

Lincoln University Digital Thesis

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- you will use the copy only for the purposes of research or private study
- you will recognise the author's right to be identified as the author of the thesis and due acknowledgement will be made to the author where appropriate
- you will obtain the author's permission before publishing any material from the thesis.

**Identification of Glucocorticoid-regulated Genes and Inferring
their Network focused on the Glucocorticoid Receptor in
Childhood Leukaemia, based on Microarray Data and Pathway
Databases**

A thesis
submitted in partial fulfilment
of the requirements for the Degree of
Doctor of Philosophy

at
Lincoln University
by
Amphun Chaiboonchoe

Lincoln University
2010

Abstract of a thesis submitted in partial fulfilment of the
requirements for the Degree of Doctor of Philosophy

Abstract

Identification of Glucocorticoid-regulated Genes and Inferring their Network focused on the Glucocorticoid Receptor in Childhood Leukaemia, based on Microarray Data and Pathway Databases

by

Amphun Chaiboonchoe

Acute lymphoblastic leukaemia (ALL) has the highest mortality rate in childhood cancer. Glucocorticoids (GCs) have been used as chemotherapeutic drugs for children with ALL for more than 50 years. GCs induce apoptosis in lymphoid cells. However, little is known about the molecular mechanism of GC-induced apoptosis and there are many controversial hypotheses about genes regulated by GCs and their gene networks. In particular, two main issues are investigated: (i) GC-regulated genes and (ii) the glucocorticoid receptor (GR) gene networks. Only few overlapping genes have been reported from previous studies. Moreover, GCs function by binding with their receptors. The underlying mechanisms of cell type specific GR gene networks are not well established.

The goal of this thesis is to understand the mechanism of the GC-induced apoptosis mechanism. The first part of this thesis presents an identification of GC-regulated genes. This study uses secondary microarray data, originating from prednisolone (glucocorticoid) treated childhood ALL samples (Schmidt et al., 2006) (B-lineage and T-lineage) that were collected before treatment and at six and twenty four hours after treatment. We replicate the authors' original study and discover more probe sets including all the probe sets from that original study. This result shows the robustness of this data. Then, we extend the data analysis and propose new criteria based on differences between T- and B-ALL patients. The results reveal the proposed GC-regulated genes. These candidate genes are grouped in order to find similar expression patterns which lead to possible co-regulated genes, or similar function and sharing networks and pathways. Four emergent clustering methods are used: Self organising maps

(SOM), Emergent self organising maps (ESOM), the Short Time series Expression Miner (STEM) and Fuzzy clustering by Local Approximation of MEmbership (FLAME). These genes are used in the following gene expression analysis step.

The second part of this thesis focuses on inferring gene networks of GC-regulated genes and GR. There are many tools available for inferring gene networks including mathematical modelling and statistical methods. Each tool has its own advantages and disadvantages. For a modelling method, how do we know that the model represents the true relationship or interaction among genes? The need to verify results from modelling still exists. Prior knowledge has been used for this purpose. In this study, we use literature knowledge-based network tools, mainly the Ingenuity Pathway Analysis software (IPA) to elucidate gene networks. First, we illustrated gene networks at three time intervals and identified the prominent genes during those time points. Second, we further elucidated GR gene networks using gene lists from STEM. Third, we investigated the behaviour of selected known genes from the apoptosis, p53 and NF κ B pathways and inferred gene networks from the selected genes. Fourth, we inferred GR gene networks using the same gene list from previous studies (Phillip et al., 2005). We also used another two network tools: the BiblioSphere Pathway Edition (BSPE), and Oncomine to enhance the reliability of the gene network. Finally, we propose a GR gene network.

In summary, we undertook a gene to gene network of GC-induced apoptosis process based on childhood leukaemia patients. This study identified novel genes and their functions, and pinpointed possible gene networks which provide information for future research.

Keywords: apoptosis; BiblioSphere Pathway Edition; childhood leukaemia; DNA microarray; emergent self organising maps; Fuzzy clustering by Local Approximation of MEmbership; gene expression; glucocorticoids; glucocorticoid receptor; Ingenuity Pathway Analysis software, Oncomine; prednisolone; short time series clustering, Short Time series Expression Miner; self organising maps

Publications

The thesis has produced the following publications:

- Chaiboonchoe, A, S. Samarasinghe and D. Kulasiri (2009). Using emergent clustering methods to analyse short time series gene expression data from childhood leukaemia treated with glucocorticoids. *Proceeding of the 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, 13-17 July 2009, Cairns, Australia, pp. 741-747. ISBN: 978-0-9758400-7-8. (Anderssen, R.S., R.D. Braddock and L.T.H. Newham (eds))*
- Chaiboonchoe, A, S. Samarasinghe and D. Kulasiri (2009). Machine learning for childhood acute lymphoblastic leukaemia gene expression data: A review. *Current Bioinformatics*, volume 5, issue 02, June 2010, pp. 118-133.
- A. Chaiboonchoe, S. Samarasinghe and D. Kulasiri (2010). Identification of GC-Regulated genes and inferring the GR Gene Network in leukaemia based on Microarray data and Pathway Databases, *the Eight Asia Pacific Bioinformatics Conference (APBC 2010), Bangalore, India, 18-21 January 2010* (Poster).

This thesis is dedicated to my family for

.....their dream and unconditional love.....

With understanding and loving kindness, we will look within ourselves.

We will find happiness, wisdom, and serenity.

We will find the “Heart of a Buddha.”

Acknowledgements

The research for this work was carried out during the years 2007-2010 at the Centre for Advanced Computational Solutions (C-fACS), Lincoln University, New Zealand. During my PhD study, it was a period of hard work and intensive learning. This journey may not have come to the end without the support from the following people.

First and foremost, my sincerest gratitude is due to my supervisor, Associate Professor Sandhya Samarasinghe, for providing invaluable advice and support throughout my PhD study. It has been a privilege and pleasure to work under your supervision. Your endless patience, dedicated time spent with me; and scientific professionalism and knowledge have taught me everything I know about doing the best research possible. I am grateful to my co-supervisor, Professor Don Kulasiri, for his continuous encouragement, suggestions and advice that inspired me along this tough path and through hard times.

Secondly, I would like to thank Caitriona Cameron and Chris Odell at Teaching and Learning Services for their great help, support and encouragement to improve my English writing skills. Thank you also to Douglas Broughton, postgraduate administrator, and Tracey Shields, secretary Department of Environmental Management, for their hard work and support in all ways and their kind responses to all problems from students.

Thirdly, my heartfelt gratitude to New Zealand, the country and the people, for the education policy which allows all international PhD students to pay tuition fees at the domestic rate, and to this country, for its reputation for environmental concern, its beauty and peace, and its serene scenery in which to live and study. I would also like to give special appreciation to the Faculty of Environment, Society and Design for their research and conference funding which provided a great opportunity for me to attend and present my research at two international conferences, in Australia and India. I would like to extend my appreciation to all my part-time workplaces, this thesis could not have been accomplished without these financial sources; especially, Max and Mirium Pascoe, the owners of Gentian flower field, which is always my summer workplace, Kaituna orchards, the Faculty of Agriculture and Life Science and others. The financial support from 2008 onwards, provided by the Lincoln University Postgraduate

Scholarship, is gratefully acknowledged as well as the scholarship from Lincoln University Venture Out.

I am grateful to all of my friends who have helped, supported, cheered me on and accompanied me, each in their own way, during times when things were hard. Special thanks go to my Thai community friends for all we have done together for the Lincoln University Thai Club; every gathering was always a cheerful, enjoyable and memorable moment. I gratefully thank Apiwan Manimmanakorn and family, Rojana Thammajinda and family, Tippawan Sethapun and family, Arporn Popa, Hannah Lee, Krittaporn Na Pombejra, Nattinee Limkitisupasin, Patchanee Boontaganon, Patcharee Suriya, Sunee Sakseau and Yaowarat Sriwaranan. I have made many good friends while playing badminton, sharing flats and sharing offices; our relationships will continue and I thank you for all encouragement and support. This is especially so, for my long time best friend, Jintra Nakkarugx who despite being far away has always been there for me, with our regular internet and phone calls to share all stories, feelings and support each other.

Finally, I would like to express my deepest gratitude to my parents, mother, Sri, and father, Teeranit Chaiboonchoe, and my brother, for their inspiration and belief in me in every decision I have made in my life.

Table of Contents

Abstract	ii
Publications	iv
Acknowledgements	vi
Table of Contents.....	viii
Table of Contents.....	viii
List of Tables.....	x
List of Figures	xi
Chapter 1 Introduction	1
1.1 Childhood Leukaemia	1
1.2 Inferring Gene Networks	3
1.3 Objectives	5
1.4 Chapter Overview and Specific Contributions	10
Chapter 2 Literature Review.....	12
2.1 Childhood Leukaemia	12
2.2 Glucocorticoid-induced apoptosis	16
2.3 Microarray Data Analysis	18
2.3.1 DNA Microarray	19
2.3.2 Microarray Normalisation	24
2.4 Clustering	31
2.4.1 Self Organising Maps (SOM)	34
2.4.2 Emergent Self Organising Maps (ESOM)	37
2.4.3 Fuzzy clustering by Local Approximation of MEmbership (FLAME)	38
2.4.4 Short Time series Expression Miner (STEM)	41
2.5 Network/Pathway Analysis	43
2.5.1 Oncomine	46
2.5.2 BiblioSphere Pathway Edition (BSPE)	47
2.5.3 Ingenuity Pathway Analysis software (IPA)	49
Chapter 3 Emergent Clustering Methods for the Identification of Glucocorticoid- induced Apoptosis Genes	51
3.1 Introduction	51
3.2 Data and Software	53
3.2.1 Dataset	54
3.2.2 Methods	54
3.3 Results and Discussion	58
3.3.1 Validation/extension of original authors' results and identification of GC- regulated genes	58
3.3.2 Extraction of intrinsic biological patterns with four emergent clustering methods applied to gene clustering	69
3.6 Conclusion	87

Chapter 4 Identification of GC-induced apoptosis gene network.....	89
4.1 Introduction	89
4.2 Methods	90
4.2.1 Dataset	90
4.2.2 Computational Methods	90
4.3 Results and Discussion	91
4.3.1 Inferring GR gene networks from GC-induced apoptosis genes	91
4.3.2 Inferring GR gene networks from selected genes from STEM	102
4.4 Conclusions	107
Chapter 5 Inferring Gene Networks from Microarray data and Pathway Databases	109
5.1 Introduction	109
5.2 Methods	113
5.2.1 Dataset	113
5.2.2 Computational Methods	113
5.3 Results and Discussion	114
5.3.1 Inferring GR gene networks from selected genes from three pathways (apoptosis, p53 and NFκB)	114
5.3.2 Inferring GR gene networks using genes found vital in previous studies	121
5.3.3 Proposed GR gene network	131
5.4 Conclusions	134
Chapter 6 Summary, Conclusions and Future Directions.....	136
6.1 Summary	137
6.2 Conclusion	142
6.3 Contributions	142
6.4 Future Research	143
6.4.1 Computational methods	143
6.4.2 Biology of Childhood Leukaemia	146
Appendix A.....	148
A.1 Comparing genes from original author with our data analysis	148
A.1.1 Comparison of the number of patients who passed selection criteria in Table S2A from original article (left) and reproduced (right) in our study	149
A.1.2 Comparison of the number of patients who passed selection criteria in Table S2B from original article (left) and reproduced (right) in our study	150
A.1.3 Comparison of the number of patients who passed selection criteria in Table S2C from original article (left) and reproduced (right) in our study	151
A.1.4 Comparison of the number of patients who passed selection criteria in Table S2D from original article (left) and reproduced (right) in our study	152
A.2 Extra probe sets from our data analysis	153
A.2.1 Extra probe sets found in this study for B-ALL at each time point	153
A.2.2 Extra probe sets found in this study for T-ALL at each time point	155
References.....	157

List of Tables

Table 2.1: Overview of acute lymphoblastic leukaemia microarray data analysis for gene selection: class discrimination and subtype discovery	14
Table 2.2: Overview of acute lymphoblastic leukaemia microarray data analysis for gene selection: treatment, drug response and relapse	15
Table 2.3: Useful URLs for accessing cancer-related data and microarray gene expression databases [adapted from Barnes (2007) and Gardiner-Garden & Littlejohn (2001)].	20
Table 2.4: Description of some pre-processing methods, adapted from Irizarry et al., (2006) and N. Jiang et al., (2008).	26
Table 2.5: Advantages and disadvantages of some selected methods for inferring gene networks.....	45
Table 3.1: Differentially expressed genes involved with GC-induced apoptosis mechanism of childhood leukaemia treated with glucocorticoids, from previous studies (adapted from Schmidt et al., 2004 and Tissing et al., 2007).	52
Table 3.2: Lists of induced and repressed probe sets from Matlab, M-values (from original and reanalysis) and R/RMAExpress	60
Table 3.3: Differentially expressed genes	64
Table 3.4: GC-regulated genes and their activity patterns	65
Table 3.5: Gene list found by original authors using the combined dataset and the presence or absence of these genes in the two sub-types of ALL from separate analysis of the subtypes	66
Table 3.6: Results on 30 probe sets found relevant to the apoptosis process from Functional annotation clustering function by DAVID.	70
Table 3.7: Gene functional classification results from DAVID	71
Table 3.8: Significant clusters and profiles from STEM for T and B-ALL patients.....	78
Table 3.9: List of probe sets in some examples of significant clusters from STEM for B-ALL patients	80
Table 3.10: Comparison between gene functional groups (9 groups) with STEM significant profiles (16 profiles) for T- and B-ALL patients.....	81
Table 3.11: Comparison of probe sets from STEM profiles (16 profiles) and FLAME cluster (15 clusters: 14 clusters and one outlier) for patient number 2 (T-ALL) ..	85
Table 3.12: Comparison between gene functional groups (nine groups) with FLAME clusters for T- ALL (six clusters and one outlier cluster) and B-ALL (eight clusters and one outlier cluster) patients.....	86
Table 4.1: GC-regulated differentially expressed genes	90
Table 4.2: B-ALL gene network based on the reduced gene set for 0-6 hours, 6-24 hours and 0-24 hours	94
Table 4.3: List of T-ALL genes in networks based on the reduced gene set for 0-6 hours, 6-24 hours and 0-24 hours	95
Table 4.4: List of top three activities of the selected genes from Tables 4.2 and 4.3 in molecular and cellular functions, canonical pathways and functions for T- and B-ALL patients at three time intervals by IPA	98
Table 4.5: List of B-ALL gene networks based on gene list including GR from STEM clustering method.....	103
Table 4.6: List of T-ALL gene networks based on gene list including GR from STEM clustering method.....	105
Table 5.1: Selected gene lists from Apoptosis, p53 and NFκB pathway	117
Table 5.2: Comparison of gene expression between the original article by Phillippe et al. (2005) and our data for T-ALL and B-ALL patients.....	124

List of Figures

Figure 2.1: Schematic representation of the Affymetrix GeneChip® expression analysis system in use.....	22
Figure 3.1: Venn diagram of glucocorticoid induced differentially expressed genes at 6 h and 24 h after treatment from Matlab, M-values and RMAExpress	60
Figure 3.2: Some examples of induced probe sets that passed the criteria at 6 h from Matlab, M-values (reanalysis) and RMAExpress.....	61
Figure 3.3: Some example of repressed probe sets that passed the criteria at time 6.h from Matlab, M-values (reanalysis) and RMAExpress.....	61
Figure 3.4: Candidate differentially expressed GC-induced apoptosis genes in Table 3.4 broken down into the four patterns of expression.	64
Figure 3.5: Self-organising map of selected T- and B-ALL patients depicting gene expression across the three time points	73
Figure 3.6: Self-organising map of ALL patients at 24 hours.....	73
Figure 3.7: ESOM: P-Matrix (density-based) plot for the two selected patients (T-ALL top panel and B-ALL bottom panel).....	75
Figure 3.8: Results from STEM for T- and B-ALL	77
Figure 3.9: Clustering results from FLAME based on Pearson correlation for one selected patient from each subtype	83
Figure 4.1: B-ALL gene networks at 0-6 hours.....	96
Figure 4.2: Proposed B-ALL GC-induced apoptosis gene network for all genes active in at least one time interval (0-6 hours, 6-24 hours and 0-24 hours).....	99
Figure 4.3: Proposed T-ALL GC-induced apoptosis gene network for all genes active in at least one time interval (0-6 hours, 6-24 hours and 0-24 hours).....	100
Figure 4.4: Glucocorticoid receptor gene networks for B-ALL at 0-6 hours.....	104
Figure 4.5: Glucocorticoid receptor gene networks for T-ALL at 0-6 hours	106
Figure 5.1: The Integrated Apoptotic Pathways	112
Figure 5.2: The three gene networks from Table 5.1	119
Figure 5.3: Existing GR gene networks from three studies (Donn et al., 2007; Miller, Komak et al., 2007; Phillip et al., 2005).....	123
Figure 5.4: GR gene network from; a) BSPE for T-ALL; b) BSPE for B-ALL; c) IPA for T-ALL; and d) IPA for B-ALL.....	127
Figure 5.5: Box-plot distribution of NR3C1 across 15 independent experiments/studies created by Oncomine	130
Figure 5.6: Proposed GC-induced apoptosis based on NR3C1 gene network	133

Chapter 1

Introduction

Cancer is a leading cause of death in the world human population. One of cancer treatment process is chemotherapy. Chemotherapeutic agents kill abnormal cells; in other words, they activate the cell death programme (the apoptosis process). The knowledge of this process plays an important role in cancer therapy and there has been intensive research into different aspects of apoptosis in relation to cancer including uncovering the underlying mechanism of the apoptosis process. This research explores the glucocorticoid-induced apoptosis mechanism in childhood leukaemia using short time series gene expression data. In this chapter, childhood leukaemia and glucocorticoids are explained in detail and microarray data analysis and inferring gene networks from microarray data are reviewed. The research objectives are outlined and an overview of the thesis chapters is presented.

