2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

Essential Gene Identification for a Microarray Data of Yersinia Pestis

Lianfen Qian[1,2], Wei Zhang[1,3] and Zhongwei Li[4]
[1] Department of Mathematical Sciences, Florida Atlantic University, Boca Raton, USA. Email: Lqian@fau.edu
[2] College of Mathematics and Information Science, Wenzhou University, Zhejiang, China
[3] Weill Cornell Medical College, New York City, USA. Email: weizhangfau@gmail.com
[4] Department of Biomedical Science, Florida Atlantic University, Boca Raton, USA. Email: zli@fau.edu

***Abstract*—This paper is motivated by a DNA microarray data obtained from a genome-wide mutation library for the bacterium Yersinia Pestis. The purpose of this study is to identify essential genes for the bacterium Yersinia Pestis. The data set contains more than four thousands genes and each gene has different number of observations with unequal number of probe observations. We propose a feature selection method for the representing three probes and a new gene level adjusted multiple statistical test to handle the problem of unequal number of observations. The proposed method is compared with two other methods based on Behrens-Fisher method and Hotelling t-square method. Our results show that our proposed method is more suitable among the three for identifying essential genes using the DNA microarray data.**

***Keywords*—Microarray; genome-wide mutagenesis; gene-level adjusted multiple t-test; number of probes; essential genes; Yersinia pestis.**

I. Introduction

DNA microarray technology is a powerful approach for genomic research due to its ability to detect and compare the abundance of tens of thousands genes and their RNA products in parallel. As expected for such high throughput analysis, microarray data are characterized with high dimensions and small sample sizes. Making statistical inference for such high dimensional data structures has been challenging. It is important to develop new statistical methods, especially those suitable for analyzing specialized microarray data.

Traditionally, microarray technology has been mainly used to analyze gene expression profiles. Massive amount of data is available to provide valuable information about gene functions, inter-gene dependencies and underlying biological processes, and opens a new avenue for discovering gene co-regulations, gene interactions, metabolic pathways and gene-environment interactions. Several data-mining methodologies, such as clustering analysis and classification techniques, have already been widely used to analyze gene expression data for identifying groups of genes sharing similar expression profiles. Variations of microarrays have been used for many other purposes, and methods for statistical analysis for these varied forms require customization and design that would better serve the specific need of the particular projects.

In this work, we have applied a custom high density oligonucleotide array to identify essential genes in the bacterial pathogen Yersinia pestis. Such genes may be potential targets for developing novel antibiotics. The experimental approach is described in Figure 1. Briefly, a mutant library of Yersinia pestis KIM5 was constructed by transposon mutagenesis so that each of the more $4,000$ genes was disrupted about 20 times in average by independent transposon insertions. After growing a limited number of generations, the mutants that have inactivated essential genes will be depleted in the library, and the abundance of the corresponding transposon sequences interrupting essential genes will be much lower than that of those interrupting non-essential genes. Probes containing transposon-inserted sequences prepared from this library were labeled with Cy3 (colored green, G) and were used to hybridize a genome-tiling microarray containing 40-nucleotide oligomers. Probes prepared from the wild type were labeled with Cy5 (colored red, R) and were used as control. Therefore, after hybridization, the color on array spots would demonstrate a normal ratio of R/G for the oligonucleotides representing non-essential genes, and a greater ratio of R/G for those oligonucleotides of essential genes. One common observation is that the distribution of R/G is skewed to the right. We suggest a transformation of the R/G ratio by logarithmic base two. The advantage of this transformation is that it produces a continuous spectrum of values for differentially survived genes.

