PREDICTION OF POLYADENYLATION SITES BY PROBE LEVEL ANALYSIS OF MICROARRAY DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

YİĞİT İLGÜNER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2013

Approval of the thesis:

PREDICTION OF POLYADENYLATION SITES BY PROBE LEVEL ANALYSIS OF MICROARRAY DATA

submitted by **YİĞİT İLGÜNER** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** _____

Assoc. Prof. Dr. Tolga Can
Supervisor, **Computer Enginering Department, METU** _____

Examining Committee Members:

Prof. Dr. Faruk Polat
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Tolga Can
Computer Engineering Department, METU _____

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Elif Erson Bensan
Department of Biological Sciences, METU _____

Assist. Prof. Dr. Aybar Can Acar
Informatics Insititute, METU _____

Date: 28/08/2013

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: YİĞİT İLGÜNER

Signature :

ABSTRACT

PREDICTION OF POLYADENYLATION SITES BY PROBE LEVEL ANALYSIS OF MICROARRAY DATA

İlgüner, Yiğit

M.S., Department of Computer Engineering

Supervisor : Assoc. Prof. Dr. Tolga Can

September 2013, 57 pages

In general, identification of polyadenylation sites in 3' untranslated regions of genes is carried out by DNA sequencing. However, there is no direct high-throughput screen to detect the polyadenylation sites which are activated under particular circumstances or in certain tissues. Since microarray manufacturers usually overlook the alternative polyadenylation events when their microarrays are produced, certain design decisions of these microarrays can be used for detecting polyadenylation sites. In this thesis, we introduce a method and a corresponding tool which investigates the hybridization levels of individual probes in a probe set of a transcript to identify differential expression of two subsets of probes to the upstream and downstream of a known polyadenylation site, respectively. For the identification of the putative polyadenylation sites, we also introduce a new method that is not based on sequence information. This technique analyzes the differential expression of every possible proximal/distal grouping in a probe set and detects statistically significant variations between groups. Such a variation is an indicator of a putative polyadenylation site in between the last nucleotide of the probe sequence of the proximal subset and the first nucleotide of the probe sequence of the distal subset. We apply our method to several microarray samples that are manufactured under different conditions. We discuss the performance of our method on these datasets. Our results show that we are able to detect polyadenylation sites that

are not in common polyadenylation databases but verified by biological experiments.

Keywords: microarray, probe-level analysis, polyadenylation, alternative polyadenylation

ÖZ

MİKROÇİP VERİSİNİN PROB SEVİYESİNDE İNCELENEREK ÇOKLU ADENİN BÖLGELERİNİN BULUNDUĞU YERLERİN TAHMİN EDİLMESİ

İlgüner, Yiğit

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Doç. Dr. Tolga Can

Eylül 2013, 57 sayfa

Genellikle çoklu adenin yerlerinin tespit edilmesi, DNA dizilim deneyleriyle gerçekleştirilir. Belli bir durumda veya dokuda aktif hale gelen çoklu adenin bölgeleri, doğrudan yüksek veri incelemesi yapılarak tespit edilmemektedir. Mikroçip üreticilerinin, mikroçiplerini üretirken alternatif çoklu adenin bölgelerinin varlığını ihmal etmesi nedeniyle çoklu adenin bölgelerinin tespitinde, belli mikrodizi tasarım kararları kullanılabilir. Bu tezde, bilinen bir çoklu adenin bölgesinin sırasıyla yukarı ve aşağı olarak iki alt kümesinde diferansiyel ifade belirlemek için bir transkriptteki prob kümesinin prob melezleme düzeylerini araştıran bir yöntem ve karşılık gelen bir araç anlatacağız. Ayrıca, varsayılan çoklu adenin bölgelerinin tanımlanması için, dizilim bilgilerine dayalı olmayan yeni bir yöntem tanıtacağız. Bu teknik, bir prob kümesindeki her türlü olası uzak / yakın gruplara ayırılmış ifadeyi analiz edip gruplar arasında istatistiksel olarak anlamlı farklılıklar tespit etmektedir. Böyle bir varyasyon, yakın alt kümesi, prob dizisinin son nükleotit ve uzak alt kümesi, prob dizisinin ilk nükleotid arasında kabul edilen bir çoklu adenin bölgesinin işaretidir. Bu yöntem, farklı durumlarda bulunan birçok mikroçip örneğine uygulanmıştır. Bu sonuçlardan ve yöntemin performansından da bahsedeceğiz. Sonuçlar, bilinen çoklu adenin veritabanlarında yer almayan fakat biyolojik deneylerle doğrulanmış çoklu adenin bölgelerini de tespit edebildiğimizi göstermektedir.

Anahtar Kelimeler: mikrodizi, prob-seviyesi analizi, çoklu adenin olayı, alternatif çoklu adenin olayı

To my parents

ACKNOWLEDGMENTS

I would like to thank my supervisor Assoc. Prof. Tolga Can for his support and guidance. It was a great chance and honor to work with him.

I would also like to thank Begüm H. Akman, Taner Tuncer, and Elif Erson-Bensan for using the application and for providing valuable feedback.

Finally, I would like to thank my family for their invalueable support.

This project is supported by ODTU (Orta Doğu Teknik Üniversitesi-METU [Middle East Technical University]) intramural interdisciplinary research funds DAP2010/2011.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

APA	Alternative Polyadenylation
APADetect	A Java application for detection of alternative polyadenylation events by probe level analysis of Affymetrix arrays
ASCII	The American Standard Code for Information Interchange
DNA	Deoxyribonucleic Acid
GEO	Gene Expression Omnibus
MAD	Median Absolute Deviaton
NCBI	National Cancer for Biotechnology Information
RNA	Ribonucleic Acid
UTR	Untranslated Region

CHAPTER 1

INTRODUCTION

1.1 Problem Definition and Motivation

Alternative polyadenylation (APA) is a mechanism in which the poly(A) tail of a transcript is added at different positions in the $3'$ untranslated regions(UTR). This process results in several isoforms of a gene. Different polyadenylation sites on $3'$-UTRs are usually identified by sequencing experiments [4, 5, 10, 11]. However, there is no direct high-throughput screen to identify which polyadenylation site is active in a certain condition or tissue. On the other hand, many microarray manufacturers generally do not take into account of the alternate transcripts of the same gene. This overlook brings an opportunity to detect alternative poly(A) events in a high-throughput manner using the microarray technology, by leveraging certain design choices of current arrays. By combining known poly(A) sites with the chip design information, it is possible to identify transcripts that could be screened for potential APA events. By detecting differential hybridization levels between proximal and distal probes of a transcript in a sample, one can detect APA events. Furthermore, by analyzing a group of samples against a group of control samples, a biologist can identify an APA event that is selectively activated in a given specific condition.

Poly(A) event detection without sequence information is a challenging problem in transcriptomics. It is an interesting question whether for a gene x, its k probe sequences are divided into two parts such that there are l probes in one part and m probes in the other part. The condition $l+m \leq k$ is satisfied, and these sub probe sets are significantly differentially expressed. Solving this combinatorial problem, one can identify previously uncharacterized novel poly(A) events.

In this thesis, we describe a method and a corresponding tool which analyzes the hybridization levels of individual probes in a probe-set of a transcript to identify differential expression of two subsets of probes to the upstream and downstream of a known polyadenylation site, respectively. Such differential expression could be an indicator of an APA event to be further tested by wet-lab experiments. APADetect is a user-friendly cross-platform application with a simple graphical interface written

in Java. APADetect provides biologists a platform to analyze all public microarray experiment series already available in NCBI GEO [3] to screen for APA events.

We also propose a novel method to identify new poly(A) sites. This method investigates each possible split of a probe set of a gene and tries to find the groups in which the difference between expressions of their distal and proximal probes are statistically significant. We validate the proposed method on real datasets and show that it is able to identify novel poly(A) sites which do not exist in the PolyA_DB [12] database.

1.2 Biological and Mathematical Background

1.2.1 Gene

Gene is the molecular unit of heredity; it keeps genetic information and helps it pass on from generation to generation. Genes are encoded in both strands of DNA. The structure of DNA is like a chain which is made from nucleotides. A nucleotide is composed of a five-carbon sugar, a phosphate group and a base. There are four kinds of nucleotides and they differ in the type of the base. Two strands rotate around each other as a spiral which is called as double helix structure. A strand consists of several nucleotides. Each strand has both $5'$ end and $3'$ end. On a double helix, each nucleotide matches with a specific nucleotide on its opposite strand. As a result, these two anti-parallel strands should be complementary to each other and $3'$ end of one strand corresponds to $5'$ prime of other strand. The nucleotides that are aligned towards the $5'$ end are called upstream. Downstream is the area which is towards the $3'$ end. The *positive strand* of DNA is the one whose RNA version of the sequence contains the instructions for building a protein. Therefore, the *negative strand* is the complementary sequence. Protein coding and non-coding RNAs can be copied from both strands. The structure of gene is given in Figure 1.1[1].

1.2.2 Transcription

Transcription is the process in which mRNA is made from DNA. A copy of gene is conveyed from DNA to mRNA by an RNA polymerase enzyme, while transcription occurs. When transcription finishes, a copy of the complementary strand which includes Uracil nucleobase instead of Timin nucleobase is created.

1.2.3 Messenger Ribonucleic Acid(mRNA)

Genetic information is transferred by messenger Ribonucleic Acid (mRNA) from DNA

[1] http://genome.wellcome.ac.uk/assets/GEN10000674.jpg

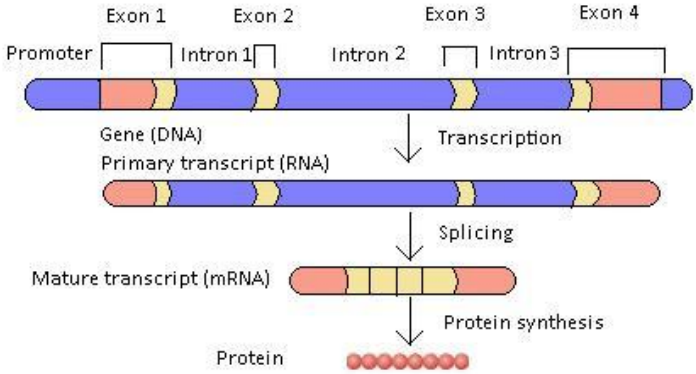


Figure 1.1: Structure of gene

to the ribosome. mRNA transcription is carried out by an RNA polymerase from DNA [8]. Most mRNAs have 5′ and 3′ untranslated regions (UTRs), exons and introns. A group of three consecutive nucleotides is called a *codon*. Exons and introns are coding and noncoding sections of mRNA, respectively. The formation of the codons on the strand of a custom mRNA is given in Figure 1.2[2]. The *5′ cap* is a modified guanine nucleotide added to the 5′ end of the pre-mRNA. It helps mRNA to recognize and attach to ribosome. Coding regions consist of codons. *Coding regions* start from the start codon and end with the stop codon. The structure of a typical human protein coding mRNA[3] is given in Figure 1.3[4] .

1.2.4 5′ Untranslated Region(5′-UTR) & 3′ Untranslated Region(3′-UTR)

5′-UTR is the section of mRNA which is placed before the start codon and it is not translated to an amino acid sequence. 3′-UTR is the noncoding or untranslated region at the 3′ end of mRNA. It begins with the nucleotide immediately following the stop codon. Untranslated regions serve several functions in gene expression, including mRNA stability and mRNA localization [8].

1.2.5 Polyadenylation

As the transcription of a gene finishes, several Adenine bases are added to the 3′ end of the mRNA. This is called polyadenylation (poly(A)) and the corresponding

[2] http://imcurious.wikispaces.com/file/view/codon.jpg/77121207/243x356/codon.jpg
[3] http://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/MRNA_structure.svg/700px-MRNA_structure.svg.png
[4] http://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/MRNA_structure.svg/700px-MRNA_structure.svg.png

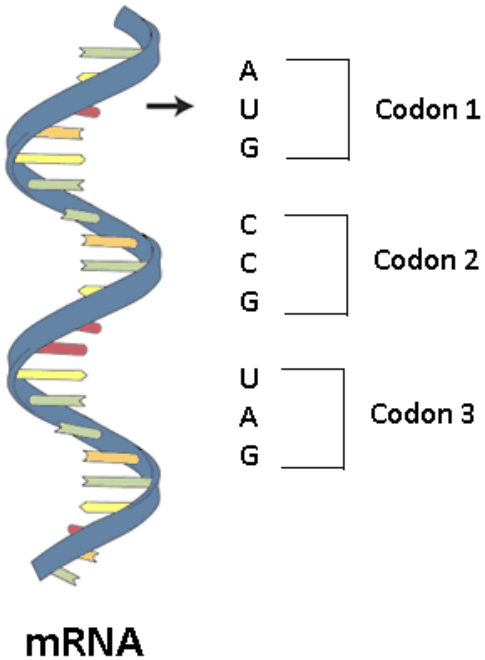


Figure 1.2: Formation on codons on mRNA

Figure 1.3: The structure of a typical human protein coding mRNA

Adenine bases are called the poly(A) tail. Sequences that belong to 3′-UTR organizes the polyadenylation. This process usually occurs at the 3′ end and involves two steps. Firstly, mRNA is cleaved at a specific site (mostly 10-30 nucleotides after the polyadenylation signal); then adenosine residues are added [9]. Proteins which participate in several operations such as stability, translation and transport of mRNA bind to the related sites of the poly(A) tail. Formation in the 3' end is related to the other transcriptional and post-transcriptional processes such as splicing and transcriptional termination. As a result, failures in 3′ end formation can have severe effects on the development, growth and viability of a cell [8]. Figure 1.4[5] shows a summary of the polyadenylation process.

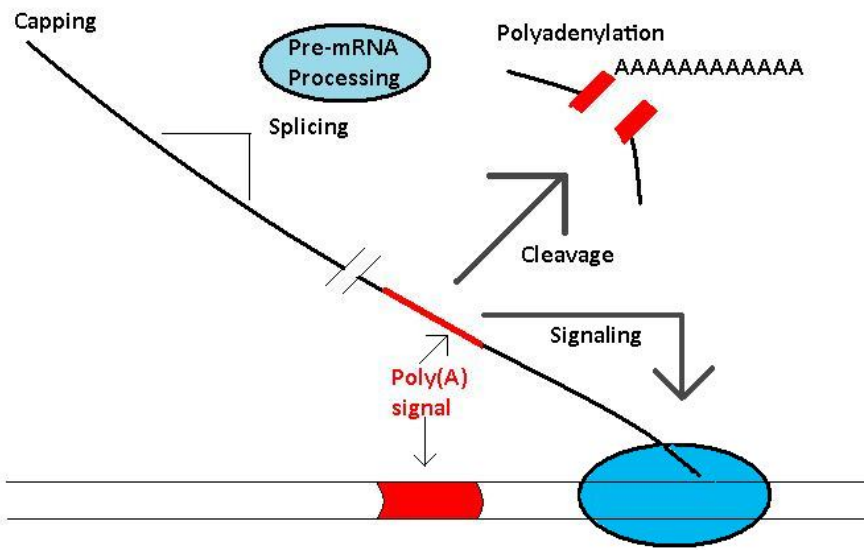


Figure 1.4: Polyadenylation process

1.2.6 Alternative Polyadenylation (APA)

Generally, the poly(A) tail is synthesized at the end of 3′-UTR.It is not unusual that sometimes more than one polyadenylation signal can be present and the poly(A) tail can be added at any of the possible sites of a mRNA. Position of polyadenylation signal may alter the reading frame, for example, in an intron. This situation may result with multiple isoforms. It causes more than one transcript to be produced from a gene [8]. Alternative polyadenylation (APA) events can lead to RNA transcripts which can differ in subtle ways such as occurrences of different length 3′-UTRs or in more serious ways, such as encoding different proteins with different domains. These variations may affect the transcript localization, stability, and transport. Even the variations in the

[5] http://www.biochemistry.ucla.edu/biochem/Faculty/Martinson/images/Slide1.jpg

untranslated regions may have serious biological effects as shown by recent studies [1]. APA sites are usually identified by sequencing studies and collected in databases such as PolyA_DB [12].

