

ANALYSIS OF MOTIFS IN MICRORNA-TRANSCRIPTION FACTOR GENE
REGULATORY NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BİLGE SÜRÜN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOINFORMATICS

AUGUST 2014

Approval of the thesis:

**ANALYSIS OF MOTIFS IN MICRORNA-TRANSCRIPTION FACTOR
GENE REGULATORY NETWORKS**

submitted by **BİLGE SÜRÜN** in partial fulfillment of the requirements for the degree
of **Master of Science in Bioinformatics, Middle East Technical University** by,

Prof. Dr. Nazife Baykal

Director, **Graduate School of Informatics**

Assist. Prof. Dr. Yeşim Aydın Son

Head of Department, **Health Informatics, METU**

Assist. Prof. Dr. Aybar Can Acar

Supervisor, **Health Informatics, METU**

Assoc. Prof. Dr. Vilda Purutçuoğlu

Co-supervisor, **Statistics, METU**

Examining Committee Members:

Assoc. Prof. Dr. Tolga Can

Computer Engineering, METU

Assist. Prof. Dr. Aybar Can Acar

Health Informatics, METU

Assoc. Prof. Dr. Vilda Purutçuoğlu

Statistics, METU

Assist. Prof. Dr. Bala Gür Dedeoğlu

Biotechnology, Ankara University

Assist. Prof. Dr. Yeşim Aydın Son

Health Informatics, METU

Date: _____

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

Name, Last Name: BİLGE SÜRÜN

Signature :

ABSTRACT

ANALYSIS OF MOTIFS IN MICRORNA-TRANSCRIPTION FACTOR GENE REGULATORY NETWORKS

Sürün, Bilge

M.S., Department of Bioinformatics

Supervisor : Assist. Prof. Dr. Aybar Can Acar

Co-Supervisor : Assoc. Prof. Dr. Vilda Purutçuoğlu

August 2014, 106 pages

MicroRNAs are small non-coding RNA molecules which contain 21-25 nucleotides, and function in post transcriptional regulation by inhibiting the translation of mRNA targets. miRNAs typically affect gene regulation by forming composite feed forward circuits (cFFCs) which also comprise a transcription factor (TF) and a target gene. By analyzing these cFFCs, the contribution of miRNAs in altering TF networks can be revealed. These contributions could either be the de-escalation of the target gene repertoire or to increase the redundancy through cFFC formation. To conduct the analysis, the connections between genes, miRNAs, and TFs are obtained using two datasets one of which is obtained from human myeloid leukemia cell line. These two datasets are also different from each other in terms of the numbers of TFs and miRNAs that are included in the networks and the significance of the predicted connections. The first dataset which contains connectivity information of a normal cell involves 83 TFs, 564 miRNAs and 5169 genes which construct 124,740 and 34,298 human-mouse conserved TF and miRNA regulatory connections, respectively. The second dataset which contains 137 miRNAs, 274 TFs and 6749 genes which are compiled from the FANTOM 4 database from which the total number of human-mouse conserved regulatory connections is identified as 6631 for miRNAs and 60969 for TFs. Then, in order to reveal the significance on a statistical level, the randomization tests are applied

to the connectivity matrix. Obtaining the significance of miRNA-based cFFCs lead us to conclusions about the effect of miRNAs in fine-tuning gene regulatory networks and the evolutionary role of miRNAs in the cell regulation.

Keywords: microRNA, transcription factor, composite feedforward circuits, regulatory network

ÖZ

MİKRORNA-TRANSKRİPSİYON FAKTÖRÜ GEN REGÜLASYON AĞLARINDA MOTİF ANALİZİ

Sürün, Bilge

Yüksek Lisans, Biyoenformatik Programı

Tez Yöneticisi : Yrd. Doç. Dr. Aybar Can Acar

Ortak Tez Yöneticisi : Doç. Dr. Vilda Purutçuoğlu

Ağustos 2014 , 106 sayfa

MikroRNAlar kendileri protein kodlaması yapmayan, fakat diğer genlerin ifade sonrası değişik oranlarda susturulmalarını sağlayan yaklaşık 22 nükleotid uzunluğunda ufak RNA molekülleridir. miRNAlar tipik olarak transkripsiyon faktörleri (TF) ile karma önbesleme devreleri (kÖBD - cFFC) oluşturarak gen regülasyonunu etkilerler. Bu cFFClerin incelenmesi, miRNAların TF-gen ağlarında regülasyon değişikliğine neden olan katkılarını ortaya çıkartılabilir. Bu katkılar, TF hedef gen repertuarının daraltılması veya cFFCler yoluyla regülasyonda yedeklilik sağlanması olarak özetlenebilir. Analizi gerçekleştirmek için TF, miRNA ve hedef genler arasındaki bağlantılar, bir tanesi miyeloid lösemi hücre dizisine ait olmak üzere iki farklı veri kümesi kullanılarak elde edilmiştir. Veri kümelerinin elde edildiği hücre dizileri tipinin farklılığının yanında, bu iki veri kümesi, içerdikleri TF ve miRNA sayıları ve tahmin edilen bağlantıların istatistiksel anlamlılığı açılarından da farklılık göstermektedir. Bağlantı bilgisi sağlıklı hücreden alınan veri kümesi toplam 124,740 korunmuş insan-fare TF ve 34,298 miRNA düzenleyici bağlantılarını oluşturan 83 TF, 564 miRNA ve 5169 gen içermektedir. İkinci veri kümesi ise FANTOM4 veritabanından elde edilen 173 miRNA, 274 TF ve 6749 genden oluşmakta olup toplam 6631 miRNA ve 60969 TF korunmuş insan-fare düzenleyici bağlantılarını içermektedir. Bu bağlantı matrisine randomizasyon testleri uygulanarak miRNA-tabanlı cFFClerin istatistiki olarak ne kadar anlamlı oldukları

ölçülmüştür. Bu ölçümler ise miRNAların gen regülasyonundaki ince ayar etkileri ve evrimsel rolleri hakkında çıkarım yapabilmeyi sağlamıştır.

Anahtar Kelimeler: mikroRNA, transkripsiyon faktör, karma önbesleme devreleri, regülasyon ağı

To my family

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Assist. Prof. Dr. Aybar Can Acar who has been a tremendous mentor for me. His constant intellectual support and stupendous guidance have allowed this thesis to come alive in the first place.

I would like to thank my co-supervisor Assoc. Prof. Dr. Vilda Purutçuođlu not only for her remarkable support throughout this work, but also for her priceless guidance with a friendly attitude since the day I set foot in METU as an undergraduate student.

I also would like to thank Assist. Prof. Dr. Yeşim Aydın Son, Assoc. Prof. Dr. Tolga Can and Assist. Prof. Dr. Bala Gür Dedeođlu for taking time out from their busy schedule to attend the examining committee and provide insightful comments.

I am grateful to my friends for all the emotional support and caring they provided and for being there always willingly to help. Their supports and useful suggestions helped me to stay focused on my study.

Last but not the least important, I would like to thank my parents, Şirin Sürün and Hilmi Sürün for their unconditional support, encouragement and infinite trust.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZ	vi
ACKNOWLEDGMENTS	ix
TABLE OF CONTENTS	x
LIST OF TABLES	xvi
LIST OF FIGURES	xviii
LIST OF ABBREVIATIONS	xxiii
CHAPTERS	
1 INTRODUCTION	1
1.1 Motivation	1
1.2 Scope and Goal	2
1.3 Contribution	2
1.4 Outline	3
2 BACKGROUND AND RELATED WORKS	5
2.1 Biological Information Transfer and Gene Regulation	5
2.1.1 The Central Dogma	5

2.1.2	Gene Regulation	6
2.1.2.1	Transcriptional Regulation	6
2.1.2.1.1	Transcription Factors	7
2.1.2.2	Post Transcriptional Regulation	8
2.1.2.2.1	microRNA	8
2.2	Biological Network Motifs	9
2.2.1	Autoregulation	9
2.2.2	Feed-Forward Loop Network Motif	10
2.2.2.1	Coherent Type 1 FFL	11
2.2.2.2	The Incoherent Type 1 FFL	12
2.2.3	Single Input Module	12
2.2.4	Bi-fans and Dense Overlapping Regulons	12
2.3	Analysis of Regulatory Networks	13
2.3.1	Analysis of Transcription Regulation Networks	13
2.3.2	Analysis of Combined Transcriptional - Post Transcriptional Regulatory Networks	14
3	RANDOMIZATION ALGORITHMS AND ANALYSIS METHODS	19
3.1	Data	19
3.1.1	Dataset Retrieved from the Study of Iwama et al.	19
3.1.1.1	Matrix Representation of the Connection Data	19
3.1.2	Dataset Retrieved from FANTOM4 EdgeExpress Database	20

	3.1.2.1	FANTOM Network Information According to Different Weight Thresholds	21
3.2		Key Terms Used in the Network Analysis	21
	3.2.1	Composite Feed Forward Circuits	21
	3.2.2	cFFC Targeted Genes	22
	3.2.3	cFFC Redundancy	24
	3.2.3.1	miRNA Derived cFFC Redundancy	24
	3.2.3.2	TF Derived cFFC Redundancy	24
3.3		Randomization Procedure	25
	3.3.1	Main Randomization Procedure	25
	3.3.1.1	Pseudocode of the Main Randomization Procedure	25
	3.3.2	Partial Randomization Procedure	30
	3.3.2.1	Pseudocode of the Partial Randomization Procedure	31
	3.3.2.1.1	miRNA - Gene Connection Only Randomization	31
	3.3.2.1.2	miRNA - TF Connection Only Randomization	32
3.4		Statistical Analysis	36
	3.4.1	Normality Tests	36
	3.4.1.1	Quantile-Quantile Plots (Q-Q Plots)	36
	3.4.1.2	Shapiro-Wilk Test	37
	3.4.2	One Sample Z - Test	38

	3.4.3	Two-Sample t-Test	38
4		EXPERIMENTAL RESULTS AND THE ANALYSIS OF SIMULATED NETWORKS	41
	4.1	Analysis of the Main Randomization Results Based on Iwama et. al. Data	41
	4.1.1	Q-Q Plots and Normality Tests	41
	4.1.2	Comparison of the Simulated Data via Real Gene Regulatory Network	43
	4.1.2.1	Random Expectations	43
	4.1.2.2	Comparison of Random Expectations with Real GRN via Z-test	44
	4.1.2.3	Random Expectations of TF Derived and miRNA Derived Redundancy	45
	4.1.2.4	Comparison of Redundancies with respect to Real GRN via Z-test	45
	4.1.2.5	Comparison of Redundancies via the t-test	46
	4.2	Analysis of Partial Randomization Results Based on Iwama et. al. Data	47
	4.2.1	Q-Q Plots and Normality Tests	47
	4.2.2	Comparison of the Simulated Data via Real Gene Regulatory Network	50
	4.2.2.1	Random Expectations	50
	4.2.2.2	Comparison of Random Expectations with Real GRNs via Z-test	50
	4.2.2.3	Comparison of Redundancies via t-test	51
	4.3	Analysis of the Main Randomization Results Based on FANTOM Data	52

4.3.1	Q-Q Plots and Normality Test	52
4.3.2	Comparison of the Simulated Data via Real Gene Reg- ulatory Network	54
4.3.2.1	Random Expectations	54
4.3.2.2	Comparison of Random Expectations with Real GRN via Z-test	55
4.3.2.3	Comparison of Redundancies via the t-test	56
4.4	Analysis of Partial Randomization Results Based on FANTOM Data	56
4.4.1	QQ Plots and Normality Tests	56
4.4.2	Comparison of the Simulated Data with the Real GRN	59
4.4.2.1	Random Expectations	59
4.4.2.2	Comparison of Random expectations with Real GRN via Z-test	60
4.4.2.3	Comparison of Redundancies via t-test . .	61
5	CONCLUSION	63
5.1	Summary	63
5.2	Discussion	64
5.3	Future Work	65
	REFERENCES	67
APPENDICES		
A	QQ PLOTS	71
A.1	QQ Plots of Main Randomization Procedure of Iwama et. al. Data	71

A.2	QQ plots of Partial Randomization Procedure of Iwama et. al. Data	78
A.3	QQ Plots of Main Randomization Procedure of FANTOM Data	82
A.4	QQ Plots of Partial Randomization Procedure of FANTOM Data	89
B	SOURCE CODES	95
B.1	Parser to Retrieve miRNA Connections	95
B.2	Parser to Retrieve TF Connections	95
B.3	Orthology Parser	96
B.4	Source Codes of Main Randomization Procedure	97
B.5	Source Codes of Partial Randomization Procedure	99
	B.5.1 miRNA-Gene Edge Randomization	99
	B.5.2 miRNA-TF Edge Randomization	102

LIST OF TABLES

TABLES

Table 3.1 Network information for different weights	22
Table 4.1 Shapiro-Wilk Test Results - Iwama et. al. Data	43
Table 4.2 Randomization Results of Iwama et. al. Data	43
Table 4.3 Comparison of Z and P Values - Iwama et. al. Data	44
Table 4.4 Randomization Results of miRNA and TF Derived Redundancy Ob- tained from Iwama et. al. Data	45
Table 4.5 Comparison of Real GRN with Random Networks in terms of miRNA Derived and TF Derived Redundancy - Iwama et. al. Data	45
Table 4.6 Results of Variance Ratio F-test	46
Table 4.7 Results of Welch's t-test	46
Table 4.8 Shapiro-Wilk Test Results of Partial Randomization - Iwama et. al. Data	50
Table 4.9 Results of Partial Randomization of Iwama et. al. miRNA Connection Data	50
Table 4.10 Comparison of Z and P Values of Partial Randomization based on Iwama et. al. miRNA Connection Data	51
Table 4.11 Results of Variance Ratio F-test - Partial Randomization of Iwama et. al. Data	52

Table 4.12 Results of two sample t-test - Partial Randomization of Iwama et. al. Data	52
Table 4.13 Shapiro-Wilk Test Results - FANTOM Data Weight 1.5	54
Table 4.14 Randomization Results of FANTOM Data	54
Table 4.15 Comparison of Z and P Values - FANTOM Data	55
Table 4.16 Results of Variance Ratio F-test	56
Table 4.17 Results of Welch's t-test	56
Table 4.18 Shapiro-Wilk Test Results of Partial Randomization - FANTOM Data	59
Table 4.19 Shapiro-Wilk Test Results after Outlier Removal - FANTOM Data .	59
Table 4.20 Results of Partial Randomization of FANTOM miRNA Connection Data	60
Table 4.21 Comparison of Z and P Values of Partial Randomization based on FANTOM miRNA Connection Data	60
Table 4.22 Difference between Iwama et. al. and FANTOM Resultings	61
Table 4.23 Results of Variance Ratio F-test - Partial Randomization of FANTOM Data	61
Table 4.24 Results of two sample t-test - Partial Randomization of FANTOM Data	61

LIST OF FIGURES

FIGURES

Figure 2.1	Central Dogma	6
Figure 2.2	Simple regulation and autoregulation.	10
Figure 2.3	Types of FFL network motif.	11
Figure 2.4	SIM network motif in which the regulator X controls three genes, $Z_1, Z_2,$ and Z_3	12
Figure 2.5	Bi-fan motif and dense overlapping regulon.	13
Figure 2.6	Statistically significant motifs found in transcriptional regulation net- work (re-printed from [28]).	14
Figure 3.1	Form of connection matrices [16]	20
Figure 3.3	A unitary cFFC.	22
Figure 3.2	Change in the initial numbers according to different weight thresholds	23
Figure 3.4	Sub-network views for multiple cFFCs [16].	25
Figure 4.1	QQ plot of normality of number of cFFCs after miRNA connection randomization	42
Figure 4.2	QQ plot of normality for only miRNA-gene connection randomization for number of cFFC targeted genes	48

Figure 4.3 QQ plot of normality for only miRNA-TF connection randomization for number of cFFC targeted genes	49
Figure 4.4 QQ plot of normality of number of cFFCs after TF connection ran- domization - FANTOM Data	53
Figure 4.5 QQ plot of normality of the number of cFFC targeted genes after only miRNA-TF connection randomization - FANTOM Data	58
Figure A.1 QQ plot of normality of number of cFFC targeted genes after miRNA connection randomization - Iwama et. al. Data	71
Figure A.2 QQ plot of normality of cFFC redundancies after miRNA connection randomization - Iwama et. al. Data	71
Figure A.3 QQ plot of normality of miRNA derived redundancy after miRNA connection randomization - Iwama et. al. Data	72
Figure A.4 QQ plot of normality of TF derived redundancy after miRNA con- nection randomization - Iwama et. al. Data	72
Figure A.5 QQ plot of normality of number of cFFCs after TF connection ran- domization - Iwama et. al. Data	73
Figure A.6 QQ plot of normality of number of cFFC targeted genes after TF connection randomization - Iwama et. al. Data	73
Figure A.7 QQ plot of normality of cFFC redundancies after TF connection randomization - Iwama et. al. Data	74
Figure A.8 QQ plot of normality of miRNA derived redundancy after TF con- nection randomization - Iwama et. al. Data	74
Figure A.9 QQ plot of normality of TF derived redundancy after TF connection randomization - Iwama et. al. Data	75
Figure A.10 QQ plot of normality of number of cFFCs after TF-miRNA both connection randomization - Iwama et. al. Data	75

Figure A.11	QQ plot of normality of number of cFFC targeted genes after TF-miRNA both connection randomization - Iwama et. al. Data	76
Figure A.12	QQ plot of normality of cFFC redundancy after TF-miRNA both connection randomization - Iwama et. al. Data	76
Figure A.13	QQ plot of normality of miRNA derived cFFC redundancy after TF-miRNA both connection randomization - Iwama et. al. Data	77
Figure A.14	QQ plot of normality of TF derived cFFC redundancy after TF-miRNA both connection randomization - Iwama et. al. Data	77
Figure A.15	QQ plot of normality of number of cFFCs after only miRNA-gene connection randomization - Iwama et. al. Data	78
Figure A.16	QQ plot of normality of cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data	78
Figure A.17	QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data	79
Figure A.18	QQ plot of normality of TF derived redundancy after only miRNA-gene connection randomization - Iwama et. al. Data	79
Figure A.19	QQ plot of normality of number of cFFCs after only miRNA-TF connection randomization - Iwama et. al. Data	80
Figure A.20	QQ plot of normality of cFFC redundancy after only miRNA-TF connection randomization - Iwama et. al. Data	80
Figure A.21	QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data	81
Figure A.22	QQ plot of normality of TF derived cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data	81
Figure A.23	QQ plot of normality of number of cFFC targeted genes after TF connection randomization - FANTOM Data	82

Figure A.24QQ plot of normality of cFFC redundancy after TF connection randomization - FANTOM Data	82
Figure A.25QQ plot of normality of miRNA derived cFFC redundancy after TF connection randomization - FANTOM Data	83
Figure A.26QQ plot of normality of TF derived cFFC redundancy after TF connection randomization - FANTOM Data	83
Figure A.27QQ plot of normality of number of cFFCs after both TF-miRNA connection randomization - FANTOM Data	84
Figure A.28QQ plot of normality of number of cFFC targeted genes after both TF-miRNA connection randomization - FANTOM Data	84
Figure A.29QQ plot of normality of cFFC redundancy after both TF-miRNA connection randomization - FANTOM Data	85
Figure A.30QQ plot of normality of miRNA derived cFFC redundancy after both TF-miRNA connection randomization - FANTOM Data	85
Figure A.31QQ plot of normality of TF derived cFFC redundancy after both TF-miRNA connection randomization - FANTOM Data	86
Figure A.32QQ plot of normality of number of cFFCs after miRNA connection randomization - FANTOM Data	86
Figure A.33QQ plot of normality of number of cFFC targeted genes after both miRNA connection randomization - FANTOM Data	87
Figure A.34QQ plot of normality of cFFC redundancy after both miRNA connection randomization - FANTOM Data	87
Figure A.35QQ plot of normality of miRNA derived cFFC redundancy after both miRNA connection randomization - FANTOM Data	88
Figure A.36QQ plot of normality of TF derived cFFC redundancy after both miRNA connection randomization - FANTOM Data	88

Figure A.37	QQ plot of normality of number of cFFC targeted genes after only miRNA-TF connection randomization - FANTOM Data	89
Figure A.38	QQ plot of normality of cFFC redundancy after only miRNA-TF connection randomization - FANTOM Data	89
Figure A.39	QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-TF connection randomization - FANTOM Data	90
Figure A.40	QQ plot of normality of TF derived cFFC redundancy after only miRNA-TF connection randomization - FANTOM Data	90
Figure A.41	QQ plot of normality of number of cFFCs after only miRNA-gene connection randomization - FANTOM Data	91
Figure A.42	QQ plot of normality of number of cFFC targeted genes after only miRNA-gene connection randomization - FANTOM Data	91
Figure A.43	QQ plot of normality of cFFC redundancy after only miRNA-gene connection randomization - FANTOM Data	92
Figure A.44	QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-gene connection randomization - FANTOM Data	92
Figure A.45	QQ plot of normality of TF derived cFFC redundancy after only miRNA-gene connection randomization - FANTOM Data	93

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger RNA
TF	Transcription factor
TFBS	Transcription factor binding site
miRNA	microRNA
miRISC	miRNA-induced silencing complex
GRN	Gene regulatory network
cFFC	Composite feedforward circuits
NCBI	National Center for Biotechnology Information
SIM	Single input module
NAR	Negative autoregulation
PAR	Positive autoregulation
FFL	Feedforward loop
CDF	Cumulative distribution function
QQ Plot	Quantile Quantile Plot

CHAPTER 1

INTRODUCTION

1.1 Motivation

With the completion of the Human Genome Project, revealing the genetic causes under the phenotypic characteristics of organisms has become the main goal of the genomic revolution. In order to achieve this goal, analysing complex biological networks, finding the interactions between the network components and understanding their biological functions are essential. By focusing on simple patterns, called network motifs, that are highly recurrent in real networks when compared with random networks, the characteristics of complex systems and their structure can be better understood. The motifs are generally shared among different organisms, implying that these mechanisms are favored by evolution. The significant biological network motifs, such as auto regulation, feed-forward loops, single input modules and bi-fans help to transfer process information by slowing or accelerating the response time, creating a pulse-like dynamics, generating temporal activation etc.

In gene regulatory networks, the information process is carried out to adjust the expression levels of genes by two distinct mechanisms; transcriptional regulation and post-transcriptional regulation. The two main players which have an important role in these mechanisms are Transcription Factors (TFs) and microRNAs (miRNAs). TFs are a group of proteins which control the system of expression by binding to cis-regulator DNA sequences and control the extent to which a specific gene will be transcribed. miRNAs are small RNA molecules that consist of $\simeq 22$ nucleotides which bind to the 3'-untranslated regions (3'-UTRs) of target transcripts in order to change their expression level by the repression or the degradation. Since they both are included in the gene regulation mechanisms, it was proposed that there is a strong possibility of an interplay between TFs and miRNAs by formation of FFLs which will prevent the production of mRNA in high concentrations by repressing its regulator TF and thus it provides a "quick-OFF-slow-ON" mechanism [48]. According to the research conducted in this field, this particular type of FFL in which the TF activates its target gene, and miRNA represses both the TF and its target is found significant and most abundant in gene regulatory networks. These can be called as "the composite feed forward circuits" (cFFCs). Although the significance of the FFLs has been discovered, the difference between the TFs and miRNAs in contribution to the GRNs, the redundancy adding role of miRNAs and the stability of TF networks to the alterations caused by miRNAs had remained unclear till the study conducted by Iwama et. al. (2010). However, the increase in the regulator-target information poses new research

questions, such as whether the revealed characteristics of miRNAs show variations in larger networks obtained from cancer lines with more significant regulator-target interactions, and how these structures and motifs are affected by disorders such as cancer.

1.2 Scope and Goal

The main objective of this study is to analyse gene regulatory networks (GRNs) based on cFFCs and cFFC dependent terms with the concern of finding the differences of TFs and miRNAs in contribution to GRNs and to see whether the evolutionary distinctions of TFs and miRNAs affect the GRNs by changing its conformation. It is also aimed to provide an insight of the behaviour of miRNAs in the GRNs that are obtained from myeloid leukemia cell line. As a first step of the analysis, the study conducted by Iwama et. al. (2010) with the same research findings needed to be verified by using the same TF-target and miRNA-target connections which consist of 83 TFs, 564 miRNAs and 5169 genes including the 83 TFs. Secondly, in order to prove that the results are not attributable to the employed network and the findings are the true characteristics of miRNAs, and to reveal the behaviour of miRNAs in the network obtained from cancer line, the imbalance between the total number of TFs and miRNAs that are included in the network should be eliminated, and the analysis should be re-conducted with regulator-target predictions that have high occurrence rate in biological processes.

In accordance with these purposes the research questions are tested using two different GRNs. One of them is retrieved from the previous study of Iwama et. al. (2010), whereas the other one is obtained from the FANTOM EdgeExpress database that does not constitute a class imbalance problem and provides a weight parameter indicating the strength of the connections for every regulator-target prediction, thus it enables obtaining more significant interactions by setting a threshold value for the weight parameter [19]. More importantly, the FANTOM data originate from the myeloid leukemia cell line; and will thus, allow the investigation of differences between healthy cells and cancer.

1.3 Contribution

The main contribution of the proposed study is to provide a framework that is applicable to different GRNs and enable the researcher to compare the behaviours of network motifs obtained from different cell lines. Using this framework, the results of a previous study, conducted by Iwama et. al. (2010), were reproduced and the code developed was verified. Subsequently, the framework is implemented to another dataset obtained from the FANTOM4 database, and it was revealed that adding redundancy role of miRNAs by forming cFFCs is not attributable to the network, thus it is a distinctive property of miRNAs. It is also uncovered that the myeloid leukemia cancer mechanism changes the behaviour of miRNAs within the network in a such way that they started to regulate their target genes directly instead of showing a tendency of being a part of cFFCs. Another minor contribution is the correction of a logic error in the basis study (Iwama et al. 2010) where self-loops were mistakenly counted as cFFCs.

1.4 Outline

This document which provides an understanding in the differences between miRNAs and TFs in terms of their contribution to the GRNs and their evolutionary behaviour, is composed of 7 chapters, including Introduction and Conclusion, that are outlined as given below.

The first chapter gives an insight to the related research field with the explanation of the motivation, scope and the goal of this study. In the second chapter, the transfer of the sequential information biological processes is summarized and the role of miRNA and TF as main players in gene regulation are explained. Additionally, significant biological network motifs are explained focusing mainly on FFLs, since this study is centered around FFLs. The related studies, based on finding significant motifs using random networks that are conducted in the similar manner of this research are also presented in the second chapter. Moreover, this chapter includes a detailed summary of the study published by Iwama et. al. (2010), which provides a basis for this thesis. Besides including a brief description of the data sources and the explanation of the properties of datasets, the third chapter also covers the simulation procedures with the corresponding pseudocode and statistical analysis methods that were used on the randomization results. In the fourth chapter, the results obtained from randomization experiments are presented and discussed for both datasets. Lastly, in the fifth chapter the conclusions are expressed and the future studies that could be an extension for this thesis are sketched out.

CHAPTER 2

BACKGROUND AND RELATED WORKS

2.1 Biological Information Transfer and Gene Regulation

2.1.1 The Central Dogma

The term, *Central Dogma*, was first expressed by Nobel Laureate Francis Crick in 1958 with the words "Once information has got into a protein it can't get out again" [10]. The arrows given in the Figure 2.1a show the flow of the sequential information and point out that the reverse transfer of the information from protein to DNA or RNA is not possible.

The unidirectional Central Dogma described by Francis Crick suggests that the transfer can be divided into three groups according to the evidence of existence. First group which has strong evidence of occurrence in all cells, consists of the information flow from

- DNA \rightarrow DNA,
- DNA \rightarrow RNA,
- RNA \rightarrow Protein,

and shown with the solid line in the Figure 2.1a [9].

The second group contains the transfers, which may occur in special circumstances, from

- RNA \rightarrow RNA,
- RNA \rightarrow DNA,
- DNA \rightarrow Protein,

and shown with the dash line in the Figure 2.1a.

Third group, which central dogma claims to be not possible, consists of unknown transfers from

- Protein \rightarrow Protein,
- Protein \rightarrow DNA,
- Protein \rightarrow RNA.

As a summary, the information flow seen in all cells follows a path such that self-replicating deoxyribonucleic acid (DNA) is used as a template the synthesis of ribonucleic acid (RNA) and protein, RNA is synthesized through transcription, and proteins are synthesized by the translation of RNA. Although most of the parts of the Central Dogma defined by Francis Crick are valid today, it is manifested that the procedure of information flow is more perplexing and the concept has evolved in time. Studies have shown that the transfers mentioned in the third group as unknown are now known facts such as post-transcriptional regulation, methylation, and inteins [3, 11, 20].

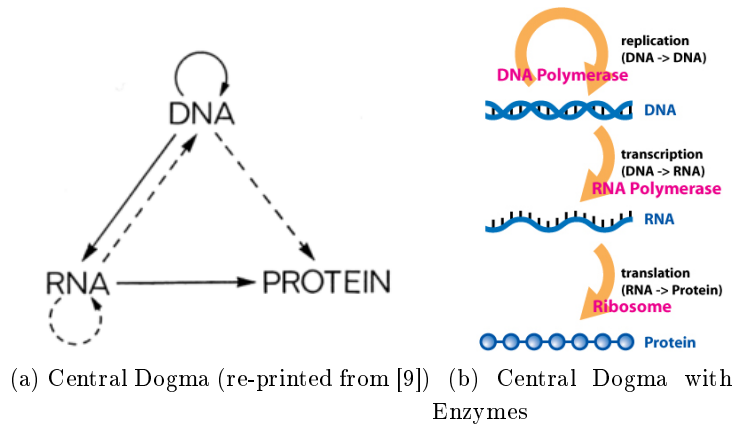


Figure 2.1: Central Dogma

2.1.2 Gene Regulation

2.1.2.1 Transcriptional Regulation

Even though the expression of eukaryotic protein coding genes are regulated at various steps such as transcription initiation, mRNA processing and translation, the main step that most of the regulation occurs is the transcription initiation. *Cis*-regulatory DNA elements included in genes transcribed by RNA polymerase II involve two main parts which are promoter consisting of a core promoter and proximal regulatory elements, and distal regulatory elements composed of enhancers, silencers, insulators and locus control regions [27].

