BIOINFORMATICS APPROACH TO CANCER RESESARCH: FROM BIOMARKER DISCOVERY TO GENOME-WIDE STUDIES

by

CELINE SEULGI HONG

(Under the Direction of Ying Xu)

ABSTRACT

The advancements in high-throughput technologies led to overwhelming amounts of biological data. By utilizing the –omics data, we have applied bioinformatics methods and applications to study various aspects of cancer, from biomarker discovery to genome-wide studies. Various techniques and approaches are presented here to address some of challenging questions in cancer. We mainly focus on enhancing biomarker discovery and utilizing –omics data to understand cancer cell behavior.

INDEX WORDS: Gastric Cancer, Stomach Cancer, Biomarker, Urinary Proteins, Urinary Biomarker, miRNA biomarker, Support-Vector Machine, Genome, Transcriptome, Apoptosis

BIOINFORMATICS APPROACH TO CANCER RESESARCH: FROM BIOMARKER DISCOVERY TO GENOME-WIDE STUDIES

by

CELINE SEULGI HONG

B.S., The University of Georgia, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

BIOINFORMATICS APPROACH TO CANCER RESESARCH: FROM BIOMARKER DISCOVERY TO GENOME-WIDE STUDIES

by

CELINE SEULGI HONG

Major Professor: Ying Xu

Committee: Liming Cai

Natarajan Kannan

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

August 2012

TABLE OF CONTENTS

CHAPTER PAGE

1 INTRODUCTION……1

2 A COMPUTATIONAL METHOD FOR PREDICTION OF EXCRETORY PROTEINS AND APPLICATION TO IDENTIFICATION OF GASTRIC CANCER MARKERS IN URINE…….6

Abstract/Introduction……7

Methods……9

Results and Discussion ……16

Figures and Tables……22

3 CHARACTERIZATION OF URINE-EXCRETORY MIRNAS AND DEMONSTRATION OF THEIR FEASIBILITY AS DISEASE DIAGNOSTIC MARKERS……40

Abstract/Introduction……41

Results…… ……43

Discussion……45

Materials and Methods……47

Concluding Remarks……49

Figures and Tables……50

4 GENOME-WIDE TRANSCRIPTION DATA ANALYSIS IN CO-LOCALIZED REGIONS OF BREAST CANCER……56

Background……56

Materials and Methods……57

Results and Discussion……60

Conclusion……106

5 CONCLUSION……108

REFERENCES……111

APPENDIX A……124

APPENDIX B……132

APPENDIX c……148

CHAPTER 1

INTRODUCTION

The field of bioinformatics has emerged and developed as high-throughput technologies started to get widely implemented in research studies. High-throughput experimental methods, such as ChIP-seq, genome sequence, RNA-seq, ChIA-PET, and microarray, which are commonly used in studies, require sophisticated computational analysis. With overwhelming amounts of data continually getting generated, application of bioinformatics methods are crucial to extract biologically meaningful information.

The focus of this work is to apply various bioinformatics methodology to study and answer challenging questions in cancer. Cancer is a complex disease that is characterized by high proliferation and metastasis as one of its traits [1]. According to the American Cancer Society Cancer Statistics 2012, it is estimated that probability for male and female to encounter cancer in their lifetime is 1/2 and 1/3, respectively. It is a disease that affects most Americans either directly or indirectly.

Recently, an approach to cancer research has been experiencing transitions as it continues to adapt and integrate new technologies. The traditional method to study cancer has been finding the oncogenes or characterizing functions of genes that may be responsible or contribute to the development of cancer. This approach, while it has been effective at understanding the function of cancer-related genes, it has not been successful at providing key insights to the fundamental problems of cancer. The question of “why” and “how” are still being investigated.

One of the challenges in cancer is an early diagnosis. The prognosis of a patient depends on the stage of the cancer in which it is diagnosed; the earlier the cancer is diagnosed, the better the prognosis is. The prognosis is measured by the five year survival rate. In gastric cancer, for example, the survival rate if diagnosed at Stage I ranges from 57%-71%. In contrast, the survival rate for Stage IV diagnosis is at 4% [2]. The survival rate plummets as the cancer is diagnosed at a later stage. This statistic is a strong indication of the importance of the early detection. However, early diagnosis is not trivial as most cancers show no obvious symptoms until it is fully metastasized and invading other organs.

There are many methods currently used to diagnose cancer. Some of diagnostic methods are MRI, CAT -Scans, endoscopy, and X-ray. However, it is often inconvenient for patients and costly, thus it is not part of routine check-ups. An effective biomarker, that is highly accurate, is desirable.

We focus our study in enhancing biomarker discovery in urine. While there are other biofluid which are good sources of biomarkers, serum, plasma, or saliva, there are several advantages of finding biomarkers in urine. It is non-invasive, thus it is easy to collect. An ample amount of sample can be collected over duration of time for monitoring purposes. And urine is relatively simple in composition, which makes detection of biomarker and the results unequivocal.

Traditionally, approaches to the discovery of biomarkers were top-down. The data is collected to find all existing biomolecules in a sample and are compared to find proteins that show differential expression. When a sample is collected, protein profiling is done. The proteomics profiling usually utilizes various platforms of Mass Spectrometry. The challenges in interpreting the data and detecting proteins that are masked by highly abundant proteins may hinder or obscure the discovery of a true biomarker. Thus we propose and utilize a target-

focused study, where we circumvent the profiling step. The target-focused discovery proposed here, is to predict proteins that can get excreted to urine among the differentially expressed genes. Thus, an antibody mediated platform such as Western Blot or Elisa can be used to test the candidate markers. This methodology has proven effective and can be applied to wide-range of studies.

Another on-going battle in cancer is to understand how a cancer is developed and why they are able to evade cell death. Even after decades of research with tremendous amounts of resources, cancer still remains elusive. It is perplexed and complicated disease that remains to be understood in numerous perspectives.

With sequencing techniques revolutionizing the cancer research, the cost of genome sequencing technology has become affordable that it enables researchers to study cancer in genomic scale. The first cancer genome study on Acute Myeloid Leukemia was published in 2008 [3]. Since its astonishing finding of over 20,000 mutations in AML, the same approach was repeated in other types of cancer [4]. This approach was successful in identifying tumor specific mutations, thus collaborative efforts to sequence large numbers of cancer genomes has been initiated, termed The Cancer Genome Atlas.

Only recently it became feasible to carry out whole genomics projects thanks to the low sequencing cost fostered by the $1000 genome competition, the Archon Genomics X PRIZE. The next generation sequencing technologies empower researchers to comprehensively characterize the mutations profile in cancer. It is a beginning step towards understanding the genetic changes responsible for the disease. Using this approach, the sequence of 5 pairs of gastric cancer tissue genomes and its corresponding healthy tissue genomes are examined comparatively to characterize single point mutation behavior in gastric cancer.

In addition to genomics, transcriptomics is a widely popular method to profile gene expression in a sample. The field of transcriptomics was pioneered by the microarray platform, which later was expanded by –seq methods [5]. The transcriptomics data can provide insights into the transcriptional behavior or the levels of gene expression in different conditions. By utilizing transcriptomics data, we aim to answer the below questions.

1) We utilize transcriptomics data to understand the ways cancer cells suppress apoptotic pathways. Apoptosis, a cell death pathway, is intricately controlled in a cell. Each cell is equipped with a mechanism to initiate a suicide when an abnormal activity is detected. In a normal cell, the activity of proliferation and the apoptosis is correlated [6,7]. This is to prevent a cell from dividing uncontrollably. However, this balance in proliferation and apoptosis is lost and the apoptotic pathway is suppressed in cancer cells[1]. The mechanism on how a cell is able to suppress apoptosis is poorly understood. With transcriptomics data available, we approach this problem to examine the apoptosis related genes in different pathway: intrinsic apoptosis, extrinsic apoptosis, and immune response triggered apoptosis.
2) Transcriptome data can also be used to gain understanding on the transcription machinery. The co-localization of distal DNA regions is a phenomenon which has been observed during transcription but has been highly under the debate [8,9]. To approach this question, genome-wide interactome data is used to complement genome-wide transcriptome data to examine behavior of transcriptional activity in interacting sites. The aim is to find whether spatial colocalization is a phenomenon spurred by transcription, or is purely out of a random chance. If it is a random phenomenon, no correlation will be observed. If it is indeed promoted by

transcription, a high correlation between the transcriptional activity and colocalized regions will be observed.

As described above, cancer research can be approached in many aspects using bioinformatics methods. The results obtained from these studies are highly encouraging. Bioinformatics approach is not only effective, but now a necessary skill in studying diseases like cancer.

CHAPTER 2

A COMPUTATIONAL METHOD FOR PREDICTION OF EXCRETORY PROTEINS AND APPLICATION TO IDENTIFICATION OF GASTRIC CANCER MARKERS IN URINE[1]

[1] Hong CS, Cui J, Ni Z, Su Y, Puett D, et al. (2011) PLoS One 6: e16875
Reprinted here with permission of the publisher

Abstract

A novel computational method for prediction of proteins excreted into urine is presented. The method is based on the identification of a list of distinguishing features between proteins found in the urine of healthy people and proteins deemed not to be urine excretory. These features are used to train a classifier to distinguish the two classes of proteins. When used in conjunction with information of which proteins are differentially expressed in diseased tissues of a specific type *versus* control tissues, this method can be used to predict potential urine markers for the disease. Here we report the detailed algorithm of this method and an application to identification of urine markers for gastric cancer. The performance of the trained classifier on 163 proteins was experimentally validated using antibody arrays, achieving > 80% true positive rate. By applying the classifier on differentially expressed genes in gastric cancer *vs.* normal gastric tissues, it was found that endothelial lipase (EL) was substantially suppressed in the urine samples of 21 gastric cancer patients *versus* 21 healthy individuals. Overall, we have demonstrated that our predictor for urine excretory proteins is highly effective and could potentially serve as a powerful tool in searches for disease biomarkers in urine in general.

Introduction

The rapid advancement of *omic* techniques in recent years has made it possible to search for biomarkers for specific human diseases in a systematic and comprehensive manner, which is substantially improving our ability to detect diseases at early stages. Most of the previous biomarker studies have been focused on serum markers [10], mainly because of the known richness of serum in containing signals for various physiological and pathophysiological conditions.

Compared to serum markers, existing urinary markers are mostly related to urinary-tract or closely associated diseases. Only within the last few years has improved proteomic

analyses of urine samples revealed that, like sera, urine is also a rich source of information for detecting human diseases such as the graft-*versus*-host disease and coronary artery disease [11,12,13]. Note that urine is formed by filtration of blood through the kidneys; hence some proteins in blood may pass through the filters and be excreted into urine. As a result, the urinary proteins not only reflect the conditions of the kidney and the urogenital tract, but also those of other organs that may be distal from the kidney, as at least 30% of the urinary proteins are not originally from the urogenital tract [14,15]. The plethora of information in urine makes it an attractive source for biomarker screening since, compared to serum, the composition of urine is relatively simple, and urine collection is easier and noninvasive.

Marker identification in urine could potentially be done through comparative proteomic analyses of urine samples of patients with a specific disease and control groups. The challenge in such searches for urinary markers in a blind fashion is twofold. (a) Urine could have a large number of proteins/peptides (in contrast to the previous understanding[16]) with relatively low abundance. (b) The dynamic range in the abundance of these proteins could span a few orders of magnitude, wider than the range typically covered by a mass spectrometer [17]. For these reasons, comparative analyses, particularly (semi)quantitative analyses, of proteomic data of urine samples can be very challenging. This might be a key reason that there are no reliable urine markers for cancer diagnosis.

Our study focuses on development of a computational method for accurately predicting proteins that are urine excretory (see Figure 1 for the outline of the approach). These proteins must have specific properties that allow them to be secreted from cells first and then to be filtered out through the glomerulus membrane in kidneys. A recent proteomic study identified more than 1,500 proteins/peptides that are excreted into urine through healthy glomerular membranes [16]. Using this set of proteins and proteins deemed not to be urine excretory, we

have identified a list of distinguishing features between these two classes of proteins and trained a support vector machine (SVM) based classifier to predict if a given protein might be excreted into urine. The prediction method was experimentally validated using antibody arrays in conjunction with Western blots, and the results are highly encouraging.

This classifier has been applied to predict proteins that might be excreted into urine based on the identified differentially expressed genes in gastric cancer *versus* reference gastric tissues; and a number of potential urine markers for gastric cancer have been identified. A key contribution made in this work is that it provides a new and effective way to guide proteomic studies of urine by suggesting candidate marker proteins, hence allowing targeted marker searches using antibody-mediated techniques like Western blots and Elisa, which are substantially more feasible than large-scale comparative proteomic analyses of urine samples without any targets with which to work. While this prediction program has been applied to gastric cancer data in this study, no gastric cancer-specific information was used in this program; hence, it can be used for urine marker searches for other diseases.

Methods

This study consists of three main components: (i) construction of a classifier for predicting urine excretory proteins; (ii) evaluation of the performance of the classifier by applying it to a set of proteins for which the excretory status of the proteins is known; and (iii) application of the validated classifier to gene-expression data of gastric cancer to demonstrate its effectiveness in solving the urine marker identification problem.

Ethics

This research was approved by the Institutional Review Board at the University of Georgia, Athens, Georgia, USA (Office of the Vice President for Research DHHS Assurance ID NO. FWA00003901, Project Number 2009-10705-1)and by the Chinese Institutional Review Board

overseeing human subjects at Jilin University College of Medicine, Changchun, China. A consent form, approved by IRB at the University of Georgia and Chinese IRB, was collected from each subject. All subjects are aware that any data from research may be used for documents or publications as stated in the consent form.

a. An algorithm for predicting excretory proteins into urine

The general understanding of protein excretion from tissues to urine is that some proteins are secreted or leaked from cells into blood circulation, and then a portion of these proteins, along with some native proteins in blood, may be excreted into urine. Our goals are first to identify distinguishing features for such urine excretory proteins and then to build a classifier based on these features to predict which proteins in cells can be excreted into urine. To the best of our knowledge, there has not been any published work aimed to solve this problem. The importance in having such a capability is that it provides an effective link in connecting *omic* analyses of tissues to marker search in urine by providing candidate markers in urine that can be studied using antibody-based approaches.

The first step in developing such a predictive capability, i.e., a classifier, is to have a training dataset containing proteins that can and that cannot be excreted into urine, based on which a set of distinguishing features could possibly be identified. Fortunately, we have found one large proteomic dataset of urine samples from healthy people in a recently published study[16], which contains more than 1,500 unique proteins of which 1,313 have SwissProt accession IDs. We have used these 1,313 proteins as the positive training data for the to-be-trained classifier. The following procedure was then used to generate a negative training set: arbitrarily select at least one protein from each Pfam family that does not contain any positive training data, and the number of selected proteins from each family is proportional to the size of

the family [18,19]. As a result, 2,627 proteins were selected and used as the negative training set.

We examined 18 physiochemical features computed from protein sequences, which are potentially useful for the classification problem based on the general understanding of urinary excretion of proteins. The details of the 18 features and the computer programs used to calculate them are listed in Table S1. Some of these features are represented by multiple feature values, e.g., the amino acid composition in a protein sequence is represented by 20 feature values; overall the 18 features are represented using 243 feature values. We then identified a subset of features values from the 243, which can distinguish between the positive and the negative training data using an SVM-based classifier. The RBF kernel was used in our SVM training, considering its capability to handle non-linear attributes [20,21].

To ascertain which of the initially considered features are actually useful, the feature selection[2] tool provided in LIBSVM [20] was used to select the most discerning features among the 243. Codes used in this are publicly available from LIBSVM website (http://www.csie.ntu.edu.tw/~cjlin/libsvm/); we also have made the relevant program accessible at http://seulgi.myweb.uga.edu/files. An F-score [20], defined as follows, is used to measure the discerning power of each feature value to our classification problem,

$$F(i) = \frac{\left(\bar{x}_i^+ - \bar{x}_i\right)^2 + \left(\bar{x}_i^- - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}\left(x_{k,i}^+ - \bar{x}_i^+\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n-}\left(x_{k,i}^- - \bar{x}_i^-\right)^2},$$

where X_k refers to the training feature values (k=1,…, m); n_+ and n_- are the number of proteins in the positive (+) and negative (-) training dataset, respectively; X_i, $X_i^{(+)}$, $X_i^{(-)}$ are the averages of the *i*th feature value across the whole training dataset, the positive dataset and the negative

[2] Other feature selection tools could possibly be used but we have considerable experience in using this tool and found it to be adequate.

dataset, respectively; and $X_{k,i}^{(+)}$ and $X_{k,i}^{(-)}$ are the *i*th feature of the *kth* protein in the positive and negative training data, respectively. Generally, the larger an F-score, the more discriminative the corresponding feature is. In our selection, all features with F-scores above a pre-selected threshold were retained and used in training the final classifier. To find an optimal F-score threshold, we considered a list of possible thresholds and then selected the best one based on the training results.

The training of our SVM-based classifier is done using a standard procedure provided in LIBSVM [20] to find values of two parameters *C* and γ that give an optimal classification on the training data, where *C* controls the trade-off between training errors and classification margins, and γ determines the width of the kernel used [20]. Our training procedure is summarized as follows:

a. Calculate the F-score for every feature value;
b. For each of the pre-selected thresholds, do the following
    i. Remove the feature values with F-scores lower than the threshold;
    ii. Randomly split the training data into a sub-training and a sub-validation sets with equal size;
    iii. Train an SVM with an RBF kernel on the sub-training set to search for optimal values of *C* and γ, and then apply it to the sub-validation data and calculate the classification error;
    iv. Repeat steps (i) – (iii) five times and calculate the average validation error;
c. Choose the threshold that gives the lowest average validation error, and keep the features with F-score above the selected threshold; and
d. Retrain an SVM based on the selected features as the final classifier.

b. Datasets used to evaluate the performance of the classifier

An independent dataset was used to assess the performance of the trained classifier for which the excretory status of each protein is known. The positive subset of this dataset has 460 human proteins found in the urine of healthy individuals by three urinary proteomics studies [22,23,24], and the negative subset contains 2,148 proteins selected using the same procedure described previously but does not overlap with the negative set used for training.

The following measures were used to assess the classification accuracies: the sensitivity, the specificity, the accuracy, the Matthew's correlation coefficient, and the AUC [25]. Table 1 summarizes the classification accuracies of the trained classifier on the both training and the test datasets [25]. From the classification accuracies on the two datasets, we believe that our trained classifier captured the key distinct features of the excretory proteins in urine.

In addition, our classifier was tested on a separate dataset, a subset of the 274 proteins fixed on a pre-made protein antibody array (the RayBio Human G-series Array 4000 (RayBiotech, Inc., Norcross, GA)). Of the 274 proteins, 111 are known to be excretory and were included in our training or independent test dataset. We applied the classifier on the remaining 163 proteins for which the excretory status was unknown (see Results and Table S2). This protein array provides the relative expression level for each protein on the array when tested on a (urine) sample, which is measured in terms of the signal intensity, quantified by the densitometry. The background of the array was used as the control to determine the actual presence of a protein in the (urine) sample. The signal intensity for a protein was considered as a true signal if it was at least 5-fold higher than that of the control, as suggested by the manufacturer's recommendation. We focused our experimental validation on confirming the positive predictions only since it is virtually impossible to prove a protein is not present in a urine sample due to limitations in detection sensitivity of the current technology when the protein is of very low concentration in the sample.

c. Urine sample collection/preparation

Urine samples from gastric cancer patients and healthy controls were collected at the Medical School of Jilin University, Changchun, China. Gastric cancer patients, from who the samples were collected from, are all late stage patients (see Table S3 for patient information). These samples were immediately lyophilized and stored at -80°C until further use after their surgical removal from the patients. They were then reconstituted and centrifuged (3,000*xg* for 25 min at 4°C) to remove cellular components. The supernatants were collected and dialyzed at 4^0C against Millipore ultra-pure water (three buffer changes followed by an overnight dialysis) using Slide-A-Lyzer Dialysis Cassettes (Thermo Fisher Scientific, Rockford, IL). Protein concentrations were measured using the Bio-Rad Protein Assay (Bio-Rad, Hercules, CA) with bovine serum albumin as a standard.

d. Identification of genes that are differentially expressed in gastric cancer and control tissues

A total of 80 gastric cancer tissues and their adjacent noncancerous tissues from 80 patients were collected at the Medical School of Jilin University. Microarray experiments were conducted on these tissues using the Affymetrix GeneChip Human Exon 1.0 ST Array, which covers 17,800 human genes. The PLIER algorithm [26] was used to summarize the probe signals to gene-level expressions. For each gene, we examined the distribution of the expression fold-change between the paired cancer and control tissues across all 80 pairs of tissues. Let $K_{exp,}$ be the number of pairs of tissues whose fold-change is at least 2. A gene is considered as *differentially expressed* if the *p*-value of the observed K_{exp} is less than 0.05. Using this criterion, a total of 715 genes were found to be differentially expressed in gastric cancer across all human genes, and the names of the 715 genes, along with the associated K_{exp} and *p*-values, are given in Table S4. A detailed study of the microarray data has been reported elsewhere [27].

e. Function and pathway enrichment analyses

The DAVID Bioinformatics Resources and the KOBAS web server [28,29] were used to do functional and pathway enrichment analysis, respectively, for all the predicted urine-excretory proteins, using the whole set of human proteins as the background. We refer the readers to [28,29] for details on the methods for functional and pathway enrichment analyses. Using DAVID Bioinformatics Resources, the enrichment score for a specified group of proteins was determined by the EASE score [28,30] KOBAS is a complementary tool to DAVID as it expands the gene annotation using KEGG Orthology (KO) terms. The KOBAS web server, along with the KO-based annotation system [29,31], was used to find statistically enriched and underrepresented pathways among the predicted urine-excreted proteins. KOBAS takes in a set of protein sequences and annotates them using the KO terms. The annotated KO terms were then compared against all human proteins as the background set for assessing if they are being enriched or underrepresented.

f. Western blots

Urinary proteins from each sample (total of 2 μg) were combined with 3x sample dye. Each tube was boiled for 5 min and loaded on SDS-PAGE gels, along with 10 μl standards and run for 1 h at 200 volts. The membrane was activated with 100% methanol, following a transfer from the gel to the membrane (100 volts for 1 h). Once the transfer was complete, the membrane was allowed to dry, rewetted in 100% methanol and washed 2X for 5 min each with Tris-Buffered Saline (TBS). The membrane was then incubated in 3% milk blocking solution for 2 h at room temperature. Next the membrane was incubated in the first antibody solution (1:200 dilution in 1.5% milk blocking) for 1 h at room temperature, and the unbound antibody was removed by washing the membrane 3X with TBS Tween-20 (TBST) solution for 10 min each. Then the membrane was incubated in a 1:10,000 dilution of the secondary antibody in 1.5% milk blocking solution for 1 h at room temperature. The membrane was washed 3X with TBST and 2X

with TBS (10 min each). Lastly, the membrane was covered completely with an equal amount of enhancer and peroxide solution from a Pierce Western Blotting kit for 5 min and exposed to the film. Each experiment was repeated multiple times to ensure reproducibility. The signal intensities were determined using the imageJ software [32]. For each membrane, the blank lane was used to normalize the signal intensities across the membranes. The performance was examined using ROC and whisker-box plot.

Results and Discussion

Signal peptide and secondary structures are key features of urine-excreted proteins

The initial list of features was carefully selected to include what we believed to be protein characteristics relevant to urinary excretion based on literature search and our current understanding of urinary proteins. For example, the negatively charged glomerular wall in kidney will allow the filtration of only positively or neutrally charged proteins. Thus, charge of a protein is one of the features we selected. Taking the available information into consideration, the total number of feature values collected initially was 243, representing basic sequence properties, motifs, physicochemical properties, and structural properties (Table S1). In identifying features that are effective in discriminating urine excretory proteins from the non-excretory ones, a simple and effective method to eliminate features that show little or no discerning power for our classification problem was employed; 74 feature values were selected using the procedure outlined in Section a of Methods (Table S5). These feature values were used to train the final classifier.

Among the selected features, the most discriminatory one was the presence of signal peptides. It is understood that proteins that are secreted through the ER have signal peptides and are trafficked to their destination according to the specific signal peptides; thus, not surprisingly, most excreted proteins have this feature. Another prominent feature was the

secondary structure type; specifically, the percentage of alpha helices in a protein sequence was ranked as the number 2 feature value among the selected 74 (Table S5). As expected, the charge of a protein was among the top ranked features for excreted proteins. This is consistent with the general understanding that charge is a factor in determining which proteins can be filtered through the glomerular membrane [33] as proteins inside glomerular membranes and podocyte slits are negatively charged, and hence negatively charged proteins will have low chances to filter through the kidneys. Indeed, the feature values of positive amino acids and charge were among the top ranked feature values.

Interestingly, however, molecular weight, which ranked at 232 out of 243, was not included in the final 74 feature values. This could be explained by the following. Proteins present in serum may have already undergone a cleavage or have been partially degraded, and thus may not be in their intact or complete form when they enter the kidney. It has, in fact, been established that the majority of proteins found in urine are extensively degraded [34]. While an intact protein may not be able to filter through the glomerulus due to its size or shape, a protein-derived peptide may easily pass through the podocyte slits. As a result, the molecular weight of the intact protein is a non-factor in predicting if the protein is urine excretory.

It should be noted that urine excretory proteins and secreted proteins share some common characteristics as some of the features used to identify blood-secreted proteins in our previous study [18] were selected in the urinary protein prediction in this study. For example, features such as solvent accessibility, polarity and signal peptides were included in both classifiers. However there is a clear difference between the features used in the two classifiers. While features such as beta-strand-content, features associated with beta-barrel transmembrane protein and protein ratio, TatP motif, transmembrane domain, protein size, and the longest disordered region were among the top features for prediction of blood-secretory

proteins [18], they were not included in the final features for the urinary protein prediction. Moreover, features related to positive charge, such as the composition of positively charged amino acids, were prominent in urinary protein prediction but not selected in the blood secretion prediction. Similarly, the alpha-helix-content and the coil-content of proteins were among the top features for urinary protein prediction, but they were not selected for the blood-secretory protein prediction. It is interesting to note that in contrast to the finding that beta-strands are a common secondary structure type among the blood secretory proteins, urinary proteins tend to have higher alpha-helix and coil content, which indicates that the urinary proteins possess properties not shared by blood secretory proteins in general.

Performance of the classifier

To determine the accuracy of the final classifier, we tested it on an independent test set, which consists of 460 experimentally validated urine excretory proteins and 2,148 non-urine excretory proteins. Our classifier has its prediction sensitivity and specificity on this independent test set at 0.78 and 0.92, respectively (Table 1).

We then ran the classifier on the 163 out of the 274 proteins fixed on the pre-made antibody array (see Methods), for which the excretory status was unknown. Of the 163 proteins, 112 proteins were predicted to be urine excretory by our classifier. To assess the performance of this prediction, antibody array-based experiments were conducted on 14 urine samples, seven from healthy individuals and seven from gastric cancer patients. Of the 112 predicted urine-excretory proteins, 92 were found in at least one of the urine samples (Table S6), giving a positive prediction rate of 0.81, which is consistent with the performance level on the first test set.

It should be noted that one limitation of this classifier is that some proteins might have been partially degraded before being excreted into urine or in urine, making it difficult for our

classifier to detect so formed peptides as it was trained on whole intact proteins. This issue will be addressed in the future through deriving feature values based on the actual proteins/peptides identified in previous urinary proteomic studies rather than their corresponding full-length proteins as done in this study. While there is clearly room for further improvement, the prediction results of the current classifier are highly encouraging.

Application of classifier to gastric cancer data

Our previous study on 160 sets of microarray gene-expression data of gastric cancer has identified 715 differentially expressed genes with at least 2-fold changes in gastric cancer *versus* control tissue samples [27]. Our classifier was applied to these 715 proteins[3], and it predicted that 201 of the 715 proteins are urine excretory. Table S7 provides the detailed information of the 201 proteins. Since it is unrealistic to check all the 201 proteins in this study to determine if they are urine excretory or not, we did analyses to narrow down this list. Specifically, we have carried out the following analyses: (i) functional and pathway enrichment analyses to gain a better understanding of the types of proteins present in urine, (ii) literature search on urinary proteins to compile information about published urinary marker proteins, (iii) examining the gene expression data to remove genes that are not substantially differentially expressed between cancer and control tissue samples; and (iv) Western blots on proteins chosen from a narrowed down list of the 201 proteins. This procedure showed a high success rate and led to an interesting discovery of potential biomarker for gastric cancer.

For (i), we have carried out functional and pathway enrichment analyses on all the 201 proteins using the DAVID [28] and KOBAS [29] servers, respectively. We found that the enriched functional groups included the extracellular matrix (ECM), cell adhesion, and development, cell

[3] While it would be preferable to have proteomic data of the tissue samples, we have only gene-expression data available in this study. Hence, gene expression data are being used as an approximation to the protein expression in this methodology-oriented study.

motility, defense response, angiogenesis, which are all known to be involved in the development of or in defense of cancer (Figure. S1A). The most enriched pathways were ECM-receptor interaction and inorganic ion transport and metabolism pathways (Figure S1B).

The following criterion was used to reduce the list of 201 proteins for steps (ii) - (iii): *the proteins have not been reported to be related to any cancer based on our extensive literature search*, which gives rise to 71 proteins. The list was further reduced based on a pre-selected cutoff on differential expressions (> 2 fold difference) and functional annotations (potentially relevant to gastric cancer rather than immune responses).

Endothelial lipase is substantially reduced in the urine samples of gastric cancer patients We chose six proteins (MUC13, COL10A1, AZGP1, LIPF, MMP3, and EL) for experimental validation from the above narrowed down list. To do this, we have collected urine samples of 21 gastric cancer patients and 21 healthy individuals. Of the six selected proteins, five proteins, MUC13, COL10A1, LIPG, AZGP1, and EL were detected by Western blots in at least one urine sample. Out of the five, MUC13, COL10A1, and EL were detected even at a very low quantity of the total urinary proteins (1-2 μg). MMP3 was not found in the samples we tested, which may be due to the low concentration of MMP3 in urine or a false prediction by our classifier.

It is particularly interesting to note that we were able to detect consistent differences in the EL abundance (encoded by *LIPG*) between the two sets of 21 urine samples. The Western blots for EL showed a substantial reduction in its abundance in urine samples of the 21 gastric cancer patients compared to the control samples. As shown in Figure. 2A, the majority of the control samples showed the presence of EL, whereas most of the gastric cancer samples had relatively low amounts of EL. This pattern was observed repeatedly.

The molecular weight of this protein has been determined to be 68 kDa [35]; thus, a homo-dimer is expected to be 134 kDa. In the Western blots, however, bands were detected at

near 100 kDa. This probably corresponds to a partially cleaved homo-dimer, an active form of which was confirmed by a previous study [36], although the possibility of a monomeric form of EL associated with another protein cannot be ruled out. The Western blots do provide semi-quantitative information based on the signal intensities. The ROC curve suggests that the EL concentration was discriminant in distinguishing the gastric cancer samples from the non-gastric cancer samples, yielding an AUC greater than 0.9 (Figure. 2B-C). Using 5,000 as a signal intensity cutoff, true positive rate and false positive rate were 85% and 9.5%, respectively.

A further study is required to assess EL as a gastric cancer biomarker. The limited sample size of 21 samples in each group is too small to accurately evaluate its potential for biomarker. Enrolling many more patients is needed to confirm the efficacy of EL as a potential biomarker for clinical purposes. Also, it would be interesting to test EL on the early stage of gastric cancer, as our samples were all from late stage gastric cancer patients. Nonetheless, our preliminary result shows highly encouraging results.

Concluding remarks

The available evidence indicates that many proteins are excreted into urine that may be good biomarker candidates for different diseases. The novel computational method developed and used herein for predicting excreted proteins may aid in identifying these and other biomarkers in urine. Our study has demonstrated that the integrated approach, coupling bioinformatics prediction with experimental validation, is an effective paradigm for identification and validation of potential urinary biomarkers. We anticipate that this approach will provide a powerful tool in the future for urinary proteomics and biomarker studies in general.

ACKNOWLEGEMENT: We thank Xizeng Mao (CSBL, UGA) for his assistance in pathway analysis.

Figures and Tables

Table 1: Classification performance by the trained classifier on the training and an independent test set

Sets	TP	TN	FP	FN	SEN	SP	ACC	MCC	AUC
Train	972	2,493	134	341	0.74	0.95	0.88	0.52	0.94
Independent	360	1,983	165	100	0.78	0.92	0.90	0.45	0.93

TP=true positive; TN=true negative; FP=false positive; FN=false negative; N= total number of proteins in dataset; SEN = TP/(TP+FN); SP = TN/(TN+FP); ACC = (TP+TN)/N; MCC =(TPxTN-FPxFN)/√((TP+FN)(TP+FP)(TN+FP) (TN+FN)); AUC is described in (37).