1.1 Childhood Leukaemia

White blood cells cancer, or leukaemia, starts in the bone marrow where the white blood cells are produced. The underlying causes of leukaemia remain unclear; however, there are risk factors indicated for some types of leukaemia including radiation, chemical, genetic problems, and smoke. The symptoms usually vary with the type of leukaemia but there are some common symptoms; for example, fever, headaches, easy bruising or bleeding, pain in the bones or joints and weight loss. There are two main groups of leukaemia: childhood and adult leukaemia. Childhood leukaemia can be divided into two types: acute (rapidly growing) or chronic (slow growing) but most childhood leukaemia is acute. There are two groups of acute leukaemia: acute lymphoblastic leukaemia (ALL) and acute myelogenous leukaemia (AML). Within these groups there are two subgroups of ALL: T-lineage and B-lineage. Chemotherapy (using drugs to kill cancer cells or stop cell division) is the most common treatment for children with ALL. The drug and dosage combinations may vary for each child. Unfortunately, chemotherapy treatments may result serious short- and long-term side effects.

Glucocorticoids are a type of steroid hormone. Synthetic glucocorticoids such as dexamethasone (Dex) and prednisolone (PRD) are the most important drugs that have been used extensively in the treatment of children with acute lymphoblastic leukaemia because of their ability to induce apoptosis (cell death) in the lymphoid cells. GCs enter the cell membrane and bind with the glucocorticoid receptor (GR), a member of the nuclear receptor subfamily 3, group C, also known as NR3C1, in order to induce apoptosis. The glucocorticoid-GR (GC-GR) complex then translocates to the nucleus, resulting either in transactivation or transrepression of the target genes. Transactivation happens when the GC-GR complex binds with glucocorticoid responsive elements (GRE) in the DNA, while in transrepression, it binds to transcription factors such as Nuclear Factor- κ B (NF- κ B) or Activator Protein-1 (AP-1). Transcriptional activation and repression cause immunosuppression, a stress response, or induction of apoptosis, depending on cell type (Tissing, Meijerink, den Boer, & Pieters, 2003).

GCs regulate diverse biological processes; for example, metabolism, development, differentiation, cell survival and apoptosis. Apoptosis is a cell death programme which varies with each multicellular organism and involves multi-steps and multi-pathways. There are two pathways for apoptosis: extrinsic and intrinsic. The extrinsic apoptosis pathway is initiated directly through the cell surface receptor while the intrinsic pathway is initiated through the mitochondria, the energy centre of the cell.

Several research groups have studied glucocorticoid-response genes in GC-induced apoptosis pathways using gene expression profiling. However, the GC-induced apoptosis mechanism is still an active research area. After reviewing the relevant literature, there are currently two issues which arise from the on-going research based on GCs (i) GC regulated genes and (ii) the glucocorticoid receptor gene network. Many studies have revealed a large number of GC-regulated genes using gene expression profiles from different cells and samples. There are more than 2000 studies on GC-induced apoptosis in lymphoid cells (Herr, Gassler, Friess, & Büchler, 2007). There are, however, only a few overlapping genes, as indicated in Chapters 2 and 3. Therefore, there is a need to identify and verify GC-regulated genes. Before GCs regulate genes to induce the apoptosis process, GCs bind to the GC receptor (GR), an intracellular receptor, and then transactivate or transrepress specific target genes.

The underlying mechanisms of cell type specific glucocorticoid receptor signals are not well understood. The understanding of GC's functional mechanisms may be enhanced by finding the GC-regulated genes and GR gene networks.

GC-induced apoptosis has been studied using microarray technology *in vivo* and *in vitro* on samples consisting of GC- treated ALL cell lines, mouse thymocytes and/or ALL patients. However, time series GC treated childhood ALL datasets are currently extremely limited. Currently, Systems Biology concepts are used to maximise the value of microarray data. Gene expression data can be used to ascertain signalling pathways and gene networks. One approach to gene networks is to present the information as a graph where the nodes represent genes and the edges represent interactions, which include activation or repression, and positive or negative feedback loops.

1.2 Inferring Gene Networks

DNA microarray technology is a technology for accessing thousands of gene expression profiles per experiment from different cells/tissues/organisms (Korenberg, 2007). This technology is widely used and leads to possible novel cancer causing genes, gene networks, and the identification of targets for cancer treatment. Specifically, there are two well-known microarrays: cDNA, and oligonucleotide arrays (Allison, Page, Beasley, & Edwards, 2006). Both arrays have been used in many studies, especially cancer studies, in search of possible novel genes for cancer diagnosis, prognosis and therapy.

DNA microarray technology has made publicly-available gene expression profiles available. However, the difficulties, time required and costs associated with collecting the data involved, mean only a few time series gene expression data are available. Time series expression data may play an essential role in understanding the underlying mechanisms of complex diseases through inferring gene networks. Studies so far are based on a wide range of biological systems (Bar-Joseph, 2004; Chan, Havukkala, Jain, Hu, & Kasabov, 2008; Ernst, Nau, & Bar-Joseph, 2005; Warren Liao, 2005) including synthetic and experimental data. Many existing clustering tools such as hierarchical clustering have been used with time series data; however, these clustering tools have not always been successful. Furthermore, Kim and Kim (2007)

describe some of the existing method restrictions that mainly focus on the similarity of gene expression and not including the time series characteristics of the data. Prior knowledge is also needed in some methods, and similarities in each cluster do not always represent a gene relationship or interaction.

There are two main types of time series, short and long. Most series (80%) are short time series, containing eight or fewer time points (Ernst et al., 2005). This presents a challenge to meaningful short time series data analysis. There are specific clustering tools for these data including STEM - the Short Time series Expression Miner, a software program specifically designed for investigating short time series gene expression data (Ernst & Bar-Joseph, 2006), and Difference-based clustering, an algorithm using differences between the first and second order for each time point (Kim & Kim, 2007). There are only limited time series data available for ALL, most are short time series data with only a few time points.

Microarray data analysis normally has four main methods: gene selection, clustering, classification and pathway analysis (S. B. Cho & Won, 2003). Many mathematical modelling, machine learning and statistical methods have been applied to microarray data analysis, especially, cancer (Berkhin, 2006; S. B. Cho & Won, 2003; Dam, Abbass, Lokan, & Yao, 2007; D. Jiang, Tang, & Zhang, 2004; Lau & Schultz, 2002; Ma, Castillo-Davis, Zhong, & Liu, 2006; J. Wang, Li, & Ruan, 2005; Y. Wang et al., 2005; Xu & Wunsch, 2005).

Networks can be studied at different levels, but in this study a gene network is the focus. Due to a lack of understanding of the actual network structures, deciphering gene networks from rapidly growing microarray expression databases has been shown to be a very promising approach in cancer research. Many tools are emerging and available for inferring gene networks. These tools include Boolean networks, Bayesian networks, Ordinary Differential Equations (ODEs), and pathway analysis. Considering their complexity, it is often difficult to evaluate or validate the performance of the available tools (Bansal, Belcastro, Ambesi-Impiombato, & di Bernardo, 2007; K. H. Cho et al., 2007; D'haeseleer, Liang, & Somogyi, 2000; de Jong, 2002; van Someren, Wessels, Backer, & Reinders, 2002). All existing tools, however, have their limitations and have not been able to infer whole gene networks (Andrecut & Kauffman, 2006). There are no conclusions about the most suitable method for

inferring gene networks as each tool may reveal different aspects of gene networks (van Someren, Wessels, Backer, & Reinders, 2002). However, inferring gene networks using gene expression profiles has the potential to reveal the underlying biological knowledge (Swain, Hunniford, Dubitzky, Mandel, & Palfreyman, 2005). In this knowledge- based era, integrating datasets from many experiments and existing databases is a difficult task. However, there are a several free tools available for non-profit use and some are licensed /commercial. This study infers gene networks through networks/pathways using databases. Some existing tools have been used for synthetic and experimental data, for example, the yeast cell cycle, however, they have not been applied for inferring gene networks from data of short time series of childhood ALL treated with glucocorticoid.

1.3 Objectives

This thesis covers gene expression analysis and the construction of gene to gene networks of childhood leukaemia. Most previous medical studies have focused on gene expression profiling which leads to the diagnostic and subtype classification for identifying novel therapeutic genes for ALL. Similarly, in the computational field, specific machine learning studies have used leukaemia data as a test set for newly developed methods. There is still a need for fundamental knowledge from both medical and computational research about childhood leukaemia. We aim to outline the essential available information through asking two questions:

- (a) What is the current situation of childhood leukaemia?
- (b) What current machine learning approaches have been used to study leukaemia?

We address this question by reviewing the existing literature using the keywords “childhood leukaemia, gene expression, machine learning”. For this review, refer to our published article, “*Machine learning for childhood acute lymphoblastic leukaemia gene expression data: a review*” (Chaiboonchoe, Samarasinghe, & Kulasiri, 2010). This review aims to serve as a starting point for those interested in microarray analysis, in general, and cancer research, in particular. In addition, Chapter 2 reviews the literature relevant to this study.

Our review led us to select this research topic, which focuses on childhood leukaemia treatment using chemotherapeutic drugs- glucocorticoids.

The primary purpose of this thesis is to better understand and to acquire more insight into the complexity of how GCs kill the malignant lymphoid cells through the GC-induced apoptosis mechanism in childhood leukaemia.

As mentioned in section 1.1, there are two main GC issues: (1) GC-regulated genes, and (2) the GR gene network. This study focuses on both issues. The first and second specific objectives discuss the first issue. The third and fourth specific objectives emphasise the second issue.

The first specific objective of this study is to identify GC-regulated candidate genes. This objective is accomplished by using the prominent short time series gene expression dataset collected by Schmidt et al. (2006). The samples were collected from childhood leukaemia patients treated with prednisolone. The original authors identified 22 glucocorticoid-response genes (also called GC-regulated genes or GC-induced apoptosis genes) using fold changes at early response (six hours after treatment). We start by investigating the selected data relating to the following three questions:

(1) How reproducible or robust are the original authors' results? We test the validity of the original these results by re-analysing them using the same dataset, method and gene selection criteria six hours after treatment. We also extend the analysis to cover 24 hours after treatment as well as between six and 24 hours.

(2) Do different platforms (software) available to normalise data have an effect on the final gene sets? We use R software as the original authors' did and add two software- Matlab and RMAExpress to normalise the raw data.

(3) Do leukaemia subtypes- T and B-ALL produce similar differentially expressed gene sets? Many studies have indicated that leukaemia subtypes-T and B-ALL have different gene expression profiles (Den Boer et al., 2009; Fulci et al., 2009; Mullighan et al., 2007; Yeoh et al., 2002). The original authors combined T and B-ALL data due to the small number of samples-13 patients. We aim to determine the validity of original authors' criteria and we propose a new gene expression analysis that separates T and B-ALL. This finding leads to our proposed GC-regulated genes.

The expected outcome from the first specific objective is a GC-induced apoptosis candidate gene set. This gene set is used throughout the thesis. We further classify GC-induced apoptosis genes based on their expression level before and after treatment and identify their cellular function. This leads to our **second specific objective, which is to identify group of GC-induced apoptosis genes that may have similar functions**. This objective can be achieved by using clustering methods.

Clustering methods have been used to cluster genes into similar groups in order to predict their functions. These clustered genes may be involved in similar networks and pathways. There are many clustering tools available and this is an on-going research area. We aim to understand how emergent clustering methods work with childhood leukaemia short time series gene expression data. We ask the following question:

(1) Do gene clusters differ when analysed by general clustering methods as opposed to using clustering methods specifically designed for short time series data?

We select four existing artificial intelligence methods: self organising maps (SOM), Emergent self organising maps (ESOM), Fuzzy clustering by Local Approximation of MEmbership (FLAME) and Short Time series Expression Miner (STEM). All gene clusters from the four methods are then compared and similar clusters are reported. This finding can be used by biologists or scientists for further investigation. Chapter 3 covers our first and second specific objectives. The work in this chapter has been published in the proceedings of 18th World IMACS Congress and MODSIM 2009.

The expected outcome from the second specific objective is a group of genes that may have a similar cellular function. This gene group does not, however, indicate an interaction between genes in the group. Which genes interact with which genes? This information can be studied through gene network inferring methods. A gene network can be constructed based on gene expression level, which represents gene to gene relationships. Microarray data can be used to elucidate gene networks through reverse engineering or inferring approaches. Currently, this is an on-going research and there are many proposed methods, for example, mathematical modelling, Boolean networks and Bayesian networks, as well as literature-based commercial and non-commercial network development tools.

The second aspect of this thesis addresses the second GCs related issue: the GR gene network. GCs are mediated by binding to, and activating, the glucocorticoid receptor (GR). Therefore, we aim to understand the relationship between the gene network of GC-induced apoptosis genes and the GR gene network. **The third specific objective is to construct gene networks of proposed GC-induced apoptosis genes.** We address this objective with the following questions:

(1) What are the possible gene networks before treatment and at six and 24 hours after treatment?

(2) What are common genes/gene hubs between these three gene networks?

The expected outcome from the third specific objective is a gene network of GC-induced apoptosis genes. Work related to this third objective is presented in Chapter 4. The Ingenuity Pathway Analysis software (IPA) is the main method used in this chapter.

We continue data analysis with **the final specific objective to elucidate the glucocorticoid receptor gene network.** This objective is accomplished by using web-based knowledge network/pathway tools. It is commonly known that the apoptosis process involves two main pathways: the extrinsic and the intrinsic. This leads to the following question:

(1) How do known genes from apoptosis pathways (extrinsic and intrinsic) behave in childhood leukaemia at different time points?

We further considered two other relevant pathways (p53 and NF κ B) because apoptosis involves multiple processes and pathways. Generally, a biological pathway represents molecule relationships which are not specific to the time when the interactions take place. We further investigate the gene networks of known genes from three selected pathways (apoptosis, p53 and NF κ B pathways).

Furthermore, we ask the question:

(2) Do different tissues (blood or liver) and drugs (prednisolone or dexamethasone) produce different GR gene networks?

We located three existing proposed networks from Phillip et al. (2005), Donn et al. (2007) and Miller et al. (2007). Networks have been proposed as literature-derived networks of biological relationships and signalling networks, respectively, by the two latter authors, and only Phillip et al. (2005) called the proposed network a regulator of GR. Phillip et al.'s (2005) study used mouse liver tissues treated with dexamethasone, while Schmidt et al. (2006) (used in our study) used blood sampling from childhood patients treated with prednisolone. The expected outcome from the fourth specific objective is a GR gene network constructed from the childhood leukaemia patients' *in vivo* dataset. This network is then compared with the previous proposed networks from the three studies mentioned above.

Next, we ask the question:

(3) What is the GR gene network obtained from the three web-based knowledge network/pathway tools?

These three tools are the Ingenuity Pathway Analysis software (IPA), the BiblioSphere Pathway Edition (BSPE) and Oncomine. IPA, a literature-based tool, has been used in many studies to construct gene networks, including the GR network, for leukaemia but not for childhood leukaemia. Oncomine is a well-established database for cancer, thus, it is an excellent source of information on genes related to cancer. Finally, we selected BSPE to validate the IPA network because BSPE is also a literature-based tool.

Finally, we proposed a GR gene network by combining all output gene networks in this objective with three existing GR gene networks; this work is presented in Chapter 5 using CellDesigner™.

This thesis sheds some light on the understanding of GC-induced apoptosis in childhood leukaemia in terms of identifying GC-regulated genes, their network, and its relationship with the GR gene network. This work is based on the analysis of gene expression data. The results of the analysis are presented and, as well, future directions for research are highlighted.