Trans-regulatory transcription factors (TFs) are proteins which regulate gene expression by controlling the transfer of the genetic information from DNA to mRNA. Through recognition sites within the *cis*-regulatory DNA elements, transcription factors bind to transcription factor binding sites (TFBS), by which they either promote or block the recruitment of RNA polymerase [16, 33] and mediate the initiation of

transcription by forming transcription initiation complex along with RNA polymerase II [32].

With the completion of Human Genome Project, it is revealed that there exist $\sim 20,000$ - $25,000$ genes with unique expression pattern, in human genome [27], whereas the total number of TFs is ~ 2000 , far less than the total number of genes [6]. The difference is explained with the fact that promoters contain multiple regulatory elements which enable combinatorial control of gene regulation, leading to high numbers of distinct expression patterns [27].

2.1.2.1.1 Transcription Factors Transcription factors can be divided into three categories such as general transcription factors, activators and coactivators. General transcription factors which are of vital importance for the transcription of all protein coding genes, can identify the TATA box which consists of ~ 25 nucleotides of TATA sequence located on the core promoter, upstream of the transcriptional start site. After recognition of the TATA box, one TF binds to the gene that is being regulated from the DNA upstream, before RNA polymerase II can bind. In order to form the transcription initiation complex which unbinds the DNA double helix for initiation of the transcription, additional TFs and RNA polymerase II are added to the DNA. As a result of this protein-protein interaction between TFs and RNA polymerase II, the initiation complex starts to read the DNA template strand and the complementary strand of RNA is produced. However, this interaction generally causes the initiation at a low level and the RNA transcripts are produced with a limited number. In eukaryotes, the mechanism that is needed for high level transcription and which controls the expression time, amount and a place of specific genes relies on the interaction of control elements which are referred as specific transcription factors that contain activators and coactivators.

Specific transcription factors which function as activators bind to enhancer regions of DNA which are located distant to the promoter and upstream, downstream or within the introns of the genes they regulates. However, in many instances, the relevant activator can also bind to a recognition site which is upstream of the core promoter in different genes. For the case that the activator binds to enhancer elements, by courtesy of DNA bending proteins, the bound activator comes close to the promoter region where it interacts with the mediator proteins, general transcription factors and RNA polymerase in order to form an active transcription initiation complex. The protein-protein interaction between the specific transcription factors and the mediator complex assists the orientation of the complete complex on the promoter and activates the initiation of RNA synthesis [32].

The activators that bind to the upstream of core promoter which is also known as transcription factor binding sites (TFBS) alter the chromatin structure to prevent its blocking effect to transcription [24]. They also provide an increase in the transcription initiation complex formation [29].

Cofactors are another group of specific transcription factors which function to alter the activators activity by forming a protein-protein interaction with the activators instead of binding to a specific sequence within DNA. Their roles in regulation are similar to activators in terms of increasing the rate of transcription initiation complex formation

and modifying the chromatin structure.

Transcription factors also contain the "repressors" group of transcription factors which are DNA binding proteins whose recognition sites are silencers within the distal control elements. Repressors inhibit the gene expression by blocking activator binding [14], or by binding regulatory DNA elements directly as a result of competition with the activator for the same binding site [32]. Repressors can also turn off the transcription by forming repressive chromatin structure which result in blocking the general transcription factors and/or activators access to the promoter region [42]. In some instances, repressors may interfere the transcription negatively by blocking RNA polymerase II, TFIIB and TFIIE assembly, even though the activators are already bound to DNA [7].

2.1.2.2 Post Transcriptional Regulation

Even though the transcription is considered to be the main step in the gene regulation, it is observed that the rate of the gene expression deviates from the levels of the corresponding mRNA implying the role of post transcriptional regulation in gene expression. Post transcriptional regulation is performed at any step between the transcription and the translation, including procedures such as alternative splicing and RNA editing by which multiple proteins are produced from a single gene. It also includes the control of RNA stability and of translation which provides a sudden change in protein levels when it is needed. Moreover, it has been also verified that microRNAs play an important role in post transcriptional regulation by distinct mechanisms such as binding the complementary mRNAs which results in their degradation or inhibition, regulating TFs, or blocking the gene by an inoperative chromatin structure [22].

2.1.2.2.1 microRNA In a study conducted in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum, it was discovered that the *lin-4* gene, which is important in terms of controlling the initiation of larval development of *C. elegans*, generates two small RNA in length of 22 nt and 61 nt, instead of coding for a protein [23]. The RNA consists of 61 nt was identified as a precursor of the 22 nt RNA as a result of its formation into a stem loop. It is also identified that *lin-4* RNAs are complementary to a sequence in the 3' UTR of *lin-14* gene, and the complementarity plays role in the repression of *lin-14* gene with negligible change in the levels of *lin-14* mRNA [45]. These findings led to other studies on genes that encodes ~ 22 nt RNAs, but there was no evidence of such regulatory small RNAs within or beyond the nematodes till the discovery that another gene which plays role in the *C. elegans* heterochronic pathway named *let-7* produces a ~ 22 nt RNA which acts as regulatory RNA in the same way of *lin-4* RNA [5]. Moreover, the homologue of *let-7* RNA was detected in wide range of species including *Homo Sapiens*, *Mus Musculus*, and the two RNAs (*lin-4*, *lin-7*) classified as small temporal RNA (stRNA) for the reason of their similar functions on the timing of developmental stages [30]. Eight years after, the discovery of *lin-4* RNA, it was reported that there exists more than one hundred genes that produce small regulatory RNAs which have similar properties with *lin-4* and *let-7* in terms of their conservation in evolution, their length, and their production from a stem loop precursor, identified from flies, worms and humans. In spite of these similarities, the term stRNA changed as microRNA (miRNA) because of the fact that the expression

of newly found ~ 22 nt RNAs are more likely to occur in specific cell types instead of being produced in the developmental stages only [5].

Because of the difficulties of finding new miRNAs experimentally, computational approaches based on miRNA gene identification with regards to homology searches, proximity search of known miRNA genes for the other stem loops in order to reveal the additional genes of a cluster, and identification of genomic segments with the potential of composing stem loops and aligning these with the known miRNA genes pairwise, are developed [5]. With the increased number of newly identified miRNA genes, a catalogue has been constructed for registry and systematic labelling [13].

It is revealed that RNA polymerase II plays a main role in the transcription of most mammalian miRNAs by producing primary miRNA (pri-miRNA) transcripts which are 5'-capped, polyadenylated, spliced and folded into a secondary hairpin structure with a stem and a terminal loop [41].

The reactions are catalysed by the two members of the RNase III family enzymes, Drosha and Dicer. In the nucleus, Drosha uses folded pri-miRNA as a substrate and this reaction results in ~ 70 nt pre-miRNA which is transferred to the cytoplasm in order to be processed by Drosha into a ~ 20 bp miRNA/miRNA* duplex. After the separation of two strands of miRNA duplex, one of them which is called the mature miRNA or the guide strand is included in a miRNA-induced silencing complex (miRISC) while the other strand is degraded. The complex recognises and targets the complementary mRNAs which results in the repression or degradation of the mRNAs [47].

2.2 Biological Network Motifs

To obtain an insight in the characteristics of biological processes, their representation which includes protein-protein, protein-DNA, and protein-metabolite interactions are termed as biological networks and "network motifs" are defined as recurrent patterns that occur significantly more frequently than those in fittingly randomized networks. The main biological network motifs can be examined into four categories which are auto-regulation, feed-forward loop, single input module, and dense overlapping regulons.

2.2.1 Autoregulation

Autoregulation is a type of motif where the gene is targeted by its own product and regulated either negatively or positively.

Negative autoregulation (NAR) is seen when the gene is inhibited by its own product [35]. It provides an initial increase of the concentration of the targeted gene and then decreases the rate of production of it after a certain threshold. In other words, the increase in the concentration of the TF lowers the production rate of the targeted gene. As a result, the gene reaches its steady state, which is slightly lower than the related threshold level, and the time between the initial and the steady state

is shorter than that of a simple regulation. Moreover, the differences of the protein levels in different cells are balanced by negative autocorrelation because of the negative correlation between the concentration of the gene and the concentration of its TF [2].

Positive autoregulation (PAR) is a type of motif in which the expression of gene is promoted by its own product [2]. Hence, there exists a positive correlation between the concentration of the gene and the concentration of its TF until reaching the steady state. In contrast to the NAR motifs, the response time, which is defined as the time that is required to get halfway to the steady state, is longer than a simple regulation. In addition, PAR motifs cause an increase in variations of the protein levels between cells due to the enhancement of the gene by its own product [17] [25].

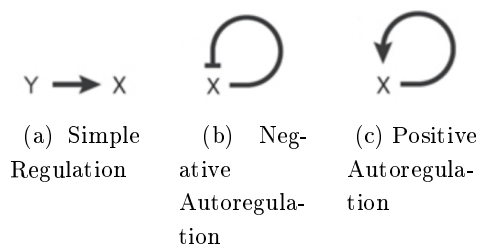


Figure 2.2: Simple regulation and autoregulation. Figure 2.2a shows the simple regulation in which the gene X is regulated by only Y. Figure 2.2b the NAR and Figure 2.2c shows the PAR.

2.2.2 Feed-Forward Loop Network Motif

The feed-forward loops (FFLs) are strong network motifs because their occurrence is much more than expected, when compared to random networks. They are 3-node subgraphs commonly seen in transcription regulation networks, involving 8 possible edge combinations where the edges indicate either activation or repression interactions. Figure 2.3 shows the all possible 8 edge combinations and the FFL types which are differentiated according to the sign of the regulation paths. FFLs consist of two parallel paths which are a direct path and an indirect path. If the overall sign is same with the sign of direct path, the FFL is considered as coherent type, and if their signs are opposite, it is termed as incoherent FFL. Among all eight FFL types, type 1 coherent and type 1 incoherent FFLs have a higher occurrence rate in the transcription networks which are shown in the Figure 2.3 with respect to the other types of FFLs [2].

Most common input functions for FFLs are an "AND gate" and an "OR gate" meaning that either both regulators are needed to control the expression of the shared target, or the binding of one regulator to the gene is enough for the process to work, respectively [26].

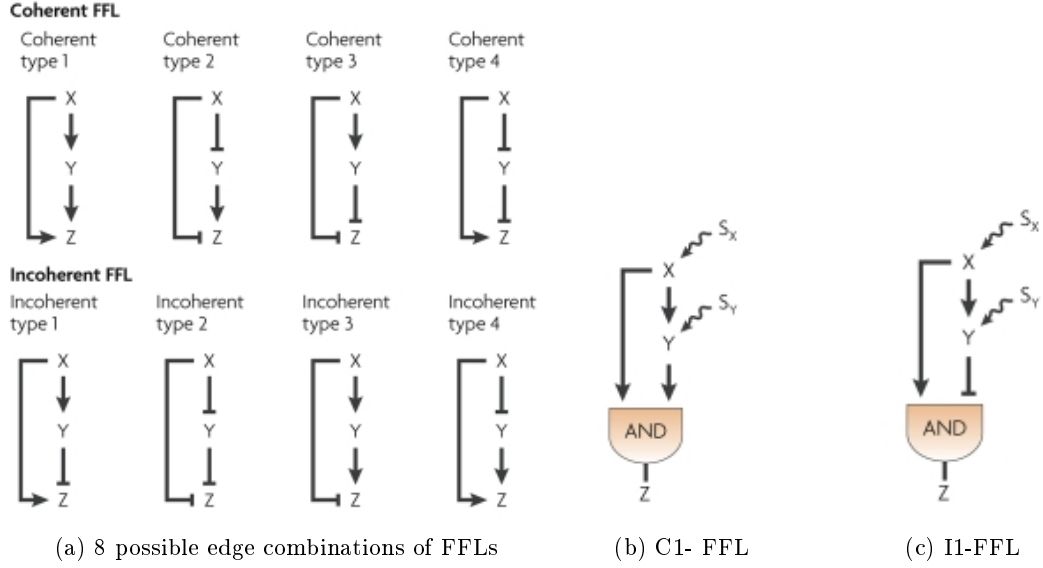


Figure 2.3: Types of FFL network motif. In the Figure 2.3a, the coherent and the incoherent types of FFLs are given. Figure 2.3b and Figure 2.3c demonstrates C1-FFL and I1-FFL with an AND input function, respectively. The explanations given in the Section 2.2.2.1 and 2.2.2.2 are based on the assumption of the existence of the signal S_Y .

2.2.2.1 Coherent Type 1 FFL

The coherent type I FFL (C1-FFL) type using an "AND" input function causes a sign-sensitive delay in the activation of the jointly targeted gene. This dynamic process starts when the activation of X is triggered by a signal S_x which results in the accumulation of Y. The production of Z begins after the concentration of Y reaches the activation threshold for the promoter of Z. Although C1-FFL generates delay in the activation step, the inactivation process occurs without delay, since the off-state of S_x inactivates X which causes a rapid deactivation of Z by the reason of the AND logic.

By courtesy of the sign sensitive delay, C1-FFL with the AND gate functions as a filter for the inconsistent pulses in the fluctuating cell environment.

The C1-FFL type with an "OR" gate function works as the opposite direction with the "AND" input, i.e. the delay does not occur in the beginning when X is in an active form but after the removal of the S_x signal, the production of Z does not end rapidly because of the fact that the accumulated Y is sufficient to continue the process. By this mechanism, type I coherent FFLs compensate for the short termed loss of signal [1].

2.2.2.2 The Incoherent Type 1 FFL

In the incoherent type I FFL (I1-FFL), while Z is activated by X directly, it is also repressed by Y through X , indirectly. That is; when X becomes active, the production of Z increases but after some time the accumulation of Y passes a threshold for the Z promoter and starts to repress it. This process results in pulse-like dynamics.

The I1-FFL also functions as a response accelerator such that before Y starts to repress Z , with a strong initial production, Z can reach the non-zero steady state which results in shortening the response time. After the shut-down of X , the production of Z rapidly decreases, and the concentration of Z shows an exponential decay with a speed based on its degradation rate [2].

2.2.3 Single Input Module

SIMs are larger network motifs where a group of genes are controlled by a superior regulator which also regulates itself usually. The coordinated regulation of these genes creates a dynamic process by generating temporal expression as a result of their different activation thresholds which stems from the variations in sequence and position of binding sites. The temporal activation ensures that a gene is not expressed before it is needed by activating the gene with the lowest threshold, then the second lowest and so forth. This kind of a motif is seen frequently in metabolic pathways to form the desired product and in damage repair systems to produce a response to a stress such as DNA damage or heat shock [34] [1].

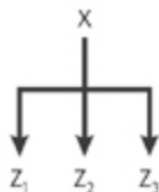


Figure 2.4: SIM network motif in which the regulator X controls three genes, Z_1 , Z_2 , and Z_3 .

2.2.4 Bi-fans and Dense Overlapping Regulons

Although there are 199 possible interactions in the 4-node subgraphs, only 2 of them, namely multi output FFL and bi-fan, are significant motifs in transcription regulation networks. In bi-fan motifs, two regulators combinatorially control the expression of two genes [1].

On the other hand, there exists a larger significant network motif called the dense overlapping regulon, which is a complex form of bi-fans where a group of genes are jointly controlled by a group of regulators. The DORs function as gateways; that the

multiple inputs are processed to control each output. Although DORs are not fully connected, i.e. a gene is not targeted by all transcription factors within the motif, the number of connections is much more than those of patterns found in a random graph [40].

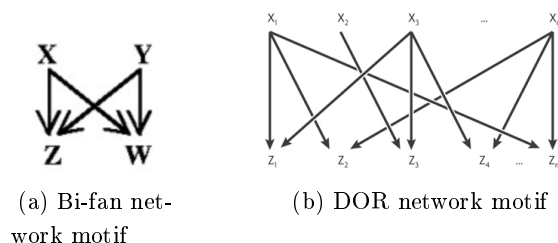


Figure 2.5: Bi-fan motif and dense overlapping regulon.

2.3 Analysis of Regulatory Networks

2.3.1 Analysis of Transcription Regulation Networks

In 2002, a study was conducted by Alon et al. in order to find significant motifs in complex networks [28]. An algorithm was developed in order to reveal the significant networks within a directed complex network by searching for n -node sub-graphs and compare the numbers with those obtained from random networks. The randomization procedure preserves the number of incoming and outgoing edges for every node in the real network in order to capture the characteristics caused by a single node such as target hubs. The randomization code is implemented using complex networks from different fields, i.e. biochemistry (transcriptional regulation network), ecology, neurobiology, and engineering.

The two transcriptional regulation networks from an eukaryote (*Saccharomyces cerevisia*) and a bacterium (*Escherichia coli*) are used where the genes are represented with nodes and the edges are directed from TF to gene that is regulated by that TF. The results obtained using 1000 randomized networks show that there exist two significant patterns which are FFL and bi-fan. The related statistical measures and the number of edges and nodes can be seen from Figure 2.6 which also enables the comparison of the random networks with the real regulatory network, i.e. while the number of FFLs are found as 40 in *E. coli*, the mean of the number of FFLs obtained from 1,000 generated random networks equals to 7 [28]. Given Z -scores implies the significance of the network motifs in *E. coli* and *S. cerevisia*.

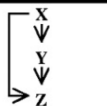

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	
Gene regulation (transcription)				Feed-forward loop			Bi-fan		
	<i>E. coli</i>	424	519	40	7 ± 3	10	203	47 ± 12	13
	<i>S. cerevisiae</i> *	685	1,052	70	11 ± 4	14	1812	300 ± 40	41

Figure 2.6: Statistically significant motifs found in transcriptional regulation network (re-printed from [28]). In the Figure, N_{real} and N_{rand} indicate the number of genes in real and randomized networks, in order. SD denotes the standard deviation and $Z - score$ shows its test statistic.

2.3.2 Analysis of Combined Transcriptional - Post Transcriptional Regulatory Networks

In a study presented by Zhou et. al. (2007), the coordinated regulation of TFs and miRNAs are studied using the targets of 168 human miRNAs and 236 human transcription factors that are obtained from PicTar and TRANSFAC, respectively. Various statistical tests, such as Fisher's exact test, chi-square test are applied to the dataset with the aim of determining the strength of interplay between the pairs of miRNA-miRNA, TF-TF, miRNA-TF which leads to the conclusion that TF-TF and miRNA-miRNA interactions are more frequent than TF-miRNA interplay. Moreover, it is revealed that within the regulators, at least one of them involves in targeting many genes and some particular TFs are interacting with nearly every miRNA. As a next step of the study, it is hypothesized that formation of a feed forward loop where the TF activates its target and miRNA represses both the TF and its target concurrently is possible if the interaction between the TF and the miRNA is strong meaning that if they have many shared targets. To test this hypothesis, Fisher's exact tests were conducted which resulted in the finding that such feed forward loops were significant among the TF-miRNA pairs that are highly interacting [48].

Another study in the area of combinatorial regulation of TFs and miRNAs is conducted by Shalgi et al.(2007). The analysis is implemented based on the data obtained from PicTar and TargetScan, which are the two databases for the miRNA target prediction. In this study, evolutionary conserved miRNAs are used, and the orthologous genes among human, mouse, rat and dog are selected from the targets of miRNAs in order to eliminate the false positive results of mapping miRNAs to genes. The adjacency matrices described in the Section 3.1.1.1 are constructed for the connection data obtained from two databases, separately and the degree preserving edge swapping randomization procedure is applied to data. First of all, the differences between original data and the randomized data are examined in terms of degree distributions, i.e the distribution of the number of targets belonging to each miRNA and the number of miRNAs belonging the each gene. It is observed that the distributions are different from each other in terms of their widths and shapes. In spite of the fact that the original distribution contains many target hubs (which are the genes such that each is regulated by a vast amount of miRNAs), in the randomized network the genes that are targeted by more than 10 miRNAs are few in number. In other words, while the distributions of the randomized data look Gaussian, the distribution of the original network is long-right tailed. These findings led to the study in a direction where dense target

hubs corresponding to the data which are above the 85th percentile of the distribution are examined using Gene Ontology (GO) and it was revealed that the developmental processes are highly regulated by these target hubs.

Secondly, it was tested to see whether miRNAs combinatorially regulate gene expression by interacting other miRNAs using a similar randomization approach applied in the degree distribution analysis. The connection matrix is randomized 1000 times in a degree preserving manner with 100,000 swapping steps. After every randomization iteration, the "Meet/Min" score given in Equation 2.1 is calculated for each pair of miRNAs by which 107 and 199 significant pairs are detected among the data obtained from TargetScan and PicTar databases, respectively. Furthermore, a hierarchical model which contains a few miRNAs with many allies at the top and miRNAs with very few allies at the bottom is identified. It is also reported that the degree distribution of the network is power law.

$$\frac{|Targets(i) \cap Targets(j)|}{\min(|Targets(i)|, |Targets(j)|)} \quad (2.1)$$

Thirdly, since an interplay between miRNAs and TFs in regulating mutual targets seems to be a strong possibility, a similar randomization procedure described above is applied to the data to confirm this possibility by revealing the co-occurrence of miRNA-TF pairs. In addition to the miRNA-target matrix, an adjacency matrix for TF connections is created. Both matrices are randomized in order to calculate co-occurrence rates and p-values. 104 significant TF-miRNA pairs in the TargetScan dataset and 916 significant pairs in the PicTar dataset are identified and the network motif analysis is conducted by using these significant pairs.

Finally, after identifying the TF-miRNA pairs that regulate the same target, it is analysed to determine whether the pairs regulate each other by forming FFLs via the randomization procedure. It is observed that the regulation of TFs by miRNAs (type II coherent FFL) and the regulation of miRNAs by TFs (type I incoherent FFL) are both significant. This conclusion, obtained from the sequence information, is supported experimentally by using the expression data to calculate the correlation coefficients between miRNAs and TFs among all pairs which result in observing high correlation between these either negatively or positively. Negative correlations are justified by considering TFs which function as repressors and the inhibitory property of miRNAs [38].

Another study, which is the primary predecessor for this thesis was conducted by Iwama et al. (2010) in order to analyse the differences between TFs and miRNAs based on their contributions and the effect of their evolutionary distinctions to the gene regulatory networks through cFFC formation (see Section 3.2.1).

The datasets are obtained from NCBI by using Build 36.3 and Build 37.1 for human and mouse, respectively. Non-overlapping genes with respect to their 8 kb upstream sequences are selected and the human-mouse orthologous gene pairs are identified which result in obtaining 5169 genes.

In order to obtain miRNA-target connections, the mature miRNA sequences were downloaded by using miRBase and the miRNA target sites of 5169 genes are identified via the PITA program based on the exact 8 nt or 7 nt match between the miRNA sequences and the 3' UTR regions of the genes. The identified human miRNA target

sites are aligned with the 3' UTR regions of mouse sequences by using ReAlignerVR and the identical ones are selected. As a result, 564 mature miRNAs with 34298 connections are obtained.

After the identification of 83 TF and 564 miRNAs, the adjacency matrices are created with the objective of conducting a partial randomization procedure which includes only miRNA-connection matrix randomization, only TF-connection matrix randomization and both TF-miRNA connection matrices randomization. Matrices are randomized using a swapping step algorithm in a degree preserving manner which is explained in detail in the Section 3.3.1.1. 1000 matrices are generated by applying 500,000 swapping steps to the original matrices in every iteration and the number of cFFCs, the number of cFFC targeted genes and the cFFC redundancy are obtained from every generated matrix in order to make a comparison with the corresponding numbers from real GRNs.

First of all, the contribution of TFs and miRNAs is analysed in terms of the cFFC formation based on the results of three randomization procedures. It is observed that while the number of cFFCs is 44373 in the real GRN, the expected number of cFFCs equals 44117.3 for only miRNA connection matrix randomization, 44161.3 for only TF connection matrix randomization and 43766.4 for both TF-miRNA connection matrices randomization. The Z-scores and p-values which are calculated under the assumption of normality led to the conclusion that there is no significant deviation in terms of cFFC formation between the real GRNs and the expectations that are obtained from all three randomization procedure. In other words, cFFCs are neither excessively nor inadequately represented in the real GRNs. Besides the number of cFFCs, the number of cFFC targeted genes are also examined for every generated matrix. The real GRNs contain 2476 genes that are targeted at least one cFFC. Although there is no significant deviation in the number of cFFC targeted genes between the real GRNs and the expectation of the TF connection matrix randomization, miRNA randomization and both TF-miRNA randomization reveal that the number of targeted genes of real GRNs is excessively reduced from the expectations and the difference is statistically significant.

Because of the existence of target hubs, cFFC redundancy which is the ratio of the total number of cFFCs to the number of cFFC targeted genes, is calculated in order to find the average number of cFFCs per gene. It is reported that the cFFC redundancy is significantly higher in the real GRN with respect to both the miRNA connection randomization and the TF-miRNA connection randomization. The significant deviation in terms of the number of cFFC targeted genes and the cFFC redundancy implies that the TF networks stay steady and do not give response to the alterations of miRNA network structure. Whereas the miRNA networks contribute to the GRNs by causing de-escalation of the target gene repertoire or increasing the redundancy through the cFFC formation. However, the excessive representation of the redundancy can be caused by the sharp reduction of the cFFC targeted genes. Hence, the miRNA derived redundancy and the TF derived redundancy are examined separately. The results show that the miRNA derived redundancy is substantially higher than random expectations in the real GRNs as the TF derived redundancy shows no significant deviation from random expectations for the miRNA-connection randomization and the TF-miRNA connection randomization. This implies that the miRNA derived redundancy is not a consequence of the reduction of the cFFC targeted genes. Although

the findings that imply steadiness of TF networks are consistent with each other, it is possible that the different number of edges of miRNAs and TFs within cFFC may exert such an effect. Therefore, one edge randomization procedure which is described in detail in Section 3.3.2 was applied in this study. The results do not only support the previous findings, but also enlighten the significant distinctness of the miRNA configuration changing in terms of the miRNA-TF connections and the miRNA-target connections. It is reported that the number of cFFCs is inadequately represented in the miRNA target connection and is excessive in the miRNA-TF connection. Moreover, the miRNA-target connection randomization shows that the miRNA derived redundancy is represented excessively whereas the TF derived redundancy is adequately presented in the real GRN.

CHAPTER 3

RANDOMIZATION ALGORITHMS AND ANALYSIS METHODS

3.1 Data

In this study, the interplay between TFs and miRNAs are analysed in the scope of cFFCs, cFFC targeted genes, and the cFFC redundancy and TF derived cFFC redundancy, miRNA derived cFFC redundancy using two different datasets that diverge from each other in terms of the number of edges and nodes that are included in, as well as the type of cell. The detailed information about the first and the second dataset are given in the Section 3.1.1 and Section 3.1.2, respectively.

3.1.1 Dataset Retrieved from the Study of Iwama et al.

First dataset whose collection procedure is reported in the Section 2.3.2 was obtained by Iwama et al. for their own study which is conducted to find the significance of the miRNAs in the composite cFFCs. The dataset is available online¹ as a two separate files in which the TF-target and miRNA-target information given as connectivity matrix. The first data file, is given in plain text format, is the TF-connection matrix of which the first row contains the HGNC symbols of 83 TFs and the first column shows the entrez ids of 5169 genes including the 83 TFs. Second data file given in PDF format includes mirBase ids of 564 miRNAs in its first row, and entrez ids of 5169 genes in the first column. There exist 124740 human-mouse conserved TF regulatory connections and 34298 human-mouse conserved miRNA regulatory connections.

3.1.1.1 Matrix Representation of the Connection Data

In the graphical view of the network, the nodes are represented with the genes and the miRNAs and the edges represent the interactions between nodes which are directed from the TF or miRNA to the gene that is regulated by them. In order to represent the interactions, the adjacency matrix consisting of zeros and ones is created according to rule given in the Equation 3.1 and Equation 3.2 [16]. For the TF connectivity matrix:

¹ <http://mbe.oxfordjournals.org/content/28/1/639/suppl/DC1>

$$M_{ij} = \begin{cases} 1, & \text{if the gene } i \text{ is targeted by the TF } j. \\ 0, & \text{otherwise.} \end{cases} \quad (3.1)$$

For the miRNA connectivity matrix:

$$\mu_{ij} = \begin{cases} 1, & \text{if the gene (or TF) } i \text{ is targeted by the miRNA } j. \\ 0, & \text{otherwise.} \end{cases} \quad (3.2)$$

As a result, two connectivity matrices where the existing connections between nodes are shown with the ones are obtained with the following form:

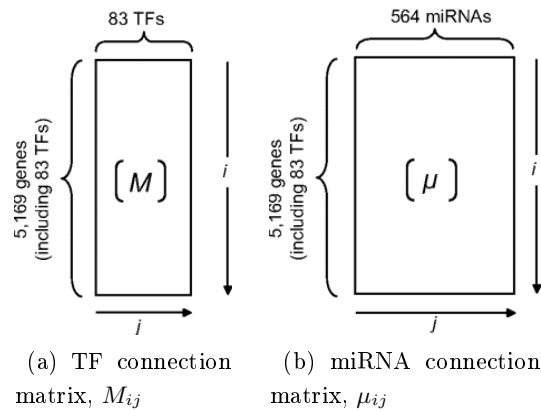


Figure 3.1: Form of connection matrices [16]

3.1.2 Dataset Retrieved from FANTOM4 EdgeExpress Database

The second dataset, will be called FANTOM network through the document, collected from the Functional Annotation of the Mammalian Genome 4 (FANTOM4) database which is an international collaborative project focusing on the transcriptional landscape in the mammalian genome with the aim of making the network predictions accessible [18].