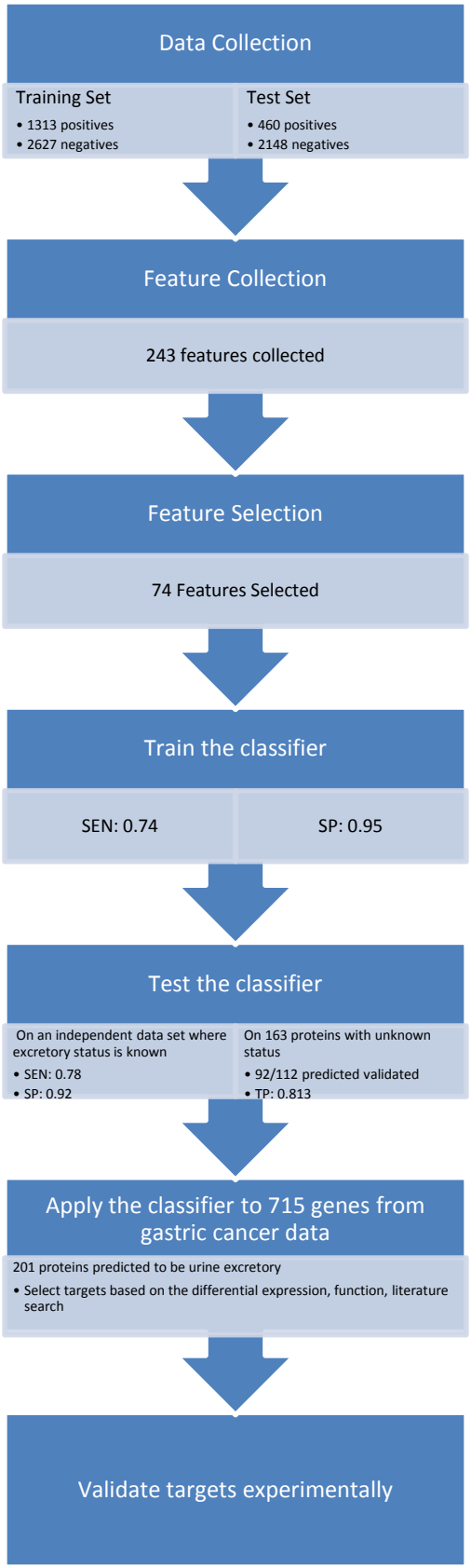


Figure 1. The outline of the study

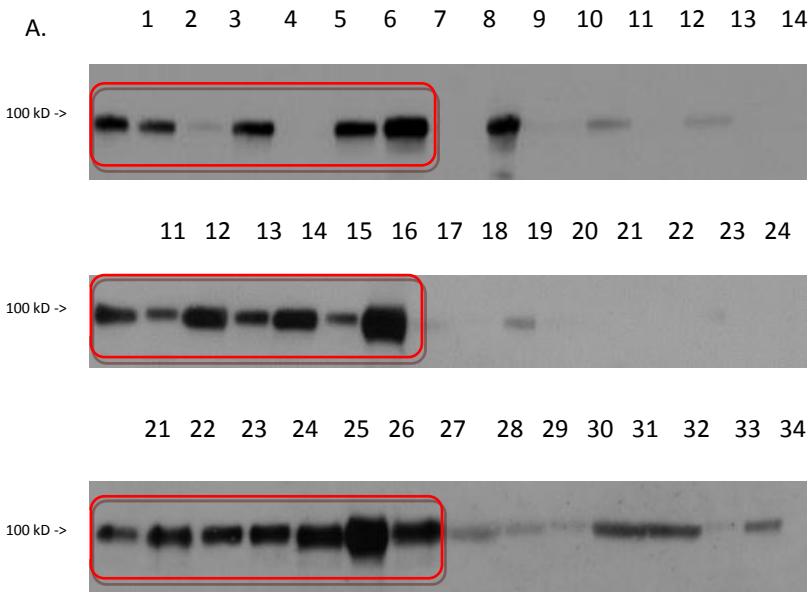


Fig. 2: A: Western blots for EL on control and gastric cancer samples. Control samples (denoted by the red lines): Lanes 1-7, 11-17, 21-27. Cancer samples: Lanes 8-14, 18-24, 28-34. B: Corresponding whisker-box plot for the signal intensities. C. ROC curve of the EL Western blot. Red line: no discrimination; blue line: ROC by EL.

Table S1: Summary of features used in the initial classification model

Feature class	Features (No. of feature values	Program used to calculate the features
Sequence features	Sequence Length (1) AA composition (20)	Fldbin [37], Profeat [38]
Physicochemical properties	Hydrophobicity (21), normalized Van der Waals volume (21), polarity (21), polarizability (21), charge (21), secondary structure (21), solvent accessibility (21), Pseudo-AA descriptor (50)	Locally calculated, Profeat [38]: using three descriptors: composition, transition, and distribution
	Unfoldability (1), charge (1), hydrophobicity (1), #	Fldbin [37], Swiss-[39], locally calculated

	of disordered regions (1), longest disordered regions (1), # of disordered residues (1), PI (1), MW (1), charge (2), percentage of disordered region (1)	
Motifs	Transmembrane domain (1), Twin-arginine signal peptide (1), transemembrane domains (alpha helix, or beta barrel) (2), Glycosylation number & presence (N&O linked) (4)	TMB-Hunt [40,41], TatP [41], phobius [42], NetOgly [43], NetNGly [44]
Structural properties	Secondary structural content (4), Radius gyration (1), Radius (1),	SSCP [45], Radius Gyration (http://www.scfbio-iitd.res.in/software/proteomics/rg.jsp, locally calculated

Table S2: Uniprot IDs of 163 proteins used for classifier performance evaluation

P03950	P13232	P09038	P10145	P22003	P01135	Q6EBC2	Q15389	P04271
O43927	P13500	P01138	O14625	P18075	P61812	P01308	Q9BY76	Q15465
P12644	P80075	P35070	P47992	Q9H2A7	P35590	Q9BQR3	Q15582	P01266
P22004	P80098	Q9NRJ3	P10147	O75509	Q02763	Q29983	Q8WXI7	
P51671	Q99616	Q9Y4X3	P13236	P21860	Q99727	Q29980	Q16790	
O00175	O75078	Q06418	Q99731	P16581	P33151	P22894	P25774	
Q9Y258	Q07325	P42830	P34130	P48023	P35968	P09238	Q16627	
P10767	Q16663	P25445	O00300	Q14627	P35916	P05121	O00585	
P21781	P78556	P08620	Q61207	P31785	P28908	P02776	P13385	
P80162	P20783	P31371	P20333	Q9HBE5	P29965	Q15109	P41271	
P39905	P55774	P09919	P19438	Q01344	P31994	Q9Y6Q6	P07585	
Q76BR7	A2NWD3	Q9UNG2	O15444	P15248	Q8N4E7	P02735	O94907	
P08833	Q92583	Q9Y5U5	P01033	P02778	O95633	Q9Y336	Q9UBP4	
P22692	P01137	P09341	P40225	P48357	P19883	P78536	P27487	
P05019	P10600	O15467	O14798	P03956	P09958	Q96D42	Q92838	
P22301	P01375	P14210	Q9UBN6	P45452	P01241	O14763	P58294	
Q14005	P01374	P32942	P15692	P55773	Q13651	Q9NP99	P04626	
P01583	Q15848	P08069	O43915	P16234	Q9GZX6	Q969D9	P78552	
P60568	O15123	P29460	Q13740	P01236	Q8IZJ0	Q9HAV5	Q9P0M4	
P05112	P15514	Q16552	P33681	O15389	Q8IU54	P02771	Q96PD4	

Table S3: Patient information for Western blot analyses

Lanes	Sample	Gender	Age Group
1	N1	M	20-30
2	N2	M	40-50
3	N3	M	20-30
4	N4	M	20-30
5	N5	M	20-30
6	N7	M	20-30
7	N17	F	30-40
8	C1	F	50-60
9	C4	F	50-60
10	C5	M	60-70
11	C7	F	70-80
12	C9	M	70-80
13	C10	M	60-70
14	C11	M	80-90

15	N11	F	40-50
16	N15	F	unknown
17	N18	M	unknown
18	N34	M	40-50
19	N19	M	unknown
20	N21	M	40-50
21	N23	F	40-50
22	C2	F	60-70
23	C3	M	40-50
24	C6	M	60-70
25	C8	M	unknown
26	C10	M	50-60
27	C11	M	60-70
28	C12	F	50-60
29	N26	M	40-50
30	N28	M	50-60

31	N32	M	40-50
32	N38	M	50-60
33	N39	F	30-40
34	N48	M	50-60
35	N37	M	50-60
36	C13	F	40-50
37	C16	unknown	unknown
38	C18	M	60-70
39	C20	M	30-40
40	C22	M	70-80-
41	C23	M	unknown
42	C30	F	40-50-

Table S4: Patient information for the gastric cancer tissues

Patient No.	Age	Gender	Stage
1	41	F	IV
2	62	F	III
3	54	F	III
4	62	F	IIIA
5	63	M	IIIB
6	56	M	IIIB
7	71	M	IIIB
8	55	F	IIIB
9	53	M	IIIB
10	-	M	IV
11	55	M	IIIB
12	51	M	IIIB
13	64	M	IIIB
14	53	F	IIIB
15	56	M	IIIB
16	54	M	III
17	53	M	III
18	71	M	III
19	57	M	IIIA
20	58	M	III
21	42	M	IB
22	73	M	IB
23	69	F	III
24	65	F	IIIA
25	50	M	III
26	47	M	IB
27	59	M	III
28	75	M	III
29	40	M	III
30	69	M	III
31	41	M	II
32	76	F	II
33	51	F	III
34	36	M	IIIA
35	67	F	IV
36	42	M	III
37	68	M	III
38	65	M	III
39	59	M	III
40	68	M	IV
41	74	M	IB
42	65	F	IIIA
43	50	M	III
44	49	M	III
45	58	M	IV
46	-	F	IV
47	53	F	IIIA
48	84	M	IV
49	60	F	IIIB
50	55	M	III
51	70	M	II
52	56	F	III
53	43	F	III

54	71	F	III
55	56	F	IV
56	81	M	III
57	65	M	III
58	55	M	III
59	56	F	II
60	76	M	II
61	78	F	III
62	55	M	III
63	65	M	III
64	68	M	III
65	63	M	IV
66	-	M	IV
67	57	F	III
68	68	F	III
69	54	M	III
70	51	M	II
71	34	M	III
72	75	F	IV
73	61	M	III
74	54	M	IV
75	55	M	III
76	67	F	II
77	62	F	IV
78	50	F	III
79	71	M	IV
80	58	M	IV

Table S6: List of 74 Features according to the rank

Rank	Description
1	presence of SP
2	Composition Secondary Structure: Helix (EALMQKRH)
3	Composition Normalized van der Waals vol. (range 0-2.78)
4	% of alpha-content
5	Transition Normalized van der Waals vol. (range 4.03-8.08)
6	Transition Secondary Structure: Coil (GNPSD)
7	Transition Polarizability value (.219-.409) KMHFRYW
8	Composition Charge. Positive (KR)
9	Composition Polarizability value (0-1.08) GASDT
10	Transition Polarizability value (0-1.08) GASDT
11	Composition Normalized van der Waals vol. (range 4.03-8.08)
12	Composition Polarizability value (.219-.409) KMHFRYW
13	% of coil-content
14	Amino acid composition G
15	Pseudo-AA descriptors
16	Amino acid composition T
17	Composition Secondary Structure: Coil (GNPSD)
18	Isoelectric point
19	Composition Charge. Neutral (ANCQGHILMFPSTWYV)
20	Transition Charge. Positive (KR)
21	Composition Hydrophobicity-neutral (GASTPHY)
22	Transition Normalized van der Waals vol. (range 0-2.78)
23	Transition Solvent Accessibility: Exposed(RKQEND)
24	Composition Polarity. Polarity Value(8.0-9.2) PATGS
25	Composition Polarity. Polarity Value(10.4-13.0) HQRKNED
26	Distribution
27	Pseudo-AA descriptors
28	Pseudo-AA descriptors
29	Distribution
30	Amino acid composition R
31	Composition secondary Structure: Strand (VIYCWFT)
32	Number of N-glyc site
33	Composition Hydrophobicity-polar (RKEDQN)
34	Composition Solvent Accessibility: Exposed(RKQEND)
35	Transition Polarity. Polarity Value(4.9-6.2) LIFWCMVY
36	Pseudo-AA descriptors
37	% of disordered region
38	Amino acid composition K
39	Amino acid composition C

40	Charge calculated
41	Distribution
42	Pseudo-AA descriptors
43	Pseudo-AA descriptors
44	Distribution
45	Amino acid composition M
46	Amino acid composition E
47	Pseudo-AA descriptors
48	Transition Charge. Neutral (ANCQGHILMFPSTWYV)
49	Distribution
50	Distribution
51	Transition Hydrophobicity-neutral (GASTPHY)
52	Transition Polarity. Polarity Value(8.0-9.2) PATGS
53	Composition Solvent Accessibility: Buried (ALFCGIVW)
54	Distribution
55	Pseudo-AA descriptors
56	Distribution
57	Composition Normalized van der Waals vol. (range 2.95-4.0)
58	Distribution
59	Transition Hydrophobicity-hydrophobic (CLVIMFW)
60	Charge
61	Pseudo-AA descriptors
62	Amino acid composition H
63	Unfoldability
64	Amino acid composition L
65	Distribution
66	Distribution
67	presence O-glyc site
68	Amino acid composition N
69	Distribution
70	Amino acid composition Y
71	Amino acid composition W
72	Pseudo-AA descriptors
73	Amino acid composition V
74	Pseudo-AA descriptors

Table S7: Experimental confirmation results of predicted urine excretory proteins (TP: true positive, FP: false positive)

Protein ID	Experiment Result
P03950	TP
P22004	TP
P21781	TP
P08833	TP
P05019	TP
P13500	TP
O75078	TP
P78556	TP
P55774	TP
P01137	TP
P10600	TP
Q15848	TP
O15123	TP
P15514	TP
P01138	TP
P35070	TP
Q9NRJ3	TP
Q06418	TP
P42830	TP
Q9UNG2	TP
Q9Y5U5	TP
P09341	TP
O15467	TP
P14210	TP
P08069	TP
P29460	TP
Q16552	TP
O14625	TP
P47992	TP
P10147	TP
P13236	TP
Q99731	TP
P34130	TP
P01033	TP
P40225	TP
O14798	TP

O43915	TP
Q13740	TP
P22003	TP
P18075	TP
Q9H2A7	TP
O75509	TP
P21860	TP
P16581	TP
Q14627	TP
P31785	TP
Q01344	TP
P15248	TP
P02778	TP
P48357	TP
P45452	TP
P55773	TP
P16234	TP
P01236	TP
O15389	TP
P01135	TP
P61812	TP
P35590	TP
Q02763	TP
P33151	TP
P35968	TP
P35916	TP
P28908	TP
Q8N4E7	TP
O95633	TP
P09958	TP
P01241	TP
Q8IZJ0	TP
Q6EBC2	TP
Q29983	TP
Q29980	TP
P22894	TP
P05121	TP
Q15109	TP
Q9Y6Q6	TP
Q9Y336	TP
Q9NP99	TP

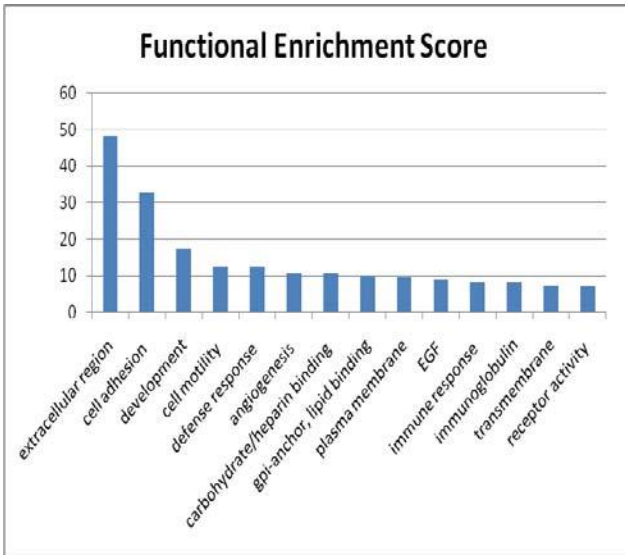
P02771	TP
Q15389	TP
Q15582	TP
Q16627	TP
O00585	TP
P07585	TP
Q9UBP4	TP
P27487	TP
P04626	TP
P78552	TP
Q9P0M4	TP
Q96PD4	TP
Q15465	TP
P01266	TP
P12644	FP
P80162	FP
Q16663	FP
P25445	FP
P32942	FP
O00300	FP
Q61207	FP
P20333	FP
P19438	FP
Q9HBE5	FP
P03956	FP
P19883	FP
Q13651	FP
P01308	FP
Q9BQR3	FP
P09238	FP
P78536	FP
O14763	FP
Q16790	FP
P25774	FP
P41271	FP

Table S8: The list of 201 genes predicted to be excretory from differentially expressed genes of gastric cancer

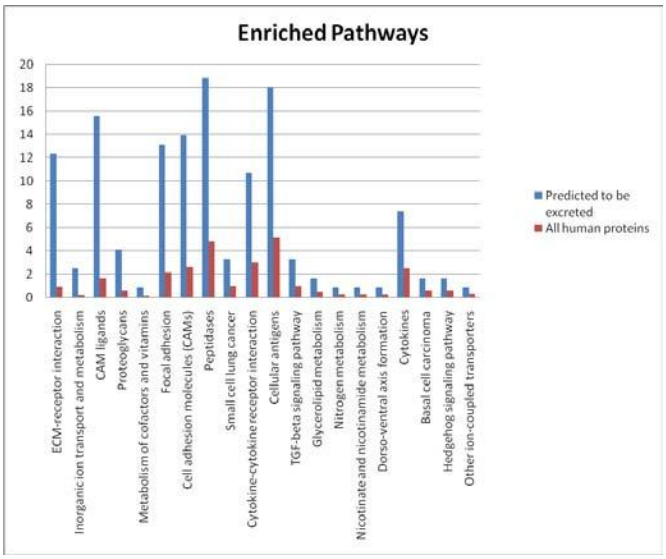
http://csbl.bmb.uga.edu/~xyn/Table_S6.pdf

Figure S1: A. Enriched functional groups as identified by DAVID. The x-axis represents the functional groups, and the y-axis represents the enrichment score. B. Enriched pathways for 201 predicted urine proteins using the KOBAS web server. Each blue bar represents the percentage of the 201 proteins; each red bar indicates all human proteins; the x-axis indicates the pathway names; and the y-axis indicates the percentage. (Red bar indicates all human proteins; the x-axis indicates the pathway names; and the y-axis indicates the percentage).

A.



B.


Enriched Pathways
20
18
16
14
12
10
8
6
4
2
0
ECM-receptor interaction
Inorganic ion transport and metabolism
CAM ligands
Proteoglycans
Metabolism of cofactors and vitamins
Focal adhesion
Cell adhesion molecules (CAMs)
Peptidases
Small cell lung cancer
Cytokine-cytokine receptor interaction
Cellular antigens
TGF-beta signaling pathway
Glycerolipid metabolism
Nitrogen metabolism
Nicotinate and nicotinamide metabolism
Dorso-ventral axis formation
Cytokines
Basal cell carcinoma
Hedgehog signaling pathway
Other ion-coupled transporters
Predicted to be excreted
All human proteins

CHAPTER 3

CHARACTERIZATION OF URINE-EXCRETORY MIRNAS AND DEMONSTRATION OF THEIR FEASIBILITY AS DISEASE MARKERS[4]

[4] Hong CS, Cui J, Xu Y Submitted to *PLoS ONE* (May, 2012)

Abstract

Urinary microRNAs (miRNAs) have recently been attracting increasing attention as biomarkers for human diseases. This is mainly because, like proteins, aberrant expressions of miRNAs can provide information about diseases of organs distant from kidney, but unlike proteins, miRNAs tend not to be processed by liver so they remain intact substantially longer than proteins. In this study, the feasibility of urinary miRNAs as potential biomarkers is demonstrated using a computational approach. Based on a selected set of features of miRNAs previously found in urine, a Support Vector Machine-based classifier to distinguish excretory miRNAs from non-excretory miRNAs was constructed. On our training and testing sets, the classifier was able to positively identify excretory miRNA at 0.9 sensitivity rate. This classifier was applied to a set of 325 miRNAs with their excretory status unknown, which predicted 138 to be urine excretory. This finding suggests that there are many more urinary miRNAs than what the previous studies have suggested. When used in conjunction with differential expression analyses of miRNAs in diseased tissues, this capability provides a powerful tool in searching for disease diagnostic markers in urine.

Introduction

MicroRNAs (miRNAs) are small, single-stranded endogenous non-coding RNAs (18-22 nt) that comprise 1-3% of the human genome [46]. These small miRNAs have been found to play a crucial role in maintaining the normal levels of mRNAs. One of the well-known functions of miRNA is post-transcriptional regulation of gene expression. They regulate gene expression by binding to the complement mRNA strand and cleaving the transcript products, thereby preventing them from being translated into proteins [47,48]. Naturally, abnormal abundances of miRNAs can cause the overexpression or repression of a gene. Due to this nature of miRNAs, their deregulated expressions can have detrimental effects on cells. As a result, miRNAs have

been implicated in various diseases. Studies have indicated that the level of miRNAs found in body fluids may reflect various pathological conditions and diseases, such as cancer [49,50,51,52,53,54]. For example, miRNAs may target a tumor repressor gene or an oncogene; aberrant expression of such miRNAs may be associated with the development of cancer. Therefore, the interest in searching for miRNAs and using them as potential biomarkers for various diseases, especially cancer, continues to grow [55,56,57].

MiRNAs have been widely studied as disease markers in body fluids. Unlike proteins or mRNAs, which are degraded relatively easily, miRNAs are stable in body fluids, including blood and urine, and are protected from RNAse degradation [58], hence making them more ideal candidates as disease biomarkers.

To facilitate effective searches for miRNA biomarker in body fluids, we have developed a computational method to predict cellular miRNAs that can be excretory to urine using a similar approach to our previous work on protein biomarkers [59,60]. While the challenge is presented with urine as its concentration of biomolecules is much more diluted compared to that of serum, nevertheless, urine has been suggested as a good source of miRNAs as potential disease markers by a number of studies [48,61,62]. The basic idea of the study can be explained as follows.

We utilized a machine learning methodology to discover features of miRNAs that have been identified in urine, which are distinct from those of the miRNAs that have not been found in urine. We then trained a Support-Vector-Machine (SVM) classifier based on the identified features, which can best distinguish between the urine excretory miRNAs from the non-excretory miRNAs [63]. The design of the study is outlined in Figure 1. We have previously showed this protocol is highly effective in suggesting reliable biomarker candidates for experimental validation when used in conjunction with differential expression data of

genes/proteins in cancer *versus* control tissues [59]. Using this classifier, we predicted a list of miRNAs that can possibly get excreted into urine.

Results

Overall 10 features (Table S1) of miRNAs were examined with the aim of identifying features, each of which shows differential distributions between miRNAs that have been found in urine and those that have not been found in urine. These 10 features are represented as 158 feature values. This analysis of differential distributions is done on all the 237 urine excretory miRNAs and the remaining miRNAs among all the 718 human miRNAs identified at the time of data collection (see Materials and Methods). At the end of such feature selection and elimination, 10 feature values from 5 features were found to show differential distributions between the excretory and non-excretory miRNAs (Table S2). Table 1 lists these feature values. The top ranking feature was the composition of guanine nucleotide at the 5' end of miRNAs. Many of the most discriminant features are in the category of the nucleotide composition, suggesting the nucleotide composition as an important feature in determining the excretory traits of miRNAs.

Using these 10 features, a classifier was trained to discriminate the excretory from non-excretory miRNAs using an SVM learning method (see Materials and Methods). The performance of the classifier was analyzed on the training and the testing set by using the following measures: TP=True Positive; TN=True Negative; FP= False Positive; FN=False Negative; sensitivity; specificity; ACC=accuracy; MCC=Matthew's Correlation Coefficient (see Materials and Methods). The performance of the trained classifier is summarized in Table 2. The classifier was able to accurately identify excretory miRNAs with sensitivity over 0.90 on the training set. When applied to an independent testing set, the classifier achieved a similar result, able to identify excretory miRNAs at a sensitivity level at 0.88. As the purpose of the classifier is to assist in

discovering excretory miRNAs for biomarker purposes, the high sensitivity is crucial (*versus* the specificity level).

The specificity of the trained classifier was relatively low compared to the high sensitivity. We believe that a key reason could be that some miRNAs in the negative (non-excretory set) dataset may be urine excretory but have not been detected yet experimentally due to the limited efforts in searching for such miRNAs. We fully anticipate that the trained classifier actually does better than its identification specificity suggests. The relatively MCC numbers of the trained classifier are related to the low specificity.

This classifier was applied to an independent set of miRNAs. This set contains 325 miRNAs that were found in serum but not in urine at the time of data collection, with their excretion status being unknown. Among the 325 miRNAs, 82 were differentially expressed in various cancer tissues (Table S3). Our trained classifier predicts that 138 of the 325 miRNAs were urine excretory, of which 11 were repeatedly implicated in cancer (Table S4).

To assess the performance of our classifier, we checked the predicted miRNAs against the published literature. Even with very limited data available about miRNAs in urine, we were able to validate a few predicted excretory miRNAs against the literature published since the time of our initial data collection. For example, has-let-7b, has-miR-1270, has-miR-1275, and has-miR548j were predicted to be excreted by our classifier. These miRNAs were experimentally detected in urine since we collected our training datasets, hence determined to be potential biomarkers. Deregulated expressions of hsa-let-7b, has-miR-1270 and has-miR-1275 have been associated with cancer and/or its progression [64,65,66,67]. Our classifier has successfully identified these miRNAs, which can hence be used as potential biomarker miRNAs in urine. The advantage in having a prediction capability like ours is that we can use antibody-based method

to fish out candidate miRNA biomarkers rather than blindly search for urinary miRNAs with differential abundances in cancer patients *versus* healthy people's urine.

An enrichment analysis of pathways targeted by the 138 miRNAs was performed to determine whether some pathways may be significantly affected by these 138 predicted excretory miRNAs. The analysis was carried out by using mirPath [ref??] to determine the enriched pathways that are annotated in KEGG [68]. Interestingly, the analysis reveals that the excretory miRNAs enrich a number of cancer-related pathways (Figure. 2). MAPK signaling was one of the most enriched pathways, as well as the p53 signaling and focal adhesion. These pathways are major contributors to the cellular changes accompanying the transformation of cells from their normal state to a cancerous state.

Discussion

Distinguishing features of urine excretory miRNAs

At the time of feature collection, we collected the features that were thought to be pertinent to the problem based on the known characteristics of miRNAs. For example, extracellular miRNAs are found in circulation in two different ways. MiRNAs can either be transported by larger biomolecular complexes such as microvesicles or exosomes, or by association with Ago proteins. A recent study showed that strand bias selection exists for miRNAs in incorporation into the RISC complex, and highly expressed strands tend to have nucleotide G-bias and U-bias at the 5' end [69]. This observation suggests that miRNAs enriched with G and U nucleotides at the 5' end are more likely to bind to the Ago2 protein, forming a RISC complex. Thus, we included the nucleotide composition for the 3' end and 5' end of the miRNAs. For the same reason, we have also considered dinucleotide compositions.

The mechanism of how miRNAs are selected for exosome loading remains to be understood [70,71]. It is clear that miRNA loading in microvesicles, however, is non-random. A

recent study showed that microparticles exhibit highly distinct binding patterns with miRNAs, suggesting the existence of a sophisticated mechanism for miRNA selection to transport certain miRNAs out of cells [72]. Hence the binding and transport mechanisms may play a pivotal role in determining whether a miRNA is deemed to be excretory or not. While the mechanism is not understood, it is possible that the binding or recognition signal could be structure-based or motif-based. Thus sequence-based features such as palindromic sequences and structure-based features such as secondary structure content were included in our feature collection.

Among the features we have examined, the most discriminating feature was the nucleotide G frequency in the first segment of miRNAs, i.e., the 6-7 nucleotides of the 5' end of miRNAs. This suggests that the Ago-binding of miRNAs could be an important factor in determining if a miRNAs is excretory traits or not. MiRNAs can be bound to Ago proteins, such as Ago2, Ago1, Ago3 and Ago4, as part of the RNAi silencing complex. Evidence suggests that the miRNA binding with the Ago complex protects miRNAs from RNAse degradation, thus providing its stability in biofluids [58,73]. It could be due to this property that the bias at the 5' end is one of the discriminant features in identifying excretory miRNAs. Other highly ranked features include palindromic patterns and structure-based features (Table 1). While it is unclear the precise mechanism how these features influence if a miRNA is excretory or not, these features may contribute to the recognition and loading unto exosomes.

Urine contains more miRNAs than expected

Our classifier predicted 138 out of the 325 miRNAs to be excretory. The overall result is encouraging for two reasons: (a) it suggests that urine contains more miRNAs than suggested by the published literature. With an ever growing number of human miRNAs in miRBase, we expect that the number of miRNAs identified in urine will continue to increase; and (b) the enriched

pathways by the targets of excretory miRNAs being highly cancer-related suggest that urine is a rich source of miRNA biomarkers for cancer.

Materials and Methods

a. Collection of datasets

The sequences of all mature miRNAs were collected from mirBase [74,75,76,77]. All miRNAs that were found in blood circulation and in urine were collected through an exhaustive literature search [48,78,79,80,81]. The positively excreted dataset is defined as the ones that are found in urine by experimental studies; the unknown dataset is defined as the ones that were found in blood circulation but not in urine; the negative dataset is defined by the ones that were neither found in blood nor in urine. The positive and negative miRNAs sets were randomly divided into the training and testing set with the ratio of 4:1 to ensure the coverage of as many miRNA types as possible in the training set. The total number of miRNAs used in the training set is 192 positives and 128 negatives, in the testing set 42 positives and 28 negatives, and unknown results in 328 miRNAs in total.

b. Feature collection & feature selection

Features based on the mature miRNA sequences were collected, and a total of 10 were collected [82,83,84,85,86,87]; the feature details are given in Table S1. The features were filtered to include relevant details and to remove redundancy. A Pearson's Correlation Coefficient was used to determine the relevance of each feature to the excretory/non-excretory classification problem. A correlation measure was calculated between features to determine the redundancy; two features with a correlation coefficient greater than 0.95 will have one removed.

Pearson's linear correlation coefficient, defined in (Equation 1), was calculated for each value and sorted in descending order. Correlation coefficient was also calculated for each feature against all other features, and the less relevant feature was removed if the correlation coefficient was > 0.95. The final list of features was selected based on a 5-fold cross validation SVM by adding one feature at a time until the average accuracy no longer improved.

$$p(x,y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$, (Equation 1)

c. Predicting excretory miRNAs

An SVM-based classifier was used to classify the excretory miRNAs from non-excretory miRNAs. The selected features were used to train a classifier using the RBF-kernel [63]. The RBF was used for its ability to classify using non-linear classification as well as its simplicity due to a relatively small number of parameters. The optimal parameter set, defined by the highest agreement between the desired values and the trained output values, was selected by screening all possible sets of parameters. Using the optimal feature list and the parameter set, a RBF-SVM classifier was trained. The performance was measured using the following values: TP=True Positive, TN=True Negative, FP= False Positive, FN=False Negative, sensitivity = TP/(FP+FN), specificity = TN/(TN+FP), Accuracy (ACC)= (TP+TN)/N, and Matthew's correlation coefficient (MCC) = (TPxTN-FPxFN)/√((TP+FN)(TP+FP)(TN+FP) (TN+FN)).

The performance of the classifier was calculated on the training set and on an independent testing test.

Concluding Remarks

We demonstrated that excretory miRNAs could be accurately predicted by a SVM-based classifier. This classifier can be used to aid in biomarker search in urine by providing a targeted search (versus blind search). Rather than using a comprehensive miRNA profiling approach, this classifier can be used to facilitate targeted searches from a list of miRNAs of interest. In doing so, a sensitive method, such as RT-PCR, could be used to detect miRNAs that may otherwise be missed by the whole-genome profiling method. It is clear from this study that a nucleotide composition, especially G and U, and palindromic sequences are important determinants in aiding in excretion. Based on our finding, we conclude that the 5' end bias existing in strand selection for RISC complex is one of the major contributors for miRNA excretion. Palindrome sequences may also be involved in exosome/microvesicle recognition and loading. Moreover, our prediction on an unknown set of miRNAs reveals that there is most likely a higher number of miRNAs in urine than the current literature suggests and that many of these miRNAs targets are enriched in disease-implicated pathways. Our study further corroborates that urine is an enriched source of miRNAs and an ideal source of biofluid for miRNA biomarker search.

Acknowledgements: We thank Dr. Dave Puett for critical reading of the manuscript and making insightful comments which have helped to improve the presentation of the paper. We thank for a seed grant from the Offices of the UGA President and Vice President for Research at the University of Georgia.

