Multi-omics Data Integration for Identifying Disease Specific Biological Pathways

Yingzhou Lu

Thesis submitted to the faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

Yue Wang, Chair

Guoqiang Yu

Tam Chantem

May 2, 2018

Arlington, Virginia

Keywords: Biological Pathways, Multi-omics Data Integration, Muscular Dystrophy, Statistical significance test, Gene set enrichment analysis

Multi-omics Data Integration for Identifying Disease Specific Biological Pathways

Yingzhou Lu

(ABSTRACT)

Pathway analysis is an important task for gaining novel insights into the molecular architecture of many complex diseases. With the advancement of new sequencing technologies, a large amount of quantitative gene expression data have been continuously acquired. The springing up omics data sets such as proteomics has facilitated the investigation on disease relevant pathways.

Although much work has previously been done to explore the single omics data, little work has been reported using multi-omics data integration, mainly due to methodological and technological limitations. While a single omic data can provide useful information about the underlying biological processes, multi-omics data integration would be much more comprehensive about the cause-effect processes responsible for diseases and their subtypes.

This project investigates the combination of miRNAseq, proteomics, and RNAseq data on seven types of muscular dystrophies and control group. These unique multi-omics data sets provide us with the opportunity to identify disease-specific and most relevant biological pathways. We first perform t-test and OVEPUG test separately to define the differential expressed genes in protein

and mRNA data sets. In multi-omics data sets, miRNA also plays a significant role in muscle development by regulating their target genes in mRNA dataset. To exploit the relationship between miRNA and gene expression, we consult with the commonly used gene library - Targetscan to collect all paired miRNA-mRNA and miRNA-protein co-expression pairs. Next, by conducting statistical analysis such as Pearson's correlation coefficient or t-test, we measured the biologically expected correlation of each gene with its upstream miRNAs and identify those showing negative correlation between the aforementioned miRNA-mRNA and miRNA-protein pairs. Furthermore, we identify and assess the most relevant disease-specific pathways by inputting the differential expressed genes and negative correlated genes into the gene-set libraries respectively, and further characterize these prioritized marker subsets using IPA (Ingenuity Pathway Analysis) or KEGG. We will then use Fisher method to combine all these p-values derived from separate gene sets into a joint significance test assessing common pathway relevance. In conclusion, we will find all negative correlated paired miRNA-mRNA and miRNA-protein, and identifying several pathophysiological pathways related to muscular dystrophies by gene set enrichment analysis.

This novel multi-omics data integration study and subsequent pathway identification will shed new light on pathophysiological processes in muscular dystrophies and improve our understanding on the molecular pathophysiology of muscle disorders, preventing and treating disease, and make people become healthier in the long term.

Multi-omics Data Integration for Identifying Disease Specific Biological Pathways

Yingzhou Lu

(GENERAL AUDIENCE ABSTRACT)

Identification of biological pathways play a central role in understanding both human health and diseases. A biological pathway is a series of information processing steps via interactions among molecules in a cell that partially determines the phenotype of a cell. Specifically, identifying disease-specific pathway will guide focused studies on complex diseases, thus potentially improve the prevention and treatment of diseases.

To identify disease-specific pathways, it is crucial to develop computational methods and statistical tests that can integrate multi-omics (multiple omes such as genome, proteome, etc) data. Compared to single omics data, multi-omics data will help gaining a more comprehensive understanding on the molecular architecture of disease processes.

In this thesis, we propose a novel data analytics pipeline for multi-omics data integration. We test and apply our method on/to the real proteomics data sets on muscular dystrophy subtypes, and identify several biologically plausible pathways related to muscular dystrophies.

Acknowledgments

I would like to take this opportunity to express my gratitude to everyone who ever helped me during my entire course of study.

First and foremost, there are not enough words to describe how thankful I am to my parents for their continued love and encouragement, they support me in all my pursuits. I wish I could let them be proud of me.

Secondly and equal importantly, I would like to express my deepest gratitude to Dr. Yue Wang for all the academically and personally support. I have been so lucky to have Dr. Wang as my supervisor and thanks to him I had the opportunity to start the integrated pathway prioritization project. His enthusiasm, motivation, and immense knowledge would have a long-term impact on me.

My sincere thanks also go to Dr. Guoqiang Yu for the academic guidance and Dr. Tam Chantem for the time and patience, their inspiration and insightful comments are very helpful to the thesis.

Last but not least, I would like to say thank you to all my labmates of Computational Bioinformatics and Bioimaging Laboratory at Virginia Tech. Thank to Yi Tan Chang for encouragement and sincere suggestions. Thank Lulu Chen and Chiung-Ting Wu, I appreciate all their contributions of time and ideas whenever I need help. The group has been a source of friendships as well as good advice and collaboration.

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

1.1 Background

miRNAs are a group of small noncoding RNAs that regulate their downstream target mRNA post-transcriptionally or proteins post-translationally. It is biologically expected that such miRNA-mRNA or miRNA-protein regulation is mainly suppression. In other words, mathematically, negative correlation is expected [1] .

Though several explorations have been made for integrative analysis, the technology restriction and biology complexity have made multi-omics data integration a challenging task[2].

The muscular dystrophies are inborn errors of metabolism resulting in muscle weakness, wasting. The philosophy here is to apply the additional more stringent criterion to assess the initial mRNA/protein-pathway association by examining whether the differentially expressed 'downstream readout' miRNA-mRNA/protein is expectedly regulated by some differentially expressed miRNA.