In FANTOM4, the predictions are made based on the expression profiles and promoter regions of a human myeloid leukemia cell line which result in obtaining the interactions of TFs and the genes with the weight parameter which is an indicator of the significance of the predicted connections [19] [43]. Additionally, siRNA perturbation edges, chromatin immunoprecipitation edges, protein-protein interaction edges and miRNA target gene edges are made available to users [18].

The FANTOM4 EdgeExpress database also offers users two types of visualization options which are the center view showing all the regulatory elements, possible interactions and the expression graphs of the queried gene and the subnet view providing a graphical network view consisting of nodes and edges for the queried set of genes (and/ or miRNAs).

3.1.2.1 FANTOM Network Information According to Different Weight Thresholds

The target information of TFs and miRNAs given within the FANTOM4 Database are obtained by downloading XML files which contain the target information of TFs and miRNAs, and parsing these files with the Python script developed (given in Appendix B.2 for TF connections and B.1 for miRNA connections) according to the different weight parameters. The target information is transferred to the CSV files separately for every resultant network, both of which consist of two columns; the first column contains the HGNC symbol of TFs and mirBase IDs of miRNAs and the second column contains the HGNC symbols of genes which are targeted by the TF and/or miRNA given in the first column.

The human-mouse orthologs of the genes that are included in the resulting networks are identified using homologue data which are available online ², using a Python script (given in Appendix B.3). The miRNA homologs are retrieved from the miRBase database and are cross matched with the same Python script to be able to take the human-mouse ortholog subset of the FANTOM miRNA connection dataset.

The total number of human-mouse conserved connections of miRNAs and TFs are provided in the Table 3.1 with the total number of miRNAs, TFs and genes that are included in the network for the given weight parameter. In addition to the network information, the changes in the original numbers of research interests which are described in detail in the Section 3.2 are shown in the Figure 3.2 according to different weights.

Although it is observed that the number of cFFCs, the number of the cFFC targeted genes, cFFC redundancy, miRNA derived cFFC and TF derived redundancy which are given in the Figure 3.2 show little deviation after the 1.2 weight threshold, in order to obtain the highly significant predictions of connections with minimum number of false positives, the weight parameter is set to 1.5 as suggested in the study conducted by Suzuki et. al. (2009) [43]. Hence, the randomization is implemented to the network which contains 137 miRNAs, 274 TFs and 6749 genes from which the total number of human-mouse conserved regulatory connections are identified as 6631 for miRNAs and 60969 for TFs.

3.2 Key Terms Used in the Network Analysis

3.2.1 Composite Feed Forward Circuits

As mentioned in the Section 2.2.2, FFLs are the most abundant and significant network motifs that found are in biological networks such as transcriptional regulator network and are initially defined as three node subgraphs consisting of one general TF which is also known as a master regulator, one specific TF and their shared target gene [28]. On the other hand, since the aim of this study is to uncover the differences between

² <ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene/current/>

Table 3.1: Network information for different weights

	number of miRNAs	number of TFs	number of genes	number of miRNA connections	number of TF connections
weight ≥ 0	197	274	10504	93782	98621
weight ≥ 0.1	197	274	10497	93782	96766
weight ≥ 0.2	197	274	10486	93782	94802
weight ≥ 0.3	197	274	10467	93782	92764
weight ≥ 0.4	197	274	10453	93782	90585
weight ≥ 0.5	197	274	10434	93782	88288
weight ≥ 0.6	192	274	9913	75423	85956
weight ≥ 0.7	184	274	9469	60277	83453
weight ≥ 0.8	173	274	8957	44593	80890
weight ≥ 0.9	167	274	8339	26307	78036
weight ≥ 1.0	160	274	7869	15720	75373
weight ≥ 1.1	156	274	7571	12858	72680
weight ≥ 1.2	153	274	7340	10603	69742
weight ≥ 1.3	141	274	7117	8956	66846
weight ≥ 1.4	139	274	6922	7635	63847
weight ≥ 1.5	137	274	6749	6631	60969

TFs and miRNAs in terms of their contribution to the GRNs, the FFL is specified as a composite feed-forward circuit (cFFC) which is also a significant biological network motif that is composed of a miRNA as a master regulator, a TF and a gene in which both TF and gene are repressed by the miRNA whereas the shared target of the TF-miRNA pair is stimulated by TF [48].



Figure 3.3: A unitary cFFC. The unitary cFFC involves miRNA, TF and gene. miRNA represses its targets concurrently, whereas TF functions as an activator for the shared target.

In this thesis, the real GRNs and the randomized networks are searched for such cFFCs, and the resulting total number of cFFCs are compared using the methods given in the Section 3.4.

3.2.2 cFFC Targeted Genes

Because of the fact that a gene can participate in more than one cFFC as illustrated in the Figure 3.4, the cFFC targeted genes are defined as the set of genes that are included in at least one cFFC [16].

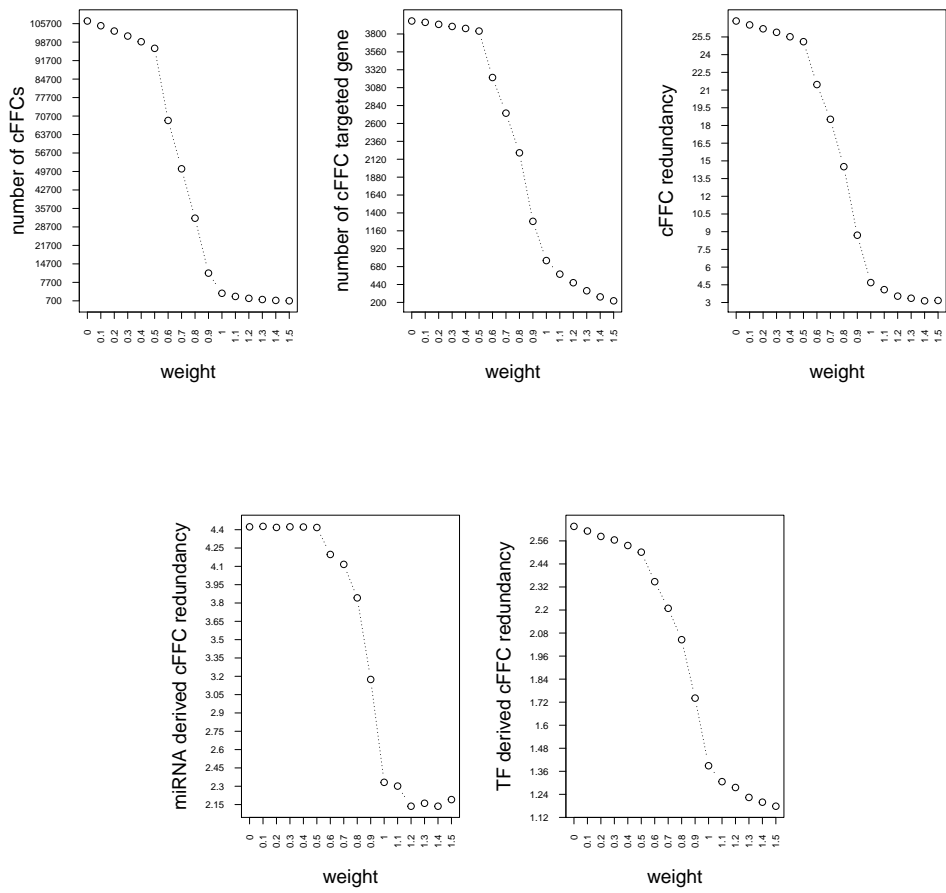


Figure 3.2: Change in the initial numbers according to different weight thresholds

Within the scope of this study, the number of the cFFC targeted genes are obtained for both real GRNs and its random permutations.

3.2.3 cFFC Redundancy

As mentioned before and illustrated in the Figure 3.4, the redundancy of the network may be increased by forming cFFCs around a shared targeted gene by additional loops of TFs and miRNAs. In other words, some genes are regulated by more than one cFFC. Therefore, cFFC redundancy is described as the average number of cFFCs that are involved in regulating a gene [16]. The cFFC redundancy is calculated by taking the ratio of total number of cFFCs to the number of cFFC targeted genes.

In order to see the main cause in the formation of cFFC redundancy, it is partitioned into two factors, namely; "miRNA derived cFFC redundancy" and "TF derived cFFC redundancy".

3.2.3.1 miRNA Derived cFFC Redundancy

The miRNA derived redundancy, whose pattern is shown in the Figure 3.4b is obtained by taking the ratio of total number of cFFCs to the total number of TF-gene connections that are included in cFFCs. The aim is to determine the redundancy which originates from cFFCs that involve the extra miRNA loops that occur around the same TF-gene edge. For example, in the Figure 3.4b, the miRNA derived cFFC redundancy equals three because of the fact that it only includes one TF-gene edge, while there exist three cFFCs that are constituted by extra miRNA loops.

3.2.3.2 TF Derived cFFC Redundancy

In a similar manner with the miRNA derived redundancy, the TF derived redundancy is defined as the ratio of the total number of cFFCs to the total number of miRNA-gene edges that are involved in cFFCs in order to determine the cFFC redundancy that is caused by additional TF loopings around the same miRNA-gene edges. In Figure 3.4c, the TF derived redundancy equals three since the module contains three cFFC with the one shared miRNA-gene connection.

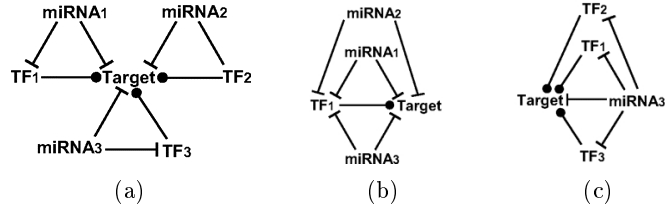


Figure 3.4: Sub-network views for multiple cFFCs [16]. Figure 3.4a, 3.4b, and 3.4c illustrate the patterns of the genes that are targeted by more than one cFFC. Figure 3.4b and Figure 3.4c also represent the sub-networks that are made redundant by miRNAs and TFs, respectively.

3.3 Randomization Procedure

3.3.1 Main Randomization Procedure

In this study, three main randomization procedures are applied to both datasets, which are; miRNA only connection matrix randomization, TF only connection matrix randomization and both TF and miRNA connection matrix randomization. Randomizations are conducted using a swapping step algorithm which results in preserving the degrees of nodes as in the real GRN. This implies that the algorithm enables the TFs and miRNAs to keep their degrees fixed while changing the destination of their outgoing edges, and genes to keep their number of incoming edges fixed while changing their regulator. The randomization of the connectivity matrices are conducted 1000 times using 500000 swapping steps for each. In every randomization, the number of cFFCs, the number of cFFC targeted genes which is defined as the number of genes that are targeted by at least one cFFC, miRNA-gene edges and TF-gene edges that participates in cFFCs are reported. Besides, the miRNA derived redundancy and the TF derived redundancy are calculated by dividing the total number of cFFCs to the total number of the cFFC participant miRNA-gene edges and the TF-gene edges for each randomization iteration.

3.3.1.1 Pseudocode of the Main Randomization Procedure

The data files that contain the TF connections and the miRNA connections are converted to hash tables (the dictionary data type) using the Algorithm 1 and 2 for Iwama et. al. data and FANTOM data, respectively. After creating dictionaries, the countforeffc function is used to calculate the number of cFFCs and the cFFC dependent terms such as cFFC targeted genes, cFFC redundancy etc. It holds the number of the cFFC targeted genes, the cFFC participant miRNA-gene edge number and the cFFC participant TF-gene edge number in the dictionaries which are genecountdict, miredgecountdict and tfedgecountdict, respectively.

The randomization of the connections are conducted using the randomizematrix function.

The source codes are provided in Appendix B.4.

Procedure 1 Python Dictionary Creator Function for IwamaEtAl Dataset

```
function READMATRIX(filename)
  open filename
  headerline  $\leftarrow$  first row
  locs  $\leftarrow$  first line [1 to length of headerline]
  mydict  $\leftarrow$  empty dictionary
  for  $\forall$  key  $\in$  locs do
    mydict[key]  $\leftarrow$  empty dictionary
  end for
  for  $\forall$  line  $\in$  filename do
    values  $\leftarrow$  line entities separated by comma
    geneid  $\leftarrow$  first entity of values
    interactions  $\leftarrow$  remaining entities of values
    for  $\forall i \in$  range of interactions' length do
      if  $i^{th}$  element of interactions equals 1 then
        set mydict[locs[i]][geneid] to 1
      end if
    end for
  end for
  close filename
  return (mydict, locs)
end function
```

Procedure 2 Python Dictionary Creator Function for FANTOM Dataset

```
function READMATRIX(filename)
  open filename
  locs  $\leftarrow$  empty list
  mydict  $\leftarrow$  empty dictionary
  for  $\forall$  line  $\in$  filename do
    values  $\leftarrow$  line entities separated by comma
    regulator  $\leftarrow$  first entity of the values
    geneid  $\leftarrow$  second entity of the values
    if regulator  $\notin$  locs then
      add regulator to locs
    end if
    if regulator is not a key of mydict then
      mydict[regulator]  $\leftarrow$  empty dictionary
    end if
    mydict[regulator][geneid]  $\leftarrow$  1
  end for
  close filename
  return (mydict, locs)
end function
```

Procedure 3 cFFC Count Function

```
function COUNTFORCFFC(mirna, mirdict, tfdict, genecountdict, miredgecountdict,
tfedgecountdict)
  targets of mir  $\leftarrow$  mirdict[mirna]
  count  $\leftarrow$  0
  for  $\forall$  t in targets do
    if t is a key of tfdict then
      for  $g \in$  tfdict[t] do
        if g is a key of targets then
          if t is not equal to g then
            increase count by 1
            if g is a key of genecountdict then
              increase genecountdict[g] by 1
            else
              set genecountdict[g] to 1
            end if
            if m is not a key of miredgecountdict then
              miredgecountdict[m]  $\leftarrow$  empty dictionary
            else
              set miredgecountdict[m][g] to 1
            end if
            if t is not a key of tfedgecountdict then
              tfedgecountdict[t]  $\leftarrow$  empty dictionary
            else
              set tfedgecountdict[t][g] to 1
            end if
          end if
        end if
      end for
    end if
  end for
  return count
end function
```

Procedure 4 Connection Randomization Function

```
function RANDOMIZEMATRIX(mydict)
  firstnode  $\leftarrow$  randomly selected key from mydict
  secondnode  $\leftarrow$  randomly selected key from mydict
  while length of mydict[firstnode] is 0 do
    reselect firstnode
  end while
  while length of mydict[secondnode] is 0 do
    reselect secondnode
  end while
  firsttarget  $\leftarrow$  randomly selected key from mydict[firstnode]
  secondtarget  $\leftarrow$  randomly selected key from mydict[secondnode]
  if mydict[secondnode] has key firsttarget or mydict[firstnode] has key secondtarget then
    return 0
  end if
  remove firsttarget from mydict[firstnode]
  add secondtarget to mydict[firstnode]
  set mydict[firstnode][secondtarget] to 1
  remove secondtarget from mydict[secondnode]
  add firsttarget to mydict[secondnode]
  set mydict[secondnode][firsttarget] to 1
  return 1
end function
```

Procedure 5 Applications of Functions

```
number of replicates  $\leftarrow$  1000
number of swapping steps  $\leftarrow$  500000
number of cFFCs  $\leftarrow$  open writable file
number of cFFC targeted genes  $\leftarrow$  open writable file
mirna derived redundancy  $\leftarrow$  open writable file
tf derived redundancy  $\leftarrow$  open writable file
(mirdict,mirlist)  $\leftarrow$  READMATRIX("mirna_connection_matrix.csv")       $\triangleright$  mirdict
holds miRNA-target information
(tfdict,tflist)  $\leftarrow$  READMATRIX("tf_connection_matrix.csv")       $\triangleright$  tfdict holds
TF-target information
for  $i \in$  range 0 to number of replicates do
  swaps  $\leftarrow$  0
  while swaps < number of swapping steps do
    swaps  $\leftarrow$  swaps + RANDOMIZEMATRIX(mirdict)
  end while
  while swaps < number of swapping steps do
    swaps  $\leftarrow$  swaps + RANDOMIZEMATRIX(tfdict)
  end while
  swaps  $\leftarrow$  0
  swaps 2  $\leftarrow$  0
  while swaps < number of swapping steps do
    swaps  $\leftarrow$  swaps + RANDOMIZEMATRIX(mirdict)
  end while
  while swaps 2 < number of swapping steps do
    swaps 2  $\leftarrow$  swaps 2 + RANDOMIZEMATRIX(tfdict)
  end while
```

} Only miRNA
connection matrix
randomization

} Only TF con-
nection matrix
randomization

} Both miRNA - TF
connection matrices
randomization

Procedure 5 Applications of Functions (Continued)

```
genecountdict ← empty dictionary           ▷ cFFC targeted genes
mirEdgeCountDict ← empty dictionary       ▷ cFFC participant miRNA-target
pairs
tfEdgeCountDict ← empty dictionary       ▷ cFFC participant TF-target pairs
total ← 0                                 ▷ total number of cFFCs
mirtargetededgecount ← 0                 ▷ cFFC participant miRNA-target edge count
tftargetededgecount ← 0                 ▷ cFFC participant TF-target edge count
for mir in mirdict do
    count ← count + COUNTFORCFFC(mir,mirdict, tfdict, genecount-
dict,mirEdgeCountDict,tfEdgeCountDict)
    total ← total + count
end for
for  $\forall x \in$  keys of mirEdgeCountDict do
    mirtargetededgecount ← mirtargetededgecount + length of mirEdgeCountDict[x]
end for
for  $\forall y \in$  keys of tfEdgeCountDict do
    tftargetededgecount ← tftargetededgecount + length of tfEdgeCountDict[y]
end for
mirnaderivedredundancy ← float(total) / (mirtargetededgecount)
tfderivedredundancy ← float(total) / (tftargetededgecount)
write length of genecountdict to file "number of cFFC targeted genes"
write total to file "number of cFFCs"
write mirnaderivedredundancy to file "mirna derived redundancy"
write tfderivedredundancy to file "tf derived redundancy"
end for
```

3.3.2 Partial Randomization Procedure

Partial randomization procedure consists of two sub-processes, which are the miRNA-TF connections randomization and the miRNA - non-TF connection randomization. This is done to reveal whether the difference between the edge numbers of TFs and miRNAs in an unitary cFFC affects the results, or not. In the miRNA-TF connection randomization, the edges between miRNAs and the genes that are not TF are kept unchanged, while in the miRNA - non-TF target randomization, the miRNA - TF connections are kept fixed. The partial randomization procedure is applied to only miRNA connection matrix since the procedure corresponds to the TF randomization in a way that it swaps one edge for miRNA or one edge for TF [16].

In the same way with the main randomization procedure, the degree preserving swapping step algorithm is used to partially randomize the miRNA connection matrix with 500,000 swapping steps to generate 1,000 random networks. In every randomization, the number of cFFCs, the number of the cFFC targeted genes, the miRNA derived cFFC redundancy and the TF derived cFFC redundancy is reported.

The partial randomization results obtained from Iwama et. al. data are discussed in Section 4.2 and the randomization results belonging to the FANTOM data are discussed in the Section 4.4. The source codes are also provided in Appendix B.5.

3.3.2.1 Pseudocode of the Partial Randomization Procedure

3.3.2.1.1 miRNA - Gene Connection Only Randomization Procedure 6 and 7 given below are used with the Procedures 1, 2 and 3, 6 in the randomization of the miRNA - non-TF target connections. The randomizematrix function is modified to not select miRNA-TF edges for swapping, and the main part in which the functions are called is revised to randomize only miRNA connections data. The source codes are presented in Appendix B.5.1.

Procedure 6 Connection Randomization Function

```
function RANDOMIZEMATRIX(mydict,tfnames)
  firstnode  $\leftarrow$  randomly selected key from mydict
  secondnode  $\leftarrow$  randomly selected key from mydict
  while length of mydict[firstnode] is 0 do
    reselect firstnode
  end while
  while length of mydict[secondnode] is 0 do
    reselect secondnode
  end while
  firsttarget  $\leftarrow$  randomly selected key from mydict[firstnode]
  secondtarget  $\leftarrow$  randomly selected key from mydict[secondnode]
  if firsttarget  $\in$  tfnames or secondtarget in tfnames then
    return 0
  end if
  if mydict[secondnode] has key firsttarget or mydict[firstnode] has key secondtarget then
    return 0
  end if
  remove firsttarget from mydict[firstnode]
  add secondtarget to mydict[firstnode]
  set mydict[firstnode][secondtarget] to 1
  remove secondtarget from mydict[secondnode]
  add firsttarget to mydict[secondnode]
  set mydict[secondnode][firsttarget] to 1
  return 1
end function
```

Procedure 7 Applications of Functions

```
number of replicates ← 1000
number of swapping steps ← 500000
number of cFFCs ← open writable file
number of cFFC targeted genes ← open writable file
mirna derived redundancy ← open writable file
tf derived redundancy ← open writable file
(mirdict,mirlist) ← READMATRIX("mirna_connection_matrix.csv")    ▷ mirdict
holds miRNA-target information
(tfdict,tflist) ← READMATRIX("tf_connection_matrix.csv")        ▷ tfdict holds
TF-target information
for i ∈ range 0 to number of replicates do
  swaps ← 0
  while swaps < number of swapping steps do
    swaps ← swaps + RANDOMIZEMATRIX(mirdict)
  end while
  genecountdict ← empty dictionary                                ▷ cFFC targeted genes
  mirEdgeCountDict ← empty dictionary                            ▷ cFFC participant miRNA-target
pairs
  tfEdgeCountDict ← empty dictionary                             ▷ cFFC participant TF-target pairs
  total ← 0                                                       ▷ total number of cFFCs
  mirtargetededgecount ← 0                                       ▷ cFFC participant miRNA-target edge count
  tftargetededgecount ← 0                                       ▷ cFFC participant TF-target edge count
  for mir in mirdict do
    count ← count + COUNTFORCFFC(mir,mirdict, tfdict, genecount-
dict,mirEdgeCountDict,tfEdgeCountDict)
    total ← total + count
  end for
  for ∀ x ∈ keys of mirEdgeCountDict do
    mirtargetededgecount ← mirtargetededgecount + length of mirEdgeCountDict[x]
  end for
  for ∀ y ∈ keys of tfEdgeCountDict do
    tftargetededgecount ← tftargetededgecount + length of tfEdgeCountDict[y]
  end for
  mirnaderivedredundancy ← float(total) / (mirtargetededgecount)
  tfderivedredundancy ← float(total) / (tftargetededgecount)
  write length of genecountdict to file "number of cFFC targeted genes"
  write total to file "number of cFFCs"
  write mirnaderivedredundancy to file "mirna derived redundancy"
  write tfderivedredundancy to file "tf derived redundancy"
end for
```

3.3.2.1.2 miRNA - TF Connection Only Randomization This, in addition to the Procedures 1, 2 and 3, and 4, involves functions that create inverse dictionaries of the miRNA connection matrix, extracts the non-TF connections from the miRNA connection matrix and merges the extracted connections with the miRNA-TF connections in order to count the number of cFFCs, cFFC targeted genes, miRNA derived cFFC redundancy and TF derived cFFC redundancy after randomization. The

pseudocode of the functions mentioned is given as Procedures 8, 9 and 10, and their implementation steps are provided in Procedure 11. The source codes are included in Appendix B.5.2.

Procedure 8 Inverse Dictionary Creator Function

```

function INVERTDICT(mydict)
  inversedict  $\leftarrow$  empty dictionary
  for all key, value pairs in mydict do
    for all subkey, subvalue pairs in mydict[key] do
      if subkey is not a key of inversedict then
        inversedict[subkey]  $\leftarrow$  empty dictionary
      end if
      set inversedict[subkey][key] to subvalue
    end for
  end for
  return inversedict
end function

```

Procedure 9 Function to Extract miRNA - TF Edges from miRNA Connection Dictionary

```

function EXTRACT(inversedict,locs,mydict)
  excludedinteractions  $\leftarrow$  empty dictionary
  allgenes  $\leftarrow$  keys of inversedict
  onlygenes  $\leftarrow$  allgenes - locs
  for all key,value pairs in mydict do
    for all subkey,subvalue pairs in mydict[key] do
      if subkey  $\in$  onlygenes then
        delete mydict[key][subkey]
        if key is not a key of excludedinteractions then
          excludedinteractions[key]  $\leftarrow$  empty dictionary
        end if
        set excludedinteractions[key][subkey] to 1
      end if
    end for
  end for
  if length of mydict[key] equals 0 then
    delete mydict[key]
  end if
end for
  return (mydict,excludedinteractions)
end function

```

Procedure 10 Function to Merge Extracted miRNA-gene Edges with the Randomized miRNA-TF Edges

```
function MERGE(mydict, excludedinteractions)
  for key in excludedinteractions do
    if key in mydict then
      if mydict[key] and excludedinteractions[key] is instances of dictionary
    then
      add entities of excludedinteractions[key] to mydict[key]
    end if
    else
      mydict[key] ← excludedinteractions[key]
    end if
  end for
  return mydict
end function
```

Procedure 11 Applications of Functions

```
number of replicates ← 1000
number of swapping steps ← 500000
number of cFFCs ← open writable file
number of cFFC targeted genes ← open writable file
mirna derived redundancy ← open writable file
tf derived redundancy ← open writable file
(mirdict,mirlist) ← READMATRIX("mirna_connection_matrix.csv")    ▷ mirdict
holds miRNA-target information
(tfdict,tflist) ← READMATRIX("tf_connection_matrix.csv")        ▷ tfdict holds
TF-target information
mirtargetdict ← INVERTDICT(mirdict) ▷ reverse miRNA connection dictionary in
which keys are miRNA targets
for i ∈ range 0 to number of replicates do
  swaps ← 0
  mergedmirdict ← empty dictionary
  (mirtfdict,mirgenedict) = EXTRACT(mirtargetdict, tflist, mirdict))
  while swaps < number of swapping steps do
    swaps ← swaps + RANDOMIZEMATRIX(mirdict)
  end while
  genecountdict ← empty dictionary                                ▷ cFFC targeted genes
  mirEdgeCountDict ← empty dictionary                            ▷ cFFC participant miRNA-target
pairs
  tfEdgeCountDict ← empty dictionary                             ▷ cFFC participant TF-target pairs
  total ← 0                                                       ▷ total number of cFFCs
  mirtargetedgedcount ← 0                                         ▷ cFFC participant miRNA-target edge count
  tftargetedgedcount ← 0                                         ▷ cFFC participant TF-target edge count
  mergedmirdict ← MERGE(mirtfdict,mirgenedict)
  for mir in mirdict do
    count ← count + COUNTFORCFFC(mir,mergedmirdict, tfdict, genecount-
dict,mirEdgeCountDict,tfEdgeCountDict)
    total ← total + count
  end for
```

Procedure 11 Applications of Functions (Continued)

```
for  $\forall x \in$  keys of mirEdgeCountDict do
    mirtargetededgecount  $\leftarrow$  mirtargetededgecount + length of mirEdgeCountDict[x]
end for
for  $\forall y \in$  keys of tfEdgeCountDict do
    tftargetededgecount  $\leftarrow$  tftargetededgecount + length of tfEdgeCountDict[y]
end for
mirnaderivedredundancy  $\leftarrow$  float(total) / (mirtargetededgecount)
tfderivedredundancy  $\leftarrow$  float(total) / (tftargetededgecount)
write length of genecountdict to file "number of cFFC targeted genes"
write total to file "number of cFFCs"
write mirnaderivedredundancy to file "mirna derived redundancy"
write tfderivedredundancy to file "tf derived redundancy"
end for
```

3.4 Statistical Analysis

3.4.1 Normality Tests

3.4.1.1 Quantile-Quantile Plots (Q-Q Plots)

The quantile-quantile (Q-Q) plot is a specialized scatter plot of two CDFs in which the quantiles of a CDF, $F(x)$ are plotted against the corresponding quantiles of a CDF, $G(x)$ in order to assess the goodness of fit and provide graphical means in estimating parameters of distributions which belong to the same location-scale family [21]. In other words, the Q-Q plot gives initial information about the underlying distribution before conducting further statistical analyses and without specifying the parameters. Its usefulness also arises from that every entry has a unique position in the plot and it is possible to represent whole dataset within the plot without the need for arbitrary categories.

To reveal the underlying distribution, the comparison can be made with two empirical CDFs, or one theoretical and one empirical CDF. If the compared distributions are identical, the quantiles are spread on a straight line in $y = x$ axis. When $F(x)$ is a linear function of $G(x)$, the quantiles still form a straight line but the slope of the line differs from 1 which indicates the distributions are mainly similar except their location and scale parameters [46].

Because of the fact that the quantiles, also known as plotting positions, are a function of rank i and the sample size n that shows rapid changes for sparse densities and slow changes for high densities, it is sensitive to the disparities in the tails of distributions [12].

The general formula to calculate plotting positions is

$$p = (i - a)/(n + 1 - a), \quad (3.3)$$

where a is a distinctive constant within the interval $[0, 0.5]$ based on the distribution

taken as reference, i.e. for normal distribution, the best suitable formula is the Blom where the a equals 0.375 [15].

Normal probability plots are a specialized type of Q-Q plots where standard normal distribution is used as a theoretical distribution in order to test the normality of empirical CDF, visually. The observed values are ranked and the corresponding plotting positions are calculated. Then, for every plotting position, each quantile from both distributions are paired to construct a normal probability plot [15]. Note that the quantiles of standard normal distribution can be obtained from a standard normal distribution table or from a statistical package which contains computationally approximated values of the inverse standard normal distribution.