1.2.7 DNA Microarray

DNA microarray is a collection of microscopic DNA spots attached to a solid surface. It is utilized to measure to expression levels of large numbers of genes simultaneously or to genotype multiple regions. Several steps such as hybridization, labeling, scanning, data processing are applied onto a microarray sample. Final form of the sample can be thought as a two-dimensional array with spots. A summary of the microarray chip production process is given in Figure1.5[6]. Each spot contains fragments of DNA or RNA of variable lengths which is called as probes. Probes are used in DNA or RNA samples to detect the presence of nucleotide sequences (the DNA target) which are complementary to the sequence in the probe.

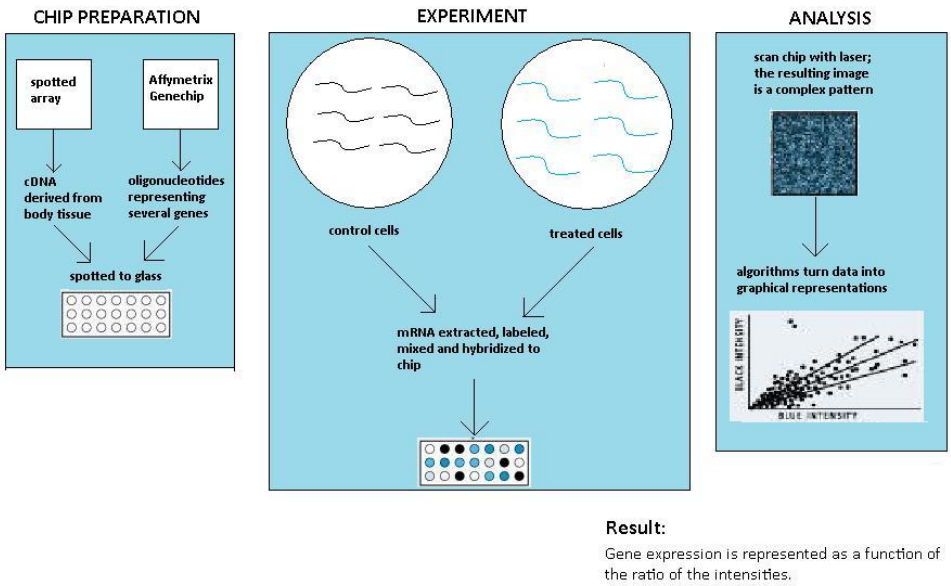


Figure 1.5: Microarray chip production process

1.2.8 Affymetrix Genechip

Affymetrix GeneChip arrays (Affymetrix, Inc., Santa Barbara, CA) are the most widely

[6] http://www.cumc.columbia.edu/publications/in-vivo/Vol1_Iss1_jan14_02/pictures/microarray-chart.jpg

used microarrays to analyze gene expression [2]. Among Affymetrix platforms, the Human Genome U133A and U133 Plus 2.0 platforms are the most commonly used platforms with 32584 samples on the U133A and 74097 samples on the U133 Plus 2.0 arrays in the NCBI GEO database (REF) as of December 12, 2012. In an Affymetrix chip, each transcript is identified by 11 to 20 unique oligonucleotide probes which are together called a *probe set*. Figure 1.6 represents a probe set with 11 probes and the distribution of those probes on the target sequence. The physical form of an Afymetrix genechip is represented in Figure1.7[7].

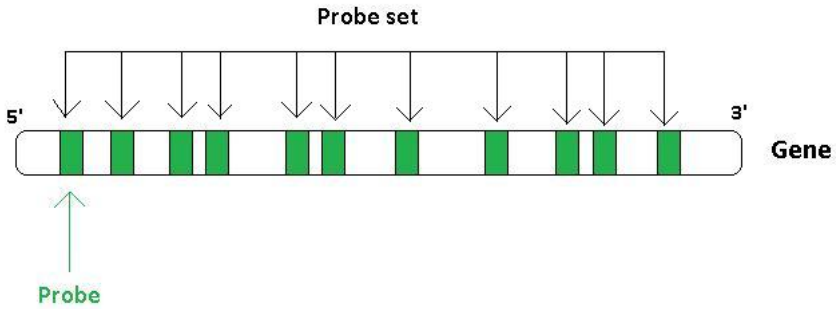


Figure 1.6: Distribution of 11 probes on a target sequence

1.2.9 Affymetrix Raw Intensity (CEL) File

In order to obtain the probe level information, the CEL file of a sample is required. A CEL file stores the results of the intensity calculations on the pixel values of an n X m probe array. This file includes an intensity value, standard deviation of the intensity, the number of pixels used to calculate the intensity value. Some of these files are in binary format, so before processing the information, binary files are required to be converted into ASCII format.

1.2.10 Distal & Proximal Probes

When a poly(A) site splits a probe set into two separate groups, the probe sequences to the upstream of the poly(A) site are called as the *proximal* probes, and the probes sequences to the downstream of the poly(A) site are called as the *distal* probes. If a poly(A) site overlaps with a probe, this probe is called neither as proximal nor as distal.

[7] http://www.igece.org/Immunology/ImmArray/Affymetrix-microarray.jpg

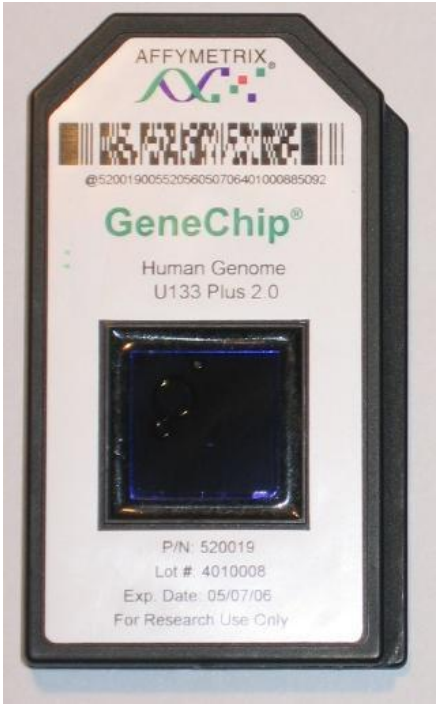


Figure 1.7: Example of an Affymetrix genechip

1.2.11 Outlier Detection

In order to get more reliable results, we try to find and discard the samples which deviate differently from the rest of the group. The reasons of the deviation can be incorrect experiments or erroneous data. Detection of potential outliers is necessary, because an outlier may be the sign of bad data and it has to be excluded from the analysis. As the outlier detection method, we used Iglewicz and Hoaglin's Median Based Outlier Detection technique. In this section, we give statistical background about this method.

1.2.11.1 Median

Median is the value of the middle element in a sorted list. It cuts the lower half of the list from the upper. We use median value instead of mean, because mean value can be affected by the values that are highly deviated from the majority and may not indicate the true distribution of the group. We use median value for eliminating the samples which relatively remain too above or too below from the generality.

1.2.11.2 Median Absolute Deviation

Median absolute deviation of a set is the median of the absolute differences between each sample and the median value of the group. For a data set with the members X_1, X_2 X_3, X_4,. . . , X_n , median absolute deviation of the ith member is formulated as

$$MAD = median_i \left(|X_i - median_j(X_j)|\right) \tag{1.1}$$

In this study, median absolute deviation is preferred because of its robustness to outliers.

1.2.11.3 Iglewicz and Hoaglin's Median Based Outlier Detection

In this method, modified z-score of the data set is calculated to label outliers. For a data set whose elements are X_1, X_2 X_3, X_4,. . . , X_n , modified z-score of the ith member is found as

$$M_i = \frac{0.6745 * (X_i - median_j(X_j))}{MAD} \tag{1.2}$$

This modified z-score can be used in outlier detection, and detected outliers can be excluded from a set of values that are to be analyzed.

1.2.12 Normalized Standard Deviation

Normalized standard deviation is computed by dividing the standard deviation to the mean. Normalized version gives us an opportunity for a more appropriate comparison between the samples with different scales of standard deviations. For a data set with the members X_1, X_2 X_3, X_4,. . . , X_n, formula of normalized standard deviation is

$$\sigma_N = \frac{\sigma}{\bar{X}} \tag{1.3}$$

where σ is standard deviation and $\bar{X}$ is mean of the set.

1.2.13 Standard Deviation

In statistics, the standard variation represents the amount of variation from the mean. A large standard deviation indicates the data points are far from the mean and a small standard deviation indicates they are clustered closely around the mean. For a data set that consists of the elements X_1, X_2 X_3, X_4,. . . , X_n, the standard deviation is computed as

$$\sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n}} \tag{1.4}$$

1.2.14 Welch's T-test

Welch's t-test is a conformation of Student's t-test. Welch's t-test is aimed for use with two samples which have have possibly unequal variances. This test defines the statistic t by the following formula

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{1.5}$$

where $\bar{X}_i$, s_i^2, N_i are the ith sample mean, sample variance and sample size, respectively.

The degrees of freedom, υ, associated with this variance estimate is approximated using the following equation

$$\upsilon = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 * \upsilon_1} + \frac{s_2^4}{N_2^2 * \upsilon_2}} = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 * (N_1 - 1)} + \frac{s_2^4}{N_2^2 * (N_2 - 1)}} \tag{1.6}$$

Here $\upsilon_i = (N_i - 1)$, indicates the degrees of freedom associated with the ith variance estimate.

Once t and υ have been computed, these statistics can be used with the t-distribution to test the null hypothesis that the two population means are equal, or the null hypothesis which one of the population means is greater than or equal to the other population mean. Especially, the p-value which is found as a result of this test may or may not give sufficient evidence to reject the null hypothesis.

1.3 Contributions

There are two contributions in this thesis.

1. We propose a new method which does not use the information of known poly(A) sites to detect polyadenylation events. Our method analyzes the differential expression of every possible proximal/distal grouping in a probe set and detects statistically significant variations between groups. Such a variation indicate a putative poly(A) site in between the last nucleotide of the probe sequence of the proximal subset and the first nucleotide of the probe sequence of the distal subset. These novel sites can further lead to other biological studies for validation of these sites.

2. We propose a new alternative polyadenylation events screening tool based on probe level analysis of microarray data which differentially denotes the probes divided by polyadenylation sites and visually explains the result in various formats.

1.4 Thesis Outline

This thesis is organized as follows: In Chapter 2, our poly(A) site detection technique which takes advantage of the design decisions of the Affymetrix chip and exploits the probe-level information is explained in details. In Chapter 3, we describe our alternative polyadenylation event detection method and the tool we developed. In Chapter 4, experimental results of the proposed method in Chapter 2 are shown. In Chapter 5, the thesis is concluded with a summary and future work.

CHAPTER 2

MATERIALS AND METHODS

In this chapter, to identify putative poly(A) sites ,we introduce a new method that does not depend on the knowledge from any poly(A) site database. This means we also are able to analyze the genes which are not reported in PolyA_DB. This method analyzes every possible split grouping in a gene and picks the ones with significantly different proximal and distal probe groups. In order to obtain robust results, we eliminate each sample which has inconsistent proximal and distal probe groups and a different variance than the rest of the samples. Finally, we apply statistical tests to leftovers and pick the significant ones. The details of this procedure are expressed in the following sections.

2.1 Extracting Coordinates on Array and on Target Sequences

The target sequence and the array coordinates on the chip for each probe are computed and listed into a text file. Each line represents a probe set and contains Unigene name, probe set ID, and number of probes. Furthermore, for every probe, corresponding location on the gene and on the microarray (x-y coordinates) are written. Each column is separated by a tab. This file is platform-specific which means every Affymetrix chip has its own unique list. The procedure for a single array platform is described below.

Firstly, Unigene names and corresponding Affymetrix probe set IDs are taken from the annotation file of the chip. Annotation file of a chip is a comma-separated file which is downloaded from Affymetrix website and contains data about the probe sets that are belong to the chip. Each line shows the probe set name, chip type, Unigene ID of the set and additional notes. There may be multiple probe sets for the same Unigene, however the probe set IDs are unique, so it is certain that a probe set ID is encountered only once.

Secondly, the probe information file is processed and probe locations are obtained for each probe in the probe sets. The probe information file includes the matching x and y coordinates of the probes in the array chip. Target location of every probe is also given in that file. Since the location in the file is the middle position of a 25mer on the

target gene, 13 is subtracted from the target location to obtain the beginning position. This procedure is done for only the probe sets that we are interested in.

Finally, the alignment file is read and the target sequence is mapped onto the genome positions. The alignment file contains information about corresponding probe positions both in the query and the target sequence. Due to the selections from different parts of the gene, a probe set may have several blocks with different block sizes and starting locations. A position array that has the same length as the query is created and initialized. Some of the positions possibly remain unmapped, because of the faulty alignment between the Affymetrix consensus/exemplar sequence and the hg19 assembly. Although mapping of the '+' strand genes occurs in the same direction as starting positions, mapping is reversed for '-' strand. To illustrate, in a probe set whose length is 637 bases with block sizes 93,144,229,70,21 respectively, mapping procedure of the '+' strand is to begin from the starting query position of the first block and assign the target positions corresponding to those blocks. For a block that starts from 34 in the query and 14361 in the target chromosome, 35th position of mapping array has the value 14362. 36th position of mapping array becomes 14363 and this incremental assignment continues until end of the blocks. Same algorithm is applied on the other blocks. For the '-' strand, assigned position for the mapping array is decided by subtracting the (query start location +1) from the query length. For a block that starts from 34 in the query and 14361 in the target chromosome, mapping begins from the 633rd (although the result of 637 - (34 + 1) is equal to 632, index notation of the array is zero-based) and has corresponding target location, 14362. As a consequence, 632nd place of the array is given 14363. This procedure goes on consecutively until there is no blocks left. Next step is to go over the mapping array again and fix the unmapped positions. Each unmapped entry is assigned the chromosome position of the last mapped member. When mapping procedure finishes, For each probe set, the poly(A) sites that are on the same strand as this probe set are selected. Every poly(A) site is looked for whether it divides the probe set into two subsets. In the case of division, the probe is considered as a proximal probe if all of it is to the upstream of the poly(A) site, however if all of it is to the downstream of the poly(A) site, that probe is marked as distal. There is a third option that there can be a probe which is cut by the poly(A) site. That probe is assumed as neither as a proximal nor a distal probe. Non-zero proximal and distal sets indicate that we have a splitter poly(A) and it is written into the split probe set file of the chip. After mapping ends, the consecutive probes which have a smaller distance than 25 polymers between them are marked and written to the end of line. This information is used for constructing possible split probe regions in further steps as a signal of the overlapping probes. To explain briefly, there are at least two probes in both of the proximal and distal groups which are generated in the artificial split probe set groups. An example of the split regions for a single transcript is shown in Figure 2.1.

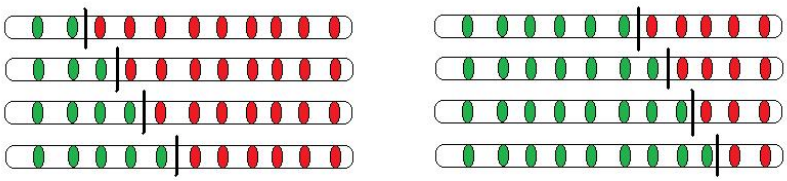


Figure 2.1: Candidate proximal and distal probe groups for a single transcript

2.2 Extracting Raw Intensities from CEL Files

In order to read and process the raw probe intensities of control samples in a rapid way, those intensities are extracted into a file that is created as ASCII format. Firstly, coordinate pairs for each probe are read from the file that is produced in Section 2.1 and stored into a hash table. Secondly, raw probe intensities of each sample are obtained from CEL files. If a probe coordinate is not occurred in the hashtable, then raw intensity belong to that probe is discarded. A probe and its raw intensities which are taken from the CEL files are put in a hash table in order to write into the output file. As the last step, for a probe set, its Unigene ID, its probe set name, target sequences and sample intensities of every probe are copied into a ASCII file. These values also correspond to the columns. Each probe intensity is delimited with a dollar ($) sign within a probe. Columns of the output file are separated by a tab character. A row consists of the Unigene name and Affymetrix probe set ID of the probe set. Furthermore, target sequences and raw intensities of the probes which are members of the probe set are components of a row.