1.2 Motivation

Previous studies have defined the molecular pathogenesis of dystrophinopathies, sarcoglycanopathies, dysferlinopathy, myotonic dystrophy, and nuclear envelop dystrophies.

RNA-sequencing (RNA-seq) is the most recent and prominent way to conduct gene expression profiling[3]. With the development of RNA-seq, biologists are able to explore genomewide profiling of RNA and protein more accurately and at the same time reduce costs.

The preliminary data identifying a novel microRNA-mRNA-protein network associated with severity of fibrosis and failed regeneration, centered on a microRNA/chaperonin-T pathway. The integration of miRNAseq, proteomics, and RNAseq data on seven types of muscle disease plus normal controls will lead to new systems models of muscle disease and predict response to therapies.

1.3 Statement of Problem

The relationships between miRNA and target genes in mRNA/protein dataset are quite complex, for example, miRNAs may target multiple mRNA/proteins[4]. To explore the miRNA-mRNA/protein pairing, we looked into miRNA- mRNA and miRNA-protein correlations. Furthermore, we integrated multi-omics data to provide relative probabilities of disease-specific induction and loss of specific pathways. Using the 40 human muscle biopsy dataset with existing RNAseq-ribozero, PacBio splicing data, microRNAseq, and SILAC proteomics, define disease-specific networks and assign a relative confidence to these networks by use of Bayesian relative probability models.

1.3.1 Dataset

The large-scale multi-omics dataset is derived from 40 human muscle biopsy tissue samples of 7 mutation-defined muscular dystrophies and normal controls. Each biopsy has had generation of deep RNAseq-ribozero, microRNAseq, and SILAC proteomics (~2,000 proteins quantitatively assayed in each biopsy), with a subset analyzed for mRNA splicing via long-range PacBio sequencing, they carefully matched histopathology and choose 5 representative and well-preserved muscle biopsies per group.

Data Type	Number of Unit
protein	692
mRNA	24835
miRNA	869

Table 1: dataset composition

Table: Disease Group	
Disease 1	DMD
Disease 2	COLVI
Disease 3	LMNA
Disease 4	RYR1
Disease 5	ANO5
Disease 6	BMD
Disease 7	CAPN3

Table 2: muscular disease type for dataset

Currently our investigation is focus on two disease DMD and LMNA. Duchenne muscular dystrophy (DMD) is a type of dystrophin deficiency associated with muscle wasting with the voluntary muscles[5].The DMD disease type is the one we interested. DMD is a progressive muscular disease, at the same time, the most prevalent muscular dystrophy in children. The DMD biopsies were all in the early stage of disease (at diagnosis at about 5 yrs of age) as noted

above, but can be considered the most severe disease, thus showing the highest number of protein differences.

LMNA (lamin A/C) is another disease type we explored in multi-omics. LMNA would cause Congenital Muscular Dystrophy (L-CMD)[6], which is a kind of genetic disorder displays in axial weakness, and may cause progressive muscle weakness and degeneration.

1.3.2 Previous Work

Several efforts have been made to accumulating genomics data and explore omics data[7], In 2013, Berger [6] et al. described some mathematics and statistical approaches for integrating genomics data. Bersanelli et al. later classify diverse approaches of multi-omics integration into two categories. First is “network-based” theory which takes advantage of the currently known relationship to draw a graph to see the interactions between different variables. Second is using the Bayesian model to calculate the posterior probability distribution for analyzing the multi-omics data.

A number of questions regarding the integration of multi-omics remain to be addressed. First, because biological processes are often complex in essence, it is difficult to consolidate various types of genomic data, the data sets may have different scales with its own characteristics[8]. Mathematical method and statistical test are necessary to make the integration reasonable. Second, there could be many noises in addition to a large proportion of missing data. Matrix completion and other pre-processing steps are required. Third, it is not easy to understand the biological meaning of pathways, and how it works on changing a cell or a certain target[9].

1.4 Contribution

A large number of existing studies in the broader literature have examined that pathway exploration are pivotal for it may lead to more personalized strategies for preventing and treating disease, and make people become healthier in the long term.

Although several research has focused on single omics data, only a few studies have combined multi-omics data, notwithstanding to apply in the muscular disease field. Our research is novel and meaning, for which can help understand the pathogenesis of muscular disease and how they respond to the therapy, and help illuminate the how gene expressed and regulated in muscle and muscular disease.

Chapter 2 – Design

To measure the regulation and correlation more preciously, more sophisticated methods are developed and make comparisons, since miRNA and mRNA or protein are two different omics, we cannot simply use numeric to define which is up-regulated and which is down-regulated and correlation between miRNA-mRNA, miRNA-protein. The methods we applied including fold change, correlation coefficient, and t-test. We also invesitage on how to reasonably integrate the multi-omics data from miRNA, mRNA and protein.

2.1 Method

2.1.1 Fold Change

A common way to measure upregulation or downregulation in biological research is fold change. Suppose we have t samples for case and control group respectively per gene case for mRNA, protein dataset, and the miRNA regulated by the gene

$Case: x_1, x_2, x_3, \dots, x_t$

$Control: y_1, y_2, y_3, \dots, y_t$

$$\overline{x} = \frac{\sum_{i=1}^{t} x_i}{t}$$

$$\overline{y} = \frac{\sum_{i=1}^{t} y_i}{t}$$

If $\overline{x} > \overline{y}$, we consider the gene pair as up-regulated, otherwise, it would be considered as down-regulated.

