITOR PROTEIN 2); trypsin inhibitor
260567_at	At2g43820	GT__GT/UGT74F2 (UDP-GLUCOSYLTRANSFERASE 74F2); UDP-glucosyltransferase/ UDP-glycosyltransferase/ transferase, transferring glycosyl groups / transferase, transferring hexosyl groups
260685_at	At1g17650	phosphogluconate dehydrogenase (decarboxylating)
260872_at	At1g21350	electron carrier/ oxidoreductase
260981_at	At1g53460	similar to zinc finger (Ran-binding) family protein [Arabidopsis thaliana] (TAIR:AT1G55040.1); similar to Zn-finger in Ran binding protein and others, putative [Oryza sativa (japonica cultivar-group)] (GB:AAX95671.1); similar to Os03g0712200 [Oryza sativa (japonica cultivar-group)] (GB:NP_001051062.1); similar to Os01g0203300 [Oryza sativa (japonica cultivar-group)] (GB:NP_001042331.1)
261105_at	At1g63000	NRS/ER__NRS/ER (NUCLEOTIDE-RHAMNOSE SYNTHASE/EPIMERASE-REDUCTASE)
261326_s_at	At1g44180 At1g44820	aminoacylase, putative / N-acyl-L-amino-acid amidohydrolase, putative
261727_at	At1g76090	SMT3__SMT3 (S-adenosyl-methionine-sterol-C-methyltransferase 3); S-adenosylmethionine-dependent methyltransferase
261924_at	At1g22550	proton-dependent oligopeptide transport (POT) family protein
262134_at	At1g77990	AST56_SULTR2;2__AST56 (sulphate transporter 2;2); sulfate transporter
262458_at	At1g11280	carbohydrate binding / kinase
262875_at	At1g64970	G-TMT_TMT1_VTE4__G-TMT (GAMMA-TOCOPHEROL METHYLTRANSFERASE)
262916_at	At1g59700	ATGSTU16__ATGSTU16 (Arabidopsis thaliana Glutathione S-transferase (class tau) 16); glutathione transferase
263553_at	At2g16430	PAP10__PAP10; acid phosphatase/ protein serine/threonine phosphatase
263714_at	At2g20610	SUR1_ALF1_HLS3_RTY_SUR1__SUR1 (SUPERROOT 1); transaminase
264052_at	At2g22330	CYP79B3__CYP79B3 (cytochrome P450, family 79, subfamily B, polypeptide 3); oxygen binding
264513_at	At1g09420	G6PD4__G6PD4 (GLUCOSE-6-PHOSPHATE DEHYDROGENASE 4); glucose-6-phosphate 1-dehydrogenase
264790_at	At2g17820	ATHK1_AHK1_ATHK1__ATHK1 (HISTIDINE KINASE 1)
264954_at	At1g77060	mutase family protein

265032_at	At1g61580	ARP2_RPL3B__ARP2/RPL3B (ARABIDOPSIS RIBOSOMAL PROTEIN 2); structural constituent of ribosome
265058_s_at	At1g52030 At1g52040	MBP2_F-ATMBP_MBP1.2__MBP2 (MYROSINASE-BINDING PROTEIN 2)
265354_at	At2g16700	ADF5__ADF5 (ACTIN DEPOLYMERIZING FACTOR 5); actin binding
265486_at	265486_at	
265611_at	At2g25510	unknown protein
265905_at	At2g25640	similar to transcription elongation factor-related [Arabidopsis thaliana] (TAIR:AT5G25520.2); similar to PHD finger protein-like [Oryza sativa (japonica cultivar-group)] (GB:BAD24999.1); similar to Os02g0208600 [Oryza sativa (japonica cultivar-group)] (GB:NP_001046260.1); contains InterPro domain Transcription elongation factor S-II, central region; (InterPro:IPR003618); contains InterPro domain SPOC; (InterPro:IPR012921)
266472_at		
266643_s_at	At2g29710 At2g29730	UDP-glucoronosyl/UDP-glucosyl transferase family protein
267078_at	At2g40960	nucleic acid binding