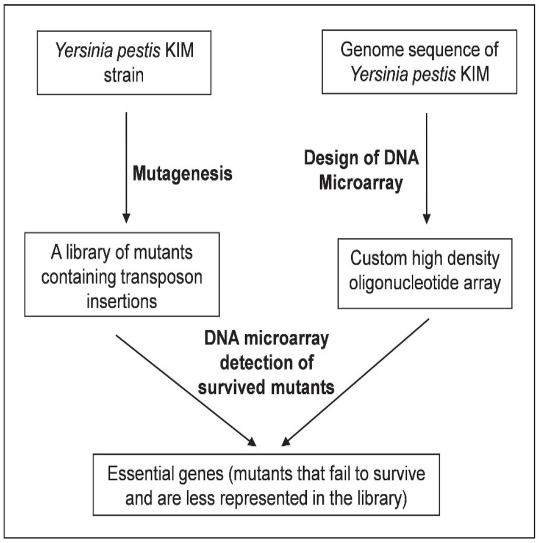
Yersinia pestis KIM strain
Genome sequence of *Yersinia pestis* KIM
Mutagenesis
Design of DNA Microarray
A library of mutants containing transposon insertions
Custom high density oligonucleotide array
DNA microarray detection of survived mutants
Essential genes (mutants that fail to survive and are less represented in the library)

Fig. 1. Outline of Yersinia pestis project

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.20

Because of the nature of the experimental approach, it is difficult to report gene essentiality based on simple R/G ratio analysis. First, the mutant library may contain mutants that survive at various levels, making no clear cut in the abundance of transposon in essential genes vs. in non-essential genes. Second, probe preparation was based on PCR amplification on total genomic DNA which may result in the presence of a large quantity of nonspecific products and hence nonspecific signals on the DNA microarray. Third, the probes were generated using restriction fragments of genomic DNA that may sometimes cover intergenic sequences next to the essential genes, producing false negative results. Fourth, each probe may cover one or more oligonucleotides on the array, and each gene may cover one or more probes, making gene identification highly complicated. A good method for data analysis would have to take into consideration of these factors and is able to deal with the dirtiness of the data.

Early analysis of microarray data relied on the fold change cut-offs to identify essential genes. Typically a two-fold change is taken as the cut-off. That is, if a gene has R/G greater than 2, then the gene is said to have differentially expressed. Shena et al. (1995) used a spiked control in mRNA samples to normalize the signals for the two fluorescent dyes and declared a gene as differentially expressed if the difference of the expression levels is more than 5 in two mRNA samples. In our work, one of the issues is that genes are covered by one or more probes and each of the probes are hybridized to one or more microarray spots, raising a problem about the efficiency of analysis. The varied lengths of genes cause varied number of observations among the genes. Hence representative spots selection became one of the major issues before conducting formal statistical analysis. Though Antipova et al. (2002) suggested a way to reduce probes for genes, it is not suitable for our data set.

To overcome the complicated data features mentioned above, we grouped the data according to the probe fragmentation and identified 3229 genes, each contains at least one probe covering only the gene under study but not intergenic regions or adjacent genes. These genes were analyzed in this work using the hybridization signals derived from their exclusive probes. We first conducted lowess smoothing, then adjusted by median to normalize the abundance of transposon. The final gene identification results were verified by comparing with experimentally proven essential genes from Escherichia coli because Y. Pestis and E. coli are closely related bacteria and they potentially share most essential genes.

This paper is organized as follows: Section 2 overviews the Yersinia Pestis DNA microarray data. Section 3 proposes the new gene-level adjusted multiple statistical test after an exploratory gene feature analysis. Two other methods are briefly introuduced for comparison purpose. Those two methods are based on Behrens-Fisher method (Shrestha an Ramachandran, 2008) and Hotelling t-square method (Lu et al., 2005). The results on the number of essential genes detected are presented in Section 4.

II. Yersinia Pestis DNA Microarray Data Overview

The dataset contains the abundance measures for both wild type (red) and mutant (green) for both positive and negative strands. There are a total of 3229 Yersinia pestis genes and 69794 40-basepairs oligonucleotide probes meansurements. On average, each gene is represented by about 25 probes. Most of the genes have less than 40 probes. The only very long gene is yapH which is cut into 273 probes. There are 14 genes relatively shorter, but still very long with the number probes between 100 and 200. They are hylA, gltB, tcaC1, clpB, rpoB, rpoC, acrB, yapA, y1697, y2267, irp2, irp1, mukB and y3333. As shown in Figure 2, the distributions of ratios are skewed to the right hence we introduce log_2 transformation. Figure 3 shows the distributions of $\log_2(R/G)$.