2.1.2 Correlation Coefficient

Correlation Coefficient can quantify both the direction and strength of the linear correlation between two variables. Several types of correlation coefficient have been proposed, like Pearson's correlation (also called Pearson's R, commonly used in linear regression); Spearman correlation(defined as r_s) which depicts monotonic relationships that are not necessarily be linear and Kendall rank correlation, evaluates the degree of similarity between two rank sets of variables. Considering the pros and cons of each correlation coefficient type, Pearson correlation coefficient was selected to measure the linear correlation.

We are interested in the relationship between the following two variables, X and Y

-X: case and control groups combined data from omics data platform1 (miRNA)

-Y: case and control groups combined data from omics data platform2 (mRNA or protein)

We will select the minimal p-value to solve one gene target with multiple miRNA problems, and calculate Pearson correlation coefficient value which ranges from -1 to 1 (1 is a total positive linear correlation, 0 is no linear correlation, and −1 is a total negative linear correlation),

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

σ_X: the standard deviation of X, σ_Y: the standard deviation of Y

μ_X : the mean of X, μ_X $=E[X]$

μ_Y : the mean of Y, μ_Y $=E[Y]$

the covariance: $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$

$$=E[(X-E[X])(Y-E[Y])]=E[XY]-E[X]E[Y]$$

$$\sigma_X = \sqrt{E[(X - E[X])^2]} = \sqrt{E[X]^2 - [E[X]]^2}$$

$$\sigma_Y = \sqrt{E[(Y - E[Y])^2]} = \sqrt{E[Y]^2 - [E[Y]]^2}$$

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{E[XY]-E[X]E[Y]}{\sqrt{E[X]^2-[E[X]]^2}\sqrt{E[Y]^2-[E[Y]]^2}}$$

the p-value is calculated using a *t*-distribution with *n* – 2 degrees of freedom

$$t = \frac{r\sqrt{n-2}}{1-r^2}$$

n: number of observations

r: correlation coefficient value

The correlation obtains for each gene in the database will show whether it positive or negative correlated in different omics and statistically significant at the certain p-value.

2.1.3 t-test

The sample size in this problem is comparatively small, only ten (five for case and five for control), so the t-test is a good option for testing the difference between samples[10].

To determine whether two omics data are significantly different from each other, I conducted a two-sample t-test on miRNA- mRNA/miRNA-protein dataset.

The null hypothesis is

$$H_0 : \mu_1 = \mu_2$$

Which means the mean of two groups of samples are the same, the null hypothesis is the pairwise difference between the two tests is equal

The alternative hypothesis

$$H_1 : \mu_1 \neq \mu_2$$

means the mean of two groups of samples are different

The $x_1, x_2, x_3, \dots, x_{n_1}$ are independent variables from dataset1,

The $y_1, y_2, y_3, \dots, y_{n_2}$ are independent variables from dataset2,

The likelihood function is

$$L(\mu_1,\ \mu_{2,}\sigma^2 | x, y) = \frac{1}{(2\pi\sigma^2)^{\frac{(n_1+n_2)}{2}}} exp\text{-}\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{i=1}^{n_2}(y_i - \mu_2)^2\right)$$

the overall maximum likelihood estimators are

unbiased estimators for the mean

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

and the variance

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

The overall maximum likelihood estimators are

$$\mu_1 = \overline{x},\ \mu_2 = \overline{y}$$

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}$$

The test statistic

$$T(x, y) = \frac{\overline{x} - \overline{y}}{\sqrt{(\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2})(\frac{(\mathrm{n}_1 - 1)s_1^2 + (\mathrm{n}_2 - 1)s_2^2}{\mathrm{n}_1 + \mathrm{n}_2 - 2})}}$$

The likelihood ratio

$$\Lambda(x, y) = \left(\frac{n_1 + n_2 - 2}{n_1 + n_2 - 2 + T(x, y)^2}\right)^{\frac{2}{(n_1 + n_2)}}$$

Where $n_1 + n_2 - 2$ aredegree of freedom,

The *t* statistic for testing if the means are significantly different

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2.2 Approach

2.2.1 Find Differential Expressed Genes

The new dataset provides us the new opportunities to explore muscular disease. To better analyze expression data and find pathways, I used some gene selection methods to find differential expressed genes. Given normalized data, differential expressed genes are defined by adjusted p-values for multi-group testing.

The first strategy is using the permutation. Given five case and five control samples for each gene,

case: $x_1, x_2, x_{3,} x_{4,} x_5$ control: $x_{1'}, x_{2'}, x_{3'}, x_{4'}, x_{5'}$

we take average for case and control group seperately

case average: $\bar{x}$ control average: $\overline{x'}$

$$\Delta t_0 = \overline{x'} - \bar{x}$$

five-combinations of ten elements set

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{10!}{5!\,5!} = 252$$

obtain the case and control for generated data

$$\Delta t_1, \Delta t_2, \ldots \Delta t_{252}$$

calculate p-value for statistical significance, and set threshold to get differentially expressed gene.

Several other approaches have been proposed for multiple group comparison. OVEPDG, which proposed by Yu (et al.,2010) is selecting the union of one-versus-everyone (OVE) phenotypic up-regulated genes (PUGs) [3] and phenotypic down-regulated genes (PDGs).