APPENDIX B

R CODE

```
#######################################################################

# ------------ #
# Pakages Used #
# ------------ #

library(limma)
library(randomForest)
library(varSelRF)
library(maps)


# ------------ #
# Reading Data #
# ------------ #

Ecodata = read.table("AtGE_ecotypes.txt", header =T, sep="\t")
Geo = read.table("Geo.txt", header =T, sep="\t")
x = read.table("EcotypesGeo.txt",sep="\t")

#######################################################################

# ------ #

# Map #

# ------ #


par(mar=rep(0, 4))

par(mfrow=c(2,1))

map("world",col="grey")

text(Geo$Longitude,Geo$Latitude,Geo$Ecotype,col="black",cex=0.8)

points(Geo$Longitude,Geo$Latitude,col= rainbow(16:20)[1:10],cex=0.7,lwd=3)

legend(120,85, Geo$Ecotype , fill = rainbow(16:20)[1:10] , cex=0.8, bty="n")

points(Geo$Longitude[7],Geo$Latitude[7],col="DarkGoldenRod ",cex=3,lwd=2)

arrows(10.5, 30, 10.5, -50, lwd=2,angle = 15,col="DarkGoldenRod  ")

text(5.5,-70, "Germany ", adj=0, cex=1.5, col="DarkGoldenRod ")


map("world", "Germany",col="DarkGoldenRod ")

text(Geo$Longitude,Geo$Latitude,Geo$Ecotype,cex=0.8,col="red")

points(Geo$Longitude,Geo$Latitude,col= rainbow(16:20)[1:10],cex=0.7,lwd=3)
```

```
# ------------------------ #
# Cluster for La4 and Al4 #
# ------------------------ #

la = x[-1,4]
la=la[1:10]
names(la)=x[-1,1][1:10]
dist(la)

hc.la <- hclust(dist(la))
plot(hc.la)

la.km <- kmeans(dist(la),4)$cluster
la.km # Cluster

al = x[-1,7]
al=al[1:10]
names(al)=x[-1,1][1:10]
dist(al)

hc.al <- hclust(dist(al))
plot(hc.al)

al.km <- kmeans(dist(al),4)$cluster
al.km # Cluster


# ---------------------- #
# Gene Expression Plot #
# ---------------------- #

genelist = la4    # changable variable
name= "La4"

genedata<-Ecodata[genelist,2:31]
gene.x1<-apply(genedata[,1:3],1,mean)
```

```
gene.x2<-apply(genedata[,4:6],1,mean)
gene.x3<-apply(genedata[,7:9],1,mean)
gene.x4<-apply(genedata[,10:12],1,mean)
gene.x5<-apply(genedata[,13:15],1,mean)
gene.x6<-apply(genedata[,16:18],1,mean)
gene.x7<-apply(genedata[,19:21],1,mean)
gene.x8<-apply(genedata[,22:24],1,mean)
gene.x9<-apply(genedata[,25:27],1,mean)
gene.x10<-apply(genedata[,28:30],1,mean)

genegexp<-data.frame(gene.x1, gene.x2, gene.x3, gene.x4, gene.x5,
gene.x6, gene.x7, gene.x8, gene.x9, gene.x10)

for (i in 1:length(genelist)){
GeneExpression.gene=t(rbind(genegexp[i,]))
matplot(GeneExpression.gene,axes=F,frame=T,type='b',pch=1)
row.names(GeneExpression.gene)<-c("Bay0", "C24", "Col0", "Cvi", "Est", "Kin0", "Ler", "Nd1", "Sha",
"Van0")
axis(1, 1:10, row.names(GeneExpression.gene))
par(new=T)
}
title(xlab="Ecotypes",main=paste(name))


# ---------------------------------------------------------- #
# Gene Expression Plot - Each picture represents one gene #
# ---------------------------------------------------------- #

genelist = la4       # changeable variable
N = 20               # the number of genes
genedata<-Ecodata[genelist,2:31]
gene.x1<-apply(genedata[,1:3],1,mean)
gene.x2<-apply(genedata[,4:6],1,mean)
gene.x3<-apply(genedata[,7:9],1,mean)
gene.x4<-apply(genedata[,10:12],1,mean)
gene.x5<-apply(genedata[,13:15],1,mean)
```

```
gene.x6<-apply(genedata[,16:18],1,mean)
gene.x7<-apply(genedata[,19:21],1,mean)
gene.x8<-apply(genedata[,22:24],1,mean)
gene.x9<-apply(genedata[,25:27],1,mean)
gene.x10<-apply(genedata[,28:30],1,mean)
genegexp<-data.frame(gene.x1, gene.x2, gene.x3, gene.x4, gene.x5,
gene.x6, gene.x7, gene.x8, gene.x9, gene.x10)
for (i in 1:N){
GeneExpression.gene=t(rbind(genegexp[i,]))
matplot(GeneExpression.gene,axes=F,frame=T,type='b',pch=1)
row.names(GeneExpression.gene)<-c("Bay0", "C24", "Col0", "Cvi",
"Est", "Kin0", "Ler", "Nd1", "Sha", "Van0")
axis(1, 1:10, row.names(GeneExpression.gene))
title(main=paste("Gene",i))
}


#----------------------#
#     Latitude - La4    #
#----------------------#

ecorep = c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9,10,10,10)
design = model.matrix(~-1+factor(ecorep))

Ala4 = design[,5]
Bla4 = design[,1]+design[,7]+design[,8]+design[,10]
Cla4 = design[,2]+design[,3]+design[,6] +design[,9]
Dla4 = design[,4]

designla4 = data.frame(Ala4, Bla4, Cla4, Dla4)
contrast.matrixla4 = makeContrasts(Ala4 – Bla4, Ala4 – Cla4, Ala4 – Dla4, Bla4 – Cla4, Bla4 – Dla4,
Cla4 – Dla4,levels=designla4)

eco.fitla4 = lmFit(Ecodata[,2:31],designla4)
eco.fit2la4 = contrasts.fit(eco.fitla4, contrast.matrixla4)
eco.ebla4 = eBayes(eco.fit2la4)
```