Figures and Tables

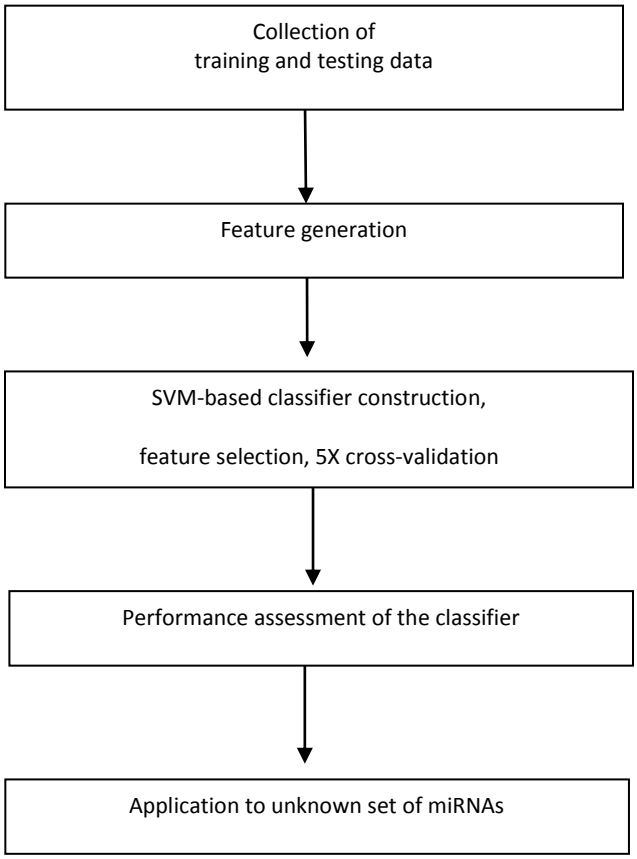


Figure 1: A schematic of our implemented study

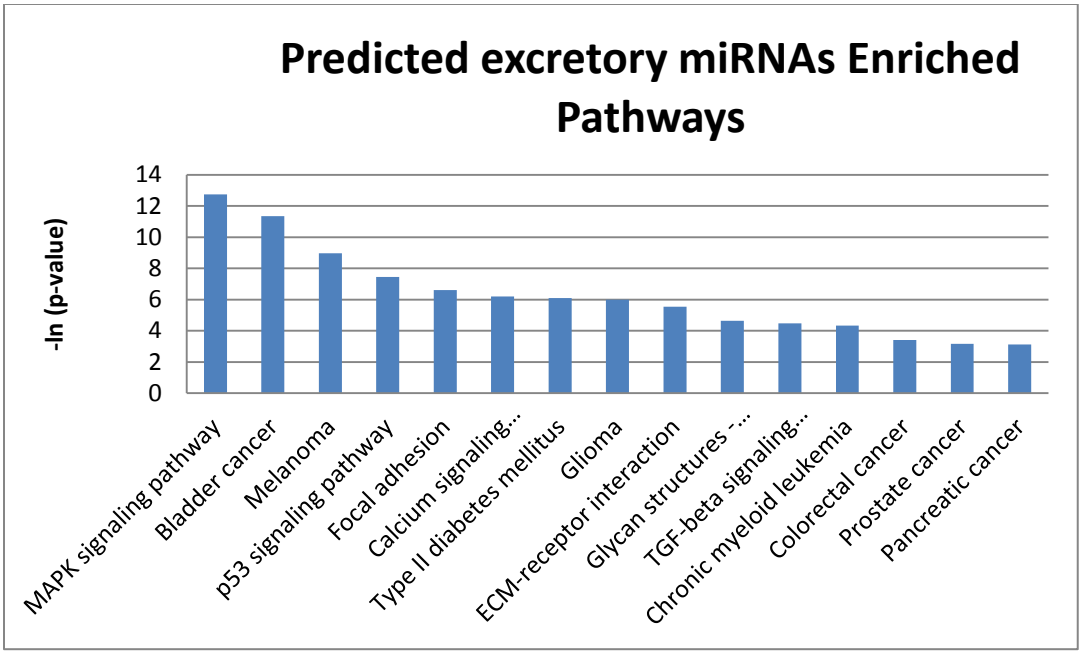


Figure 2: The miRNA target pathway enrichment analysis. The KEGG pathway enrichment analysis of the positively predicted set of miRNAs using DIANA-mirPath. X-axis indicates the pathways modulated by miRNAs; Y-axis indicates the enrichment value in each KEGG pathway, which is represented by the negative natural logarithm of the p-value.

Table 1: The list of features used in training the classifier.

Rank	Feature	Description
1	Window 1 compG	G composition in 5’ end
2	Branches	Secondary structure
3	Palindrome >3 nc	No. of palindromes
4	Di-nucleotide composition	GU
5	Mature miRNA composition	Composition of C
6	Di-nucleotide composition	CU
7	Length	Length of sequence
8	Maxhelix	Secondary structure

9	Di-nucleotide composition	CC
10	Di-nucleotide composition	GG

The first column indicates the rank in order of features; the second and the third columns represent the feature and its description

Table 2: Performance of the trained classifier.

ID	TP	TN	FP	FN	SEN	SP	ACC	MCC
TRAIN SET	180	65	63	12	0.937	0.5078	0.76	0.26
TEST SET	37	15	13	5	0.88	0.53	0.74	0.2

TP: true positive; TN: true negative; FP: false positive; FN: false negative; SEN: sensitivity; SP: specificity; ACC: accuracy; MCC: Matthew's correlation coefficient.

Table S1. Features Collected.

Feature Number	Description	Information
1	Nucleotide Composition (1)	Locally Calculated
2	Length(1)	Locally Calculated
	Dinucleotide Composition(16)	
3	Restriction Site (1)	EMBOSS 6.3.1:restrict[88]
4	Heat Energy Required for Denature (101)	ViennaRNA 1.8.4[85]
5	Palindrome (1)	EMBOSS 6.3.1:palindrome[88]
6	Composition of nucleotide at 5', mid, 3' ends (12)	Locally Calculated
7	Melting Profiles (4)	DINAMELT[87]
8	Pairing Bases (2)	RNAFOLD[89]
9	Free Energy of Structure (1)	RNAFOLD[89]
10	Secondary Structure (14)	MCFOLD[86]

Table S2. The Performance of Classifiers in Feature Selection.

Number of Features Used in Cross Validation	Average Accuracy	C	G
5	72.2772	512	0.078125
9	72.2772	8	0.03125
10	75.7426	32	0.03125
11	75.2475	32	0.03125
12	75.2475	8	0.3125
14	74.7525	32768	0.00048
15	74.25	32768	0.00048
20	73.76	8	0.03125
25	72.77	8	0.03125

Performance of models based on selected features with optimal set of parameters. (C: cost parameter; G: gamma parameter in kernel function)

Table S3. Differentially Expressed miRNAs in Cancer.

let-7	hsa-miR-198	hsa-miR-9
hsa-miR-101-1	hsa-miR-199a	hsa-miR-92
hsa-miR-106a	hsa-miR-19a	hsa-miR-95
hsa-miR-106b	hsa-miR-19b	hsa-miR-96
hsa-miR-10b	hsa-miR-200a	
hsa-miR-124a-1	hsa-miR-200b	
hsa-miR-124a-3	hsa-miR-200c	
hsa-miR-125a	hsa-miR-203	
hsa-miR-125b	hsa-miR-205	
hsa-miR-125b1	hsa-miR-20a	
hsa-miR-126	hsa-miR-20b	
hsa-miR-133b	hsa-miR-21	
hsa-miR-135b	hsa-miR-210	
hsa-miR-139-5p	hsa-miR-212	
hsa-miR-140	hsa-miR-214	
hsa-miR-141	hsa-miR-219-1	
hsa-miR-143	hsa-miR-220	
hsa-miR-145	hsa-miR-221	
hsa-miR-145b	hsa-miR-222	
hsa-miR-146	hsa-miR-224	

hsa-miR-15	hsa-miR-24-2	
hsa-miR-150	hsa-miR-27b	
hsa-miR-155	hsa-miR-28	
hsa-miR-15a	hsa-miR-299-5p	
hsa-miR-15b	hsa-miR-29b-2	
hsa-miR-16	hsa-miR-30a-5p	
hsa-miR-17	hsa-miR-31	
hsa-miR-17-3p	hsa-miR-32	
hsa-miR-17-5p	hsa-miR-33	
hsa-miR-17-92	hsa-miR-340	
hsa-miR-18	hsa-miR-342	
hsa-miR-182	hsa-miR-368	
hsa-miR-183	hsa-miR-372	
hsa-miR-18a	hsa-miR-373	
hsa-miR-18b	hsa-miR-378	
hsa-miR-191	hsa-miR-421	
hsa-miR-192	hsa-miR-497	
hsa-miR-195	hsa-miR-658	
hsa-miR-197	hsa-miR-763-3p	

Table S4. MiRNAs Predicted to be Excreted by the Classifier.

hsa-let-7b	hsa-miR-330-3p	hsa-miR-586	hsa-miR-548k
hsa-let-7c	hsa-miR-338-3p	hsa-miR-587	hsa-miR-1294
hsa-let-7f	hsa-miR-339-3p	hsa-miR-548b-5p	hsa-miR-1303
hsa-miR-17	hsa-miR-346	hsa-miR-588	hsa-miR-1304
hsa-miR-19b	hsa-miR-196b	hsa-miR-550	hsa-miR-1244
hsa-miR-20a	hsa-miR-422-5p	hsa-miR-592	hsa-miR-1250
hsa-miR-32	hsa-miR-424	hsa-miR-595	hsa-miR-1254
hsa-miR-33a	hsa-miR-449a	hsa-miR-602	hsa-miR-1255a
hsa-miR-92a	hsa-miR-450a	hsa-miR-605	hsa-miR-1262
hsa-miR-196a	hsa-miR-433	hsa-miR-608	hsa-miR-1266
hsa-miR-199a-3p	hsa-miR-453	hsa-miR-612	hsa-miR-1270
hsa-miR-208a	hsa-miR-409-5p	hsa-miR-614	hsa-miR-1275
hsa-miR-139-3p	hsa-miR-489	hsa-miR-616	hsa-miR-1276
hsa-miR-10b	hsa-miR-202	hsa-miR-621	hsa-miR-1278
hsa-miR-182	hsa-miR-493	hsa-miR-630	hsa-miR-1255b
hsa-miR-199b-3p	hsa-miR-432	hsa-miR-650	hsa-miR-1308
hsa-miR-219-5p	hsa-miR-193b	hsa-miR-661	hsa-miR-664

hsa-miR-219-1-3p	hsa-miR-512-3p	hsa-miR-662	hsa-miR-1306
hsa-miR-140-5p	hsa-miR-519e	hsa-miR-411	hsa-miR-1307
hsa-miR-140-3p	hsa-miR-519d	hsa-miR-654-5p	hsa-miR-1538
hsa-miR-153	hsa-miR-517b	hsa-miR-658	hsa-miR-1908
hsa-miR-125a-3p	hsa-miR-517c	hsa-miR-320b	
hsa-miR-129-3p	hsa-miR-522	hsa-miR-1468	
hsa-miR-154	hsa-miR-501-3p	hsa-miR-1301	
hsa-miR-190	hsa-miR-503	hsa-miR-1185	
hsa-miR-193a-3p	hsa-miR-508-3p	hsa-miR-886-5p	
hsa-miR-206	hsa-miR-509-3p	hsa-miR-450b-5p	
hsa-miR-320a	hsa-miR-532-3p	hsa-miR-874	
hsa-miR-106b	hsa-miR-559	hsa-miR-190b	
hsa-miR-219-2-3p	hsa-miR-563	hsa-miR-885-3p	
hsa-miR-299-5p	hsa-miR-567	hsa-miR-760	
hsa-miR-301a	hsa-miR-569	hsa-miR-939	
hsa-miR-99b	hsa-miR-574-5p	hsa-miR-942	
hsa-miR-363	hsa-miR-574-3p	hsa-miR-1179	
hsa-miR-302b	hsa-miR-576-3p	hsa-miR-1180	
hsa-miR-369-5p	hsa-miR-577	hsa-miR-548j	
hsa-miR-372	hsa-miR-579	hsa-miR-1285	
hsa-miR-373	hsa-miR-583	hsa-miR-1287	
hsa-miR-375	hsa-miR-584	hsa-miR-1291	

CHAPTER 4

GENOME-WIDE TRANSCRIPTION DATA ANALYSIS IN CO-LOCALIZED REGIONS OF BREAST CANCER

BACKGROUND

Recent development in sequencing technologies lead to another field denoted as transcriptomics [5]. The idea is to sequence the whole RNAs extracted from a cell of interest. There are many sequencing approaches to obtain transcriptomics. The first and perhaps the most commonly used technique is called the RNA-seq [90]. Instead of sequencing the DNA of a genome, the high-throughput sequencing technology was used to sequence the cDNA to obtain a comprehensive profiling of transcriptional activity [90]. By sequencing the whole transcripts, the RNA-seq allowed the researchers to identify the activity of gene expression.

In past few years, a new development in transcriptomics emerged. The Global Run-On Sequencing (GRO-seq) was developed to sequence the nascent RNAs that are transcriptionally engaged with RNA polymerase[91]. Since its introduction, this methodology has been widely used in transcriptome studies [92,93,94,95]. Unlike the traditional –seq methods, GRO-seq provides an insights to the actively transcribing DNA regions, thus it is tremendously useful in observing the changes of transcriptional activity after an inducement.

Using GRO-seq technique, the transcriptomics was published that compared the different time points of transcriptional activity after estrogen (E2) inducement in breast cancer cell line, MCF-7[95]. A lifetime exposure to estrogen is speculated to be risk factor in breast cancer development, and once breast cancer is established, it relies on estrogen to grow [96,97,98]. Naturally, examining the gene transcriptional activity triggered by the E2 is an interest to many researchers.

With publicly available GRO-seq datasets, our interest is to examine the relationships between the co-localized genomic regions and its transcriptional activity after E2 inducement. Two distal genomic regions can co-localize or come in close proximity [99]. This phenomenon has been highly debated. A study published in 2008 found two genes, TFF1 and GREB1, on two distal genomic locations to co-localize after E2 inducement [8].This study has documented two interchromosomal locations migrating within 15-60 minutes after E2 addition. It not only revealed long-range chromosome pairing interactions, but also raised questions on how and what might trigger the specific pairing. The study was repeated by an independent lab but could not obtain the same results; they could not observe any evidence for colocalization [9].

We decided to investigate this matter. Utilizing the publicly available GRO-seq transcriptome and the ChIA-PET interactome data for MCF-7 breast cancer cell lines, a genome-wide colocalization sites are examined in respect to their transcription activity [95,100]. As aforementioned, GRO-seq provides sensitive measurements on transcription activities occurring at the time. In the interactome study, ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag sequencing) was used to extract genome-wide co-localized regions in MCF-7 breast cancer cell lines after E2 inducement. By examining and analyzing the two datasets, we hope to find evidence to link the transcriptional activity and DNA-DNA interaction in genomic regions which spatially co-localize.

Materials & Methods

Data Processing

Datasets were collected from the two studies of genome-wide interactome and transcriptome data in breast cancer cells [95,100]. For transcriptome data, only the time points 0, 10 minutes, and 40 minutes after E2 minutes are considered. This is due to other secondary indirect effects that may be influencing the transcriptional activity after 40 minutes of E2

inducement. Each genomic location was extended by 150bp and the peak score was calculated by counting the tags. The data were normalized by the total read count for each replicate between the different time points. The differential expression was calculated by the Equation 1.

$$Fold\ Change = log_2\left(\frac{Peak\ Score_{time}}{Peak\ Score_{control}}\right),$$ (Equation 1)

To examine how data looks, bins were categorized into the following categories:

1. Decrease: the expression continues to decrease
2. Increase: the expression continues to increase
3. Peak at 10m: the expression is at peak at 10 minutes and decrease
4. Peak at 40m: the expression is comparable at time 0-10m followed by a sudden increase at time 40m

For interactome data, there were 2185 duplex intra chromosomal interaction available and 827 duplex inter chromosomal interaction available. Bad mapped interactions were removed. An interaction was considered to be badly mapped if there were unusually high number of interactions mapped to the region. This could be contributed by the similarity in sequences, such as in repeated regions. 875 duplex-intra-chromosomal interactions and 3 duplex inter chromosomal interacts were left after filtering out the badly mapped interactions. Due to highly false positives occurring in interactome data, only high quality interactions are considered initially. For example, the top 50 intra interaction sites have a 90% reproducible rate. The reproducible rate decreases thereafter.

Data Analysis

To examine the overall transcriptional patter in co-localized regions, interaction sites were categorized into the four categories mentioned above. The Pearson's correlation coefficient and covariance were calculated for E2 regulated genes in interaction sites using R

[101,102]. Box-dot-plot was produced and graphed using R. The correlation coefficient from all E2 regulated genes were used as the background for comparison.

$$Pearson'sCorrelation\ Coefficient = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S_x}\right)\left(\frac{Y_i-\bar{Y}}{S_y}\right), \text{(Equation 2)}$$

$$Covariance(j,k) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij}-\bar{X_j}\right)\left(X_{ik}-\overline{X_k}\right), \text{(Equation 3)}$$

Data Visualization

The GRO-seq and interactome data were visualized in UCSC genome browser. In addition estrogen binding sites and bidirectional promoters were visualized to aid in selecting candidates for FISH experiments [103,104]. The overall method of the study is shown in Figure 4.1.

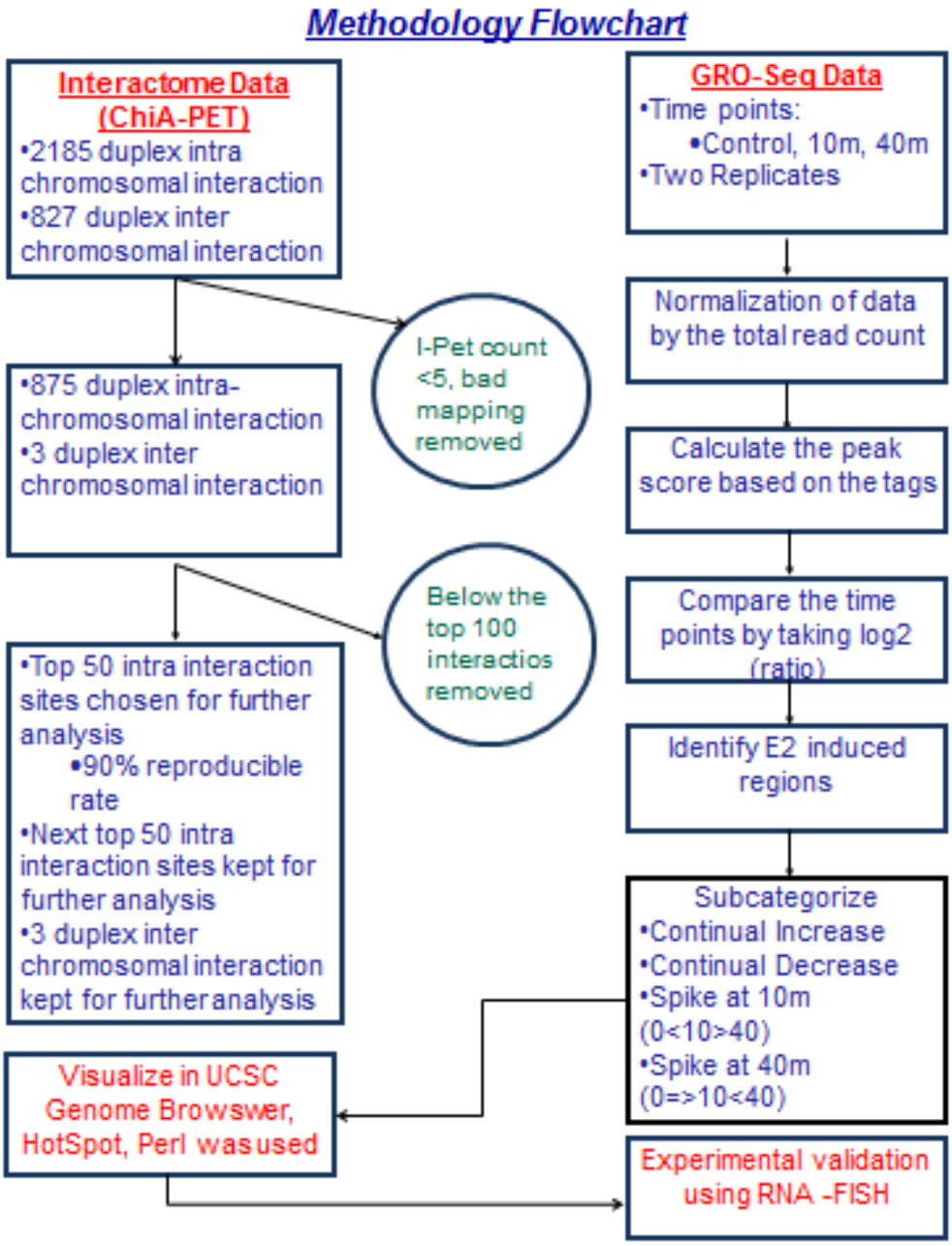


Figure 4.1. The schematic of the study.

Results and Discussion

The fixed sized bins (k=150 bp) were categorized according to their transcriptional activity pattern. The four categories are: decrease, increase, peak at 10m, peak at 40m. For

each bin, the changes in folds were calculated at each time point. The number of bins and interactions belonging to each category is shown in Table 4.1, and interactions sites corresponding to each category are listed in Table 4.2. As shown in the table, the ratio of decrease to one of the increase categories is 1:2; about 1/3 of bins decrease in transcription activity after E2 inducement. Interestingly, when interactions are categorized into the four categories, the ratio significantly changes. The ratio of number of interaction belonging to the decrease category to all other increasing categories is now 1:19; about 0.05 percent of all interactions belong to the decrease. A significantly small number of interactions belonging to the decrease category imply that the interaction regions favor the increase in transcriptional activity.

Table 4.1. Number of bins and interactions belonging to each category

Category	Number of Bins	Number of Interactions
Decrease	62901	32
Increase	37489	362
Peak at 10m	91295	132
Peak at 40m	33545	89

Table 4.2. Interaction sites according after categorizing into four categories

Head Interacting Site	Tail Interacting Site	Category
chr1:65219516-65221912	chr1:65277879-65283042	decrease
chr1:22665859-22669771	chr1:22670528-22673199	decrease
chr1:17719192-17721280	chr1:17728418-17732870	decrease
chr1:19778477-19781668	chr1:19795532-19798359	decrease

chr1:12465711-12470461	chr1:12581327-12583778	decrease
chr1:22670528-22673199	chr1:22824281-22826284	decrease
chr2:39259876-39262045	chr2:39263515-39267466	decrease
chr1:15213276-15218882	chr1:15222288-15227780	decrease
chr1:65219516-65221912	chr1:65233200-65235435	decrease
chr1:42013969-42019783	chr1:42105247-42106998	decrease
chr7:101310946-101316944	chr7:101396989-101407861	decrease
chr1:17719192-17721280	chr1:17775395-17779390	decrease
chr2:39204943-39207877	chr2:39259876-39262045	decrease
chr1:19668674-19671623	chr1:19778477-19781668	decrease
chr1:12436697-12439369	chr1:12465711-12470461	decrease
chr16:79203338-79209489	chr16:79211403-79213650	decrease
chr10:8022489-8031646	chr10:8046939-8049366	decrease
chr16:52676661-52677676	chr16:52680319-52687875	decrease
chr2:39256229-39259095	chr2:39259876-39262045	decrease
chr1:111855438-111862813	chr1:112319359-112324942	decrease
chr1:12436697-12439369	chr1:12581327-12583778	decrease
chr1:19668674-19671623	chr1:19795532-19798359	decrease
chr14:99548435-99554130	chr14:99671435-99675188	decrease
chr2:20534583-20541346	chr2:20645838-20650636	decrease
chr1:15122431-15130822	chr1:15222288-15227780	decrease
chr20:46733623-46739871	chr20:46781649-46788160	decrease
chr1:39380155-39381763	chr1:39414607-39417502	decrease

chr22:28527515-28532184	chr22:28537719-28540961	decrease
chr1:65277879-65283042	chr1:65286306-65289756	decrease
chr1:44299083-44303587	chr1:44305838-44307340	decrease
chr16:52671766-52674563	chr16:52676661-52677676	decrease
chr2:20438279-20438467	chr2:20440000-20442206	decrease
chr17:70265182-70270225	chr17:70276046-70280128	increase
chr2:11555104-11557182	chr2:11586403-11593981	increase
chr16:83877229-83882374	chr16:83895874-83898916	increase
chr21:42658085-42660840	chr21:42663417-42671366	increase
chr15:69175757-69177516	chr15:69180640-69187680	increase
chr2:11586403-11593981	chr2:11596581-11600908	increase
chr8:128879215-128885807	chr8:128950906-128954180	increase
chr2:11555104-11557182	chr2:11596581-11600908	increase
chr6:17494016-17499167	chr6:17500244-17504521	increase
chr8:128939256-128943235	chr8:128950906-128954180	increase
chr16:87514941-87519254	chr16:87520274-87524762	increase
chr16:23054621-23059485	chr16:23066456-23070163	increase
chr17:23873481-23878426	chr17:23886721-23889360	increase
chr20:48775798-48781531	chr20:48810476-48818053	increase
chr19:43479316-43483518	chr19:43499912-43508780	increase
chr9:96580325-96589069	chr9:96792489-96799734	increase
chr1:200339730-200344708	chr1:200346543-200351157	increase
chr12:122112647-122113987	chr12:122116392-122122223	increase

chr10:5737585-5739161	chr10:5741457-5743077	increase
chr10:42839272-42844751	chr10:42924555-42931130	increase
chr16:22997546-23000820	chr16:23054621-23059485	increase
chr8:128879215-128885807	chr8:128939256-128943235	increase
chr6:125560684-125562058	chr6:125563178-125566002	increase
chr6:125516530-125518785	chr6:125563178-125566002	increase
chr3:151918822-151928396	chr3:151935909-151940113	increase
chr14:73312072-73315362	chr14:73317152-73324844	increase
chr5:167653169-167658235	chr5:167674465-167680178	increase
chr16:83895874-83898916	chr16:83900662-83905640	increase
chr8:128943566-128950388	chr8:128950906-128954180	increase
chr19:45661219-45664274	chr19:45675979-45678352	increase
chr15:62599360-62602112	chr15:62701413-62705724	increase
chr5:167674465-167680178	chr5:167710174-167714488	increase
chr8:128939256-128943235	chr8:128943566-128950388	increase
chr8:128950906-128954180	chr8:128990566-128998191	increase
chr6:11159420-11164769	chr6:11173471-11178964	increase
chr3:61762411-61770900	chr3:62136464-62139861	increase
chr6:17494016-17499167	chr6:17580219-17583025	increase
chr11:70283686-70286919	chr11:70328345-70332566	increase
chr8:102547109-102552895	chr8:102581037-102592087	increase
chr10:80557068-80562441	chr10:80584844-80589864	increase
chr2:42863509-42865755	chr2:42867753-42870500	increase

chr7:101374616-101375557	chr7:101396989-101407861	increase
chr3:151935909-151940113	chr3:151953792-151958833	increase
chr7:105003924-105010948	chr7:105061509-105070579	increase
chr8:67596037-67598761	chr8:67784698-67793992	increase
chr3:61646663-61654175	chr3:61762411-61770900	increase
chr6:17500244-17504521	chr6:17580219-17583025	increase
chr5:167674465-167680178	chr5:167707201-167708562	increase
chr6:11152507-11156363	chr6:11173471-11178964	increase
chr17:35856526-35858959	chr17:35964006-35967134	increase
chr3:157742385-157746505	chr3:157835259-157842050	increase
chr17:35716155-35722642	chr17:35729116-35732770	increase
chr16:71503233-71508583	chr16:71513521-71522704	increase
chr2:11555104-11557182	chr2:11558373-11561761	increase
chr19:44312064-44313253	chr19:44330864-44336081	increase
chr16:1275400-1280338	chr16:1287022-1289690	increase
chr5:137170484-137174446	chr5:137176962-137179745	increase
chr3:126033404-126037547	chr3:126038317-126041223	increase
chr1:22665859-22669771	chr1:22670528-22673199	increase
chr10:80557068-80562441	chr10:80590597-80593531	increase
chr17:36095645-36097936	chr17:36108561-36113883	increase
chr1:17719192-17721280	chr1:17728418-17732870	increase
chr14:92567176-92571022	chr14:92578147-92588479	increase
chr21:39200012-39202137	chr21:39207458-39211164	increase

chr19:44312064-44313253	chr19:44316667-44320000	increase
chr17:23883960-23884728	chr17:23886721-23889360	increase
chr6:125563178-125566002	chr6:125567720-125568384	increase
chr1:143697537-143702547	chr1:143703052-143706219	increase
chr8:13054836-13062491	chr8:13176300-13179309	increase
chr1:200339730-200344708	chr1:200366479-200371260	increase
chr14:90926707-90934021	chr14:90952186-90955402	increase
chr2:159286988-159290429	chr2:159303838-159312929	increase
chr5:122206250-122214710	chr5:122215269-122219252	increase
chr3:157742385-157746505	chr3:158012384-158015061	increase
chr3:61665340-61671802	chr3:61762411-61770900	increase
chr6:42149141-42152949	chr6:42195851-42198811	increase
chr14:99626832-99634367	chr14:99671435-99675188	increase
chr5:122206250-122214710	chr5:122230593-122234872	increase
chr21:39203085-39206024	chr21:39207458-39211164	increase
chr2:216552501-216554104	chr2:216555422-216557629	increase
chr1:205137158-205147892	chr1:205149178-205154179	increase
chr11:72170902-72172855	chr11:72173349-72177985	increase
chr2:42421463-42425712	chr2:42867753-42870500	increase
chr4:6331147-6334796	chr4:6465909-6469914	increase
chr10:115674527-115677778	chr10:115805084-115809855	increase
chr17:35729116-35732770	chr17:35856526-35858959	increase
chr8:128879215-128885807	chr8:128943566-128950388	increase

chr6:125516530-125518785	chr6:125560684-125562058	increase
chr9:131556967-131558501	chr9:131574091-131576710	increase
chr19:10457604-10458591	chr19:10473578-10475399	increase
chr4:169830088-169836009	chr4:169846225-169857107	increase
chr15:69160813-69167342	chr15:69175757-69177516	increase
chr22:27538984-27541518	chr22:27548815-27551943	increase
chr22:35912035-35916405	chr22:35922647-35924484	increase
chr16:303976-309825	chr16:315996-318754	increase
chr17:35958309-35959828	chr17:35964006-35967134	increase
chr11:76157078-76162868	chr11:76426226-76431153	increase
chr4:169846225-169857107	chr4:170077278-170081669	increase
chr5:131640518-131643446	chr5:131739627-131745350	increase
chr6:15143373-15147673	chr6:15198010-15201188	increase
chr1:200346543-200351157	chr1:200366479-200371260	increase
chr17:23873481-23878426	chr17:23883960-23884728	increase
chr22:36287951-36293997	chr22:36296054-36297703	increase
chr15:90978286-90984847	chr15:90986477-90990979	increase
chr21:37561273-37564533	chr21:37736741-37744268	increase
chr9:96440662-96443519	chr9:96580325-96589069	increase
chrX:153301406-153302065	chrX:153370700-153375930	increase
chr9:131288028-131291856	chr9:131359774-131364751	increase
chr14:67011332-67016954	chr14:67066349-67074179	increase
chr8:128939256-128943235	chr8:128990566-128998191	increase

chr14:73291670-73295238	chr14:73317152-73324844	increase
chr14:76909743-76914885	chr14:76932784-76938037	increase
chr11:76157078-76162868	chr11:76176400-76177724	increase
chr9:4840884-4844204	chr9:4856244-4859533	increase
chr11:74721316-74725121	chr11:74735567-74741012	increase
chr22:28103390-28104340	chr22:28111014-28114366	increase
chr11:66544138-66546946	chr11:66551032-66553478	increase
chr1:109579322-109585508	chr1:109588381-109590897	increase
chr2:109454293-109458498	chr2:109460732-109465397	increase
chr20:23284038-23287470	chr20:23289686-23293206	increase
chr8:102541194-102545884	chr8:102547109-102552895	increase
chr10:80584844-80589864	chr10:80590597-80593531	increase
chr15:71787013-71791926	chr15:72018744-72021011	increase
chr9:96580325-96589069	chr9:96805363-96806393	increase
chr4:6465909-6469914	chr4:6683508-6687732	increase
chr4:88072636-88076276	chr4:88260182-88264966	increase
chr3:157835259-157842050	chr3:158012384-158015061	increase
chr9:33093829-33100131	chr9:33222508-33226248	increase
chr2:219970238-219974687	chr2:220049111-220052817	increase
chr2:238083184-238086077	chr2:238131347-238132889	increase
chr1:15175555-15179306	chr1:15222288-15227780	increase
chr19:40450747-40458628	chr19:40498454-40503026	increase
chr1:27109896-27113690	chr1:27144914-27150227	increase

chr5:131739627-131745350	chr5:131771887-131776878	increase
chr16:83895874-83898916	chr16:83912782-83917177	increase
chr13:112654175-112659318	chr13:112668798-112675426	increase
chr8:67587205-67590778	chr8:67596037-67598761	increase
chr1:15213276-15218882	chr1:15222288-15227780	increase
chr1:112319359-112324942	chr1:112326739-112333033	increase
chr3:46674965-46679578	chr3:46681338-46685325	increase
chr8:102581037-102592087	chr8:102593776-102598474	increase
chr5:131733539-131737958	chr5:131739627-131745350	increase
chr17:35723063-35727627	chr17:35729116-35732770	increase
chr19:43499912-43508780	chr19:43896802-43899641	increase
chr16:4357220-4365265	chr16:4603278-4609308	increase
chr14:73135644-73139716	chr14:73317152-73324844	increase
chr9:96580325-96589069	chr9:96751601-96755373	increase
chr3:157874538-157877642	chr3:158012384-158015061	increase
chr17:35716155-35722642	chr17:35856526-35858959	increase
chr3:157742385-157746505	chr3:157874538-157877642	increase
chrX:153301406-153302065	chrX:153420383-153424387	increase
chrX:153420383-153424387	chrX:153530976-153534094	increase
chr17:23886721-23889360	chr17:23982094-23990154	increase
chr10:42839272-42844751	chr10:42895336-42898467	increase
chr3:57998158-58004569	chr3:58054688-58060515	increase
chr3:157835259-157842050	chr3:157874538-157877642	increase

chr15:69160813-69167342	chr15:69180640-69187680	increase
chr1:65219516-65221912	chr1:65233200-65235435	increase
chr13:112668798-112675426	chr13:112684702-112688859	increase
chr6:125560684-125562058	chr6:125567720-125568384	increase
chr22:29015786-29019678	chr22:29024309-29024671	increase
chr3:66605003-66613332	chr3:66616293-66616819	increase
chr6:158379993-158387395	chr6:158390270-158391123	increase
chr20:36200993-36203290	chr20:36206119-36209987	increase
chr20:46744865-46750910	chr20:46752274-46757426	increase
chr10:95182962-95187682	chr10:95188858-95193292	increase
chr19:4486206-4489017	chr19:4489613-4491267	increase
chr11:70328345-70332566	chr11:70869833-70871817	increase
chr1:200366479-200371260	chr1:200569643-200577503	increase
chr14:90809562-90814294	chr14:90952186-90955402	increase
chr17:23873481-23878426	chr17:23982094-23990154	increase
chr7:101310946-101316944	chr7:101396989-101407861	increase
chr16:22997546-23000820	chr16:23066456-23070163	increase
chr1:17719192-17721280	chr1:17775395-17779390	increase
chr3:155364019-155365910	chr3:155418123-155423115	increase
chr1:15122431-15130822	chr1:15175555-15179306	increase
chrX:153370700-153375930	chrX:153420383-153424387	increase
chr4:15569762-15575070	chr4:15616460-15619901	increase
chr10:105637321-105640235	chr10:105680688-105685424	increase