If control samples are more than 10 samples, because of the Java heap memory restriction, they are split into the groups which consist of 10 samples and raw intensities for every group are extracted to the different files group by group. After putting an end to writing the raw intensities of each group, corresponding rows of every output file are merged and write into a text file. To illustrate, in the result files of each group, first line that contains raw probe intensities of the control samples is assumed to be belong to Probeset_A. The first lines of all files are read and combined with the sample intensities. Moreover, a larger file which contains raw probe intensities of all samples is constructed for Probeset_A. This procedure is applied for every probe set in the result file of each 10-sampled group. Computational complexity of this procedure is $\mathbf{O}(\boldsymbol{S})$ * $\mathbf{O}(\boldsymbol{P})$, where S and P indicate the number of the samples in the group and the number of probes in the microchip, respectively.

2.3 Labeling and Eliminating the Outlier Samples

Samples with unusual bias may cause misleading and incorrect results. We use Iglewicz and Hoaglin Median Based Outlier Detection in our technique so that we obtain more reliable results. We experimented three different outlier detection procedures. First method checks and calculates the modified z-score of every possible distal/proximal group, then it discards the samples with very distinct z-scores from either distal or proximal probe set depending on the identified group. Second method assumes there is only one probe set instead of two different subgroups called distal and proximal. The outlier samples are tried to be labeled and excluded from the dataset by testing the whole probe set. Final method computes proximal and distal averages sample by sample and applies Iglewicz and Hoaglin's method on these results. Details about these three types of techniques are given in the following three sections.

2.3.1 Model 1

First design applies outlier detection method for every possible proximal/distal probe set pairs and calculate average probe intensities for each probe in the probe set. In this model, we do not have two different groups named as control and treatment, but one virtual group. All scientific computations and the outlier detection algorithm are performed on that group. As a result, those average probe intensities are written into a text file in order to apply the Welch's t-test.

Firstly, ASCII-formatted file, which is constructed in Section 2.2, is opened and raw probe intensities obtained from the samples for every probe that is belong to the probe set are read. Proximal and distal probe groups are constructed heuristically by regarding two conditions:

1. Each group should have at least two probes.

2. Two consecutive probes belong the same probe set should not share common alignment positions on target gene. If that is the case, we are unable to divide these probes into two different groups. For example, a probe set is assumed to contain 11 probes and in that probe set, fourth and fifth probes are assumed to be overlapped. Since those probes cannot be taken into apart, there is no candidate proximal probe group that has four probes and no candidate distal group that consists of seven probes for that probe set.

After constructing the candidate proximal/distal group, outlier samples are detected. Outlier detection has two steps. In the first step, samples that are significantly different within the dataset are found and eliminated. The average intensities of proximal and distal probe subsets of each sample are calculated. After that procedure, by using

the percentage of the average proximal intensity to average distal intensity for each sample, it is determined whether this percentage is significantly different than the other samples in the same sample set.

Iglewicz and Hoaglin's Median Based Outlier Based Method is applied as explained in Section 1.2.11.3. For proximal groups, median value is calculated as median(intensity$_p$) where intensity$_p$ is list of the sample intensities for that proximal probe. Median absolute deviation is computed as described in Section 1.2.11.2. The modified z-score of a sample is then found as mentioned in Section 1.2.11.3. Following the author's suggestion[1], probes with absolute values of z-score greater than 3.5 are identified as outliers. If only one probe is not marked as outlier or none of the probes are labeled as outlier, that sample is removed from subsequent analyses. Same procedure is also carried out for the distal probes.

In the second step, the consistency of the probe level intensities of this transcript is tested. By consistency we mean that the intensity of the distal probes cannot be significantly larger than the intensities of proximal probes, they can either be almost equal or the intensities of the proximal probes can be significantly larger than the intensities of the distal probes. Also, the probe level intensities, both proximal and distal, should be consistent within each subset. If the variance of proximal intensities or the variance of distal intensities is significantly different, this is also considered as an inconsistency. If there is inconsistency, the transcript is identified as an erroneous read and no result is reported for this transcript.

In order to detect if there are inconsistencies within a subset, the normalized standard deviation of proximal and distal probe intensities of samples are used. The normalized standard deviation is computed as described in Section 1.2.12. If the normalized standard deviation of any single sample for this transcript for either the proximal or the distal subset is higher than or equal to 1.0, this reading is labeled as inconsistent and any result is not explained for this transcript. Also, if the average of the distal probe intensities is larger than the average of the proximal probe intensities, and the difference between these averages is greater than a **specific rate** of the proximal intensity average, this reading is also identified as inconsistent and no result is given for this transcript. If there are not any consistent samples left for a probe set, then that probe set is not added into the result file.

As a result, average probe intensities are computed for consistent samples and written to an output file. This file includes Unigene name, probe set name, total amount of the probes which are the members of the probe set. It also contains information about how much probes are put into the candidate proximal/distal groups, start location on the target sequence and average probe intensity for every probe. Each column is separated by a tab character. The computational complexity of this model is $\mathbf{O}(\boldsymbol{S})$ * $\mathbf{O}(\boldsymbol{P})$ * $\mathbf{O}(P_{Pi}{}')$ * $\mathbf{O}(\boldsymbol{prox}\{P_i\} + \boldsymbol{dist}\{P_i\})\log(\boldsymbol{prox}\{P_i\} + \boldsymbol{dist}\{P_i\})$. Here, S

[1] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

and P indicate number of the samples in the group and number of probe sets in the microchip, respectively. ${P_{P_i}}'$ denotes for number of probes in Probeset$_i$. Moreover, $prox\{P_i\}$ and $dist\{P_i\}$ are number of the proximal probes and number of the distal probes in Probeset$_i$.

2.3.2 Model 2

In this version of outlier detection method is two-phased. In the first part, Iglewicz and Hoaglin's median based outlier detection method is applied for each probe of a probe set. However, different from Section 2.3.1, our assumption is based on the existence of only one group rather than proximal and distal groups. In the second phase, after exclusion of highly deviated samples, we generate every possible split region for each probe set and identify the inconsistent samples for every possible split probe set. Unlike the usual control-treatment group separation, in this process, there is only one virtual group that can be assumed as the control group.

2.3.2.1 Phase 1

As the first step, raw probe intensities which are extracted from the related CEL files and written probe set by probe set in text file are read. This ASCII file is created as explained in Section 2.2. In the second step, significantly different samples in every sample group are marked and excluded from the result. The mean intensities of sub-groups of the proximal and distal probe are calculated for each sample. Consecutively, the ratio of the mean proximal intensity to the mean distal intensity for each sample is checked in order to identify whether this ratio is significantly different than the rest of the samples within the group.

We use Iglewicz and Hoaglin's Median Based Outlier Based Method as we mention in Section 1.2.11.3. Along a probe set, median value is calculated as median(intensity$_p$) where intensity$_p$ denotes the list of raw intensities of all samples for that probe. Median absolute deviation is computed as described in Section 1.2.11.2. The modified z-score of a sample is then found as mentioned in Section 1.2.11.3. As mentioned in Section 2.3.1, probes with absolute values of z-score greater than 3.5 are identified as outliers. If only one probe is not marked as outlier or none of the probes are labeled as outlier, that sample is removed from subsequent analyses.

Finally, for each probe set, Unigene ID and probe set ID of the probe set, total number of probes in that probe set are written into the result file. A typical row also includes target locations of the probes and raw probe intensities for each non-outlier sample. Raw intensities for the samples which are not marked as outliers are separated by a dollar ($) sign within a probe. Information in this file is utilized in Phase 2.

2.3.2.2 Phase 2

By using probe information from the file generated in Section 2.3.2.1, proximal and distal probe groups are constructed heuristically by regarding the conditions mentioned in Section 2.3.1.

After constructing the candidate proximal/distal group, we perform Iglewicz and Hoaglin's method as in Section 1.2.11.3. After outlier samples are discarded, we make consistency check for the rest of the samples as described in Section 2.3.1. We also follow the same steps with Section 2.3.1, when we write the calculated average probe intensities for consistent samples.

The computational complexity of Model 2 is $\mathbf{O}(\boldsymbol{S})$ * $\mathbf{O}(\boldsymbol{P})$ * $\mathbf{O}(P_{Pi}{}')$ * $\mathbf{O}(\boldsymbol{prox}\{P_i\}$ + $\boldsymbol{dist}\{P_i\})\log(\boldsymbol{prox}\{P_i\}$ + $\boldsymbol{dist}\{P_i\})$. S and P show the number of the samples in the group and the number of probe sets in the microchip, respectively. $P_{Pi}{}'$ represents the number of probes in Probeset$_i$. Moreover, $prox\{P_i\}$ and $dist\{P_i\}$ are the number of the proximal probes and the number of the distal probes in Probeset$_i$.

2.3.3 Model 3

In the third model, the average intensities of proximal and distal probe groups are calculated sample by sample. It is assumed there is only one group and outlier detection method is just applied for that group.

The procedure starts with reading the raw probe intensities from an ASCII file which is created as a result of Section 2.2. Proximal and distal probe groups are generated as explained in Section 2.3.1.

After constructing the candidate proximal/distal group, outlier samples are identified and excluded from subsequent analysis. Outlier detection algorithm is a two-step process. In the first step, samples which are outliers among the control samples are eliminated by using the ratio of the average proximal intensity to average distal intensity for each sample. We first computed average intensities of proximal and distal probe subsets of each sample. This leads the ratio of the average proximal intensity to average distal intensity for each sample. It is checked if there is a huge difference between this ratio and the rest of the samples in the group.

We use Iglewicz and Hoaglin's Median Based Outlier Based Method as we mention in Section 1.2.11.3. Along a probe set, median values are calculated as median(average$_p$) and median(average$_d$) where average$_p$ and average$_d$ are the average intensities for proximal and distal probes, respectively. Median absolute deviation and modified z-score of a sample is then found as mentioned in Section 1.2.11.2 and Section 1.2.11.3, respectively. Rest of the outlier detection procedure, inconsistency check and getting the final list of the average probe intensities for every probe are same as the ones

in Section 2.3.1. The computational complexity of the model is $\mathbf{O}(\boldsymbol{S})$ * $\mathbf{O}(\boldsymbol{P})$ * $\mathbf{O}(P_{Pi}{}')$ * $\mathbf{O}(\boldsymbol{prox}\{P_i\} + \boldsymbol{dist}\{P_i\})\log(\boldsymbol{prox}\{P_i\} + \boldsymbol{dist}\{P_i\})$. S and P denote for the number of the samples in the group and the number of probe sets in the microchip, respectively. $P_{Pi}{}'$ is the number of probes in Probeset$_i$. $prox\{P_i\}$ and $dist\{P_i\}$ are the number of the proximal probes and the number of the distal probes in Probeset$_i$.

2.4 Applying Welch's T-Test With Multiple Testing Correction

After the identification and discard of the highly distributed samples from the sample group, average raw intensities for each probe are calculated. We apply Welch's t-test on each probe set in order to test the following hypothesis: Mean value of proximal probes is greater than the mean value of the distal probes. When we try to remove the false positives, we want to keep true proximal/distal probe groups as much as possible. Because of this reason, Benjamini and Hochberg False Discovery Rate control method is chosen as the multiple testing correction method. P-values are corrected altogether by Benjamini and Hochberg FDR control technique. When multiple testing correction procedure finishes, final list of proximal/distal probe set pairs is formed of the significantly split probe sets.

2.4.1 Benjamini and Hochberg False Discovery Rate

Multiple testing correction helps to tune the p-values in order to eliminate the false positives. A false positive is determined by the test as a member of the result set, but the fact that it should not be included to the group. In our analysis, thousands of potential split probe sets are tested and each probe set is considered independently from others. As a multiple testing correction method, Benjamini and Hochberg False Discovery Rate control method is applied onto the p-values of potential split probe sets that are identified as the result of the analysis.

Benjamini and Hochberg False Discovery Rate control algorithm is as follows:

1. The p-values of all potential split probe sets are ranked from the smallest to the largest.

2. The largest p-value is not modified.

3. The second largest p-value is multiplied by the total number of split probe sets in the list divided by its rank. Suppose the threshold is 0.05, and every corrected p-value which is less than this threshold is considered as significant. Corrected p-value is equal to p-value*(n/n-1). If the corrected p-value remains below the threshold, than this split probe set is seen in the result list.

4. The third p-value is multiplied as in step 3. Corrected p-value is equal to p-value*(n/n-2). Candidate split probe set is considered as significant, if and only if the corrected p-value is less than 0.05

This process continues until the smallest p-value is checked. To illustrate, there are 1000 candidate split probe sets in the list and threshold is set to 0.05. First three largest p-values are supposed to be 0.1, 0.06, 0.04 for probe sets A, B, C, respectively. For Probeset A, there is not any p-value correction. Since the p-value of this probe set is 0.1, it is not included in the result. Probeset B has the rank 999. After the p-value correction, the p-value of Probeset B becomes 0.06 * (1000/999) = 0.0606, so this probe set is also not significant. After the correction, Probeset C has the p-value, 0.04008 which makes the probe set as a member of the result list.

This procedure runs in $\mathbf{O}(\boldsymbol{P})$ * $\mathbf{O}(\boldsymbol{prox}\{P_i\} + \boldsymbol{dist}\{P_i\} + \boldsymbol{n} * \log n)$ where P denote for the number of probe sets in the microchip $prox\{P_i\}$ and $dist\{P_i\}$ are the number of the proximal probes and the number of the distal probes in Probeset$_i$. In addition, n is the number of possible split probe set groups. The sorting algorithm of Java Collections takes $\mathbf{O}(\boldsymbol{n} * \log n)$ time[2].

[2] http://docs.oracle.com/javase/7/docs/api/java/util/Collections.html#sort(java.util.List)

CHAPTER 3

APADETECT

We used the poly(A) sites reported in PolyA_DB [12] in the UCSC Genome Browser [6] Database and identified target regions of transcripts that are split into two by the poly(A) site, resulting in proximal and distant probe sets. Note that if a poly(A) site is just in the middle of a probe sequence, that probe sequence is not considered as a proximal or distal probe, since it is not completely upstream or downstream of the poly(A) site. There are 18677 genes with reported poly(A) sites in PolyA_DB [12]. 10550 genes have multiple poly(A) sites. For example, for the U133 Plus 2.0 chip, there are 3067 distinct probe sets which are split into proximal and distal subsets by 4444 poly(A) sites.

After identification of probe sets which are split into distal and proximal probe subsets by a poly(A) site, we analyze the expression levels of distal and proximal probe sets on a microarray experiment to screen differentially expressed distal/proximal probe sets between control and treated samples. For a given experiment series, our tool downloads the related CEL files for each sample from NCBI GEO automatically. There are two levels of differential expression. Within a single microarray sample, the distal and the proximal probes of a probe set may be differentially expressed. We look for differential expression in which the proximal probes are highly expressed compared to the distal probes. Such a differential expression indicates that, within that sample, a shorter $3'$-UTR of the transcript is observed. Note that, the differential expression can only be detected if the probe set on this Affymetrix chip is designed for a longer $3'$-UTR of the same transcript.