1.4 Chapter Overview and Specific Contributions

This thesis comprises six chapters.

Chapter 1 focuses on childhood leukaemia and its chemotherapeutic drugs- glucocorticoids- because of their capability to induce apoptosis in lymphoid and malignant lymphoid cells. Then, the motivation and objectives of this thesis are described.

Chapter 2 presents a review of the relevant literature. We address two relevant questions (i) the current situation of childhood leukaemia and (ii) current machine learning approaches that have been used to study leukaemia. The contribution from this chapter is an original and comprehensive published review article which provides essential basic knowledge in this area.

In Chapter 3, differentially expressed genes or novel GC-induced apoptosis genes, are identified and grouped according to their similarities in expression by using four emergent clustering methods. The first and second specific objectives are addressed in this chapter. The specific contribution from this chapter is (i) to emphasise different subtypes, sample types treated with different chemotherapeutic drugs may share common response; but there may still be unique patterns and the final genes discovered can vary. We extend the investigation further from the original Schmidt et al. (2006) study and propose new criteria to select novel genes in B-ALL and T-ALL subtypes. More genes were found than in the original research that combined the two subgroups in the analysis. Most available childhood leukaemia gene expression is from short time series. Different clustering methods produce different final gene clusters from the same dataset, and there is no conclusion about the best clustering method. Therefore, (ii) we addressed the comparison of gene clusters from four selected clustering methods with short time series data and evaluated the results with gene functional groups. Clustering revealed possible co-regulated or co-expressed genes but not the details of interactions between genes in the same cluster. Therefore, the next chapter is focused on inferring gene networks from the differentially expressed genes.

In Chapter 4, the gene networks of the candidate GC-induced apoptosis genes from Chapter 3 are elucidated by using network/pathway tools. The third specific objective is addressed in this chapter. The contribution of this chapter is demonstrating a combination of gene networks from three time intervals in order to minimise the relevant genes or identify novel genes for further study. We select the most common gene (node) and propose GC-induced apoptosis gene networks for T- and B-ALL, separately.

In Chapter 5, a GR gene network is proposed. The fourth specific objective is addressed in this chapter. In order to activate apoptosis process, GCs are binding to and activating the glucocorticoid receptor (GR). In Chapter 4, we have already illustrated the gene network for GC-regulated genes; in this chapter, we add the gene network that is involved with GR. This leads to understanding of the whole picture of GC-induced apoptosis process. We manually combined three existing GR gene networks from previous studies with our three inferred gene networks from the selected genes using CellDesignerTM. This network is a starting point for scientists conducting further investigations. The main contribution of our study is to present an approach to utilise publicly available gene expression data and pathway databases to identify candidate GC-regulated genes and their network which we believe expands the current knowledge about GC-induced apoptosis in childhood leukaemia. This network might lead to an understanding of the underlying mechanisms and better clinical treatment.

Finally, in Chapter 6, an overview and the most important findings of this research, including its contribution to the overall understanding of GC-induced genes and mechanisms, and suggestions for potential future research are defined.

Chapter 2

Literature Review

This research is part of the field called biomedical informatics which integrates different disciplines including medicine, genomics and informatics. As a result, this chapter provides an overview of the relevant background information. Section 2.1 starts with an extensive review of previous research on childhood leukaemia. This research focuses on glucocorticoids, which are used extensively in treatment of children with acute lymphoblastic leukaemia. Hence, section 2.2 is a discussion on how GCs kill the malignant lymphoid cells, as well as on GC responsive genes and pathways. To understand GC-induced cell death mechanism, microarray technology has been extensively used to identify the relevant genes and gain insight into the underlying biological process. Therefore, section 2.3 provides a review of microarray data analysis focusing on microarray technology and data processing. Furthermore, the gene sets extended from microarray data analysis are increasingly being used for further investigation on how genes work in particular conditions and tissues. Therefore, clustering techniques and networks/pathways analysis are presented in sections 2.4 and 2.5, respectively.

2.1 Childhood Leukaemia

Leukaemia is a cancer of blood cells; especially the abnormal proliferation of white blood cells. Normal blood stem cell development takes place in the bone marrow (BM) and then it divides and differentiates into platelets, red blood cells, and different types of lymphoids and myeloid cells. Consequently, leukaemia can be divided into four groups: acute myeloid leukaemia (AML), acute lymphoblastic leukaemia (ALL), chronic myeloid leukaemia (CML), and chronic lymphocytic leukaemia (CLL). The most common childhood leukaemia is acute lymphoblastic leukaemia (ALL), which can be further divided into two subgroups: T-lineage and B-lineage according to whether T- or B- lymphocytes are involved.

To date, many studies have focused on diagnostic and subtype classifications, identification of novel drug targets and identification of risk stratification for ALL patients. Golub et al. (1999) started the first leukaemia microarray data analysis (Golub et al., 1999), which was followed by many other groups (Armstrong et al., 2002; Cheok et al., 2003; Chiaretti et al., 2004; Chiaretti et al., 2005; Choi et al., 2007; DeAngelo, 2005; Dunphy, 2006; Moos et al., 2002; Mullighan et al., 2007) and work in several regions in the world including Canada and Europe, i.e. Germany and the United Kingdom (Herold, von Stackelberg, Hartmann, Eisenreich, & Henze, 2004; Ramanujachar et al., 2007; Rogers et al., 2007; Schröder et al., 2006). World-wide, there are at least nine well-known leukaemia research groups including the Boston, Austrian, Utah, Memphis, Japan, Munich, Stanford, Copenhagen and Netherland groups (Knudsen, 2006). An overview of some of those studies (including groups of investigators, methods used and outcomes achieved) that focus on childhood ALL is given in Tables 2.1 and 2.2. Basically, ALL research can be classified as follows:

- Overview of childhood leukaemia research, treatment and future research directions (Bhojwani, Moskowitz, Raetz, & Carroll, 2007; Carroll, Bhojwani, Min, Moskowitz, & Raetz, 2005; Carroll et al., 2003; Dunphy, 2006; Howell, Ward, Austin, Young, & Woods, 2007; Pui, 2004; Pui, Schrappe, Ribeiro, & Niemeyer, 2004)
- Identification and classification of leukaemia subtypes (Andersson, Edén et al., 2005; Andersson, Olofsson et al., 2005; Andersson et al., 2007; De Pitta et al., 2005; Moos et al., 2002; Ross et al., 2003; Willman, 2004; Yeoh et al., 2002).
- Identification of genetic determinants and aberrants (Kuiper et al., 2007; Kustanovich, Savitskaja, Bydanov, Belevtsev, & Potapnev, 2005; Mullighan et al., 2007; Sinnett, Labuda, & Krajinovic, 2006).
- Identification of novel genes that enhance the diagnosis and prognosis, drug response, poor drug response or relapse and effective therapies development for leukaemia treatment (Bhojwani et al., 2006; Estes, Lovato, Khawaja, Winter, & Larson, 2007; Flotho et al., 2007; Holleman et al., 2004; Holleman et al., 2006; Kirschner-Schwabe et al., 2006; Lugthart et al., 2005; Smedmyr & Heyman, 2006; Tissing et al., 2007; Willenbrock, Juncker, Schmiegelow, Knudsen, & Ryder, 2004).

Table 2.1: Overview of acute lymphoblastic leukaemia microarray data analysis for gene selection: class discrimination and subtype discovery

Authors	Methods	Findings
Golub et al. (1999)	Clustering: Self Organising Map (SOM) Classification : Neighbourhood analysis and weight voting	50 informative genes which distinguish ALL from AML
Moo et al. (2002)	Clustering: Hierarchical clustering using t-test and info-score	20 best discriminating genes for ALL vs AML and B-lineage vs T-lineage ALL
Yeoh et al. (2002)	Clustering: unsupervised hierarchical clustering Classification <ul style="list-style-type: none"> • <u>Genes selection</u> : Chi-square, t-statistics, and CFS (Correlation-based Feature Selection), and SOM/DAV (Self Organising Map/Discriminant Analysis with Variance) • <u>Supervised learning algorithms</u> : k-nearest neighbours (k-NN), support vector machine (SVM), neural networks (NN), weight voting, and prediction by collective likelihood of emerging patterns (PCL) 	six distinct subgroups of ALL : T-ALL, BCR-ABL, E2A-PBX1, TEL-AML1, MLL gene rearrangement, and hyperdiploid>50 chromosomes
Ross et al. (2003)	This is continuation of Yeoh et al.'s research with different Affymetrix array (HG-U133 instead of HG-U95Av2 array). Seven distinct subgroups of ALL: T-ALL, BCR-ABL, E2A-PBX1, TEL-AML1, MLL gene rearrangement, hyperdiploid>50 chromosomes and other (novel, hyperdiploid, normaldiploid and pseudodiploid). Comparing between top 100 selected genes; about 60% of the genes were not selected before and, thus, are new class discriminators. An ANN supervised learning algorithm was used with the top 50 identified genes, 97% of overall prediction accuracy was achieved.	
Willman (2004)	Clustering : VxInsight Software (http://www.cs.sandia.gov/projects/VxInsight.html) Classification : Bayesian networks and support vector machines with recursive feature elimination (RFE) (SVM-RFE), VxInsight/ANOVA and TnoM (Threshold number of misclassification))	This study uses gene expression profiling for class discovery and class prediction of ALL data. It highlights the possibility of finding potentially novel diagnostic and therapeutic targets by using microarray technology.
Andersson et al. (2005)	Clustering : Hierarchical clustering analyses and principal component analyses (PCAs) Classification : k-nearest-neighbours This study compares normal hematopoietic and cells leukemic cells. In addition, it identifies the gene-expression signatures of normal subpopulations of different lineages and maturations. There is 77-86% of differentially expressed genes overlap when using ALL and AML datasets by Ross <i>et al.</i> (2003). A high accuracy (98.2%) was retrieved when using k nearest neighbours' classifier to predicted genetic subtype among B lineage ALLs.	
De Pitta et al. (2005)	Significance analysis of Microarray (SAM), Predictive Analysis of Microarray (PAM), Principal component analysis, Hierarchical cluster analysis, k-means and profile similarity searching	This study identified 30 genes that best discriminate three subtypes: T-ALL, B-ALL and B-ALL with MLL/AF4 rearrangement.

Table 2.2: Overview of acute lymphoblastic leukaemia microarray data analysis for gene selection: treatment, drug response and relapse

Authors	Methods	Findings
Cheok et al. (2003)	Linear Discriminant Analysis (LDA), ANOVA, Support Vector Machine (SVM), k-nearest neighbour, artificial neural network and empirical Bayesian	Identified 124 genes that accurately discriminated among the four treatments
Holleman et al. (2004)	Wilcoxon rank-sum test, t-test, Bagging algorithms, Cox proportional-hazards regression, Fine and Gray's estimator accounting for competing events, Fisher's exact test and Hierarchical-clustering	124 genes were identified with resistance to four drugs: prednisolone, vincristine, asparaginase and daunorubicin.
Willenbrock et al. (2004)	Hierarchical clustering, k-nearest neighbour, Nearest centroid, LDA, SVM and Maximum Likelihood	This study shows the high percentage of classification accuracy (78%) achieved by using DNA microarray to predict relapse and treatment response in childhood ALL.
Lugthart et al. (2005)	Hierarchical clustering, Principal component analysis (PCA), Spearman's rank correlation.	Identified 45 genes associated with cross-resistance to four mechanistically distinct anti-leukaemic agents and 139 genes significantly related to a novel phenotype of discordant resistance to vincristine and asparaginase.
Bhojwani et al. (2006)	Robust multiarray analysis (RMA), Cluster and TreeView software, VxInsight, Multiple supervised analysis and SAM	Identified different pathways between the timing of disease recurrence (early and late relapse).
Kirschmer-Schwabe et al. (2006)	t-test, Fisher's exact test, least angle regression and nearest shrunken centroid	Identified 83 genes differentially expressed in very early relapsed ALL compared to late relapse.
Schmidt et al. (2006)	Fold change	22 genes as novel genes for glucocorticoids-induced apoptosis
Flotho et al. (2007)	Analysis-of-Variance (ANOVA), t-test, Spearman correlation, simple linear regression, Kruskal-Wallis test and Wilcoxon Mann-Whitney tests	A set of 40 genes that predicted clinical outcome and 14 were involved in the regulation of cell proliferation and associated with minimal residual disease (MRD) during early remission induction therapy.

The review above provides a perspective on childhood leukaemia research. One of the open questions is how to cure the children who do not respond well to existing childhood leukaemia treatments. An understanding of how chemotherapeutic drugs kill the immature lymphoid cells will lead to better childhood leukaemia treatment. Treatment of childhood acute lymphoblastic leukaemia is based on the patient's specific risk group. Unfortunately, it is difficult and expensive to identify accurately a patient's risk group. Therefore, gene expression profiling has been used to identify possible novel gene response in childhood leukaemia treated with chemotherapy. The most widely used therapeutic drugs for treating children with ALL are glucocorticoids (GCs) such as dexamethasone or prednisolone. GCs induce apoptosis (cell death) and inhibit proliferation in lymphoid and malignant lymphoid cells, as described in detail in the following section.

2.2 Glucocorticoid-induced apoptosis

Tumour cells are killed by induced apoptosis during chemotherapy treatment (Igney & Krammer, 2002). Glucocorticoids have been used as chemotherapeutic drugs for children with ALL. GCs induce apoptosis and G1 (Gap 1, an early stage of cell division between the synthesis and mitosis phases) cell cycle arrest in malignant lymphoid cells. In fact, little is known about the molecular mechanism of GC-induced apoptotic signal transduction pathways and there are many controversial hypotheses about both the genes regulated by GCs and the potential molecular mechanism of GC-induced apoptosis (Schmidt et al., 2006). Therefore, an understanding the mechanism of this drug should lead to better prognostic factors (treatment response), more targeted therapies and prevention of side effects. To understand the GC-induced apoptosis mechanism, this section reviews three relevant issues: known genes, known pathways and the relationship between cancer, apoptosis and cell cycle arrest.

Previous studies have focused on identifying genes involved in GC-induced apoptosis. A review by Schmidt et al. (2004) identified 31 common genes from seven previous studies (Schmidt et al., 2004). Each study found a different set of genes, for example, Tonko et al. (2001) found eight genes that were differentially regulated: Leucine zipper, integrin alpha6 (ITGA6), GR, ESTs, SOCS1/JAKbp, YAF2, LDH A and Arylsulfatase C. (Tonko, Ausserlechner, Bernhard, Helmberg, & Kofler, 2001).

Recent research by Schmidt et al. (2006) proposed 22 novel differentially expressed genes. In particular, they proposed that the GC-induced apoptosis mechanism is activated by several GR-induced genes. GR regulates transcription of its target genes which can be either activated or inhibited; for example, it activates p53 and represses AP-1, NF- κ B, and c-myc. Previous studies defined the interplay between GR, AP-1, NF- κ B, and c-myc (De Bosscher, Vanden Berghe, & Haegeman, 2003). The mechanism of GC inhibition of NF- κ B depends on cell type. A number of studies have reported that GR directly interacts with the transcription factors AP-1 and NF- κ B, as part of the GC-induced apoptosis (Greenstein, Ghias, Krett, & Rosen, 2002).

An issue that should be taken into account when identifying the novel genes is that treatment with different types of glucocorticoid drugs in different clinical settings (*in vivo*, *in vitro* and human samples) provides different differentially expressed novel gene sets. For example, Cario et al.'s (2008) study of prednisone response in childhood leukaemia found 72 out of 104 differentially expressed genes in common with Schmidt et al. (2006) who also used prednisolone. The novel glucocorticoid-response genes from *in vivo* and *in vitro* prednisolone treated paediatric ALL studies found only five genes in common (FKBP5/FKBP51, SNF1LK, ZBTB16, ZFP36L2 and SOCS-1) (Tissing et al., 2007). An *in vivo*, gene list by Schmidt et al (2006) claimed 22 essential candidate genes with only three common with the 31 identified from numerous previous studies (FKBP5/FKBP51, DDIT4/Dig2, and SOCS-1). Tissing et al.'s (2007) *in vitro* study highlighted 57 probe sets (51 differentially expressed genes), of which 22 are found in previous studies of lymphoid cell lines. Therefore, identification of the specific genes involved in GC-induced apoptosis mechanism is complex and is an on-going research issue.