chr11:66551032-66553478	chr11:66583771-66587004	increase
chr1:205102259-205108474	chr1:205137158-205147892	increase
chr16:83877229-83882374	chr16:83900662-83905640	increase
chr1:22670528-22673199	chr1:22684454-22688887	increase
chr6:32628540-32631562	chr6:32642545-32644356	increase
chrX:109291728-109293212	chrX:109301654-109305434	increase
chr14:75052772-75061906	chr14:75068230-75073709	increase
chr19:43887066-43892916	chr19:43896802-43899641	increase
chr20:45404907-45410383	chr20:45413870-45418180	increase
chrX:109295844-109298368	chrX:109301654-109305434	increase
chr9:33218637-33219688	chr9:33222508-33226248	increase
chr5:167674465-167680178	chr5:167682130-167685760	increase
chr19:40496365-40496707	chr19:40498454-40503026	increase
chr17:35961425-35963582	chr17:35964006-35967134	increase
chr19:40498454-40503026	chr19:40733754-40737102	increase
chrX:153301406-153302065	chrX:153530976-153534094	increase
chr16:87520274-87524762	chr16:87747331-87750012	increase
chr1:200346543-200351157	chr1:200569643-200577503	increase
chr10:115805084-115809855	chr10:115987950-115989714	increase
chr11:116557680-116562609	chr11:116673683-116682940	increase
chr8:128879215-128885807	chr8:128990566-128998191	increase
chr17:35856526-35858959	chr17:35958309-35959828	increase
chrX:130452333-130455391	chrX:130550406-130557104	increase

chr15:99522272-99526803	chr15:99605868-99607181	increase
chr14:76852758-76855131	chr14:76932784-76938037	increase
chr18:55798474-55803210	chr18:55871263-55872469	increase
chr2:238066094-238070177	chr2:238131347-238132889	increase
chr18:53544792-53548856	chr18:53607997-53614251	increase
chr5:95031882-95035154	chr5:95085764-95092518	increase
chr8:128898752-128905209	chr8:128950906-128954180	increase
chr11:70236259-70240546	chr11:70283686-70286919	increase
chr8:128943566-128950388	chr8:128990566-128998191	increase
chr9:96751601-96755373	chr9:96792489-96799734	increase
chr8:128898752-128905209	chr8:128939256-128943235	increase
chr16:83877229-83882374	chr16:83912782-83917177	increase
chr11:70833474-70839681	chr11:70869833-70871817	increase
chr11:66551032-66553478	chr11:66578156-66582809	increase
chr5:95031882-95035154	chr5:95057371-95060966	increase
chr10:5741457-5743077	chr10:5764548-5768922	increase
chr1:149220299-149225319	chr1:149244108-149249099	increase
chr2:42421463-42425712	chr2:42444057-42448415	increase
chr7:140315326-140320365	chr7:140335680-140339415	increase
chr2:11570268-11571171	chr2:11586403-11593981	increase
chr6:158379993-158387395	chr6:158401691-158405050	increase
chr2:11579465-11584603	chr2:11596581-11600908	increase
chr17:23873481-23878426	chr17:23890393-23892854	increase

chr5:122215269-122219252	chr5:122230593-122234872	increase
chr19:44316667-44320000	chr19:44330864-44336081	increase
chr21:42663417-42671366	chr21:42681657-42687605	increase
chr5:167665520-167667029	chr5:167674465-167680178	increase
chr21:42651984-42658015	chr21:42663417-42671366	increase
chr1:997293-999494	chr1:1004688-1005563	increase
chr14:49526287-49532923	chr14:49537978-49543669	increase
chr6:157182696-157186027	chr6:157189512-157192314	increase
chr6:11152507-11156363	chr6:11159420-11164769	increase
chr11:70866261-70867120	chr11:70869833-70871817	increase
chr8:96289274-96289951	chr8:96291343-96294367	increase
chr14:95047849-95049151	chr14:95050529-95054385	increase
chr11:66578156-66582809	chr11:66583771-66587004	increase
chr4:2390323-2392039	chr4:2392761-2394512	increase
chr1:17719192-17721280	chr1:17721952-17726774	increase
chr17:35856526-35858959	chr17:36823232-36829181	increase
chr11:70283686-70286919	chr11:70869833-70871817	increase
chr10:115428111-115433256	chr10:115805084-115809855	increase
chr17:40473722-40478327	chr17:40745619-40746951	increase
chr10:115428111-115433256	chr10:115674527-115677778	increase
chr1:200339730-200344708	chr1:200569643-200577503	increase
chr2:159070857-159077247	chr2:159286988-159290429	increase
chr6:15143373-15147673	chr6:15356645-15363710	increase

chr1:204931386-204934859	chr1:205137158-205147892	increase
chr2:232091582-232094425	chr2:232279080-232285057	increase
chr15:72895937-72898474	chr15:73015386-73018732	increase
chr11:70236259-70240546	chr11:70328345-70332566	increase
chr11:69331590-69335112	chr11:69412113-69418355	increase
chr8:128950906-128954180	chr8:129029676-129035098	increase
chr1:182715912-182722849	chr1:182780749-182783978	increase
chr5:167653169-167658235	chr5:167710174-167714488	increase
chr6:125516530-125518785	chr6:125567720-125568384	increase
chr1:6890914-6893843	chr1:6940226-6945256	increase
chr21:36481142-36483133	chr21:36528577-36529608	increase
chr21:31784870-31788441	chr21:31821302-31825492	increase
chr15:72865363-72871145	chr15:72895937-72898474	increase
chr9:96580325-96589069	chr9:96611821-96614337	increase
chr13:40464751-40466920	chr13:40489292-40490985	increase
chr19:7367152-7370010	chr19:7392203-7393077	increase
chr17:70251055-70255560	chr17:70276046-70280128	increase
chr18:9772034-9776657	chr18:9793979-9795848	increase
chr19:7372948-7375321	chr19:7392203-7393077	increase
chr1:22651239-22654020	chr1:22670528-22673199	increase
chr17:46087118-46091948	chr17:46106738-46112172	increase
chr19:4918797-4924710	chr19:4939259-4945664	increase
chr11:72161598-72162447	chr11:72173349-72177985	increase

chr2:47034927-47040394	chr2:47049037-47052438	increase
chr12:1807718-1811994	chr12:1818935-1821780	increase
chr7:47538553-47541568	chr7:47548161-47551134	increase
chr16:2319314-2321846	chr16:2328373-2334199	increase
chr4:88072636-88076276	chr4:88082654-88084903	increase
chr9:96429844-96435253	chr9:96440662-96443519	increase
chr21:36481142-36483133	chr21:36487431-36489961	increase
chr5:138995575-138998409	chr5:139001733-139006892	increase
chr10:115805084-115809855	chr10:115812588-115814168	increase
chr2:11550764-11552425	chr2:11555104-11557182	increase
chr3:32121185-32124140	chr3:32126801-32128481	increase
chrX:109291728-109293212	chrX:109295844-109298368	increase
chr15:87448704-87453452	chr15:87455773-87462425	increase
chr2:238066094-238070177	chr2:238072441-238080972	increase
chr12:120093011-120093992	chr12:120096180-120101942	increase
chr1:19791579-19793396	chr1:19795532-19798359	increase
chr12:38110966-38115375	chr12:38117184-38125316	increase
chr7:98808118-98812450	chr7:98814259-98820402	increase
chr14:99671435-99675188	chr14:99676721-99680406	increase
chr20:29721379-29726745	chr20:29727484-29734255	increase
chr17:35716155-35722642	chr17:35723063-35727627	increase
chr11:66551032-66553478	chr11:69412113-69418355	increase
chr11:75180545-75183581	chr11:76157078-76162868	increase

chr10:115428111-115433256	chr10:115987950-115989714	increase
chr1:6940226-6945256	chr1:7427756-7431867	increase
chr11:72173349-72177985	chr11:72614518-72618412	increase
chr7:84001349-84004475	chr7:84336549-84339808	increase
chr15:71787013-71791926	chr15:72102189-72105188	increase
chr6:43879347-43883736	chr6:44183738-44189637	increase
chr9:130995805-130998554	chr9:131288028-131291856	increase
chr17:4097759-4101656	chr17:4379738-4387645	increase
chr2:159070857-159077247	chr2:159303838-159312929	increase
chr11:69221008-69223547	chr11:69412113-69418355	increase
chr14:74455942-74461985	chr14:74604826-74608975	increase
chr1:12436697-12439369	chr1:12581327-12583778	increase
chr17:57153739-57159960	chr17:57290288-57299016	increase
chr17:35723063-35727627	chr17:35856526-35858959	increase
chr1:19668674-19671623	chr1:19795532-19798359	increase
chr9:94466010-94469457	chr9:94565159-94571120	increase
chr1:15122431-15130822	chr1:15222288-15227780	increase
chr12:1818935-1821780	chr12:1908978-1918253	increase
chr16:1275400-1280338	chr16:1366512-1369301	increase
chr3:33733145-33737717	chr3:33814399-33815689	increase
chr7:101310946-101316944	chr7:101374616-101375557	increase
chr2:217903362-217905958	chr2:217958303-217962121	increase
chr19:2476020-2477534	chr19:2529492-2535166	increase

chr2:238072441-238080972	chr2:238131347-238132889	increase
chr4:6465909-6469914	chr4:6519870-6525333	increase
chr19:4918797-4924710	chr19:4967763-4975923	increase
chr5:131739627-131745350	chr5:131787135-131793057	increase
chr5:167674465-167680178	chr5:167719166-167720496	increase
chr2:216514914-216518270	chr2:216555422-216557629	increase
chr16:1366512-1369301	chr16:1403867-1406718	increase
chr1:39380155-39381763	chr1:39414607-39417502	increase
chr14:103228644-103231155	chr14:103263654-103266443	increase
chr20:61930017-61933457	chr20:61964069-61967585	increase
chr18:59036780-59041858	chr18:59070463-59074412	increase
chr3:4765215-4771825	chr3:4798861-4800954	increase
chr6:3662510-3665065	chr6:3691036-3696114	increase
chr14:99641566-99645903	chr14:99671435-99675188	increase
chr17:35856526-35858959	chr17:35881481-35886344	increase
chr10:123444-127405	chr10:147667-151376	increase
chr10:42899881-42904508	chr10:42924555-42931130	increase
chr12:49702444-49706536	chr12:49725907-49728215	increase
chr6:42149141-42152949	chr6:42170069-42174022	increase
chr3:15287847-15291893	chr3:15308316-15310012	increase
chr11:68470476-68473432	chr11:68489383-68491060	increase
chr14:90908656-90911499	chr14:90926707-90934021	increase
chr5:167674465-167680178	chr5:167693155-167694693	increase

chr16:71513521-71522704	chr16:71533808-71539065	increase
chr15:87464067-87468262	chr15:87479201-87485979	increase
chr3:15294206-15298881	chr3:15308316-15310012	increase
chr18:9785273-9788283	chr18:9793979-9795848	increase
chr12:1807718-1811994	chr12:1815214-1817479	increase
chr3:15301503-15305143	chr3:15308316-15310012	increase
chr1:18564868-18565692	chr1:18568635-18570952	increase
chr1:94485080-94486233	chr1:94488474-94491456	increase
chr10:78907987-78912138	chr10:78914352-78916080	increase
chr2:238072441-238080972	chr2:238083184-238086077	increase
chr11:74731623-74733647	chr11:74735567-74741012	increase
chr19:1116041-1118251	chr19:1120096-1122738	increase
chr15:87455773-87462425	chr15:87464067-87468262	increase
chr11:66551032-66553478	chr11:66555102-66558100	increase
chr17:35958309-35959828	chr17:35961425-35963582	increase
chr19:18456629-18458303	chr19:18459824-18462402	increase
chr16:87354782-87360148	chr16:87361656-87365136	increase
chr10:99319836-99322316	chr10:99323694-99327194	increase
chr3:158012384-158015061	chr3:158016228-158021463	increase
chr8:102547109-102552895	chr8:102553779-102557717	increase
chr9:131284942-131287475	chr9:131288028-131291856	increase
chr1:178109255-178115369	chr1:178115868-178123648	increase
chr17:70265182-70270225	chr17:70276046-70280128	peak10

chr6:17494016-17499167	chr6:17500244-17504521	peak10
chr1:42013969-42019783	chr1:42027418-42029615	peak10
chr9:96580325-96589069	chr9:96792489-96799734	peak10
chr14:73312072-73315362	chr14:73317152-73324844	peak10
chr1:32067731-32071566	chr1:32080738-32082901	peak10
chr22:27538984-27541518	chr22:27543299-27546082	peak10
chr8:102547109-102552895	chr8:102581037-102592087	peak10
chr1:65219516-65221912	chr1:65277879-65283042	peak10
chr7:101374616-101375557	chr7:101396989-101407861	peak10
chr20:46733623-46739871	chr20:46744865-46750910	peak10
chr1:111815184-111819855	chr1:111855438-111862813	peak10
chr16:82529897-82535542	chr16:82537482-82539139	peak10
chr8:67596037-67598761	chr8:67784698-67793992	peak10
chr3:58054688-58060515	chr3:58062786-58068506	peak10
chr17:35716155-35722642	chr17:35729116-35732770	peak10
chr1:22665859-22669771	chr1:22670528-22673199	peak10
chr1:32080738-32082901	chr1:32116549-32120632	peak10
chr1:51545124-51547096	chr1:51557587-51562827	peak10
chr1:17719192-17721280	chr1:17728418-17732870	peak10
chr8:13054836-13062491	chr8:13176300-13179309	peak10
chr1:19778477-19781668	chr1:19795532-19798359	peak10
chr1:162982906-162984871	chr1:162986866-162995233	peak10
chr1:12465711-12470461	chr1:12581327-12583778	peak10

chrX:130452333-130455391	chrX:130526546-130528331	peak10
chr8:102516727-102520182	chr8:102547109-102552895	peak10
chr2:42087390-42094106	chr2:42115715-42119827	peak10
chr7:872491-875249	chr7:886639-890353	peak10
chr11:72170902-72172855	chr11:72173349-72177985	peak10
chr1:22670528-22673199	chr1:22824281-22826284	peak10
chr8:128879215-128885807	chr8:128943566-128950388	peak10
chr22:35912035-35916405	chr22:35922647-35924484	peak10
chr16:303976-309825	chr16:315996-318754	peak10
chr15:71787013-71791926	chr15:71795040-71796360	peak10
chr1:32067731-32071566	chr1:32116549-32120632	peak10
chr7:156721390-156723331	chr7:156739575-156742841	peak10
chr1:200346543-200351157	chr1:200366479-200371260	peak10
chr15:71769861-71776295	chr15:71787013-71791926	peak10
chrX:153301406-153302065	chrX:153370700-153375930	peak10
chr14:73291670-73295238	chr14:73317152-73324844	peak10
chr14:76909743-76914885	chr14:76932784-76938037	peak10
chr11:66544138-66546946	chr11:66551032-66553478	peak10
chr1:27109896-27113690	chr1:27144914-27150227	peak10
chr17:73918138-73920441	chr17:73926213-73932067	peak10
chr1:15213276-15218882	chr1:15222288-15227780	peak10
chr17:35723063-35727627	chr17:35729116-35732770	peak10
chr1:111849786-111854537	chr1:111855438-111862813	peak10

chr16:313940-315235	chr16:315996-318754	peak10
chr16:4357220-4365265	chr16:4603278-4609308	peak10
chr9:96580325-96589069	chr9:96751601-96755373	peak10
chr3:157742385-157746505	chr3:157874538-157877642	peak10
chr20:48030333-48032926	chr20:48126553-48127855	peak10
chr11:116183124-116185187	chr11:116212001-116213579	peak10
chr1:65219516-65221912	chr1:65233200-65235435	peak10
chr20:46744865-46750910	chr20:46752274-46757426	peak10
chr14:90809562-90814294	chr14:90952186-90955402	peak10
chr1:117524084-117528068	chr1:117609655-117617462	peak10
chr7:101310946-101316944	chr7:101396989-101407861	peak10
chr19:52252747-52254251	chr19:52274522-52276665	peak10
chr7:872491-875249	chr7:892302-894804	peak10
chr14:75052772-75061906	chr14:75068230-75073709	peak10
chr22:38016253-38020700	chr22:38023425-38029441	peak10
chr15:88165846-88178521	chr15:88180895-88185967	peak10
chr1:112229666-112233034	chr1:112233881-112239852	peak10
chr20:47935418-47940792	chr20:48126553-48127855	peak10
chr6:44183738-44189637	chr6:44313324-44314903	peak10
chr1:19668674-19671623	chr1:19778477-19781668	peak10
chr8:128879215-128885807	chr8:128990566-128998191	peak10
chr8:128943566-128950388	chr8:128990566-128998191	peak10
chr9:96751601-96755373	chr9:96792489-96799734	peak10

chr11:72173349-72177985	chr11:72210141-72213141	peak10
chr11:70833474-70839681	chr11:70869833-70871817	peak10
chr1:12436697-12439369	chr1:12465711-12470461	peak10
chr1:42003075-42006521	chr1:42027418-42029615	peak10
chr1:997293-999494	chr1:1004688-1005563	peak10
chr6:11152507-11156363	chr6:11159420-11164769	peak10
chr1:51557587-51562827	chr1:51565833-51569781	peak10
chr1:42027418-42029615	chr1:42032094-42033901	peak10
chr1:17432031-17432949	chr1:17435368-17441467	peak10
chr16:79203338-79209489	chr16:79211403-79213650	peak10
chr1:21450666-21451945	chr1:21453632-21454537	peak10
chr1:17719192-17721280	chr1:17721952-17726774	peak10
chr16:82537482-82539139	chr16:83051605-83054630	peak10
chr2:232091582-232094425	chr2:232279080-232285057	peak10
chr1:7427756-7431867	chr1:7523358-7524562	peak10
chr6:151936060-151937788	chr6:152017702-152023133	peak10
chr1:162949216-162956291	chr1:162982906-162984871	peak10
chr17:70251055-70255560	chr17:70276046-70280128	peak10
chr22:28520554-28525767	chr22:28542817-28544969	peak10
chr1:10716505-10718762	chr1:10733693-10735370	peak10
chr19:4918797-4924710	chr19:4939259-4945664	peak10
chr20:45404907-45410383	chr20:45418551-45425811	peak10
chr1:51533369-51537474	chr1:51545124-51547096	peak10

chr2:10075711-10082330	chr2:10088071-10093335	peak10
chr3:4756865-4759745	chr3:4765215-4771825	peak10
chr21:36481142-36483133	chr21:36487431-36489961	peak10
chr22:30304777-30307819	chr22:30310338-30312724	peak10
chr2:238066094-238070177	chr2:238072441-238080972	peak10
chr16:82537482-82539139	chr16:82541322-82546754	peak10
chr12:52657708-52659469	chr12:52661318-52665017	peak10
chr7:98808118-98812450	chr7:98814259-98820402	peak10
chr17:35716155-35722642	chr17:35723063-35727627	peak10
chr1:6940226-6945256	chr1:7427756-7431867	peak10
chr1:111855438-111862813	chr1:112233881-112239852	peak10
chr1:12436697-12439369	chr1:12581327-12583778	peak10
chr17:57153739-57159960	chr17:57290288-57299016	peak10
chr1:19668674-19671623	chr1:19795532-19798359	peak10
chr1:32067731-32071566	chr1:32176807-32177623	peak10
chr9:94466010-94469457	chr9:94565159-94571120	peak10
chr1:15122431-15130822	chr1:15222288-15227780	peak10
chr20:47935418-47940792	chr20:48030333-48032926	peak10
chr1:42013969-42019783	chr1:42100238-42102437	peak10
chr1:12465711-12470461	chr1:12530298-12532792	peak10
chr7:101310946-101316944	chr7:101374616-101375557	peak10
chr20:46733623-46739871	chr20:46781649-46788160	peak10
chr11:64901630-64906518	chr11:64941140-64944528	peak10

chr3:4765215-4771825	chr3:4798861-4800954	peak10
chr12:49702444-49706536	chr12:49725907-49728215	peak10
chr1:51545124-51547096	chr1:51565833-51569781	peak10
chr17:73904593-73907748	chr17:73926213-73932067	peak10
chr11:67547140-67549348	chr11:67559639-67567074	peak10
chr9:138861798-138864031	chr9:138871522-138877081	peak10
chr22:43483227-43486717	chr22:43493916-43495873	peak10
chr1:10726501-10727876	chr1:10733693-10735370	peak10
chr1:65277879-65283042	chr1:65286306-65289756	peak10
chr22:36612649-36614193	chr22:36617322-36619828	peak10
chr5:176838976-176839344	chr5:176841717-176846185	peak10
chr15:71781577-71784754	chr15:71787013-71791926	peak10
chr4:765730-771170	chr4:772751-777622	peak10
chr16:87354782-87360148	chr16:87361656-87365136	peak10
chr8:102547109-102552895	chr8:102553779-102557717	peak10
chr9:131284942-131287475	chr9:131288028-131291856	peak10
chr15:69175757-69177516	chr15:69180640-69187680	peak40
chr2:11555104-11557182	chr2:11596581-11600908	peak40
chr1:200339730-200344708	chr1:200346543-200351157	peak40
chr3:151918822-151928396	chr3:151935909-151940113	peak40
chr8:128943566-128950388	chr8:128950906-128954180	peak40
chr17:78161205-78165880	chr17:78167625-78174019	peak40
chr8:128950906-128954180	chr8:128990566-128998191	peak40

chr3:61762411-61770900	chr3:62136464-62139861	peak40
chr11:70283686-70286919	chr11:70328345-70332566	peak40
chr8:102547109-102552895	chr8:102581037-102592087	peak40
chr1:65219516-65221912	chr1:65277879-65283042	peak40
chr11:76152983-76156793	chr11:76157078-76162868	peak40
chr3:61646663-61654175	chr3:61762411-61770900	peak40
chr16:71503233-71508583	chr16:71513521-71522704	peak40
chr2:11555104-11557182	chr2:11558373-11561761	peak40
chr1:22665859-22669771	chr1:22670528-22673199	peak40
chr1:51545124-51547096	chr1:51557587-51562827	peak40
chr1:17435368-17441467	chr1:17513466-17514972	peak40
chr1:200339730-200344708	chr1:200366479-200371260	peak40
chr14:99626832-99634367	chr14:99671435-99675188	peak40
chr5:122206250-122214710	chr5:122230593-122234872	peak40
chr1:205137158-205147892	chr1:205149178-205154179	peak40
chr15:69160813-69167342	chr15:69175757-69177516	peak40
chr1:200346543-200351157	chr1:200366479-200371260	peak40
chr15:90978286-90984847	chr15:90986477-90990979	peak40
chr14:76909743-76914885	chr14:76932784-76938037	peak40
chr20:23284038-23287470	chr20:23289686-23293206	peak40
chr8:102541194-102545884	chr8:102547109-102552895	peak40
chr1:205134029-205137021	chr1:205137158-205147892	peak40
chr19:40450747-40458628	chr19:40498454-40503026	peak40

chr3:46674965-46679578	chr3:46681338-46685325	peak40
chr8:102581037-102592087	chr8:102593776-102598474	peak40
chr20:19210869-19213723	chr20:19509438-19513352	peak40
chr8:128950906-128954180	chr8:129186792-129196403	peak40
chr15:69160813-69167342	chr15:69180640-69187680	peak40
chr10:95182962-95187682	chr10:95188858-95193292	peak40
chr1:200366479-200371260	chr1:200569643-200577503	peak40
chr7:101310946-101316944	chr7:101396989-101407861	peak40
chr1:15122431-15130822	chr1:15175555-15179306	peak40
chr20:19172659-19178666	chr20:19210869-19213723	peak40
chr1:205102259-205108474	chr1:205137158-205147892	peak40
chr1:22670528-22673199	chr1:22684454-22688887	peak40
chr20:45404907-45410383	chr20:45413870-45418180	peak40
chr20:19172659-19178666	chr20:19509438-19513352	peak40
chr19:40498454-40503026	chr19:40733754-40737102	peak40
chr1:200346543-200351157	chr1:200569643-200577503	peak40
chr8:128898752-128905209	chr8:128950906-128954180	peak40
chr11:70236259-70240546	chr11:70283686-70286919	peak40
chr8:128943566-128950388	chr8:128990566-128998191	peak40
chr11:72173349-72177985	chr11:72210141-72213141	peak40
chr1:12436697-12439369	chr1:12465711-12470461	peak40
chr6:157182696-157186027	chr6:157189512-157192314	peak40
chr1:51557587-51562827	chr1:51565833-51569781	peak40

chr2:42867753-42870500	chr2:42873448-42876345	peak40
chr1:17432031-17432949	chr1:17435368-17441467	peak40
chr17:40473722-40478327	chr17:40745619-40746951	peak40
chr1:200339730-200344708	chr1:200569643-200577503	peak40
chr1:204931386-204934859	chr1:205137158-205147892	peak40
chr11:70236259-70240546	chr11:70328345-70332566	peak40
chr1:6890914-6893843	chr1:6940226-6945256	peak40
chr1:22651239-22654020	chr1:22670528-22673199	peak40
chr10:8022489-8031646	chr10:8046939-8049366	peak40
chr20:45404907-45410383	chr20:45418551-45425811	peak40
chr1:22670528-22673199	chr1:22678208-22681481	peak40
chr3:170812594-170817945	chr3:170822939-170827484	peak40
chr7:30164220-30170466	chr7:30175261-30180318	peak40
chr12:52657708-52659469	chr12:52661318-52665017	peak40
chr14:99671435-99675188	chr14:99676721-99680406	peak40
chr5:122206250-122214710	chr5:122486889-122491065	peak40
chr22:36612649-36614193	chr22:36724447-36728300	peak40
chr1:12465711-12470461	chr1:12530298-12532792	peak40
chr15:88105019-88109081	chr15:88165846-88178521	peak40
chr4:6465909-6469914	chr4:6519870-6525333	peak40
chr3:71195821-71201666	chr3:71240811-71245663	peak40
chr2:216514914-216518270	chr2:216555422-216557629	peak40
chr3:62101257-62105892	chr3:62136464-62139861	peak40

chr14:99641566-99645903	chr14:99671435-99675188	peak40
chr21:15271632-15277381	chr21:15296497-15301065	peak40
chr1:51545124-51547096	chr1:51565833-51569781	peak40
chr15:69158323-69159769	chr15:69175757-69177516	peak40
chr8:107705153-107706915	chr8:107721554-107728553	peak40
chr2:74977811-74983660	chr2:74997568-75002703	peak40
chr6:44298887-44302276	chr6:44313324-44314903	peak40
chr1:202363702-202369127	chr1:202378993-202382439	peak40
chr1:65277879-65283042	chr1:65286306-65289756	peak40
chr2:238072441-238080972	chr2:238083184-238086077	peak40
chr11:74731623-74733647	chr11:74735567-74741012	peak40
chr20:48614965-48619298	chr20:48620630-48625199	peak40
chr8:102547109-102552895	chr8:102553779-102557717	peak40

Further evidence that co-localized regions and transcription activity may be dependent is obtained from the covariance and correlation analyses. The correlation and covariance were calculated for the following: All E-2 regulated genes; E2 regulated genes in top 100 interaction sites; E2 regulated genes in top 50 interaction sites; all genes in top 875 interaction sites; E2 regulated genes in the top 875 interaction sites. The values are shown in Table 4.3 and Table 4.4.

Table 4.3. Covariance Analysis

Covariance Analyses					

Statistics for	All E2-Regulated genes	Genes in Top 100 Interaction Sites	Genes in Top 50 interaction sites	All Genes in top 875 interactions (include non-E2 reg genes)	E2 Regulated Genes in top 875 interactions
Mean	2.32E+02	30981.028	44789.3477	2778.614	11226.2163
Standard deviation	4.99E+04	55496.796	76386.7606	16161.56	3224188.86
Min	-9.89E+06	-28802.547	-95.8695	-300824.2	-30082415%
25%	-7.74E+02	1347.07	3828.6164	-7.403635	342.0675
50%	-6.29E+00	9376.807	14509.4568	10.25886	231429%
75%	7.78E+02	38535.019	54165.4704	553.3879	9965.3723
Maximum	5.85E+07	510129.071	510129.0715	758902.2	721399.0778

Table 4.4. Correlation Coefficient Analyses

stats	All E2-Regulated genes	Genes in Top 100 Interaction Sites	Genes in Top 50 interaction sites	All Genes in top 875 interactions (include non-E2 reg genes)	E2 Regulated Genes in top 875 interactions
mean	0.0211	0.72	0.7859	0.2451327	0.6144589
Standard deviation	0.0805	0.4572	0.3168	0.6142017	0.5242827

Min	-1	-0.998	-0.2566	-0.9999883	-0.9999617
25%	-0.824	0.6958	0.7042	-0.3018181	0.5489359
50%	-0.027	0.954	0.9633	0.3795782	0.8596274
75%	0.905	0.99	0.993	0.8259218	0.9557231
Maximum	1	1	0.999	1	0.9999937

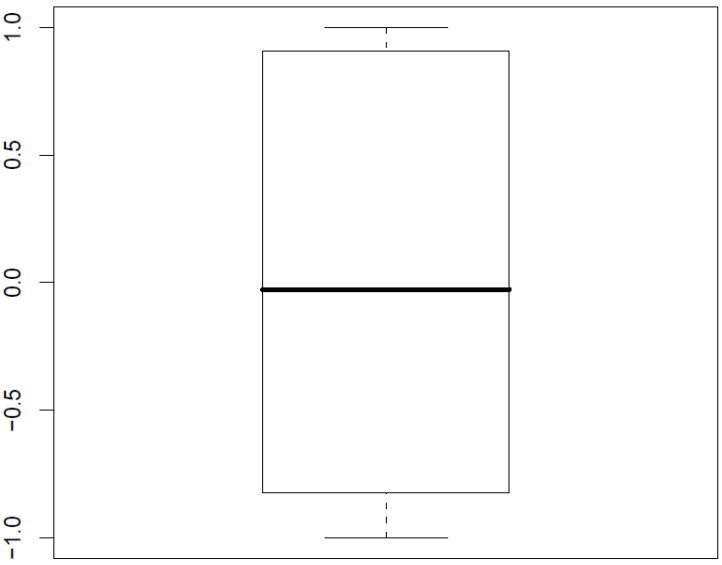
The correlation and covariance was produced by the following. For each gene of interest, its correlation and covariance value was calculated in respect to all other genes in the group, thus producing a matrix of dimension N, where N is a number of genes of interest.

As shown in Figure 4.2, the correlation coefficients of all E2 regulated genes (background) is evenly distributed between -1 and 1, with a mean of approximately 0. In comparison, when the genes in the top 100 interaction sites are examined, the mean of correlation coefficient is 0.72, a very strong positive correlation. When we narrow further and examine the genes in the top 50 interaction sites, the correlation coefficient is at 0.78 (Table 4.4).

A correlation study is also calculated for all genes in the top 875 interaction sites as well as E2 regulated genes in 875 interaction sites. The mean correlation for all genes in interactions examined is 0.245. Contrast to the background, all E2 regulated genes (cc=0.021), where no correlation was observed, a weak correlation is discerned (Table 4.4). This suggest that in this data, some of the interaction sites found E2 inducement that are associated with non-E2 regulated genes may be sites which are co-localized by chance. Movements induced by E2 binding sites may cause other irrelevant genomic locations to co-localize. Because the interactome data does not distinguish which interactions are truly co-localizing to carry out a function, the relatively low correlation may be due to such sites.

When all E2 regulated genes in 875 interactions are examined, a high correlation is observed (cc=0.614) (Table 4.4). While it is still high, the correlation is relatively lower than the correlation calculated from the high quality interaction sites, e.g. Top 50 and Top 100. The reproducible rate for interaction sites decrease after top 100 sites. When false positives of interaction sites are considered, 0.614 is significantly high correlation. The correlation study shows a general trend of E2 regulated genes to be up-regulated in interaction regions after E2 inducement.

A



B

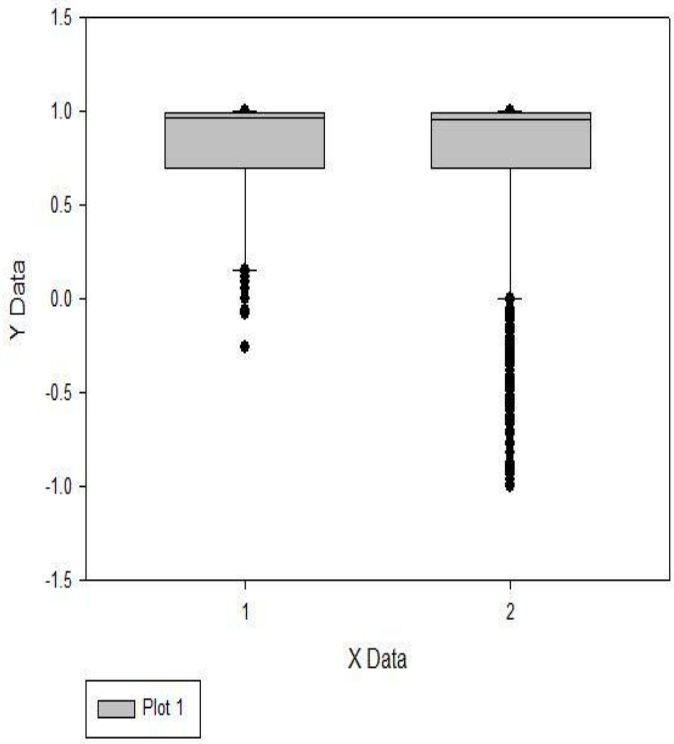


Figure 4.2. Graphs of correlation coefficient values. A(top): Correlation coefficient graph for the background. B(bottom): Correlation coefficient for the top 50 interaction sites (left) and top 100 interaction sites(right).