In this study, the normal probability plots are constructed using R for miRNA connection only randomization, TF connection only randomization, and both TF - miRNA connection randomization based on Iwama et. al. and the FANTOM dataset to be able to get an insight of whether the normality assumption holds, or not. The results are discussed in Section 4.1.1 and 4.3.1.

3.4.1.2 Shapiro-Wilk Test

Besides assessing the goodness of fit of the datasets to the normal distribution using probability plots, the datasets are also tested to detect if there exists departure from normality using a powerful parametric approach that is the Shapiro-Wilk's W statistic which was originally constructed by Shapiro and Wilk and applicable to the datasets of which the sample size, n , is within the interval [3, 50] [39]. The W statistic is defined for a random sample $y_1 < y_2 < \dots < y_n$ as

$$W = \frac{(\sum_{i=1}^n a_i y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3.4)$$

where y_i is the i^{th} order statistic and \bar{y} is the sample mean. The calculation of a_i given in the Equation 3.5 is conducted according to the expected values of the order statistics of normally and identically distributed independent random variables, denoted by m and the corresponding covariance matrix, represented as V [31].

$$a_i = (a_1, a_2, \dots, a_n) = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m}}. \quad (3.5)$$

Since the test has disadvantages such as not being appropriate for the samples larger than 50, and being ill-suited for computer implementation, Royston introduced an extension to the W statistic by revising the weight approximation, and assigning a transformation that normalize the distribution of the W statistic by which the calculation of the exact p-value for the sample sizes within the interval [3, 5000] becomes convenient [37].

The value of the W statistic is always between 0 and 1, and the obtaining a W statistic which is close to 1 indicates that the dataset comes from a normal distribution [31].

In this study, the normality of the datasets which are obtained from the real GRN by applying randomization procedures is tested using Shapiro and Wilk's W statistic, and the results are discussed in Section 4.8 and 4.18.

3.4.2 One Sample Z - Test

One sample Z-test, which is also known as a one sample location test is a hypothesis testing method which is conducted in order to reveal the difference between the population mean from which the observations are randomly sampled and a particular value which is stated in the null hypothesis, under the assumption of the normality with known variance [4]. In the case that the variance is unknown, it can be substituted by its estimator obtained from the sample when it is large enough. Under its assumptions, the Z-test is more convenient than the t-test because of the fact that it provides particular significance levels for every Z-statistic that is calculated from the sample while t-test uses one critical value based on the sample size and specified alpha level.

In this study, the Z-test is conducted to test whether there exists significant deviation between the numbers obtained from real GRNs and the randomized networks based on the number of cFFCs, the number of the cFFC targeted genes, the cFFC redundancy, the miRNA derived cFFC redundancy and the TF derived redundancy and its formula is given in Equation 3.6.

$$Z = \frac{(\text{value obtained from real GRNs} - \text{mean})}{\text{standard deviation}}, \quad (3.6)$$

where the mean and the standard deviation obtained from random networks which are provided in Table 4.2, 4.9 and Table 4.14, 4.20 and the results of the Z-test are discussed in Section 4.1.2.2, 4.2.2.2 and Section 4.3.2.2, 4.4.2.2 for Iwama et. al. and the FANTOM datasets, respectively.

3.4.3 Two-Sample t-Test

The two-sample t-test is a hypothesis test which is conducted under several assumptions to reveal whether there exists a difference between the means of the populations from which the samples consisting of the randomly selected and independent observations are drawn. Besides the requisite of randomness and independence of the observations, the normality assumption must be satisfied in order to compare the equality of means using the two-sample t-test. In other words the distribution of the populations that is samples are selected from should be normal density. Homogeneity of variance implying the equality of the variances of the populations is an another assumption that is required to conduct two-sample t-test [8]. The usage of the two-sample t-test in the case of the violation of the homogeneity of variance assumption is introduced by Welch with an approximation method that considers degrees of freedom, ν as a random variable and estimates its distribution from the Pearsonian Type III Curve [44].

The calculation of the t-statistic depending on the conditions related to the population

variances are given in Equation 3.7 by [36]

$$t = \begin{cases} \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, & \text{for known variances.} \\ \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, & \text{for equal and unknown variances.} \\ \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, & \text{for unequal and unknown variances.} \end{cases} \quad (3.7)$$

In this study, the samples consisting of independent and random observations are tested for the normality and the homogeneity of variances assumptions using the Shapiro-Wilk test and the Variance Ratio F-test, respectively. Then, the two-sample t-test for unequal variances, also known as Welch's t-test, is conducted in order to reveal the differences between the population means of the miRNA derived redundancy and the TF derived redundancy obtained from the same randomization procedure and the results are discussed in Section 4.1.2.5, 4.2.2.3, 4.3.2.3 and 4.4.2.3.

CHAPTER 4

EXPERIMENTAL RESULTS AND THE ANALYSIS OF SIMULATED NETWORKS

4.1 Analysis of the Main Randomization Results Based on Iwama et. al. Data

4.1.1 Q-Q Plots and Normality Tests

In order to conclude whether the data come from a normal distribution, the quantiles of the data are scaled as the ratio of the values to the maximum value of the data set, and are plotted against the quantiles of the normal distribution. Getting a plot in which the data points are spread through a straight line on the $y = x$ axis is a strong indicator of the normality. According to the graphs, there exist deviations in the tails, and the datasets do not perfectly spread through the $y = x$ axis which is a sign of non-normality.

The Q-Q plot belonging to the number of cFFCs obtained after the miRNA connection randomization is given in Figure 4.1. Q-Q plots belonging to the number of cFFCs, the number of the cFFC targeted genes, the cFFC redundancy, the miRNA derived cFFC redundancy and the TF derived cFFC redundancy after the randomization procedure can be seen in Appendix A.

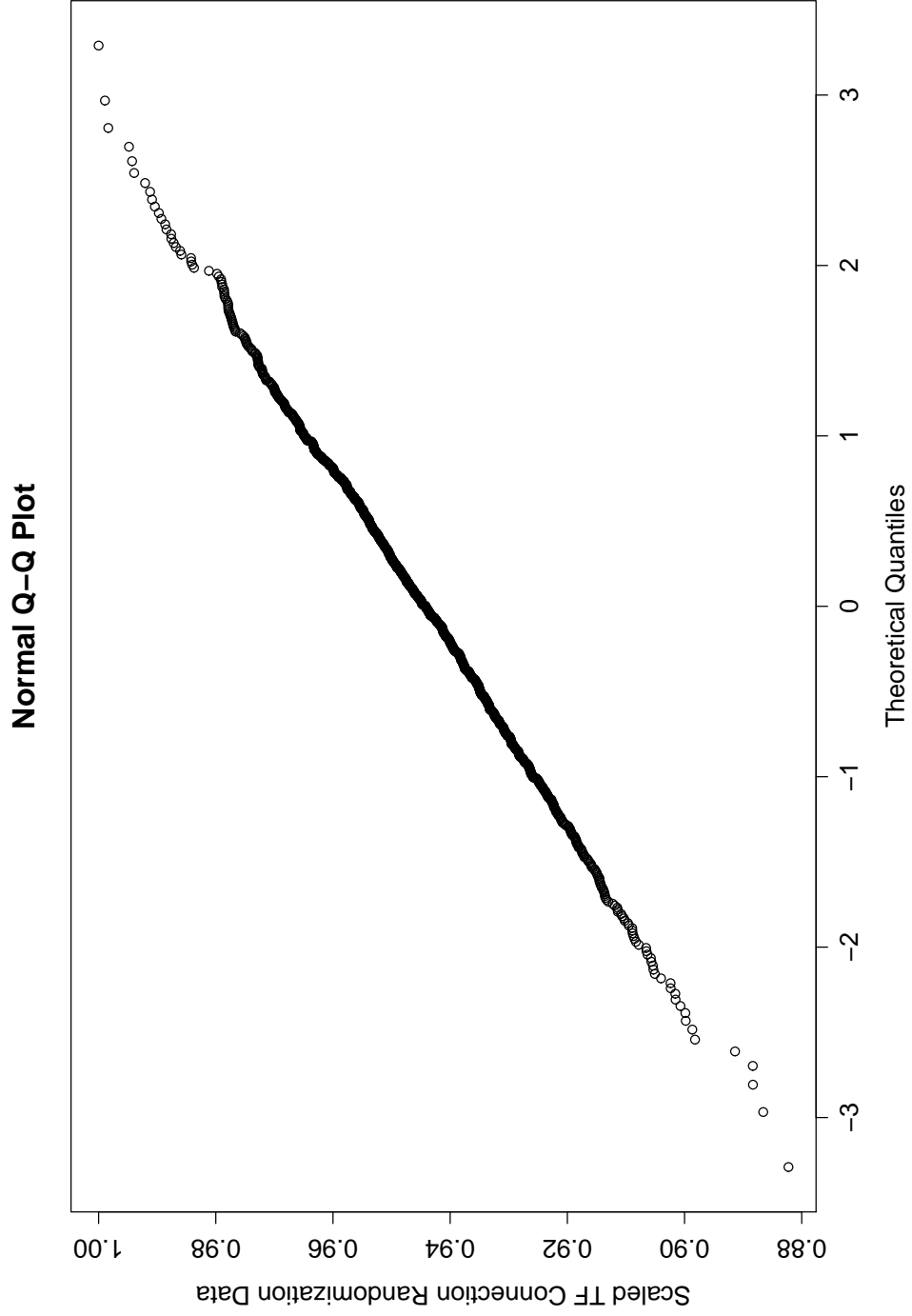


Figure 4.1: QQ plot of normality of number of cFFCs after miRNA connection randomization

As an alternative of Q-Q plot, a parametric test can be used to test normality and the Shapiro-Wilk test is the most well known test for this purpose. Thereby, this test is conducted in order to verify that the data of the number of cFFCs, the number of the cFFC-targeted genes, the cFFC redundancies, the miRNA derived cFFC redundancy and the TF derived cFFC redundancy obtained after the randomization step, are normally distributed. The results are given in Table 4.8.

Table 4.1: Shapiro-Wilk Test Results - Iwama et. al. Data

	mir connection randomization		tf connection randomization		tf-mir both connection randomization	
	W	p-value	W	p-value	W	p-value
No. of cFFCs	0.9986	0.6264	0.9988	0.7451	0.999	0.862
No. of cFFC-targeted genes	0.9981	0.3354	0.9982	0.397	0.9981	0.3047
cFFC redundancy	0.9985	0.5873	0.9988	0.7287	0.9986	0.5897
miRNA derived redundancy	0.999	0.8609	0.9974	0.1084	0.9979	0.2451
TF derived redundancy	0.9991	0.9019	0.998	0.2683	0.9988	0.7254

According to the Shapiro-Wilk test results, all p-values are greater than 0.01 indicating that the normality assumption holds. The following analyses are based on this assumption.

4.1.2 Comparison of the Simulated Data via Real Gene Regulatory Network

4.1.2.1 Random Expectations

The comparison of the simulated data with the real gene regulatory network is conducted in three main aspects, which are the number of cFFCs, the number of cFFC targeted genes and the cFFC redundancy. The results obtained after each randomization procedure are given in Table 4.2. The numbers given in the first column are obtained from real GRN. The rest of the columns shows the numbers which belong to simulated data after the miRNA connection only randomization, the TF connection only randomization and both the miRNA-TF connection randomization, respectively. The mean and the standard deviations of the simulated data are also provided in Table 4.2.

Table 4.2: Randomization Results of Iwama et. al. Data

	Real GNR Observed	miRNA Connection Randomized Expected \pm Standard Deviation	TF Connection Randomized Expected \pm Standard Deviation	Both miRNA and TF Connection Randomized Expected \pm Standard Deviation
Number of cFFCs	43481	42997.65 \pm 872.0966	43544.15 \pm 232.0001	43094.92 \pm 871.8188
Number of genes targeted by cFFCs	2497	2687.458 \pm 14.47494	2500.146 \pm 11.69853	2684.651 \pm 15.56006
cFFCs redundancy	17.41	15.99948 \pm 0.3180161	17.41698 \pm 0.1177703	16.05251 \pm 0.3197785

4.1.2.2 Comparison of Random Expectations with Real GRN via Z-test

To be able to compare the simulated data with the real GRN, and make inferences about research questions, Z-statistics are calculated according to mean and standard deviation, and the corresponding p-values are obtained, under the assumption of normality (Table 4.3).

Table 4.3: Comparison of Z and P Values - Iwama et. al. Data

	mir randomization		tf randomization		tf-mir randomization	
	Z	p-value	Z	p-value	Z	p-value
Number of cFFCs	0.5542333	0.2897096	-0.2722024	0.3927332	0.442842	0.32894
Number of cFFC-targeted genes	-13.15778	$7.676671 * 10^{-40}$	-0.2689227	0.3939946	-12.05979	$8.611052 * 10^{-34}$
cFFC redundancy	4.445757	$4.379151 * 10^{-6}$	-0.03128567	0.4875209	4.255421	$1.043278 * 10^{-5}$

The calculated Z-statistics and corresponding p-values are used to test whether there exist significant deviations between the numbers obtained from real GRN and the random expectations which are given in Table 4.2. The values in the columns are related to the miRNA-connection randomization, the TF-connection randomization and both TF-miRNA connection randomization, respectively.

Firstly, the hypothesis that whether the number of cFFCs belonging to real GRN is equal to the number of cFFCs of that random expectations is tested. These values indicate that there is no significant deviation between the number of cFFCs of simulated data and the real GRN meaning that the number of cFFCs in the real GRN are neither over-represented nor under-represented.

Secondly, the hypothesis mentioned above is examined in terms of the number of the cFFC targeted genes. The second row contains the Z-statistics and p-values which belong to the number of cFFC targeted genes calculated from simulated data. Small p-values, regarding miRNA-connection and both TF-miRNA connection randomization results point out that there is a significant deviation between the number of the cFFC targeted genes of the real GRN and the simulated data which implies that the number of cFFC targeted genes decreases sharply from the random expectations of the miRNA-connection randomization and both TF-miRNA connection randomization.

Thirdly, the mentioned hypothesis is tested regarding the cFFC redundancy. Since the p-values of the cFFC redundancy calculated after miRNA-connection randomization and both TF-miRNA connection randomization are very small, it leads to the conclusion that there exists a significant deviation between the cFFC redundancy of simulated data and the real GRN. It is also concluded that the cFFC redundancy in the real GRN is represented excessively according to the miRNA-connection randomization and the TF-miRNA connection both randomization. This conclusion puts forward another research question because of the fact that the increased cFFC redundancy might be an after-effect of the sharp decrease in the cFFC targeted gene numbers.

4.1.2.3 Random Expectations of TF Derived and miRNA Derived Redundancy

In order to reveal the effect of miRNAs and TFs in the increase of the cFFC redundancy and to prove that the over-representation of the cFFC redundancy in real GRN is not caused by the sharp reduction of the cFFC targeted genes, the miRNA derived redundancy and the TF derived redundancy are analysed separately. By that, it is tested that whether the redundancy adding role is a distinguishing property of miRNAs.

As mentioned in Section 3.2.3, the miRNA derived cFFC redundancy is defined as the ratio of the total number of cFFCs to the total number of the TF-gene edges that is included in cFFCs. In a similar manner, the TF derived cFFC redundancy is described as the ratio of the total number of cFFCs to the total number of miRNA-gene edges that participates in cFFCs. The miRNA derived cFFC redundancy and the TF derived cFFC redundancy obtained from real GRN are given in Table 4.4 with the means and standard deviations of the simulated data obtained from three different randomization procedures.

Table 4.4: Randomization Results of miRNA and TF Derived Redundancy Obtained from Iwama et. al. Data

	Real GNRs Observed	miRNA Connection Randomized Expected \pm Standard Deviation	TF Connection Randomized Expected \pm Standard Deviation	Both miRNA and TF Connection Randomized Expected \pm Standard Deviation
miRNA derived redundancy	3.12	2.13 \pm 0.027	3.13 \pm 0.014	2.14 \pm 0.027
TF derived redundancy	2.03	1.98 \pm 0.046	2.02 \pm 0.009	1.99 \pm 0.046

4.1.2.4 Comparison of Redundancies with respect to Real GRN via Z-test

To be able to compare the random expectations given in the Table 4.4 with the real GRN statistically, the Z-scores and the corresponding p-values which are shown in the Table 4.5, are calculated.

Table 4.5: Comparison of Real GRN with Random Networks in terms of miRNA Derived and TF Derived Redundancy - Iwama et. al. Data

	mir randomization		tf randomization		tf-mir randomization	
	Z	p-value	Z	p-value	Z	p-value
miRNA derived cFFC redundancy	36.40619	1.698901 $\times 10^{-290}$	-0.5866912	0.2787055	35.61502	4.101235 $\times 10^{-278}$
TF derived cFFC redundancy	0.8839155	0.1883709	0.1814938	0.42799	0.7814491	0.2172692

According to the Z-test results given in Table 4.5, it is clearly seen that the miRNA derived cFFC redundancy in real GRN shows significant deviation from random networks that are created by the miRNA connection randomization (p-value = $1.698901 \times 10^{-290}$) and the TF-miRNA connection both randomization (p-value = $4.101235 \times 10^{-278}$) implying that the real GRN is made redundant by miRNAs, while there is no significant

difference between real GRN and random networks of all three randomization procedures in terms of the TF derived cFFC redundancy. Concordant with the previous results, TF networks again remain steady by showing no alterations to the conformational changes in the miRNA networks.

4.1.2.5 Comparison of Redundancies via the t-test

Besides the comparison of the results which are produced from randomization with that of real GRN, the TF derived redundancy and the miRNA derived redundancy that are obtained from the same randomization procedure are also compared by conducting the Welch’s t-test because of the fact that the assumption of homogeneity of variance which is a requirement of the two-sample t-test is not satisfied according to the variance ratio F-test of which the results are given in Table 4.6.

Table 4.6: Results of Variance Ratio F-test

	miRNA Connection Randomization		TF Connection Randomization		Both TF-miRNA Connection Randomization	
	F-statistic	p-value	F-statistic	p-value	F-statistic	p-value
miRNA Derived cFFC Redundancy	0.351	$< 2.2 \times 10^{-16}$	2.3605	$< 2.2 \times 10^{-16}$	0.3535	$< 2.2 \times 10^{-16}$
TF Derived cFFC Redundancy						

Table 4.7: Results of Welch’s t-test

	miRNA Connection Randomization		TF Connection Randomization		Both TF-miRNA Connection Randomization	
	t-statistic	p-value	t-statistic	p-value	t-statistic	p-value
miRNA Derived cFFC Redundancy	81.0665	$< 2.2 \times 10^{-16}$	2076.938	$< 2.2 \times 10^{-16}$	88.7143	$< 2.2 \times 10^{-16}$
TF Derived cFFC Redundancy						

The hypothesis that there exists no significant difference between the means of the miRNA derived redundancy and the TF derived redundancy is tested for the miRNA connection only randomization, the TF connection only randomization and both the TF-miRNA connection randomization, separately, and the results are given in Table 4.7. Although the result of the t-test between miRNA derived redundancy and TF derived redundancy is reported as non-significant for the dataset obtained from only TF randomization procedure in the study conducted with the same real GRNs by Iwama et. al. (2010), the test statistics obtained in this thesis are large and the corresponding p-values are smaller than 0.01 for all three randomization procedures. Hence, the null hypothesis is rejected and it is concluded that the population means of the miRNA derived redundancy and the TF derived redundancy are different for only miRNA connection randomization, only TF connection randomization and both TF-miRNA connection randomization. It is also revealed that the miRNA derived redundancy is more frequent than TF derived redundancy which leads to the conclusion that causing redundancy is a distinct property of miRNAs.

Although the results that are obtained so far revealed the differences of miRNAs and TFs in terms of their contribution to the GRNs, there is a possibility that the results could be affected by the difference between the out-degrees of miRNAs and TFs that

is included in an unitary cFFC. In other words, since a unitary cFFC contains two miRNA edges and one TF edge and accordingly the miRNA connection matrix includes both genes and TF as its targets, the randomization procedure shuffles two edges in miRNA randomization. Thereby, the partial randomization procedures miRNA-TF targets only randomization and miRNA-gene targets only randomization are applied to dataset in order to control whether the imbalance between the edge numbers in a unitary cFFC creates biased results while randomizing the real GRN.

4.2 Analysis of Partial Randomization Results Based on Iwama et. al. Data

4.2.1 Q-Q Plots and Normality Tests

The Q-Q plots of the number of cFFCs, the number of the cFFC targeted genes, cFFC redundancies, miRNA derived cFFC redundancies and TF derived cFFC redundancies which are obtained from the simulated data as a result of partial randomization procedures which is conducted to the miRNA connection matrix by randomizing only miRNA-gene edges and only miRNA - TF edges, separately, are provided in Appendix A.

The plots share similarities with the Q-Q plots of simulations which are generated by using main randomization procedure in terms of the deviations from the straight line in tails. In order to conclude that whether this deviations cause non-normality, the Shapiro - Wilk test is conducted and the test statistics are given in Table 4.8 with the corresponding p-values.

The Q-Q plot of the number of the cFFC targeted genes from the miRNA-gene connection randomization and the miRNA-TF connection randomization are given in Figure 4.2 and Figure 4.3, respectively, as an example of the departure from the normality line at the tails.

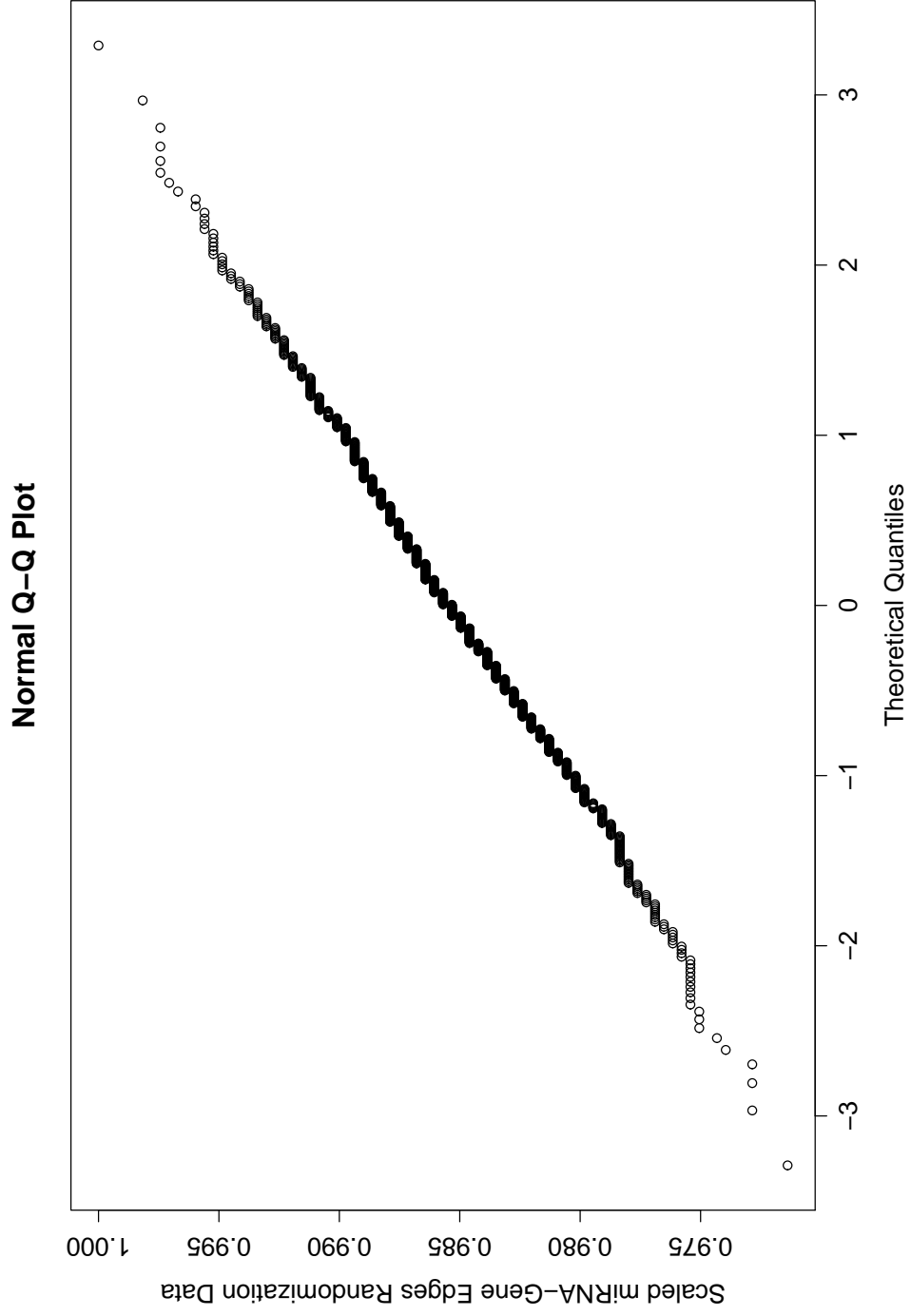


Figure 4.2: QQ plot of normality for only miRNA-gene connection randomization for number of cFFC targeted genes

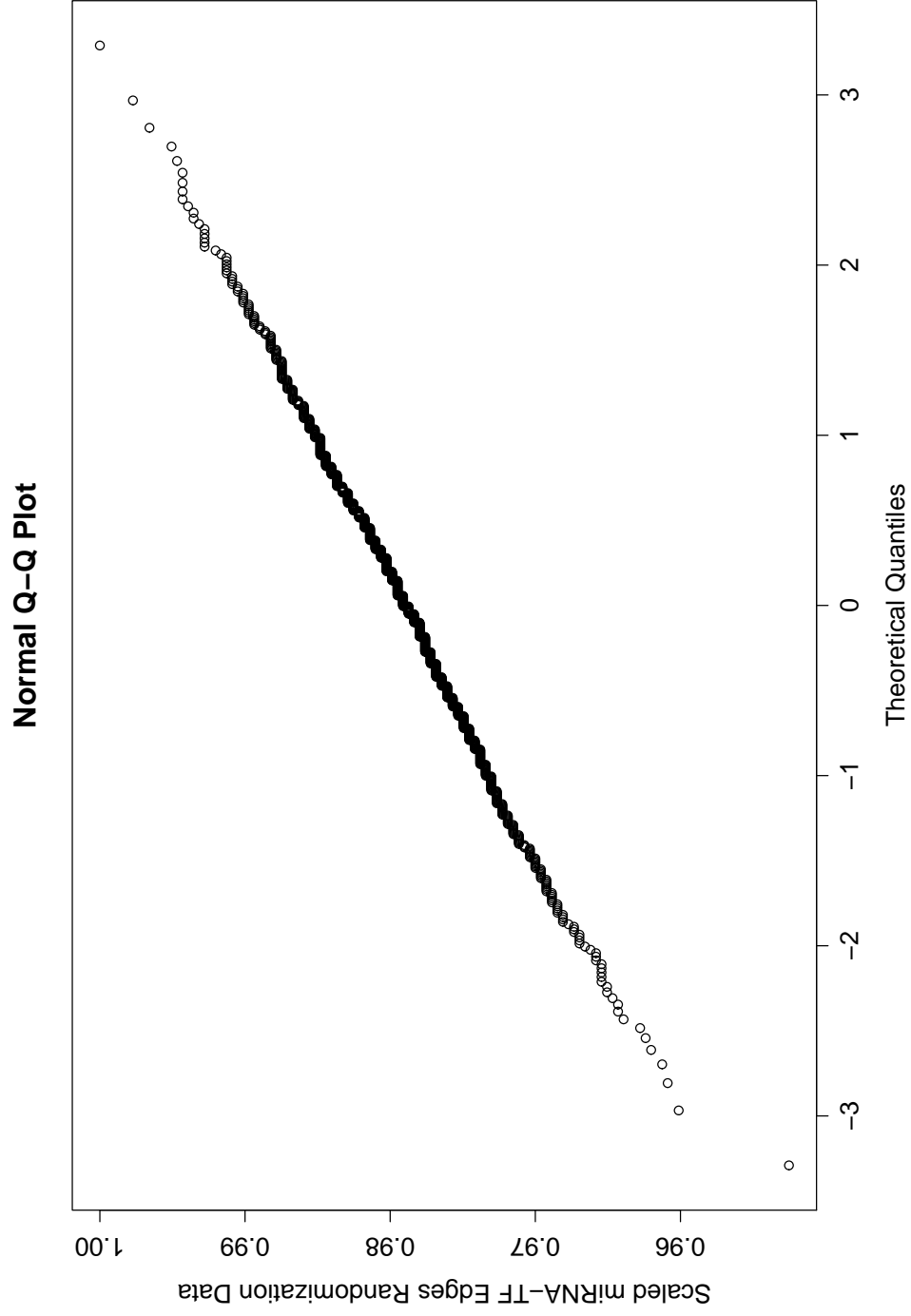


Figure 4.3: QQ plot of normality for only miRNA-TF connection randomization for number of cFFC targeted genes

Table 4.8: Shapiro-Wilk Test Results of Partial Randomization - Iwama et. al. Data

	mir-gene connection randomization		mir-tf connection randomization	
	W	p-value	W	p-value
No. of cFFCs	0.9985	0.5464	0.9989	0.8295
No. of cFFC-targeted genes	0.9973	0.09949	0.9973	0.09224
cFFC redundancy	0.998	0.2963	0.9983	0.4133
miRNA derived redundancy	0.9991	0.9019	0.9986	0.5891
TF derived redundancy	0.9985	0.5543	0.9972	0.07987

According to the Shapiro - Wilk test results, the datasets are distributed normally since all the p-values are greater than 0.01 which enables the usage of parametric approaches such as the Z-test.

4.2.2 Comparison of the Simulated Data via Real Gene Regulatory Network

4.2.2.1 Random Expectations

The numbers observed from real GRN are provided in the first column of the Table 4.9. The second and third columns contain the random expectations and the corresponding standard deviations belonging to the number of cFFCs, the number of cFFC targeted genes, the cFFC redundancies, the miRNA derived redundancy and the TF derived redundancy which are calculated from the simulated data that are generated from the miRNA-gene connection randomization and the miRNA-TF connection randomization. To be able to make statistical inferences about the difference between the observed values and the random expectations, the Z-test is conducted, for which results are discussed in the next section.