OVEDEG=OVEPUG+OVEPDG

OVEPDG computes one-versus-everyone PDG-statistic and estimates its null distribution by weighted permutation scheme. There are several advantages for OVEPDG, first, it can greatly

control false discovery rate (FDR). In addition, the DEG found by OVEPDG method is highly consistent with the unsupervised CAM with empirical cutoff[11].

2.2.2 Measure Upregulation/Downregulation

There is a wide choice of measuring data change. Fold changes, which are defined directly in terms of ratios, is frequently used in genomics.

Compare a disease and a normal sample, then the fold change ratio for the *i*th gene would be

$$FC = \frac{Disease}{Norm}$$

There are some disadvantages of fold change, which treating up-regulated and down-regulated genes differently[4]. Biostatisticians proposed an alternative method to measure changes-using transformation of the ratio is the logarithm base 2, which would equally treat up-regulated and down-regulated genes and make value spectrum continuous[12].

$$log_2 FC = log_2 Disease - log_2 Norm$$

The algorithm would be compared with 0, if the result is positive, then we define it is as up-regulated otherwise it is down-regulated.

2.2.3 Multiomics Data Comparison

Matched microRNA and mRNA/protein profiling were obtained from miRTarVis to visualize the miRNA and protein interactions (Jung et al. 2015). After knowing their interactions, we start analyzing their profiling under various conditions.

There are three conditions in miRNA- mRNA / miRNA-protein pairing,

(1) One-to-One miRNA- mRNA /protein pair

(2) Multiple-to-One miRNA- mRNA/protein pair

(3) One-to-Multiple miRNA- mRNA/protein pair

Condition (1) is comparatively easy to justify, use fold change or correlation coefficient to see whether it is anti-correlation or not.

Condition (2) is more complicated. Bioinformatics analysis indicates that a specific miRNA can regulate expression of up to thousand mRNAs through miRNA-mRNA association, and a specific mRNA can be regulated by multiple miRNAs. We will look into its correlation in miRNA- mRNA / miRNA-protein seperately, and select one pair which we are most confident for each gene.

For condition (3) because we only consider mRNA /protein 'readout' as the unit of pathway enrichment analysis, this scenario is the same as (1) after ungrouped.

In chapter 2.2, I explained the Pearson correlation coefficient method, using it to measure the linear correlation between two variables X and Y:

-X: case(LMNA) and control(Norm) group's combined data for gene-related miRNA

-Y: case(LMNA) and control(Norm) group's combined data for the gene in mRNA/Protein,

If the correlation coefficient value is negative, the pair would be counted as negative correlated. By conducting the same operation to entire dataset, we can find those genes which are of negative correlation.

2.2.4 Gene Set enrichment Analysis

In order to find the biological implications of muscle or muscular dystrophies gene expression changes, enrichment analysis was implemented on DEGs(differentially expressed genes). A commonly used method to identify pathways is gene set enrichmentanalysis. Pathways can be

found by summarizing the significant genes we found and then querying such gene lists to prior knowledge gene-set libraries. Most commonly used gene-set libraries we use are KEGG, WikiPathways and Reactom. Submit only the members of the selected paired units to IPA for further assessing the top pathways.

The gene-set libraries used three criteria to ranking pathways: p-value, z-score and combinef score. P-value computed by using the Fisher exact test, the null hypothesis is the pathway is randomly selected, if the p-value is very small, it means the pathway has many genes overlap with our gene list. we can reject the null hypothesis, and the possibility of the pathway being randomly selected is very small.

There are three criterion for evalute pathways. P-value, z-score and combined value. Binomial distribution and independence for the probability of any gene belonging to any set. Z-score computed by assessing the deviation from the expected rank.

Combined score considers both the p-value from the Fisher exact test and z-score of the deviation from the expected rank

$$c = \log(p) \cdot z$$

2.2.5 Fisher's Method to Combine P-values

Consider a set of k independent tests, each of these to test a certain null hypothesis. For each test, a significant p-value p_i, is obtained.

Fisher's method combines p-values into one test statistics χ^2

$$\chi_{2v}^2 \sim -2\sum_{i=1}^{k} \ln(p_i)$$

χ^2 is a chi-squared distribution with $2v$ degrees of freedom, assume $Z_1, Z_2, \dots, Z_v$, are independent and obey normal distribution[13],

then χ^2 equals to the sum of their squares

$$\chi^2 = Z_1^2 + Z_1^2 + \cdots + Z_v^2$$

Where v: degree of freedom, $v > 0$

The probability density function would be

$$f_v(t) = \frac{1}{2^{\frac{v}{2}}\,\Gamma(\frac{v}{2})}\; t^{\frac{v-2}{2}} e^{-\frac{v}{2}}$$

where $\Gamma(\cdot)$ is the Gamma Function[14]

computes chi-square Cumulative Distribution Function $P(\chi^2|v)$ at each of the values in χ^2 using the corresponding degrees of freedom in v [15]

$$P(\chi^2|v) = \int_0^{\chi^2} f_v(t)dt = \int_0^{\chi^2} \frac{t^{\frac{v-2}{2}}\, e^{-\frac{t}{2}}}{2^{\frac{v}{2}}\;\Gamma(\frac{v}{2})}\; dt$$

The upper tail $Q(\chi^2|v)$

$$Q(\chi^2|v) = \int_{\chi^2}^{+\infty} f_v(t)dt = 1 - P(\chi^2|v)$$

Chapter 3 – Implementation

This experimental design was employed upon the idea in chapter 2. Statistical analyses were performed with different methods. The main focus of the experiments was to find diffential expressed genes within all the dataset we have, figuring out the up-regulation or down-regulation tendency for each gene, and calculating correlation coefficient across various omics data for each gene pair.