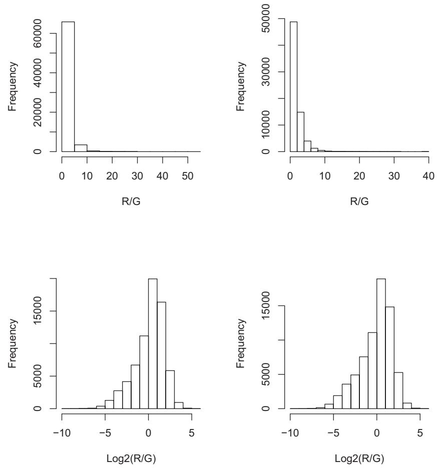

Fig. 2. Distributions of R/G ratios of positive (left) and negative (right) strands.

III. Gene Classification Methods

In this section we propose our gene-level adjusted multiple statistical test, and compare the proposed methods with two existing methods: Behrens-Fisher method and Hotelling t-square test. The proposed method is detailed in Section III-A, while Behrens-Fisher and Hotelling's t-square based methods are briefly introduced in Sections III-B and III-C, respectively, for the completeness of the paper.

A. Gene-level adjusted multiple statistical test

Recall that in our dataset, there are unequal number of probes for each gene. For each gene, two strands are present: positive strand and negative strand. The logratio of red versus green fluorescence abundances of transposons are calculated and used to construct statistic. To be more precise, let X be the lowess normalization followed by median adjusted logratio of red versus green fluorescence abundances. Let n_j be the number of the probes for gene j. Denote X_{ij}^P and X_{ij}^N as the lowess normalization and median adjustment logratio fluorescence abundance for the ith probe of gene j from the positive and negative strands, respectively. The normalized and adjusted logratio vectors for gene j for the positive and negative strands are

$$X_j^P = (X_{1j}^p, X_{2j}^p, \ldots, X_{n_j j}^p) \text{ and } X_j^N = (X_{1j}^N, X_{2j}^N, \ldots, X_{n_j j}^N).$$

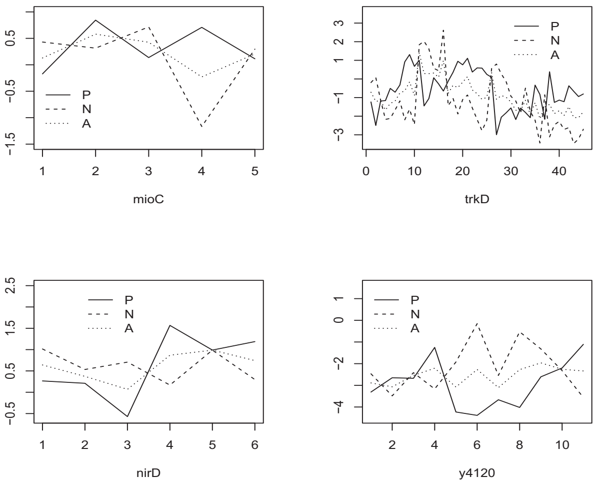

Fig. 3. Log-ratios against the probe location of genes mioC, nirD, trkD and y4120