Table 4.9: Results of Partial Randomization of Iwama et. al. miRNA Connection Data

	Real GNR Observed	Randomization with Fixed miRNA-TF Edges Expected \pm Standard Deviation	Randomization with Fixed miRNA-Gene Edges Expected \pm Standard Deviation
Number of cFFCs	43481	43807.27 \pm 117.2996	42324.5 \pm 511.6457
Number of genes targeted by cFFCs	2497	2685.918 \pm 12.71436	2577.867 \pm 16.04771
cFFCs redundancy	17.41	16.31034 \pm 0.08750752	16.41923 \pm 0.2356964
miRNA derived cFFCs redundancy	3.12	2.267732 \pm 0.008714465	2.217392 \pm 0.03349587
TF derived cFFCs redundancy	2.03	2.040665 \pm 0.005687421	2.000262 \pm 0.03349587

4.2.2.2 Comparison of Random Expectations with Real GRNs via Z-test

Being consistent with the previous results, the partial randomization revealed that there exists a significant difference between the real GRN and the random networks in

terms of the cFFC targeted genes and the cFFC redundancy for both the miRNA-TF connection randomization and the miRNA-gene connection randomization. Accordingly, the cFFC targeted genes are represented inadequately whereas the cFFC redundancy is over-represented, in the real GRN. Additionally, it is also observed that the difference between real GRN and the random networks is statistically significant in terms of the miRNA derived redundancy which is also over-represented in real GRNs for both partial randomization procedure.

In contrast with the over representation of miRNA derived redundancy, the TF derived redundancy is inadequately represented in real GRN compared with the random networks for miRNA-gene connection randomization and it shows no deviation between the real GRN and the random networks for miRNA-TF randomization.

As opposed to the previous results, in which the number of cFFCs in random networks show no significant deviation from that of real GRN, one sees two different characteristics in partial randomizations. For the miRNA-TF connection randomization, the observed value is significantly different from the random expectation of the number of cFFC in the favor of real GRN. In contrast to over-representation in miRNA-TF connection randomization, it shows a decrease from the random expectation in miRNA-gene connection randomization which implies that miRNAs tend to regulate their targeted gene indirectly through TFs by reducing the cFFC formation while the cFFC formation is preferred in the direct effects of miRNAs on TFs.

Table 4.10: Comparison of Z and P Values of Partial Randomization based on Iwama et. al. miRNA Connection Data

	Randomization with Fixed miRNA-TF Edges		Randomization with Fixed miRNA-Gene Edges	
	Z	p-value	Z	p-value
Number of cFFCs	-2.781475	0.002705627	2.260347	0.01189985
Number of cFFC-targeted genes	-14.85863	3.058015e-50	-5.039163	2.337862e-07
cFFC redundancy	12.6042	1.000975e-36	4.217594	1.234613e-05
miRNA derived cFFC redundancy	97.57376	0	26.88819	1.509447e-159
TF derived cFFC redundancy	-2.586778	0.004843903	0.8112788	0.2086028

4.2.2.3 Comparison of Redundancies via t-test

In order to test the difference between the miRNA derived redundancy and the TF derived redundancy obtained from randomization with fixed miRNA-TF edges, the Welch's t-test is applied to the results since the equality of variances assumption is not satisfied according to the variance ratio F-test whose p-value is given in Table 4.11. Hence all assumptions which are prerequisite of the two sample t-test are satisfied, it is used to test the difference between the population means of miRNA derived redundancy and TF derived redundancy. The variance ratio F-test and the t-test results are provided in Table 4.11 and Table 4.12, respectively.

Table 4.11: Results of Variance Ratio F-test - Partial Randomization of Iwama et. al. Data

	Randomization with Fixed miRNA-TF Edges		Randomization with Fixed miRNA-Gene Edges	
	F-statistic	p-value	F-statistic	p-value
miRNA Derived cFFC Redundancy	2.3477	$< 2.2 * 10^{-16}$	1.1189	0.07608
TF Derived cFFC Redundancy				

Table 4.12: Results of two sample t-test - Partial Randomization of Iwama et. al. Data

	Randomization with Fixed miRNA-TF Edges		Randomization with Fixed miRNA-Gene Edges	
	t-statistic	p-value	t-statistic	p-value
miRNA Derived cFFC Redundancy	2.3477	$< 2.2 * 10^{-16}$	148.958	$< 2.2 * 10^{-16}$
TF Derived cFFC Redundancy				

According to the t-test results, the population means of the miRNA derived redundancy and the TF derived redundancy are not equal and the miRNA derived cFFC redundancy is greater than the TF derived cFFC redundancy for both partial randomization procedure.

Given the fact that the results obtained from the main randomization procedure are confirmed with the partial randomization results regarding the under representation of cFFC targeted genes, over representation of the cFFC redundancy and the miRNA derived cFFC redundancy, and obtaining lower TF derived redundancy than miRNA derived redundancy, it is concluded that the different edge numbers of TFs and miRNAs in a cFFC do not create bias in the results.

4.3 Analysis of the Main Randomization Results Based on FANTOM Data

4.3.1 Q-Q Plots and Normality Test

The quantiles of the data are plotted against the quantiles of the normal distribution in order to decide whether the underlying distribution of the datasets fit to a normal distribution. According to the graphs, all datasets spread through the $y = x$ axis, which means the compared distributions are almost identical with normal distribution, although there exist some weak deviations in the tails. The Q-Q plot of number of cFFCs obtained by randomizing only the TF connection matrix of FANTOM data is given in Figure 4.4 as an example of the common pattern that is observed in the Q-Q plots of all other features, which are provided in Appendix A.3.

Besides the visualization method that shows the deviations of the experimental data from the normal distribution, the Shapiro-Wilk test which is a parametric approach is used to test whether the deviations in the tails that are observed in the Q-Q plots cause non-normality. The results of the Shapiro-Wilk test are given in Table 4.18.

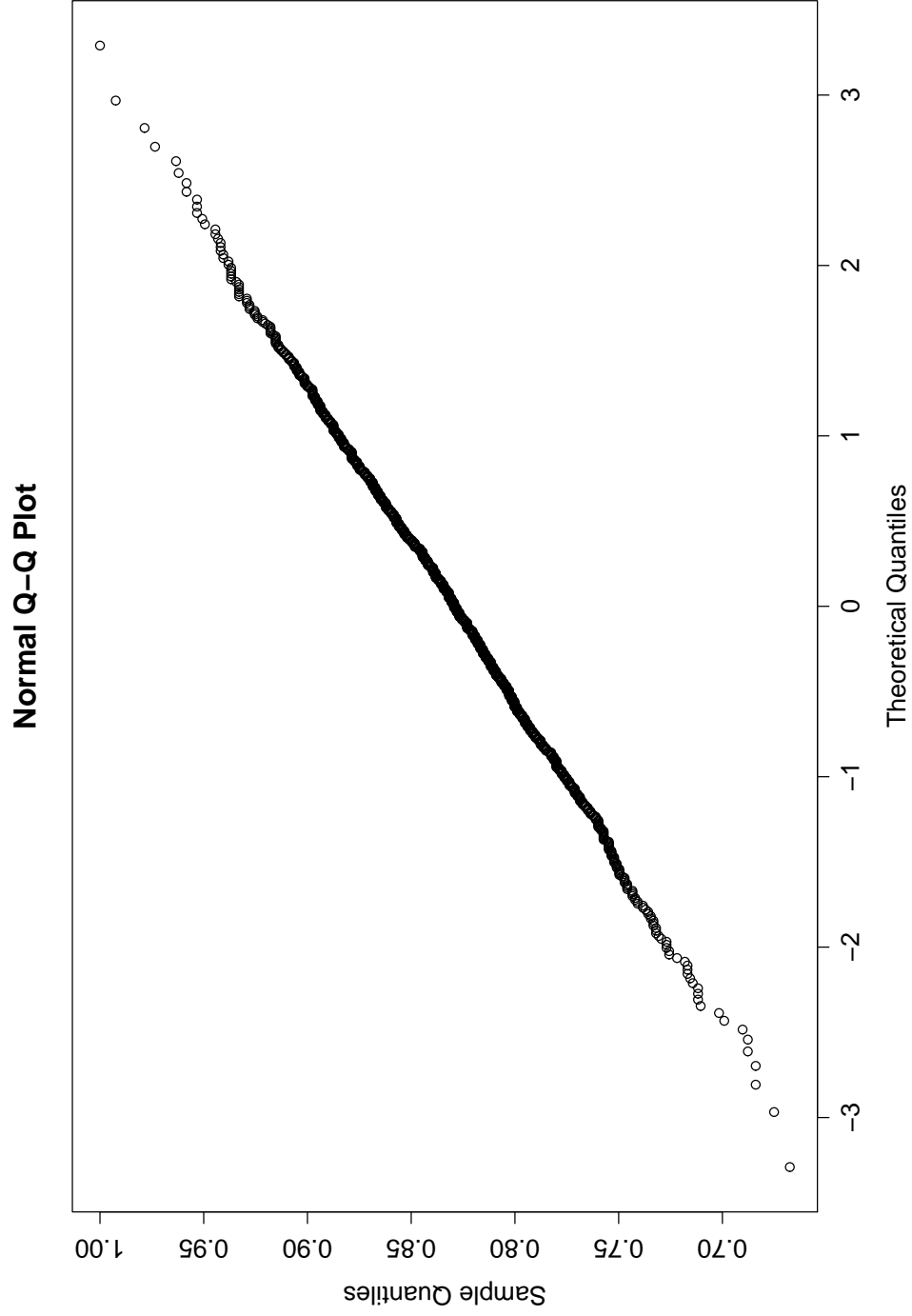


Figure 4.4: QQ plot of normality of number of cFFCs after TF connection randomization - FANTOM Data

Table 4.13: Shapiro-Wilk Test Results - FANTOM Data Weight 1.5

	mir connection randomization		tf connection randomization		tf-mir both connection randomization	
	W	p-value	W	p-value	W	p-value
No. of cFFCs	0.9985	0.5552	0.9989	0.8392	0.9972	0.08425
No. of cFFC-targeted genes	0.9972	0.08628	0.9977	0.1816	0.9983	0.4073
cFFC redundancy	0.9969	0.0453	0.9973	0.09985	0.9976	0.1578
miRNA derived redundancy	0.9976	0.1553	0.998	0.2726	0.9981	0.3499
TF derived redundancy	0.9984	0.5053	0.9984	0.4726	0.997	0.05967

According to the Shapiro-Wilk test results, all p-values are greater than 0.01 which indicate that the datasets are distributed normally, despite of the fact that there exist some weak deviations in the tails, and this enables the use of the Z-test and the two-sample t-test for further analyses.

4.3.2 Comparison of the Simulated Data via Real Gene Regulatory Network

4.3.2.1 Random Expectations

The random expectations and the corresponding standard deviations of the five variables obtained from three different connection randomizations are provided in Table 4.14 with the original counts that are observed from the real GRN. These numbers are used to calculate the Z-statistics and their p-values in order to make statistical inferences about the significance of the differences between the real GRN and the randomized networks in the next section.

As mentioned in Section 3.1.2, it should be noted that the number of cFFCs and the number of cFFC targeted genes obtained from original FANTOM network are far below the numbers observed from the IwamaEtAl network because of the weight threshold which is set to 1.5 in order to obtain significant regulator-target connections as suggested in the study conducted by Suzuki et. al. [43].

Table 4.14: Randomization Results of FANTOM Data

	Real GRN Observed	miRNA Connection Randomized Expected \pm Standard Deviation	TF Connection Randomized Expected \pm Standard Deviation	Both miRNA and TF Connection Randomized Expected \pm Standard Deviation
Number of cFFCs	701	682.408 \pm 32.11362	656.642 \pm 42.3214	680.747 \pm 40.58591
Number of genes targeted by cFFCs	221	309.832 \pm 12.51121	215.26 \pm 10.17536	313.801 \pm 13.4616
cFFCs redundancy	3.17	2.203255 \pm 0.07821949	3.051475 \pm 0.1558457	2.169462 \pm 0.09210835
miRNA derived cFFC redundancy	2.190625	1.308385 \pm 0.02740297	2.118064 \pm 0.08798926	1.263141 \pm 0.02941212
TF derived cFFC redundancy	1.178151	1.176775 \pm 0.02399627	1.167871 \pm 0.02727139	1.178153 \pm 0.0256823

4.3.2.2 Comparison of Random Expectations with Real GRN via Z-test

The hypotheses that are mentioned in Section 4.1.2.2 and Section 4.1.2.3 are tested using the initial numbers of the real FANTOM GRN with its random expectations via Z-test, for which the Z-statistics and the corresponding p-values are given in Table 4.15.

Table 4.15: Comparison of Z and P Values - FANTOM Data

	mir randomization		tf randomization		tf-mir randomization	
	Z	p-value	Z	p-value	Z	p-value
Number of cFFCs	0.5789443	0.2813134	1.048122	0.1472912	0.4990155	0.3088842
Number of cFFC-targeted genes	-7.100194	6.229089e-13	0.5641078	0.2863404	-6.893758	2.716867e-12
cFFC redundancy	12.38426	.589992e-35	0.7730124	0.2197575	10.88374	6.889215e-28
miRNA derived cFFC redundancy	32.19506	1.03463e-227	0.8246546	0.2047839	31.53408	1.482003e-218
TF derived cFFC redundancy	0.05735273	0.4771321	0.3769724	0.3530971	-5.079743e-05	0.4999797

According to Table 4.15, it is clearly seen that the interpretations that are made based on the Z-test results of the FANTOM data are consistent with that of obtained from Iwama et. al. data.

The difference between the number of cFFCs of the real GRN and the randomized networks is not statistically significant since the p-values are bigger than 0.01 for all three different connection randomization implying that the number of cFFCs is neither over-represented nor under-represented in the real GRN when compared to the randomized networks.

While the difference between the real GRN and the randomized networks in terms of the number of cFFC targeted genes is insignificant in the TF connection only randomization, it is under-represented in the real GRN when compared to the miRNA connection randomization and both the TF-miRNA connection randomization.

The cFFC redundancy is overly represented in the real GRN when compared to the randomized networks of miRNA connection only randomization and both TF-miRNA connection randomization. As expected, the cFFC redundancy obtained after the TF connection randomization shows no deviation from that of real GRN. The comparison in terms of the miRNA derived cFFC redundancy has similar behaviour as the cFFC redundancy that it is represented more highly in the real GRN than random networks of the miRNA connection only randomization and both the TF-miRNA connection randomization indicating that the redundancy adding role is attributable to the miRNAs and the over representation of the cFFC redundancy is not just a consequence of the reduction in the number of the cFFC targeted genes. This is because of the fact that the TF derived redundancy shows no deviation from the random expectations of three connection randomization procedure. As seen in the previous results, the TF network preserves its stable nature and remains unchanged to the alterations caused by miRNAs.

4.3.2.3 Comparison of Redundancies via the t-test

Since the homogeneity of variances assumption does not hold according to the variance ratio F-test of which the test statistics and the corresponding p-values are given in Table 4.16, the Welch’s two-sample t-test is conducted to see whether there exists a significant difference between the population means of the miRNA derived redundancy and the TF derived redundancy for both three types of connection randomization.

Table 4.16: Results of Variance Ratio F-test

	miRNA Connection Randomization		TF Connection Randomization		Both TF-miRNA Connection Randomization	
	F-statistic	p-value	F-statistic	p-value	F-statistic	p-value
miRNA Derived cFFC Redundancy	1.3041	2.805e-05	10.4099	< 2.2e - 16	1.3115	1.882e-05
TF Derived cFFC Redundancy						

The t-test results which are given in Table 4.17 indicate that the population means of the miRNA derived cFFC redundancy and the TF derived cFFC redundancy are significantly different. Moreover, the mean of the miRNA derived cFFC redundancy exceeds the mean of the TF derived cFFC redundancy which supports the decision that the increase in the cFFC redundancy is not manipulated by the decrease in the cFFC targeted genes.

Table 4.17: Results of Welch’s t-test

	miRNA Connection Randomization		TF Connection Randomization		Both TF-miRNA Connection Randomization	
	t-statistic	p-value	t-statistic	p-value	t-statistic	p-value
miRNA Derived cFFC Redundancy	114.26	< 2.2e - 16	326.1855	< 2.2e - 16	68.8294	< 2.2e - 16
TF Derived cFFC Redundancy						

Although the conclusions that are made based on the FANTOM data are compatible with the Iwama et. al. results heretofore, to see whether the imbalance between the degrees of miRNAs and TFs in an unitary cFFC reveals different behaviours, partial randomization procedure is implemented using the FANTOM network and the results are provided in the following sections.

4.4 Analysis of Partial Randomization Results Based on FANTOM Data

4.4.1 QQ Plots and Normality Tests

Since parametric tests require normality of the populations from which the samples are drawn, the datasets are examined visually using Q-Q plots to see if there exist departures from normality. The Q-Q plots of FANTOM datasets based on partial randomization procedure, which are provided in Appendix A.4, point out that the deviations from the straight line in the tails may cause a departure from the normality.

To obtain more certain results, the Shapiro-Wilk test is conducted and the results are given in Table 4.18.

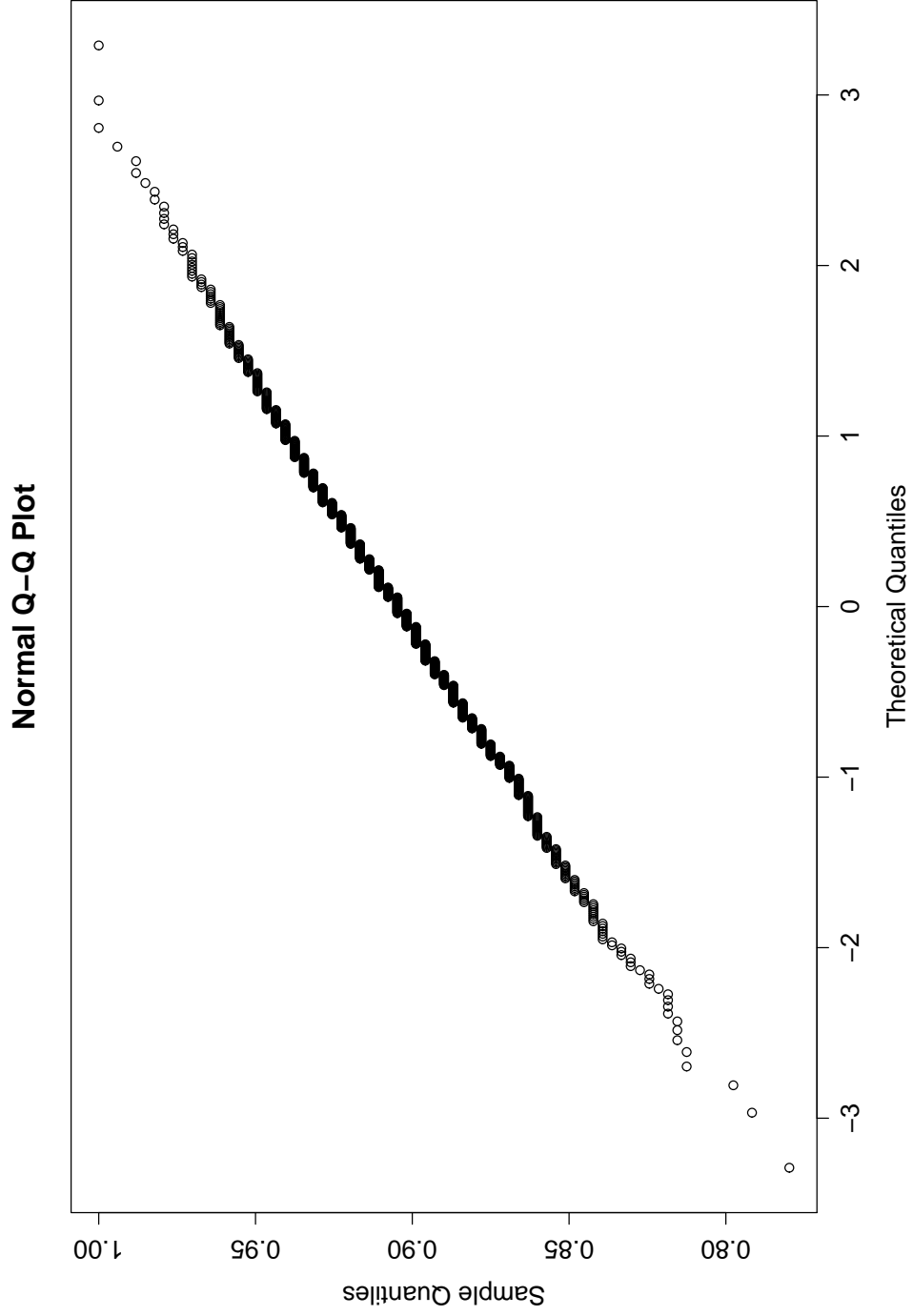


Figure 4.5: QQ plot of normality of the number of cFFC targeted genes after only miRNA-TF connection randomization - FANTOM Data

Table 4.18: Shapiro-Wilk Test Results of Partial Randomization - FANTOM Data

	mir-gene connection randomization		mir-tf connection randomization	
	W	p-value	W	p-value
No. of cFFCs	0.9981	0.3276	0.998	0.2897
No. of cFFC-targeted genes	0.9981	0.3098	0.9973	0.09895
cFFC redundancy	0.9973	0.08772	0.9977	0.1716
miRNA derived redundancy	0.9986	0.6112	0.9863	4.731e-08
TF derived redundancy	0.9982	0.3898	0.9974	0.1114

According to the Shapiro-Wilk test results that are given in Table 4.18, the p-values are greater than 0.01 except the one that belongs to the miRNA derived cFFC redundancy obtained from miRNA connection randomization with fixed miRNA-non-TF gene edges. Since the normality assumption must be satisfied to be able to conduct parametric tests such as the Z-test and the two-sample t-test, the dataset is searched for outliers that may cause a failing Shapiro-Wilk result. It is observed that after the removal of the outlier that corresponds to the maximum value of the dataset and identified using the "outlier" function that is covered in "outliers" package of R, the miRNA derived cFFC redundancy dataset shows no significant deviation from normality.

The test results after the removal of the outlier of the miRNA derived cFFC redundancy and its corresponding entity in TF derived cFFC redundancy are provided in Table 4.19.

Table 4.19: Shapiro-Wilk Test Results after Outlier Removal - FANTOM Data

	mir-tf connection randomization	
	W	p-value
miRNA derived redundancy	0.9986	0.6353
TF derived redundancy	0.9982	0.3975

4.4.2 Comparison of the Simulated Data with the Real GRN

4.4.2.1 Random Expectations

The initial numbers which are retrieved from the original FANTOM network and the random expectations that are obtained from one edge randomization of the FANTOM data are presented in Table 4.20 of which the second column contains the results obtained by randomizing the only miRNA - non-TF gene edges while third column involves the results of only miRNA - TF edge randomizations. Both partial randomization procedures are applied to only miRNA connection dataset, and the resultants are compared using the Z-test and the t-test in the following sections.

Table 4.20: Results of Partial Randomization of FANTOM miRNA Connection Data

	Real GRNs Observed	Randomization with Fixed miRNA-TF Edges Expected \pm Standard Deviation	Randomization with Fixed miRNA-Gene Edges Expected \pm Standard Deviation
Number of cFFCs	701	704.516 \pm 21.89075	682.483 \pm 24.35041
Number of genes targeted by cFFCs	221	302.187 \pm 10.02899	303.902 \pm 11.63468
cFFCs redundancy	3.17	2.332553 \pm 0.06850075	2.247277 \pm 0.07721423
miRNA derived cFFCs redundancy	2.190625	1.386886 \pm 0.02500251	1.386803 \pm 0.02487623
TF derived cFFCs redundancy	1.178151	1.172991 \pm 0.016067	1.173009 \pm 0.01606435

4.4.2.2 Comparison of Random expectations with Real GRN via Z-test

The real GRN is compared with random expectations via the Z-test of which the results are provided in the Table 4.21. As observed before, the test results revealed that the number of cFFC targeted genes are below the random expectations while the cFFC redundancy is more abundant in the real GRN. Besides, since the miRNA derived cFFC redundancy is over-represented in the real GRN, it is concluded that the cFFC redundancy is not a consequence of under representation of the cFFC targeted genes.

Despite of the fact that the conclusions of the Z-test are the same for two different datasets in terms of number of the cFFC targeted genes, the cFFC redundancy, the miRNA derived cFFC redundancy and the TF derived cFFC redundancy, there exists one important distinction between the results of partial randomization conducted using the Iwama et. al. network and the FANTOM network: The number of cFFCs shows no significant deviations from the random expectations of FANTOM data. This differentiation may be caused by the difference in the cell lines from the two networks which are built from. In other words, it is proposed that since FANTOM data are obtained from a myeloid leukemia cell line, there is the likely possibility that the regulator-target information and accordingly the orientation of the network are differentiated from the Iwama et al. network which contains targeting information in a healthy cell. Thus, while miRNAs are likely to regulate their targeted gene through TFs indirectly in a normal cell, they show different behaviour in the cancerous cell and may start regulating the target gene directly.

Table 4.21: Comparison of Z and P Values of Partial Randomization based on FANTOM miRNA Connection Data

	Randomization with Fixed miRNA-TF Edges		Randomization with Fixed miRNA-Gene Edges	
	Z	p-value	Z	p-value
Number of cFFCs	-0.1606158	0.436198	0.7604388	0.2234962
Number of cFFC-targeted genes	-8.095233	2.85774e-16	-7.125422	5.188113e-13
cFFC redundancy	7.410697	6.281885e-14	11.97537	2.391638e-33
miRNA derived cFFC redundancy	12.08813	6.101705e-34	32.31284	2.308682e-229
TF derived cFFC redundancy	-0.973776	0.1650839	0.3200963	0.3744476

Table 4.22: Difference between Iwama et. al. and FANTOM Resultings

	miRNA-TF edges randomization		miRNA-gene edges randomization	
	Z-statistic	p-value	Z-statistic	p-value
Iwama et. al. number of cFFCs	2.260347	0.01189985	-2.781475	0.002705627
FANTOM number of cFFCs	0.7604388	0.2234962	-0.1606158	0.436198
Iwama et. al. TF derived redundancy	0.8112788	0.2086028	-2.586778	0.004843903
FANTOM TF derived redundancy	0.3200963	0.3744476	-0.973776	0.1650839

4.4.2.3 Comparison of Redundancies via t-test

Because of the fact that the homogeneity of the variances assumption is not held according to the variance ratio F-test whose results are provided in Table 4.23, the Welch t-test is applied in order see whether the population means of the miRNA derived cFFC redundancy and the TF derived cFFC redundancy are different. The Welch's two sample t-test results which are represented in Table 4.24 point out that there exists a significant difference between the compared datasets which supports the decision that the reduction in the repository of the cFFC targeted genes does not constitute the increase of the cFFC redundancy. Hence, adding redundancy role is a distinct property of miRNAs.

Table 4.23: Results of Variance Ratio F-test - Partial Randomization of FANTOM Data

	Randomization with Fixed miRNA-TF Edges		Randomization with Fixed miRNA-Gene Edges	
	F-statistic	p-value	F-statistic	p-value
miRNA Derived cFFC Redundancy	2.4216	$< 2.2e - 16$	2.398	$< 2.2e - 16$
TF Derived cFFC Redundancy				

Table 4.24: Results of two sample t-test - Partial Randomization of FANTOM Data

	Randomization with Fixed miRNA-TF Edges		Randomization with Fixed miRNA-Gene Edges	
	t-statistic	p-value	t-statistic	p-value
miRNA Derived cFFC Redundancy	227.5907	$< 2.2e - 16$	228.1949	$< 2.2e - 16$
TF Derived cFFC Redundancy				

CHAPTER 5

CONCLUSION

5.1 Summary

The co-regulation between miRNAs and TFs in gene regulatory networks is analysed in terms of cFFCs and cFFC dependent terms, namely; cFFC targeted genes, cFFC redundancy, miRNA derived cFFC redundancy and TF derived cFFC redundancy. This was done in order to uncover the difference between TFs and miRNAs in terms of their contribution and the effect of their evolutionary distinctions to the regulatory networks. cFFCs are significant network motifs which comprise a miRNA which is a master regulator, an intermediary TF, and a gene as shared target. A gene within the network is defined as a cFFC targeted gene if it is regulated by at least one cFFC. The cFFC redundancy is termed as the average number of cFFCs that involves in targeting one gene. In order to see the effects of miRNAs and TFs in cFFC redundancy, it is divided into two factors which are the miRNA derived cFFC redundancy and the TF derived cFFC redundancy. The miRNA derived cFFC redundancy are calculated as a ratio of the total number of cFFCs to the total number of the cFFC participant TF-gene edges. Similarly, the TF derived cFFC redundancy is obtained by dividing the total number of cFFCs to the total number of the miRNA-gene edges that are involved in cFFCs.

The analyses are conducted based on two different GRNs one of which was obtained from human myeloid leukemia cell line experimentally and provided in the FANTOM4 EdgeExpress database, while the other one contains the human mouse conserved targeting information of a healthy cell which is collected computationally by Iwama et. al. using miRBase and TRANSFAC databases. These two datasets differ from each other in terms of the number of edges and nodes that are involved in the network. FANTOM4 also provides the weight parameter that indicates the significance of the regulator-target prediction which is set to 1.5 to obtain more significant connections with minimum number of false positives [43]. The FANTOM4 dataset is used after the identification of human-mouse conserved regulatory connections of the network which is retrieved using 1.5 as the weight threshold.