3.1 R Implementation of Finding DEG

For permutation method,

	Dataset	LMNA
P=0.05	mRNA	6243
P=0.01	mRNA	1883
P=0.05	protein	415
P=0.01	protein	117

Table 3: genes found using permutation

The finding Differential Expressed Genes are based on OVEPUG. The OVEPUG are modified to OVEPDG to find genes significantly lower expressed in one disease group than other normal and other disease groups. P-value and Z score are calculated for each gene and adjusted for multiple testing corrections. When p-value is very small (like smaller than 0.05), it represents the gene is

significantly higher or lower in one group than another groups, and it belongs to the expected differential expressed genes. Q values are p-values processed with false discovery.

For each gene, the data are divided into three groups, DMD, normal, 6-other subtype diseases analysis. 238 genes within the protein are identified as differentially expressed with q<0.01 while 442 genes within miRNA are identified as differentially expressed with q<0.055.

False Discovery Rate

False discovery is likely to happen in multiple testing. To address this problem, the false discovery rate suggested by Benjamini and Hochberg are applied to the project.

The proportion of errors of false discoveries among the discoveries (rejections of the null hypothesis),

$$Q = \frac{V}{V + S}$$

Where V means when the test is declared significant, the number of null hypothesis are rejected while null hypothesis is true, S refers to the number of the null hypothesis should and correctly be rejected,

False Discovery Rate Q_e can then be calculated with

$$Q_e = E(Q) = E\left(\frac{V}{V + S}\right)$$

A two-tailed t-test was performed on our RNA-seq data with the default settings of random seed number through 1000 permutations to get the p-values, then tested with false discovery rate (FDR) analysis to get the q-values.

	Total	ANO5	BMD	Capn3	ColVI	DMD	LMNA	Norm	Ryr1
p=0.01	610	44	35	39	72	236	75	57	52
p=0.05	2625	133	211	62	642	928	435	83	131
p=0.1	4223	226	325	81	1216	1436	619	122	198

Table 4: Differential Expressed Genes Found via OVEPUG

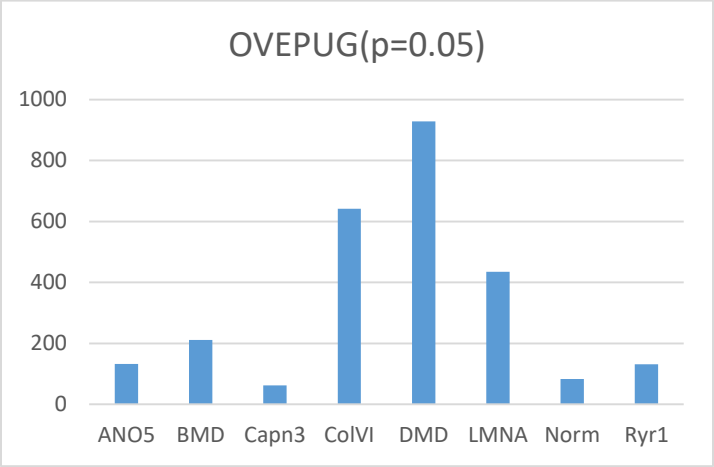


Figure 1: Histogram for OVEPUG (p=0.05)

3.2 Matlab Implementation of Measuring Up-regulation/Down-regulation

The most common approach is calculating fold change between treatment and control group for each gene's expression.