From the previously known GC-induced apoptosis pathways, GC-induced apoptosis is generally activated through the regulation of caspase by two major pathways: intrinsic and extrinsic. Of the possible pathways involved in GC induced cell death, Tissing et al (2007) pointed out the link between three pathways (MAPK pathway, NF- κ B signalling and carbohydrate metabolism) while Herr et al (2007) proposed 12 molecules and pathways: mitochondria, death receptor signalling, Bcl-2 family, caspases, c-myc, I κ B, Granzyme A, TDAG8, lysosomes, proteasomal degradation, stress pathway and other modulators such as PKC, IL-6 and T-cell receptors (Herr et al., 2007). Detailed studies of each of these pathways

have been conducted including TRAIL death receptor signalling (Finnberg & El-Deiry, 2008; Ndebele et al., 2008; Press & Reminder, 2008), lysosomes (Guicciardi, Leist, & Gores, 2004), and Bcl-2 family (Adams & Cory, 2007; Youle & Strasser, 2008). Furthermore, a gene signalling pathway of GC-regulated genes was proposed by Miller et al. (2007) who used CEM cell lines (human leukaemia cultured cell lines) as samples. A glucocorticoid receptor gene network was presented by Phillip et al. (2005) based on liver tissue. At present, no GCs or their relevant gene networks have been reported using patient (human) samples. There are studies confirming differences in gene expression profiling between cell lines corresponding to normal or tumour tissues, with the degree of difference varying with the type of tissue (Ertel, Verghese, Byers, Ochs, & Tozeren, 2006; Leupin et al., 2006).

GC-induced apoptosis is a very complicated process. To understand GC-induced apoptosis, it is vital to understand it at a genetic level. Microarray technology enhances the possibility of investigating the thousands of gene expressions in one experiment. The next section describes microarray technology and data processing.

2.3 Microarray Data Analysis

In this section, an overview of microarray data analysis is presented. Section 2.3.1 explains DNA microarrays, focusing on Affymetrix GeneChip® HG-U133 Plus 2. Section 2.3.2 introduces the data normalisation process with details of the Robust Multichip Average (RMA).

Microarray technology is an essential source of data that promises to pave the way for better cancer prediction and diagnosis and to identify target drugs for cancer treatment. DNA microarray technology has been used to study human cancer including breast cancer, prostate cancer and leukaemia (Russo, Zegar, & Giordano, 2003). Since its initial introduction, the number of microarray applications has expanded. The technology involving the production including experimental design and preparation and image processing is beyond the scope of this review. The general procedure of microarray data analysis starts from experimental design including the design of experiments and extraction of mRNA samples. mRNA is

allowed to hybridise with a gene chip containing a strand of all the genes in the human genome (in the case of HG-U133 Plus 2). Genes represented in the sample and on the chip hybridise and the level of hybridisation is measured through image analysis, providing the raw data. Next, the raw data are further processed and normalised (more details are given in section 2.3.2) to be used in the next step, to identify candidate genes. This is followed by pattern discovery based on various clustering methods (see more details in section 2.4). The final microarray data analysis process is biological modelling. Reverse engineering, an approach used to construct biological networks from data and to get new biological insights, is discussed in section 2.5. Microarray experiments are expensive and generate an overwhelming volume of data; therefore, this has led to the creation of publicly available databases. A list of databases for cancer-related microarray data and gene expression data is given in Table 2.3. As of 30 July 2009, three selected databases were used to identify ALL datasets by using keyword “acute lymphoblastic leukaemia”. There are 97 datasets retrieved from Gene Expression Omnibus (GEO) (Barrett et al., 2007; Barrett et al., 2009), 85 experiments and 11,966 assays were obtained from Array Express, for Oncomine, there are 103 datasets from 18 studies. Specifically, 19 experiments and 1,298 assays were found in Array Express, there are eight datasets on GEO, and only one study was found in Oncomine when using the key word “childhood acute lymphoblastic leukaemia”.

2.3.1 DNA Microarray

Human organisms consist of cells with 24 chromosomes. Chromosomes constitute deoxyribonucleic acid (DNA), whose structure was illustrated by Jame D. Watson and Franscis H.C Crick in 1953 (Watson & Crick, 1953). DNA is recognised as a double helix. The double helix is formed by base pairs of hydrogen bonds: adenine (A) binds to thymine (T) and cytosine (C) binds to guanine (G). A gene is a part of DNA; its expression proceeds from transcription of genetic information from the DNA level to ribonucleic acid (RNA) level, then translation into the protein level. This process is represented as the central dogma of molecular biology. The regulation of gene expression may be controlled at many levels: RNA processing (transcription) and transport, RNA translation and post-translational modification of proteins, but the most well-established study is in the regulation of DNA transcription.

Table 2.3: Useful URLs for accessing cancer-related data and microarray gene expression databases [adapted from Barnes (2007) and Gardiner-Garden & Littlejohn (2001)].

Databases	Resource/Organisation	URL
	General Resources Cancer Genome Anatomy Project Cancer Genome Project National Cancer Institute Centre for Bioinformatics	http://www.cgap.nci.nih.gov http://www.sanger.ac.uk/genetics/CGP http://ncicb.nci.nih.gov
Amad	Gene Expression Stanford University/University of California at Berkeley, University of California at San Francisco (UCSF)	http://derisilab.ucsf.edu/data/microarray/software.html
ArrayExpress	European Bioinformatics Institute (EBI)	http://www.ebi.ac.uk/microarray-as/ae/
BASE	Lund University	http://base.thep.lu.se/
ChipDB	Whitehead Institute for Biomedical Research/MIT Centre for Genome Research	http://chipdb.wi.mit.edu/chipdb/public/
Dragon	Johns Hopkins University	http://pevsnerlab.kennedykrieger.org/dragon.htm
ExpressDB	Harvard University	http://arep.med.harvard.edu/ExpressDB/
GeneDirector	BioDiscovery	http://www.biodiscovery.com/
GEO	National Cancer for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/geo/
GXD	The Jackson Laboratory	http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml
Longhorn	UT Austin	http://www.longhornarraydatabase.org/
mAdb	National Cancer Institute (NCI)	http://madb.nci.nih.gov/
maxdSQL	The University of Manchester	http://www.bioinf.man.ac.uk/microarray/maxd/
Oncomine	Compendia Bioscience	http://www.oncomine.org
PUMAdb	Princeton University	http://puma.princeton.edu/
SMD	Stanford University	http://smd.stanford.edu/
SAGE	Johns Hopkins Oncology Centre	http://www.sagenet.org/
vMGV	Ecole Normale Supérieure	http://www.transcriptome.ens.fr/ymgv/

The interactions between DNA, RNA and proteins are complex and can be divided into three broad categories: metabolic, signalling and regulatory networks (Lewin, 2008; Lodish et al., 2003).

Gene expression involves the complex process of transcription of DNA (genes) to messenger ribonucleic acid (mRNA) and the translation of mRNA into proteins. In gene expression, it is the relative amount of mRNA which represents the activity of a specific gene. The knowledge about gene expression, specifically, which genes are differentially expressed, where and when in the cell, helps in understanding the function of cells and their development at the molecular level in an organism. The regulation of gene expression happens via genetic regulatory networks consisting of the interaction of genes (DNA), RNA, proteins and small molecules and their mutual regulatory interaction. DNA microarrays are essential tools for the analysis the expression of many genes simultaneously using information at the transcriptional level. Microarrays can be manufactured by two major approaches: complementary DNA (cDNA) microarray and Oligonucleotide chips. cDNA arrays, developed by Stanford University, are fabricated by robotic spotting on glass slides; and oligonucleotide arrays, produced by Affymetrix, are fabricated by photolithographic chemistry and light-directed chemical synthesis on small glass plates (Allison et al., 2006).

Affymetrix GeneChip high density oligonucleotide gene expression arrays (Affymetrix® and GeneChip®, <http://www.affymetrix.com/index.affx>) are one of the most well-known arrays and have been widely used in biomedical research. Each gene is represented on this array by 11-20 different probe pairs called a ‘probe set’ (Figure 2.1). Each probe consists of 25 nucleotide bases and each probe pair has two components: perfect match (PM), which is designed to match the specific sequence and mismatch (MM), located at the 13th base and intended to measure noise caused by non-specific binding (NSB) (see Figure 2.1). The expression level of a gene comes from the whole probe set, as the average difference between PM and MM.

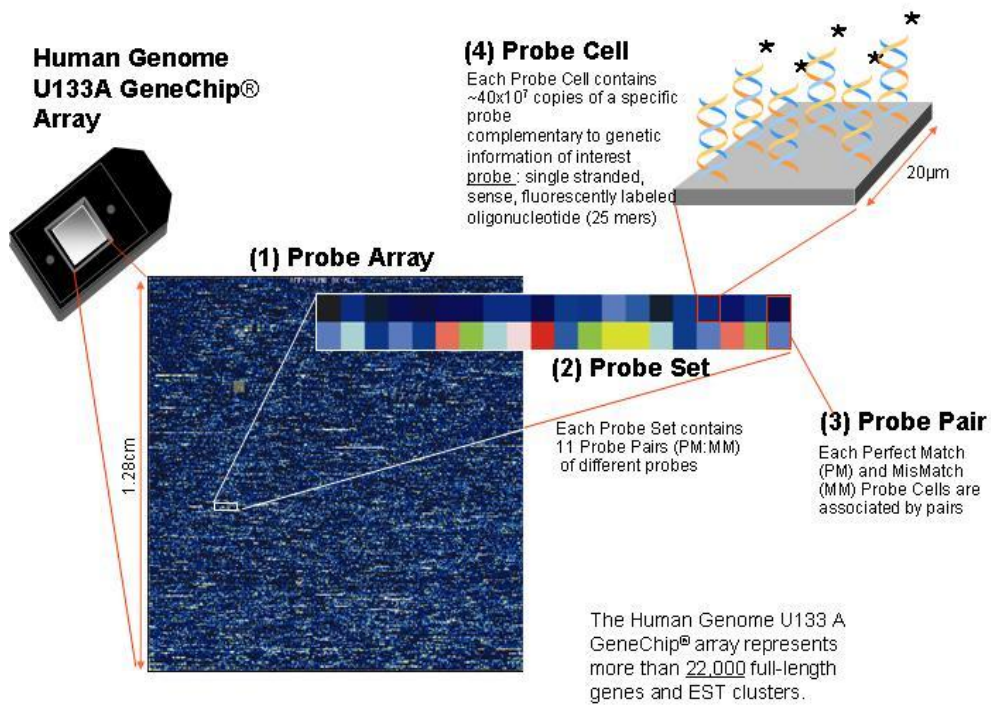


Figure 2.1: Schematic representation of the Affymetrix GeneChip® expression analysis system in use

The Probe Array contains ~22,000 Probe Sets on a surface of $\sim 1.2 \text{ cm}^2$.

Source: <http://www.weizmann.ac.il/>

Specifically, the average difference is the intensity of the whole probe set and is calculated as the average of the differences between the intensities of perfect match and mismatch for each probe pair, as shown in equation 2.1.

$$\sum_{j=1}^J PM_j - MM_j \quad (2.1)$$

where j is the number of the probe pair.

Many human genome arrays have been used in leukaemia research, of which the Affymetrix GeneChip® Human Genome U133 Plus 2 array is the focus of this study. The size of this array is 11 μm and it comprises 11 probe pairs per probe set and more than 54,000 probe sets. These represent more than 47,000 transcripts corresponding to 38,500 known human genes. Some genes are referred to more than once on the chip.

The process involved in measuring gene expression starts from GeneChip® or arrays containing oligonucleotide probes. Then the labelled cDNA or cRNA targets are hybridised to the array and washed and scanned by laser. Finally, the expression of the genes is measured in the form of the fluorescent intensity of the scanned image of the hybridised probes. The intensity levels of gene expression vary from the lowest to the highest, represented by dark blue, blue, light blue, green, yellow, orange, red and white (Figure 2.1).

The Affymetrix GeneChip® is analysed by the Affymetrix GeneChip Operating System Software (GCOS), which generates the cell intensity (.CEL) file from the image data file. This CEL file provides information about position and intensity of each probe on GeneChip®.

Microarray data (intensity images) are analysed and presented as gene expression matrices whose rows denote genes and columns denote samples. Gene expression represents the interested condition, for instance, healthy or diseased. A time series gene expression matrix is represented as $n \times m$ matrix where n represents genes and m represents time points: matrix

$M = [g_{ij}]$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$, where g_{ij} denotes the expression level of i^{th} gene at j^{th} time point is shown below (Bandyopadhyay, Maulik, & Wang, 2007):

$$M = \begin{bmatrix} g_{11} & g_{12} & \dots & g_{1m} \\ g_{21} & g_{22} & \dots & g_{2m} \\ \dots & \dots & \dots & \dots \\ g_{n1} & g_{n2} & \dots & g_{nm} \end{bmatrix} \quad (2.2)$$

The raw gene expression matrix retrieved from the imaging process consists of noise, missing values and systematic variations from biological experimental process. Therefore, data pre-processing is needed. In the next section, data pre-processing and the identification of differential gene expression are discussed.

2.3.2 Microarray Normalisation

After the experiment has been conducted and the raw intensity data obtained, data pre-processing needs to be carried out to reduce the systematic sources of non-biological variation (Amaratunga & Cabrera, 2004). Pre-processing, which amalgamates multiple probe signals into a single expression measure, called normalisation, is a common first step in data processing (Lim, Wang, Lefebvre, & Califano, 2007).

Microarray normalisation is one of the current active research areas that have produced an increasing number of available methods. Normalisation has three main steps: (i) background correction, which removes background noise from signal intensities, (ii) normalisation, which is intended to remove non-biological variability between arrays and make distributions across arrays identical, and (iii) summarisation, which gives a single expression measure to each probe set on the array.

The most commonly used methods are MAS5.0, RMA (Robust Multichip Average) and GCRMA (GeneChip RMA). MAS5.0 uses MM probes to adjust the PM probes for probe-specific nonspecific binding for background correction, uses a baseline array and scales all the other arrays to have the same mean intensity for normalisation and uses tukey biweights for summarisation (Affymetrix, 2002). RMA (Irizarry, Hobbs et al., 2003) uses a global correction, quantile normalisation and median polish summarisation. The GCRMA (GeneChip RMA) (Z. Wu & Irizarry, 2004) was developed from RMA and differs from the RMA in its uses of the probe sequence information for background correction. There are many more available pre-processing methods and new ones are being developed. Some of them are summarised in Table 2.4. The question now is how to select the most appropriate and reliable method to address a particular biological question from a specific dataset. A criterion has been introduced to use as an indicator; it is the capability to detect differentially expressed genes. It can be described in two terms: the precision (specificity/variance) and accuracy (sensitivity/bias) (Irizarry, Wu, & Jaffee, 2006). To help find the best method for each application, there is a benchmark; a web-based tool to assess several methods based on the same data and to obtain a summary statistical report, which helps scientists to choose the best pre-processing method for their tasks (Cope, Irizarry, Jaffee, Wu, & Speed, 2004).

There still are many questions with regard to the normalisation process: what is the most important normalisation stage? One study emphasised that the main factor that makes each method different is background correction (Irizarry et al., 2006). In contrast, recent research argues that different background corrections have insignificant effects on the correlation between methods (N. Jiang et al., 2008). What is the best normalisation method? Jiang et al (2008) indicated that GCRMA and MAS 5.0 performed poorly compared with others, while Wu et al (2004) concluded that GCRMA performed better for genes with lower expressions and RMA for higher ones. Lim et al (2007) pointed out that GCRMA and RMA show better performance than MAS5.0. However, GCRMA creates artificial correlations during the normalisation process. Recent research has indicated that the positional dependent nearest neighbour (PDNN) (L. Zhang, Miles, & Aldape, 2003) is the best normalisation method among seven different methods. They used sensitivity, reproducibility and consistency as the criteria to evaluate the performance of the seven normalisation methods (N. Jiang et al., 2008).

Table 2.4: Description of some pre-processing methods, adapted from Irizarry et al.,(2006) and N. Jiang et al., (2008).

Method	Background correction	Normalisation	Summarisation	Citation
dChip	MM intensities are subtracted	Spline fitted to rank invariant s	A multiplicative model is fitted	Li and Wong, 2001
GL	None	Loess fitted to subset	As RMA	Freudenberg, 2005
GCRMA	Based on probe sequence	As RMA	As RMA	Wu et al., 2004
MAS5.0	Spatial effect and MM subtracted	Scale normalisation	A robust average (Tukey biweight)	Affymetrix, 2002
MBEI (PM-MM)	MM intensities are subtracted	Invariant set	Multiplicative model	Li and Wong, 2001
MMEI	None	Linear mixed model	A linear mixed model is fitted	Deng et al., 2005
PDNN	Model is fitted accounting for background and specific signal	Quantile	Specific and non-specific binding effects are estimated using free energy model	Zhange et al., 2003
RMA	A global correction	Quantile	A robust linear model (median polish)	Irizarry et al., 2003

Another issue that needs to be taken into account is that data, which are retrieved from different labs using the same platform, are different (Irizarry et al., 2005). Moreover, using a single platform with different normalisation methods led to large variability in the results (Stafford & Tak, 2008). Different studies show different conclusions and there is no consensus at present on the most suitable method for all data. Nonetheless, all differentially expressed gene lists from each method are unique but significant (N. Jiang et al., 2008). Normalisation has a consequential effect on the final microarray data analysis results (Irizarry et al., 2005). Normalisation methods alter in how the correlation structure from the data is revealed and, in turn, affect the accuracy of inference of cellular networks (Lim et al., 2007). Genes that have been determined to be differentially expressed are highly dependent on the normalisation method (Steinhoff & Vingron, 2006). This research uses RMA, as this method was selected by the original authors.