Two randomization procedures which generate 1,000 random networks, each from the real GRNs using 500,000 edge swapping steps for each in a degree preserving manner is implemented to the networks. The first one, which is referred to as the main the randomization procedure throughout the thesis contains three sub-procedures which are the miRNA connection only randomization, the TF connection only randomization

and both the miRNA-TF connection randomization. The second one, which is called as the partial randomization procedure, is conducted to reveal whether the imbalance between the edge numbers of TFs and miRNAs in unitary cFFCs affects the results that are obtained from main randomization procedure, and it is implemented on only miRNA connections. The partial randomization procedure contains two sub-processes which are miRNA - non-TF gene edges randomization and miRNA-TF edges randomization. The cFFC dependent terms are calculated from the randomly generated networks from both main randomization procedure and partial randomization procedure to be compared with the corresponding values that are obtained from original networks using the Z-test. Moreover, the miRNA derived cFFC redundancy and the TF derived cFFC redundancy that are generated from the same sub-process are also compared with each other by using the t-test in order to reveal the difference of the TFs and miRNAs in contribution to the cFFC redundancy.

The study conducted by Iwama et al.(2010) is reproduced by implementing the randomization procedures to the dataset which is collected for the same study. It should be mentioned that, the total number of cFFCs obtained from original network during the reproduction of the results of the study conducted by Iwama et al. is not the same with the one that is published by them. In the original paper, the total number of cFFCs is reported as 44,373, while in this study, it is identified as 43,481. The difference between these numbers are caused by the elimination of the loops which consist of one miRNA and one TF in which the miRNA regulates the autoregulatory TF that clearly does not function as FFLs.

The reproduction of the results of the study enables to verify the randomization code, and build a framework that is applicable to different GRNs. Moreover, the implementation of this framework to the FANTOM4 dataset not only provides justification to the results of Iwama et al. obtained by the main randomization procedure, but also reveals the different behaviour of miRNAs in the networks of cancerous cells by the partial randomization procedure.

The results are discussed in detail in the following section.

5.2 Discussion

Network analyses which are conducted with two different datasets revealed that miRNAs change the network conformation by forming comprehensive modifications in the GRNs, while the TF networks preserve their stability to the changes that are caused by miRNAs. The differences between TFs and miRNAs in terms of their contribution to GRNs are examined in four main aspects. Firstly, it is observed that the changes in the miRNA network revealed an increase in the cFFC redundancy through the cFFC formation when compared with the random networks. Secondly, the alterations in the miRNA network also uncovered a sharp decrease in the number of the cFFC targeted genes from the random expectations, that results in the de-escalation of the target gene repertoire. Thirdly, by examining the cFFC redundancy in terms of two factors which are the miRNA derived cFFC redundancy and the TF derived cFFC redundancy, it is identified that the increase in the cFFC redundancy is caused by the additional loop formation of miRNAs around the same TF-gene edge. Hence, it is

concluded that, adding redundancy role is a distinctive property of miRNAs. Lastly, the contrast between miRNA networks and the TFs networks based on their tendency to remain indifferent to the structural changes of the regulatory networks is exposed. As mentioned before, the TF network remains neutral to conformational changes in the network.

Partial randomization results are also uncovered that although miRNAs tend to involve in regulatory connections through the cFFC formation, they exhibit different behaviour in the network that is obtained from myeloid leukemia cell line and instead of regulating the targeted gene indirectly by forming cFFCs with TFs, they may start to regulate their targeted genes directly because of the elimination of the TF-target edge as a result of the cancer mechanism.

5.3 Future Work

Recently, FANTOM5 which includes the regulatory models of different cell types, cancer lines and tissues was released. Implementation of our framework to the updated target predictions which may obtained from FANTOM5, and comparing the regulatory networks of different different cell lines may provide a deeper insight in the understanding of the mechanism of miRNAs in cancer, cell differentiation, or in developmental stages.

Additionally, the procedure and steps of analysis described in this study can be generalized for other types of motif structures to investigate their significances and effects of their components in biological framework.

The results based on normality assumption can be also extended for student-t and long-tailed symmetric distribution family in order to get more robust conclusions. The reason is that the Q-Q plots of datasets mostly indicate variations in the tail which is the indication of the long-tailed family.

REFERENCES

- [1] U. Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.
- [2] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.
- [3] P. Anderson and N. Kedersha. Rna granules: post-transcriptional and epigenetic modulators of gene expression. *Nature reviews Molecular cell biology*, 10(6):430–436, 2009.
- [4] L. J. Bain and M. Engelhardt. *Introduction to probability and mathematical statistics*, volume 4. Duxbury Press Belmont, CA, 1992.
- [5] D. P. Bartel. Micrnas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.
- [6] A. H. Brivanlou and J. E. Darnell. Signal transduction and the control of gene expression. *Science*, 295(5556):813–818, 2002.
- [7] L. Chen and J. Widom. Mechanism of transcriptional silencing in yeast. *Cell*, 120(1):37–48, 2005.
- [8] T. Coladarci, C. D. Cobb, E. W. Minium, and R. C. Clarke. *Fundamentals of statistical reasoning in education*. John Wiley & Sons, 2010.
- [9] F. Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.
- [10] F. H. Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.
- [11] V. de Lorenzo. From the selfish gene to selfish metabolism: revisiting the central dogma. *BioEssays*, 36(3):226–235, 2014.
- [12] J. D. Gibbons and S. Chakraborti. *Nonparametric statistical inference*. Springer, 2011.
- [13] S. Griffiths-Jones. The microrna registry. *Nucleic acids research*, 32(suppl 1):D109–D111, 2004.

- [14] M. B. Harris, J. Mostecky, and P. B. Rothman. Repression of an interleukin-4-responsive promoter requires cooperative bcl-6 function. *Journal of Biological Chemistry*, 280(13):13114–13121, 2005.
- [15] D. R. Helsel and R. M. Hirsch. *Statistical methods in water resources*, volume 49. Elsevier, 1992.
- [16] H. Iwama, K. Murao, H. Imachi, and T. Ishida. MicroRNA networks alter to conform to transcription factor networks adding redundancy and reducing the repertoire of target genes for coordinated regulation. *Molecular biology and evolution*, 28(1):639–646, 2011.
- [17] S. Kalir, S. Mangan, and U. Alon. A coherent feed-forward loop with a sum input function prolongs flagella expression in escherichia coli. *Molecular systems biology*, 1(1), 2005.
- [18] H. Kawaji, J. Severin, M. Lizio, A. R. Forrest, E. van Nimwegen, M. Rehli, K. Schroder, K. Irvine, H. Suzuki, P. Carninci, et al. Update of the fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic acids research*, page gkq1112, 2010.
- [19] H. Kawaji, J. Severin, M. Lizio, A. Waterhouse, S. Katayama, K. M. Irvine, D. A. Hume, A. R. Forrest, H. Suzuki, P. Carninci, et al. The fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome biology*, 10(4):R40, 2009.
- [20] E. V. Koonin. Does the central dogma still stand. *Biol Direct*, 7(1):27–27, 2012.
- [21] M. Kratz and S. I. Resnick. The qq-estimator and heavy tails. *Stochastic Models*, 12(4):699–724, 1996.
- [22] D. Latchman. *Gene regulation*. Taylor & Francis, 2012.
- [23] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene *lin-4* encodes small rnas with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, 1993.
- [24] B. Lemon and R. Tjian. Orchestrated response: a symphony of transcription factors for gene control. *Genes & development*, 14(20):2551–2569, 2000.
- [25] Y. T. Maeda and M. Sano. Regulatory dynamics of synthetic gene networks with positive feedback. *Journal of molecular biology*, 359(4):1107–1124, 2006.
- [26] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.
- [27] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

- [28] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
- [29] G. Orphanides, T. Lagrange, and D. Reinberg. The general transcription factors of rna polymerase ii. *Genes & development*, 10(21):2657–2683, 1996.
- [30] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna. *Nature*, 408(6808):86–89, 2000.
- [31] N. M. Razali and Y. B. Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.
- [32] J. Reece, L. Urry, M. Cain, S. Wasserman, and P. Minorsky. *Campbell Biology*. Pearson Benjamin Cummings, 2011.
- [33] R. G. Roeder. The role of general initiation factors in transcription by rna polymerase ii. *Trends in biochemical sciences*, 21(9):327–335, 1996.
- [34] M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the national academy of sciences*, 99(16):10555–10560, 2002.
- [35] N. Rosenfeld, M. B. Elowitz, and U. Alon. Negative autoregulation speeds the response times of transcription networks. *Journal of molecular biology*, 323(5):785–793, 2002.
- [36] S. M. Ross. *Introduction to probability and statistics for engineers and scientists*. Academic Press, 2009.
- [37] P. Royston. Remark as r94: A remark on algorithm as 181: The w-test for normality. *Applied Statistics*, pages 547–551, 1995.
- [38] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel. Global and local architecture of the mammalian microrna–transcription factor regulatory network. *PLoS computational biology*, 3(7):e131, 2007.
- [39] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, pages 591–611, 1965.
- [40] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

- [41] H. Siomi and M. C. Siomi. Posttranscriptional regulation of microRNA biogenesis in animals. *Molecular cell*, 38(3):323–332, 2010.
- [42] L. Srinivasan and M. L. Atchison. Yy1 dna binding and pcg recruitment requires ctbp. *Genes & development*, 18(21):2596–2601, 2004.
- [43] H. Suzuki, A. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M. J. de Hoon, et al. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature genetics*, 41(5):553–562, 2009.
- [44] B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, pages 350–362, 1938.
- [45] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *c. elegans*. *Cell*, 75(5):855–862, 1993.
- [46] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis for the analysis of data. *Biometrika*, 55(1):1–17, 1968.
- [47] J. Winter, S. Jung, S. Keller, R. I. Gregory, and S. Diederichs. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature cell biology*, 11(3):228–234, 2009.
- [48] Y. Zhou, J. Ferguson, J. T. Chang, and Y. Kluger. Inter-and intra-combinatorial regulation by transcription factors and microRNAs. *BMC genomics*, 8(1):396, 2007.

APPENDIX A

QQ PLOTS

A.1 QQ Plots of Main Randomization Procedure of Iwama et. al. Data

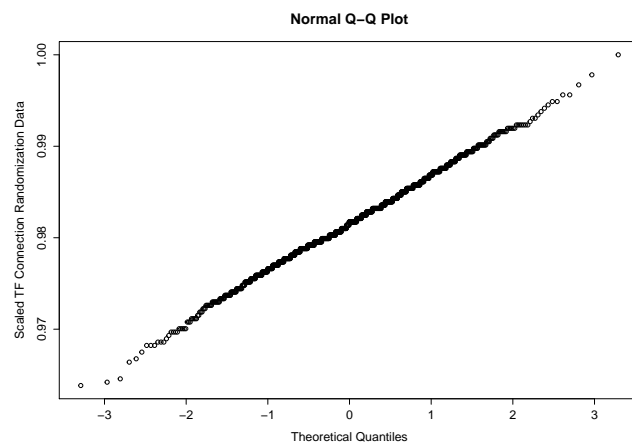


Figure A.1: QQ plot of normality of number of cFFC targeted genes after miRNA connection randomization - Iwama et. al. Data

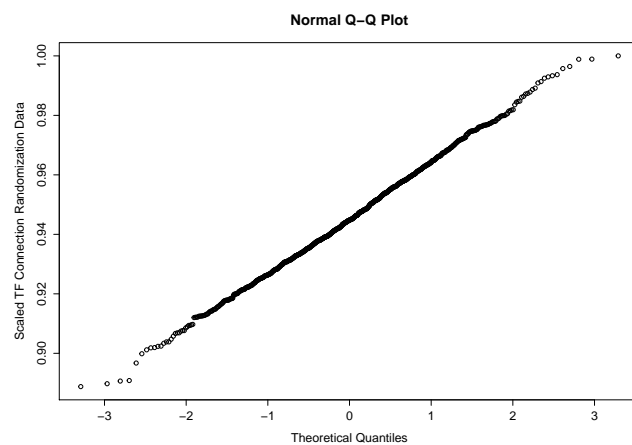


Figure A.2: QQ plot of normality of cFFC redundancies after miRNA connection randomization - Iwama et. al. Data

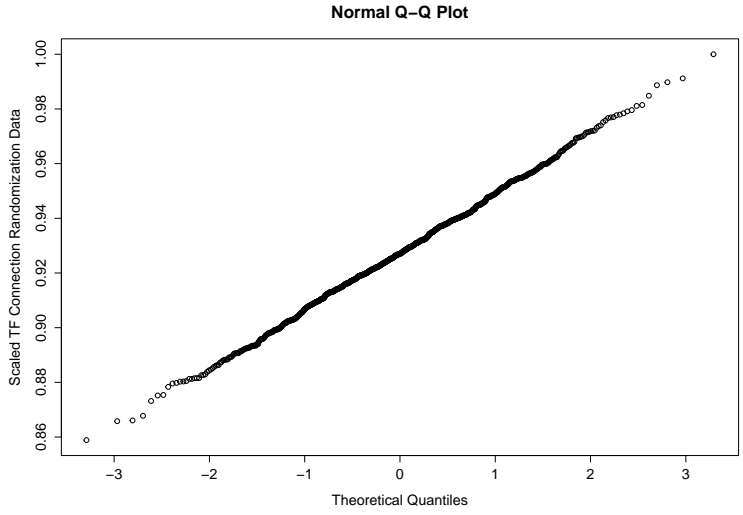


Figure A.3: QQ plot of normality of miRNA derived redundancy after miRNA connection randomization - Iwama et. al. Data

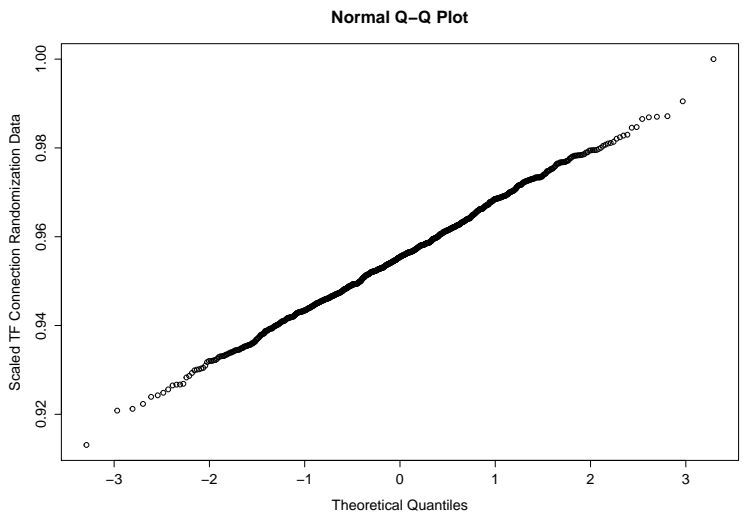


Figure A.4: QQ plot of normality of TF derived redundancy after miRNA connection randomization - Iwama et. al. Data

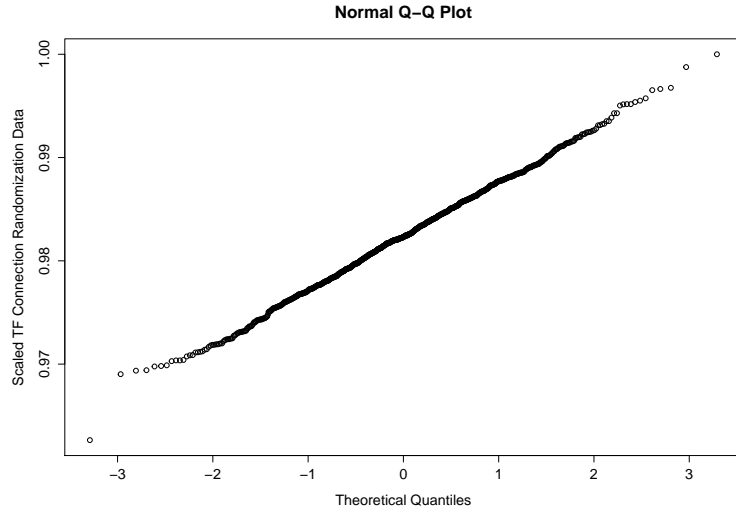


Figure A.5: QQ plot of normality of number of cFFCs after TF connection randomization - Iwama et. al. Data

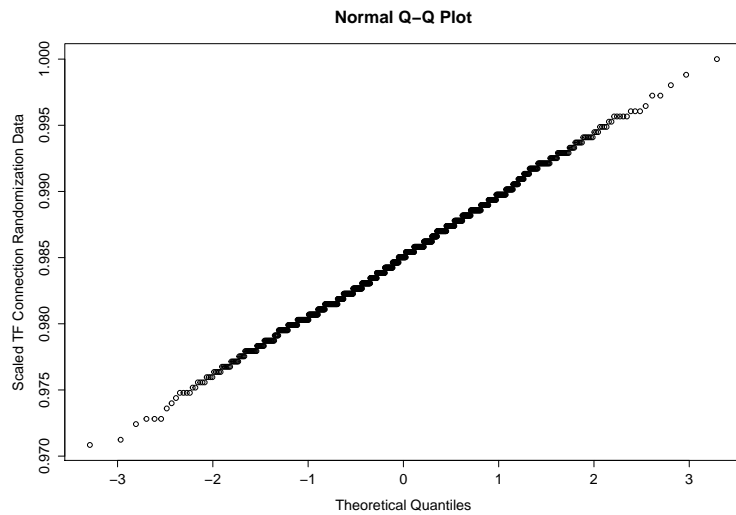


Figure A.6: QQ plot of normality of number of cFFC targeted genes after TF connection randomization - Iwama et. al. Data

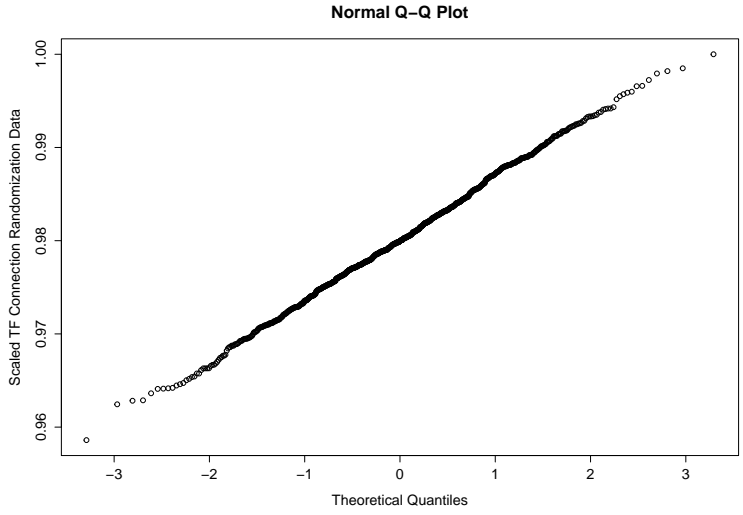


Figure A.7: QQ plot of normality of cFFC redundancies after TF connection randomization - Iwama et. al. Data

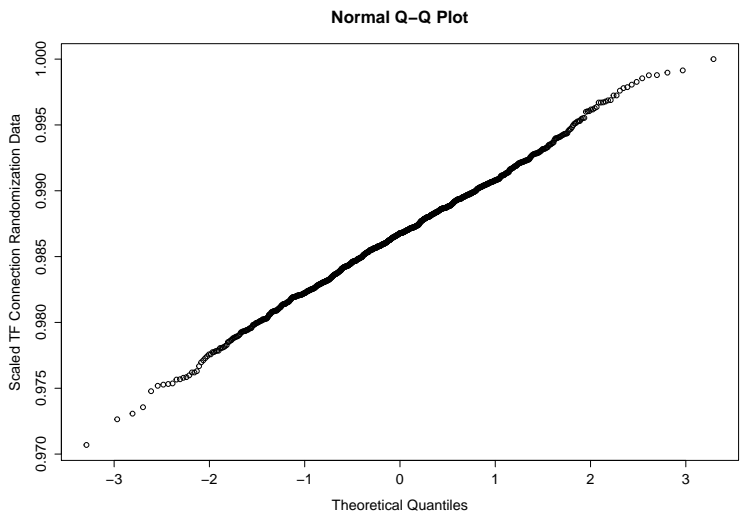


Figure A.8: QQ plot of normality of miRNA derived redundancy after TF connection randomization - Iwama et. al. Data

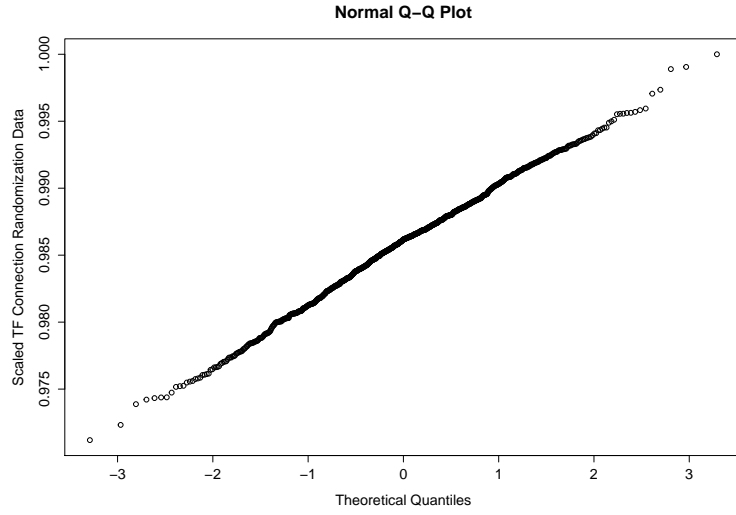


Figure A.9: QQ plot of normality of TF derived redundancy after TF connection randomization - Iwama et. al. Data

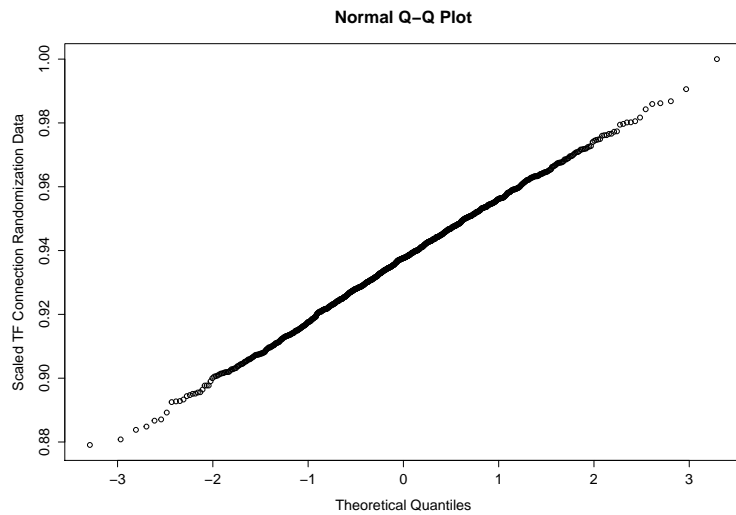


Figure A.10: QQ plot of normality of number of cFFCs after TF-miRNA both connection randomization - Iwama et. al. Data

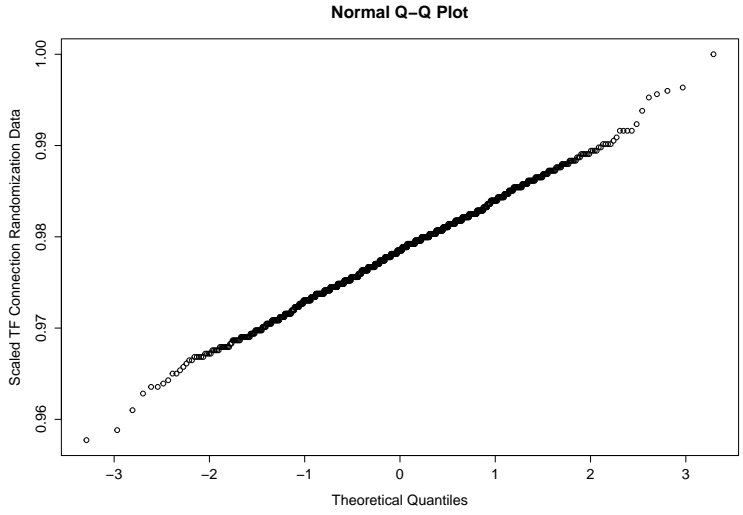


Figure A.11: QQ plot of normality of number of cFFC targeted genes after TF-miRNA both connection randomization - Iwama et. al. Data

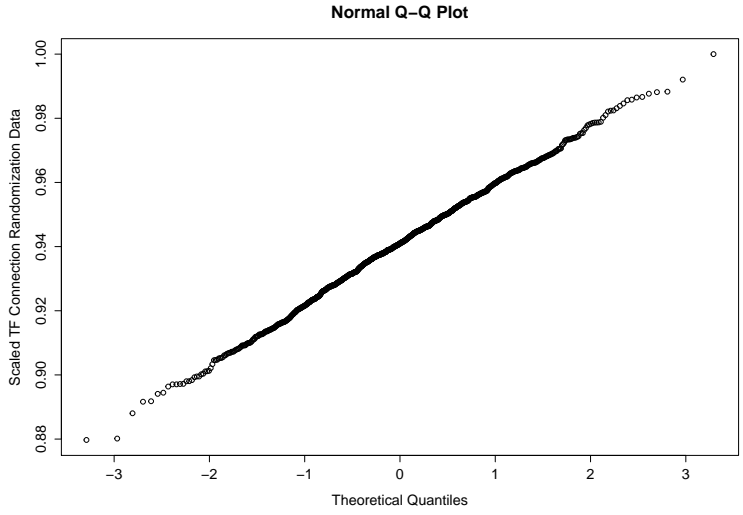


Figure A.12: QQ plot of normality of cFFC redundancy after TF-miRNA both connection randomization - Iwama et. al. Data

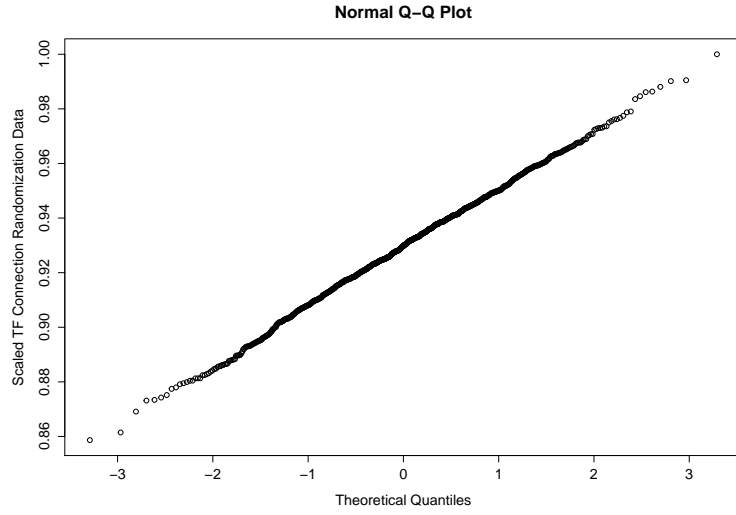


Figure A.13: QQ plot of normality of miRNA derived cFFC redundancy after TF-miRNA both connection randomization - Iwama et. al. Data

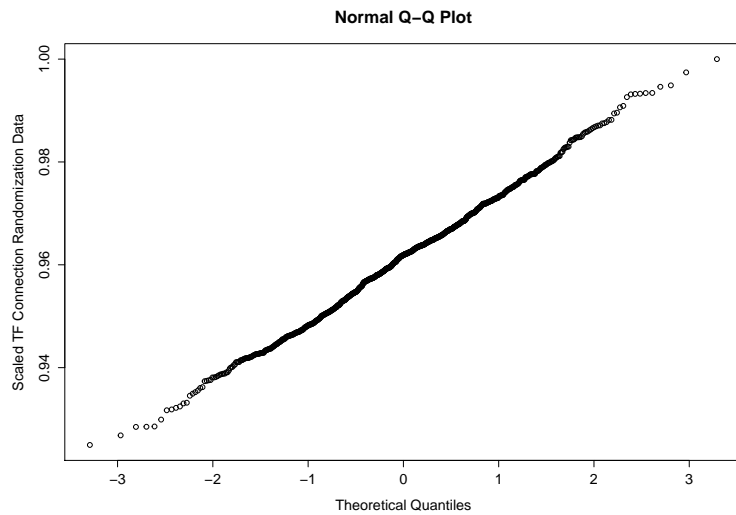


Figure A.14: QQ plot of normality of TF derived cFFC redundancy after TF-miRNA both connection randomization - Iwama et. al. Data

A.2 QQ plots of Partial Randomization Procedure of Iwama et. al. Data

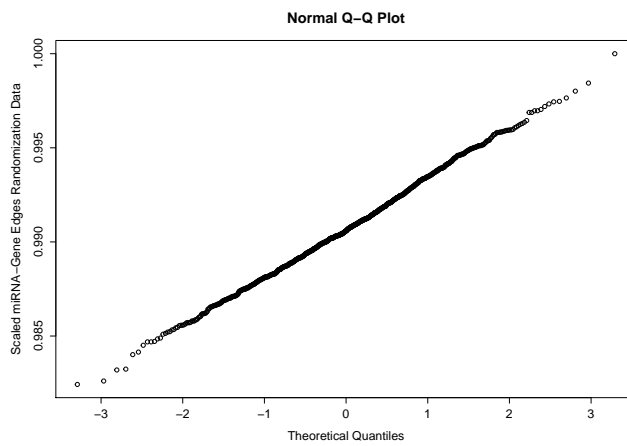


Figure A.15: QQ plot of normality of number of cFFCs after only miRNA-gene connection randomization - Iwama et. al. Data