3.3 Matlab Implementation of Measuring Anti-Correlation

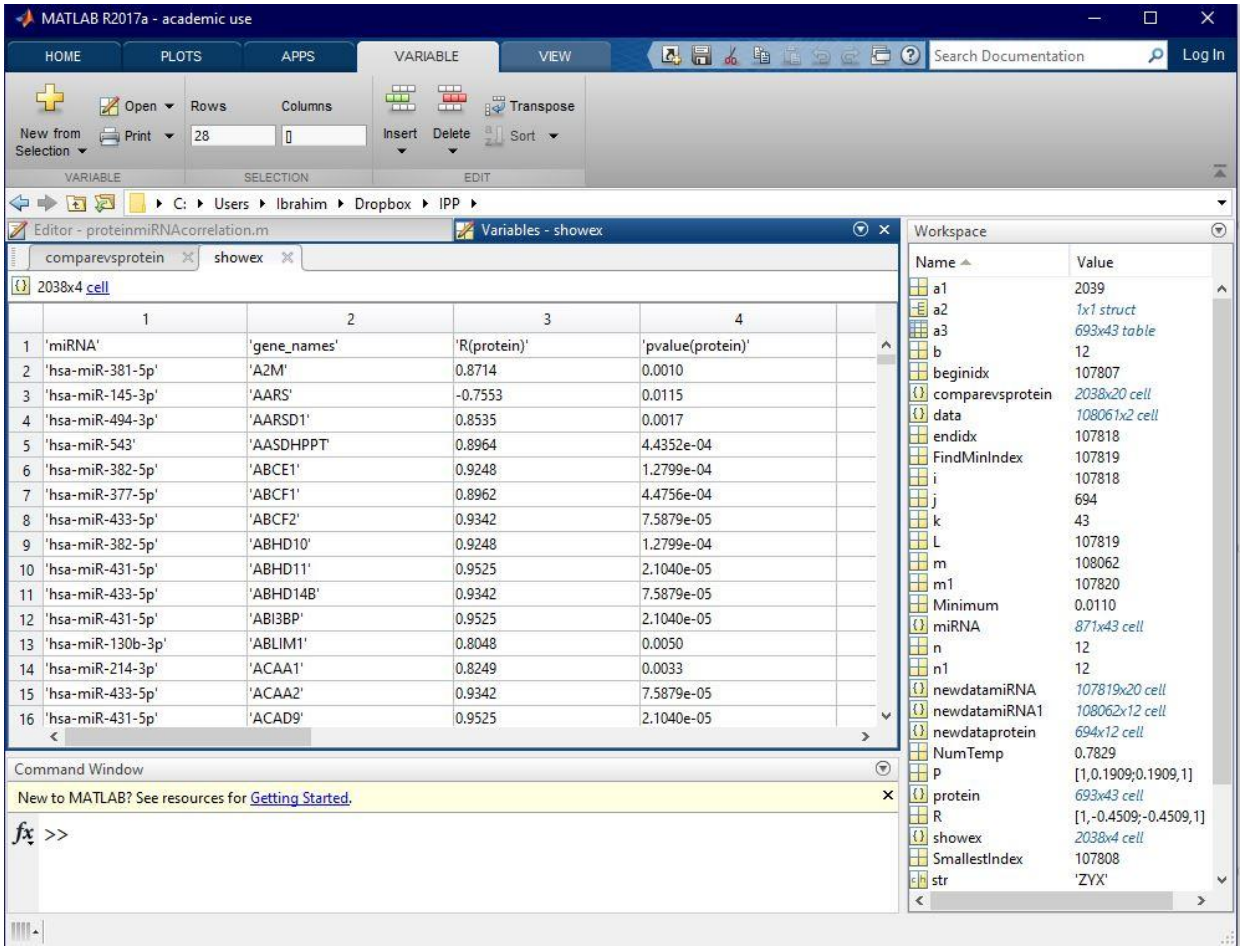
For one gene in mRNA regulated by multiple miRNA problem, we have two options,

(1) selecting the pair that has the largest correlation coefficient value,

(2) selecting the pair that has the smallest p-value,

The first option means the pair has the most obvious correlation coefficient, the second option means we are most confident in its correlation coefficient value. The second one is preferred because we define all the pairs whose correlation coefficient are less than zero as negative correlation, and we want to find the one which is statistically significant.

For one mRNA regulated by multiple miRNA problem, select the pair with the smallest p-value. Then, compare DEG set with genes that are anti-correlated in miRNA-mRNA/protein dataset separately and find the intersection.



	1	2	3	4
1	'miRNA'	'gene_names'	'R(protein)'	'pvalue(protein)'
2	'hsa-miR-381-5p'	'A2M'	0.8714	0.0010
3	'hsa-miR-145-3p'	'AARS'	-0.7553	0.0115
4	'hsa-miR-494-3p'	'AARSD1'	0.8535	0.0017
5	'hsa-miR-543'	'AASDHPPT'	0.8964	4.4352e-04
6	'hsa-miR-382-5p'	'ABCE1'	0.9248	1.2799e-04
7	'hsa-miR-377-5p'	'ABCF1'	0.8962	4.4756e-04
8	'hsa-miR-433-5p'	'ABCF2'	0.9342	7.5879e-05
9	'hsa-miR-382-5p'	'ABHD10'	0.9248	1.2799e-04
10	'hsa-miR-431-5p'	'ABHD11'	0.9525	2.1040e-05
11	'hsa-miR-433-5p'	'ABHD14B'	0.9342	7.5879e-05
12	'hsa-miR-431-5p'	'ABI3BP'	0.9525	2.1040e-05
13	'hsa-miR-130b-3p'	'ABLIM1'	0.8048	0.0050
14	'hsa-miR-214-3p'	'ACAA1'	0.8249	0.0033
15	'hsa-miR-433-5p'	'ACAA2'	0.9342	7.5879e-05
16	'hsa-miR-431-5p'	'ACAD9'	0.9525	2.1040e-05

Name	Value
a1	2039
a2	1x1 struct
a3	693x43 table
b	12
beginidx	107807
comparevsprotein	2038x20 cell
data	108061x2 cell
endidx	107818
FindMinIndex	107819
i	107818
j	694
k	43
L	107819
m	108062
m1	107820
Minimum	0.0110
miRNA	871x43 cell
n	12
n1	12
newdatamiRNA	107819x20 cell
newdatamiRNA1	108062x12 cell
newdataprotein	694x12 cell
NumTemp	0.7829
P	[1,0.1909;0.1909,1]
protein	693x43 cell
R	[1,-0.4509;-0.4509,1]
showex	2038x4 cell
SmallestIndex	107808
str	'ZYX'

Figure 2: Matlab implementation to find negative correlated genes

Chapter 4 – Result and Analysis

In this chapter, I will illustrate and analyze some experimental results. Key findings include all the anti-correlated genes in our dataset as well as pathways found which are related to muscular diseases.

4.1 Significantly Anti-correlated Genes

The analysis of gene expression to identify those differential expressed genes among which the gene pairs are negative correlated. Additionally, the significance testing is based on q-value for false discovery which is equal to or less than 0.006.

	mRNA	Protein
Original Gene Numbers	24386	2171
DEG	604	391

Table 5: differential expressed genes

	miRNA-mRNA	miRNA-Protein
Negative Correlated Gene Numbers	4701	501

Table 6: negative correlated genes

4.2 Pathway Found Related to Muscular Disease

After obtaining the differential expressed genes, the gene list can be used as input for conducting gene set enrichmentanalysis into gene-set libraries. Many gene-set libraries have been developed, after doing some researches, I chose Enrichr for gene set enrichment analysis.