Such a drastic contrast is observed between the correlations from genes involved in interactions *vs* all genes induced by E2. Genes involved in interactions are highly correlated and are highly transcribed. This corroborates the hypothesis that the interactions may occur to enhance the transcription efficiency. The findings from this study, that the genes involved in interactions are highly transcribed, deduce that genomic regions co-localize to transcribe genes. The purpose of this study is to find genomic locations that are co-localized by analyzing the expression data and the interactome data. Thus, we calculated the correlation coefficient for all

875 interaction sites and checked the regions that are highly induced or repressed. Table 4.4 shows negatively correlated interaction site (Table 4.4 can be found at the end of the chapter).

To find the candidates for RNA FISH experiment, a careful selection was made where interaction sites which exhibit significantly increased or decreased gene expressions after E2 inducement were chosen. The data were visualized in UCSC genome browser and each candidate sites were examined manually [105,106]. The following data were visualized in UCSC genome browser to aid in the candidate selection: GRO-seq transcription data; ChIA-PET data; bidirectional promoter data; ER binding sites [95,100,103,104].

Bidirectional promoter can initiate transcriptions in both strands. We checked whether there are any bidirectional promoter that can be found in between the interaction sites and genes. Table 4.5 lists the bidirectional promoters that are found between genes or at least one gene in the interaction sites.

Table 4. 6. Genes involved in bidirectional promoter

Location	Genes involved in bidirectional promoters	
chr10:115603904-115604170	DCLRE1A	NHLRC2
chr10:115604172-115604257	DCLRE1A	NHLRC2
chr10:115604259-115604534	DCLRE1A	NHLRC2
chr11:46595132-46595892	HARBI1	ATG13
chr11:46595447-46595448	HARBI1	ATG13
chr1:148125028-148125493	HIST3HBE	HIST2H2AC
chr1:150029094-150029647	TDRKH	

chr1:150029788-150030181	TDRKH	
chr1:154519237-154519596	SMG5	TMEM79
chr11:73559854-73560091	C2D3	PPME1
chr1:181707735-181708137	DKFZP564C196	SMG7
chr1:215871023-215871320	GPATCH2	SPATA17
chr12:4517899-4518373	C12orf4	RAD51AP1
chr12:4518002-4518025	C12orf4	RAD51AP1
chr12:81276287-81276616	CCDC59	C12orf26
chr13:47473491-47473877	SUCLA2	
chr15:47700145-47700441	C15ORF33	DTWD1
chr15:47700443-47700663	C15ORF33	DTWD1
chr19:53558782-53559356	THEM143	SYNGR4
chr20:61965742-61966518		TPD52L2
chr2:220145054-220145465	OBSL1	INHA
chr22:40194969-40195597	PHF5A	ACO2
chr2:85683347-85683858	TMEM150A	USP39
chr3:53855123-53855943	CHDH	IL17RB
chr4:88074784-88075258	C4orf36	AFF1
chr4:88074900-88074954	C4orf36	AFF1
chr7:99864336-99864652	ZCWPW1	MEPCE
chr8:125620315-125620621	TATDN1	NDUFB9
chr8:144170666-144171510	LOC100133669	LY6E
chr9:130124536-130124950	TRUB2	COQ4

chrX:100549245-100549789	GLA	HNRNPH2

After the examination of potential sites, 34 interactions that show positive correlation and 1 interaction that show negative correlation are selected for FISH experiment. We did not find convincing evidence to pursue interchromosomal interaction; thus they were filtered out during the process of examination. Figure 4.3-5 shows visualizations of FOXN1, KCNK6, and TFF1 interactions.

FOXN1's significant increase in transcription activity is evident in interaction with *KIAA0100*. KIAA0100 is also induced (Figure 4.3). The genomic distance between the two genes interaction spans over 100k bases, makes an ideal candidate for FISH experiment. The similar pattern can be observed for genes KCNK6 and ACTIN 4 interacts as well as TFF1 and TMPRSS3 genes (Figure 4. 4-5). TFF1 and TMPRSS3 genes are a special interest, as both genes are oncogenes.

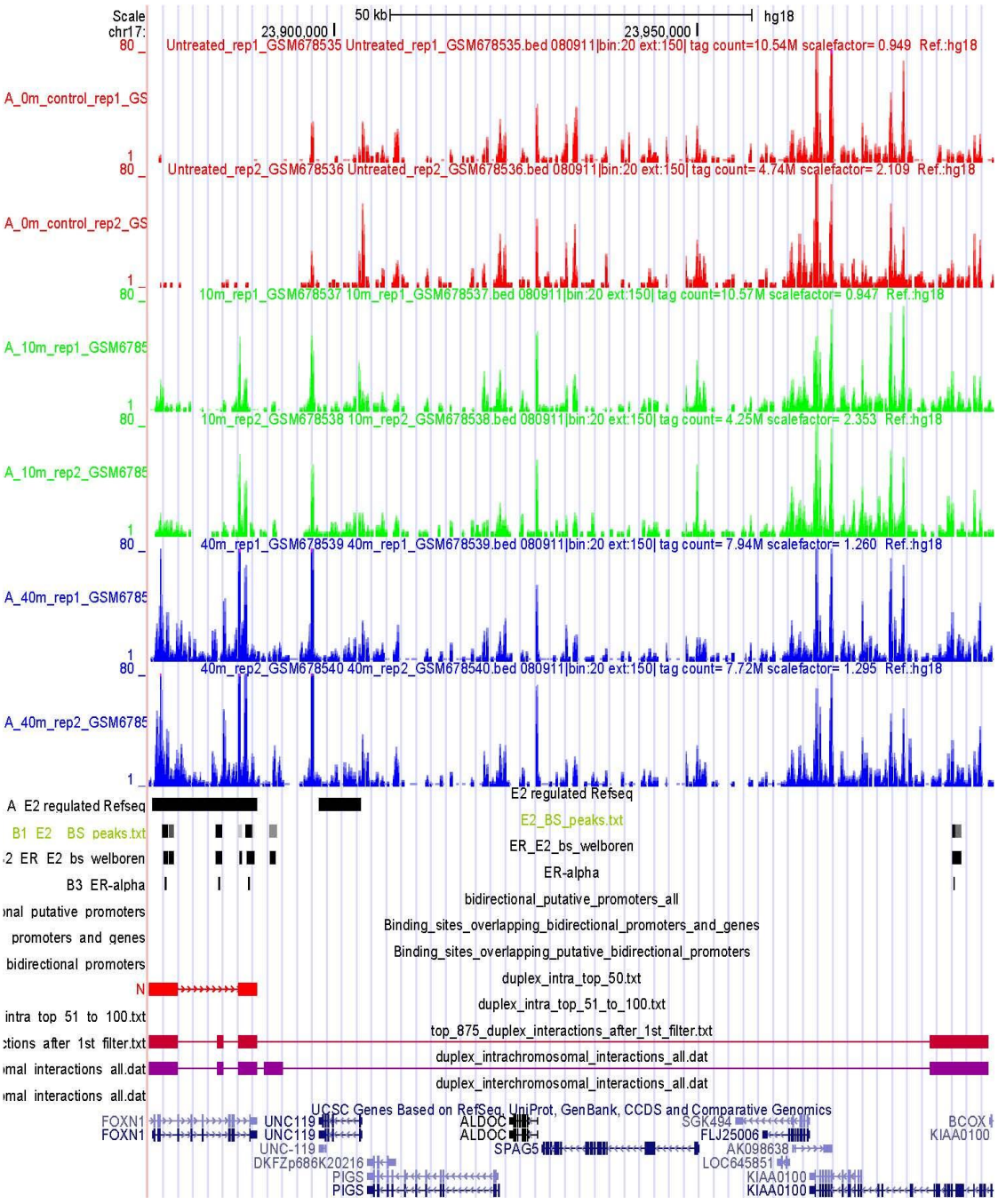


Figure 4.3. The visualization of FOXN1 interaction with KIAA0100

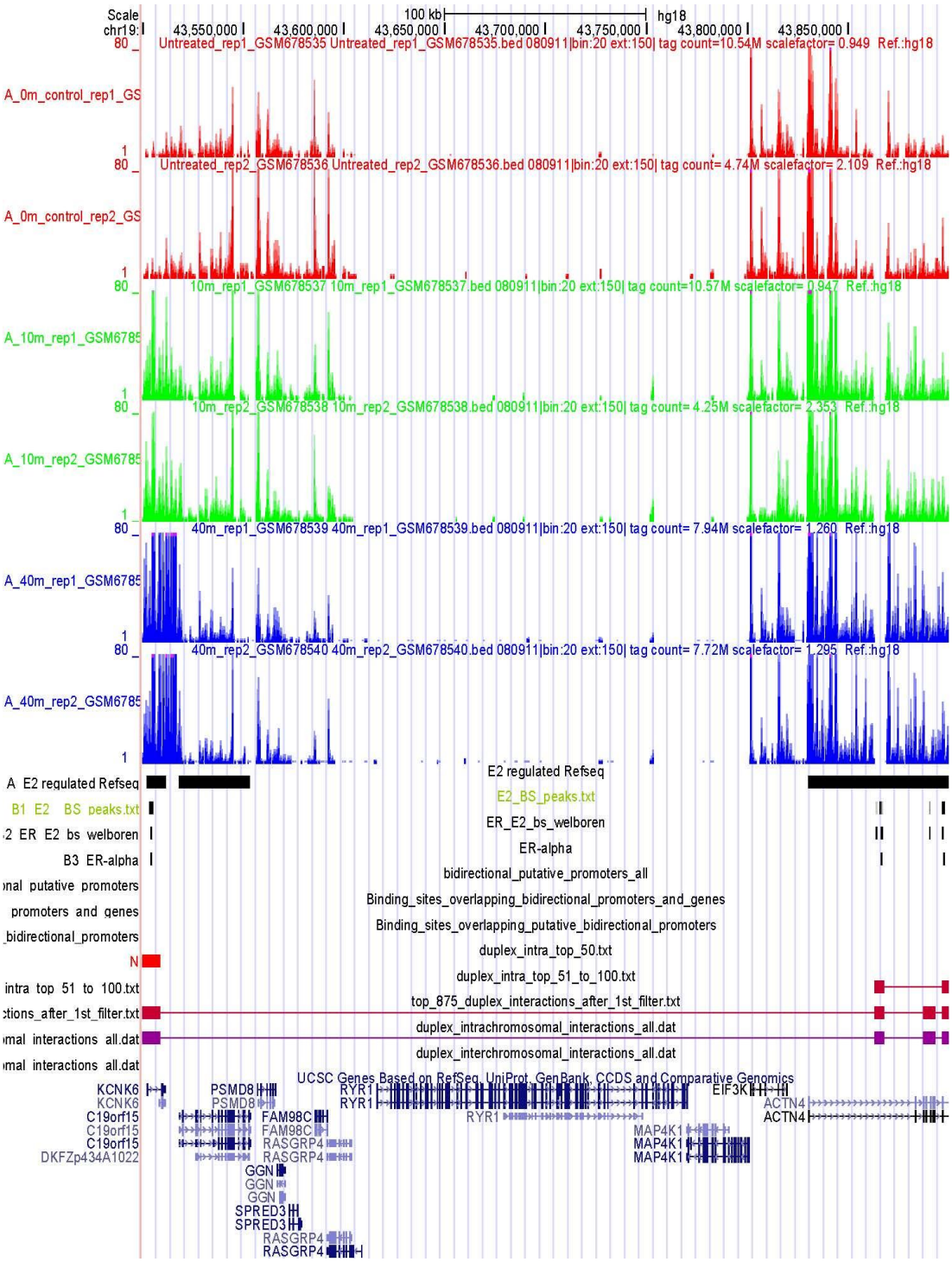


Figure 4.4. The visualization of KCNK6 interaction with ACTIN4

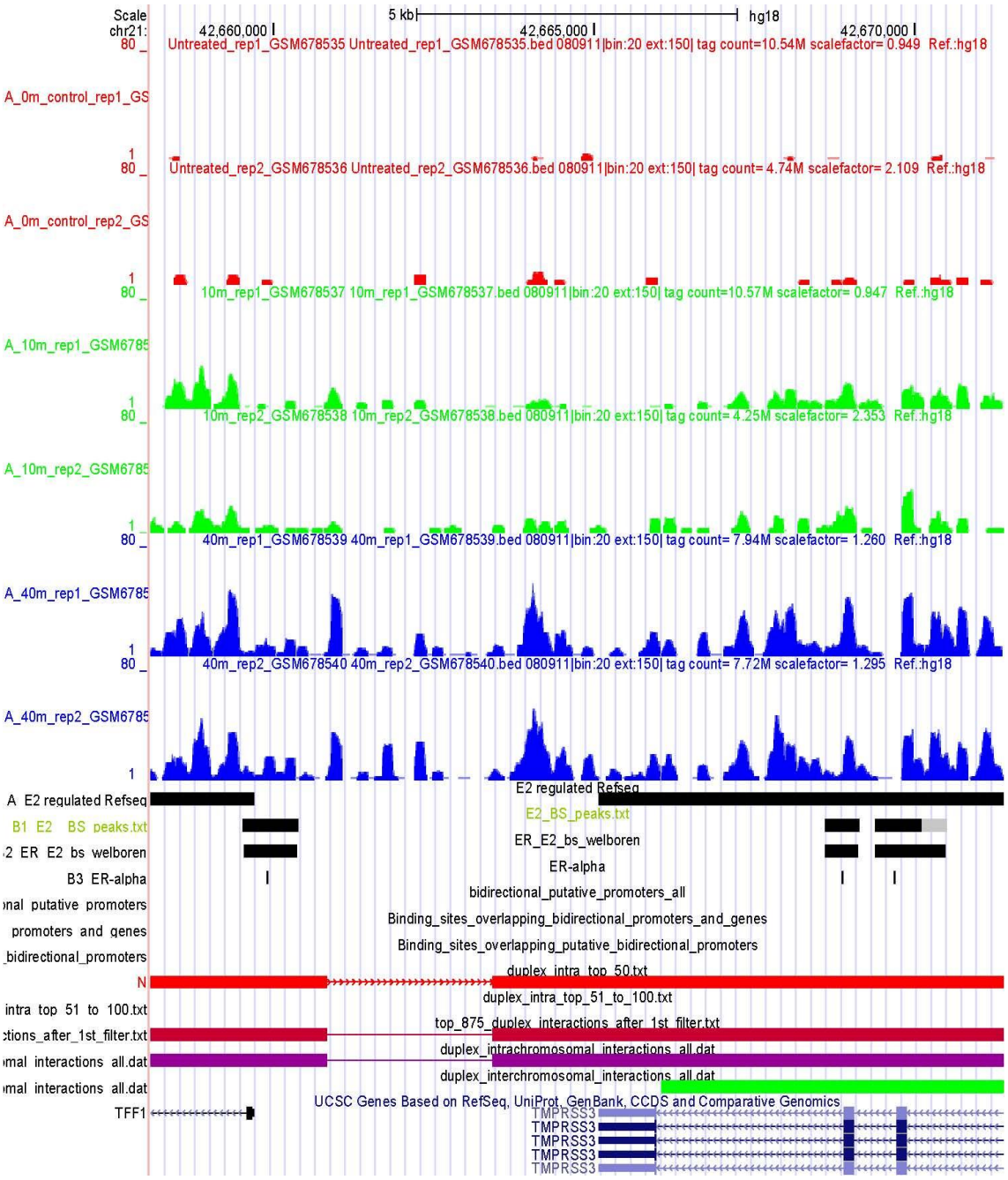


Figure 4.5. The visualization of TFF1 interaction with TMPRSS3

Table 4.4. Interaction regions that show negatively correlated expressions between the time points 0, 10m, 40m.

Notes	head clust	tail clust	UCSC	Gen	head	head	head	head	tail V	tail 1	tail 4	tail 1	Corr Coef	Corr Coef	Inter

	er	er	view range	omic span		10	40	160		0	0	60	f(0,10,40)	f(0,10,40,160)	action No.
	chr2:543362-547038	chr2:553689-560668	chr2:543362-560668	17306	0	2.648445	17.086881	2.241233	3.676903	3.380989	1.595625	1.999083	-0.999919756	-0.742574299	196
	chr16:82517816-82519668	chr16:82537482-82539139	chr16:82517816-82539139	21323	5.679446	3.606395	5.528114	1.999083	4.128311	9.410422	4.404724	3.981667	-0.999819173	-0.177842907	483
	chr18:41634594-41641082	chr18:41692272-41695371	chr18:41634594-41695371	60777	0	3.014717	3.310946	3.984134	11.613939	8.227064	7.745282	9.956634	-0.999435986	-0.712091779	125
	chr16:23054621-23059485	chr16:23456828-23459484	chr16:23054621-23459484	40463	16.061684	106.9799	409.0721	141.4059	10.644494	9.466745	6.592281	11.947747	-0.997971747	-0.810452933	513
	chr5:117371051-117375024	chr5:117498126-117507386	chr5:117371051-117507386	13635	3.676903	3.240124	2.778863	4.727921	0.451408	1.690494	3.813418	1.244579	-0.999091588	-0.652834254	458
	chr1:271098963711690	chr1:271449142715027	chr1:271098962715027	40331	70.189334	66.718114	83.570339	76.660665	252.09063	252.98323	230.7288	261.15757	-0.987045492	-0.633172113	568
	chr10:115668367-115677414	chr10:115731838-115733525	chr10:115668367-115733525	66883	84.453359	35.725279	42.894224	39.5712	0	2.056766	2.217168	1.742904	-0.979752655	-0.965791805	785

	6	0	0												
VTCN1/intergenic	chr1:117524084-117528068	chr1:117609655-117617462	chr1:117524084-117617462	9378	19.09847	35.894887	46.057113	12.693772	145.951	72.353309	56.758202	34.602676	-0.976324057	-0.166205598	775
	chr8:68455784-68459030	chr8:68809093-68812450	chr8:68455784-68812450	356666	4.259585	4.479804	5.498502	1.494976	65.880222	50.827447	35.681404	23.647201	-0.937842623	0.503878223	741
	chr17:16921555-16924903	chr17:16928506-16937356	chr17:16921555-16937356	15801	28.0.061	44.882858	50.018661	34.359776	61.563604	55.786331	47.150991	50.281449	-0.916272021	-0.686895445	850
	chr2:39204943-39207877	chr2:39259876-39262045	chr2:39204943-39262045	57102	152.1555	203.8171	239.595966	167.0487	35.9399	22.624343	22.880551	24.400066	-0.906477478	-0.724000742	792
	chrX:130402905-130407976	chrX:130526546-130528331	chrX:130402905-130528331	1254426	14.904552	10.509224	9.9920061	28.632772	6.4509089	12.650555	19.6584	22.159543	-0.894728381	0.456995505	766
	chr5:105603538-105607413	chr5:105608794-105611763	chr5:105603538-105611763	8225	0	1.324222	0.561695	0	13.681757	0.732544	2.778863	0	-0.894529607	-0.465555975	507
not a significant expressio	chr1:124655711-124700461	chr1:125813327-125833778	chr1:124655711-125833778	1180567	526.2689	336.6618	344.0673	325.896	368.9683	413.1287	453.0925	283.83232	-0.863165854	-0.022541491	313

n															
	chr1:60836236084236	chr1:61381636142145	chr1:60836236142145	58522	0.9684453	1.6904494	12.061617	8.959984	70.788812	50.207669	33.907212	30.873395	-0.861644 03	-0.869596 69	789
	chr1:65219516-65221912	chr1:65233200-65235435	chr1:65219516-65235435	15919	100.84 36	89.088894	176.8088	107.7949	119.8845	113.0377	103.6128	53.777 17	-0.849307 047	-0.007814 393	702
	chr8:128990566-128998191	chr8:129213857-129228383	chr8:128990566-129228383	2378 17	231.9062	256.8704	223.8979	366.206	6.836676	3.888125	17.500 45	8.218664	-0.822586 579	-0.273424 304	748
	chr8:129136752-129142040	chr8:129213857-129228383	chr8:129136752-129228383	91631	137.1197	177.4172	125.7845	310.7106	6.836676	3.888125	17.500 45	8.218664	-0.808359 752	-0.307563 993	663
	chr6:33887694-33894569	chr6:33970322-33977178	chr6:33887694-33977178	89484	0.4514 08	0	0	0.746254	0.968453	24.174 06	70.031 41	50.040 97	-0.757875 248	-0.172273 909	662
	chr8:128950906-128954180	chr8:129186792-129196403	chr8:128950906-129196403	245497	10811.144	11 06.6 82	2038.095	523.0304	307.7335	205.085	164.4794	540.7025	-0.734950 139	-0.844881 255	631
	chrX:130402905-130407976	chrX:130452333-130455391	chrX:130402905-130455391	52486	14.904 52	10.509 24	9.9920 61	28.632 72	18.901 56	21.751 02	38.991 99	53.769 77	-0.682359 932	0.715876 417	208
	chr1	chr1	chr1	13	29	25	27	18	72	13	15	44	-	0.57	59

	7:739181 38-7392 0441	7:739262 13-7393 2067	7:739181 38-7393 2067	929	.234 41	.920 87	.846 93	.673 62	.593 75	4.59 14	6.85 45	.809 79	0.64 1038 025	4029 424	3
	chr1: 4202 7418 - 4202 9615	chr1: 4210 5247 - 4210 6998	chr1: 4202 7418 - 4210 6998	7958 0	29 .2 92	28 .0 05 86	53 .6 83 97	30 .8 66 55	49 .8 10 33	34 .3 17 02	30 .6 55 4	45 .8 06 44	-0.61 3255 937	-0.64 4291 717	12 1
	chr1 1:67 5388 77-6754 5596	chr1 1:67 5596 39-6756 7074	chr1 1:67 5388 77-6756 7074	28 197	6. 58 22 63	2. 42 30 38	13 .1 25 34	2. 48 91 58	37 .3 59 76	52 .6 87 05	41 .6 52 41	56 .2 53 99	-0.60 0720 064	-0.71 4925 644	83 2
	chr1 9:40 4507 47-4045 8628	chr1 9:40 4984 54-4050 3026	chr1 9:40 4507 47-4050 3026	52 279	17 1. 52 46	13 6. 64 81	14 3. 46 17	12 0. 49 6	28 .2 65 96	83 .1 72 27	22 0. 05 67	13 .4 44 91	-0.58 0952 377	0.01 0935 917	56 2
not a signi ficant expr essio n	chr8: 1291 8679 2-1291 9640 3	chr8: 1292 1385 7-1292 2838 3	chr8: 1291 8679 2-1292 2838 3	41 591 1	30 7. 73 35	20 5. 08 5	16 4. 47 94	54 0. 70 25	6. 83 67 6	3. 88 81 25	17 .5 00 45	8. 21 86 64	-0.56 0357 481	-0.29 3841 518	14 6
	chr1 2:96 4850 96-9648 8297	chr1 2:96 6171 57-9662 0495	chr1 2:96 4850 96-9662 0495	13 53 99	49 6. 13 42	32 9. 81 53	36 4. 16 61	29 1. 53 62	3. 22 54 94	6. 25 48 4	14 .7 51 2	10 .4 52 49	-0.55 0544 092	-0.57 1228 927	36 0
	chr4: 1061 0964 4-1061 1489 2	chr4: 1062 6649 7-1062 7058 7	chr4: 1061 0964 4-1062 7058 7	16 09 43	17 .2 27 2	6. 67 74 35	11 .5 29 09	7. 96 58 01	2. 38 83 15	4. 70 52 11	7. 15 39 76	8. 95 99 84	-0.52 6020 847	-0.62 5176 674	40 5
	chr1: 4201 3969	chr1: 4210 5247	chr1: 4201 3969	93 02 9	13 3. 22	11 8. 70	20 1. 93	14 7. 62	49 .8 10	34 .3 17	30 .6 55	45 .8 06	-0.51 4565	-0.48 8737	77 4

	-42019783	-42106998	-42106998		98	08	44	14	33	02	4	44	213	698	
	chr10:115668367-115674146	chr10:115805084-115809855	chr10:115668367-115809855	1414	88	84.453	35.725	79	42.894	24	39.571	2	2.708449	8.227064	61.725	36	27.626	2	-0.454390	058	-0.449557	64	646
	chr10:115668367-115674146	chr10:115674527-115677778	chr10:115668367-115677778	94	11	84.453	35.725	79	42.894	24	39.571	2	2.257041	3.747266	19.658	4	4.730388	-0.447766	388	-0.334661	753	108
	chr15:71787013-71791926	chr15:71795040-71796360	chr15:71787013-71796360	93	47	65.822	63	77.988	03	96.104	42	55.011	88	38.204	99	58.124	81	29.502	4	35.847	34	-0.402327	712	-0.100651	522	395
	chr3:157835259-157842050	chr3:157874538-157877642	chr3:157835259-157877642	42	383	2.774	086	4.479	804	40.913	04	11.694	6	634.93	26	783.318		648.6132		504.453		-0.389872	707	-0.127717	489	687
	chr3:157742385-157746505	chr3:157874538-157877642	chr3:157742385-157877642	13	52	57	90.0	754	5	95.4	284	4	196.0	537	199.1	672	634.93	26	783.318		648.6132		504.453		-0.385033	426	-0.662650	395	647
	chr1:205102259-205108474	chr1:205137158-205147892	chr1:205102259-205147892	45	633	58.0	75	34.1	76	16	39.9	37	71	12.6	9866	248.477		281.1291		944.4626		132.1931		-0.326025	982	0.254876	295	814
	chr5:149209136314	chr5:149209136314	chr5:149209136314	57	228	12.9	10	14.7	63	13.1	54	13.9	38	0.9028		2.7893		22.229		4.9783		-0.316183		-0.341188		793

	9-149211349	6-149266367	9-149266367		03	63	95	3	16	1	98	17	324	47	
	chr2:27538984-27541518	chr2:27548815-27551943	chr2:27538984-27551943	12959	498.78838	630.72231	791.30441	469.07069	564.30336	413.24111	509.49333	180.73666	-0.305101703	0.449527804	387
	chr1:65219516-65221912	chr1:65277879-65283042	chr1:65219516-65283042	63526	100.84836	89.088894	176.8088	107.79749	775.36333	448.09335	489.27751	251.67227	-0.280272517	-0.0506327 8	156
	chr2:159286988-159290429	chr2:159290641-159294315	chr2:159286988-159294315	7327	3.7425 4	11.382 65	37.750 15	5.7270 38	1.9369 07	0.3662 72	1.0937 78	0	-0.255383755	0.072792052	45
	chr10:127312037-127313983	chr10:127367619-127371807	chr10:127312037-127371807	59770	1.3542 25	0	0.5616 95	0	1.4198 62	8.4524 71	46.618 82	10.702 89	-0.240793997	-0.104614996	790
	chr5:149209139-149211349	chr5:149236857-149240061	chr5:149209139-149240061	30922	12.9 10 03	14.7 63 63	13.1 54 95	13.9 383	1.8056 33	4.3389 39	16.347 45	1.9908 33	-0.235660452	-0.290042836	100
	chr5:116849019-116851218	chr5:117371051-117375024	chr5:116849019-117375024	52605	7.8052 14	9.4104 22	8.2181 42	9.4607 76	3.6769 03	3.2401 24	2.7788 63	4.7279 21	-0.232418214	0.442704291	512
	chr8:102516 72	chr8:102547 10	chr8:102516 72	3616 8	221. 52	313. 98	219. 17	138. 90	154. 28	319. 27	661. 38	227. 28	-0.218952	0.137794014	323

	7-102520182	9-102552895	7-102552895		38	1	02	94	13	79	32	68	13		
	chr2:159290641-159294315	chr2:159303838-159312929	chr2:159290641-159312929	22288	1.9369907	0.3662672	1.0937378	0	0.4514408	3.747226	19.303098	12.445479	-0.205408396	-0.326962428	297
	chr16:14986216-14994538	chr16:14998177-14998270	chr16:14986216-14998270	12054	79.176001	54.377755	76.3285	75.438428	0	0.9579551	1.6850585	4.7303888	-0.182992531	0.165923603	849
	chr7:83992963-83999669	chr7:84001349-84004475	chr7:83992963-84004475	11512	3.742254	1.9159001	2.1875556	1.9908833	5.5481772	4.1135332	9.8736215	9.4607676	-0.143282509	-0.301057425	118
	chr3:157874538-157877642	chr3:158012384-158015061	chr3:157874538-158015061	140523	634.9326	783.318	648.6132	504.4533	55.284881	87.567353	154.554775	60.736445	-0.115372193	0.279400994	642
	chr5:55998981-56001529	chr5:56013700-56017687	chr5:55998981-56017687	18706	0	1.324222	0.532083	1.749204	0	0	1.625861	0.746254	-0.112660386	0.017253428	700
	chr10:127300051-127305221	chr10:127312037-127313983	chr10:127300051-127313983	13932	6.968034	5.663161	37.0404	2.989951	1.354225	0	0.561695	0	-0.061294856	0.163120876	178
	chr3:190971312-190971313	chr3:191039477-191039478	chr3:190971312-190971313	74762	2.257041	3.747226	2.690028	0	12.38493	24.2586	75.174558	10.204056	-0.058325258	0.387816263	547

	190976134	191046074	191046074												
	chr14:96104693-96115360	chr14:96169875-96172621	chr14:96104693-96172621	67928	784.5964	504.1333	605.3923	474.2266	0.9684 53	0	8.1885 31	0.4983 25	-0.05 0641 981	0.14 5515 697	370
	chr19:10 4576 04-1045 8591	chr19:10 4735 78-1047 5399	chr19:10 4576 04-1047 5399	17795	12.5 8184	9.69 2152	19.9 832	8.70 9587	41.1 1035	90.8 9208	76.0 9161	58.5 125	-0.04 5764 393	0.09 4978 503	382
	chr4: 6465 909-6469 914	chr4: 6683 508-6687 732	chr4: 6465 909-6687 732	22 18 23	22.0 0383	29.3 864	82.7 4312	39.5 8601	13 0.85 76	11 1.65 72	12 2.59 2	61.2 3478	-0.03 1511 048	0.08 3341 127	520
	chr9: 1039 328-1041 422	chr9: 1326 122-1327 537	chr9: 1039 328-1327 537	28 82 09	28.8 4864	59.4 4903	144. 4981	25.8 8824	0.9684 53	1.32 42 22	1.06 41 67	0.4983 25	-0.00 2442 202	0.41 9542 003	516

Conclusion

Examinations of transcription activity after E2 inducement in respect to interaction sites have revealed a strong correlation between the spatial localization and the transcription activity. From the data, it is unequivocal that spatially localized E2 regulated genes are driven by a mechanism; it is not an event that occurs by a chance. Also, we were able to discern that this behavior is distinctively exhibited by E2 regulated sites. When we examined non E2 regulated genes in co-localized sites, the correlation was significantly weaker. This indicates the existence of co-localized genomic regions by a random chance that was captured by the ChIA-PET experiment. Nonetheless, the layers have datasets utilized in this study have led to a high

quality interaction sites to be examined via RNA-FISH and provided a good insights into the behavior of E2 regulated genes in interaction sites.

CHAPTER 5

CONCLUSION

Bioinformatics approach in conjunction with experimental validation provides a powerful combination to carry out a research. We have applied different techniques and approaches to confront some of challenging issues in cancer research. A wide range of questions were addressed.

We examined genomics, proteomics, and transcriptomics data, which are now highly accessible. Proteomics data were used to train classifier that could accurately predict the excreted proteins. The idea here was to develop a new protocol that will allow us to carry out a target-focused study rather than having to comprehensively check for all proteins in urine to find a potential biomarker. Thus, by predicting the excreted proteins among highly differentially expressed proteins in gastric cancer, we are only left with a small set of proteins to carry out a Western-blot validation.

The result was a highly accurate and highly specific SVM-based classifier that can be applied to any set of proteins to find out whether a protein can be found in urine. This approach was found to be extremely effective and helpful in narrowing down the list of proteins to examine for potential biomarker. By doing so, not only the whole profiling methodology is evaded, but also low abundant proteins, ordinarily masked by profiling methodology, can be examined as well. As a result, we found a highly reproducible potential gastric cancer biomarker, the endothelial lipase. This study led to a follow-up study where the efficacy of EL as a biomarker is now being tested on hundreds of patients.

We applied the same methodology to characterize excreted miRNAs. It allowed us to identify the potentially excreted miRNAs that have not been experimentally found yet. While there is a room for an improvement, this study unveiled some of the properties of miRNAs that may play a pivotal role in allowing the miRNA to get excreted.

Another study that we carried out was a whole-genome comparative study in gastric cancer. It is widely popular to sequence the gastric cancer tissue and compare it to its corresponding healthy tissues. Various aspects of genomic changes can be examined: single nucleotide variants, insertions/deletions, genomic rearrangements, copy number variants, translocations, etc. The single nucleotide variants are examined between the groups of gastric cancer vs healthy tissues. The results from this study can be found in Appendix A.