Robust Multichip Average (RMA)

The Robust Multichip Average or Robust Multi-array Average (RMA) (Irizarry, Bolstad et al., 2003; Irizarry, Hobbs et al., 2003) is one of the normalisation approaches that convert probe level data (CEL files) into a gene expression measure. RMA starts with the background correction, a non-linear correction on each chip, where the background signal (BG) intensity (optical noise and non-specific binding) is subtracted from probe level signal- perfect-match (PM) values. Quantile normalisation is then applied across chips in order to equalise probe intensities and, finally, the summarisation process creates a single expression measure for each probe set. RMA uses the median polish summarisation process.

The RMA algorithm can be formulated as:

$$\log_2 (PM_{ij} - BG) = e_i + \alpha_j + \varepsilon_{ij}, i = 1, \dots, I \text{ and } j = 1, \dots, J \quad (2.2)$$

where i represents array and j represents probes, e_i denotes the log scale expression of array i and α_j denotes affinity probe effect and ε_{ij} denotes an independent identically distributed error term with mean zero.

For any probe-pair, perfect match (PM) intensity consists of signal and noise and is defined by:

$$PM = O_{PM} + N_{PM} + S \quad (2.3)$$

where O denotes optical noise, N denotes Non specific binding (NSB) noise, both O and N follow a log-normal distribution with mean μ and standard deviation σ , S denotes signal component or a quantity proportional to RNA expression (the quantity of interest) in the form of an exponential distribution.

Quantile normalisation, a non-parametric method introduced by Terry Speed's group with the basic assumption that all samples have almost the same gene abundant distribution. Each chip is normalised by computing the quantile value of the distribution of probe intensities and then transforming this value to reference the chip quantile's value. Equation (2.4) depicts the quantile transformation:

$$x_{norm} = F_2^{-1}(F_1(x)) \quad (2.4)$$

where F_1 defines the distribution function of the actual chip intensities and F_2 defines the distribution function of the reference chip.

Median polish is a statistical method proposed by John W. Tukey (Tukey, 1977). It is an additive-fit model for a two-way layout. Tukey's median polish decomposes data into:

$$\text{Data} = \text{all effect} + \text{row effect} + \text{column effect} + \text{residual} \quad (2.5)$$

This method subtracts the median of each row from the row values as well as from each column from the column values and repeats this until convergence occurs (median = 0); these are called row effect and column effect variables.

RMA can be performed through at least three available software packages: affy package (Bioconductor project), RMAExpress and Matlab. Affy is a software package implemented in R language, which can be downloaded from the Bioconductor project website (<http://www.bioconductor.org>). Bioconductor is a freely available source for users and developers to analyse genomic data. RMAExpress 1.0.2 release is cross-platform software to obtain gene expression summary values using the Robust Multichip Average expression from Affymetrix Genechip® CEL files. This software is open source and can be freely downloaded from the website (<http://rmaexpress.bmbolstad.com/>). Matlab, developed by the Mathworks, is a numerical computing environment that enables researchers to perform computationally intensive applications, including matrix manipulation, plotting and implementation of algorithms, faster than with traditional programming languages. The Bioinformatics add-on toolbox in Matlab provides RMA commands for pre-processing Affymetrix microarray data.

After the normalisation process, the next step is to address cancer research problems by identifying differentially expressed genes that may likely be involved in a particular biological mechanism and could become targets for therapeutic intervention. The process of identifying differentially expressed genes can be carried out by two main steps: first, set up threshold criteria for identifying novel genes by using fold change and statistical methods and, second, cluster novel genes into similar expression patterns (clustering or classification); this is explained in detail in the following section. Normally, a gene with more than two fold changes is considered significant or differentially expressed. Fold changes and traditional statistical methods have been used to identify differentially expressed genes, for example, t-test, significance analysis of microarrays (SAM) (Chu, Narasimhan, Tibshirani, & Tusher, 2002) analysis of variance (ANOVA) and data clustering. Identifying which method to use is essential because it may affect the novel gene lists. Comparisons between some of these methods can be found in Jeffery, Higgins, and Culhane, (2006) and Tibshirani and Witten (2007).

In the analysis of differentially expressed genes, a subset of genes, also called a novel gene set, is selected from the expression matrix, which is strongly associated with the samples. Identification of differentially expressed genes can be undertaken through analysis of gene expression levels. Differentially expressed genes can be very informative but they do not reveal the whole underlying biological mechanism and process. Most studies focus on finding differential expression patterns between healthy and cancer samples, different stages of tumour or before and after treatment. Genes with similar expression patterns may be called co-regulated genes, which imply that they are regulated together and these genes may help in understanding the underlying biological process. Normally, there are two main basic patterns which are called *underexpression* (down-regulation) and *overexpression* (up-regulation). Overexpressed genes are genes that, on the one hand, have higher expression values when two samples are compared; for example, cancer (target) and healthy. On the other hand, underexpressed genes have lower expression values in target than reference samples (Dubitzky, Granzow, Downes, & Berrar, 2003).

In summary, differentially expressed genes can be detected by using different pre-processing processes followed by traditional statistical methods but the results from each method contain false gene sets and true gene sets. It is difficult to produce consistent groups of differentially expressed genes even from the same platform but different subsamples. This issue is one of the limitations of high-throughput technology that should be taken into account. Gene expression analysis is still an on-going research field; there is no one suitable method for all data but a combination or evolution of existing/new statistical and clustering methods may help understand the global mechanisms behind biological phenomena. Understanding how genes work in a human cell still remains a major challenge for scientists to overcome. There are many ways to extract biological information from microarray data including identifying differentially expressed genes, identifying global patterns of gene expression and determining the biological meaning from each gene and their network (Santos & Liu, 2007). Therefore, in the next section, clustering methods are described, explaining how to identify similar patterns of differentially expressed genes.

2.4 Clustering

There are two main groups of gene classification processes: supervised classification and unsupervised clustering. Normally, clustering has been applied to find gene expression patterns after obtaining the set of novel genes. Clustering gene expression data helps to increase the understanding of gene function, gene regulation, cellular processes and subtype of diseases, etc. (D. Jiang et al., 2004). Clustering methods have been used in biomedical applications, including cancer, in several aspects: gene expression, sequences analysis, gene networks and protein-protein interactions. Clustering or unsupervised classification approaches give a cluster of distinct and highly similar genes (or co-expressed genes) but without a predefined cluster. Co-expressed genes may have similar functions and be active in the same cellular process. Furthermore, a strong correlation between gene expression patterns can define co-regulated genes. Clustering has four major parts including feature selection or extraction, clustering algorithm design or selection and implementation, cluster validation and interpretation of results (Xu & Wunsch, 2005).

There are many reviews of previous clustering methods for microarray data (D. Jiang et al., 2004; Kerr, Ruskin, Crane, & Doolan, 2007; Madeira & Oliveira, 2004; Shamir & Sharan, 2001; Tibshirani et al., 1999; Xu & Wunsch, 2005; Yin, Huang, & Ni, 2007), specifically, there are many recently developed clustering tools including neural networks and fuzzy clustering (Bandyopadhyay, Mukhopadhyay, & Maulik, 2007; Herrero, Valencia, & Dopazo, 2001; Madeira & Oliveira, 2004; Pal, Aguan, Sharma, & Amari, 2007; L. Wang, Chu, & Xie, 2007; Y. P. Wang et al., 2008). There are many unsupervised clustering and supervised classification methods. Evaluation and comparison of gene clustering methods can be found in Chou, Zhou, Kaufmann, Paules and Bushel, (2007) and Thalamuthu, Mukhopadhyay, Zheng, and Tseng (2006). Thalamuthu et al. (2006) evaluated and compared six clustering methods including hierarchical clustering, K-Means, partitioning around medoids (PAM), self organising maps (SOM), model-based clustering and tight clustering (Tseng & Wong, 2005) for cluster synthetic and experimental datasets. They introduced the weighted Rand index as the criterion to evaluate the performance of the six clustering methods with simulated data and annotation prediction and functional prediction accuracy for a real dataset. Tight clustering and model-based clustering perform better than other clustering methods in

simulated and experimental data, whereas, hierarchical clustering and SOM perform worse than the other methods.

Clustering methods and the distance measure used can vary the final novel gene set in each cluster. Different distance measures represent different aspects of rgw data. Clustering approaches have been applied to first microarray analysis ALL data by Golub (1999), specifically, who identified 50 informative genes by using SOM clustering to differentiate ALL from AML. In addition, identifying ALL subgroups using neural networks including support vector machines (Yeoh et al., 2002), identifying genes that distinct four treatment types (Cheok et al., 2003); predicting relapse and treatment response (Willenbrock et al., 2004); and other studies are shown in Tables 2.1 and 2.2.

The focus of the research reported in this study is on clustering methods for short time series gene expression data. Complex gene networks govern and orchestrate the biological systems of a cell. Understanding the underlying of these networks is now possible using a microarray time series datasets that measure at several time points. Knowledge obtained through this process may help to understand the underlying mechanisms of specific diseases, especially cancer, and may lead to the identification of novel chemotherapeutic gene targets. Previous studies have used time series data for particular purposes including: (i) meaningful temporal gene expression patterns in the data, (ii) specify genes that belong to each pattern, (iii) relationships between gene groups, and (iv) model development to depict gene groups' relationships (Famili et al., 2004).

Time series gene expression is the measurement of gene expression of particular samples at particular time points. The co-expressed genes may have similar gene expression patterns over time. This group of genes can be used for further analysis of inferring gene networks. Time series data have two main attributes: order and interdependency, but these two essential points are normally disregarded by conventional clustering methods. Time series expression data can be used to enhance the understanding of the underlying biological mechanisms through the process of modelling gene networks. The first applications of time series clustering methods used the Unweighted Pair Group Method with Arithmetic Mean

(UPGMA) with Pearson's correlation to cluster gene expression in the budding yeast, *Saccharomyces cerevisiae* (Eisen, Spellman, Brown, & Botstein, 1998).

Many conventional clustering methods have also been applied to time series gene expression analysis. Novel algorithms have also been proposed to investigate time series gene expression; for example, a novel pattern based clustering method by Phan et al., (2007). Recently, some methods have been developed especially for time series data which apply or modify the existing methods to analyse time series data (Corduas & Piccolo, 2008; Das, Kalita, & Bhattacharyya, 2009; Douzal-Chouakria, Diallo, & Giroud, 2009; Dzeroski, Gjorgjioski, Slavkov, & Struyf, 2007; Kriegel, Kroger, Pryakhin, Renz, & Zherdin, 2008; Magni, Ferrazzi, Sacchi, & Bellazzi, 2008; Savvides, Promponas, & Fokianos, 2008; Summa, Steyaert, Vautrain, & Weitkumat, 2007). Many clustering algorithms have been applied to these unique microarray time series data and they include: the Merge SOM (MSOM) (Hammer, Micheli, Neubauer, Sperduti, & Strickert, 2005; Strickert & Hammer, 2005), Growing Recurrent Self Organising Map (GRSOM) (Yeloglu, Heywood, & Malcolm, 2007) and Self Organising Maps and Particle Swarm Optimisation (Xiao, Dow, Eberhart, Miled, & Oppelt, 2003). A two-step regression-based approach called maSigPro (microarray Significant Profiles) has been introduced to identify differentially expressed genes in time series profiles (Conesa, Nueda, Ferrer, & Talon, 2006). Difference-based clustering has also been proposed (J. Kim & Kim, 2007). In addition, hybrid principal component and neural networks (PCA-NN) (Ao & Ng, 2006), multi-step approaches (Amato et al., 2006), and clustering software such as TimeClust have been used. TimeClust is freely available to download from its website (Magni et al., 2008). A review of time series microarray data analysis can be found in Androulakis, Yang, and Almon, (2007), Bar-Joseph, (2004) and Warren Liao (2005).

Analysis of short time series has recently emerged and previous clustering methods, including the methods that have been developed for non time series and long time series, data do not perform well on short time series data with limited time points (Xuewei, Ming, Zheng, & Chan, 2008). Most currently available methods focus on long time series data and, at the time of writing, only Short Time series Expression Miner or STEM was specifically designed for short time series data.

As mentioned above, there are many clustering methods with different advantages. However, there is no single method that is able to visualise the patterns of all data, assign one gene to multiple groups that is specifically designed for short time series data. This research aims to extract intrinsic biological patterns underlying time series data and find meaningful biological knowledge from the combination of four prominent and emergent clustering methods. The selected clustering methods used in this research include Self organising maps (SOM), Emergent self organising maps (ESOM), Fuzzy clustering by Local Approximation of MEMbership (FLAME) and Short time series expression miner (STEM), as explained in detail in the next section.

2.4.1 Self Organising Maps (SOM)

Self organising maps were introduced by Coonan (Coonan, 1997). This method projects high dimensional data into a one or two-dimensional rectangular or hexagonal grids and organises data on similarity basis. The SOM algorithm starts with a predefined map topology (rectangular or hexagonal) and then input vectors (gene expression profiles each containing expression values of selected genes) that are linked by weights to the map topology based on a distance measure. Initial weights are random and are adjusted after each iteration. The weights of the node closest to the input vector (the winner) and its neighbours are updated until convergence. The principles of SOM are that weight vectors are initially set randomly for each neuron and, during learning, input data (gene expression profiles) are compared with neuron weight vectors using a distance measure such as Euclidean distance, and weights of the best match neuron and its neighbours are updated until nearby neurons represent genes with similar expression profiles. The weight of the winner neuron (with the smallest distance to the input) and its nearest neighbour neurons are updated after presentation of each input vector (or a batch of input vectors) until the map is converged when weight change is negligible (Samarasinghe, 2006). A disadvantage of SOM is the user needs to predefine many parameters, especially the number of nodes.

SOM algorithm (X. Liu & Kellam, 2003):

1. Initialise the topology and size of the output map.
2. Initialise random weight values over the interval $[0,1]$ as well as the learning rate η and the neighbourhood size, r .
3. For each input node:
 - Present new input vector, v .
 - Calculate the Euclidean distance between input vector and weight vector of each node in the output map. The shortest distance (minimum distance) declares the winner node, c .

$$c = \min \sqrt{\sum_{k=1}^N (v_k - w_{jk})^2}, j = 1, 2, \dots, M \quad (2.12)$$

where N is number of inputs (e.g. samples, patients) in the input vectors (e.g. gene expression profile), M is number of nodes.

- Modify weight. w , learning rate η and N_c , the neighbourhood surrounding the winner node c . For each node $j \in N_c$, new weight can be calculated as follows:

$$w_j^{new} = w_j^{old} + \eta S(d, t) [v_i - w_j^{old}] \quad (2.13)$$

- Incrementally decrease both the neighbourhood size and learning rate $h(t)$ and repeat Step 3 until convergence.

Data clusters (density of data) from SOM can be visualised by U-Matrix. U-Matrix (unified distance matrix) is a coloured map representing the pair wise distances between neuron weights; the colour of the map varies according to distance. There are two distinct areas on this map where the clusters are separated by a dark gap (the largest distance between neurons) and each light area represents a cluster (a smaller distance between neurons). SOM also can be visualised by maplets, whose colour represents the spectrum of values of a specific input variable (e.g. expression of a gene across all patients). In some program outputs, dark dots inside the neurons denote the number of input vectors (expression profiles) represented by each neuron. No dot means no input falling into that particular node.

SOM have been used to analyse gene expression data (Fernandez & Balzarini, 2007; Nikkilä et al., 2002; Tamayo et al., 1999; Toronen, Kolehmainen, Wong, & Castren, 1999). There are limited numbers of public software and tools for gene expression analysis by SOM. For example, the SOM_PAK and SOM_Toolbox 2.0 (for Matlab) programs available at the Laboratory of Computer and Information Science (CIS), Department of Computer Science and Engineering, Helsinki University of Technology (<http://www.cis.hut.fi/somtoolbox/>) (Vesanto, Himberg, Alhoniemi, & Parhankangas, 1999). Some commercial SOM software is available including Matlab with its neural network toolbox and Synapse (<http://www.peltarion.com/products/synapse/>). SOM are used to extract information from large-scale data but SOM only represent an overview of gene expression data and, therefore, further investigation of the map is needed. SOM has some limitations, being restricted to a predefined map structure, therefore, more powerful versions of SOM have been developed, specifically, Emergent Self Organising Maps (ESOM) (Ultsch & Morchen, 2005). Maps in ESOM can freely follow data and define its own form.