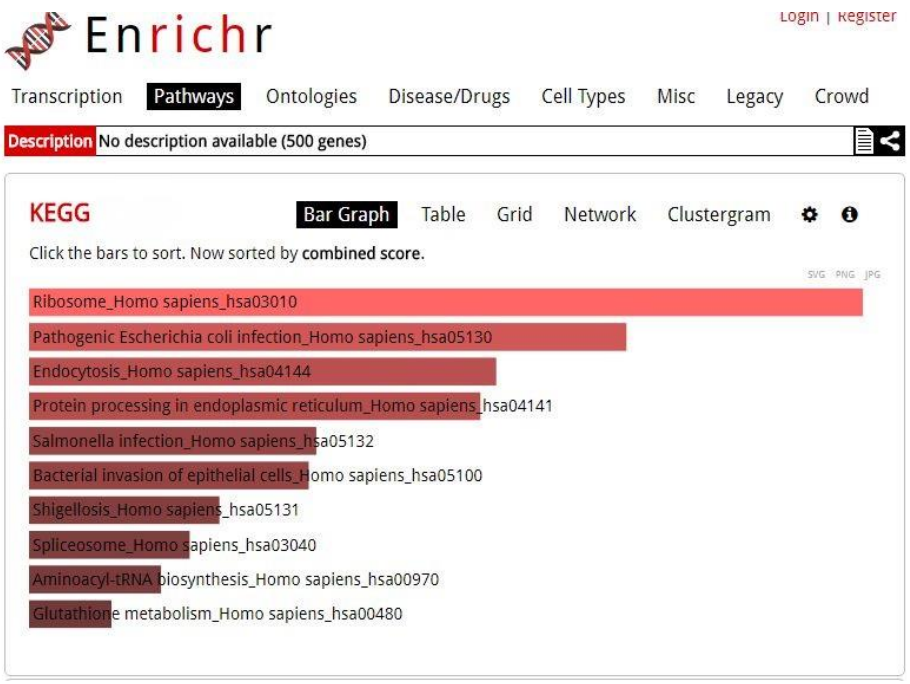


Figure 3: biological pathways search via Enrichr

Pathway Name: Striated Muscle Contraction_Homo sapiens	
Gene Num in Pathway:	38
Gene Num found in Eric's dataset:	18
Negative Correlation Gene Num:	12

Table 7: information for candidate pathway

Pathway Function: Striated muscle contraction is a process whereby force is generated within striated muscle tissue, resulting in a change in muscle geometry, or in short, increased force being exerted on the tendons.

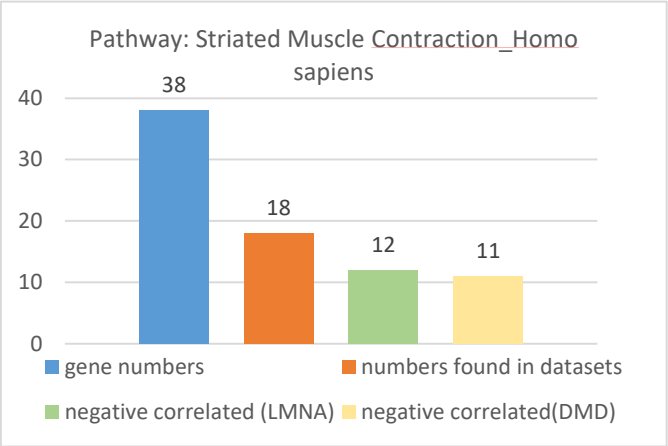


Figure 4: histogram for candidate pathway

The reason why we select this pathway for testing is that the pathway has more genes which can be found in our datasets so that we can testify our idea. For other candidate pathways we found, they contain more than 100 genes in each pathway, however, less than 20 percent of genes can be found expression level on our datasets. So it become much harder to use real-data to validate the experiment.

Looking into the two disease groups: DMD and LMNA. In DMD (compared to normal), from table 8 and table 9, we can see that 18 genes out of 38 genes in the pathway could be found expression in our dataset among which more than half of them showing negative correlation tendency. Blue refers to negative correlated genes while orange represents positive correlated genes. In LMNA(compared to normal), slightly fewer (12) genes show negative correlated tendency. For table 8 and 9, blue represents negative correlated while yellow means positive correlated.

Genes found in our dataset (DMD vs Norm)					
ACTN2	MYBPC1	MYL1	TMOD1	TNNT1	TPM2
DES	MYH3	MYL4	TNNC1	TNNT2	TTN
DMD	MYH8	NEB	TNNI1	TPM1	VIM

Table 8: Correlation of genes in candidate pathway (DMD vs Norm)

Genes found in our datasets (LMNA vs Norm)					
ACTN2	MYBPC1	MYL1	TMOD1	TNNT1	TPM2
DES	MYH3	MYL4	TNNC1	TNNT2	TTN
DMD	MYH8	NEB	TNNI1	TPM1	VIM

Table 9: Correlation of genes in candidate pathway (LMNA vs Norm)

The p-values for the pathway Striated muscle contraction contains two p-values from differential expressed genes and mRNA and protein dataset respectively, and two from correlation in miRNA-mRNA, miRNA-protein.