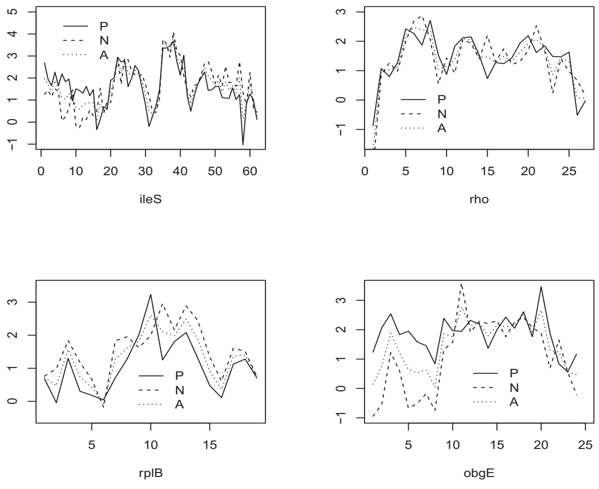

Fig. 4. Log-ratio against the prob location for genes: ileS, rho, rplB and obgE

Our proposed method is based on the feature analysis of the logratio against the location of the probes for each gene. After examining a random sampled genes, we discover two features for most of genes: (1) the average of the adjusted logratio is less than 1 and the overall average is close to 0; (2) the adjusted logratio shows opposite trend for the positive strand and the negative strand: that is, if in positive strand, X_{ij}^P is less than $X_{i+1,j}^P$ or $X_{i-1,j}^P$, then it is highly possible that in negative strand, X_{ij}^N is greater than $X_{i+1,j}^N$ or $X_{i-1,j}^N$. To illustrate these features, we randomly select four genes and report the graphs of the logratios in Figure 4.

For feature comparison, we randomly select four experimentally tested E. coli essential genes, also presented in the Yersinia Pestis and examine the logratio features in Figure 5. Figure 5 shows high values for the logratio of red versus green abundance for E. coli essential gene. This is expected since the DNA microarrray data were collected for detecting essential genes under normal growth. One also observes that essential genes don't follow the opposite-trend rule as regular genes do. For example for gene ileS, the average logratio of gene ileS is greater than 1 at many positions and gene ileS significantly breaks the opposite-trend rule at probe 23, probe 38, and probe 41. These three probes happen to be the positions where the positive logratio has 3 peak values. Similarly for gene rplB, the average logratio of gene rplB is greater than 1 at most positions but significantly breaks the opposite-trend rule at probe 10, probe 13, and probe 18. Also notice that these three probes happen to be the positions where the positive logratio has 3 peak values. Similarly features for the rest chosen genes in Figure 5.

Therefore, if we define statistic which can involve both the position information and breaking rule behavior for each individual gene, we are able to distinguish those genes whose average logratios are significantly different from the other genes. Based on the feature analysis, we propose two statistics to detect essential genes: the position statistic and the trend statistic. For gene j, the position statistic is defined as a_j, the average of the the logratios. The trend statistic is defined as below:

1. Top three probes: Let m_1 be the index of the probe where X_{ij}^P attains maximum, m_2 for the second peak and m_3 for the third peak.
2. Let $I(\cdot)$ be an indicator function. If $1 < m_1 < n_j$, let

   $D_{1j}^P = (X_{m_1,j}^P - X_{m_1-1,j}^P)$, $D_{1j}^N = (X_{m_1,j}^N - X_{m_1-1,j}^N)$

   and

   $D_{2j}^P = (X_{m_1,j}^P - X_{m_1+1,j}^P)$, $D_{2j}^N = (X_{m_1,j}^N - X_{m_1+1,j}^N)$.

   Then, we define our statistic as:

   $$Z_{1j} = \frac{Z_{11j} + Z_{12j}}{2},$$

   where

   $$Z_{1ij} = \begin{cases} \left|D_{ij}^P\right| + \left|D_{ij}^N\right|, & \text{if } D_{ij}^P D_{ij}^N \geq 0, \\ \left|D_{ij}^P + D_{ij}^N\right|, & \text{if } D_{ij}^P D_{ij}^N < 0 \end{cases}$$

   for $i = 1, 2$. For $m_1 = 1$ and $m_1 = n_j$, we define $D_{1j}^P = D_{2j}^P$, $D_{1j}^N = D_{2j}^N$ and $D_{2j}^P = D_{1j}^P$, $D_{2j}^N = D_{1j}^N$, respectively.
3. Repeat the same process for second peak and third peak, which gives us two other statistics. Denote them by Z_{2j} and Z_{3j} respectively.
4. We choose the maximum of the three statistics:

   $$Z_j = max(Z_{1j}, Z_{2j}, Z_{3j}).$$
5. Outlier adjustment: Among the three probes, denote the number of probes at which the fluorescence abundance breaks the opposite trend rule by k_j. Then we utilize an outlier adjustment multiplier $O_{k_j} = 2^{k_j - 1}$ and let $Z_j^* = O_{k_j} Z_j$.