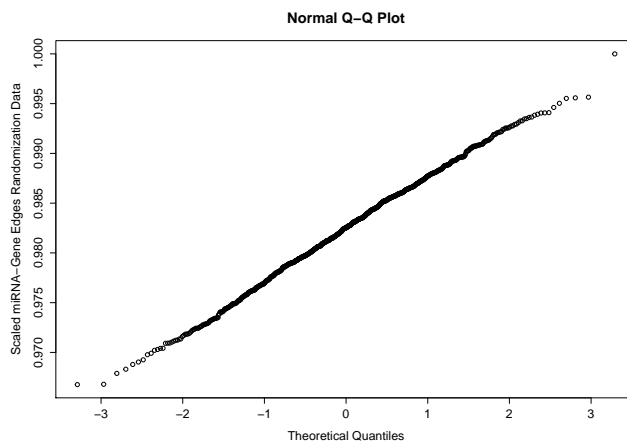


Figure A.16: QQ plot of normality of cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data

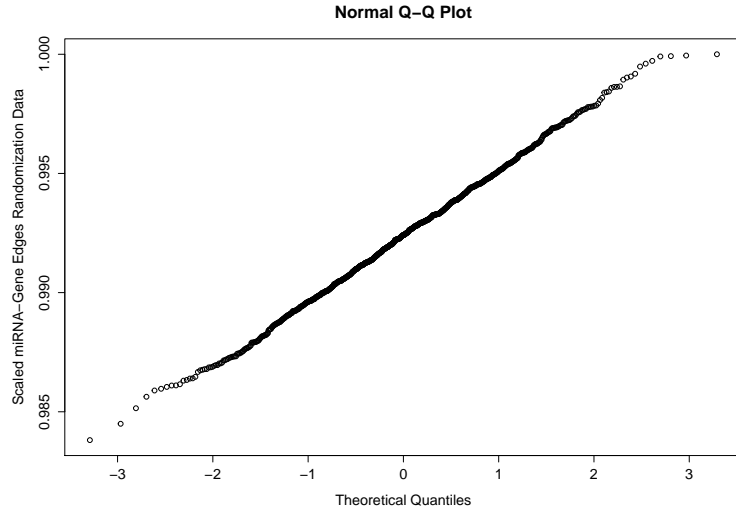


Figure A.17: QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data

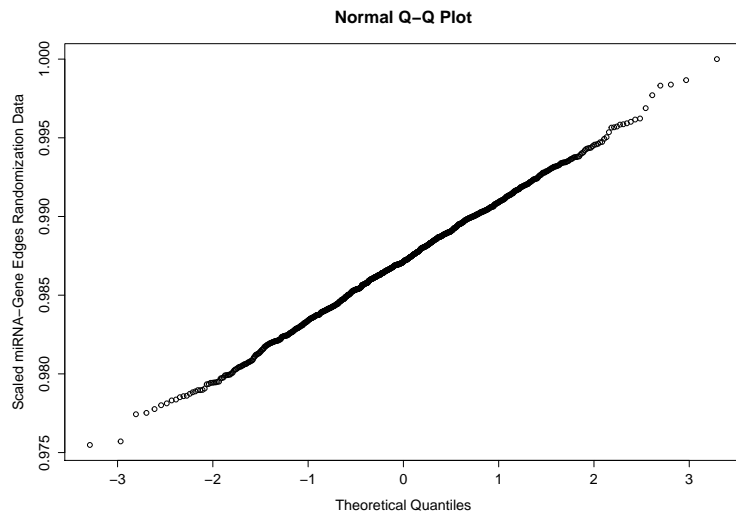


Figure A.18: QQ plot of normality of TF derived redundancy after only miRNA-gene connection randomization - Iwama et. al. Data

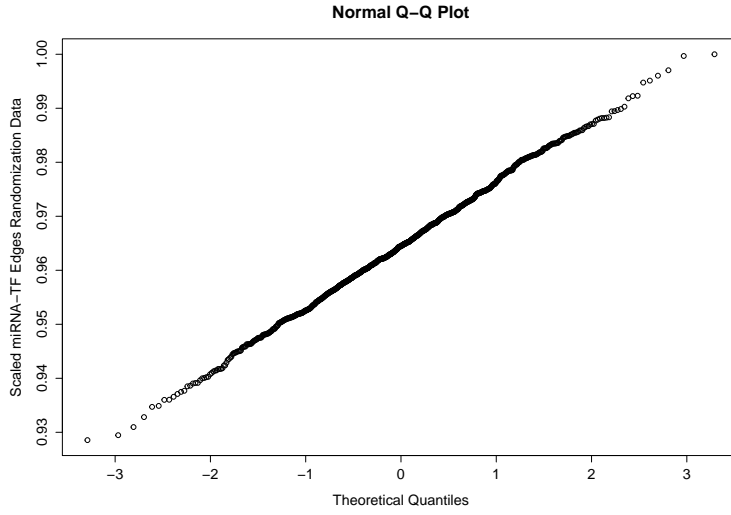


Figure A.19: QQ plot of normality of number of cFFCs after only miRNA-TF connection randomization - Iwama et. al. Data

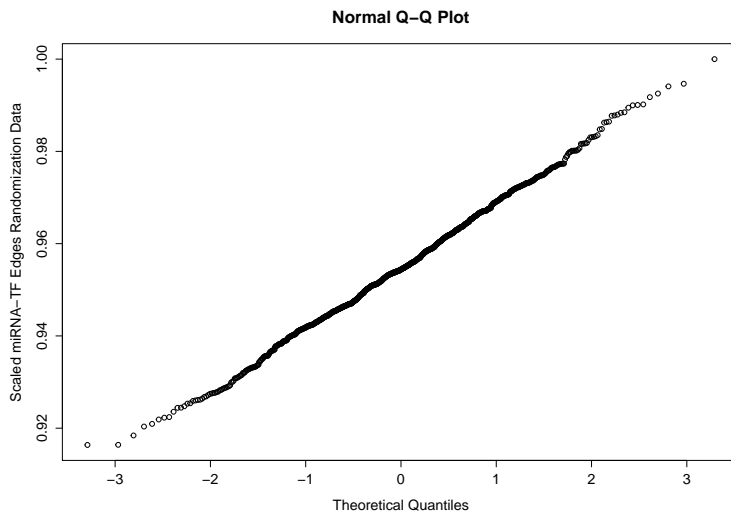


Figure A.20: QQ plot of normality of cFFC redundancy after only miRNA-TF connection randomization - Iwama et. al. Data

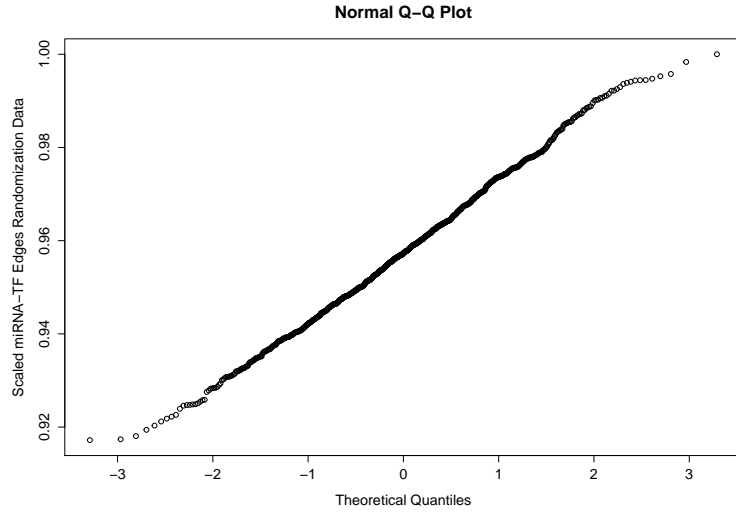


Figure A.21: QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data

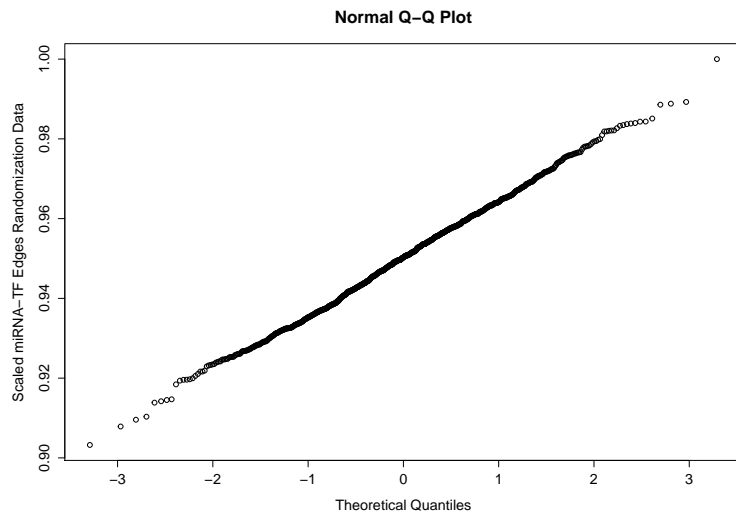


Figure A.22: QQ plot of normality of TF derived cFFC redundancy after only miRNA-gene connection randomization - Iwama et. al. Data

A.3 QQ Plots of Main Randomization Procedure of FANTOM Data

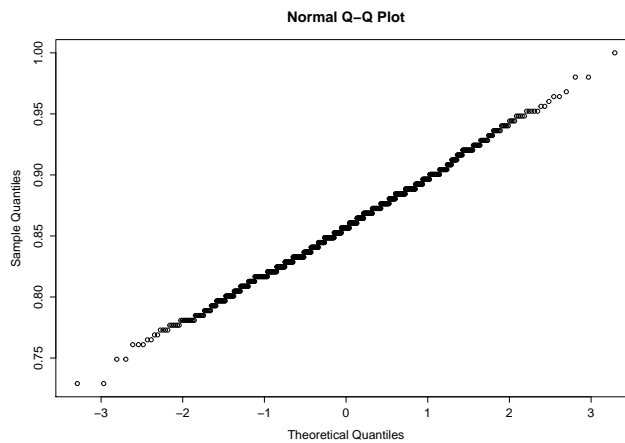


Figure A.23: QQ plot of normality of number of cFFC targeted genes after TF connection randomization - FANTOM Data

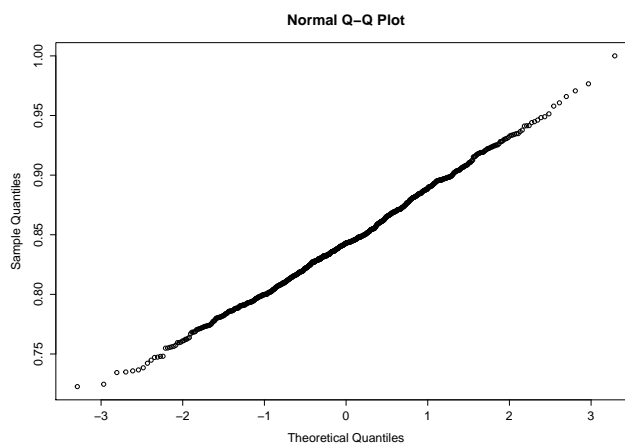


Figure A.24: QQ plot of normality of cFFC redundancy after TF connection randomization - FANTOM Data

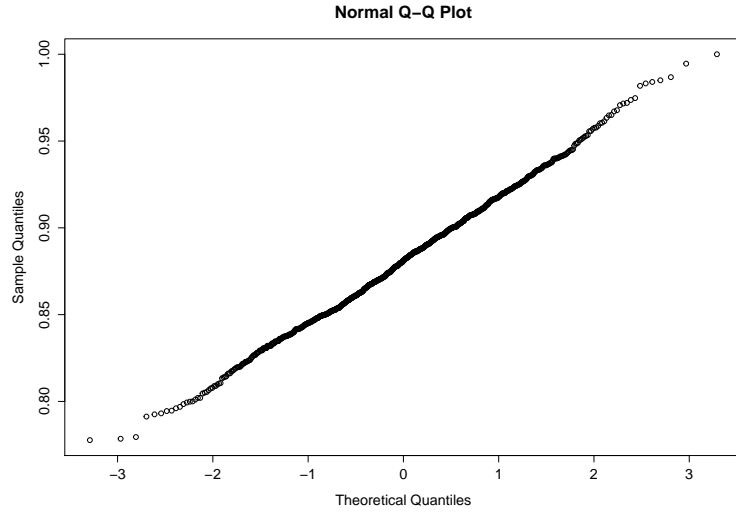


Figure A.25: QQ plot of normality of miRNA derived cFFC redundancy after TF connection randomization - FANTOM Data

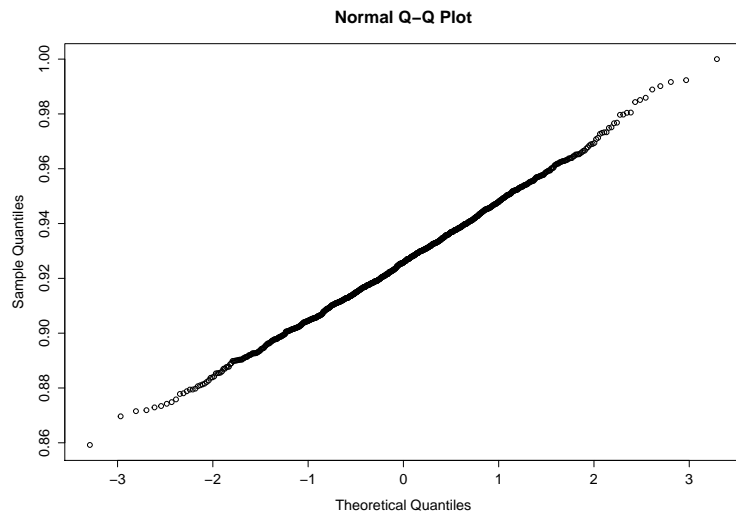


Figure A.26: QQ plot of normality of TF derived cFFC redundancy after TF connection randomization - FANTOM Data

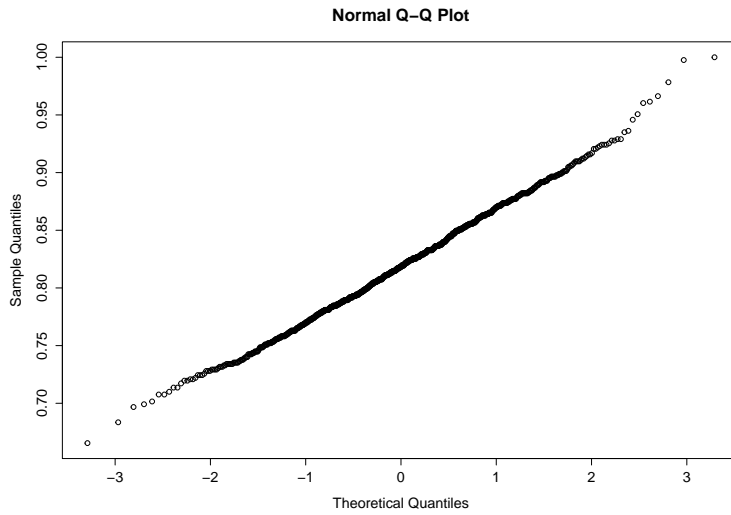


Figure A.27: QQ plot of normality of number of cFFCs after both TF-miRNA connection randomization - FANTOM Data

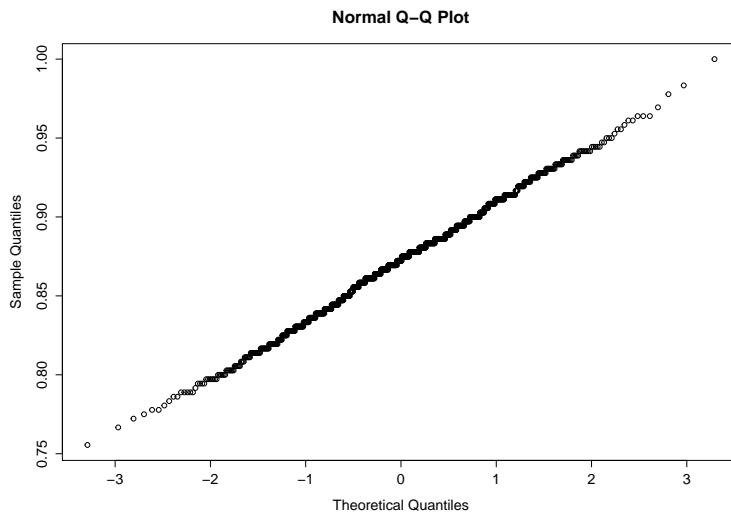


Figure A.28: QQ plot of normality of number of cFFC targeted genes after both TF-miRNA connection randomization - FANTOM Data

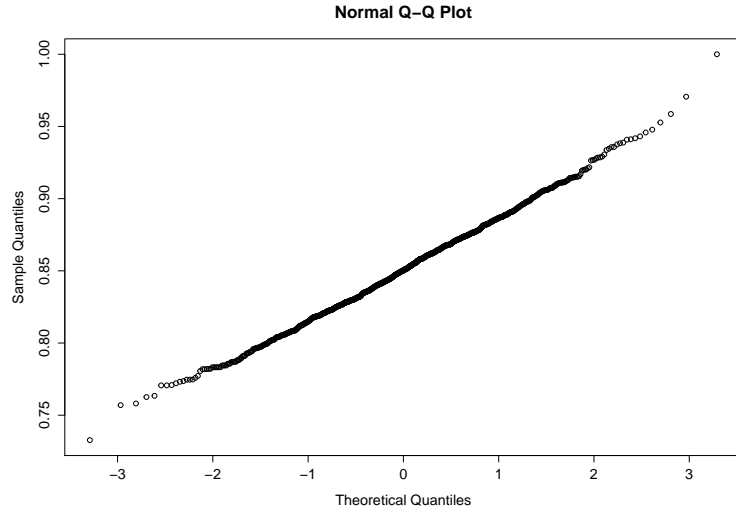


Figure A.29: QQ plot of normality of cFFC redundancy after both TF-miRNA connection randomization - FANTOM Data

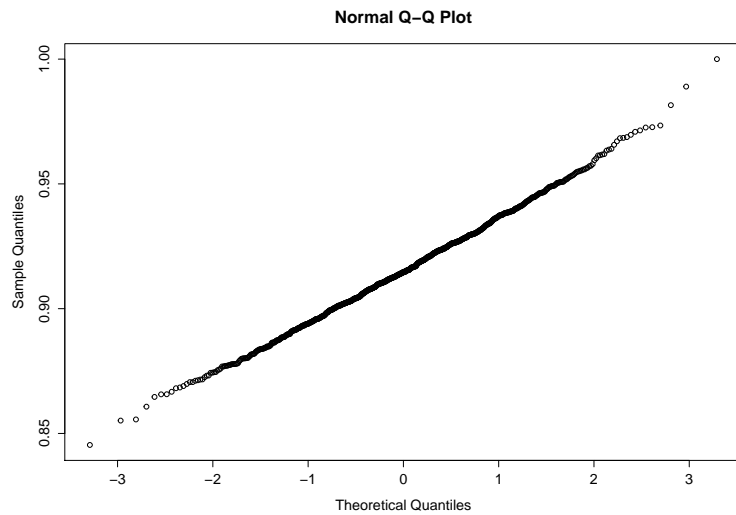


Figure A.30: QQ plot of normality of miRNA derived cFFC redundancy after both TF-miRNA connection randomization - FANTOM Data

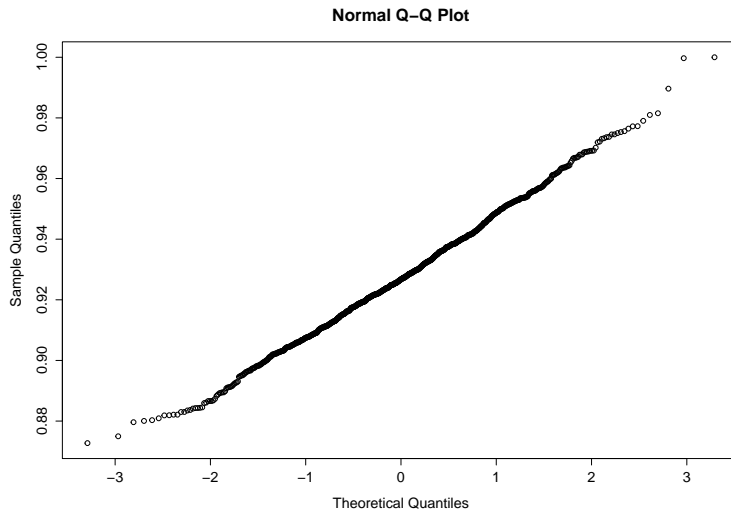


Figure A.31: QQ plot of normality of TF derived cFFC redundancy after both TF-miRNA connection randomization - FANTOM Data

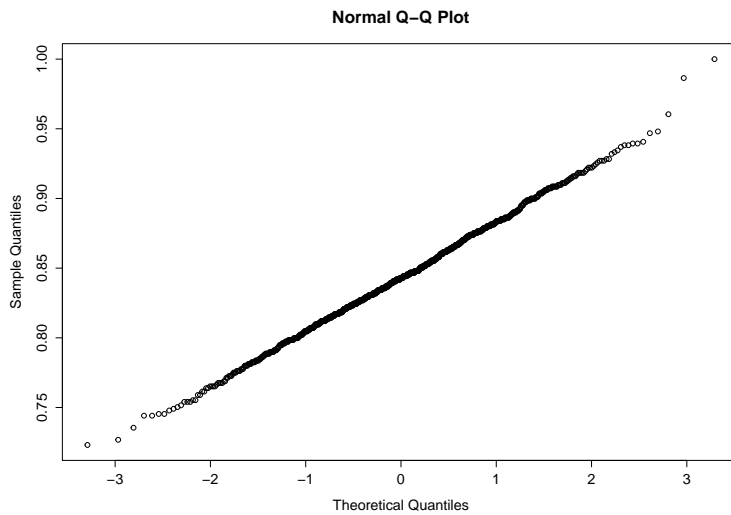


Figure A.32: QQ plot of normality of number of cFFCs after miRNA connection randomization - FANTOM Data

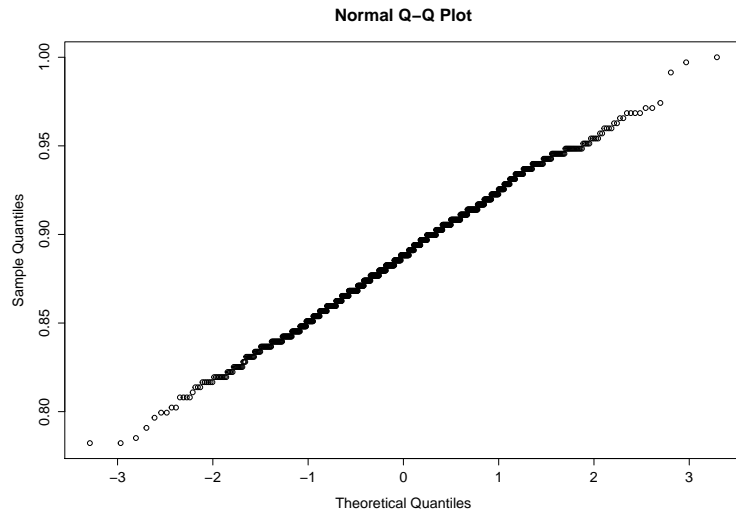


Figure A.33: QQ plot of normality of number of cFFC targeted genes after both miRNA connection randomization - FANTOM Data

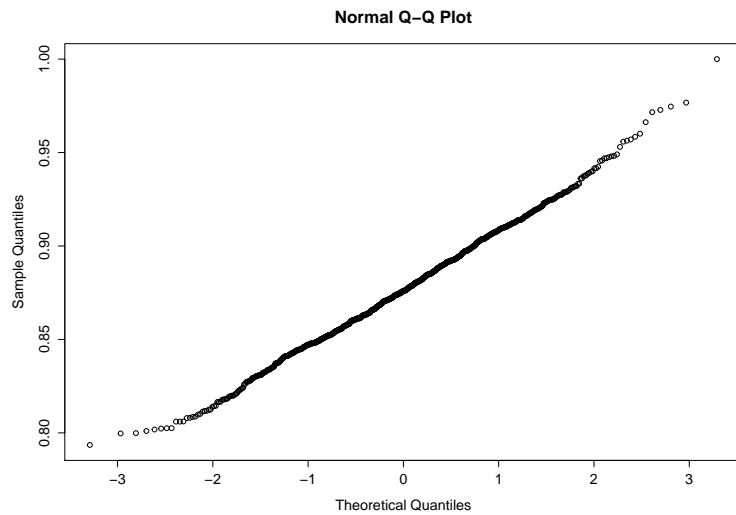


Figure A.34: QQ plot of normality of cFFC redundancy after both miRNA connection randomization - FANTOM Data

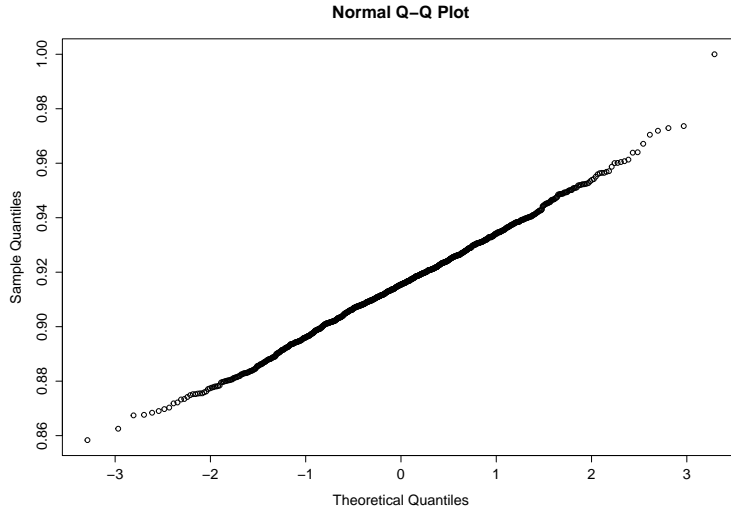


Figure A.35: QQ plot of normality of miRNA derived cFFC redundancy after both miRNA connection randomization - FANTOM Data

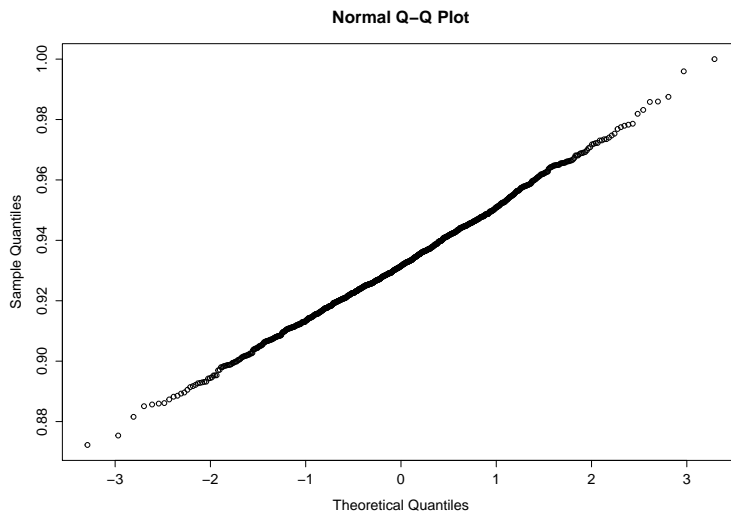


Figure A.36: QQ plot of normality of TF derived cFFC redundancy after both miRNA connection randomization - FANTOM Data

A.4 QQ Plots of Partial Randomization Procedure of FANTOM Data

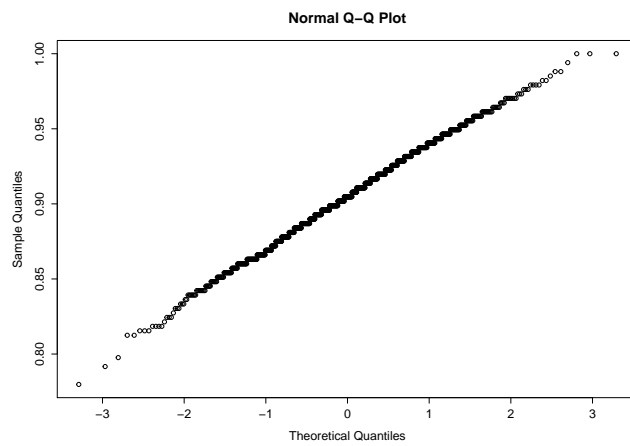


Figure A.37: QQ plot of normality of number of cFFC targeted genes after only miRNA-TF connection randomization - FANTOM Data

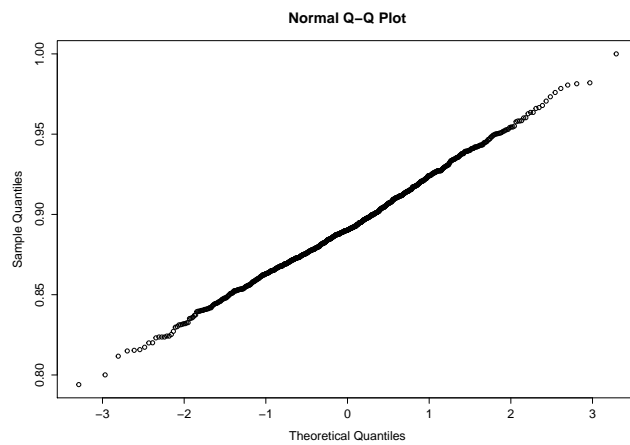


Figure A.38: QQ plot of normality of cFFC redundancy after only miRNA-TF connection randomization - FANTOM Data