#= decideTests =#

clasla4 = decideTests(eco.ebla4, method= "nestedF", adjust.method= "fdr", p=0.05)

rownames(clasla4) = Ecodata[,1]

kla4 = rowSums(abs(clasla4))

select the genes which are significant at least in one contrast

c1.la4 = clasla4[,1]

c2.la4 = clasla4[,2]

c3.la4 = clasla4[,3]

c4.la4 = clasla4[,4]

c5.la4 = clasla4[,5]

c6.la4 = clasla4[,6]

la4.c1 = which(c1.la4 == 1 | c1.la4 ==-1)

la4.c2 = which(c2.la4 == 1 | c2.la4 ==-1)

la4.c3 = which(c3.la4 == 1 | c3.la4 ==-1)

la4.c4 = which(c4.la4 == 1 | c4.la4 ==-1)

la4.c5 = which(c5.la4 == 1 | c5.la4 ==-1)

la4.c6 = which(c6.la4 == 1 | c6.la4 ==-1)

la4.all = unique(c(la4.c1,la4.c2,la4.c3,la4.c4,la4.c5,la4.c6))

#= Look decideTests in different way =#

la4k0 = length(which(kla4==0))

la4k1 = length(which(kla4==1))

la4k2 = length(which(kla4==2))

la4k3 = length(which(kla4==3))

la4k4 = length(which(kla4==4))

la4k5 = length(which(kla4==5))

la4k6 = length(which(kla4==6))

la4k = c(la4k0, la4k1, la4k2, la4k3, la4k4, la4k5, la4k6)

names(la4k)=c(0:6)

la4bar = barplot(la4k,space=1.5,col= c("yellow","red","blue","lightblue", "mistyrose", "lightcyan", "lavender"),legend=la4k, xlab="number of significant contrasts", main="la4")

la4row = which(kla4>=4)