A second level of differential expression may be observed between microarray samples. The distal and the proximal probes of a probe set which is differentially expressed in one sample may not be differentially expressed in the other sample. Such an observation indicates that the shorter $3'$-UTR is observed in one sample, the longer $3'$-UTR is observed in the other sample. Our tool is able to screen for such second level differential expression between different samples. Such a differentiation indicates selective activation of an APA event. For the first level differential expression, we first compute the average the intensities of the probes in a proximal/distal subset. We also average over the control or treatment replicate samples. For each gene in a microarray

experiment, we compute the fold change between the distal and proximal expression. This fold change gives us the first level differential expression. By computing the ratio of fold changes between control and treatment samples, we identify differentially expressed distal and proximal probe sequences between samples at the second level.

In our tool, we provide a list of all genes sorted by second level differential expression. The list can be exported as a Microsoft Excel document. We also provide a scatter plot of proximal to distal expression ratio for treated samples versus proximal to distal ratio for control samples. It is also possible to obtain detailed sample level intensities for genes of interest. For each such gene of interest, a separate text file that includes calculated proximal and distal probe set intensities of each sample is created. A more extensive description of the functions and user interface is given in the following sections.

3.1 Functionality

3.1.1 Constructing Split Probe Set Files

A split probe set file includes information about proximal and distal probes of a probe set which is took apart by a poly(A) site. Each line represents a probe set and contains Unigene name, probe set ID, poly(A) site name, number of proximal and distal probes. Furthermore, corresponding probe locations on the gene and on the microarray (x-y coordinates) are written. Each column is separated by a tab. The procedure for a single platform is described below.

In order to construct the split probe set file, PolyA_DB table, annotation file of the microarray chip, Affymetrix alignment table and probe info file is required. Initially, the PolyA_DB table is read and poly(A) information and Unigene names are kept in a hashtable. The PolyA_DB file has Unigene name, strand, number of poly(A) sites and names of those sites with starting positions on the gene which is attached with a dollar ($) symbol at the end. PolyA_DB file is obtained from UCSC Genome Browser [6].

Secondly, Unigene names and corresponding Affymetrix probe set IDs are taken from the annotation file of the chip. Annotation file of a chip is a comma-separated file which is downloaded from Affymetrix website[1] and contains data about the probe sets that are belong to the chip. Each line shows the probe set name, chip type, Unigene ID of the set and additional notes. There may be multiple probe sets for the same Unigene, however the probe set IDs are unique, so it is certain that a probe set ID is encountered only once.

A further step is to parse the probe information file and get probe locations for each

[1] http://www.affymetrix.com/

probe in the probe sets. The probe information file includes matching x and y coordinates of the probes in the array chip. Target location of every probe is also given in that file. Since the location in the file is the middle position of a 25mer on the target gene, 13 is subtracted from the target location to obtain the beginning position. This procedure is just done for the probe sets that we are interested in.

Finally, the alignment file is read and the target sequence is mapped onto the genome positions. This procedure is similar to the one mentioned in Section 2.1. The alignment file contains information about corresponding probe positions in the query and target sequence. Due to selections from different parts of the gene, a probe set may have several blocks with different block sizes and starting locations. A position array that has the same length as the query is created and initialized. Some of the positions may remain unmapped, because of the imperfect alignment between the Affymetrix consensus/exemplar sequence and the hg19 assembly. For each strand, target locations of genes are mapped from the upstream to the downstream. Following step is to go over the mapping array again and fix the unmapped positions. Each unmapped entry is assigned the chromosome position of the last mapped member.

3.1.2 Computing the Difference Between Perfect Match and Mismatch Probes

In this stage, absolute value of difference between the average intensities of a perfect match and mismatch probes is computed. Firstly, split probe set information file is parsed, both perfect match probe coordinates and mismatch probe coordinates are kept in a hash table. A mismatch probe has the same x position as its perfect match counterpart, however its y coordinate exceeds one position the y coordinate of the corresponding perfect match probe. Secondly, for every proximal and distal probe, average probe intensities from the control and treatment samples are calculated. Final step is to find out the absolute value of difference between average intensities of mismatch and perfect match versions for each probe from control samples and write into an intensity file. Same procedure is applied to treatment samples. Computational complexity of this procedure is $\mathbf{O(S}_C\mathbf{)} * \mathbf{O(P}_{Ch}\mathbf{)} + \mathbf{O(S}_T\mathbf{)} * \mathbf{O(P}_{Ch}\mathbf{)} + \mathbf{O(P}_A\mathbf{)} * \Big(\mathbf{O}\big(\mathbf{prox(P}_i\mathbf{)}\big) * \mathbf{O(S}_C + \mathbf{S}_T\mathbf{)} + \mathbf{O(dist}(P_i)\mathbf{)} * \mathbf{O(S}_C + \mathbf{S}_T\mathbf{)}\Big)$. Here P_A, S_C, S_T, P_{Ch} denote the number of poly(A) sites matched from PolyA_DB, the number control samples, the number of treatment samples, and the number of probes in the arraychip, respectively. Moreover, $prox(P_i)$ and $dist(P_i)$ are the number of proximal and distal probes in the Probeset$_i$.

3.1.3 Computing Probe-Level Differential Expression

Proximal and distal probe intensities of both control and treatment samples are read from the intensity file and processed gene by gene. Since there may be some samples which are inconsistent with the remainder of the set, samples with intensity values which have higher deviations from the majority of group are identified and discarded. There are two criteria to find the outlier samples in a group. Firstly, inconsistent probes are found by Iglewicz and Hoaglin's median base outlier detection method as mentioned in Section 1.2.11.3. After computing two average values (i.e. proximal and distal) for each sample, proximal to distal ratio is computed. ratio_{c_i} as the ratio avg_{c_ip} / avg_{c_id}. At the end of this process, there are q different ratio values for control samples and r different ratio values for treatment samples. Median values of these two sets of ratios are calculated as median(ratio_c) and median(ratio_t). Median absolute deviation for each control and treatment group is found separately as the median of the absolute differences of individual ratios from median(ratio_c) and median(ratio_t) as mentioned in Section 1.2.11.2. The modified z-score of a control sample is then computed as mentioned in Section 1.2.11.3.

The z-score of a treatment sample is computed similarly by using the median and the MAD of the treatment samples. Probes with absolute values of z-score greater than a user specific threshold (default value for this threshold is 3.5) are identified as outliers. If only one probe is not marked as outlier or none of the probes are labeled as outlier, that sample is removed from subsequent analyses.

Secondly, for a sample that is not outlier, it is checked that whether the difference between average distal intensities and average proximal intensities is larger than %20 per cent of the average proximal intensities. If it is, then that sample is labeled as an outlier and removed when computing the overall averages. This filter has a biological motivation that we expect differential expression by means of higher proximal expression and lower distal expression; lower proximal expression and higher distal expression does not make sense biologically.

After discarding the outlier samples, if at least one sample remains in both groups, average proximal to distal ratio for treatment samples (RT), average(ratio_{t_i} = avg_{t_ip} / avg_{t_id}) and control samples (RC), average(ratio_{c_i} = avg_{c_ip} / avg_{c_id}). Proportion of these ratios (RT/RC) are calculated as

$$\frac{average\,(ratio_{t_i} = avg_{t_ip}/avg_{t_id})}{average\,(ratio_{c_i} = avg_{c_ip}/avg_{c_id})} \tag{3.1}$$

Furthermore, median value of proximal to distal ratio for treatment and control and proportion of these two values are found. The transcripts are sorted in descending order of this final ratio in the results file. The computational complexity of this procedure is $\mathbf{O(P}_A\mathbf{)}$ * $\Big(\mathbf{O(S}_C\mathbf{)}$ * $\mathbf{O}\Big(\mathbf{prox(P}_i\mathbf{)} + \mathbf{dist(P}_i\mathbf{)}\Big)$ * $\mathbf{log}\Big(\mathbf{prox(P}_i\mathbf{)} + \mathbf{dist(P}_i\mathbf{)}\Big)$ +

$\mathbf{O(S_T)\ *\ O\ \Big(prox(P_i)\ +\ dist(P_i)\Big)\ *\ log\Big(prox(P_i)\ +\ dist(P_i)\Big)\Big)}$. Here P_A, S_C and S_T represent the number of poly(A) events matched from PolyA_DB, the number control samples, the number of treatment samples, respectively. Moreover, $prox(P_i)$ and $dist(P_i)$ are the number of proximal and distal probes in the Probeset$_i$.

The methods described in this section and previous section run totally about 10-20 seconds per sample for Affymetrix Human Genome U133 Plus 2.0 platfrom. For Affymetrix Human Genome U133A, it takes almost half of the runtime for Affymetrix HG-U133 Plus 2.0 to complete.

3.1.4 Analyzing a Single Gene

Another facility of the tool is gene based analysis. For each sample, this analysis gives detailed information about numbers of proximal and distal probes, individual proximal and distal ratios, proximal to distal ratio, median values of both proximal and distal groups with their deviation.

Proximal and distal probe intensities of both control and treatment samples are read from the intensity file which is created processed gene by gene. A similar analysis described in previous section is applied. However, a sample which has a higher difference between its average distal intensities and its average proximal intensities than %70 per cent of the average proximal intensities is marked as inconsistent.

After discarding the outlier samples, if at least one sample remains in both proximal and distal sets, average intensities, median and standard deviation of both groups are computed.

3.2 Description of the Tool Windows

3.2.1 Main Panel

Main Panel which is represented in Figure 3.1 is the primary window of the application. User is able to select the platform of the microarray samples from Annotation Box. Analysis can only be carried on for samples that belong to the same platform. HG-U133A and HG-U133 Plus 2.0. are the offered platforms.

Users can add samples either by giving a valid series ID (i.e. GSE number) or sample number (i.e. GSM number). There are two alternatives to select samples by their GSM numbers. Firstly, GSM IDs which are separated by commas are typed into the text box. After typing all samples, user clicks the “Add Sample(s) to Sample List” button. Secondly, by clicking "Select Sample List From File" button, a file dialog occurs and the file that includes the sample IDs is selected. A critical point is that file should be

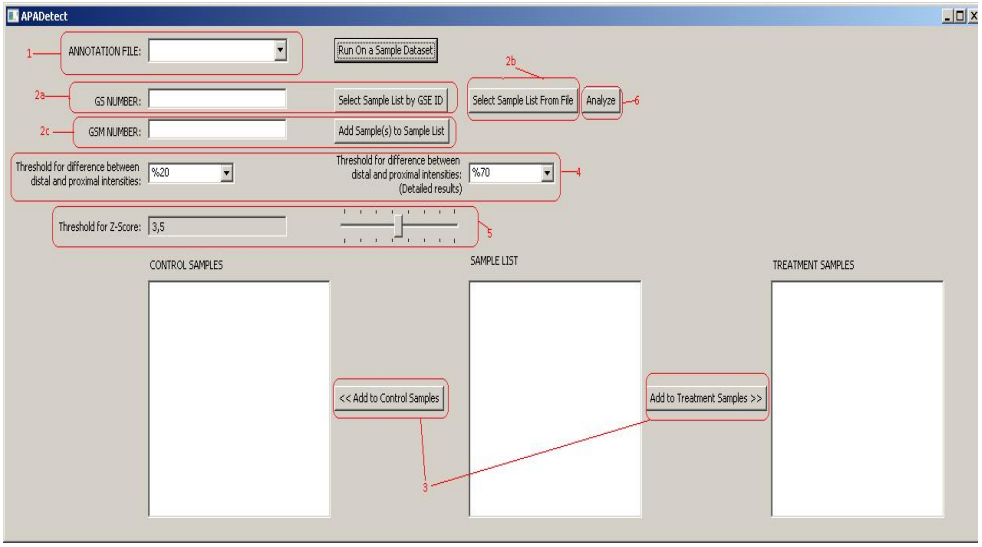


Figure 3.1: Main panel of APADetect

either a text file with .txt extension or an Excel file with .xls/.xlsx extension. Each line/row should contain only one sample number.

Threshold value for difference between distal and proximal intensities can also be decided by user interface. Distal intensities can be at most %5, %20, %50 or %70 higher than proximal intensities. This threshold can be different for analysis and gene based detailed report. A further threshold is to detect the outlier samples by applying Iglewicz and Hoaglin's median based outlier detection method in Section 1.2.11.3. Range of the value is from 0.1 to 7.0.

When candidate samples for analysis are added to sample list, three lists become visible. List boxes that stand left hand side and right hand side identifies the control and treatment samples respectively. In the middle, the main list contains the samples which are not marked as neither control nor treatment. To assign a sample to a group, its GSM number is selected from the sample list. After that, "Add to Control Samples" button or "Add to Treatment Samples" button is clicked. Platform of the sample is compared with the selected platform type. Unless it matches with the selected platform, a message box warns the user about the platform type mismatch. In order to carry out an analysis, both control and treatment groups have to include at least one sample. If analysis terminates successfully, each list box is cleared and prepared for the next analysis.

If a group of samples which are elements of a specific series is analyzed, picking by GS number is the best option. GEO accession of the series is typed into the GS Number textbox and procedure is initiated. Firstly, application looks for the file, GSE_NUMBER.txt, locally under a directory is specified by us. Unless the file is

found, it is downloaded from NCBI GEO website. URL for a series is **http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE_NUMBER&targ=self&form=text&view=quick**

GSE_NUMBER denotes the given GSE ID. Parameter targ=self gets the relevant data for the record such as its title, its platform ID, platform of the organism and IDs of the samples which are the members of the series. Extracting this data from a text file would be more effortless, so parameter from=text was chosen. Data table of the series was not necessary at this stage, so obtaining the quick version of file would be enough. If the platform ID of the series matches with the selected annotation file, GEO accession of the samples are represented in the Sample Group box. A sample in the Sample Group box can be classified as either a control or a treatment sample. When a sample is put in a group APADetect automatically searches the information file, GSM_NUMBER.txt, locally under a pre-defined directory. Unless the file is found, it is downloaded from NCBI GEO website. Download address for a sample information file is **http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM_NUMBER&targ=self&form=text&view=quick**

GSM_NUMBER indicates the given GSM ID. Query parameter targ=self is preferred to obtain the relevant info for the entry such as its platform ID, download URL of the supplementary file. These data is read rapidly and easily from an ASCII file, so from=text is decided as the file format.

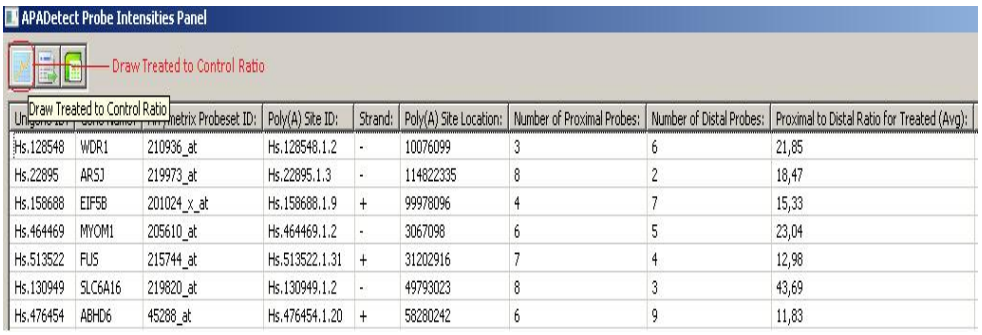
3.2.2 Probe Intensities Panel

This panel pops up, if no error occurs in the analysis procedure. This screen contains three menus with images at the top-left side and a table that visualizes the result. The table consists of 14 main columns and several additional columns which are adjusted by the number of probes in the probe set. Main columns give information about Unigene ID, gene name, Affymetrix probe set ID, poly(A) site ID, strand, poly(A) site starting position, number of proximal probes, number of distal probes, average proximal to distal ratio for treated samples, average proximal to distal ratio for control samples, treated to control ratio (RT/RC), median proximal to distal ratio for treated samples, median proximal to distal ratio for control samples, and median treated to control ratio. Starting locations of probes on target sequence and raw probe intensities of each sample are written into the additional columns.

Leftmost menu helps user to visualize proximal to distal ratio for both control and treatment samples by plotting the scatter chart. The graph panel includes many opportunities such as zooming in a certain area and saving the chart as an image.