	P-value
Pathway from miRNA-mRNA Negative Correlated Genes	$6.998e^{-1}$
Pathway from miRNA-protein Negative Correlated Genes	$3.820e^{-5}$
Pathway from mRNA Differential Expressed Genes	$3.525e^{-1}$
Pathway from protein Differential Expressed Genes	$1.333e^{-5}$

Table 10: p-value for candidate pathway

Using Fisher's method to combine P-values, we will obtain the Fisher's Method combined P-value. The P-value is very small which indicates the pathway we have is significant.

Combined P-value	$2.8364e^{-7}$

Table 11: combined p-value using Fisher's method

Chapter 5 - Summary and Future Work

In this chapter, I would like to discuss several aspects to improve our work. In future research, more detailed research is needed to apply and test in the procedure of finding biological pathways, for example, when one mRNA regulated by multiple miRNA problem, is there any alternative choice for justifying when it is negative correlated or positive correlated for each gene in different omics datasets. Further studies should also investigate in the application of the method proposed in chapter 2.

5.1 Refine Correlation Measurement

For one gene targeted by multiple miRNA problem, we can see from the figure,since the up-regulation and down-regulation are not equally distributed, we may need some strategy for adjusting pairing method.

database	down-regulation number	total number	down-regulation percentage	up-regulation percentage
miRNA	375	870	0.431	0.569
mRNA	3853	24387	0.158	0.842
Protein Dataset 1	183	2171 (1327 without missing value)	0.138	0.862
Protein Dataset 2	24	692 (579 without missing value)	0.041	0.959

Table 12: regulation proportion of dataset dataset

-down-regulation : Norm>LMNA

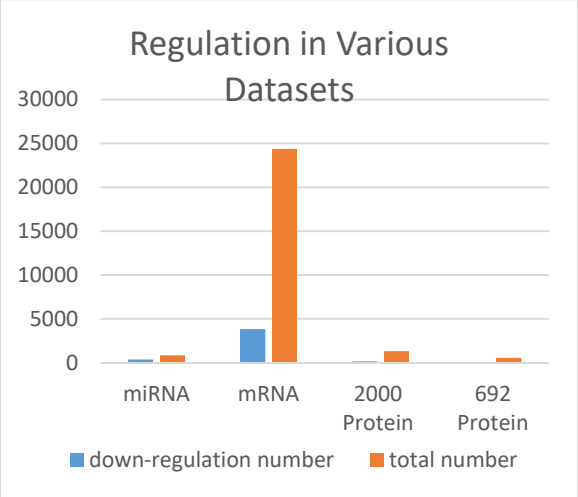


Figure 5: histogram for the proportion of

-up-regulation: LMNA>Norm

5.2 Apply on miR-29b Gene Set

miRNA has been found to regulate on their targeted gene in mRNA or protein datasets by conducting mRNA degradation or inhibiting protein translation. The mir-29 family has been found targeting a large group of functionally related genes, which effect on kidney, muscle and other organs. There are three members in the miR-29 family, miR-a, miR-b, and miR-c. MiR-29b is the one that results in multiple types of muscle atrophy. Our research can further be applied on miR-29b.

In addition, future work can also consider dealing with missing data, missing data has been a troublesome problem in bioinformatics research, a large number of missing data could occur when the signal is not detected, in our protein dataset, nearly half genes have missing data in sample's expression. If we can solve this problem, it can definitely help us to make full use of the dataset. We can try methods like matrix completion to enhance our dataset and apply our methods to design more experiment.

Reference

[1] Ambros, Victor, et al. "A uniform system for microRNA annotation." *Rna* 9.3 (2003): 277-279.

[2] Bersanelli, Matteo, et al. "Methods for the integration of multi-omics data: mathematical aspects." *BMC bioinformatics* 17.2 (2016): S15.

[3] https://cofactorgenomics.com/advantages-rna-seq-over-microarray-technology/

[4] Peter, M. E. "Targeting of mRNAs by multiple miRNAs: the next step." *Oncogene* 29.15 (2010): 2161.

[5] Roberts, Thomas C., et al. "Multi-level omics analysis in a murine model of dystrophin loss and therapeutic restoration." *Human molecular genetics* 24.23 (2015): 6756-6768.

[6] https://en.wikipedia.org/wiki/LMNA-related_congenital_muscular_dystrophy.

[7] Berger B, Peng J, Singh M. Computational solutions for omics data.Nat Rev Genet. 2013;14(5):333–46. doi:10.1038/nrg3433.

[8] Roberts, Thomas C., et al. "Multi-level omics analysis in a murine model of dystrophin loss and therapeutic restoration." *Human molecular genetics* 24.23 (2015): 6756-6768.https://en.wikipedia.org/wiki/Fisher%27s_method

[9] Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease." *Genome biology* 18.1 (2017): 83.

[10] http://math.arizona.edu/~jwatkins/ttest.pdf

[11] Yu, Guoqiang, et al. "Matched gene selection and committee classifier for molecular classification of heterogeneous diseases." *Journal of Machine Learning Research* 11.Aug (2010): 2141-2167.

[12] Quackenbush, John. "Microarray data normalization and transformation." *Nature genetics* 32 (2002): 496.

[13] https://en.wikipedia.org/wiki/Chi-squared_distribution

[14] http://www.netlib.org/math/docpdf/ch15-03.pdf

[15] https://brainder.org/2012/05/11/the-logic-of-the-fisher-method-to-combine-p-values/