6. Length adjustment: Since some gene only has less than 5 probes while some others have more than 100 probes, we propose to do a probe length adjustment in order to make our result more reliable. Here we only do adjustment for those genes with extremely short length or long length.

$$\begin{aligned} Z_j^{'} &= \left(\frac{Z_j^*}{c_1\sqrt{n_j}}\right) I(0 \leq n_j < 10) + Z_j^* I(10 \leq n_j < 100) \\ &\quad + \left(\frac{Z_j^*}{c_2\sqrt{n_j}}\right) I(n_j \geq 100). \end{aligned}$$

We choose the constants c_1 and c_2 by minimizing the mean of differences between the statistics from the different strands of the gene according to the probe length. This criterion is intended to minimize the probe length dependency. Other methods of choosing c's are shown in Gao et al. (2011).

7. Position statistic and trend statistic combination: Finally we multiply the position statistic by the trend statistic to define the proposed statistic: $\tilde{Z}_j = a_j * Z_j^{'}$. Therefore, if one gene not only has significant high logratio, but also breaks the opposite-trend rule, we can easily detect it out by right-tail test; if one gene doesn't have significant high logratio, but still breaks the opposite-trend rule, we can also detect it out if the trend statistic is extremely high compared to the average. This is also true if the the gene doesn't significantly breaks the opposite-trend rule while it does have extremely high average logratio; at last, if the gene has low average value and doesn't break the opposite-trend rule, this statistic will be very small, hence is classified as nonessential.

The computing algorithm for each gene can be visualized through the flow chart in Figure 6.

Using the proposed statistic, we conduct right-tail tests according to the purpose of the microarray data. Due to the multiple hypothesis testing problem, we choose to control both FWER (the probability of at least one false positive in the set of significant genes) and FDR. For FWER, we use the Holm's step-down procedure (1979). Figure 6 shows the histogram of unadjusted p-values and the volcano plot. The plotted along the x-axis is a measure of effect size (here, the mean fold-change) and along the y-axis is a measure of significance (here, the negative logarithm of the p-value). One can find that some of the genes have a large average fold-change, but low significance, and high significance does not always imply large effect size. From Figure 6, one observes a number of genes with very low p-values (which correspond to essential genes) and a whole range of insignificant p-values.

It is well known that FWER is greater than FDR in general. Therefore, FDR control is less strict than the FWER control, hence it has more power. They are equal only if all null hypotheses are true. Instead of controlling the FWER, Benjamini and Hochberg (1995) proposed a method to control the FDR. We utilize Benjamini and Hochberg (1995) procedure as implemented in R package *multtest* to control the FDR.

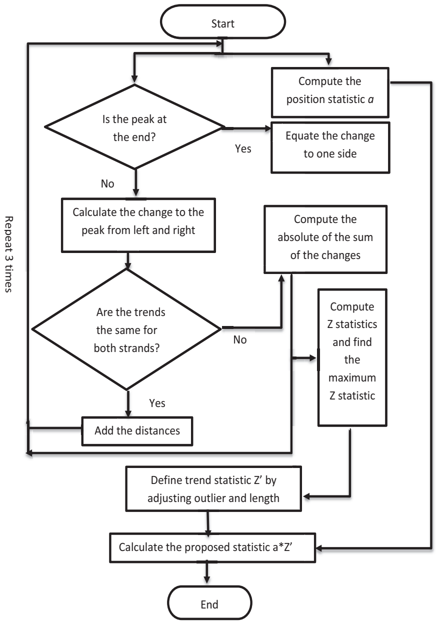
Start
Compute the position statistic *a*
Is the peak at the end?
Yes
Equate the change to one side
No
Calculate the change to the peak from left and right
Compute the absolute of the sum of the changes
Repeat 3 times
Are the trends the same for both strands?
No
Compute Z statistics and find the maximum Z statistic
Yes
Add the distances
Define trend statistic Z' by adjusting outlier and length
Calculate the proposed statistic a*Z'
End