2.4.2 Emergent Self Organising Maps (ESOM)

Emergent Self Organising Maps (ESOM) is based on the concept of emergence. ESOM allows the emergence of network structure making it possible to observe overall data structure in high level of detail. ESOM topology is a toroid map, and it is used to solve border effects of classical SOM. ESOM can be visualised in three patterns: distance-based visualisation (U-Matrix), density-based visualisation (P-Matrix), and distance- and density-based visualisation (U*-Matrix). P-Matrix is used in this study. P-Matrix (Ultsch, 2003a, 2003b) works well for gradually changing density and overlapping clusters. This density-based measure represents the density in data space sampled using the Pareto Density Estimation (PDE) (Ultsch, 2003b).

U-Matrix shows the distance relationship of the input data in the data space using the average distance to neighbours of each neuron. A large U denotes a larger distance to neighbouring neurons while a smaller U reveals shorter distances. P-Matrix is compatible with U-Matrix: U-Matrix uses local distances to give insights into the distance structure; in contrast, P-Matrix uses density values of neurons (number of input vectors represented by each neuron) to reveal the underlying density structures of high dimensional data. P-Matrix in ESOM uses data density calculated from Pareto Density Estimation (PDE) as P-value at the coordinates of neuron, n_i . Dense areas in a map contain neurons with large P-values while small P-values represent sparse areas. For neuron, n , the density of data space can be defined as:

$$P(n)=p(w(n), X), \quad (2.14)$$

where $p(x, X)$ denotes an empirical density estimation at point x (i.e $w(n)$) in the data space X .

For each neuron, P-Matrix shows the number of input vectors in a Pareto radius of hypersphere (Pareto sphere). Under the assumption of multivariate mutually independent Gaussian standard normal density distribution (MMI), for at least two clusters in the data, the average radius can be calculated by:

$$r_u = \frac{1}{2} cd\chi_d^2(p_u) \quad (2.15)$$

where $cd\chi_d^2$ denotes the Chi-square cumulative distribution function for degrees of freedom d , and p_u is the average probability. For one dimensional data, the 18th percentile is the closet percentile of distance to r_u . This is called the Pareto radius, r_p . “The density measured at point x using the number of points inside a hypersphere with radius r_p (Pareto radius) is called Pareto (probability) Density Estimation PDE(x)” (p. 2, (Ultsch, 2003b)).

The ESOM method is a promising knowledge discovery approach that has the potential to be applied to microarray data analysis. ESOM has been used for clustering real data and the performance is superior to traditional SOM (J. Poelmans, P. Elzinga, S. Viaene, M. M. Van Hulle, & G. Dedene, 2009b). For example, a study by Ultsch and Morchen (2005) applied ESOM with an ALL dataset by Golub et al. (1999) and concluded that ESOM discovered new leukaemia sub-classes.

2.4.3 Fuzzy clustering by Local Approximation of MEMbership (FLAME)

Most of the traditional clustering methods are hard clustering methods which allow one gene to belong to only one cluster. Fuzzy clustering was introduced to overcome this limitation, where one gene can be clustered to more than one cluster. Each gene is assigned a cluster membership which indicates the degree of belonging in each cluster. Well-known methods include Fuzzy C-Means (Bezdek & Ehrlich, 1984). Fuzzy C-Means has been shown to give better clustering results even when the data contained outliers and overlapping areas, when compared with SOM, K-Means and hierarchical clustering (Mingoti & Lima, 2006).

A recently developed method is Fuzzy clustering by Local Approximations of MEMberships (FLAME) (Fu & Medico, 2007). FLAME is implemented by Gene Expression Analysis Studio (GEDAS) (<http://sourceforge.net/projects/gedas>) software.

The FLAME algorithm has three main steps (Fu & Medico, 2007):

1. Calculating similarities of expression patterns using Pearson's correlation, and then creating a connected graph of all K-Nearest Neighbours (KNN) of each expression vector. For each expression pattern, density, which indicates the number of neighbours within a specified distance, is calculated, classifying it into one of three groups:

- (i) Cluster supporting object (CSO) which has higher density than their neighbours,
- (ii) Outlier, which has lower density than its neighbours and a predefined threshold, and
- (iii) The Rest.

2. Starting with assigning initial memberships to each group. Each CSO, referring to a cluster, is given fixed and full membership to itself. All outliers making another cluster are given the fixed and full membership in the outlier group. The Rest are given the equivalent or same memberships to all cluster and the outlier groups. This step is called Local Approximations of Memberships. Local Approximation of fuzzy membership of the three groups is determined and each object is updated by a linear combination of the fuzzy memberships of its nearest neighbours or Local Approximations of Memberships. A membership degree of vector, x in cluster, i is $p_i(x)$ and is given by:

$$x: p(x) = (p_1(x), p_2(x), \dots, p_M(x)), \quad (2.16)$$

where $0 \leq p_i(x) \leq 1$; $\sum_{i=1}^M p_i(x) = 1$ and $M = |X_{cso}| + 1$, X_{cso} is the set of clusters supporting objects with Local Maximum Density.

Membership vector of each object is updated by the summation of x 's nearest neighbours' memberships, as in Equation (2.17).

$$p^{t+1}(x) \approx \sum_{y \in KNN(x)} w_{xy} p^t(y) \quad (2.17)$$

where $p^t(x)$ fuzzy membership vector of object x at iteration t and w_{xy} is the weight vector between objects x and y .

Each iteration process attempts to reduce the Local (Neighbourhood) Approximation Error $E(\{p\})$, the difference between the approximation of membership vectors in the current and previous iterations, defined by:

$$E(\{p\}) = \sum_{x=X-CSO-Outlier} \left\| p(x) - \sum_{y \in KNN(x)} w_{xy} p(y) \right\|^2 \quad (2.18)$$

After finishing the calculation of the Local Approximation of fuzzy membership, the process moves to step three.

3. Clusters can be contracted into two categories based on fuzzy memberships: one gene to one cluster or one gene to multiple clusters.

2.4.4 Short Time series Expression Miner (STEM)

STEM is a software program for analysing short time series gene expression data (Ernst & Bar-Joseph, 2006). Many clustering algorithms have been applied to time series data in general and more details can be found in Wang et al., (2008). Among available methods, the Short Time series Expression Miner (STEM) (Ernst & Bar-Joseph, 2006) was particularly created to analyse short time series gene expression data. This method uses the change in direction and magnitude of the inputs with time. STEM is a Java-based program; STEM (<http://www.cs.cmu.edu/~jernst/stem/>) version 1.3.4 was used in this study.

STEM algorithm

1. Constructing model profiles. The first step of STEM is to create reference model profiles to represent all possible gene expression changes over time in terms of numbers of possible units of change. This can be done by discrete changes between two consecutive time points, as no change, up or down (e.g. ± 1 , ± 2 , ± 3 etc.). The first time point is always zero, and then it can stay the same or increase or decrease. The maximum changing units is c . The number of all possible model profiles ($|P|$) can be then formulated as:

$$|P| = (2c + 1)^{n-1} \quad (2.19)$$

where n denotes the number of time points in a profile. For three time points, $n = 3$, and for $c = 3$ (i.e., successive time points either go up or down for a maximum of three units), there are 49 distinct model profiles. If the number of possible model profiles are too many to be viewed, for example, $n = 5$ and $c = 3$ which produces 2,401 possible model profiles, in the next step, possible model profiles need to be reduced by selecting m distinct profiles. The distinct model profile is selected, based on maximise the minimum the pair wise distance (d) between these distinct profiles as expressed in equation 2.20.

$$\max_{R \subset P, |R|=m} \min_{p_1, p_2 \in R} d(p_1, p_2) \quad (2.20)$$

2. Enumerating statistically significant profiles. The experimental time series gene expression profile of each selected gene is assigned to the closest reference based on correlation distance, and the total (actual) number of genes represented by each reference profile is obtained. The number of expected genes in each profile is then calculated. The number of expected genes is the possible number of genes when data are random and computed by random permutations of the gene expression values. The actual number of genes allocated versus the number of expected genes is then used to identify statistical significance. 3. Clustering significant model profiles. After obtaining the significant gene profiles, they are grouped according to their similarity. The idea of grouping is based on the premise that significant profiles that should be grouped together could have been separated due to noise and other effects. This clustering step finds true and robust groups so that genes with similar gene expression profiles are grouped together.

In conclusion, clustering results from different clustering methods on the same dataset can vary. Clustering methods may force the data to cluster even when there is no similar group in a given dataset and these create false positives and distort the structure of existing clusters. Therefore, clustering is only the first step in the analysis of gene expression data, and careful interpretation and further in-depth analysis is needed. For this purpose, using several approaches to clustering and comparing the results can be useful in assessing the robustness of gene sets. Furthermore, clusters of differentially expressed genes containing upregulated or downregulated genes can be used to identify complex patterns in the form of gene networks. Network/pathway analysis is explained in the following section in order to shed light on how gene networks are constructed.

2.5 Network/Pathway Analysis

After gene expression analysis, the next question is how to extract biological meaning from differentially expressed genes. The apoptotic pathway is an essential pathway in understanding cancer treatment. There are two main mechanisms for apoptosis pathways: the extrinsic (death-receptor pathway) and the intrinsic (mitochondrial pathway) (Igney & Krammer, 2002). The extrinsic pathway is initiated when a cell responds to an external apoptotic signal through a death receptor, while the intrinsic pathway is initiated with a trigger of cellular damage or stress (e.g. DNA damage and heat shock) through the mitochondria. Cells are normally tightly regulated by these pathways; however, in cancer, there are many important malfunctioning genes in the pathway (Folarin & Bioinformatics, 2003). This is essential and important knowledge in a clinical study as most chemotherapy induces apoptosis through DNA damage; and resistance to therapy often involves resistance to apoptosis (Folarin & Bioinformatics, 2003). Transformation of normal cells to cancer cells is caused by a complex interaction series of multiple networks and pathways, in particular, apoptosis pathways. There are many genes and gene products involved in these pathways. Unravelling these pathways (identifying genes and inferring their interaction through their gene networks) may lead to a better understanding of cancer treatment through apoptosis pathways.

How to gain a better understanding of gene networks is currently a huge challenge for scientists due to a lack of understanding about network structures. Microarray technology has had fast paced growth which, in turn, has created a large amount of publicly available data. Deciphering gene networks from the rapidly growing microarray expression databases has been shown to be a very promising approach in cancer treatment. A current research trend in bioinformatics is to construct “physical” networks, for example, gene regulation pathways, from “conceptual” networks; for example, co-expression information (Benson & Breitling, 2006). These tasks require a computational systems biology approach. Pathways and biological gene networks can be inferred from the analysis of microarray data by grouping genes (also called “modules” or “gene sets”). The concept is to find common patterns from similar microarray experiments. The changes in gene expression are controlled by gene networks. Gene networks can be identified by using time series gene expression data which indicates genes that are turned on or turned off at particular times in specific conditions and tissues.

Recently, inferring gene networks from gene expression profiles has become a new challenge for scientists. Various statistical and machine learning methods for inferring gene networks have been introduced only in the last decade. Therefore, it is difficult to evaluate or validate the performance of these approaches that have already been proposed (Bansal, Belcastro, Ambesi-Impiombato, & di Bernardo, 2007; K.-H. Cho et al., 2007; D'haeseleer, Liang, & Somogyi, 2000; de Jong, 2002; van Someren et al., 2002). There are two main approaches to developing gene network inference algorithms: physical interaction, which aims to identify gene-to-sequence interaction (interaction between target genes and transcription factors) and influence interaction, which infers gene-to-gene interaction (finds the relationship among expressed genes) (Bansal et al., 2007). Many methods have been used to model/infer genetic pathways and interactions and these include Bayesian networks (Friedman, Linial, Nachman, & Pe'er, 2000; Schäfer & Strimmer, 2005), Dynamic Bayesian networks (Husmeier, 2003; Zou & Conzen, 2005), Boolean networks (Kauffman, Peterson, Samuelsson, & Troein, 2003), S-systems (Kimura et al., 2005), ordinary differential equations (ODEs) (Chen, Wang, Tseng, Huang, & Kao, 2005), Neural networks (Lee & Yang, 2008), Petri nets (Heiner, Koch, & Will, 2004), and Graphical Gaussian models (Toh & Horimoto, 2002). A more detailed overview of these methods can be found in K.-H. Cho et al., (2007), Christensen, Thakar, and Albert, (2007), Lee and Tzou, (2009), Markowitz and Spang, (2007), Schlitt and Brazma, (2007), Styczynski and Stephanopoulos, (2005) and van Someren et al., (2002). Some advantages and disadvantages of some selected methods are shown in Table 2.5.

Apart from the methods mentioned above, there are software programs from academic institutions and commercial enterprises that provide complex bioinformatics tools to construct networks/pathways from genes of interest. There are almost 170 online pathway databases for a range of biological processes, which can be divided into four main categories: metabolic, signalling, protein interaction and gene regulation (Cary, Bader, & Sander, 2005). A pathway database system was created for storing, managing, analyzing, visualising and querying biological pathways at multiple levels of detail. Research by Krishnamurthy et al. (2003) separated the databases into three groups of biological pathway: metabolic and biochemical; transcription, regulation and protein synthesis; and signal transduction.

Table 2.5: Advantages and disadvantages of some selected methods for modelling/infering gene networks

Methods	Advantages	Disadvantages
Mathematical Modelling		
Boolean Networks	<ul style="list-style-type: none"> • Computationally simple • Binary model assumes gene to be either on or off • Networks are inherently dynamic • Can explore large scale networks 	<ul style="list-style-type: none"> • Neglect intermediate transitions that are well-known and proven.
Bayesian Networks	<ul style="list-style-type: none"> • Statistical reference • Handle noisy data • Used with incomplete data • Able to add prior knowledge 	<ul style="list-style-type: none"> • Increase in computational complexity • Networks are inherently static but can be overcome by dynamic Bayesian networks; however, limited to small datasets
Nonlinear Ordinary Differential Equations (ODEs)	<ul style="list-style-type: none"> • More accurate models; describe network in great detail 	<ul style="list-style-type: none"> • Lack of <i>in vivo</i> and <i>in vitro</i> measurement of the kinetic parameters in the rate equations • Implicit assumptions not valid at molecular level
Machine Learning approaches		
Neural Networks	<ul style="list-style-type: none"> • Handle complex problems from discrete to continuous • Handle many variables and non-linear interactions • Flexible and adaptive learning • Reliable network prediction • Ability to retrieve all possible interactions between predictor variables • Can combine multiple training algorithms 	<ul style="list-style-type: none"> • Computational cost depends on chosen topology and learning algorithms • Optimal network topology is difficult to define • Neural networks is implicit and difficult to explicitly identify possible causal relationships
Genetic Algorithms	<ul style="list-style-type: none"> • Robust optimisers 	<ul style="list-style-type: none"> • GAs are very slow • Do not find the exact solution but always find the best possible solution

Examples of well-known pathway databases are: KEGG (www.genome.ad.jp/keg) (Kanehisa & Goto, 2000) and Biocarta (www.biocarta.com). Each database has different conceptualisations, and which one is the best for a problem depends on the purpose of the study. Network/pathway analysis software available include ASIAN (Aburatani, Goto, Saito, Toh, & Horimoto, 2005), Cytoscape (Shannon et al., 2003), GeneNet (Ananko et al., 2005), GeneNetwork (C. C. Wu, Huang, Juan, & Chen, 2004), Oncomine, The BiblioSphere Pathway Edition (BSPE) and Ingenuity Pathway Analysis software (IPA). The latter three are used in this research and are explained in detail in the following sections.

2.5.1 Oncomine

Oncomine (Rhodes et al., 2007; Rhodes et al., 2004) is a knowledge-based database curated from the existing literature of human cancer gene expression profiles and an integrated data-mining platform. The differentially expressed genes were analysed using t-statistics and corrected for measure of significance using false discovery rates. As of 22 June, 2009, there were 41 cancer types with 392 studies and 28,880 microarray experiments available for further analysis with integration another 18 bioinformatics resources including GEO, SMD, KEGG pathways, Gene ontology and Biocarta. Oncomine (i) collects microarray data from the original authors' website or downloads from them publicly available databases and then (ii) all datasets are transformed to log scale and median-centred for each array and normalised to one standard deviation, and next (iii) stores all data in the Oncomine database by re-naming all datasets to FirstAuthor_TissueTypeProfiled. All data are grouped into analyses of interest, for example, cancer tissue versus normal tissue, several molecular subtypes and treatment responses. Then the t-test is used to identify differentially expressed genes for two classes of expression profiles, while Pearson's correlation is used for multiclass comparisons. Both statistical methods are implemented using the R statistical computing package (<http://www.r-project.org>).