The menu in the middle lets user to carry on a detailed intensity analysis which includes the computation of average intensities, median values, standard deviations of median

value for both distal and proximal groups, on a single gene or multiple genes. Genes that are going to be analyzed are selected by clicking on the names. Rightmost menu prepares a Microsoft Excel 2000 or Microsoft Excel 2007 version of the result table. Each row in the document corresponds with the exact counterpart in the table. The menus and an example of the result table is given in Figure 3.2.



Un	Gene Name	Affymetrix Probeset ID:	Poly(A) Site ID:	Strand:	Poly(A) Site Location:	Number of Proximal Probes:	Number of Distal Probes:	Proximal to Distal Ratio for Treated (Avg):
Hs.128548	WDR1	210936_at	Hs.128548.1.2	-	10076099	3	6	21,85
Hs.22895	ARSJ	219973_at	Hs.22895.1.3	-	114822335	8	2	18,47
Hs.158688	EIF5B	201024_x_at	Hs.158688.1.9	+	99978096	4	7	15,33
Hs.464469	MYOM1	205610_at	Hs.464469.1.2	-	3067098	6	5	23,04
Hs.513522	FUS	215744_at	Hs.513522.1.31	+	31202916	7	4	12,98
Hs.130949	SLC6A16	219820_at	Hs.130949.1.2	-	49793023	8	3	43,69
Hs.476454	ABHD6	45288_at	Hs.476454.1.20	+	58280242	6	9	11,83

Figure 3.2: Drawing scatter chart

3.2.3 Scatter Chart Panel

On this window shown in Figure 3.3, the graphical representation of proximal to distal ratio for treated samples (RP) versus proximal to distal ratio for control samples (RN) is shown for the current analysis. Particular data can be observed by either zooming into the related area or getting the vertical/horizontal cross-section. In addition, chart area can be dragged towards any direction. Furthermore, the snapshot of the chart can be saved into the computer as an image.

3.3 User Guide

As it is described in previous section, our tool has several functions. In this section, we explain the instruction steps that user follows for each facility.

3.3.1 Analyzing distal and proximal probe sequences in the control and treatment samples

As shown in Figure 3.1, to conduct an analysis

1. Select the type of the Affymetrix chip(U133 or U133 plus 2.0) from ANNOTATION FILE combo box.

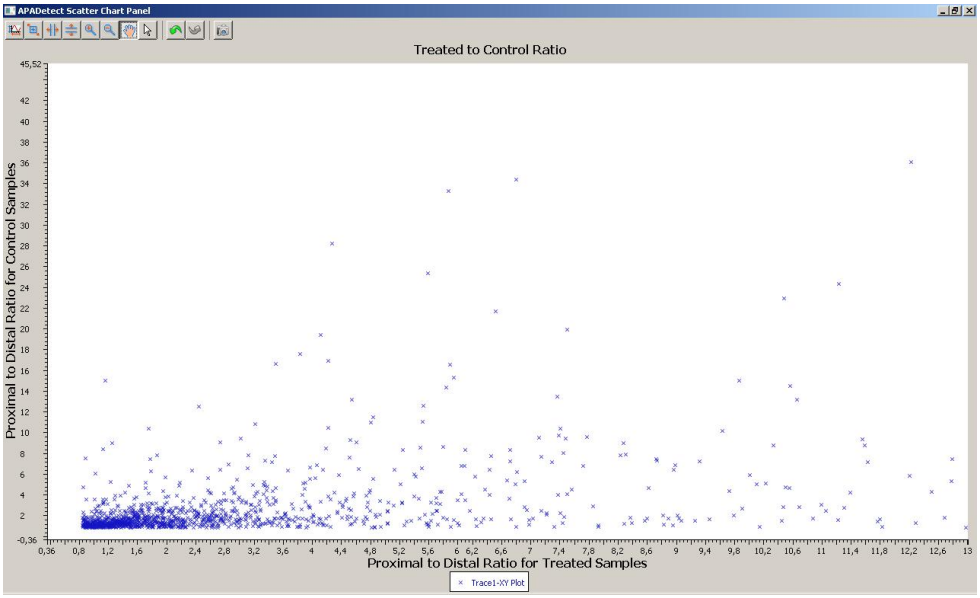


Figure 3.3: A custom scatter chart panel

2. Generate the sample list for control and treatment groups. There are three different ways:
   - User may type the GS identifier into the GS NUMBER text box. There must be only one identifier and its format can be either GSEXYZ or XYZ which XYZ denotes the serial number of the GSE series.
   - A local text or Excel file can be selected from the user's computer. For the text file, each line should start with the sample identifier which is going to be added to list. For the Excel document, each sample identifier should be written to the first cell of each row.
   - The last way is to type the sample numbers separating with comma(,).
3. Divide the samples in the sample list into the two groups: Control and Treatment.
4. Determine the threshold value for difference between distal and proximal intensities.
5. Pick the threshold value for Iglewicz and Hoaglin's median based outlier detection.
6. Click the 'Analyze' button to start the procedure after grouping the samples.

For each sample file, application looks at a certain directory whether this file exists or not. If that file is not found, then it is downloaded from NCBI GEO website, unzipped and converted into ASCII format which makes easier to read and process the information.

After the analysis is completed, a new panel is opened. Result is represented on a table, in that panel. Each row in the table has the following columns: Unigene identifier, gene name, Affymetrix probe set name, poly(A) site name, strand, poly(A) site location on target sequence, number of proximal probes, number of distal probes, proximal to distal ratio for treated samples (RP), proximal to distal ratio for control samples (RN), treated to control ratio (RP/RN), positions and average intensities(for control and treatment samples) of both proximal & distal probes. On the left of the panel, there is a toolbar menu which includes three items.

3.3.2 Drawing scatter plot of RP to RN

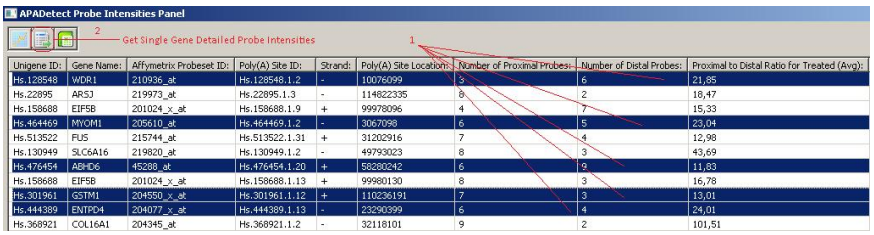
To draw the scatter plot of proximal to distal ratio for treated samples (RP) versus proximal to distal ratio for control samples (RN), user clicks to 'Draw Treated to Control Ratio' menu item which is the leftmost toolbar item. The position of the menu is pointed in Figure 3.2. An example of the chart screen is given in Figure 3.3.

3.3.3 Calculating the average intensities of distal/proximal probes in a gene

To obtain the average intensities of distal and proximal probe sequences in a single gene:

1. Select the gene(s) from the table.
2. Click to the 'Get Single Gene Detailed Probe Intensities' menu item which is in the middle of the toolbar.

For each selected gene, a text file that includes calculated proximal and distal probe set ratios of this gene is opened. Figure 3.4 explains the steps for gene based analysis.



Unigene ID:	Gene Name:	Affymetrix Probeset ID:	Poly(A) Site ID:	Strand:	Poly(A) Site Location:	Number of Proximal Probes:	Number of Distal Probes:	Proximal to Distal Ratio for Treated (Avg):
Hs.128548	WDR1	210936_at	Hs.128548.1.2	-	10076099	3	6	21,85
Hs.22895	ARSJ	219973_at	Hs.22895.1.3	-	114822335	8	2	18,47
Hs.158688	EIF5B	201024_x_at	Hs.158688.1.9	+	99978096	4	7	15,33
Hs.464469	MYOM1	205610_at	Hs.464469.1.2	-	3067098	6	5	23,04
Hs.513522	FUS	215744_at	Hs.513522.1.31	+	31202916	7	4	12,98
Hs.130949	SLC6A16	219820_at	Hs.130949.1.2	-	49793023	8	3	43,69
Hs.476454	ABHD6	45288_at	Hs.476454.1.20	+	58280242	6	9	11,83
Hs.158688	EIF5B	201024_x_at	Hs.158688.1.13	+	99980130	8	3	16,78
Hs.301961	GSTM1	204550_x_at	Hs.301961.1.12	+	110236191	7	3	13,01
Hs.444389	ENTPD4	204077_x_at	Hs.444389.1.13	-	23290399	6	4	24,01
Hs.368921	COL16A1	204345_at	Hs.368921.1.2	-	32118101	9	2	101,51

Figure 3.4: Steps to conduct a detailed gene analysis

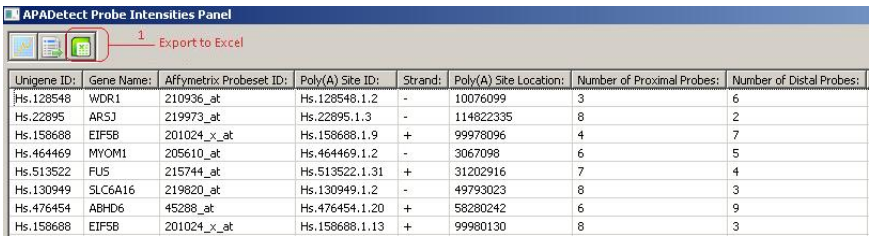
3.3.4 Exporting the results into an Excel document

Analysis result can be kept in an Excel document by following these steps:

1. Click to the 'Export to Excel' button which is the rightmost toolbar item.

2. Select the file and then click to 'OK'.

Figure 3.5 and Figure 3.6 describe the steps of saving the result as an Excel file.



Unigene ID:	Gene Name:	Affymetrix Probeset ID:	Poly(A) Site ID:	Strand:	Poly(A) Site Location:	Number of Proximal Probes:	Number of Distal Probes:
Hs.128548	WDR1	210936_at	Hs.128548.1.2	-	10076099	3	6
Hs.22895	ARSJ	219973_at	Hs.22895.1.3	-	114822335	8	2
Hs.158688	EIF5B	201024_x_at	Hs.158688.1.9	+	99978096	4	7
Hs.464469	MYOM1	205610_at	Hs.464469.1.2	-	3067098	6	5
Hs.513522	FUS	215744_at	Hs.513522.1.31	+	31202916	7	4
Hs.130949	SLC6A16	219820_at	Hs.130949.1.2	-	49793023	8	3
Hs.476454	ABHD6	45288_at	Hs.476454.1.20	+	58280242	6	9
Hs.158688	EIF5B	201024_x_at	Hs.158688.1.13	+	99980130	8	3

Figure 3.5: The location of Export to Excel Menu on Probe Intensities Panel

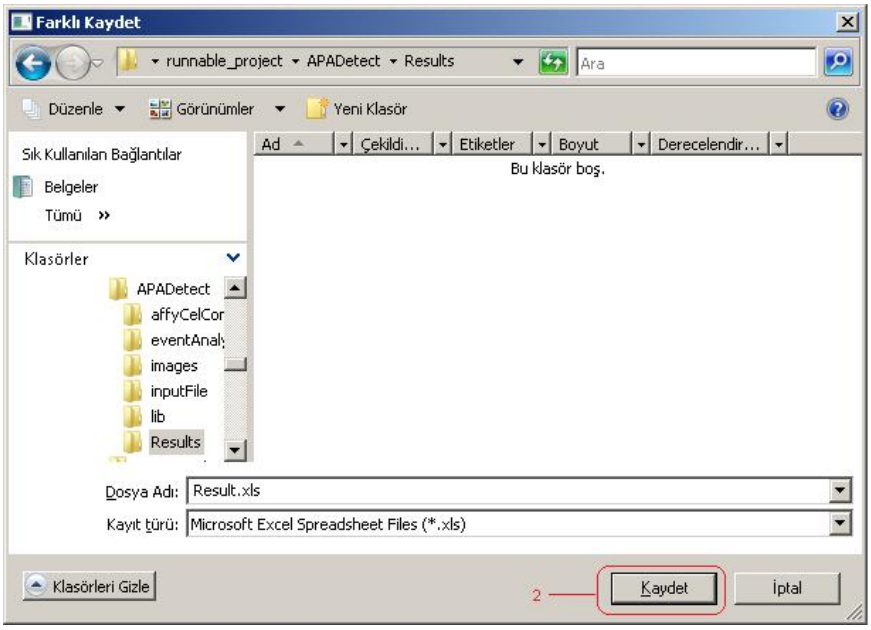


Figure 3.6: Saving as an Excel Document

3.3.5 Running Demo on Sample Dataset

A demonstration of the six samples is also available. By clicking 'Run On a Sample Data' button, the application automatically selects the Affymetrix chip type, determines the sample list via GS NUMBER and GSM NUMBER. After grouping the samples, the analysis is started.

CHAPTER 4

RESULTS

In this chapter, we give information about the experiments conducted on several datasets and discuss the results. The technique mentioned in previous chapter is applied to sixteen GEO series from five different cell lines. Our method has three versions which have different outlier detection algorithms. In Model 1, the Iglewicz and Hoaglin's outlier detection method is applied on each possible distal and proximal sub groups. In Model 2, we consider the whole probe set as one group rather than distal and proximal sub groups and label the highly deviated samples. In Model 3, we compute the average intensities of proximal and distal probe groups sample by sample, then apply the Iglewicz and Hoaglin's method to these averages. These results point to the incompleteness of PolyA_DB: Several poly(A) events are currently unreported in PolyA_DB.

Model 1, Model 2 and Model 3 consume nearly equal time, however Model 2 usually runs slightly longer then the Model 1. Runtime of the models increases proportionally with the number of probe sets in the chip platform. While, on Affymetrix Human Genome U133A platform, the execution time is almost same as the execution time on Affymetrix Human Genome U133B, elapsed time for Affymetrix Human Genome U133 Plus 2.0 is longer than those two platforms. For Affymetrix HG-U133A and Affymetrix HG-U133B, models run approximately 2-3 minutes per sample, however for Affymetrix HG-U133 Plus 2.0 it takes 5-6 minutes to end. Moreover, Model 2 finds more split groups than Model 1. We could not say anything about Model 3, because the trend of its results differ even within a GEO series. When, it yields the highest number of proximal/ distal probe set pairs for a group, it may give the lowest number of proximal/distal probe set groups for another group. In Affymetrix HG-U133 Plus 2.0, Split groupings found by the three methods match between %60-%80. Moreover, in Affymetrix HG-U133A and Affymetrix HG-U133B platforms, %70-%90 of distal/proximal probe set groups appear in the results of the three methods. This difference may be caused by having more than two-fold probe sets than Affymetrix HG-133A and Affymetrix HG-133B. Although no model dominates the each dataset in terms of robustness, Model 1 gives the most credible results in 71 percent of the sample groups. Afterwards, Model 3 comes with %28 of the datasets. Finally, in one

group (%1), Model 2 yields the most trustworthy result. We compute the robustness score as the rate of the matched poly(A) events from the four poly(A) databases over the number of identified proximal-distal probe groups.

4.1 Datasets

Poly(A) events from three different papers are examined. Because of the mismatch between their microarray platforms and ours, we are unable to use all of their datasets. However, similar GEO series prepared from the cell lines that are experimented in these papers are obtained from NCBI GEO [3].

The study by Yoon *et al.* [11] searches the effects of alternative polyadenylation on gene expression and focuses on the positioning of $3'$ end of mRNA. It is suggested that examining $3'$ end leads the way of detecting alternative polyadenylation. This study reports 11455 poly(A) sites[1] and there are 2152 poly(A) sites which are not submitted to PolyA_DB. Target positions are mapped from hg18 to hg19 by using LiftOver[2] tool. They work on immune cells obtained from people who are unrelated to each other. The GEO Accession ID of the data is GSE33154, however the platform (Illumina Genome Analyzer II) of the data does not match with ours. Instead of that series, immune cell samples with platforms U133A, U133B and U133 Plus 2.0 are used. GEO Accession IDs of these series are GSE11058, GSE22356, GSE22886, GSE28490, GSE28491 and GSE43177. Platform type of GEO series except GSE22886 is Affymetrix Human Genome U133 Plus 2.0. Samples in GSE22886 belong to either the Affymetrix Human Genome U133A or the Affymetrix Human Genome U133B.