Fig. 5. The flow chart of the proposed algorithm

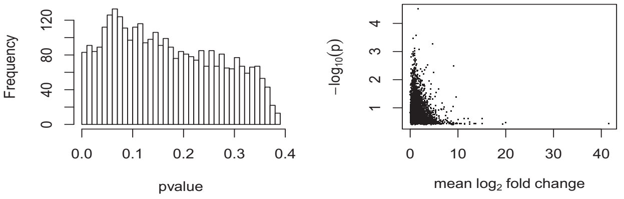

Fig. 6. The histogram of the unadjusted p-values (left) and the volcano plot (right) of the tests

B. Behrens-Fisher method

Since positive and negative strands are from two populations with unequal variances, in this case neither a pivotal statistic nor an exact confidence interval procedure exist, see Efron, Tibshirani and Tusher (2001) for reference. To solve the inhomogeneous variance problem, Shrestha and Ramachandran (2008) introduced Behrens-Fisher method. For completeness, we brief introduce the method as below:

For each gene, let μ_0^P and μ_0^N be the prior means of the logratios, m and n be the number of probes, $\bar{x}^P$ and $\bar{x}^N$ be the sample means, s_P^2 and s_N^2 be the sample variances, respectively. Then the Behrens-Fisher statistic is defined as

$$B = \frac{\mu_0^P - \mu_0^N - (\bar{x}^P - \bar{x}^N)}{\sqrt{s_P^2/m + s_N^2/n}} = B_P cos\theta - B_N sin\theta,$$

where,

$$tan\theta = \frac{s_N/\sqrt{n}}{s_P/\sqrt{m}}, \quad 0 \leq \theta \leq \frac{\pi}{2},$$

TABLE I. PRE-CLASSIFICATION OF GENES

Number of probes	Number of genes
$(0, 3)$	83
$[3, 10)$	708
$[10, 20)$	1015
$[20, 30)$	719
$[30, 40)$	328
$[40, 50)$	166
$[50, 60)$	89
$[60, 70)$	46
$[70, 300)$	75

and

$$B_P = \frac{(\mu_0^P - \bar{x}^P)}{s_P/\sqrt{m}},\ B_N = \frac{\mu_0^N - \bar{x}^N}{s_N/\sqrt{n}}.$$

It has been shown in Shrestha and Ramachandran (2008) that the statistic B is asympotically distributed as the Behrens-Fisher (BF) distribution with υ_m and ω_n degrees of freedom, where $\upsilon_m = m + \mu_0^P - 1$, $\omega_n = n + \mu_0^N - 1$.

Due to the complexity of the pdf of the BF-distribution, it is very hard to compute the corresponding probabilities, especially due to the possibility of the fractional degrees of freedom. Because of this, various types of approximations have been proposed. We shall adopt Patil's t-approximation as in Shrestha and Ramachandran (2008): $\frac{B}{a} \sim t_{(b)}$, where $a^2 = \frac{(b-2)}{b} f_1, b^2 = 4 + \frac{f_1^2}{f_2}$, and

$$f_1 = \left(\frac{\omega_n}{\omega_n - 2}\right) cos^2\theta + \left(\frac{\upsilon_m}{\upsilon_m - 2}\right) sin^2\theta,$$

$$f_2 = \frac{\omega_n^2}{(\omega_n - 2)^2(\omega_n - 4)} cos^4\theta + \frac{\upsilon_m^2}{(\upsilon_m - 2)^2(\upsilon_m - 4)} sin^4\theta.$$

Therefore, we can calculate the p-value for each gene, followed by the multiple testing procedure controling FWER or FDR.