#----------#
Latitude - Al4
#----------#

ecorep = c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9,10,10,10)
design = model.matrix(~-1+factor(ecorep))

Aal4 = design[,9]
Bal4 = design[,7]
Cal4 = design[,1]+design[,2]+design[,6]+design[,8]
Dal4 = design[,3]+design[,4]+design[,5]+design[,10]
designal4 = data.frame(Aal4, Bal4, Cal4, Dal4)
contrast.matrixal4 = makeContrasts(Aal4-Bal4,Aal4-Cal4,Aal4-Dal4,Bal4-Cal4, Bal4-Dal4,Cal4-Dal4,levels=designal4)

eco.fital4 = lmFit(Ecodata[,2:31],designal4)
eco.fit2al4 = contrasts.fit(eco.fital4,contrast.matrixal4)
eco.ebal4 = eBayes(eco.fit2al4)

#= **decideTests** =#

clasal4 = decideTests(eco.ebal4, method= "nestedF", adjust.method= "fdr", p=0.05)
kal4 = rowSums(abs(clasal4))

select the genes which are significant at least in one contrast

c1.al4 = clasal4[,1]
c2.al4 = clasal4[,2]
c3.al4 = clasal4[,3]

c4.al4 = clasal4[,4]

c5.al4 = clasal4[,5]

c6.al4 = clasal4[,6]

al4.c1 = which(c1.al4 == 1 | c1.al4 ==-1)

al4.c2 = which(c2.al4 == 1 | c2.al4 ==-1)

al4.c3 = which(c3.al4 == 1 | c3.al4 ==-1)

al4.c4 = which(c4.al4 == 1 | c4.al4 ==-1)

al4.c5 = which(c5.al4 == 1 | c5.al4 ==-1)

al4.c6 = which(c6.al4 == 1 | c6.al4 ==-1)

al4.all = unique(c(al4.c1,al4.c2,al4.c3,al4.c4,al4.c5,al4.c6))

#= Look decideTests in different way =#

al4k0 = length(which(kal4==0))

al4k1 = length(which(kal4==1))

al4k2 = length(which(kal4==2))

al4k3 = length(which(kal4==3))

al4k4 = length(which(kal4==4))

al4k5 = length(which(kal4==5))

al4k6 = length(which(kal4==6))

al4k = c(al4k0, al4k1, al4k2, al4k3, al4k4, al4k5, al4k6)

names(al4k)=c(0:6)

al4bar = barplot(al4k,space=1.5,col= c("yellow","red","blue","lightblue", "mistyrose", "lightcyan", "lavender"),legend=al4k, xlab="number of significant contrasts", main="al4")

al4row = which(kal4>=5)

#----------#

 # Cvi vs. the other 8 Ecotypes (without Sha) #

#----------#

```
ecorep = c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9,10,10,10)
design = model.matrix(~-1+factor(ecorep))

designcvi = design
colnames(designcvi)<-c("Bay0", "C24", "Col0", "Cvi", "Est", "Kin0", "Ler", "Nd1", "Shakdara", "Van0")

contrast.matrixcvi<-makeContrasts(Cvi -Bay0/8 - C24/8 - Col0/8 – Est/8 - Kin0/8 – Ler/8 - Nd1/8 -
Van0/8 ,levels=designcvi)

eco.fitcvi = lmFit(Ecodata[,2:31],designcvi)
eco.fit2cvi = contrasts.fit(eco.fitcvi, contrast.matrixcvi)
eco.ebcvi = eBayes(eco.fit2cvi)

clascvi = decideTests(eco.ebcvi, method= "nestedF", adjust.method= "fdr", p=0.05)
kcvi = rowSums(abs(clascvi))


#= Toptable =#  selecting the first 500 genes from toptable

num=500
cvi = topTable(eco.ebcvi, genelist= eco.ebcvi $genes, coef=1, n=num, adjust="fdr")
d.cvi = read.csv("cvinumber.csv")
n.cvi = d.cvi[,1]


#-----------------------------------------#
# Sha  vs. the other 8 Ecotypes (without Cvi)  #
#-----------------------------------------#

designsha = design
colnames(designsha)<-c("Bay0", "C24", "Col0", "Cvi", "Est", "Kin0", "Ler", "Nd1", "Shakdara", "Van0")

contrast.matrixsha<-makeContrasts(Shakdara -Bay0/8 - C24/8 - Col0/8 – Est/8 - Kin0/8 – Ler/8 - Nd1/8 -
Van0/8,levels=designsha)

eco.fitsha = lmFit(Ecodata[,2:31],designsha)
```