In Appendix A, mutation analysis of gastric cancer genome is discussed. The general pattern is observed across all samples; gastric cancer tissues have more deleterious mutations compared to its corresponding healthy tissue. For example, there are two different categories of mutations: synonymous mutation and nonsynonymous mutation. Synonymous mutation encodes for the same protein, thus the protein which is encoded by the gene is not affected. It still produces the fully functional protein. In contrast, nonsynonymous mutations can have wide ranges of effect, from benign to detrimental effect. We observe that within the healthy cells, the synonymous to nonsynonymous ratio is higher than the gastric cancer cells. When we delve into different types of nonsynonymous mutations, we see the same behavior in healthy cells; healthy cells have higher percentage of missense mutation as opposed to the nonsense or frame-shift mutations, which can results in deleterious mutation. Through comparative studies, we can observe the general pattern of behavior in gastric cancer.

Lastly, we used transcriptomics data to carry out two types of studies. First, we used it to examine the transcriptional behavior of co-localized regions. Second we utilize the

transcriptomics data in various types of cancer to examine the apoptosis pathways; specifically in aim to understand how apoptosis pathways are suppressed in cancer (Appendix B).

The result of analyzing the transcriptomics data in interaction sites strongly corroborates the hypothesis of transcription-driven co-localization. There is a strong correlation and evidence that interaction sites are spatially co-localized from distal sites by a driving force. It is unambiguous and unequivocal that the data suggests a non-random behavior in co-localized regions. The experimental validation should corroborate the findings.

In Appendix C, the results from the protein arrays, which were used to validate the excretory proteins in biomarker study, is discussed. The seven samples of gastric cancer urine samples and normal healthy samples were used in validation study. We show differentially expressed proteins that were found in the two groups that were not included in the published work.

As presented here, the bioinformatics approaches can be used to study various types of biological problem. A machine-learning, data mining based approach can be used to make biological predictions, which leads to a better understanding and characterization of the set being used. Transcriptomics provide insights to the gene expression at a given condition. Thus, when layered with different datasets derived from the same condition, it can provide even profound details into the biological significance. It is expected that the whole genome comparative studies will become even more common than now as the sequencing technologies advance to sequence the whole genome at even a lower cost. With –omics field expanding as ever, we highly anticipate that this will lead to a better understanding of cancer.

REFERENCES

1. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100: 57-70.
2. Msika S, Benhamiche AM, Jouve JL, Rat P, Faivre J (2000) Prognostic factors after curative resection for gastric cancer. A population-based study. Eur J Cancer 36: 390-396.
3. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66-72.
4. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, et al. (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463: 184-190.
5. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.
6. O'Connor DS, Schechner JS, Adida C, Mesri M, Rothermel AL, et al. (2000) Control of apoptosis during angiogenesis by survivin expression in endothelial cells. Am J Pathol 156: 393-398.
7. Kerr JF, Wyllie AH, Currie AR (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer 26: 239-257.
8. Hu Q, Kwon YS, Nunez E, Cardamone MD, Hutt KR, et al. (2008) Enhancing nuclear receptor-induced transcription requires nuclear motor and LSD1-dependent gene networking in interchromatin granules. Proc Natl Acad Sci U S A 105: 19199-19204.
9. Kocanova S, Kerr EA, Rafique S, Boyle S, Katz E, et al. (2010) Activation of estrogen-responsive genes does not require their nuclear co-localization. PLoS Genet 6: e1000922.
10. Ludwig JA, Weinstein JN (2005) Biomarkers in Cancer Staging, Prognosis and Treatment Selection. Nature Reviews Cancer 5: 845-856.
11. Pang JX, Ginanni N, Dongre AR, Hefta SA, Opitek GJ (2002) Biomarker discovery in urine by proteomics. J Proteome Res 1: 161-169.
12. Weissinger EM, Schiffer E, Hertenstein B, Ferrara JL, Holler E, et al. (2007) Proteomic patterns predict acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation. Blood 109: 5511-5519.
13. Zimmerli LU, Schiffer E, Zurbig P, Good DM, Kellmann M, et al. (2008) Urinary proteomic biomarkers in coronary artery disease. Mol Cell Proteomics 7: 290-298.
14. Barratt J, Topham P (2007) Urine proteomics: the present and future of measuring urinary protein components in disease. CMAJ 177: 361-368.
15. Decramer S, Gonzalez de Peredo A, Breuil B, Mischak H, Monsarrat B, et al. (2008) Urine in clinical proteomics. Mol Cell Proteomics 7: 1850-1862.
16. Adachi J, Kumar C, Zhang Y, Olsen JV, Mann M (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. Genome Biol 7: R80.
17. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389: 1017-1031.
18. Cui J, Liu Q, Puett D, Xu Y (2008) Computational prediction of human proteins that can be secreted into the bloodstream. Bioinformatics 42.
19. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281-288.
20. Chang C-C, Lin C-J (2001) LIBSVM: a library for support vector machines.

21. Platt JC (1999) Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press.
22. Castagna A, Cecconi D, Sennels L, Rappsilber J, Guerrier L, et al. (2005) Exploring the hidden human urinary proteome via ligand library beads. J Proteome Res: 1917-1930.
23. Pieper R, Gatlin C, McGrath A, Makusky A, Mondal M, et al. (2004) Characterization of the human urinary proteome: a method for high-resolution display of urinary proteins on two-dimensional electrophoresis gels with a yield of nearly 1400 nearly protein spots. Proteomics: 1159-1174.
24. Wang L, Li F, Sun W, Wu S, Wang X, et al. (2006) Concanavalin A captured glycoproteins in healthy human urine. Mol Cell Proteomics: 560-562.
25. Hand DJ, Till RJ (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45: 171-186.
26. Affymetrix (2005) Alternative Transcript Analysis Methods for Exon Arrays.
27. Cui J, Chen Y, Chou J, Sun L, Chen L, et al. (2010) Integrated Transcriptomic and Proteomic Analysis for Biomarker Identification for Gastric Cancer. Nucleic Acids Res, in press.
28. Dennis GJ, Sherman B, Hosack D, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4: P3.
29. Wu J, Mao X, Cai T, Luo J, Wei L (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res: W720-W724.
30. Huang D, Sherman B, Lempicki R (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4: 44-57.
31. Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated Genome Annotation and Pathway Identification Using the KEGG Orthology (KO) As a Controlled Vocabulary. Bioinformatics: 3787-3793.
32. Rasband WS (1997-2005) ImageJ, U.S. . National Institutes of Health, Bethesda, Maryland, USA.
33. Gilmore RE (1973) Glomerular filtration. In: RW OJB, editor. Handbook of Physiology. Washington DC: American Physiological Society. pp. 185-248.
34. Osicka TM, Panagiotopoulos S, Jerums G, Comper WD (1997) Fractional clearance of albumin is influenced by its degradation during renal passage. Clin Sci (Lond) 93: 557-564.
35. Rader DJ, Jaye M (2000) Endothelial lipase: a new member of the triglyceride lipase gene family. Curr Opin Lipidol 11: 141-147.
36. Griffon N, Jin W, Petty TJ, Millar J, Badellino KO, et al. (2009) Identification of the Active Form of Endothelial Lipase, a Homodimer in a Head-to-Tail Conformation. J Biol Chem 284: 23322-23330.
37. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, et al. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21: 3435-3438.
38. Li ZR, Lin HH, Han LY, Jiang L, Chen X, et al. (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 34: W32-37.
39. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784-3788.
40. Garrow AG, Agnew A, Westhead DR (2005) TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. Nucleic Acids Res 33: W188-192.
41. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S (2005) Prediction of twin-arginine signal peptides. BMC Bioinformatics 6: 167.

42. Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res 35: W429-432.
43. Julenius K, Molgaard A, Gupta R, Brunak S (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 15: 153-164.
44. Gupta R, Jung E, Brunak S (2004) Prediction of N-glycosylation sites in human proteins.
45. Eisenhaber F, Imperiale F, Argos P, Froemmel C (1995) Prediction of Secondary Structural Content of Proteins from Their Amino Acid Comosition Alone Utilizing Analytic Vector Decomposition.
46. Zhao Y, Srivastava D (2007) A developmental view of microRNA function. Trends Biochem Sci 32: 189-197.
47. Pillai RS (2005) MicroRNA function: multiple mechanisms for a tiny RNA? Rna-a Publication of the Rna Society 11: 1753-1761.
48. Weber JA, Baxter DH, Zhang S, Huang DY, Huang KH, et al. (2010) The microRNA spectrum in 12 body fluids. Clin Chem 56: 1733-1741.
49. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, et al. (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A 103: 2257-2261.
50. Roldo C, Missiaglia E, Hagan JP, Falconi M, Capelli P, et al. (2006) MicroRNA expression abnormalities in pancreatic endocrine and acinar tumors are associated with distinctive pathologic features and clinical behavior. J Clin Oncol 24: 4677-4684.
51. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, et al. (2005) MicroRNA gene expression deregulation in human breast cancer. Cancer Res 65: 7065-7070.
52. Meng F, Henson R, Wehbe-Janek H, Ghoshal K, Jacob ST, et al. (2007) MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. Gastroenterology 133: 647-658.
53. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, et al. (2006) Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 9: 189-198.
54. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, et al. (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 99: 15524-15529.
55. Taylor DD, Gercel-Taylor C (2008) MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. Gynecol Oncol 110: 13-21.
56. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, et al. (2008) Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A 105: 10513-10518.
57. Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, et al. (2004) The Microprocessor complex mediates the genesis of microRNAs. Nature 432: 235-240.
58. Turchinovich A, Weiz L, Langheinz A, Burwinkel B (2011) Characterization of extracellular circulating microRNA. Nucleic Acids Res 39: 7223-7233.
59. Hong CS, Cui J, Ni Z, Su Y, Puett D, et al. (2011) A computational method for prediction of excretory proteins and application to identification of gastric cancer markers in urine. PLoS One 6: e16875.
60. Cui J, Liu Q, Puett D, Xu Y (2008) Computational prediction of human proteins that can be secreted into the bloodstream. Bioinformatics 24: 2370-2375.
61. Chen X, Ba Y, Ma L, Cai X, Yin Y, et al. (2008) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res 18: 997-1006.

62. Kosaka N, Iguchi H, Ochiya T (2010) Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. Cancer Sci 101: 2087-2092.
63. Chang C-C, C-J L (2001) LIBSVM: a library for support vector machines.
64. Weng L, Wu X, Gao H, Mu B, Li X, et al. (2010) MicroRNA profiling of clear cell renal cell carcinoma by whole-genome small RNA deep sequencing of paired frozen and formalin-fixed, paraffin-embedded tissue specimens. Journal of Pathology 222: 41-51.
65. Keller A, Leidinger P, Borries A, Wendschlag A, Wucherpfennig F, et al. (2009) miRNAs in lung cancer - studying complex fingerprints in patient's blood cells by microarray experiments. BMC Cancer 9: 353.
66. Ozata DM, Caramuta S, Velazquez-Fernandez D, Akcakaya P, Xie H, et al. (2011) The role of microRNA deregulation in the pathogenesis of adrenocortical carcinoma. Endocrine-Related Cancer 18: 643-655.
67. Kahlert C, Klupp F, Brand K, Lasitschka F, Diederichs S, et al. (2011) Invasion front-specific expression and prognostic significance of microRNA in colorectal liver metastases. Cancer Sci 102: 1799-1807.
68. Papadopoulos GL, Alexiou P, Maragkakis M, Reczko M, Hatzigeorgiou AG (2009) DIANA-mirPath: Integrating human and mouse microRNAs in pathways. Bioinformatics 25: 1991-1993.
69. Hu HY, Yan Z, Xu Y, Hu H, Menzel C, et al. (2009) Sequence features associated with microRNA strand selection in humans and flies. BMC Genomics 10: 413.
70. Valadi H, Ekstrom K, Bossios A, Sjostrand M, Lee JJ, et al. (2007) Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9: 654-659.
71. Hunter MP, Ismail N, Zhang X, Aguda BD, Lee EJ, et al. (2008) Detection of microRNA expression in human peripheral blood microvesicles. PLoS One 3: e3694.
72. Diehl P, Fricke A, Sander L, Stamm J, Bassler N, et al. (2012) Microparticles: major transport vehicles for distinct microRNAs in circulation. Cardiovasc Res 93: 633-644.
73. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, et al. (2011) Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. Proc Natl Acad Sci U S A 108: 5003-5008.
74. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34: D140-144.
75. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36: D154-158.
76. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39: D152-157.
77. Griffiths-Jones S (2004) The microRNA Registry. Nucleic Acids Res 32: D109-111.
78. Hanke M, Hoefig K, Merz H, Feller AC, Kausch I, et al. (2010) A robust methodology to study urine microRNA as tumor marker: microRNA-126 and microRNA-182 are related to urinary bladder cancer. Urol Oncol 28: 655-661.
79. Yamada Y, Enokida H, Kojima S, Kawakami K, Chiyomaru T, et al. (2011) MiR-96 and miR-183 detection in urine serve as potential tumor markers of urothelial carcinoma: correlation with stage and grade, and comparison with urinary cytology. Cancer Sci 102: 522-529.
80. Nilsson J, Skog J, Nordstrand A, Baranov V, Mincheva-Nilsson L, et al. (2009) Prostate cancer-derived urine exosomes: a novel approach to biomarkers for prostate cancer. Br J Cancer 100: 1603-1607.

81. Corsten MF, Dennert R, Jochems S, Kuznetsova T, Devaux Y, et al. (2010) Circulating MicroRNA-208b and MicroRNA-499 reflect myocardial damage in cardiovascular disease. Circ Cardiovasc Genet 3: 499-506.
82. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, et al. (1994) Fast Folding and Comparison of Rna Secondary Structures. Monatshefte Fur Chemie 125: 167-188.
83. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9: 133-148.
84. McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29: 1105-1119.
85. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429-3431.
86. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452: 51-55.
87. Markham NR, Zuker M (2005) DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res 33: W577-581.
88. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.
89. Hofacker IL, Stadler PF (2006) Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics 22: 1172-1176.
90. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, et al. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45: 81-94.
91. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322: 1845-1848.
92. Seila AC, Core LJ, Lis JT, Sharp PA (2009) Divergent transcription: a new feature of active promoters. Cell Cycle 8: 2557-2564.
93. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, et al. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474: 390-394.
94. Min IM, Waterfall JJ, Core LJ, Munroe RJ, Schimenti J, et al. (2011) Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev 25: 742-754.
95. Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, et al. (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell 145: 622-634.
96. Horwitz KB, Koseki Y, McGuire WL (1978) Estrogen control of progesterone receptor in human breast cancer: role of estradiol and antiestrogen. Endocrinology 103: 1742-1751.
97. Chlebowski RT, Hendrix SL, Langer RD, Stefanick ML, Gass M, et al. (2003) Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women: the Women's Health Initiative Randomized Trial. JAMA 289: 3243-3253.
98. Pike MC, Spicer DV, Dahmoush L, Press MF (1993) Estrogens, progestogens, normal breast cell proliferation, and breast cancer risk. Epidemiol Rev 15: 17-35.
99. Strey HH, Podgornik R, Rau DC, Parsegian VA (1998) DNA--DNA interactions. Curr Opin Struct Biol 8: 309-313.
100. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462: 58-64.
101. Rodgers J, Nicewander W (1988) Thirteen ways to look at the correlation coefficient. The American Statistician 42: 59-66.
102. (2002) Oxford Dictionary of Statistics: Oxford Univeresity Press. 104 p.
103. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, et al. (2004) An abundance of bidirectional promoters in the human genome. Genome Res 14: 62-66.

104. Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, et al. (2009) ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. EMBO J 28: 1418-1428.
105. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.
106. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39: D876-882.
107. KA W DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program Available at: www.genome.gov/sequencingcosts.
108. NCI N, National Institutes of Health (2012) Cancers Selected for Study.
109. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719-724.
110. Brenner S, Stretton AO, Kaplan S (1965) Genetic code: the 'nonsense' triplets for chain termination and their suppression. Nature 206: 994-998.
111. Petitjean A, Achatz MI, Borresen-Dale AL, Hainaut P, Olivier M (2007) TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene 26: 2157-2165.
112. Schwartz S, Jr., Yamamoto H, Navarro M, Maestro M, Reventos J, et al. (1999) Frameshift mutations at mononucleotide repeats in caspase-5 and other target genes in endometrial and gastrointestinal cancer of the microsatellite mutator phenotype. Cancer Res 59: 2995-3002.
113. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, et al. (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35: W169-175.
114. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
115. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.
116. Demetris AJ (1998) Immune cholangitis: liver allograft rejection and graft-versus-host disease. Mayo Clin Proc 73: 367-379.
117. Weetman AP, McGregor AM (1994) Autoimmune thyroid disease: further developments in our understanding. Endocr Rev 15: 788-830.
118. de Visser KE, Eichten A, Coussens LM (2006) Paradoxical roles of the immune system during cancer development. Nat Rev Cancer 6: 24-37.
119. Hirohashi S, Kanai Y (2003) Cell adhesion system and human cancer morphogenesis. Cancer Sci 94: 575-581.
120. Tanaka K, Iwamoto S, Gon G, Nohara T, Iwamoto M, et al. (2000) Expression of survivin and its relationship to loss of apoptosis in breast carcinomas. Clin Cancer Res 6: 127-134.
121. Gouas DA, Villar S, Ortiz-Cuaran S, Legros P, Ferro G, et al. (2012) TP53 R249S mutation, genetic variations in HBX and risk of hepatocellular carcinoma in The Gambia. Carcinogenesis 33: NP.
122. Han ES, Moyer MP, Naylor S, Sakaguchi AY (1991) Mutation in the TP53 gene in colorectal carcinoma detected by polymerase chain reaction. Genes Chromosomes Cancer 3: 313-317.
123. Foulkes WD, Stamp GW, Afzal S, Lalani N, McFarlane CP, et al. (1995) MDM2 overexpression is rare in ovarian carcinoma irrespective of TP53 mutation status. Br J Cancer 72: 883-888.

124. Ngan HY, Cheung AN, Liu SS, Yip PS, Tsao SW (1999) Abnormal expression or mutation of TP53 and HPV in vulvar cancer. Eur J Cancer 35: 481-484.
125. Soussi T, Beroud C (2001) Assessing TP53 status in human tumours to evaluate clinical outcome. Nat Rev Cancer 1: 233-240.
126. Rasheed BK, McLendon RE, Herndon JE, Friedman HS, Friedman AH, et al. (1994) Alterations of the TP53 gene in human gliomas. Cancer Res 54: 1324-1330.
127. Haldar S, Negrini M, Monne M, Sabbioni S, Croce CM (1994) Down-regulation of bcl-2 by p53 in breast cancer cells. Cancer Res 54: 2095-2097.

APPENDIX A

SINGLE NUCLEOTIDE VARIANT ANALYSIS FROM GENOMIC DATA

In this Appendix, the methods to approach mutational analysis are shown. Thanks to the advancements in genome sequencing technologies, it is now feasible to sequence a whole human genome at a reasonable cost. The cost to sequence human genome has come a long way, marking an initial cost at 3 billion dollars from the Human Genome Project at the National Human Genome Research Institution to under $5,000 in current market. It is incredible that the cost to sequence 1 base has dropped exponentially over the past decade [107].

Due to a low cost in sequencing genomes, an approach to cancer study has been to sequence the whole genome of the cancer genome and compare it to the normal human genome. By using a comparative approach, it is possible to discover the genetic changes that may drive the cancer development and progression. The first cancer genome studies were successful at providing insights to genetic changes such as single nucleotide variants, copy number changes, and translocations [3,4]. Currently there is a massive collaboration effort in sequencing various cancer genomes, termed The Cancer Genome Atlas initiated by the National Health Institute.

The goal of TCGA is to sequence more than 20 cancer type genomes to comprehensively characterize the genomic changes occurring in each type of cancer [108]. The information that may be available from such collective effort to understand cancer at a genome-wide level has been sparked all around the world.

In a standard genome sequencing protocol, the reads produced from the sequencing machines are mapped to the human reference genome. Different platforms of sequencing

machines, such as 454 (Roche Corp), illumina (Illumina Inc), and SOLiD (Life Technology), have different pipelines and software to systematically retrieve the critical information from the DNA sequence.

In an attempt to understand the genomic changes occurring in gastric cancer patients, a single point mutational analysis was performed on five pairs of gastric cancer genome *vs* its corresponding healthy genome. The five samples were from different stages of gastric cancer, from the beginning stage to the progressively well advanced metastasized stage. The single point mutations are examined in a comparative manner.

Single Nucleotide Variants Examined

A mutation occurring at a single point can be either synonymous or nonsynonymous mutation. Synonymous mutation refers to mutations that do not change the nucleic acids. Thus when it is translated, the protein sequence is not affected. For example, nucleic acid Glycine can be encoded by four different tri-nucleotides; GGU, GGC, GGA, GGG. If a sequence containing GGU undergoes a mutational change at the 3rd nucleotide place and becomes GGC, the resulting nucleic acid is the same even if the sequence has a mutation at a DNA level. When this occurs, it is referred to as a synonymous mutation.

Unlike the synonymous mutation, nonsynonymous mutation results in a sequence change at a protein level. If the same aforementioned nucleotide was mutated at the first position with a nucleotide Adenine, the resulting DNA sequence will be AGU. When this is transcribed and translated into a protein, the protein sequence is not changed from Glycine to Serine. This may have an undesired effect on the protein.

To examine how much of single point mutational changes are synonymous and nonsynonymous, we counted the total number of synonymous and nonsynonymous mutations

in each sample. Table 1 summarizes the total number of mutations for each sample in each category.

It is evident that the number of nonsynonymous mutations is relatively smaller than the synonymous mutations. Due to its potential detrimental effect, a nonsynonymous mutation may be more aggressively repaired by the cell's innate repair system versus the synonymous mutations.

Figure AA.1. The synonymous and nonsynonymous mutation count and percentage of five gastric cancer genomes and its corresponding healthy genome (N: normal; T: gastric cancer

Sample	Synonymous	Synoymous%	Nonsynonymous	Nonsynonymous %
1N	20563	0.532830638	18029	0.467169362
1T	20260	0.529756302	17984	0.470243698
2N	20857	0.53229717	18326	0.46770283
2T	20656	0.526026281	18612	0.473973719
3N	20856	0.536654401	18007	0.463345599
3T	20106	0.532778632	17632	0.467221368
4N	20933	0.532998931	18341	0.467001069
4T	20474	0.532289933	17990	0.467710067
5N	20902	0.536636714	18048	0.463363286
5T	20872	0.535303019	18119	0.464696981

When comparing the groups of normal *vs* the tumor genomes, the difference in mutation ratio is easily observed. The mutation is calculated by the below equation, where N_{SYN}=

number of synonymous mutations, N_{NON}= number of nonsynonymous mutations, and N_{TOTAL}= number of total mutations.

$$Syn\ to\ Nonsyn\ Mutation\ Ratio = \frac{N_{SYN}/N_{TOTAL}}{N_{NON}/N_{TOTAL}}, \text{ (Equation 1)}$$

The ratio calculated for each sample is graphed on Figure AA.1. The higher synonymous to nonsynonymous ratio is observed for all five cases. The probability that this observation is P(all five cases having higher ratio for gastric cancer)=$(0.5)^5$= 0.03125. This implies the higher occurrence of relatively higher nonsynonymous mutations in cancer samples are not due to a random chance. Considering that cancer genomes undergo significant, it is not surprising to find higher rate of nonsynonymous mutations in cancer [109].

To understand the type of nonsynonymous mutations occurring in gastric cancer samples, we examined the categories of nonsynonymous mutations: missense, nonsense, frameshift. Missense mutations are mutations that results in a different protein sequence. Nonsense mutations are mutation that results in one of three stop codons: UAA, UAG, UGA [110]. Frameshift mutations are results of a single nucleotide insertions or deletions. An inserted or a deleted codon (indels) results in a shift in reading frames. The product of indels is a protein that may be drastically different in sequence from the original sequence.

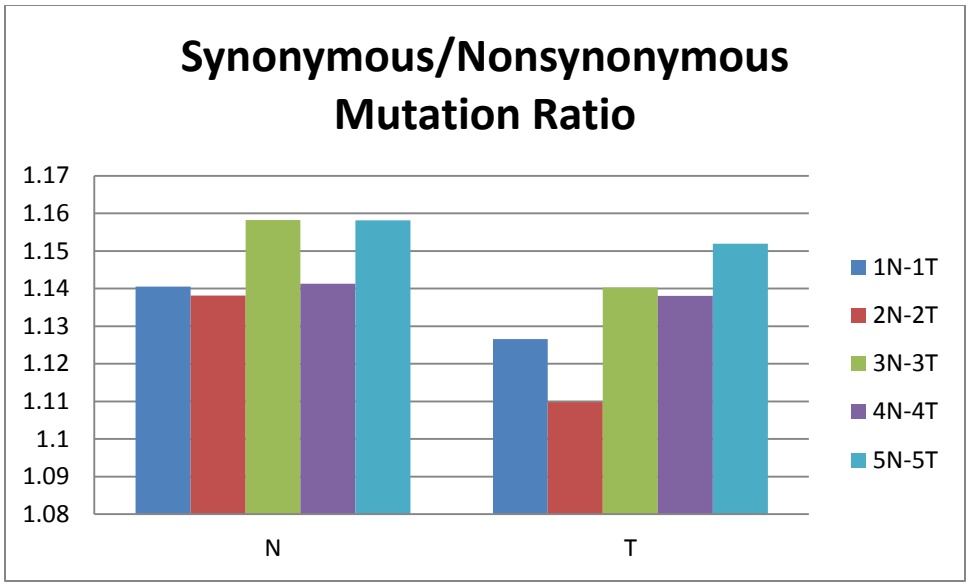


Figure AA.1. The ratio of synonymous/nonsynonymous mutation of five gastric cancer genomes and its corresponding healthy genome (N: normal; T: gastric cancer tumor)

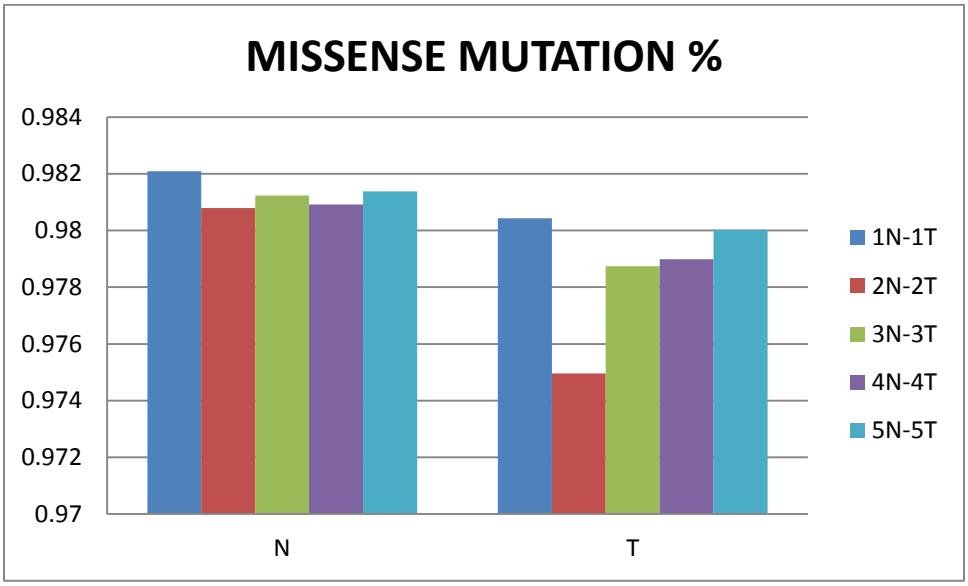
Unlike the missense mutations, the nonsense and frameshift mutations can have a detrimental effect on a cell. While missense mutations affect only one nucleic acid, a nonsense mutation can result in a non-function protein. When a stop codon is introduced in the beginning of the protein, the resulting translated product will be nonfunctional. For example, a gene *TP53* plays an important role in proofreading the DNA as well as regulating cell proliferation and apoptosis. A nonsense mutation in *TP53* may lead to a development of cancer [111]. As mentioned before, a frameshift also can cause a significant change to a protein. For example, frameshift mutations in caspase-5 were found to play a role in nonpolyposis colorectal cancers, gastrointestinal and endometrial tumors [112].

Each type of nonsynonymous mutation is examined. The values are calculated by the percentage for each type of mutations (Eqn2) and are graphed.

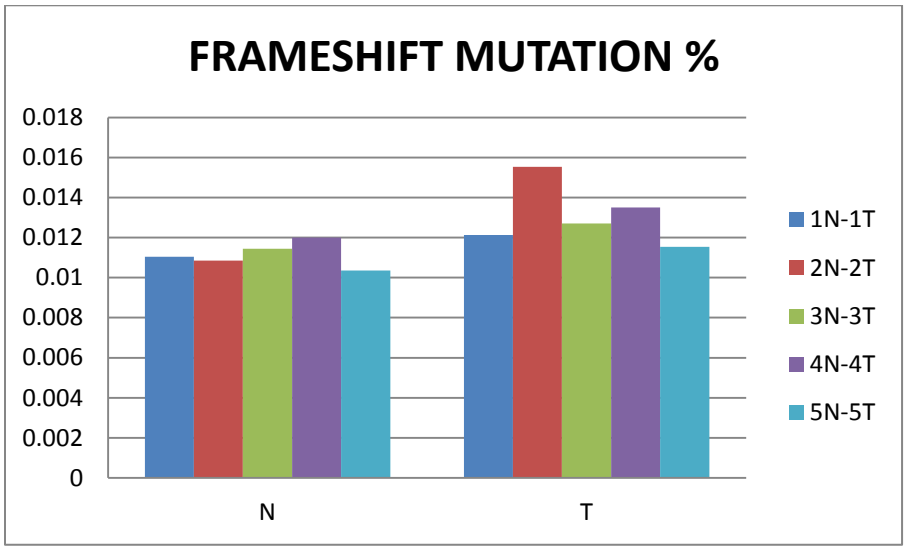
$$Percentage = \frac{N_{Mut}}{N_{Total}}, \text{ (Equation 2)}$$

In Figure AA.2A, percentage of missense mutations in normal groups compared to the gastric cancer groups are higher. However, it's the reverse case for the percentage of frameshift mutations and nonsense mutations. Figure AA.2B-C shows that all cancer genomes have higher percentage mutations in both frameshift and nonsense categories. This indicates that while the number of overall mutations occurring are comparable or exhibiting no specific trend in the groups of samples, as shown in Figure AA.3, the rate of deleterious mutations in cancer genomes are consistently higher. These deleterious mutations in cancer may be the characteristics of cancer genomes and may be contributing to the development of cancers. From this study, it is evident that the gastric cancer genomes are signature by higher percentage of deleterious mutations.

A



B



C

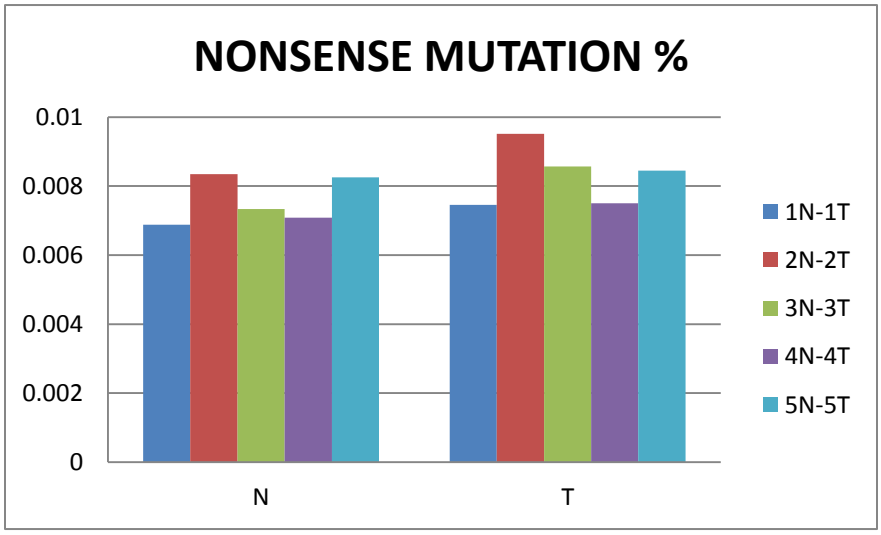


Figure AA.2. The percentage of different types of nonsynonymous mutations. A: Percentages of missense mutation occurrence. B: Percentages of frameshift mutations. C: Percentages of nonsense mutations. (N: normal; T: gastric cancer tumor)

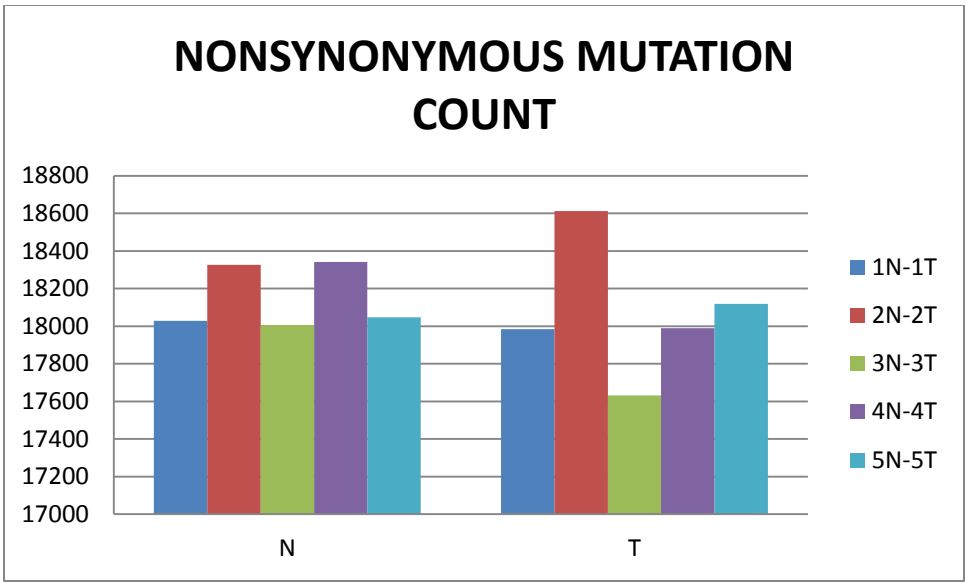


Figure AA.3. The total count of mutations in samples (N: normal; T: gastric cancer

Another analysis that was carried out was the mutation density analysis. Based on the number of mutations in a given window, a mutation density can be calculated in fixed sized windows throughout the whole genome. This was to examine whether gastric cancer genome exhibits an unusually high mutated regions. We used a window size of 1,000 kb. For every window, the number of mutations occurring in that region was counted. This was calculated for all chromosomes. Figure AA.4 shows the density graph.