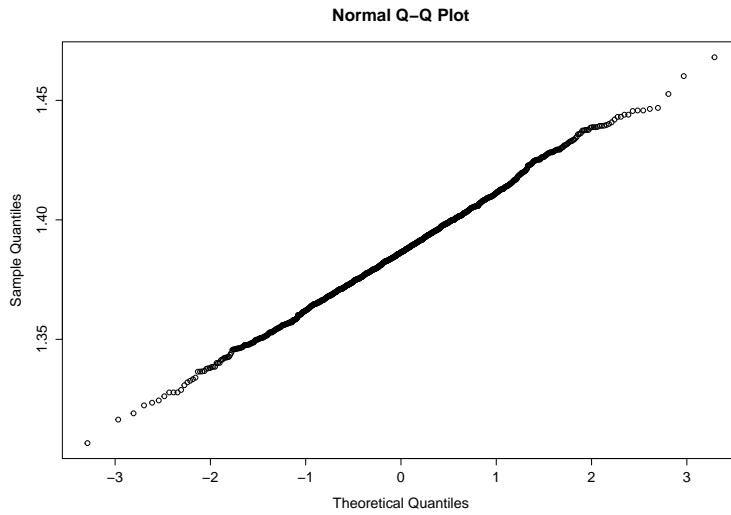


Figure A.39: QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-TF connection randomization - FANTOM Data

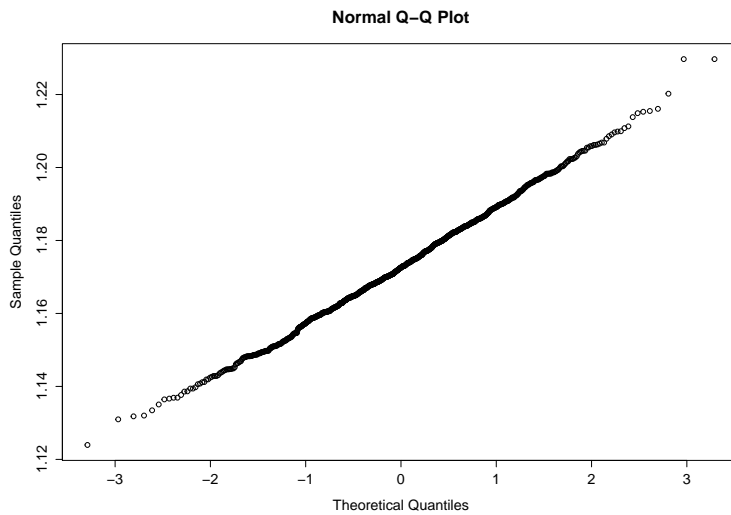


Figure A.40: QQ plot of normality of TF derived cFFC redundancy after only miRNA-TF connection randomization - FANTOM Data

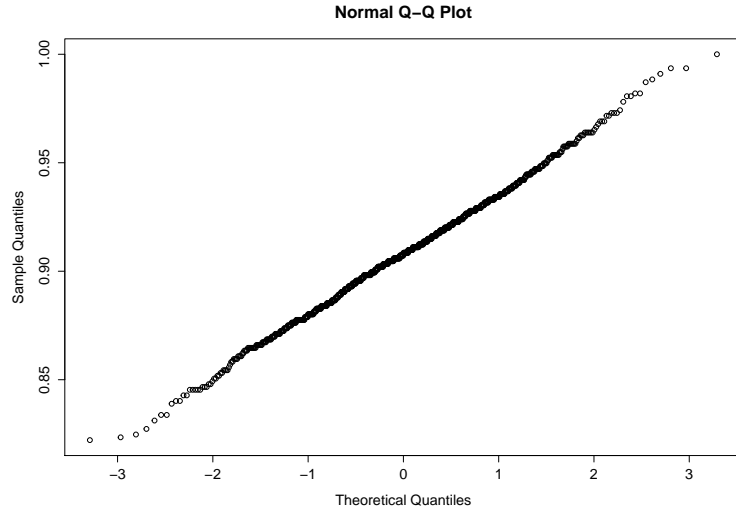


Figure A.41: QQ plot of normality of number of cFFCs after only miRNA-gene connection randomization - FANTOM Data

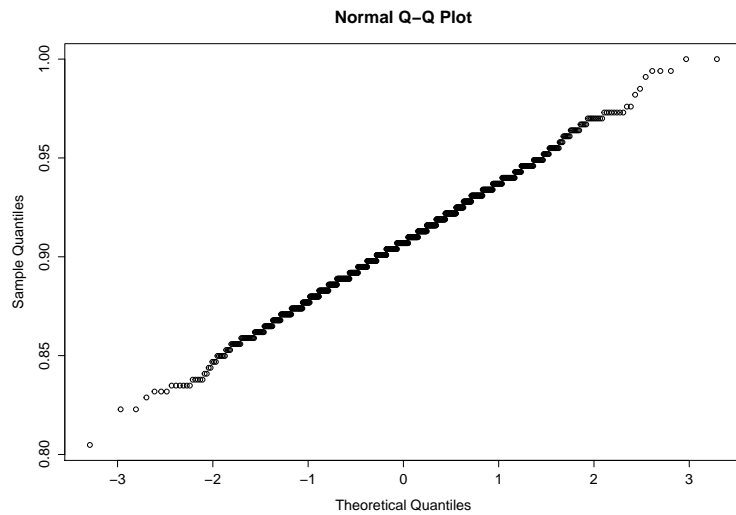


Figure A.42: QQ plot of normality of number of cFFC targeted genes after only miRNA-gene connection randomization - FANTOM Data

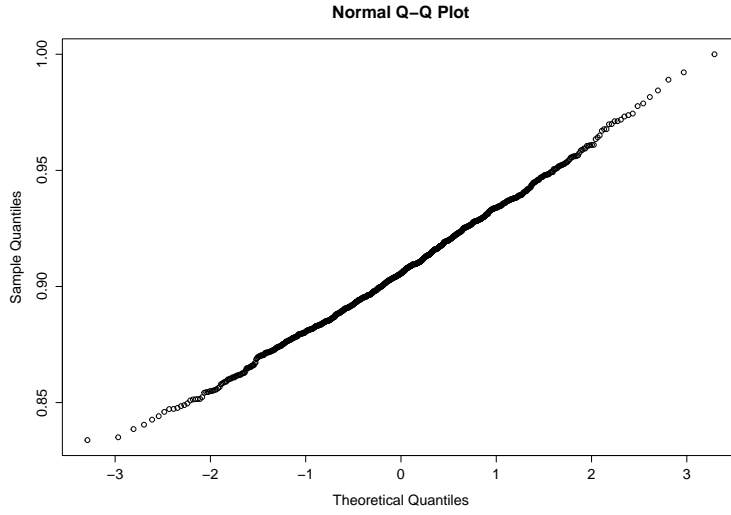


Figure A.43: QQ plot of normality of cFFC redundancy after only miRNA-gene connection randomization - FANTOM Data

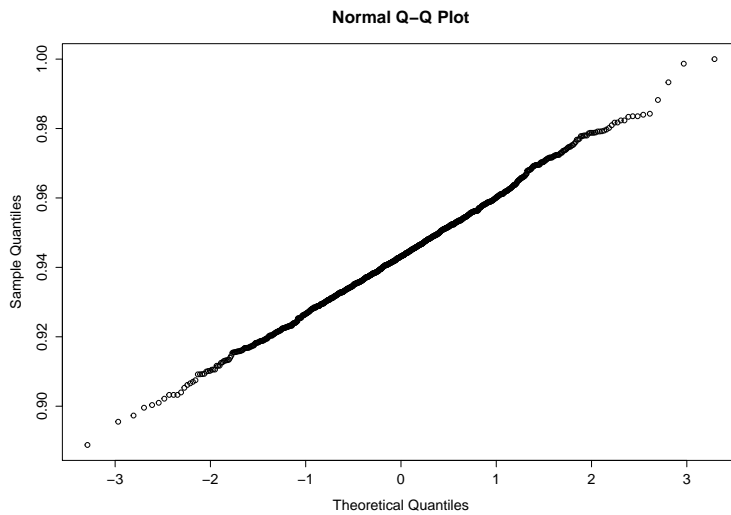


Figure A.44: QQ plot of normality of miRNA derived cFFC redundancy after only miRNA-gene connection randomization - FANTOM Data

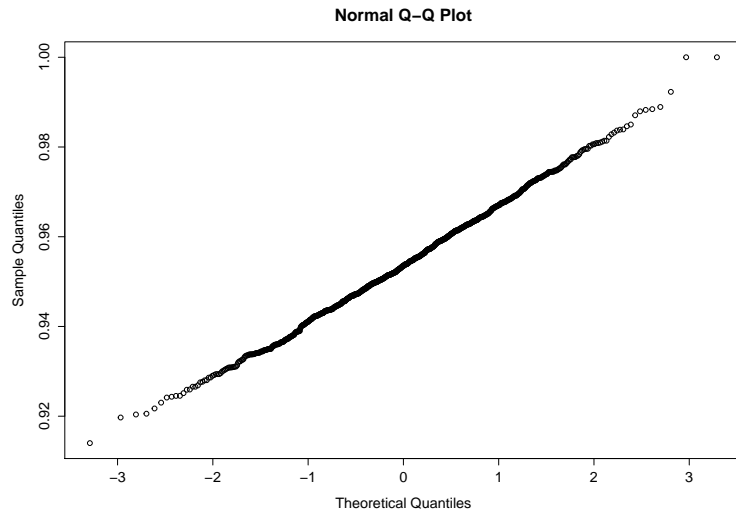


Figure A.45: QQ plot of normality of TF derived cFFC redundancy after only miRNA-gene connection randomization - FANTOM Data

APPENDIX B

SOURCE CODES

B.1 Parser to Retrieve miRNA Connections

```
import xml.dom.minidom
import sys
filelist = open(sys.argv[1])
weight_limit = -100
if len(sys.argv) > 2:
    weight_limit = float(sys.argv[2])
counter = 0
for xmlfile in filelist:
    dom = xml.dom.minidom.parse(xmlfile.strip())
    counter += 1
    if counter % 1000 == 0:
        sys.stderr.write('%d \n' % (counter) )
    root_element = dom.getElementsByTagName('EEDB')[0]
    feature_element = None
    mirna_element = None
    for node in root_element.childNodes:
        if node.nodeName == 'feature':
            feature_element = node
        if node.nodeName == 'miRNA_edges':
            mirna_element = node
    from_id = feature_element.getAttribute('desc')
    for node in mirna_element.childNodes:
        if node.nodeName == 'link_to':
            if float(node.getAttribute('weight')) >= weight_limit:
                print "%s, %s, %s" % (from_id, node.getAttribute('name'),
                                     node.getAttribute('weight'))
```

B.2 Parser to Retrieve TF Connections

```
import xml.dom.minidom
import sys
```

```

filelist = open(sys.argv[1])
weight_limit = -100
if len(sys.argv) > 2:
    weight_limit = float(sys.argv[2])
counter = 0
for xmlfile in filelist:
    dom = xml.dom.minidom.parse(xmlfile.strip())
    counter += 1
    if counter % 1000 == 0:
        sys.stderr.write('%d \n' % (counter) )
    root_element = dom.getElementsByTagName('EEDB')[0]
    feature_element = None
    tfbs_element = None
    for node in root_element.childNodes:
        if node.nodeName == 'feature':
            feature_element = node
        if node.nodeName == 'tfbs_predictions':
            tfbs_element = node
    from_id = feature_element.getAttribute('desc')
    for node in tfbs_element.childNodes:
        if node.nodeName == 'link_to':
            if float(node.getAttribute('weight')) >= weight_limit:
                print "%s, %s, %s" % (from_id, node.getAttribute('name'),
                                     node.getAttribute('weight'))

```

B.3 Orthology Parser

```

import csv
file1 = list();
with open('input filename1', 'rb') as f:
    reader = csv.reader(f, delimiter=' ')
    for row in reader:
        file1.append(row);
file2 = list();
with open('input filename2', 'rb') as f:
    reader = csv.reader(f, delimiter=',')
    for row in reader:
        file2.append(row);
with open('output file', 'wb') as csvfile:
    writer = csv.writer(csvfile, delimiter=',')
    for row1 in file1:
        for row2 in file2:
            if row1[0].lower().strip() == row2[1].lower().strip():
                writer.writerow(row2)

```


B.4 Source Codes of Main Randomization Procedure

```
import random
import csv

def readmatrix(filename):
    f = open(filename)
    headerline = f.readline().strip().split(',')
    locs = headerline[1:]
    mydict = {}\lbrace \rbrace$
    for x in locs:
        mydict[x] = \lbrace \rbrace
    for line in f:
        values = line.strip().split(',')
        geneid = values[0];
        interactions = values[1:]
        for i in range(len(interactions)):
            if (interactions[i] == '1'):
                mydict[locs[i]][geneid] = 1
    f.close()
    return mydict

def invertdict(mydict):
    inversedict = {}
    for key, val in mydict.items():
        for subkey, subval in val.items():
            if not inversedict.has_key(subkey):
                inversedict[subkey] = {}
            inversedict[subkey][key] = subval
    return inversedict

def countForMir(m, mirdict, tfdict, genecountdict, mirEdgeCountDict, tfEdgeCountDict):
    targets = mirdict[m]
    count = 0
    for t in targets:
        if tfdict.has_key(t):
            for g in tfdict[t]:
                if targets.has_key(g):
                    if not t == g:
                        count = count + 1
                    if genecountdict.has_key(g):
                        genecountdict[g] = genecountdict[g] + 1
                    else:
                        genecountdict[g] = 1
                if not mirEdgeCountDict.has_key(m):
                    mirEdgeCountDict[m] = {}
                mirEdgeCountDict[m][g] = 1
            if not tfEdgeCountDict.has_key(t):
```

```

        tfEdgeCountDict[t] = {}
        tfEdgeCountDict[t][g] = 1
    return count

def randomizeMatrix (mirdict):
    firstmir = random.sample(mirdict,1)[0]
    secondmir = random.sample(mirdict,1)[0]
    while (len(mirdict[firstmir]) == 0): firstmir = random.sample(mirdict,1)[0]
    while (len(mirdict[secondmir]) == 0): secondmir = random.sample(mirdict,1)[0]
    firstgene = random.sample(mirdict[firstmir],1)[0]
    secondgene = random.sample(mirdict[secondmir],1)[0]
    if (mirdict[secondmir].has_key(firstgene) or
        mirdict[firstmir].has_key(secondgene)):
        return 0
    mirdict[firstmir].pop(firstgene)
    mirdict[firstmir][secondgene] = 1
    mirdict[secondmir].pop(secondgene)
    mirdict[secondmir][firstgene] = 1
    return 1

mirdict = readmatrix("Final.csv")
tfdict = readmatrix("TF_Numbered.csv")
repcount = 1000
randcount = 500000
cffc = open("number_of_cFFCs.csv", "w")
cffcTargetedGene=open("number_of_cFFC_targeted_genes.csv","w")
mirnaDerivedRedundancy = open("miRNA_derived_redundancy.csv","w")
tfDerivedRedundancy = open("TF_derived_redundancy.csv","w")
genecountdict = {}
mirEdgeCountDict = {}
tfEdgeCountDict = {}
total = 0
mirtargetedgecount = 0
tftargetedgecount = 0
numrandomize = 0
for mir in mirdict:
    count = countForMir(mir,mirdict, tfdict, genecountdict,mirEdgeCountDict,
                        tfEdgeCountDict)

    total = total + count
for x in mirEdgeCountDict.keys():
    mirtargetedgecount = mirtargetedgecount + len(mirEdgeCountDict[x])
for y in tfEdgeCountDict.keys():
    tftargetedgecount = tftargetedgecount + len(tfEdgeCountDict[y])
print "Initial number of cFFCs is:", total
print "Initial number of cFFC targeted genes is:", len(genecountdict)
print "Initial miRNA derived redundancy:", float(total) / (mirtargetedgecount)
print "Initial TF derived redundancy:", float(total) / (tftargetedgecount)
print "Initial mir target edge count", mirtargetedgecount
print "Initial tf target edge count", tftargetedgecount

```

```

for i in range(0,repcount):
    genecountdict = {}
    mirEdgeCountDict = {}
    tfEdgeCountDict = {}
    numrandomize=0
    while (numrandomize < randcount): #for only miRNA connection matrix randomization
        numrandomize = numrandomize + randomizeMatrix(mirdict) #end

    while (numrandomize < randcount): # for only TF connection matrix randomization
        numrandomize = numrandomize + randomizeMatrix(tfdict) #end

    numrandomize = 0 #for both TF-miRNA connection matrices randomization
    numrandomize2 = 0
    while (numrandomize < randcount):
        numrandomize = numrandomize + randomizeMatrix(mirdict)
    while (numrandomize2 < randcount):
        numrandomize2 = numrandomize2 + randomizeMatrix(tfdict) # end
    total=0
    mirtargetedgecount = 0
    tftargetedgecount = 0
    for mir in mirdict:
        count = countForMir(mir,mirdict, tfdict, genecountdict,mirEdgeCountDict,
                             tfEdgeCountDict)

        total = total + count
    for x in mirEdgeCountDict.keys():
        mirtargetedgecount = mirtargetedgecount + len(mirEdgeCountDict[x])
    for y in tfEdgeCountDict.keys():
        tftargetedgecount = tftargetedgecount + len(tfEdgeCountDict[y])
    mirnaderivedredundancy = float(total) / (mirtargetedgecount)
    tfderivedredundancy = float(total) / (tftargetedgecount)
    cffc.write("list of gene count, %d\n" % (len(genecountdict)))
    cffcTargetedGene.write("list of loop count, %d\n" %(total))
    mirnaDerivedRedundancy.write("mirna derived redundancy,
                                  %.5f\n" % (mirnaderivedredundancy))
    tfDerivedRedundancy .write("tf derived redundancy,
                                  %.5f\n" % (tfderivedredundancy))

    print "Replicate is: ", i
    print "randomized count is:", total
    print "randomized number of targeted genes is:", len(genecountdict)
    print "mirna redundacy:", mirnaderivedredundancy
    print "tf redundacy:", tfderivedredundancy

```

B.5 Source Codes of Partial Randomization Procedure

B.5.1 miRNA-Gene Edge Randomization

```
# -*- coding: utf-8 -*-
```

```

import random
import csv
def readmatrix(filename):
    f = open(filename)
    mydict = {}
    locs = []
    for line in f:
        values = line.strip().split(',')
        keys = values[0].strip()
        if keys not in locs:
            locs.append(keys)
        geneid = values[1].strip()
        if not mydict.has_key(keys):
            mydict[keys]={}
        mydict[keys][geneid]=1
    f.close()
    return (mydict,locs)
def invertdict(mydict):
    inversedict = {}
    for key, val in mydict.items():
        for subkey, subval in val.items():
            if not inversedict.has_key(subkey):
                inversedict[subkey] = {}
            inversedict[subkey][key] = subval
    return inversedict
def writedict(filename, m):
    f = open(filename, "w")
    for x in m.keys():
        f.write(x + ": "+ ", ".join(m[x].keys()))
        f.write("\n")
    f.close()
def countForMir(m, mirdict, tfdict, genecountdict, mirEdgeCountDict, tfEdgeCountDict):
    targets = mirdict[m]
    count = 0
    for t in targets:
        if tfdict.has_key(t):
            for g in tfdict[t]:
                if targets.has_key(g):
                    if not t == g:
                        count = count + 1
                    if genecountdict.has_key(g):
                        genecountdict[g] = genecountdict[g] + 1
                    else:
                        genecountdict[g] = 1
            if not mirEdgeCountDict.has_key(m):
                mirEdgeCountDict[m] = {}
            mirEdgeCountDict[m][g] = 1
            if not tfEdgeCountDict.has_key(t):
                tfEdgeCountDict[t] = {}

```

```

        tfEdgeCountDict[t][g] = 1
    return count
def randomizeMatrixMirGene (mirdict,tfnames):
    firstmir = random.sample(mirdict,1)[0]
    secondmir = random.sample(mirdict,1)[0]

    while (len(mirdict[firstmir]) == 0): firstmir = random.sample(mirdict,1)[0]
    while (len(mirdict[secondmir]) == 0): secondmir = random.sample(mirdict,1)[0]
    firstgene = random.sample(mirdict[firstmir],1)[0]
    secondgene = random.sample(mirdict[secondmir],1)[0]
    if (firstgene in tfnames or secondgene in tfnames):
        return 0
    if (mirdict[secondmir].has_key(firstgene) or
        mirdict[firstmir].has_key(secondgene)):
        return 0
    mirdict[firstmir].pop(firstgene)
    mirdict[firstmir][secondgene] = 1
    mirdict[secondmir].pop(secondgene)
    mirdict[secondmir][firstgene] = 1
    return 1

from Counter import *
(mirdict,mirnames) = readmatrix("mirnaConnections.csv")
(tfdict,tfnames) = readmatrix("tfConnections.csv")

repcount = 1000
randcount = 500000
cffc = open("number_of_cFFCs.csv", "w")
cffcTargetedGene=open("number_of_cFFC_targeted_genes.csv","w")
mirnaDerivedRedundancy = open("miRNA_derived_redundancy.csv","w")
tfDerivedRedundancy = open("TF_derived_redundancy.csv","w")
genecountdict = {}
mirEdgeCountDict = {}
tfEdgeCountDict = {}
total = 0
mirtargetededgecount = 0
tftargetededgecount = 0
numrandomize = 0
    for mir in mirdict:
        count = countForMir(mir,mirdict, tfdict, genecountdict,mirEdgeCountDict,
                            tfEdgeCountDict)

        total = total + count
for x in mirEdgeCountDict.keys():
    mirtargetededgecount = mirtargetededgecount + len(mirEdgeCountDict[x])
for y in tfEdgeCountDict.keys():
    tftargetededgecount = tftargetededgecount + len(tfEdgeCountDict[y])
print "Initial count is:", total
print "number of targeted genes is:", len(genecountdict)
print "initial mirna redundancy:", float(total) / (mirtargetededgecount)

```

```

print "initial tf redundancy:", float(total) / (tftargetedgecount)
print "initial mir target edge count", mirtargetedgecount
print "initial tf target edge count", tftargetedgecount

for i in range(0,repcount):
    genecountdict = {}
    mirEdgeCountDict = {}
    tfEdgeCountDict = {}
    numrandomize=0
    while (numrandomize < randcount):
        numrandomize = numrandomize + randomizeMatrixMirGene(mirdict,tfnames)
    total=0
    mirtargetedgecount = 0
    tftargetedgecount = 0
    for mir in mirdict:
        count = countForMir(mir,mirdict, tfdict, genecountdict,mirEdgeCountDict,
                             tfEdgeCountDict)

        total = total + count
    for x in mirEdgeCountDict.keys():
        mirtargetedgecount = mirtargetedgecount + len(mirEdgeCountDict[x])
    for y in tfEdgeCountDict.keys():
        tftargetedgecount = tftargetedgecount + len(tfEdgeCountDict[y])
    mirnaderivedredundancy = float(total) / (mirtargetedgecount)
    tfderivedredundancy = float(total) / (tftargetedgecount)
    cffc.write("list of gene count, %d\n" % (len(genecountdict)))
    cffcTargetedGene.write("list of loop count, %d\n" %(total))
    mirnaDerivedRedundancy.write("mirna derived redundancy,
                                  %.5f\n" % (mirnaderivedredundancy))
    tfDerivedRedundancy .write("tf derived redundancy,
                                %.5f\n" % (tfderivedredundancy))

    print "Replicate is: ", i
    print "randomized count is:", total
    print "randomized number of targeted genes is:", len(genecountdict)
    print "mirna redundacy:", mirnaderivedredundancy
    print "tf redundancy:", tfderivedredundancy

```

B.5.2 miRNA-TF Edge Randomization

```

# -*- coding: utf-8 -*-
import random
import csv
def readmatrix(filename):
    f = open(filename)
    mydict = {}
    locs = []
    for line in f:
        values = line.strip().split(',')
        keys = values[0].strip()

```

```

        if keys not in locs:
            locs.append(keys)
        geneid = values[1].strip()
        if not mydict.has_key(keys):
            mydict[keys]={}
        mydict[keys][geneid]=1
    f.close()
    return (mydict,locs)
def invertdict(mydict):
    inversedict = {}
    for key, val in mydict.items():
        for subkey, subval in val.items():
            if not inversedict.has_key(subkey):
                inversedict[subkey] = {}
            inversedict[subkey][key] = subval
    return inversedict
def mirtfconnectiondict(inversedict, locs, mydict):
    excludedinteractions = {}
    allgenes = inversedict.keys()
    onlygenelist = set(allgenes).difference(locs)
    for key, gene in mydict.items():
        for subkey,subval in gene.items():
            if subkey in onlygenelist:
                del mydict[key][subkey]
                if not excludedinteractions.has_key(key):
                    excludedinteractions[key] = {}
                excludedinteractions[key][subkey] = 1
    if (len(mydict[key]) == 0):
        del mydict[key]
    return (mydict,excludedinteractions)
def merge(mydict, excludedinteractions, path=None):
    if path is None: path = []
    for key in excludedinteractions:
        if key in mydict:
            if isinstance(mydict[key], dict) and
                isinstance(excludedinteractions[key], dict):
                merge(mydict[key], excludedinteractions[key], path + [str(key)])
            elif mydict[key] == excludedinteractions[key]:
                pass # same leaf value
            else:
                raise Exception('Conflict at %s' % '.'.join(path + [str(key)]))
        else:
            mydict[key] = excludedinteractions[key]
    return mydict
def writedict(filename, m):
    f = open(filename, "w")
    for x in m.keys():
        f.write(x + ": "+ " ".join(m[x].keys()))
        f.write("\n")

```

```

        f.close()
def countForMir(m, mirdict, tfdict, genecountdict, mirEdgeCountDict, tfEdgeCountDict):
    targets = mirdict[m]
    count = 0
    for t in targets:
        if tfdict.has_key(t):
            for g in tfdict[t]:
                if targets.has_key(g):
                    if not t == g:
                        count = count + 1
                    if genecountdict.has_key(g):
                        genecountdict[g] = genecountdict[g] + 1
                    else:
                        genecountdict[g] = 1
                    if not mirEdgeCountDict.has_key(m):
                        mirEdgeCountDict[m] = {}
                    mirEdgeCountDict[m][g] = 1
                    if not tfEdgeCountDict.has_key(t):
                        tfEdgeCountDict[t] = {}
                    tfEdgeCountDict[t][g] = 1

    return count
def randomizeMatrix(mirdict):
    firstmir = random.sample(mirdict,1)[0]
    secondmir = random.sample(mirdict,1)[0]
    while (len(mirdict[firstmir]) == 0): firstmir = random.sample(mirdict,1)[0]
    while (len(mirdict[secondmir]) == 0): secondmir = random.sample(mirdict,1)[0]
    firstgene = random.sample(mirdict[firstmir],1)[0]
    secondgene = random.sample(mirdict[secondmir],1)[0]
    if (mirdict[secondmir].has_key(firstgene) or
        mirdict[firstmir].has_key(secondgene)):
        return 0
    mirdict[firstmir].pop(firstgene)
    mirdict[firstmir][secondgene] = 1
    mirdict[secondmir].pop(secondgene)
    mirdict[secondmir][firstgene] = 1
    return 1

from Counter import *
(mirdict,mirlist) = readmatrix("mirnaConnections.csv")
(tfdict,tflist) = readmatrix("tfConnections.csv")
repcount = 1000
randcount = 500000
cffc = open("number_of_cFFCs.csv", "w")
cffcTargetedGene=open("number_of_cFFC_targeted_genes.csv","w")
mirnaDerivedRedundancy = open("miRNA_derived_redundancy.csv","w")
tfDerivedRedundancy = open("TF_derived_redundancy.csv","w")
genecountdict = {}
mirEdgeCountDict = {}
tfEdgeCountDict = {}

```



```

total = 0
mirtargetedgecount = 0
tftargetedgecount = 0
numrandomize = 0
for mir in mirdict:
    count = countForMir(mir,mirdict, tfdict, genecountdict,mirEdgeCountDict,
                        tfEdgeCountDict)

    total = total + count
for x in mirEdgeCountDict.keys():
    mirtargetedgecount = mirtargetedgecount + len(mirEdgeCountDict[x])
for y in tfEdgeCountDict.keys():
    tftargetedgecount = tftargetedgecount + len(tfEdgeCountDict[y])
print "Initial count is:", total
print "number of targeted genes is:", len(genecountdict)
print "initial mirna redundancy:", float(total) / (mirtargetedgecount)
print "initial tf redundancy:", float(total) / (tftargetedgecount)
print "initial mir target edge count", mirtargetedgecount
print "initial tf target edge count", tftargetedgecount
mirtargetdict = invertDict(mirdict)
for i in range(0,repcount):
    genecountdict = {}
    mirEdgeCountDict = {}
    tfEdgeCountDict = {}
    numrandomize = 0
    mergedmirdict = {}
    (mirtfdict,mirgenedict) = mirtfconnectionDict(mirtargetdict, tflist, mirdict)
    while (numrandomize < randcount):
        numrandomize = numrandomize + randomizeMatrix(mirtfdict)
    total = 0
    mirtargetedgecount = 0
    tftargetedgecount = 0
    mergedmirdict = merge(mirtfdict,mirgenedict)
    for mir in mirdict:
        count = countForMir(mir,mergedmirdict, tfdict, genecountdict,
                            mirEdgeCountDict,tfEdgeCountDict)
        total = total + count
    for x in mirEdgeCountDict.keys():
        mirtargetedgecount = mirtargetedgecount + len(mirEdgeCountDict[x])
    for y in tfEdgeCountDict.keys():
        tftargetedgecount = tftargetedgecount + len(tfEdgeCountDict[y])
    mirnaderivedredundancy = float(total) / (mirtargetedgecount)
    tfderivedredundancy = float(total) / (tftargetedgecount)
    cffc.write("list of gene count, %d\n" % (len(genecountdict)))
    cffcTargetedGene.write("list of loop count, %d\n" %(total))
    mirnaDerivedRedundancy.write("mirna derived redundancy,
                                %.5f\n" % (mirnaderivedredundancy))
    tfDerivedRedundancy.write("tf derived redundancy,
                               %.5f\n" % (tfderivedredundancy))
    print "Replicate is: ", i

```

```
print "randomized count is:", total
print "randomized number of targeted genes is:", len(genecountdict)
print "mirna redundancy:", mirnaderivedredundancy
print "tf redundancy:", tfderivedredundancy
```