A method developed by Fu *et al.* [5] detects new poly(A) sites by extracting the sequence information of $3'$-UTR. In addition to finding poly(A) sites, this method identifies the switching events in APA sites. Furthermore, Fu *et al.* [5] compare $3'$-UTR lengths of two breast cancer cell lines, MCF7, MB231, and a normal mammary epithelial cell, MCF10A. They also examine switching of APA sites in these three cell lines. 11478 poly(A) sites[3] are reported in this paper and 1946 of them are different from the ones in PolyA_DB. Since the platforms of the data used in this study does not comply with ours, we were unable to use the dataset. As the MCF7 cell line dataset, GSE10890 and GSE48433 (GSM1178359, GSM1178360, GSM1178361, GSM1178362, GSM1178363, GSM1178364, GSM1178365, GSM1178366, GSM1178367, GSM1178368, GSM1178369, GSM1178370, GSM1178371, GSM1178372, GSM117873, GSM117874) are used. GSE7307 (GSM175968, GSM175969, GSM175970, GSM175971, GSM175972) and GSE21837 comprise the MB231 cell line data group. Each dataset consists of the samples that are elements of Affymetrix Human Genome U133 Plus 2.0 Array.

[1] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3420953/bin/pgen.1002882.s007.xlsx
[2] http://genome.ucsc.edu/cgi-bin/hgLiftOver
[3] http://genome.cshlp.org/content/suppl/2011/03/09/gr.115295.110.DC1/Supplementary_Table_8.xls

The study made by Lin *et al.* [7] is based on investigating the poly(A) sites and different forms of 3′-UTR. They also create a poly(A) map of the human genome. Genomic features of poly(A) sites and poly(A) patterns of genes are evaluated. 105 of 117 poly(A) sites[4] which are reported by them are included in PolyA_DB. In our study, we use four (MCF7, HeLa, HepG2 and K562) of six cancer cell lines that are analyzed by Lin *et al.* [7] In addition to MCF7 dataset mentioned above, GSE2735, GSE32108 and GSE33050 series are used as HeLa data. HepG2 data consists of GSE6494, GSE6869, GSE12939, and GSE30240. As K562 cell line data, GSE43998, GSE1922, and GSE12056 are experimented. Affymetrix platform version of GSE12939 and GSE1922 is Affymetrix Human Genome U133A. Microarray samples are created as Affymetrix Human Genome U133 Plus 2.0 Array.

When we determine the sample groups, we take into consideration the GEO sample information. Each outlier detection method is separately run onto every group. Inconsistency check parameter is set to 0.2 and 0.7 for each run. Split probe sets whose adjusted p-values are less than 0.05 are identified as valid probe sets and are reported in result. Target positions within a probe set are sorted from the upstream to the downstream. We only take the split probe sets which are found in each of six experiment results within a group. Cutting coordinates between the proximal-distal probe sets are looked into both PolyA_DB and poly(A) site document of the paper that uses the related cell line. If strands of both the probe set and the poly(A) site are positive, two conditions are checked for splitting by a poly(A) site. Firstly, the poly(A) site should be closer to the downstream than the final target position of the last proximal probe. Furthermore, the poly(A) site should be located before the starting target position of the first distal probe. If both the probe set and the poly(A) site have negative strands, then last target position of the last distal probe should come before the poly(A) site. In addition to this rule, the poly(A) site should be farther away from 3′-end than the starting location of the first proximal probe.

GEO series in each dataset were examined in different subgroups. These subgroups were determined depending on the cell types in each series. Names of microarray samples and the list of split probe sets for each sample group are given in the Appendix A.

4.2 Results

4.2.1 Results for the MCF7 Dataset

We applied our technique onto GSE10890 and GSE48433. 130 putative poly(A) sites are found in different genes for sixteen groups of samples. Detailed information can

[4] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3458571/bin/supp_gks637_nar-02798-f-2011-File003.xls

be obtained from Section 4.3.

MCF7 samples in GSE10890 were analyzed in two different groups. Each group was divided into six subgroups. For neither the first group nor the second group, poly(A) sites reported by Fu *et al.* [5] and Lin *et al.* [7] were found. In the first group, 13 poly(A) sites in Group 1, 9 poly(A) sites in Group 2, 10 poly(A) sites in Group 3, 4 poly(A) sites in Group 4, 7 poly(A) sites in Group 5 and 8 poly(A) sites in Group 6 are detected. Two poly(A) sites appeared in all subgroups. The list of identified poly(A) sites in the second group was 13 poly(A) sites in Group 1, 15 poly(A) sites in Group 2, 7 poly(A) sites in Group 3, 5 poly(A) sites in Group 4, 6 poly(A) sites in Group 5 and 7 poly(A) sites in Group 6. There is no common poly(A) sites between these subgroups.

MCF7 samples in GSE48433 were split into four groups. In both Group 1 and Group 2, 5 poly(A) sites were found. There are 10 poly(A) sites in Group 3 while the number of poly(A) sites found in subgroup Group 4 is equal to six. None of the poly(A) sites occured within all subgroups. Therefore, the identified poly(A) events may vary within a group of samples or between sample groups.

4.2.2 Results for the MB231 Dataset

The MB321 dataset consists of samples from two GEO series, GSE7307 and GSE21834. We found poly(A) sites from PolyA_DB in each series. Although, all poly(A) sites identified by our method were from PolyA_DB in GSE7307, we also detected some poly(A) sites which are submitted by Fu *et al.* [5] in GSE21834. Nevertheless, those poly(A) sites are also included in PolyA_DB. 36 poly(A) sites were found from two poly(A) event databases. Names and positions of these poly(A) sites are included in the Section 4.3.

Samples which belong to GSE7307 were partitioned into the two groups. 7 poly(A) sites were detected in Group 1. 2 of them also occured in Group 2. There are three subgroups for GSE21834. In the first group, we identified 3 poly(A) sites and they were submitted in PolyA_DB. We found 16 poly(A) sites from PolyA_DB and 8 poly(A) sites from study by Fu *et al.* [5] in the second group. In the last group, one of the 8 poly(A) sites was reported by Fu *et al.* [5] These groups commonly had two poly(A) sites. However, different poly(A) events indicate that a particular poly(A) site is possibly remained undetected in a sample group while it can be detected in another sample group.

4.2.3 Results for the HeLa Dataset

There are samples from GSE2735, GSE32108 and GSE33051 in HeLa dataset. All poly(A) sites detected by our method occured in PolyA_DB. We found 69 putative poly(A) sites for nine sample groups. Detailed results are given in the Section 4.3.

There was only one sample group for GSE2735 and 21 poly(A) sites are found. Samples from GSE32108 were examined in five groups. There were 5 poly(A) sites in both Group 1 and Group 5. We identified 8 poly(A) sites in Group 2 and Group 4. In Group 3, three poly(A) sites were detected. Only one poly(A) appeared in each of these five groups. Sample groups in GSE33051 consists of three groups. There were 8 poly(A) sites, 5 poly(A) sites and 6 poly(A) sites in Group 1, Group 2 and Group 3, respectively. Although two poly(A) sites are common in three groups, observed poly(A) events may differ from a sample to another.

4.2.4 Results for the HepG2 Dataset

The HepG2 cell line dataset consists of samples from three GEO series GSE6878, GSE12939 and GSE30240. Totally, we had 19 sample groups and 156 putative poly(A) sites. We only found the poly(A) sites that are submitted in PolyA_DB. Identified poly(A) sites and the list of split probe set groups are included in the Section 4.3.

GSE6878 was analyzed in 10 sample groups. In Group 1, 8 poly(A) sites were detected. 12 poly(A) sites were found in Group 2. In Group 3, 14 poly(A) sites were identified. Number of poly(A) sites in Group 4 and Group 5 are 8 and 9, respectively. 11 poly(A) sites were found in both Group 6 and Group 8. There were 12 poly(A) sites, 8 poly(A) sites and 10 poly(A) sites in Group 7, in Group 9 and in Group 10, respectively. One poly(A) site appeared in all ten groups.

Four groups were formed in GSE12939. Group 1 and Group 4 had one poly(A) site. This site was also detected in other two groups. There were two more poly(A) sites and one more poly(A) site in Group 2 and Group 3, respectively.

We partitioned the samples in GSE30240 into five groups. Number of poly(A) sites in these groups were 4, 5, 2, 6 and 8. One poly(A) site was occurred in all groups.

By regarding the identified poly(A) sites, we infer that different poly(A) sites can be found under different conditions or samples.

4.2.5 Results for K562 Dataset

The K562 dataset consists of samples from three GEO series, GSE1922, GSE12056 and GSE43998. In three series, all poly(A) sites that were found by us are only from

PolyA_DB. GSE1922 has 16 groups. No poly(A) sites were detected in Group 9, in Group 11 and in Group 15. There were 4 poly(A) sites in Group 1, 3 poly(A) sites in Group 2, 3 poly(A) sites in Group 3 and 5 poly(A) sites in Group 4. We found 1 poly(A) site, 4 poly(A) sites, 2 poly(A) sites and 5 poly(A) sites in groups 5, 6, 7 and 8, respectively. Number of poly(A) sites in groups 10, 12, 13, 14 and 16 were 4, 1, 3, 4, 2, respectively. There was not any poly(A) site which commonly occurs in these groups. Samples in GSE12056 were studied in two groups. The first group had 22 poly(A) sites and the second group includes 7 poly(A) sites. 6 poly(A) sites are appeared in both groups. Samples in GSE43998 were split into two groups. There are 7 poly(A) sites in Group 1 and 10 poly(A) sites in Group 2. Six of the 17 detected poly(A) sites were common in both groups. Uncommon poly(A) sites demonstrates the variaty of the poly(A) events under different circumstances. Please refer to the Section 4.3, for detailed information about poly(A) sites.

4.2.6 Result for the Immune Cell Dataset

Six GEO Series were examined. Since GSE22886 contains members of Affymetrix Human Genome U133A and Affymetrix Human Genome U133B arrays, we analyzed the members of each platform separately. There were five poly(A) sites which were not submitted in PolyA_DB, but reported by Yoon *et al.* [11] Because of mapping the coordinates given by Yoon *et al.* [11] from Hg18 to Hg19, shifting in position may occur for the same poly(A) site. In order to handle this situation, we checked the difference between the positions of poly(A) sites which split the same probe set and are the members of two poly(A) list. If the difference is less than 10, then the poly(A) site in PolyA_DB is assumed to be equal with the poly(A) site reported by Yoon *et al.* [11] To give an example, a poly(A) site which partitions the probe set 52285_f_at into two groups from the negative strand is detected and it appears in both PolyA_DB and the list provided by Yoon *et al.* [11] The target location of this poly(A) site is given as 12673095 in PolyA_DB and 12673096 in the other list. Since the difference between these positions is less than 10, we consider these two poly(A) sites as one site. If the position of this poly(A) site had been 12673005, then they would have been thought as different poly(A) sites.

GSE11058 consists of eight groups. We found several poly(A) sites from PolyA_DB and poly(A) events listed by Yoon *et al.* [11] However these poly(A) sites were same as the ones in PolyA_DB. One poly(A) site was appeared in all of these eight groups. 2 of 4 poly(A) sites in Group 1 were appeared in poly(A) data of Yoon *et al.* [11] There were 2 poly(A) sites detected in Group 2 and one of them was also an element of the list reported by Yoon *et al.* [11]. Two of 11 poly(A) sites in Group 3 occurred in both of two databases. The number of poly(A) sites which were detected in Group 4 is 6, one of them was also found by Yoon *et al.* [11]. There were 8 poly(A) sites in Group 5 and Group 6. For both Group 7 and Group 8, we identified 5 poly(A) sites and two

of them were the same in each poly(A) list. One of the determined poly(A) sites was the common element in all groups.

GSE22356 is examined in four groups. We detected poly(A) sites from PolyA_DB as well as from Yoon *et al.* [11]'s list. In group 1, 101 probe set groups were found. In addition, 5 poly(A) sites from PolyA_DB and 5 poly(A) sites from the list given by Yoon *et al.* [11]. In Group 2, 7 poly(A) events matched with PolyA_DB and 5 poly(A) events occurred in the paper. In Group 3, we detected two poly(A) sites from the database of Yoon *et al.* [11] were also included in PolyA_DB. 131 probe set groups complied with 8 poly(A) sites from PolyA_DB and 3 poly(A) sites from the related paper [11]. As Table 4.1 shows, one of the poly(A) sites was only submitted by Yoon *et al.* [11]. It does not appear in PolyA_DB. This poly(A) site is on the negative

Table 4.1: Summary of matched poly(A) sites for GSE22356

Group No	# of proximal-distal probe set groups	# of identified Poly(A) sites occur in PolyA_DB	# of identified Poly(A) sites occur in Yoon et al. study	# of identified Poly(A) sites which only occur in Yoon et al. study
Group 1	101	5	5	**1**
Group 2	124	7	5	**1**
Group 3	175	8	2	0
Group 4	131	8	3	**1**

strand of chr9 and located at 100844072. These groups do not share any common poly(A) sites, that indicates observation of poly(A)sites is not same for every sample and sample group.

Half of the samples in GSE22886 belong to Affymetrix Human Genome U133 and the other half is in Affymetrix Human Genome U133B, so we separated these samples by their platform types into two groups. Each group consists of 13 subgroups. In the first group (Affymetrix Human Genome U133A), there were one poly(A) site from PolyA_DB in each of Group 1, Group 6 and Group 13. For Group 3, Group 5 and Group 8, number of detected poly(A) events that occurred in PolyA_DB was two. Three poly(A) sites were identified in each of Group 4, Group 9, Group 11 and Group 12. In this sample set that belongs to Affymetrix HG-133A, we found one poly(A) site which did not occur PolyA_DB, but in the list prepared by Yoon *et al.* [11]. This poly(A) site was reported at position 45176137 on the positive strand of chr21. Split region belongs to the Probeset ID, 218018_at. There are six proximal probes and four distal probes on the transcript. Figure 4.1 and Figure 4.2[5] give visual explanation of

[5] http://www.ensembl.org/Homo_sapiens/Location/View?db=core;r=21:45176071-45176421

the poly(A) site and probe locations. Each probe is denoted by a rectangle with a color of black, red, yellow or blue. Grey line represents the location of the poly(A) site.