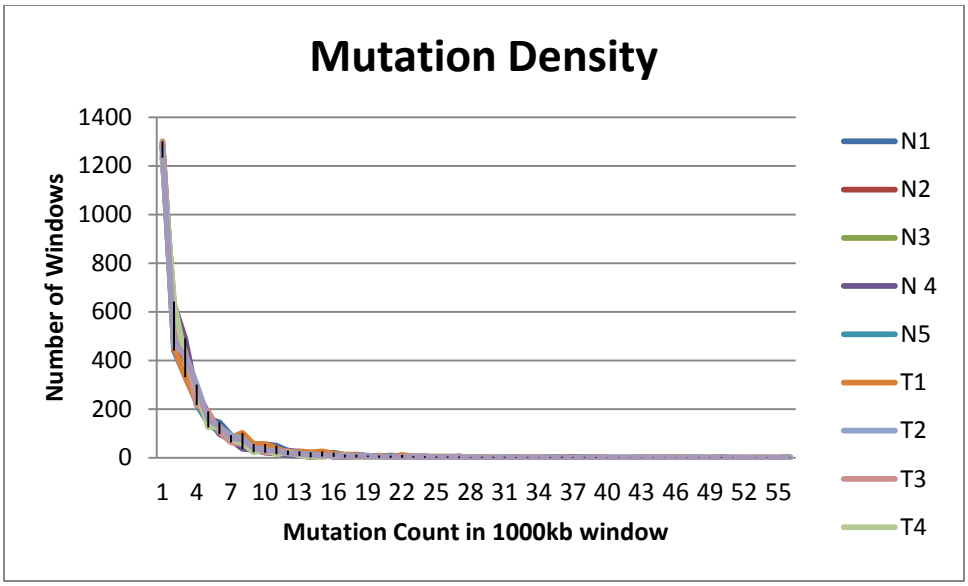


Figure AA.4. Mutation density graph. X axis shows the number of mutations that was calculated from the 1000kb windows; Y axis shows the number of windows having specified number of mutations (N: normal; T: gastric cancer tumor)

From Figure AA.4, most windows have mutation count from 0-10. While there seems to be variations from the graph between the counts 1-10, a further examination shows no discrete difference between the tumor and normal samples.

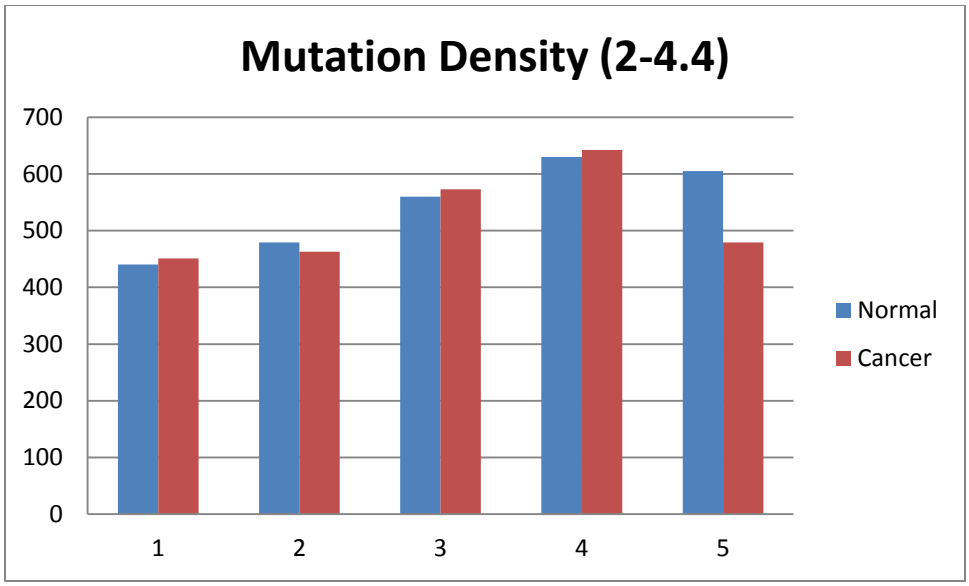


Figure AA.5. Mutation density graph for counts 2-4.44. X-axis: each sample; Y-axis: number of windows that have mutation count range of 2-4.44.(N: normal; T: gastric cancer tumor)

No distinct pattern was observed in the mutation density. The mutation count ranged from 0-117. We examined whether the highest containing regions were only discernible in gastric cancer genome. However, they were randomly distributed between all samples. Thus we observed no evidence in the difference of mutation density.

Next, we examined the mutated genes in tumor samples. A gene can be mutated by a chance, but if it is consistently mutated across tumor samples, it may contribute to the cancer progression. We filtered out to find the tumor-specific mutated genes that are mutated across all tumor samples. There were 54 such genes (Table AA.2).

Table AA.2. Names of genes that are mutated

Gene Name		
AMAC1L3	OR2T8	KCNJ12

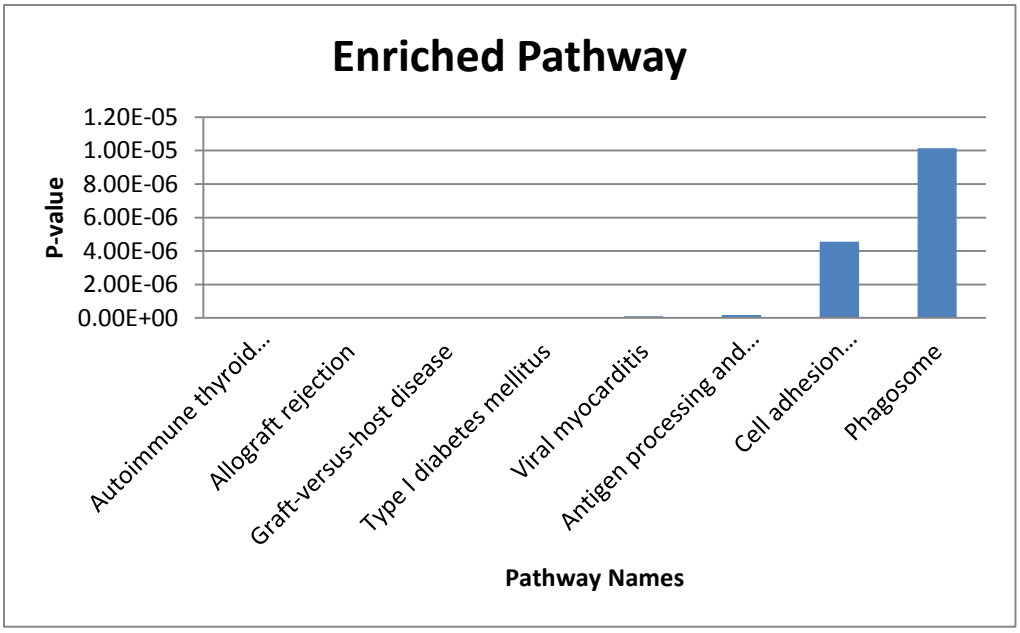
C14orf73	OR4C3	LAMA5
C16orf68	OR8U1	LOC100129307
CDC27	OR9G1	LOC100129979
CLCNKB	PDE4DIP	LOC100288568
CTBP2	PDHX	LOC100505879
EIF3CL	PKD1L2	LOC100506072
EPPK1	PLIN4	LOC100506287
FAM38A	PRIM2	LOC100507545
FCGBP	RP1L1	LRRC56
FLG	SCARF1	MRPS34
GOLGA6L2	SIK1	MUC12
HAS1	TAS2R46	MUC3A
HLA-A	TPTE	MUC4
HLA-B	TRIOBP	MUC6
HLA-C	TYSND1	MYO15A
HLA-DQA1	ZFHX4	NEFH
HLA-DQB1	ZNF469	NISCH
HLA-DRB1	ZNF517	OBSCN
IFNA17	ZNF595	OR2C1
ISCU	ZNF676	
JRK	ZNF717	

To better understand the role of these genes, the enrichment analyses were carried out. The functional and pathway enrichment was performed using DAVID database using human

genes as the background [113,114,115]. The enriched pathways and functions are determined by the p-values, and the p-values of less than 1x10E-5 are considered enriched.

The enrichment analysis shows that some of these genes are implicated in other diseases. As shown in Fig AA.6.A and B, the most enriched genes are associated with autoimmune thyroid disease and allograft rejection disease. Immune system plays an important role in both diseases, thus these genes may be linked to the immune response [116,117]. Immune response in cancer also plays an important role. An alteration in immune response can cause prolonged immune response where it creates a favorable environment to cancer development [118]. These genes that were found to be consistently mutated in tumor samples may be involved in an immune response activity where the loss of function contributes to the enhanced environment for the cancer development.

A



B

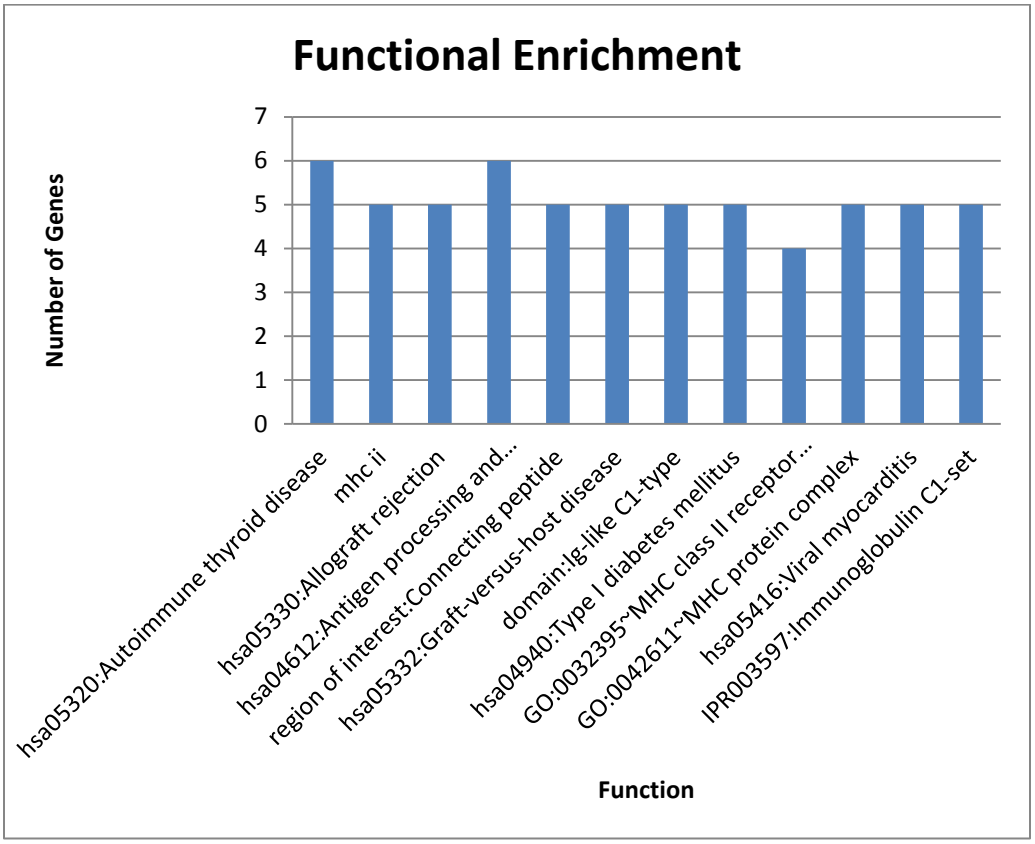


Figure AA.6. Pathway and Functional Enrichment Analysis for the 54 genes that were tumor specifically mutated in five gastric cancer genomes. A: Pathway enrichment

Another characteristic that is observed in the enrichment analysis is genes involved in cell adhesion pathway. Cell adhesion is an important factor in maintaining the cell order in tissues. When a cancer cell develops, the reduced cell adhesiveness allows a cancer cell to escape and metastasize at other sites of the body [119]. It is one of the factors that characterize the cancer cells.

Conclusion

Using a comparative approach, insightful information was extracted using bioinformatics approach. By examining the mutations at the nucleotide level, different mutation types were analyzed. The results show that the gastric cancer genomes have higher rate of nonsynonymous mutations, as well as higher rate of lethal types of nonsynonymous mutations, such as nonsense and frameshift as opposed to the missense mutations. We were able to find 54 genes that were repeatedly mutated in gastric cancer samples. The follow up analysis for these genes show that these genes are associated with inflammation and/or adhesions, which are both critical factors in development of cancer

APPENDIX B

ANALYSIS OF APOPTOSIS PATHWAY USING TRANSCRIPTOMICS DATA

Cancer cells are characterized by several traits. Six characteristics of cancer has been termed as the hallmarks of cancer: ability to proliferate; evade growth suppressors; immune to cell death; ability to replicate; angiogenesis; invasion and metastasis[1]. Apoptosis play a crucial role in developments of cancer cells. In normal cells, there exists an intricate and complicated network to monitor a cell from acting abnormal. Thus, when a rate of proliferation goes up, the rate of cell death correlates [120] .

The understanding of exact mechanism in which a cell is able to suppress the apoptosis pathway is limited. To address this question, we wanted to examine whether there is a specific type of apoptotic pathway that is suppressed in cancers. The cell death can be induced as a result of intracellular, extracellular, or immune-response activated pathways. We collected the list of genes from three different pathways and examined the expression across various types of cancer. Table AB.1 shows the list of apoptotic genes examined.

Table AB.1. List of apoptotic genes and their category

Gene Name	Pathway
CASP8	Extrinsic
TNFRSF1A	Extrinsic
TRADD	Extrinsic
FADD	Extrinsic
TRAF2	Extrinsic

TNFRSF10B	Extrinsic
TNFRSF10A	Extrinsic
TNFRSF21	Extrinsic
RIPK1	Extrinsic
CASP2	Extrinsic
CASP10	Extrinsic
CRADD	Extrinsic
LTBR	Extrinsic
TRAF3	Extrinsic
NGFR	Extrinsic
NGFRAP1	Extrinsic
CD70	Extrinsic
CD27	Extrinsic
TNFSF18	Extrinsic
TNFSRF25	Extrinsic
FAF1	Extrinsic
FLASH	Extrinsic
DAXX	Extrinsic
PTPN13	Extrinsic
RIPK1	Extrinsic
BCL10	Intrinsic
BAX	Intrinsic
BAK1	Intrinsic

BAD	Intrinsic
BCL2L11	Intrinsic
BIK	Intrinsic
BLK	Intrinsic
CYCS	Intrinsic
APAF1	Intrinsic
CASP9	Intrinsic
GZMB	Immune Response
CASP1	Immune Response
CASP4	Immune Response
CASP5	Immune Response

To characterize a distinctive behavior in cancer, the The Cancer Genome Atlas transcriptomics data was collected and examined. The following cancer types were examined: breast invasive carcinoma; colon adenocarcinoma; kidney renal clear cell carcinoma; kidney renal papillary cell carcinoma; brain lower grade glioma; lung adenocarcinoma; lung squamous cell carcinoma; ovarian serous cystadenocarcinoma; rectum adenocarcinoma; uterine corpus endometrioid carcinoma.

The average expression level for all samples in each category was calculated and graphed (Fig AB. 1). When the list of genes is examined by individually by the type of cancer, the genes that are suppressed or overexpressed are discerned. It is notable that the inflammation induced apoptosis pathway is overexpressed, whereas the intrinsic pathway is suppressed.

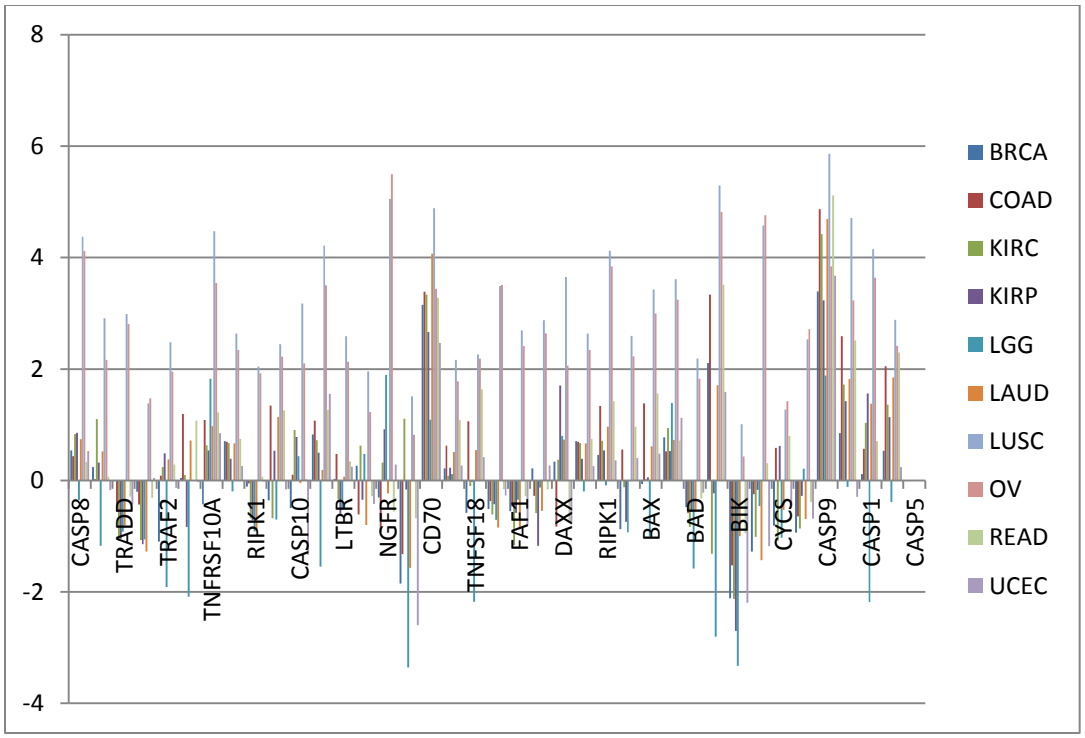
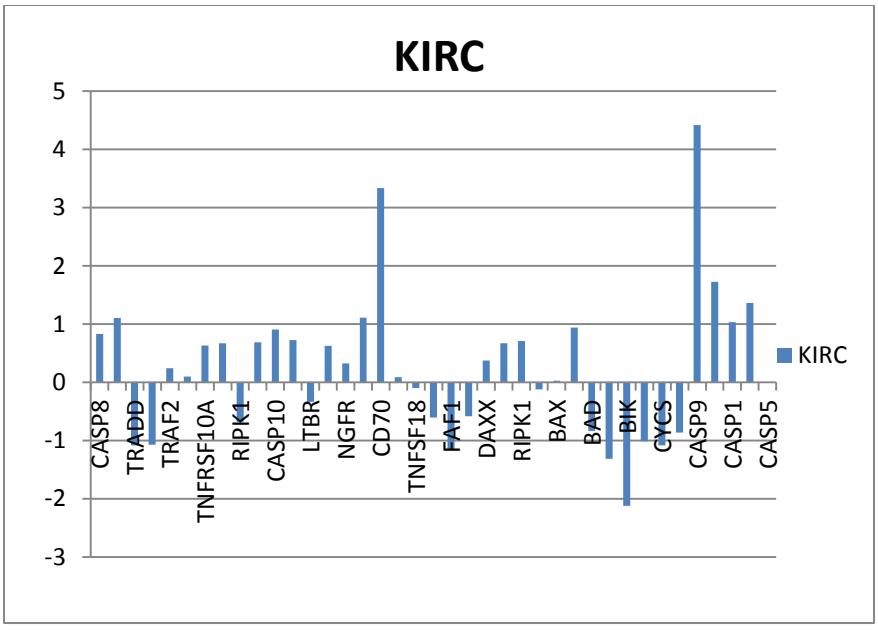
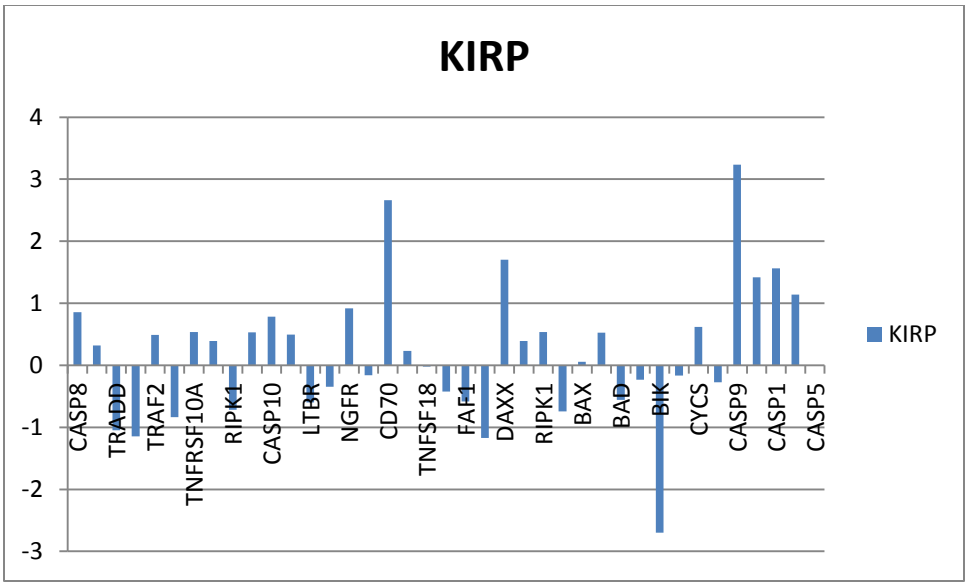


Figure AB. 1. The mean value for each apoptosis genes across various types of cancer

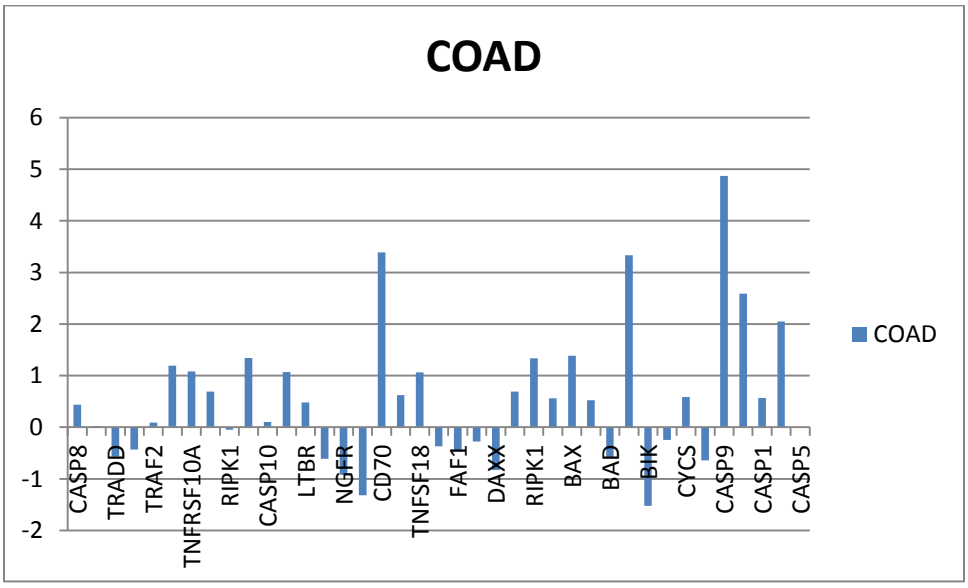
A



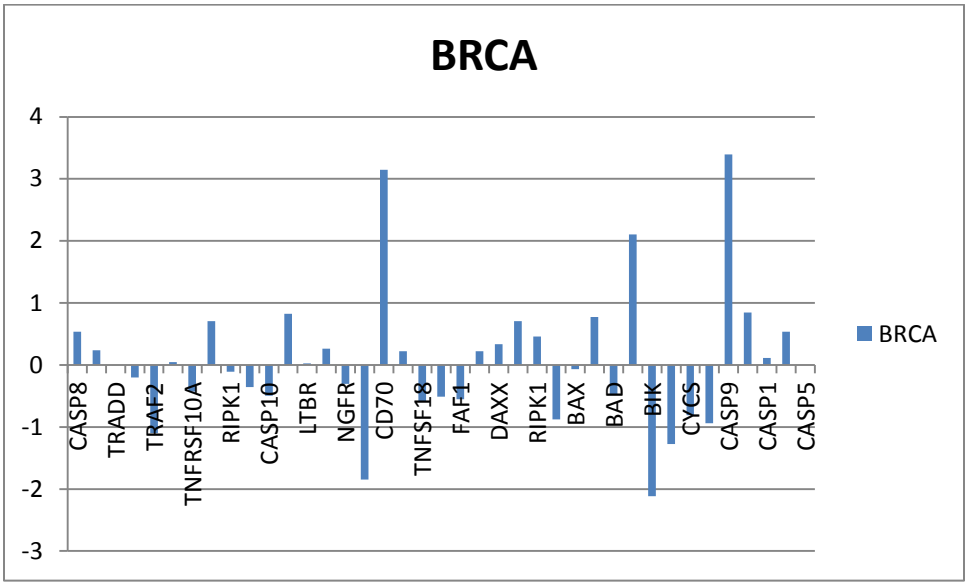
B


KIRP
4
3
2
1
0
-1
-2
-3
CASP8
TRADD
TRAF2
TNFRSF10A
RIPK1
CASP10
LTBR
NGFR
CD70
TNFSF18
FAF1
DAXX
RIPK1
BAX
BAD
BIK
CYCS
CASP9
CASP1
CASP5
KIRP

C


COAD
6
5
4
3
2
1
0
-1
-2
CASP8
TRADD
TRAF2
TNFRSF10A
RIPK1
CASP10
LTBR
NGFR
CD70
TNFSF18
FAF1
DAXX
RIPK1
BAX
BAD
BIK
CYCS
CASP9
CASP1
CASP5
COAD

D



E

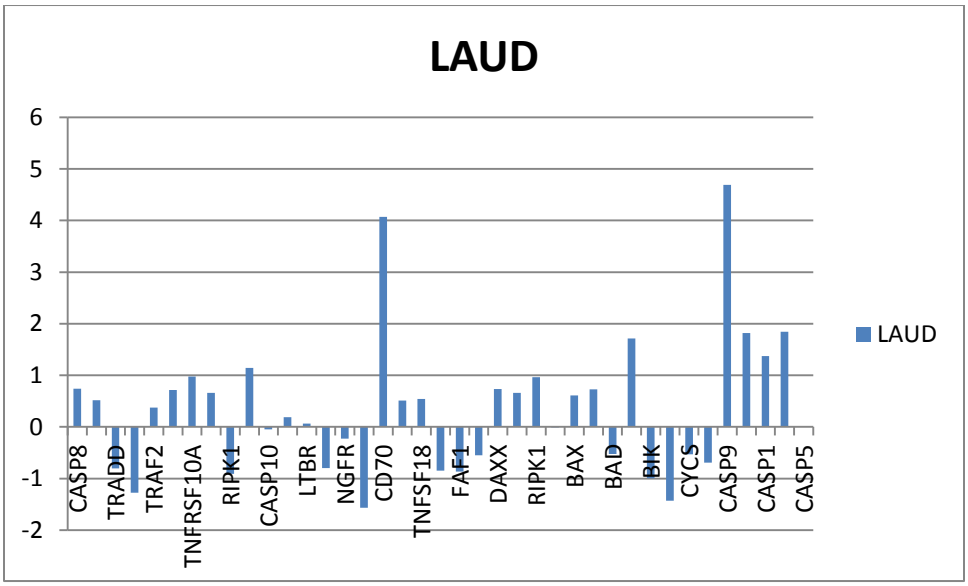


Figure AB.2. The Log2 fold change of genes which are involved in cell death. From A-E: Kidney renal clear cell carcinoma, Kidney renal papillary cell carcinoma; Colon adenocarcinoma; Breast invasive carcinoma; Lung adenocarcinoma. The immune response genes are induced across the cancers (the last four genes).

The individual observation in cancer shows that the cancers suppress the intrinsic pathways while the immune response apoptotic pathways are induced. This implies that cancer cells have lost their ability to induce their own cell death. Most of the genes that are involved in intrinsic apoptotic pathways are consistently repressed. This could be contributed by several reasons; the cancer cells may have acquired mutations in these genes, epigenetic changes could suppress the expression of these genes, or the post-transcriptional regulation due to miRNAs.

Next, stress-response genes are examined in respect to the apoptosis genes. Due to many unknown functions of proteins and the diverse involvement and role that it may have, we examined to distinguish which genes are highly correlated with the apoptotic genes. Moreover, we attempt to identify new genes that may be implicated in cell death. In total, the expression levels of 2168 genes were examined.

We did not find any consistent pattern that was common across all types of cancers. However, a set of genes were strongly correlated (with a correlation coefficient value > 0.9) with the extrinsic apoptotic pathways in rectum adenocarcinoma. Table AB.2 shows the correlation values for genes that were found to be strongly correlated.

Table AB2. Highly correlated genes in rectum adenocarcinoma. The table shows that the genes that were found to be highly correlated with apoptosis genes are specific to the rectum adenocarcinoma.

apoptotic classification	genes	RCA	OAD	IRC	IRP	GG	UAD	USC	V	EAD	CEC

XTRINSIC	RAD18,TRADD	0.004972324	.34428907474	0.224565413	0.364778555	.12743434126	0.088784749	0.100600165	0.083207463	.93183768358	0.058059559
XTRINSIC	RAD18,TRAF2	0.018316039	.32264470202	.16116573535	0.315552396	0.040009473	0.001038166	0.054006689	0.053215178	.93864310606	.02556187272
XTRINSIC	RAD18,TNFRSF10A	0.051108871	.32748748811	0.169475811	0.260435154	0.172374658	0.219539222	.26364532323	.14470797777	.93743024375	0.102275962
XTRINSIC	RAD18,NGFRAP1	0.022380815	.32103463443	0.240035251	.25901672121	.09316882828	.17075350101	0.155538883	0.068616783	.91100171818	0.332548258
XTRINSIC	BAG1,TRADD	0.028550688	.49931631641	0.371575208	0.474947035	.04181727	0.126119179	0.092805856	0.054339992	.93996872626	0.224495129
XTRINSIC	BAG1,TRAF2	0.005309	.4688566	0.053911	0.406341	.0272231	.0405810	0.058357	0.037189	.9442182	0.016641

		658	69	024	643	35	17	843	414	75	564
XTRINSIC	BAG1,TNFRSF10A	0.032770684	.47887238383	0.373818333	0.356760012	0.274753354	0.147831686	.12137983737	.104894476	.944096506	.078757872
XTRINSIC	BAG1,NGFRAP1	.039050083	.47941481313	0.028795243	.3270778 96	.0551830 35	.1986495 08	0.126838789	0.043357 78	.917456444	0.131963995
XTRINSIC	BAG1,FAF1	.0077930 21	.525865103	.0408145 7	0.230482175	.2178891 33	.0313433 74	0.097416476	0.067481813	.907420731	.173916937
XTRINSIC	SLC12A6, TRADD	.0154941 4	.19746029 5	0.101238172	0.239064367	0.025265156	0.126336825	0.086005045	0.069421955	.929971866	0.214587393
XTRINSIC	SLC12A6, TRAF2	0.036002097	.16832851	.1269497 79	0.468018555	.0746866 34	.0405931 02	0.041813244	0.035211172	.934132824	.1065935 57
XTRI	SLC12A6,	0.00	.173	0.04	0.15	0.32	0.14	.190	.172	.940	.016

NSIC	TNFRSF10A	7114618	388233	6064156	8411411	06991	5747586	09813	05996	463282	303136
XTRINSIC	SLC12A6, NGFRAP1	0.015540344	.15094566	0.206529246	.44520703	.02886267	.20179732	0.116262249	0.053309907	.90411904	0.267059005
XTRINSIC	FTH1,TNFRSF10A	0.039385912	.48469742	0.086101338	0.310950975	0.344794035	0.211931764	.10192623	.17904426	.90728181	.00092419
XTRINSIC	TRIO,TNFRSF10A	0.031323789	.53968172	0.307718	0.176810277	0.286122938	0.220947266	.24570008	.06863106	.90622250	.15093264
XTRINSIC	ILK,TRADD	0.026227951	.65639581	0.104005189	0.514710406	0.045321572	.03791679	0.088543484	0.035417586	.94996973	0.055426692
XTRINSIC	ILK,TRAF2	0.048001228	.61498395	.16427202	0.456681777	.08668888	0.069625256	0.050122188	0.002923723	.95140042	0.026605553
	I										

XTRINSIC	LK,TNFRSF10A	.005352194	.627010891	0.094910579	.059302872	0.311693492	0.184498604	.194454868	.165490057	.956300044	.012810801
XTRINSIC	ILK,NGFRAP1	0.01799291	.632097345	0.142337625	.193630757	.008360271	.001526786	0.13476017	0.025472381	.921061734	0.207958408
XTRINSIC	ILK,FAF1	.007206257	.690120777	.076558541	0.245763019	.257132856	.019314818	0.100377597	0.050700378	.912368746	0.040536279
NTRINSIC	ILK,BAD	.028458526	.723130666	.064252085	0.499405185	.303685206	0.20069852	0.054792966	0.036682949	.90200934	0.24194428
XTRINSIC	MYB,TRADD	.67E-06	.540433597	0.260994802	0.385124542	.016820209	0.067519153	0.09111173	0.091829615	.909712938	0.169885464
XTRINSIC	MYB,TRAF2	.006405921	.52783007	.123485196	0.424950685	.048966067	0.015574149	0.039550219	0.076266212	.916189171	.052358506

XTRINSIC	MYB,TNFRSF10A	0.015617336	.534845982	0.21132314	0.19501365	0.29146274	0.23031775	.275919282	.075539113	.918625565	0.06783934
XTRINSIC	ITGB1,TRADD	0.00436900	.455821227	0.22448288	0.13961449	.005076335	0.12741299	0.05134034	0.05595329	.920290586	0.04822972
XTRINSIC	ITGB1,TRAF2	0.05666776	.439919465	.068588082	0.11209652	.055603413	.042556678	.010851392	0.02534414	.919262607	0.00240319
XTRINSIC	ITGB1,TNFRSF10A	0.04592270	.448423105	0.22053147	0.01421901	0.30664818	0.14013828	.369242894	.155934383	.929295227	0.24353831
XTRINSIC	CIDEA,TNFRSF10A	0.04718579	.247617897	0.18010548	0.14193904	0.22965564	0.21583906	.287198591	.238383877	.912158439	0.22703775
XTRINSIC	ZNF3,TNFRSF10A	0.059219	.4563217	0.192492	0.084457	0.201625	0.226462	.2297573	.1185825	.9009361	0.132934

		276	91	364	155	007	159	98	4	7	51
XTRINSIC	LRRC8D,TNFRSF10A	0.036974146	.588397666	0.303200367	0.184830411	0.290356831	0.213712837	.14874602121	.14021110303	.90787055	0.132490133

The set of genes that were highly correlated with apoptosis genes in rectum adenocarcinomna were *RAD18, BAG1, SLC12A6, FTH1, TRIO, ILK, MYB, ITGB1, CIDEA, ZNF3, LRRC8D.* Some of these genes have unknown function and are not known to be involved in apoptotic pathway. It would be an interest to find out how these genes are related to and possibly contribute to cell death.