eco.fit2sha = contrasts.fit(eco.fitsha, contrast.matrixsha)

eco.ebsha = eBayes(eco.fit2sha)

#= **Toptable** =# selecting the first 500 genes from toptable

sha = topTable(eco.ebsha, coef=1, **n=num**, adjust="fdr")

d.sha = read.csv("shanumber.csv")

n.sha = d.sha[,1]

#----------#

 # Highly Variation - geneselect #

#----------#

vars=apply(AtGE, 1, var)

sortvars=sort(vars,decreasing = TRUE)

geneselect=sortvars[1:number]

gs = names(geneselect)

gs = as.numeric(gs)

Ecodata[gs,1]

#----------#

 # Randomly Selection - ran #

#----------#

x=runif(number, min=1, max=22810)

ran=as.integer(x)

#----------#

 # RandomForest #

#----------#

rfgenes = n.cvi # changeable variable

```
rfname = "n.cvi"

library(randomForest)

eco1=t(Ecodata[rfgenes,2:31]) ## select "number" genes
econames=rep(c("Bay0", "C24", "Col0", "Cvi", "Est", "Kin0", "Ler", "Nd1", "Shakdara", "Van0"),each=3)

colnames(eco1)=Ecodata[rfgenes,1]
ecotype1=data.frame(eco1,econames) ##  Data which we want ##
ecotype.rf = randomForest(econames ~ ., data=ecotype1, ntree=100,
                  keep.forest=TRUE, importance=TRUE)
ecotype.rf

imp = importance(ecotype.rf)
plot(sort(imp[,11]),type="h",ylab="Importance Score", main = rfname)
# see Accuraacy



#-------------------------#
# ntree vs. OOB error rate #
#-------------------------#

ntree=300
nrf=10       # number of boostrap
m = matrix(rep(0,ntree*nrf),nrow=ntree)

for (j in 1:nrf){
for(i in 1:ntree){
ecotype.rf = randomForest(econames ~ ., data=ecotype1, ntree=i, mtry=sqrt(22810),
                  keep.forest=TRUE, importance=TRUE)
m[i,j]=ecotype.rf$err.rate[i,1]
matplot(m,type="l",col="grey",lty=1,
     xlab="number of trees",ylab="OOB error rate",ylim=c(0,1),frame.plot=F)
axis(1, seq(0,ntree,by=50),col = "#EE9A00", col.axis="blue", lwd = 2)
axis(2, seq(0,1,by=0.2),col = "#EE9A00", col.axis="blue", lwd = 2)
}
```

```
par(new=T)
}
mmean = apply(m, 1, mean)
mean = ifelse(mmean =="NaN", 1, mmean)

op = par(new=T)
par(op)
plot(mmean,type="l",cex=1,col="red",lwd=2,
    xlab="number of trees",ylab="OOB error rate",ylim=c(0,1),frame.plot=F,axes=F)
axis(1, seq(0,ntree,by=50),col = "#EE9A00", col.axis="blue", lwd = 2)
axis(2, seq(0,1,by=0.2),col = "#EE9A00", col.axis="blue", lwd = 2)

par(op)
mini = min(mean)
a = which(mean==mini)     # a is the number of trees which we want
text(a,0.4,paste("ntree",a),adj=1,cex=1.2,col="dark green")
points(a,mean[a],col="dark green",cex=1.2,lwd=3)
```