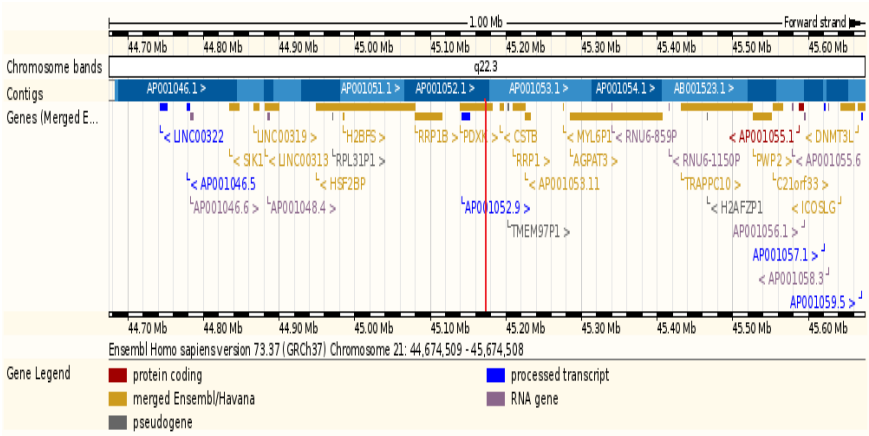


Figure 4.1: Position of probeset 218018_at on the chromosome

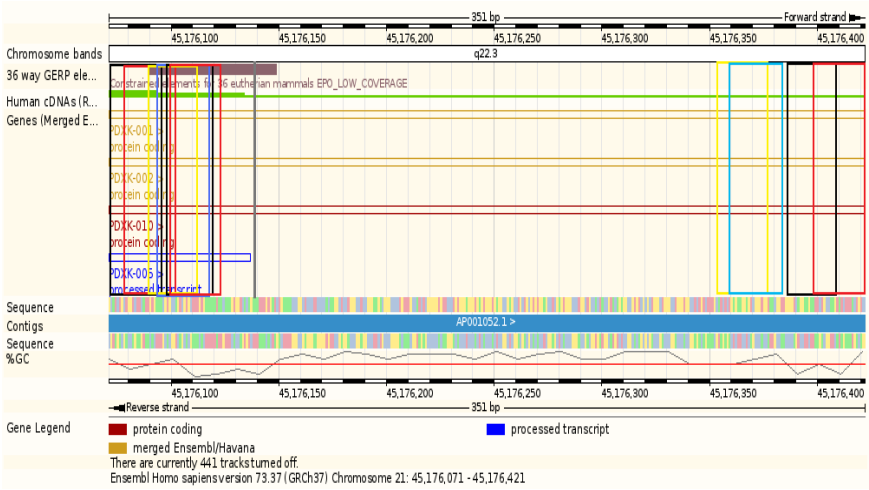


Figure 4.2: Positions of the probes and poly(A) site on chr21

Number of split probe sets and poly(A) sites which partition these probe sets are displayed in Table 4.2.

For the samples which are belong to Affymetrix U133B, our method did not find any poly(A) sites for Group 3, Group 6, Group 9 and Group 11. One poly(A) site was detected for each of other groups.

Experimented samples in GSE28490 are divided into eight groups. There were 6 poly(A) sites, 7 poly(A) sites, 8 poly(A) sites and 2 poly(A) sites from PolyA_DB,

Table 4.2: Summary of matched poly(A) sites for GSE22886 (Affy HG-U133A)

Group No	# of proximal-distal probe set groups	# of identified Poly(A) sites occur in PolyA_DB	# of identified Poly(A) sites occur in Yoon et al. study	# of identified Poly(A) sites which only occur in Yoon et al. study
Group 1	56	1	0	0
Group 2	102	4	3	0
Group 3	91	2	1	0
Group 4	65	3	1	0
Group 5	106	2	2	**1**
Group 6	67	1	1	0
Group 7	68	5	3	0
Group 8	58	2	2	0
Group 9	62	3	0	0
Group 10	33	0	0	0
Group 11	76	3	3	0
Group 12	72	3	3	0
Group 13	51	1	1	0

in Group 1, Group 4, Group 5, Group 7, respectively. In each of Group 2, Group 3, Group 6 and Group 8, three poly(A) events from PolyA_DB were detected. For Group 9, only one poly(A) site was identified from PolyA_DB. In groups 1, 2 and 6, we identified two different poly(A) sites which were reported by Yoon *et al.* [11], but not appeared in PolyA_DB. A poly(A) site which was located on the positive strand of chr4 at 120440802 was occurred in both Group 1 and Group 6. Another poly(A) site whose position was 56506649 on the positive strand of chr12. This poly(A) site is split Probeset 208676_s_at into two groups in which proximal group contains three probes and distal group has 7 probes. Visual analysis of the poly(A) site and probe locations is given in Figure 4.3 and Figure 4.4[6]. Each probe is denoted by a rectangle with a color of black, red, yellow or blue. Position of the poly(A) site is shown by the grey line.

One poly(A) site was the member of every group. This means some poly(A) events are only detected under specific conditions. Table 4.3 indicates the numbers of proximal & distal groups, the poly(A) sites which split these probe sets for each group.

There are seven experiment groups for GSE28491. Although we detected a couple of poly(A) sites from Yoon *et al.* [11] in groups 2, 4 and 7; they were also occurred in

[6] http://www.ensembl.org/Homo_sapiens/Location/View?db=core;r=12:56504775-56506871

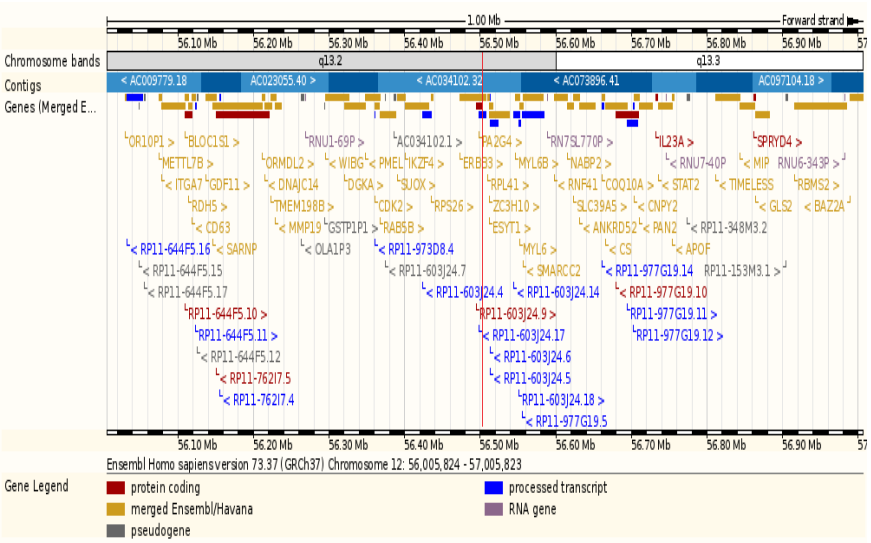


Figure 4.3: Position of probeset 208676_s_at on the chromosome

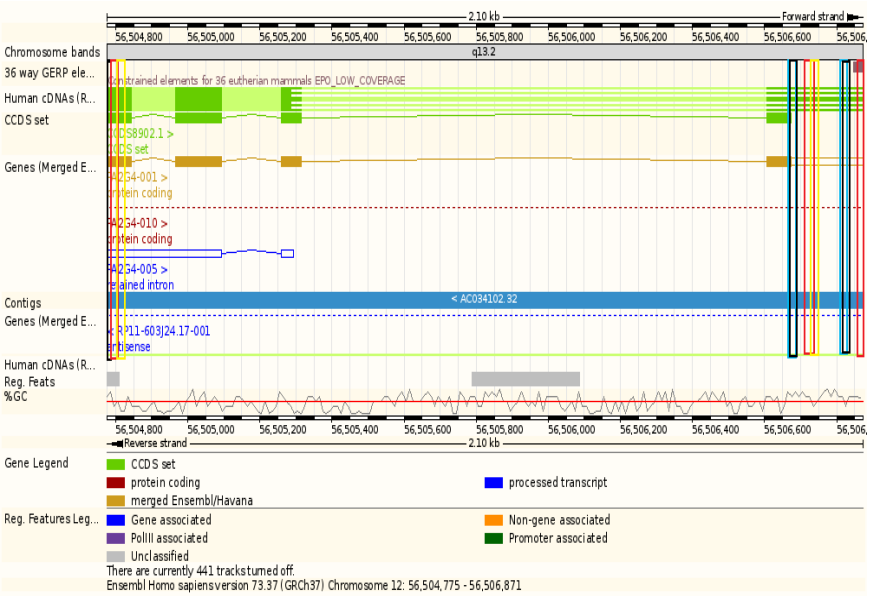


Figure 4.4: Positions of the probes and poly(A) site on chr12

PolyA_DB. We could not find any of the poly(A) sites listed by Yoon *et al.* [11] in Group 1, Group 3, Group 5 and Group 6. We found 5 poly(A) sites in Group 1 and 7 poly(A) sites in Group 2. There were 1 poly(A) site in both Group 3 and Group 5. Number of poly(A) sites in Group 4, Group 6 and Group 7 are 3. These seven groups did not have any common poly(A) site.

Table 4.3: Summary of matched poly(A) sites for GSE28490

Group No	# of proximal-distal probe set groups	# of identified Poly(A) sites occur in PolyA_DB	# of identified Poly(A) sites occur in Yoon et al. study	# of identified Poly(A) sites which only occur in Yoon et al. study
Group 1	125	6	3	**1**
Group 2	131	3	2	**1**
Group 3	95	3	1	0
Group 4	123	7	1	0
Group 5	159	8	2	0
Group 6	169	3	2	**1**
Group 7	45	2	0	0
Group 8	98	3	1	0
Group 9	91	1	0	0

GSE43177 is analyzed in two groups. For Group 1, number of identified poly(A) events from the list by Yoon *et al.* [11] was five, whereas four of them also occurred in PolyA_DB. In group 2, 5 poly(A) events matched with the ones in PolyA_DB and two of them were submitted by Yoon *et al.* [11]. We found one poly(A) site which was only included in the study by Yoon *et al.* [11]. This poly(A) site was appeared on the positive strand of chr2, at 101622571. Three poly(A) sites were common in both groups. Amount of proximal-distal probe sets, and poly(A) sites were given in Table 4.4.

Table 4.4: Summary of matched poly(A) sites for GSE43177

Group No	# of proximal-distal probe set groups	# of identified Poly(A) sites occur in PolyA_DB	# of identified Poly(A) sites occur in Yoon et al. study	# of identified Poly(A) sites which only occur in Yoon et al. study
Group 1	214	7	5	**1**
Group 2	143	5	3	**1**

By regarding the found poly(A) sites in each GEO series, it is shown that different poly(A) sites can be observed under different conditions or samples. Please refer to the Section 4.3, for detailed information about poly(A) sites.

Similar genes are found in different datasets. To illustrate, four variations of 48030_i_at located at chr5 were detected in Immune cell, MCF7, MB231, HeLa and K562 datasets. Moreover, two different split groups of 225367_at were occurred in both MB231 and K562 datasets. We also came across 200654_at in HeLa, HepG2, K562 datasets. Please refer to the Section 4.3, for complete list for common genes in different datasets.

To conclude, we investigated sixteen GEO series in total and compared our results with four poly(A) databases. Because of the microarray platform type mismatch between ours and theirs, we could not experiment their samples. Instead, we looked for the equivalents of those datasets in NCBI GEO [3] and run our method on the samples we found. Although, our results which are obtained from the MCF7, MB231, HeLa, HepG2 and K562 datasets are insufficient to evaluate the extent and adequacy of PolyA_DB, findings of the immune cell dataset indicates that PolyA_DB is not fully comprehensive and several poly(A) sites are currently out of this database. Our results also point to the convenience of the statistical significance between split probe sets as a detecting mechanism for the poly(A) sites. Although three models run for almost equal amount of time, Model 2 has longer elapsed time than Model 1. In addition, run time of these models increases with the number of probe sets in the platform. Model 1 yields the most credible results in the majority of the datasets. Since the observations of poly(A) events differ under different circumstances or samples, we do not detect exactly the same poly(A) sites for sample groups belong to the same GEO series. In order to make clearer statements about the validity and robustness of our technique, other split probe sets which are found by our method should also be analyzed.

4.3 Supplementary Data

We provide the results for each dataset in in a separate supplemental document. This document includes the list of the matched poly(A) sites found by our method described in Chapter 2. Moreover, the identified distal/proximal probe groups are added for each dataset. You may download the supplementary data from `http://www.ceng.metu.edu.tr/~e1560762/Supplementary_Data.docx`.

CHAPTER 5

CONCLUSION

5.1 Conclusion

In this thesis, as an alternative poly(A) site identification technique, we suggested a novel method which does not rely on sequencing studies. This method checks the differences between distal and proximal probes within every differentially expressed split probe set group and selects the ones that remain under a threshold. Initially, all possible split probe sets whose proximal and distal groups consist of at least two probes are formed with the positions on the chip and on gene of each probe. Sequence information was obtained from the UCSC Genome Browser [6]. For each platform, we extracted the location of each probe from the related Affymetrix annotation file. While examining the differential expression of the proximal and distal groups, we used the split probe set file which is prepared for that array platform. Three different outlier detection methods are applied on each sample group. Finally, we tested the findings of these methods by Welch's t-test, adjusted the p-values by Benjamini-Hochberg False Discovery Rate control method and accepted the split probe sets with error rate 0.05. We used *tTest* function in Apache Commons Math[1]. Although detailed wet-lab experiments are required to check the biological validity of our results, our study is a promising way to analyze and detect the alternative polyadenylation sites.

We also introduced a new tool which screens the differentially expressed probe sets that are split by polyadenylation sites. The current version of APADetect is able to analyze and report microarray experiments conducted on Human Genome U133A and Human Genome U133 Plus 2.0 arrays. It also graphically represents the trend of proximal to distal ratios of control and treatment samples in a scatter chart. In addition, APADetect lists detailed information for a specified gene. In order to perform probe level analysis, another requirement is the availability of raw CEL files. APADetect was first used by Akman *et al.* [1] on three separate GEO experiment series conducted in 2007 and 2008. The experiments were conducted on estrogen treated breast cancer cell lines for investigation of different biological aspects. By analyzing these microarray datasets from a different perspective, APADetect was able to provide candidate genes

[1] http://commons.apache.org/proper/commons-math/

that showed selective activation of APA events. The gene CDC6 which consistently appeared in the APADetect outputs of all the three datasets was further validated by wet-lab experiments [1].

5.2 Future Work

First extension is to detect the split regions which possibly point the same poly(A) site. In our study, we found several different split probe set groups that are partitioned by the same site. Detecting these kind of groups helps us to search the poly(A) site within a single range rather than separate ranges. For the proximal/distal probe groups that belong to the same probe set, if either coordinates of the last proximal or coordinates of the first distal probes of these groups are close to each other, then we may obtain the maximum range from the combination of border positions and examine the poly(A) site within that area.

Secondly, we need to do further experiments about the split probe sets given in the Supplementary Material for the possibility of new poly(A) sites. Since PolyA_DB is not a complete database, undiscovered polyadenylation sites may exist.

Thirdly, we will look for the poly(A) signals within the every possible poly(A) site region to be certain about poly(A) event.

Finally, in our tool described in Chapter 3, we only use the technique that leverage the knowledge of the poly(A) site positions. The first technique can be integrated into our tool APADetect as another analysis option.

REFERENCES

[1] B. H. Akman, T. Can, and A. E. Erson-Bensan. Estrogen-induced upregulation and 3′-utr shortening of cdc6. *Nucleic Acids Research*, 40(21):10679–10688, 2012.

[2] H. Auer, D. L. Newsom, and K. Kornacker. Expression profiling using affymetrix genechip microarrays. *Methods in Molecular Biology*, 509:35–46, 2009.

[3] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, and P. M. Sherman. Ncbi geo: archive for functional genomics data sets-10 years on. *Nucleic Acids Research*, 39:1005–1010, 2011.

[4] Y. Cheng, R. M. Miura, and B. Tian. Prediction of mrna polyadenylation sites by support vector machine. *Bioinformatics*, 22:2320–2325, 2006.

[5] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen, and A. Xu. Differential genome-wide profiling of tandem 3′ utrs among human breast cancer and normal cells by high-throughput sequencing. *Genomic Research*, 21:741–747, 2011.

[6] P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B.J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent. The uscs genome browser database: update 2011. *Nucleic Acids Research*, 39(Database issue):D876–D882, 2010.

[7] Y. Lin, Z. Li, F. Ozsolak, S. W. Kim, G. Arango-Argoty, T. T. Liu, S. A. Tenenbaum, T. Bailey, P. Monaghan, P. M. Milos, and B. John. Widespread shortening of 3′utrs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Nucleic Acids Research*, 40(17):8460–8471, 2012.

[8] C. S. Lutz. Alternative polyadenylation: a twist on mrna 3′ end formation. *ACS Chemical Biology*, 3:609–617, 2008.