Next, we examined the expression of genes in respect to TP53 expression. A tumor suppressor gene, TP53, has been repeatedly found to be mutated in various types of cancer [121,122,123,124]. It is involved in the DNA check point and regulation of apoptosis by its association with Bcl-2 [125,126,127]. The expression of TP53 was examined in each cancer (Table AB.3).

Table AB.3. The number of samples according to the TP53 expression. BRCA: breast invasive carcinoma; COAD: colon adenocarcinoma; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LGG: brain lower grade glioma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; OV: ovarian serous cystadenocarcinoma; READ: rectum adenocarcinoma; UCEC: uterine corpus endometrioid carcinoma

Cancer Type	Number of Samples with overexpression of TP53	Number of samples with repressed expression of TP53

BRCA	605	2
COAD	76	105
GBM	1115	38
KIRC	2	72
KIRP	0	16
LGG	5	22
LUAD	0	35
LUSC	295	0
OV	1210	1
READ	34	44

BRCA, GBM, LUSC, and OV cancers had induced expression of TP53, whereas KIRC, KIRP, LGG, and LUAD had decreased level of TP53 activity. We further examined READ and COAD cancers where the samples were either induced or repressed in TP53 expression level. The TP53 were examined to see whether there exists a correlation with the set of apoptosis genes, and to elucidate which apoptotic genes it directly regulates.

Interestingly, no correlation between apoptotic genes and TP53 were observed in COAD as shown in Figure AB.3. However, expression correlations are observed in READ for genes TNFRS10A, LTBR, and BLK, which are all involved in extrinsic apoptotic pathways (Figure AB.4). This finding may indicate that P53 could potentially influence the expression level of apoptotic genes in READ in indirect way. The future direction of this study is to perform in-depth analysis to characterize the function of eleven genes (*RAD18, BAG1, SLC12A6, FTH1, TRIO, ILK, MYB, ITGB1, CIDEA, ZNF3, LRRC8D)*, which were highly correlated with extrinsic apoptotic genes, in cell death.

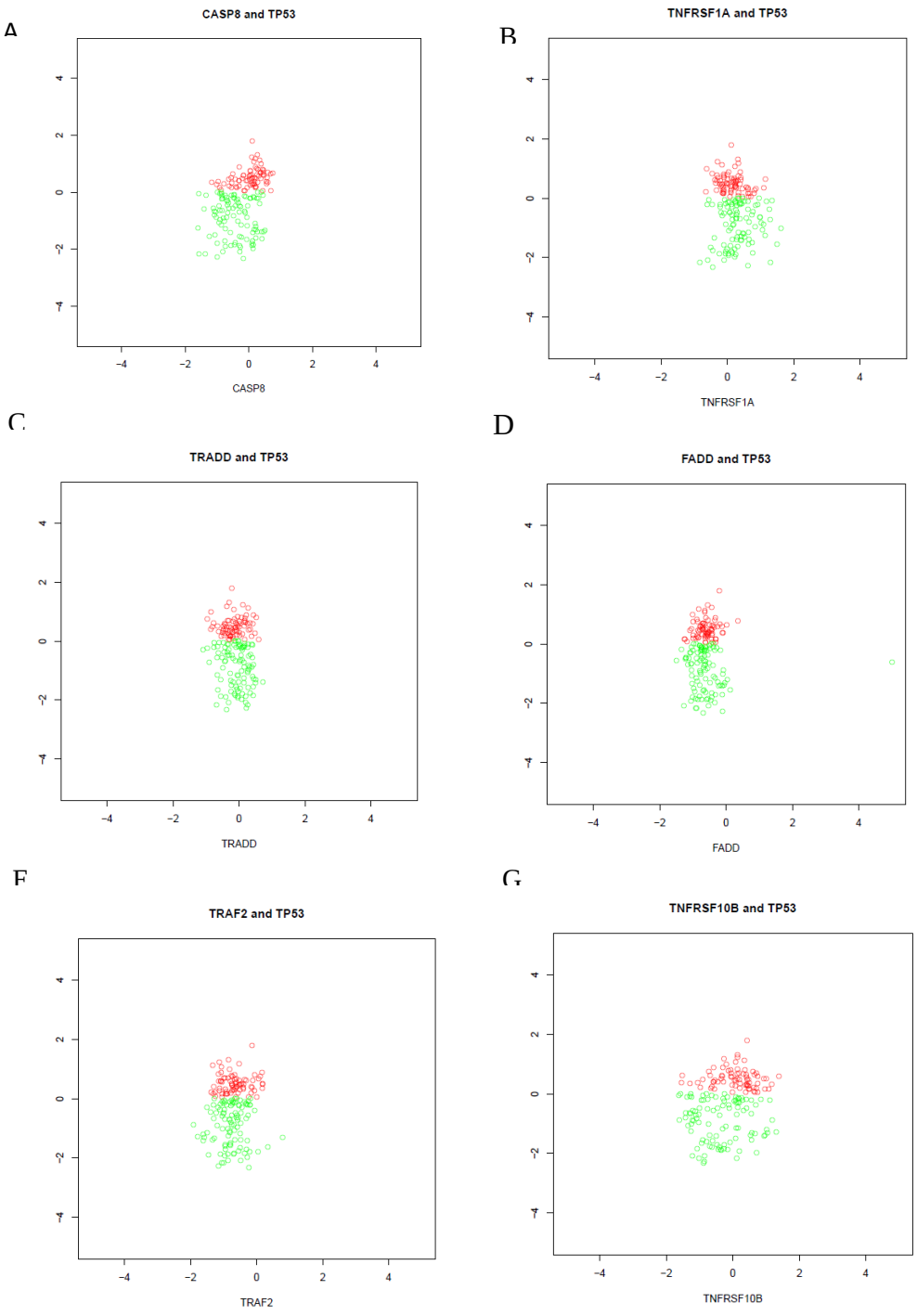
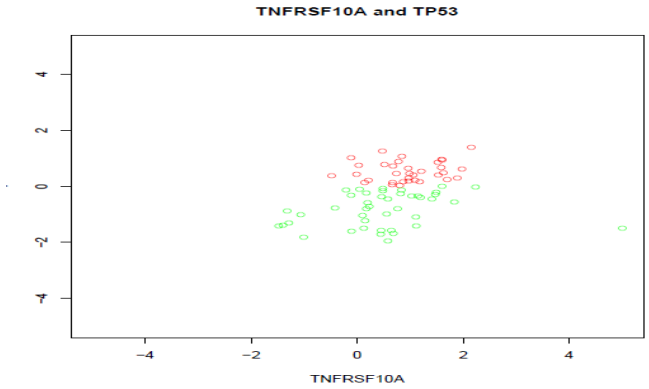
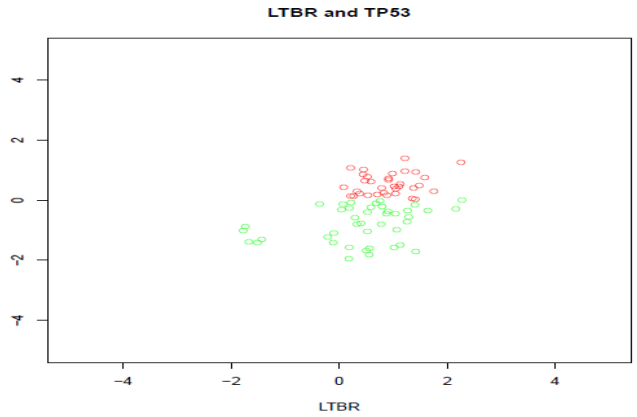


Figure AB. 3. Scatter plot of apoptosis genes *vs* TP53. The x-axis is the expression level of apoptotic genes and the y-axis is expression level of TP53. Red: Samples that show overexpressed TP53; Green: Samples that show repressed TP53.

A



B



C

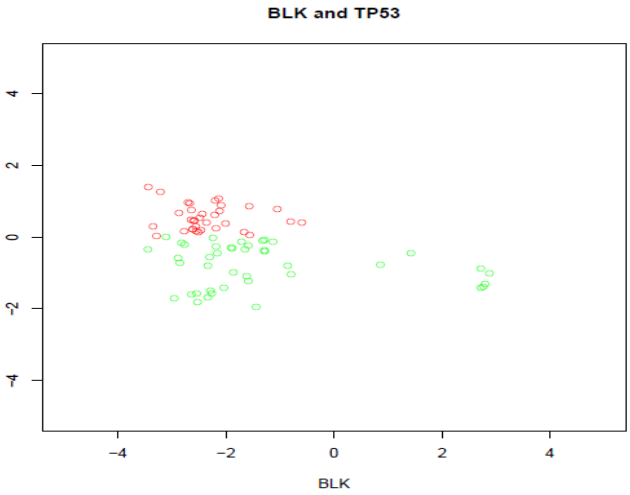


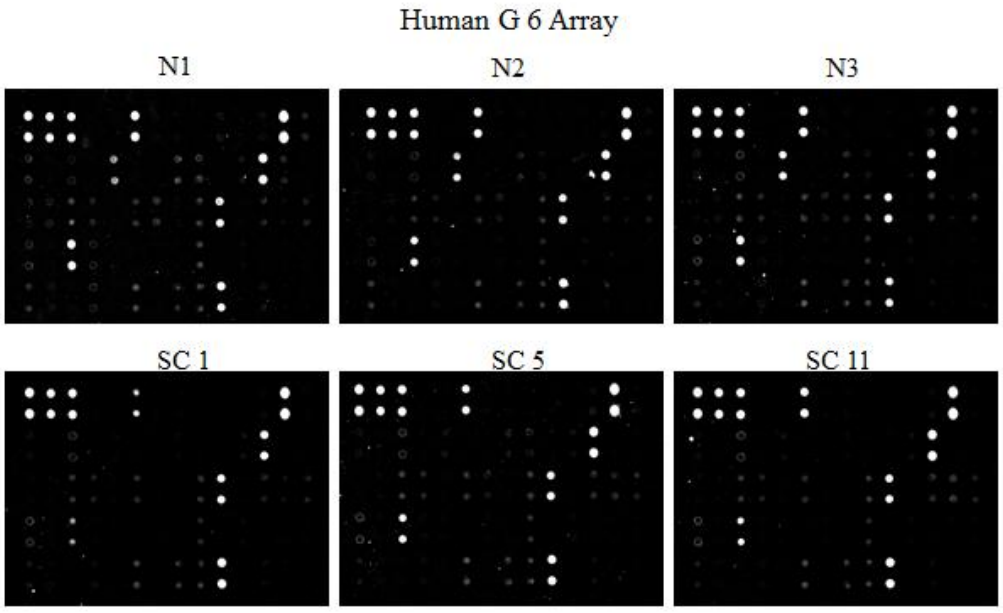
Figure AB. 4. Scatter plot of apoptosis genes *vs* TP53 that shows correlated expression values. The x-axis is the expression level of apoptotic genes and the y-axis is expression level of TP53. Red: Samples that show overexpressed TP53; Green: Samples that show

Appendix C

Protein Array

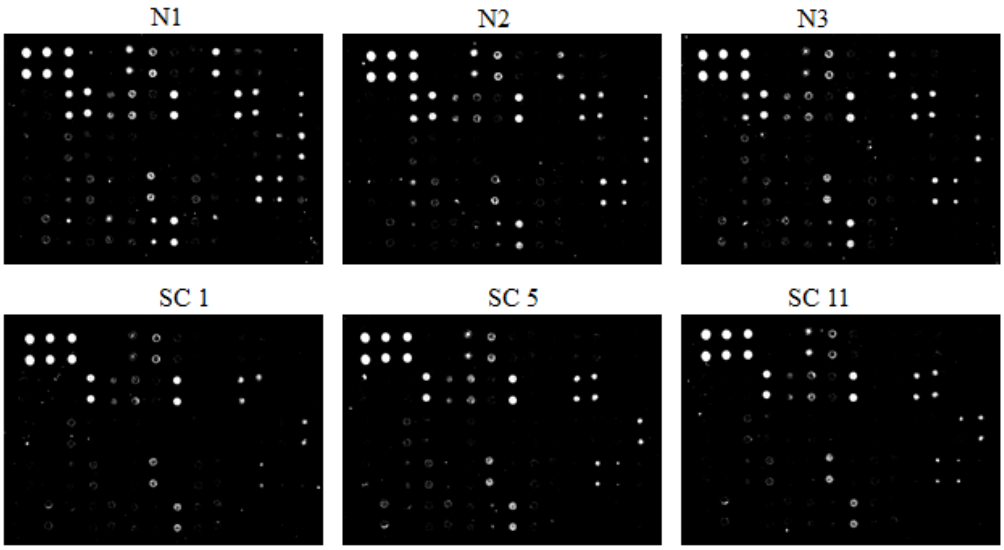
Protein array was used to validate the prediction of excretory proteins. RayBio Human G series Array 4000 (Raybiotech, Norcross, GA) was used to validate the excretory status of 163 proteins. The total of fourteen samples was tested to ensure the quality of the validation; seven were from the gastric cancer patients and seven were from normal healthy people. The classifier predicted that out of 163 proteins, 112 were excreted. Of the 112 proteins, 92 were experimentally validated using protein array (Fig AC. 1).

A



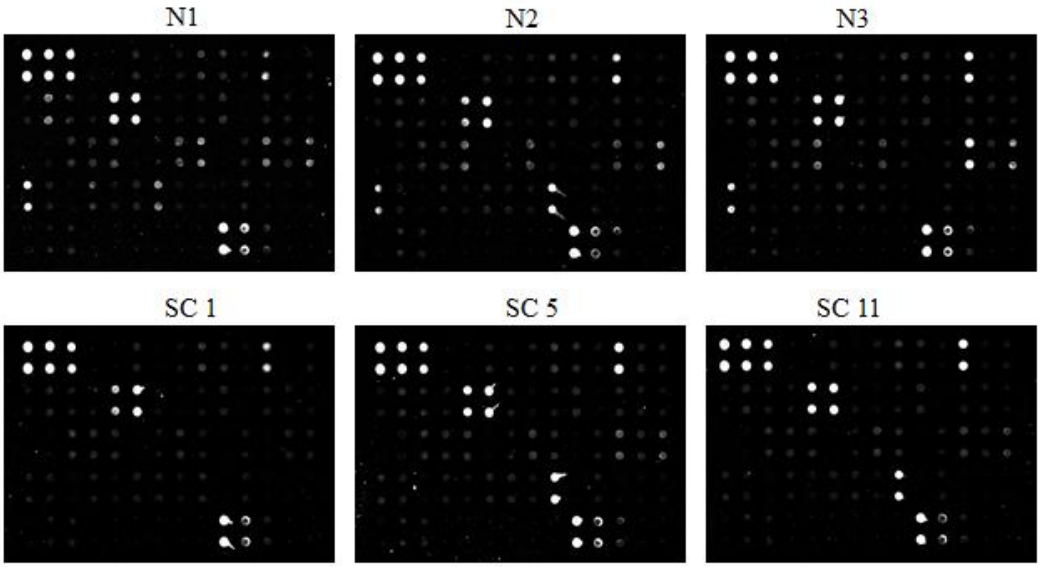
B

Human G 7 Array



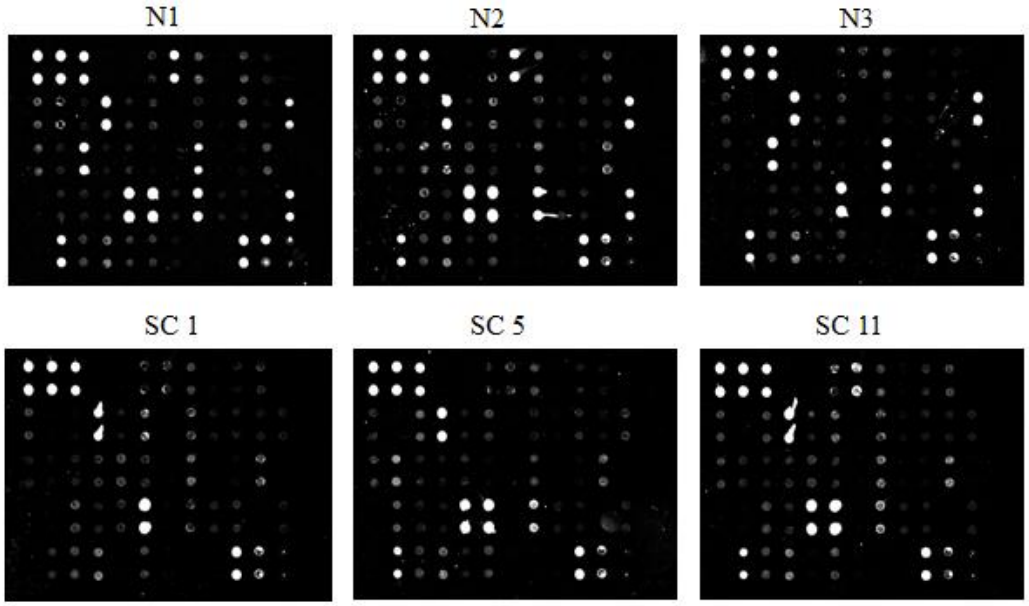
C

Human G 8 Array



D

Human G 9 Array



E

Human G 10 Array

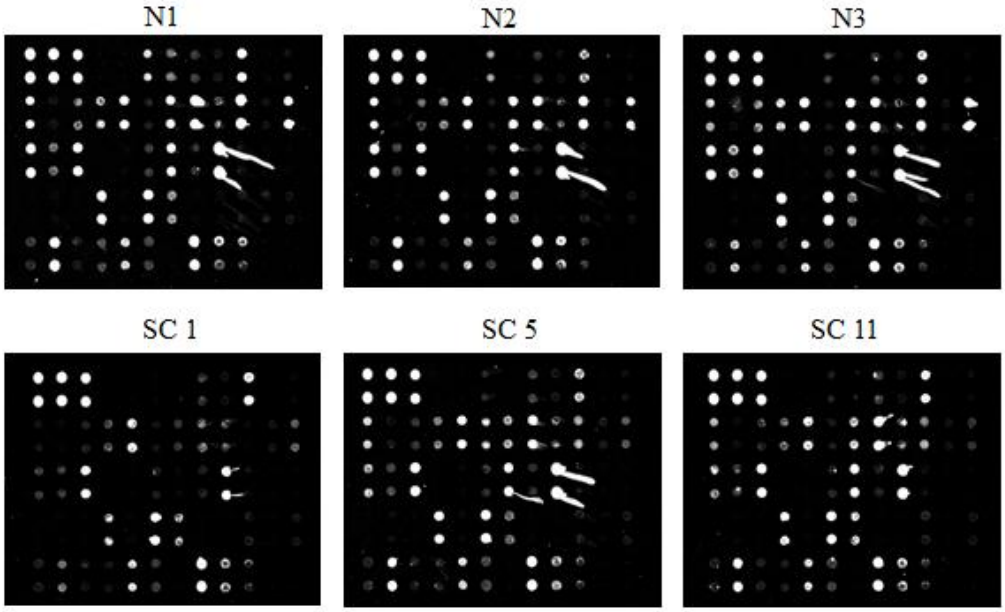
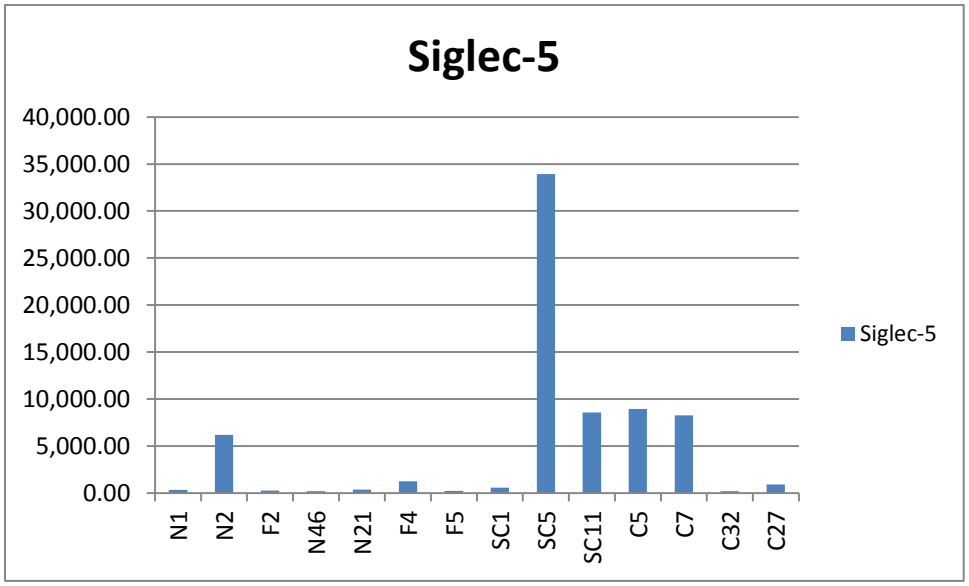


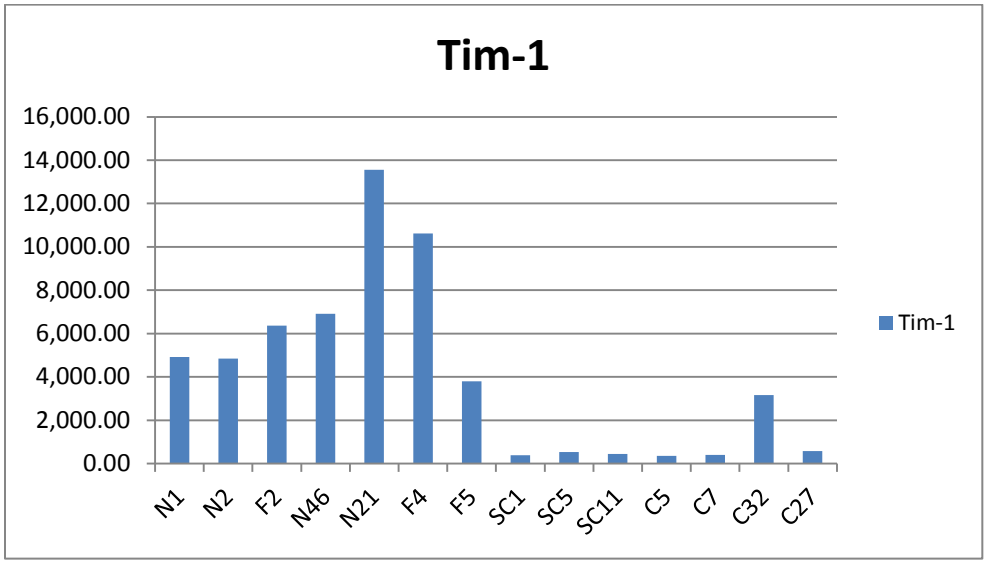
Figure AC. 1. Antibody mediated protein array for excretion validation. N1, N2, N3: Normal Samples; SC1, SC5, SC11: Gastric cancer patient samples. The data for three pairs of samples are shown.

Raw intensity values were normalized across the samples (not shown). As the validation work was performed on seven gastric cancer samples and seven normal samples, differentially expressed proteins were observed that were not mentioned in the published work. Figure AC.2 shows the differentially expressed proteins found in urine in validation study.

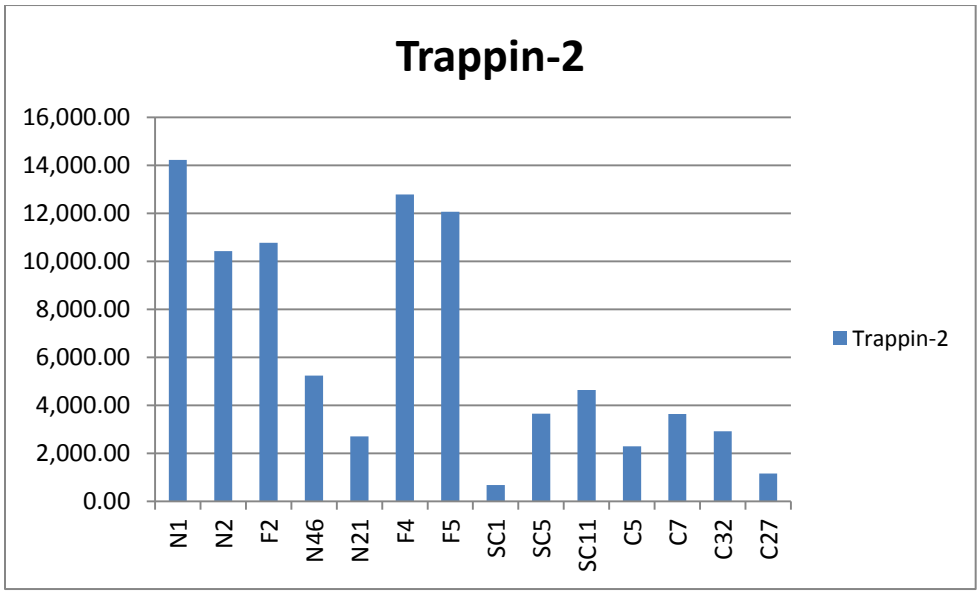
A



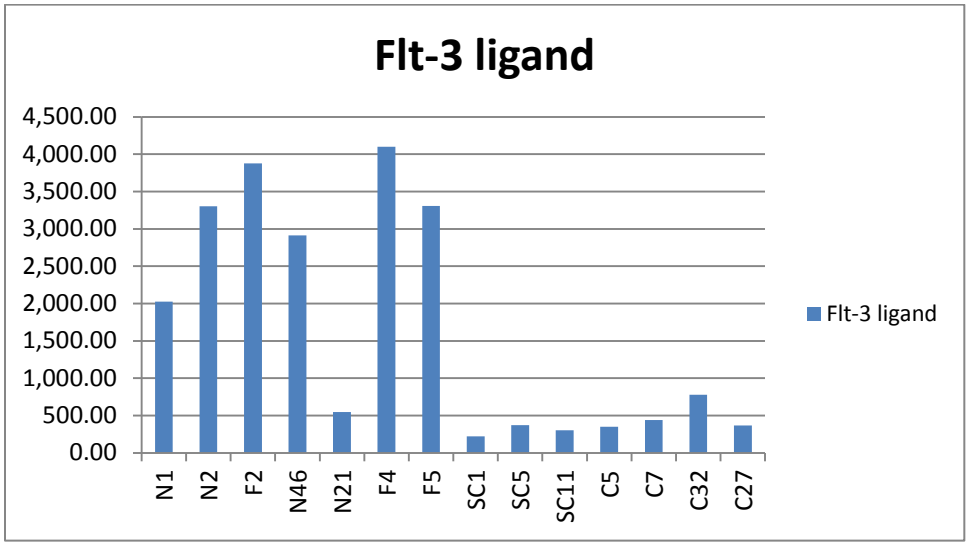
B


Tim-1
16,000.00
14,000.00
12,000.00
10,000.00
8,000.00
6,000.00
4,000.00
2,000.00
0.00
N1
N2
F2
N46
N21
F4
F5
SC1
SC5
SC11
C5
C7
C32
C27
Tim-1

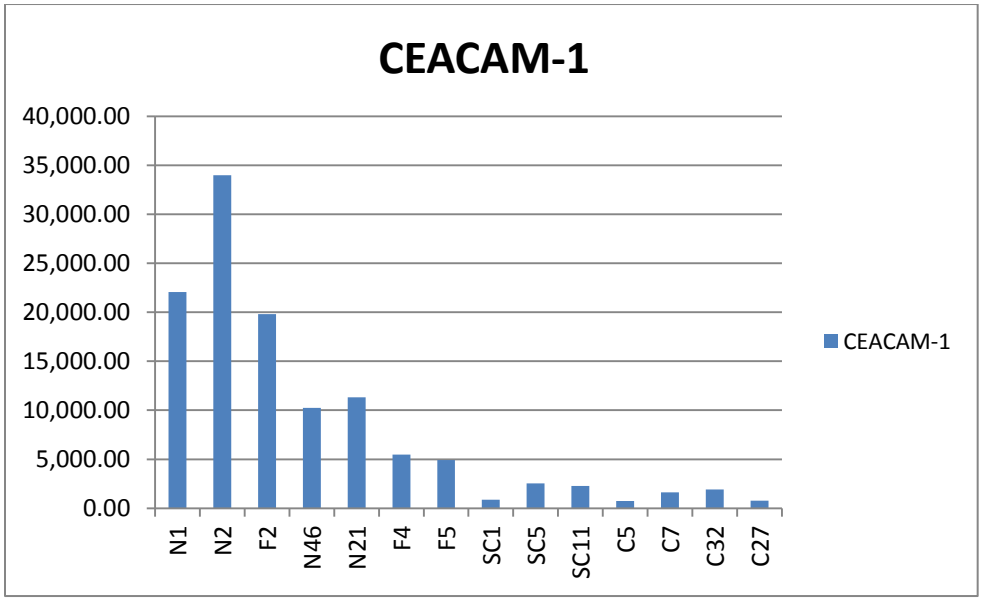
C


Trappin-2
16,000.00
14,000.00
12,000.00
10,000.00
8,000.00
6,000.00
4,000.00
2,000.00
0.00
N1
N2
F2
N46
N21
F4
F5
SC1
SC5
SC11
C5
C7
C32
C27
Trappin-2

D


Flt-3 ligand
4,500.00
4,000.00
3,500.00
3,000.00
2,500.00
2,000.00
1,500.00
1,000.00
500.00
0.00
N1
N2
F2
N46
N21
F4
F5
SC1
SC5
SC11
C5
C7
C32
C27
Flt-3 ligand

E


CEACAM-1
40,000.00
35,000.00
30,000.00
25,000.00
20,000.00
15,000.00
10,000.00
5,000.00
0.00
N1
N2
F2
N46
N21
F4
F5
SC1
SC5
SC11
C5
C7
C32
C27
CEACAM-1

F

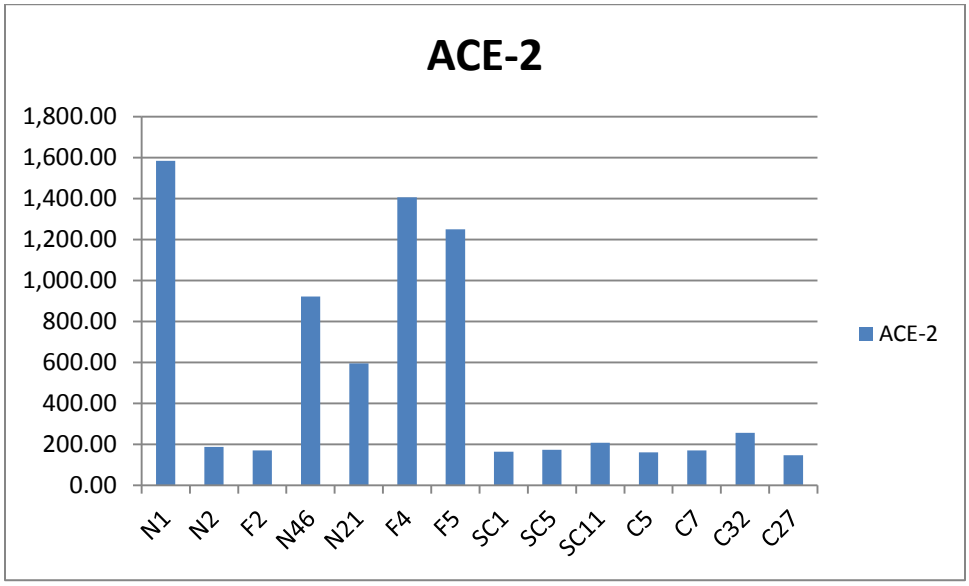


Figure AC.2. Differentially expressed proteins found in validation study. First seven samples are from healthy people (N1, N2, F2, N46, N21, F4, F5) and the last seven samples are from gastric cancer patients (SC1, SC5, SC11, C5, C7, C32, C27).