#= After finding "optimal ntree" =#

```
ecotype.rf = randomForest(econames ~ ., data=ecotype1, ntree=a,
                  keep.forest=TRUE, importance=TRUE)
ecotype.rf

econames=as.factor(econames)
e = ecotype1[,-501]
rf.eco <- varSelRF(e, econames, ntree = 210, mtry=4)
rf.eco

plot(rf.eco)
```

#----------#

 # mtry vs. OOB error rate #

#----------#

```
rf.mtry=170
nrf=10      # number of boostrap
numtree=a

m = matrix(rep(0,rf.mtry*nrf),nrow=rf.mtry)

for (j in 1:nrf){
for(i in 130:rf.mtry){
ecotype.rf = randomForest(econames ~ ., data=ecotype1,ntree=numtree, mtry=i,
                  keep.forest=TRUE, importance=TRUE)
m[i,j]=ecotype.rf$err.rate[numtree,1]

matplot(m,type="l",col="grey",lty=1,
     xlab="number of mtry",ylab="OOB error rate",ylim=c(0,1),xlim=c(130,170),frame.plot=F)
axis(1, seq(130,170,by=2),col = "#EE9A00", col.axis="blue", lwd = 2)
axis(2, seq(0,1,by=0.2),col = "#EE9A00", col.axis="blue", lwd = 2)
}
par(new=T)
}

mmean = apply(m, 1, mean)
#mean = ifelse(mmean =="NaN", 1, mmean)

op = par(new=T)
par(op)
plot(mmean,type="l",cex=1,col="red",lwd=2,
   xlab="number of mtry",ylab="OOB error rate",ylim=c(0,1),xlim=c(130,170),frame.plot=F,axes=F)
axis(1, seq(130,170,by=2),col = "#EE9A00", col.axis="blue", lwd = 2)
axis(2, seq(0,1,by=0.2),col = "#EE9A00", col.axis="blue", lwd = 2)

par(op)
mini = min(mmean[131:170])
b = which(mmean==mini)    # a is the number of trees which we want
text(b,0.4,paste("mtry",b),adj=0,cex=1.2,col="dark green")
points(b,mean[b],col="dark green",cex=1.2,lwd=3)
```

```
#= After finding "optimal ntree" & "optimal mtry" =#

ecotype.rf = randomForest(econames ~ ., data=ecotype1, ntree=a,mtry=b,
                    keep.forest=TRUE, importance=TRUE)
ecotype.rf



#------------------------------------------------ #
# Select the number of Genes from RandomForest  #
#------------------------------------------------#

rfgenes=n.cvi
eco1=t(Ecodata[rfgenes,2:31])
econames = rep(c("Bay0", "C24", "Col0", "Cvi", "Est", "Kin0", "Ler", "Nd1",
        "Shakdara", "Van0"),each=3)

colnames(eco1)= Ecodata[rfgenes,1]
ecotype1 = data.frame(eco1,econames)

econames=as.factor(econames)
e = ecotype1[,-501]
rf.eco <- varSelRF(e, econames, ntree=a , mtry=b)
rf.eco

plot(rf.eco)
```

Geo.txt

Ecotype	Latitude	Longitude
Bay-0	49.56	11.34
C24	40.2	8.25
Col-0	43.01251667	-70.05
Cvi	16.00208056	-24.05
Est	59	25.04
Kin-0	42.46638889	-84.46
Ler	48.2	10.52
Nd-1	50.77777778	8.03
Shakdara	37.18333333	73.166
Van-0	49.85049722	-123.11