[9] C. Mayr and D. P. Bartel. Widespread shortening of 3′utrs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138:673–684, 2009.

[10] B. Tian, J. Hu, H. Zhang, and C. S. Lutz. A large-scale analysis of mrna polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201–212, 2005.

[11] O. K. Yoon, T. Y. Hsu, J. H. Im, and R. B. Brem. Genetic and regulatory impact of alternative polyadenylation in human b-lymphoblastoid cells. *PLOS Genetics*, 8(8):e1002882, 2012.

[12] H. Zhang, J. Hu, M. Recce, and B. Tian. Polya_db: a database for mammalian mrna polyadenylation. *Nucleic Acids Research*, 33(Database issue):D116–120, 2005.

APPENDIX A

SAMPLE GROUPS

A.1 The Immune Cell Dataset

A.1.1 GSE11058

Group 1: GSM279592, GSM279593, GSM279594. Group 2: GSM279589, GSM279590, GSM279591. Group 3: GSM279595, GSM279596, GSM279597. Group 4: GSM279598, GSM279599, GSM279600. Group 5: GSM279601, GSM279602, GSM279603. Group 6: GSM279604, GSM279605, GSM279606. Group 7: GSM279607, GSM279608, GSM279609. Group 8: GSM279610, GSM279611, GSM279612.

A.1.2 GSE22356

Group 1: GSM556433, GSM556434, GSM556435, GSM556436, GSM556437, GSM556438, GSM556439, GSM556440. Group 2: GSM556441, GSM556442, GSM556443, GSM556444, GSM556445, GSM556446, GSM556447, GSM556448, GSM556449, GSM556450. Group 3: GSM556413, GSM556414, GSM556415, GSM556416, GSM556417, GSM556418, GSM556419, GSM556420, GSM556421, GSM556422. Group 4: GSM556423, GSM556424, GSM556425, GSM556426, GSM556427, GSM556428, GSM556429, GSM556430, GSM556431, GSM556432.

A.1.3 GSE22886

A.1.3.1 Set 1

Group 1: GSM565308, GSM565309, GSM565310, GSM565311, GSM565312, GSM565313, GSM565314. Group 2: GSM565287, GSM565288, GSM565289, GSM565290, GSM565291, GSM565292. Group 3: GSM565273, GSM565274, GSM565275, GSM565276, GSM565277, GSM565278, GSM565279 ,GSM565280, GSM565281, GSM565282, GSM565283, GSM565284, GSM565285, GSM565286. Group 4: GSM565269, GSM565270, GSM565271, GSM565272. Group 5: GSM565366, GSM565367, GSM565368, GSM565369, GSM565370, GSM565371,

GSM565372, GSM565373, GSM565374, GSM565375, GSM565376, GSM565377. Group 6: GSM565315, GSM565316, GSM565317, GSM565318. Group 7: GSM565319, GSM565320, GSM565321, GSM565322. Group 8: GSM565330, GSM565331, GSM565332, GSM565333, GSM565334, GSM565335, GSM565336, GSM565337, GSM565338, GSM565339, GSM565340, GSM565341, GSM565342, GSM565343, GSM565344, GSM565345, GSM565346, GSM565347, GSM565348, GSM565349, GSM565350, GSM565351, GSM565352, GSM565353, GSM565354, GSM565355, GSM565356, GSM565357, GSM565358, GSM565359, GSM565360, GSM565361, GSM565362, GSM565363, GSM565364, GSM565365. Group 9: GSM565378, GSM565379, GSM565380, GSM565381, GSM565382. Group 10: GSM565293, GSM565294, GSM565295, GSM565296, GSM565297, GSM565298, GSM565299, GSM565300, GSM565301, GSM565302, GSM565303, GSM565304, GSM565305, GSM565306, GSM565307. Group 11: GSM565326, GSM565327, GSM565328, GSM565329. Group 12: GSM565323, GSM565324, GSM565325. Group 13: GSM565269, GSM565270, GSM565271, GSM565272, GSM565273, GSM565274, GSM565275, GSM565276, GSM565277, GSM565278, GSM565279 ,GSM565280, GSM565281, GSM565282, GSM565283, GSM565284, GSM565285, GSM565286, GSM565287, GSM565288, GSM565289, GSM565290, GSM565291, GSM565292, GSM565293, GSM565294, GSM565295, GSM565296, GSM565297, GSM565298, GSM565299, GSM565300, GSM565301, GSM565302, GSM565303, GSM565304, GSM565305, GSM565306, GSM565307, GSM565308, GSM565309, GSM565310, GSM565311, GSM565312, GSM565313, GSM565314, GSM565315, GSM565316, GSM565317, GSM565318, GSM565319, GSM565320, GSM565321, GSM565322, GSM565323, GSM565324, GSM565325, GSM565326, GSM565327, GSM565328, GSM565329, GSM565330, GSM565331, GSM565332, GSM565333, GSM565334, GSM565335, GSM565336, GSM565337, GSM565338, GSM565339, GSM565340, GSM565341, GSM565342, GSM565343, GSM565344, GSM565345, GSM565346, GSM565347, GSM565348, GSM565349, GSM565350, GSM565351, GSM565352, GSM565353, GSM565354, GSM565355, GSM565356, GSM565357, GSM565358, GSM565359, GSM565360, GSM565361, GSM565362, GSM565363, GSM565364, GSM565365, GSM565366, GSM565367, GSM565368, GSM565369, GSM565370, GSM565371, GSM565372, GSM565373, GSM565374, GSM565375, GSM565376, GSM565377, GSM565378, GSM565379, GSM565380, GSM565381, GSM565382.

A.1.3.2 Set 2

Group 1: GSM566025, GSM566026, GSM566027, GSM566028, GSM566029, GSM566030, GSM566031. Group 2: GSM566004, GSM566005, GSM566006, GSM566007, GSM566008, GSM566009. Group 3: GSM565990, GSM565991, GSM565992, GSM565993, GSM565994, GSM565995, GSM565996 ,GSM565997, GSM565998, GSM565999, GSM566000, GSM566001, GSM566002, GSM566003. Group 4: GSM565986, GSM565987, GSM565988, GSM565989. Group 5: GSM566083, GSM566084, GSM566085, GSM566086, GSM566087, GSM566088, GSM566089, GSM566090, GSM566091, GSM566092, GSM566093, GSM566094. Group 6: GSM566032, GSM566033, GSM566034, GSM566035. Group 7: GSM566036, GSM566037, GSM566038, GSM566039. Group 8: GSM566047, GSM566048, GSM566049, GSM566050, GSM566051, GSM566052, GSM566053, GSM566054, GSM566055, GSM566056, GSM566057,

GSM566058, GSM566059, GSM566060, GSM566061, GSM566062, GSM566063, GSM566064, GSM566065, GSM566066, GSM566067, GSM566068, GSM566069, GSM566070, GSM566071, GSM566072, GSM566073, GSM566074, GSM566075, GSM566076, GSM566077, GSM566078, GSM566079, GSM566080, GSM566081, GSM566082. Group 9: GSM566095, GSM566096, GSM566097, GSM566098, GSM566099. Group 10: GSM566010, GSM566011, GSM566012, GSM566013, GSM566014, GSM566015, GSM566016, GSM566017, GSM566018, GSM566019, GSM566020, GSM566021, GSM566022, GSM566023, GSM566024. Group 11: GSM566043, GSM566044, GSM566045, GSM566046. Group 12: GSM566040, GSM566041, GSM566042. Group 13: GSM565986, GSM565987, GSM565988, GSM565989, GSM565990, GSM565991, GSM565992, GSM565993, GSM565994, GSM565995, GSM565996 ,GSM565997, GSM565998, GSM565999, GSM566000, GSM566001, GSM566002, GSM566003, GSM566004, GSM566005, GSM566006, GSM566007, GSM566008, GSM566009, GSM566010, GSM566011, GSM566012, GSM566013, GSM566014, GSM566015, GSM566016, GSM566017, GSM566018, GSM566019, GSM566020, GSM566021, GSM566022, GSM566023, GSM566024, GSM566025, GSM566026, GSM566027, GSM566028, GSM566029, GSM566030, GSM566031, GSM566032, GSM566033, GSM566034, GSM566035, GSM566036, GSM566037, GSM566038, GSM566039, GSM566040, GSM566041, GSM566042, GSM566043, GSM566044, GSM566045, GSM566046, GSM566047, GSM566048, GSM566049, GSM566050, GSM566051, GSM566052, GSM566053, GSM566054, GSM566055, GSM566056, GSM566057, GSM566058, GSM566059, GSM566060, GSM566061, GSM566062, GSM566063, GSM566064, GSM566065, GSM566066, GSM566067, GSM566068, GSM566069, GSM566070, GSM566071, GSM566072, GSM566073, GSM566074, GSM566075, GSM566076, GSM566077, GSM566078, GSM566079, GSM566080, GSM566081, GSM566082, GSM566083, GSM566084, GSM566085, GSM566086, GSM566087, GSM566088, GSM566089, GSM566090, GSM566091, GSM566092, GSM566093, GSM566094, GSM566095, GSM566096, GSM566097, GSM566098, GSM566099

A.1.4 GSE28490

Group 1: GSM705297, GSM705298, GSM705299, GSM705300, GSM705301. Group 2: GSM705302, GSM705303, GSM705304, GSM705305, GSM705306 Group 3: GSM705312, GSM705313, GSM705314, GSM705315, GSM705316. Group 4: GSM705317, GSM705318, GSM705319, GSM705320. Group 5: GSM705321, GSM705322, GSM705323, GSM705324, GSM705325. Group 6: GSM705287, GSM705288, GSM705289, GSM705290, GSM705291, GSM705292, GSM705293, GSM705294, GSM705295, GSM705296. Group 7: GSM705326, GSM705327, GSM705328. Group 8: GSM705307, GSM705308, GSM705309, GSM705310, GSM705311. Group 9: GSM705329, GSM705330, GSM705331, GSM705332, GSM705333.

A.1.5 GSE28491

Group 1: GSM705402, GSM705403, GSM705404, GSM705405, GSM705406. Group 2:

GSM705412, GSM705413, GSM705414, GSM705415, GSM705416. Group 3: GSM705417, GSM705418, GSM705419, GSM705420, GSM705421. Group 4: GSM705422, GSM705423, GSM705424. Group 5: GSM705407, GSM705408, GSM705409, GSM705410, GSM705411. Group 6: GSM705430, GSM705431, GSM705432, GSM705433, GSM705434. Group 7: GSM705425, GSM705426, GSM705427, GSM705428, GSM705429.

A.1.6 GSE43177

Group 1: GSM1057943, GSM1057944, GSM1057945, GSM1057946, GSM1057947, GSM1057948, GSM1057949, GSM1057950, GSM1057951, GSM1057952. Group 2: GSM1057953, GSM1057954, GSM1057955, GSM1057956, GSM1057957, GSM1057958, GSM1057959, GSM1057960, GSM1057961.

A.2 The MCF7 Cell Line Dataset

A.2.1 GSE10890

A.2.1.1 Set 1

Group 1: GSM275978. Group 2: GSM276046, GSM276047, GSM276048. Group 3: GSM276049, GSM276050, GSM276051. Group 4: GSM276052, GSM276053, GSM276054. Group 5: GSM276055, GSM276056, GSM276057. Group 6: GSM276058, GSM276059.

A.2.1.2 Set 2

Group 1: GSM275978. Group 2: GSM276071, GSM276072, GSM276073. Group 3: GSM276074, GSM276075, GSM276076. Group 4: GSM276077, GSM276078, GSM276079. Group 5: GSM276080, GSM276081, GSM276082. Group 6: GSM276083, GSM276084, GSM276085.

A.2.2 GSE48433

Group 1: GSM1178359, GSM1178360. Group 2: GSM1178361, GSM1178362, GSM1178363, GSM1178364. Group 3: GSM1178365, GSM1178366, GSM1178367, GSM1178368, GSM1178369. Group 4: GSM1178370, GSM1178371, GSM1178372, GSM1178373, GSM1178374.

A.3 The MB231 Cell Line Dataset

A.3.1 GSE7307

Group 1: GSM175968, GSM175969, GSM175970. Group 2: GSM175971, GSM175972, GSM175997.

A.3.2 GSE21834

Group 1: GSM543397, GSM543398, GSM543399. Group 2: GSM543400, GSM543401, GSM543402. Group 3: GSM543403, GSM543404, GSM543405.

A.4 The HeLa Cell Line Dataset

A.4.1 GSE2735

Group 1: GSM139730, GSM139731, GSM139732.

A.4.2 GSE32108

Group 1: GSM796020, GSM796021, GSM796022. Group 2: GSM796023, GSM796024, GSM796025. Group 3: GSM796026, GSM796027, GSM796028. Group 4: GSM796029, GSM796030, GSM796031. Group 5: GSM796032, GSM796033, GSM796034.

A.4.3 GSE33051

Group 1: GSM818832, GSM818833, GSM818834 Group 2: GSM818835, GSM818836, GSM818837 Group 3: GSM818838, GSM818839, GSM818840

A.5 The HepG2 Cell Line Dataset

A.5.1 GSE6878

Group 1: GSM149381, GSM149382, GSM149383. Group 2: GSM149384, GSM149385, GSM149386. Group 3: GSM149387, GSM149388. Group 4: GSM149389, GSM149390. Group 5: GSM149391, GSM149392, GSM149393. Group 6: GSM158373, GSM158374, GSM158375. Group 7: GSM158376, GSM158377, GSM158378. Group 8: GSM158382,

GSM158383, GSM158384. Group 9: GSM158379, GSM158380, GSM158381. Group 10: GSM154182, GSM154183, GSM154184.

A.5.2 GSE12939

Group 1: GSM324527, GSM324528, GSM324529, GSM324530, GSM324531, GSM324532.
Group 2: GSM324533, GSM324534, GSM324535, GSM324536, GSM324537, GSM324538.
Group 3: GSM324539, GSM324540, GSM324541, GSM324542, GSM324543, GSM324544.
Group 4: GSM324545, GSM324546, GSM324547, GSM324548, GSM324549, GSM324550.

A.5.3 GSE30240

Group 1: GSM748830, GSM748831, GSM748832. Group 2: GSM748833, GSM748834, GSM748835. Group 3: GSM748836, GSM748837, GSM748838. Group 4: GSM748839, GSM748840, GSM748841. Group 5: GSM748842, GSM748843, GSM748844.

A.6 The K562 Data Cell Dataset

A.6.1 GSE1922

Group 1: GSM34039, GSM34040, GSM34041. Group 2: GSM34042, GSM34043, GSM34044. Group 3: GSM34134, GSM34140, GSM34146. Group 4: GSM34162, GSM34163, GSM34164. Group 5: GSM34909, GSM34910, GSM34911. Group 6: GSM34912, GSM34913, GSM34914. Group 7: GSM34915, GSM34916, GSM34917. Group 8: GSM34918, GSM34919, GSM34920. Group 9: GSM34921, GSM34922, GSM34923. Group 10: GSM34924, GSM34925, GSM34926. Group 11: GSM34927, GSM34928, GSM34929. Group 12: GSM34930, GSM34931, GSM34932. Group 13: GSM34933, GSM34934, GSM34935. Group 14: GSM34936, GSM34937, GSM34938. Group 15: GSM34939, GSM34940, GSM34941. Group 16: GSM34942, GSM34943, GSM34944.

A.6.2 GSE12056

Group 1: GSM304303, GSM304304, GSM304479, GSM304498, GSM304480, GSM304481, GSM304482, GSM304483, GSM304484, GSM304486. Group 2: GSM304487, GSM304488, GSM304489, GSM304490, GSM304491, GSM304492, GSM304493, GSM304494, GSM304495, GSM304496.

A.6.3 GSE43998

Group 1: GSM1076321, GSM1076322, GSM1076323, GSM1076324. Group 2: GSM1076317, GSM1076318, GSM1076319, GSM1076320.