C. Hotelling's t-square method

The Hotelling's t-square statistic is a natural multidimentional extension of the t-statistic that is currently a widespread approach for detecting essential genes in testing individual genes, see Lu et al. (2005) and Song et al. (2006). But it requires the equal number of observations for all genes.

Therefore, before we adopt the Hotelling's t-square method, we do a pre-classification for our dataset. Genes with less than 3 probes go to the first group, genes with probe numbers greater or equal to 3 and less than 10 go to the second group, and so on, as shown in Table 1.

There are two strands for each gene, and these two strands serve as our two populations. Genes might be correlated to each other, so we use Hotelling's t-square test. In this section, we adopt the multiple forward search algorithm from Lu et al. (2005). For completeness, we brief introduce this algorithm as below:

We consider K genes from some group in Table 1 and denote $X_{i.}^P = (X_{i1}^P, ..., X_{iK}^P)'$ and $X_{i.}^N = (X_{i1}^N, ..., X_{iK}^N)'$ for the ith observation of the positive and negative strands, respectively. Let $\bar{X}_j^P$ and $\bar{X}_j^N$ be the sample average of positive and negative for gene j, $j = 1, ..., K$. Let n be the smallest number of probes for the K genes. Note that since the group interval length is at most ten (except the last class interval in Table 1), the truncated portion may not cause big loss of information. Let the average fluorescence abundance vector for K genes be $\bar{X}^P = (\bar{X}_1^P, ..., \bar{X}_K^P)'$ and $\bar{X}^N = (\bar{X}_1^N, ..., \bar{X}_K^N)'$, the corresponding sample variance-covariance matrix be S^P and S^N for positive and negative strands, respectively. Then the pooled variance-covariance matrix is

$$\begin{aligned} S &= \frac{(n-1)S^P + (n-1)S^N}{2n-2} \\ &= \frac{1}{2n-2}\left[\sum_{i=1}^{n}\left(X_{i.}^P - \bar{X}^P\right)\left(X_{i.}^P - \bar{X}^P\right)' \right. \\ &\quad \left. + \sum_{i=1}^{n}\left(X_{i.}^N - \bar{X}^N\right)\left(X_{i.}^N - \bar{X}^N\right)'\right]. \end{aligned}$$

Therefore, Hotelling's T^2 statistic for fluorescence abundance is then defined as,

$$T^2 = \frac{n}{2}\left(\bar{X}^P - \bar{X}^N\right) S^{-1} \left(\bar{X}^P - \bar{X}^N\right)'.$$

This statistic combines information from the mean and dispersion of all genes being tested in microarray data. The central limit theorem implies that

$$\frac{2n - K - 1}{(2n-2)K} T^2 \sim F(K, 2n - K - 1).$$

Then we use the following multiple forward searh algorithm (Lu et al. 2005) to detect essential genes:

Step 1. Calculate T^2 statistics for each gene and locate gene j_1 that maximizes T^2, denoting $T_{j_1}^2$.

Step 2. Keep gene j_1 as the first essential gene. Then find the group that gene j_1 falls. To be more specific, if p-value of $T_{j_1}^2 < \alpha$ (a predefined significance level), calculate T^2 statistic for two genes: one is the gene j_1 and the other is one of the remaining genes in the group gene j_1 falls. If the number of probes of the second gene is smaller than the first, we truncate gene j_1 by several probes; Otherwise we do permutation to make sure their probe length match. Since they are in same group, the probe length will not differ more than 9, in this sense we reduce the possible error since otherwise we may have large difference between two probe lengths. Then we find the gene j$_2$ that maximizes T^2 combining with the gene j_1, denoting T_{j_1,j_2}^2.

Step 3. Continue to add genes one at a time to compute T^2 as in step 2, until the p value of the T^2 is increases or $2n - K - 1 < 0$.

Step 4. Exclude those genes detected from steps 1-3. Repeat steps 2 and 3 for the gene in the step 1 with the second largest T^2 statistic. Stop this loop until we get to the genes with p-value of $T^2 \geq \alpha$ (a predefined significance level), and stop searching.