EcotypesGeo.txt

Ecotype	Array	Location	Latitude	Longitude	Altitude
Bay-0	AtGE_111_A, B, C	Bayreuth, Germany	49.56	11.34	350
C24	AtGE_112_A, C, D	Coimbra, Portugal	40.20	8.25	179
Col-0	AtGE_113_A, C, D	Columbia University (U.S.)	43.01	-70.05	49
Cvi	AtGE_114_A, B, C	Cape Verde Islands	16.00	24.05	43
Est	AtGE_115_A, B, D	Estonia	59.00	25.04	15
Kin-0	AtGE_116_A, B, C	Kinneville, MI	42.47	-84.46	273
Ler	AtGE_117_B, C, D	Landsberg, Germany	48.20	10.52	628
Nd-1	AtGE_118_A, B, C	Niederzunzheim, Germany	50.78	8.03	250
Shakdara	AtGE_119_A, C, D	Pamiro-Alay, Tadjikistan	37.18	73.17	3400
Van-0	AtGE_120_A, B, C	University of British Columbia	49.85	-123.11	50

AtGE_ecotypes.txt

Which are available on ***WEIGEL WORLD*** website:

http://www.weigelworld.org/resources/microarray/AtGenExpress/

BIBLIOGRAPHY

BIBLIOGRAPHY

A. Liaw, M. Wiener (2002), Classification and regression by randomForest, R News 2/3: 18–22.

C. Strobl, A. Boulesteix, A. Zeileis, T. Hothorn (2007), Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution, BMC Bioinformatics: 8-25.

C.V. Zwan, S.Brodie, J.J. Campanella (2000), The intraspecific phylogenetics of Arabidopsis thaliana in worldwide populations, Systematic Botany 25: 47–59.

D. W. Meinke, et al. (1998), Arabidopsis thaliana: a model plant for genome analysis, Science, Vol.282.

Furlanello et al. (2003), GIS and the Random Forest Predictor: Integration in R for Tick-borne Disease Risk Assessment, Proceedings of the DSC-03 International Workshop on Distributed Statistical Computing, Vienna, Austria.

H. Pang, A. Lin, M. Holford, B.E. Enerson, B. Lu, M. P. Lawton, E. Floyd, H. Zhao (2006), Pathway analysis using random forests classification and regression, Bioinformatics, Vol.22.

Jinwook Seo, Ben Shneiderman (2002), Interactively Exploring Hierarchical Clustering Results, IEEE Computer, Vol 35: 80-86.

K. Apel, H. Hirt (2004), Reactive oxygen species: Metabolism, oxidative stress and signal transduction, Annual Review Plant Biology, Vol.55: 373–399

L. Breiman (2001), Random Forests, In Machine Learning, Vol.45: 5-32

L. Breiman, A. Cutler, Random Forests.
URL:http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_papers.htm

M.Schmid, T.S.Davison, S.R.Henz, U.J.Pape, M.Demar, M.Vingron, B.Schölkopf, D.Weigel, and J.U.Lohmann (2005), A gene expression map of Arabidopsis development. Nature Genetics, Vol37: 501-506.

R. Diaz-Uriarte (2004), Variable Selection from Random Forests: Application to Gene Expression Data, Spanish Bioinformatics Conference 2004: 47-52.

R. Diaz-Uriarte (2005), GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest, CNIO, Spain.

R. Diaz-Uriarte, S. Alvarez de Andres (2006), Gene selection and classification of microarray data using random forest, BMC Bioinformatics 2006, 7:3.

S. Karpinski, H. Reynolds, B. Karpinska, G. Wingsle, G. Creissen, P. Mullineaux (1999), Systemic signaling and acclimation in response to excess excitation energy in Arabidopsis. Science Vol.284: 654–657.

Smyth, G. K. (2004), Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology* **3**, No. 1, Article 3.

T.K. Ho (1995), Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada: 278-282.

U. Johanson, et al. (2000), Molecular Analysis of FRIGIDA, a Major Determinant of Natural Variation in Arabidopsis Flowering Time, Science Vol.290.

Y. Truong, X. Lin, C. Beecher (2004), Learning a complex metabolomic dataset using random forests and support vector machines, KDD (Knowledge Discovery and Data Mining): 835-840.