The top 50% of genes are then selected by using average linkage hierarchical clustering to represent coexpressed genes. Oncomine also analyses microarray data in terms of molecular concept analysis. Data can be retrieved from the Oncomine website for gene signatures of the top 1%, 5% or 10% of overexpressed and underexpressed genes from a selected analysis.

Genes in those lists are ranked using p values. For interaction network analysis, cancer networks are identified from known protein interaction databases or Human Protein Reference Database (HPRD) (<http://www.hprd.org>). Oncomine can be freely accessed for academic research through the website <http://www.oncomine.org>.

2.5.2 BiblioSphere Pathway Edition (BSPE)

The BiblioSphere Pathway Edition (BSPE) (Genomatix Software, Munich, Germany, <http://www.genomatix.de>) is a software program to analyse gene relationship networks and is claimed to be the only software that uses curated information from literature analysis with proprietary genome annotation and promoter analysis. The main database for BSPE is PubMed, combined with other data sources including Gene Ontology, MeSH, KEGG Pathway and Biocarta. When a user inputs data into BSPE and all possible results are retrieved, a z-score is used to identify meaningful genes. The z-score indicates the over- or underrepresented annotation from the input gene set based on the number of observed and expected annotations of the specific term.

The z-score defines the distance and direction of annotation term from its distribution mean based on standard deviation, which can be expressed as:

$$Z = \frac{(\text{observed} - \text{expected})}{\text{std.deviation}(\text{observed})}$$

In BSPE, the z-score is given by:

$$z = \frac{\left(r - n \frac{R}{N}\right)}{\sqrt{n \left(\frac{R}{N}\right) \left(1 - \frac{R}{N}\right) \left(1 - \frac{n-1}{N-1}\right)}} \quad (2.21)$$

where N denotes the total number of annotated genes, R is the number of genes meeting the filter criterion, n is the total number of genes in the set analysed, and r is the number of genes meeting the filter criterion in the analysis set.

The BiblioSphere Pathway View can be used to elucidate the optimum and relevant networks and pathways for the input gene set including known metabolic and transduction pathways. The input gene set can be viewed in its bibliographical environment, which includes input genes and genes co-cited with input genes. The information about relationships from the literature is also reported as a graphical network. Network nodes can denote genes and edges represent the relationships between genes. Generally, a network node can take various forms, for example, it can be a transcription factor, a part of a metabolic pathway or a part of a Genomatix signal transduction pathway. Furthermore, in the case of genes, the colour of a node indicates the over or under expression of a gene in Bibliosphere. Red nodes indicate overexpressed genes while blue nodes indicate underexpressed genes. Likewise, connections and relationships are described by the type of arrow head, for instance, an open arrowhead is regulation, a filled arrowhead is activation, a blocked arrowhead is inhibition and a red arrowhead is an enzymatic modification. Networks can be viewed at six levels including *Abstract level* (two genes are mentioned in the abstract of an article), *Sentence level* (two genes are mentioned in the same sentence), *Function Word level* (two genes are mentioned in the sentence with a functional term, for example, inhibit), *Gene-Function Word (GFG)* (two genes are mentioned in order in the sentences with a functional term, for example, E2F1 activates TP53), *and expert level* (two genes are hand annotated from sentences by experts) and *signal transduction associations* (two genes are mentioned in the sentences with signal transduction information (pathway associated-term)).

2.5.3 Ingenuity Pathway Analysis software (IPA)

Ingenuity Pathway Analysis software (IPA) (Ingenuity® Systems, Redwood City, CA, USA, <http://www.ingenuity.com>) is a web-based application that integrates a systems biology approach to solve various biological problems. The knowledge base of IPA comes from journal articles, textbooks and other data sources. This software program has many applications; but only functional analysis of genes and their networks have been used in this study. The p-value defines the significance of a gene's function in a network as well as gene to gene relationships, and a p-value less than 0.05 signifies a statistically significant and non-random association. The right-tailed Fisher Exact Test is used to calculate the p-values. IPA presents the relationships and connections between a gene and curated genes in *Network explorer* and *Canonical Pathways* features. The node shape and colour indicate the different types of gene function; for example, squares represent cytokines, and diamonds represent enzymes. Moreover, for up-regulated genes, nodes are red while green nodes indicate down-regulated genes. Grey means neither up nor down-regulated and white defines the non input genes which are added to the network through the connection with other genes. Similar to BSPE, IPA uses arrows to indicate relationships; for example, a filled arrowhead indicates acts on, an unfilled arrowhead represents translocates to and an open arrowhead represent a reaction.

IPA uses a six step algorithm to create the networks. Step one: the focus genes are connected with other genes using the interconnectedness triangles concept. Step two: the networks will grow as the number of triangles increases. The top ranked genes (most connected) are selected by using a metric called “specific connectivity” which is defined as follows:

$$\text{Specific connectivity} = \frac{\text{Number of genes in intersection of the neighbourhood and network}}{\text{Number of genes in union of neighbourhood and network}}$$

The selected networks are added to a maximum network size of 35 connections. As a network grows, the overlapping genes are selected and added to the existing network. Step three: in the case of smaller networks, this method tries to reveal as many relationships as possible; therefore, small networks are merged together with “linker” or common genes between those

networks. This step by step process continues until the network size can be displayed, otherwise, it continues to step four. For networks with fewer than 35 genes, this algorithm will add extra genes to create triangle connectivity. A gene is selected based on the highest gene expression value or its top rank order in the neighbourhood of genes. Step five: all networks are merged to one single network with roughly 35 genes. The final step: p-scores are used to rank networks. The P-score can be described by:

$$\text{p-score} = -\log_{10}(\text{p-value})$$

The p-value represents the probability of finding the focus genes in the global molecular network in the database.

In summary, three network/pathway analysis tools are used in this study. IPA is the main network/pathway tool used in Chapter 4 and 5. BSPE was used to verify the result from IPA in Chapter 5 because this network/pathway tool is based on the same concept as IPA. Both tools use curated information from literature/journal articles. Oncomine was selected because it represented the knowledge-based databases from the existing literature of human cancer gene expression profiles. This tool is used to add information on GR expression levels in other existing childhood leukaemia data.

Chapter 3

Emergent Clustering Methods for the Identification of Glucocorticoid-induced Apoptosis Genes

3.1 Introduction

Glucocorticoids are the most important drug for treating leukaemia, most notably in children with acute lymphoblastic leukaemia. Generally, the mechanism of GCs is mediated by binding to the glucocorticoid receptor, a ligand-activated transcription factor that exerts a pivotal role in inducing apoptosis in malignant lymphoids. It is localised in the cytoplasm and translocated to the nucleus. The GC-GR complex activates or represses the target genes. There are more than 2000 studies on GC-induced apoptosis in lymphoid cells (Herr et al., 2007), of which some studies focus on identifying GC-regulated genes by using gene expression profiles from different cells, drugs and samples, as shown in Table 3.1. However, only a few overlapping genes have been reported (Schmidt et al., 2004). The most prominent study by Schmidt et al. (2006), contains the only time series dataset that were collected from patients, defined novel GC-induced apoptosis genes from childhood ALL patients at early response treatment. This gene expression data still has potential for further analysis.

Gene expression data can be used for four classes of analysis: class comparison, class prediction, class discovery and pathway analysis (Simon et al., 2003). Class discovery attempts to find groups in the samples (patients) or genes. Class comparison aims to identify differentially expressed genes between at least two groups of different biological processes; for instance, disease state and treatment group. Class prediction uses gene expression profiles to predict group membership of a sample. Pathway analysis adds information about functional annotation of differentially expressed genes and gives a picture of genes working as a cascade network. There are at least three pathway analysis methods: cluster analysis, reverse engineering and pathway databases and tools (Leung & Cavalieri, 2003). Pathway analysis can help to understand underlying mechanisms of a selected process. Therefore, the main focus of this chapter is on identifying GC-regulated genes, potential functional gene clusters and pathway analysis on the basis of differentially expressed genes clusters. A detailed description of gene clustering is provided in the following section.

Table 3.1: Differentially expressed genes involved with GC-induced apoptosis mechanism of childhood leukaemia treated with glucocorticoids, from previous studies (adapted from Schmidt et al., 2004 and Tissing et al., 2007).

Genes and Criteria	Drugs and Time	Cell line studies, Gene chip and Tools	References
39 up-regulated genes and 21 down-regulated genes > 2.5 fold for up-regulated and > 2 fold for down-regulated	Dexamethasone 24, 48, 72 and 96 hours	Childhood ALL cell lines-CEM (sensitive and resistance clone) HG_U95Av2 Affymetrix GeneChip suite 4.0 and GeneSpring™	Medh et al. (2003)
98 genes (23 up-regulated genes and 75 down-regulated genes) Six criteria, including at least 2 fold for up-regulated and at least 0.5 fold for down-regulated	Dexamethasone three and eight hours	T-ALL cell lines: Jurkat and CEM-C7 Hu6800/HuGeneFL GeneChip	Obexer et al. (2001)
113 genes > 3 fold	Triamcinolone acetonide four and eight hours	697- a human pre-B leukemic cell line derived from childhood ALL HG U95A array Affymetrix Microarray Suite (MAS 5.0) and the NetAffx website	Planey et al. (2003)
22 genes > 0.7 fold at least 6 of 13 patients and subtract cell cycle genes	Prednisolone 0,6/8 and 24 hours	Childhood leukaemia in vivo treatment HG U133 Plus 2 Affymetrix GCOS software and R packages	Schmidt et al. (2006)
163 GC-regulated genes and 66 G1/G0 genes > 2 fold	Dexamethasone 0, 2 and 8 hours	CCRF-CEM Incyte Gemomics	Tonko et al. (2001)
51 genes (39 up-regulated genes and 12 down-regulated genes) p-value < 0.001 and FDR < 10%	Prednisolone three hours and eight hours	<i>In-vitro</i> paediatric leukaemia cell lines HG U133A GeneChip R packages	Tissing et al. (2007)
121 genes (93 up-regulated genes and 28 down-regulated genes) > 3 fold, one or more time points	Dexamethasone 3, 6, 24 and 48 hours compared with non-treated sampled	697- a pre-B ALL cell lines HG U95A array GeneChip analysis suit software ver.3.3 (Affymetrix)	Yoshida et al. (2002)
39 up-regulated genes and 21 down-regulated genes at least 2.5 fold for up-regulated and at least 2 fold for down-regulated, for two out three experiments	Dexamethasone 20 hours	Childhood ALL cell lines-CEM (sensitive and resistance clone) HG_U95Av2 Affymetrix Microarray Suite (MAS 5.0) and GeneSpring™	Thomson and Johnson (2003)

Clustering or unsupervised classification approaches give a cluster of distinct and highly similar genes (or co-expressed genes) without a predefined cluster. ALL time series data are extremely limited, which leads to the key challenge of how to analyse this limited time series gene expression data in order to maximise the utilisation of invaluable data. As reported in Chapter 2, there are many techniques developed for clustering gene expression and more methods are being developed. Among the existing methods, three types of technique have been selected. (i) Neural networks, a well-known and widely used clustering method that has been used in cancer gene expression analysis (Golub et al., 1999). (ii) Time series clustering (many clustering algorithms have been applied to time series data and more details can be found in Wang et al., (2008)). (iii) Fuzzy clustering: fuzzy clustering was introduced to overcome crisp clustering where one gene belongs to one cluster. Each gene is assigned a cluster membership which indicates the degree of belonging in each cluster. This study uses four specific methods: Self-Organising Map (SOM), Emergent Self Organising Maps (ESOM), Short Time Series Expression Miner (STEM) and Fuzzy clustering by Local Approximations of MEMberships (FLAME) to analyse short time series gene expression data extracted from prednisolone (glucocorticoid) treated childhood leukaemia patients. The results from this study are used to identify gene clusters responsive to GCs and to further infer gene networks and pathways.

3.2 Data and Software

The data used in this study were collected by Schmidt et al. (2006) from childhood leukaemia patients treated with GCs in a study investigating GC-induced apoptosis. The data were collected over a period of 24 hours after treatment, making the dataset excellent for reconstructing gene networks.

Prior to gene clustering, differentially expressed genes must be identified. Our goal is to first verify the original authors' findings on differentially expressed genes, and then explore gene clusters to infer gene networks from short time series expression data analysis in childhood leukaemia. It should be noted that a common problem with microarray data is the curse of dimensionality, i.e., small number of samples with a large number of genes.

3.2.1 Dataset

Raw data in the format of CEL files and normalised microarray data were obtained online from the Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>). Raw data, comprising gene expression measurements for 13 patients (three T-ALL patients and ten B-ALL patients) were collected at three time points: 0 hour, 6/8 hours and 24 hours. The initial analysis of time series gene expression data for differentially expressed gene identification followed the method used by Schmidt et al. (2006). Specifically, raw data was reprocessed, starting from normalisation and selection of differentially expressed genes, with the same log ratio threshold as in the original article. All 39 files were processed and normalised by Robust Multi-array Average (RMA) in R as in the original study; however, our study not only used R as in the original article, but also additionally used RMAExpress and Matlab to calculate the gene expression matrix. Furthermore, our study extended the original authors' work to compare gene expression patterns at six hours and 24 hours. The original authors' mainly focused on gene expression before and six hours after treatment.

3.2.2 Methods

Computational Methods

There are many existing software and tools for RMA calculation available from both commercial and free sources. This study selected three software programs: Matlab, R, and RMAExpress, with the following characteristics, to investigate differentially expressed genes:

- Matlab (<http://www.mathworks.com/>) uses `affyRMA` command from bioinformatics toolbox to calculate Robust Multi-array Average (RMA).
- R (<http://www.r-project.org/>) is software for statistical computing and graphics developed by BioConductor Project (<http://www.bioconductor.org/>) which is freely available source for users and developers to analyse genomic data based on the R programming language.
- RMAExpress (<http://rmaexpress.bmbolstad.com/>) is a web-based tool used to compute RMA normalisation of Affymetrix Genechip® data and does not require R or any component of the BioConductor project and runs on Windows (and Linux).

Identifying patterns from time series gene expression data could provide insights into the underlying gene function and gene networks. Therefore, clustering methods were used in this research. The four selected methods are: SOM, ESOM, STEM, and FLAME. Each method was selected according to their strengths; they are all user friendly software that has been recently developed and still has potential for further development. They have never been used with childhood leukaemia time series data. A summary of these methods is given below:

Self organising maps (SOM), based neural networks, is a nonlinear clustering method with attractive features of data visualisation and dimensionality reduction. SOM can be visualised based on U-Matrix that indicates the distance between each neuron and its neighbours. Peltarion Synapse, a commercial neural networks software, was used and the trained map neurons were clustered using ward clustering.

Emergent self organising maps (ESOM) (Ultsch & Morchen, 2005) is an extension of SOM which allows a map to grow from the initial map. This analysis uses P-Matrix to analyze the density of the gene expression data. P-Matrix is a matrix with entries denoting data density of neurons in their respective neighbourhoods. A neuron with a large P is located in a dense data region while a small P indicates sparse data regions in the data space. Data can be analyzed by using publicly available ESOM software (Databionics ESOM Tool) developed by Ultsch and Morchen (2005). This software is available to download from <http://databionics-esom.sourceforge.net/>.

Short Time series Expression Miner (STEM) (Ernst & Bar-Joseph, 2006) was specifically created to analyse short time series gene expression data. This method uses the change in direction and magnitude of the inputs with time. The operation of the STEM algorithm can be divided into three major steps: (i) selecting reference model profiles (gene expression pattern) that are constructed before analysis, and use all possible gene expression log-ratio increments (from ± 1 to ± 3 , for example) for the two time steps of 6 h and 24 h with respect to time 0 h, (ii) identifying significant reference model profiles that match actual gene expression patterns according to p-values by using a permutation method, and (iii) grouping significant profiles. STEM is a Java-based program; STEM version 1.3.4 was used and downloaded from (<http://www.cs.cmu.edu/~jernst/stem/>).

Fuzzy clustering by Local Approximations of MEmberships (FLAME) (Fu & Medico, 2007). This algorithm starts by calculating similarities of expression patterns using Pearson's correlation and then creating a connected graph of all K-Nearest Neighbours (KNN) of each expression vector. For each expression pattern, the density, which indicates the number of neighbours within a specified distance, was calculated to classify it into one of three groups: (i) cluster supporting object (CSO), (ii) outlier, and (iii) the rest. In the next step, the local approximation of fuzzy membership of the three groups is determined and each object is updated by a linear combination of the fuzzy memberships of its nearest neighbours. Finally, the clusters can be contracted, based on fuzzy membership, into two categories: one gene to one cluster or one gene to multiple clusters. FLAME is integrated with Gene Expression Data Analysis Studio (GEDAS). The software is freely available to download from <http://sourceforge.net/projects/gedas>.