IV. RESULTS

In this section, we report the detected number of essential genes and number of non-essential genes using all three

TABLE II. Results of essential genes detected: 1=Gene-level adjusted multiple t-test, 2=Behrens-Fisher method, 3=Hotelling's t-square, (1,2,3)=common in three methods.

Method	Essential	Nonessential	Genes in Ecoli list
1	507	2722	117
2	699	2530	42
3	306	2923	35
$(1, 2, 3)$	72	2369	

TABLE III. Comparison of Gene-level adjusted t-test using FDR and FWER criteria

Criterion	Essential	Nonessential	Genes in Ecoli list
FWER	507	2722	117
FDR	533	2696	128
Both	496	2698	113

methods. One notices from Table 2 that Behrens-Fisher method detects a large number of essential genes, followed by gene-level adjusted multiple t-test, and then Hotelling's t-square test. Among those genes, only 72 genes were detected essential by all three methods. In fact, out of the 3229 Yersinia Pestis genes studied in this work, there are only 193 genes presented in the reference E. coli essential gene list. Notice that the newly proposed gene-level adjusted multiple test detects about 61% of essential genes in the reference E. coli essential gene list, while Behrens-Fisher method and Hotelling's t-square method only pick up about 22% and 18%, respectively, as shown in the last column of Table 2.

Table 3 shows the comparison of gene-level adjusted multiple t-test controling both FDR and FWER criteria. From Table 3, one observes that the two controling criteria give rise reliable list of essential genes, with a little higher number by controlling FDR than controling FWER. Among the common detected 496 essential genes, 113 genes, about 59% in the reference E. coli essential gene list, are classified as essential by both controling criteria.

V. Conclusion and Discussion

Based on the feature analysis of the gene hybridization signals along the probe location, we subjectively select three top probes as the representatives for each gene. Utilizing those three selected probes, we define our gene-level adjusted multiple t-test statistic. The benefit of this method is the essential feature extraction from each gene and so that the essentiality of the gene is detected. The limitation of this method is the subjective number of probes selected. The question is how many representative probes one should select for such kind of data with highly varying number of observations for all genes? Similarly, on the outlier adjustment and length bias adjustment, our choice of mupliers are subjective. It is necessary to study the reliability of the essential genes detected through knock out experimental study.

Acknowledgement

The authors thank Xin Gong for experimental work which provided the data for the current study.

Funding: This work was funded by US Army Research, Development & Engineering Command (RDECOM) through University of South Florida Center for Biological Defense to ZL, Contract W911SR-07-C-0084 Subagreement #6415-1012-61-A.

References

[1] Antipova, A., Tamayo, P. and Golub, T.R. (2002). A strategy for oligonucleotide microarray probe reduction. *Genome Biology*, 3: 1186-2002.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B 57 (1): 289300.

[3] Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of American Statistical Association*, 96: 1151-1160.

[4] Gao, L., Fang, Z., Zhang, K. Zhi, D. and Cui, X. (2011). Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*, 27: 662-669.

[5] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 (2): 6570.

[6] Huang, J., Wei, W., Zhang, J., Liu, G. (2004). Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics*, 1: 287-299.

[7] Lu, Y., Liu, P. Y., Xiao, P., and Deng, H. W. (2005). Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, 21: 3105-3113.

[8] Patil, V. H. (1965). Approximation to the Behrens-Fisher distributions. *Biometrika*, 52, 267-271.

[9] Shrestha, N. K. M. and Ramachandran, K. M. R. (2008). Behrens-Fisher's distribution for selecting differentially expressed genes.*Neural, Parallel and Scientific Computations*, 16: 147-164.

[10] Song, S and Black, M. (2006). Principal Coordinates and Hotelling's T^2 for the analysis of microarray data. http://www.bioconductor.org/packages/release/bioc/html/pcot2.html.

[11] Shena, M., Shalon D., Davis R. W. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray.*Science*, 270: 467-470.