Each program allows the user to adjust parameters. Synapse (SOM) used a map sized 15×15, with 100,000 epochs (number of batch iterations) in training, with initial and final learning rates of 0.5 and 0.001, respectively, and 14,000 epochs during clustering, starting with 0.1 learning rate that reached 0.001 at completion. It uses a Gaussian neighbourhood function with a linear decay function. The ESOM was trained using online learning of a map sized 50×82, 20 training epochs, a correlation distance function and a linear function with cooling strategy for radius and learning rate. STEM was clustered with a 0.7 minimum correlation coefficient and the 0.05 significance level. FLAME was used with Pearson's Correlation for 10-Nearest Neighbours with 0 and 50% thresholds.

The last tool used in this analysis is the Database for Annotation, Visualisation and Integrated Discovery (DAVID). DAVID (<http://david.abcc.ncifcrf.gov/>) is a web-based program for functional annotation and bioinformatics microarray analysis (Huang et al., 2007). DAVID has many functions, of which functional annotation clustering, gene functional classification and gene ID conversion were used in this study for validating the cluster obtained from the above cluster analysis. DAVID uses the EASE score or a modified Fisher Exact P-value to rank the biological significance of gene groups/functions. The user uploads a gene list to DAVID on the web and then chooses criteria for functional annotation clustering and gene functional classification. We used the default setting for both functions.

- **Functional annotation clustering**

Functional annotation clustering clusters relevant annotations for an input gene list in order to find similar gene annotation terms, reduce redundant terms and group heterogeneous gene annotation. The hypothesis for clustering is that similar annotations should contain similar gene members. The level of common genes between two annotations can be calculated using Kappa statistics (a measure of the degree of agreement between two groups of data, which varies from 0 to 1 (weak to strong)). Then fuzzy heuristic multiple linkage clustering is used to cluster a group of similar annotations (similar kappa values). Users can define criteria for functional annotation clustering, for example, three similarity terms overlap at the 0.50 similarity threshold with classification of three initial group memberships, three final group memberships and 0.50 multiple linkage thresholds.

- **Gene functional classification**

Gene function classification can help the interpretation of biological meaning from a large input gene list. This tool calculates a gene-to-gene similarity matrix that shares functional annotations. The fuzzy heuristic multiple linkage clustering method is also used in gene functional classification to cluster functionally related gene according to Kappa values. Users can define criteria for functional annotation clustering, for example, four similarity terms overlap at the 0.35 similarity threshold with classification of four initial group memberships, four final group memberships and 0.50 multiple linkage thresholds.

To summarise, these steps were used in this study:

1. Raw data in the format of CEL files were retrieved from the website of the National Centre for Biotechnology Information (NCBI). Raw data comprised data for 13 patients (3 T-ALL patients and 10 B-ALL patients) and were collected at three time points: 0 hour, 6/8 hours, and 24 hours.
2. All 39 files were processed using the same approach, called Robust Multi-array Average (RMA), but using different software: Matlab, R and RMAExpress to calculate a gene expression matrix.
3. All gene expression matrices were retrieved and then the log ratio was calculated. The original authors' gene expression matrix, called M-values, was downloaded from the NCBI website.

4. The two sets of criteria used to select differentially expressed genes in the original article were log ratios of more than or equal to 0.7 and 1.0 or less than or equal to -0.7 and -1.0 in at least six out of 13 patients. Our study also carried out the analysis with the same ratios.
5. In our study, a further analysis for differentially expressed genes for childhood ALL subgroups T- and B-ALL, were considered at a fold change equal to ± 1.0 with a change of at least five out of ten patients for B-ALL and at least two out of three patients for T-ALL. This was to identify genes that differentiate the two subgroups of ALL.
6. Then, the genes that passed the criteria were defined as differentially expressed genes and were clustered by the four clustering methods (SOM, ESOM, STEM and FLAME) in order to find similar gene expression patterns and pathway relationships. Next, DAVID was used to verify clusters from the selected clustering methods according to gene function.
7. Finally, we compared the resulting gene list with that extracted by the original authors as well as other lists presented in previous studies and the GC-regulated genes.

3.3 Results and Discussion

3.3.1 Validation/extension of original authors' results and identification of GC-regulated genes

The raw data from original authors' study were used to analyse their reproducibility or robustness. We re-processed this data starting from normalisation then selected differentially expressed genes with the same log ratio threshold as in the original article. Specifically, all 39 files were processed and normalised by Robust Multi-array Average (RMA), as in the original study and using the R program. We further investigated the effect on the final gene set of using different platforms (software) to normalise data. We added two platforms: Matlab, and RMAExpress, and compared the number of differentially expressed genes with that of the original authors, as shown in Table 3.2. In the original paper, Schmidt et al., (2006) combined the data for T-ALL and B-ALL in the analysis and selected differentially expressed genes under two conditions: (i) log ratio of ± 0.7 or higher (ii) log ratio of ± 1.0 or higher, for at least six out of the thirteen patients.

The authors' results contained 62 probe sets (25 induced+37 repressed) (49 genes) for early response (6 hours) and 66 probe sets (28 induced+38 repressed) (55 genes) for late response (24 hours). (A gene is represented by a probe set that needs to be converted to a gene symbol and some probe sets are repeated more than once). Schmidt et al. (2006), however, mainly focused on 0-6 hours in their subsequent analysis. Schmidt et al. (2006) found only 32 probe sets (22 genes) for 0-6 hours after deleting cell cycle genes.

The results in Table 3.2 show that we have found more probe sets including all the probe sets from the original paper. More detail of the probe sets, comparing the original authors' results with our data analysis may be found in Appendix A.1 and extra probe sets are reported in Appendix A.2. We re-analysed M-values, gene regulation values, and compared the number of genes that passed the criteria and were indicated by the original author. We found more differentially expressed genes from M-value (re-analysis) than the one reported in the original paper. R and RMAExpress are presented together because both produced the same results.

We compared differentially expressed genes selected from three different methods and the results are shown in Figure 3.1. R/RMExpress showed the most consistency because it gave genes which overlap with other methods. In this study, we then used gene expression values retrieved from R for further analysis.

We selected some probe sets that passed the criteria from each platform to illustrate the variation from different platforms on the final selected gene set. At early response, namely six hours, some selected, induced and repressed probe sets (x-axis) are presented in Figures 3.2 and 3.3 with respect to the number of patients (y-axis). As can be seen in both figures, the probe sets that passed the criteria from each method varied. Matlab and R/RMAExpress produced similar results. Not all probe sets that passed the criteria in one platform passed them in the other two platforms. Sometimes only one patient made the difference between a pass or fail in criteria (six or five patients out of thirteen). In Figure 3.2, while eleven probe sets passed the criteria, when using M-values (reanalysis), these probe sets failed when using Matlab and R/RMAExpress.

Table 3.2: Lists of induced and repressed probe sets from Matlab, M-values (from original and reanalysis) and R/RMAExpress

	6 h		24 h	
	Induced	Repressed	Induced	Repressed
Original authors' (M-values)	25	37	28	38
Re-analysis M-values	58	66	212	258
Matlab	56	75	226	266
R/RMAExpress	56	73	213	258

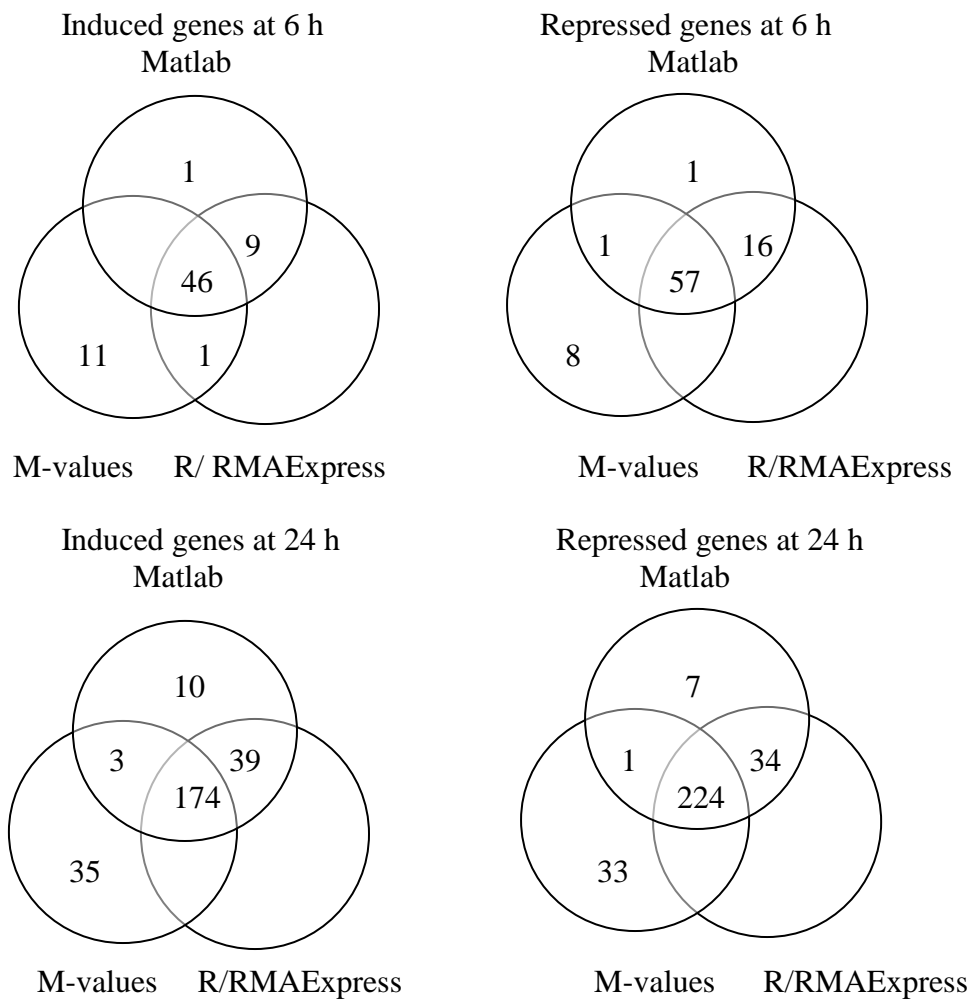


Figure 3.1: Venn diagram of glucocorticoid induced differentially expressed genes at 6 h and 24 h after treatment from Matlab, M-values and RMAExpress

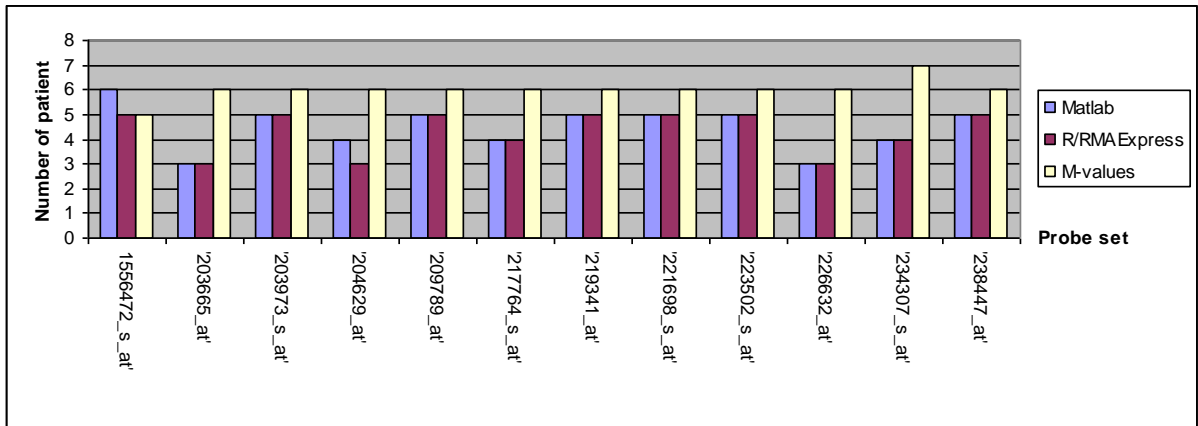


Figure 3.2: Some examples of induced probe sets that passed the criteria at 6 h from Matlab, M-values (reanalysis) and RMAExpress

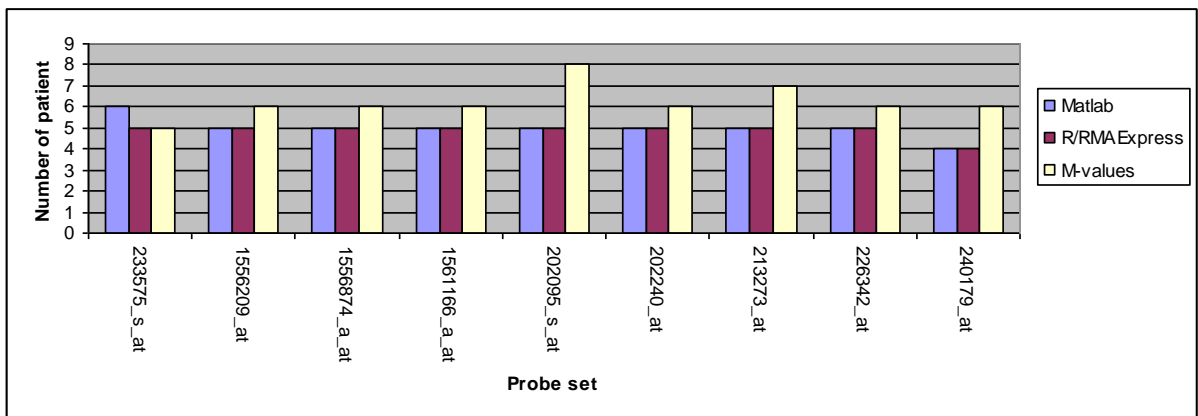


Figure 3.3: Some example of repressed probe sets that passed the criteria at time 6.h from Matlab, M-values (reanalysis) and RMAExpress

We found that combining B-ALL and T-ALL data compromised the accuracy of selection of differentially expressed genes in both T-ALL and B-ALL. It was possible that the selected differentially expressed genes came entirely from B-ALL patients, as there were only three T-ALL patients. We looked at the 22 proposed candidate genes from the original authors and found that there were some genes common to both subtypes, but some genes were found only in one subtype, as indicated in Table 3.3. Therefore, we separated the two types of patients and a new set of differentially expressed genes was selected for T-ALL and B-ALL for each time point. The new criteria used were a log ratio of ± 1 or higher for at least five out of ten B-ALL patients and two out of three T-ALL patients. We also analysed data for early response (6 hours) and late response (24 hours), but we added an analysis for response between 6 and 24 hours because this could give more information about gene activity at different times. The results are shown in Table 3.3.

Table 3.3 shows the number of differentially expressed genes six hours after treatment, between 6 hours and 24 hours and 24 hours after treatment (before and after deleting cell cycle genes). The final set had 237 probe sets (203 unique probe sets after removing repeats) for T-ALL for the combined time points and 257 probe sets (207 unique probe sets) for B-ALL and these were then combined into one set. The final set contained 380 unique probe sets (24 unique probe sets were common to T-ALL and B-ALL, of which three probes were not found in the original paper). These were converted from probe set ID to gene symbol using DAVID. The cell cycle genes were then deleted from this dataset (the cell cycle gene list was retrieved from KEGG, Cell cycle database, and the original article). After deleting the cell cycle genes, T-ALL contained 222 probe sets (172 unique probe sets) and B-ALL contained 190 probe sets (155 unique probe sets) for the combined time points. The final set had 327 unique probe sets (304 genes) responsive to GCs (19 unique probe sets were common to T-ALL and B-ALL). These genes were then assessed for their time of activation (0 h, 6 h, and 24 h or in between).

Considering the possibility of activation, these patterns can be classified into four groups with some examples of gene lists shown in Figure 3.4 and Table 3.4. (1) Genes differentially expressed only at 0-6 or 0-24 h, for example, included 15 induced genes and nine repressed genes from T-ALL. (2) Genes differentially expressed at 0-6, 6-24 and 0-24 h, for example, included three induced genes from B-ALL. (3) Genes differentially expressed at 0-6 h and 0-24 h, for example, included nine induced genes and two repressed genes from B-ALL. (4) Genes differentially expressed at 6-24 and 0-24, for example, included 13 induced genes and seven repressed genes from T-ALL.

We then compared our differentially expressed genes from T-ALL and B-ALL with Schmidt et al. (2006). The original authors had compared their results with those reported in a previous review paper of Schmidt et al. (2004) which reported genes regulated by GCs from previous studies. The comparison results are explained, as follows:

- There are 31 genes reported as GC-regulated genes from different systems, from a previous review of Schmidt et al. (2004). None of the top seven genes (LDH-A, GPR65/TDAG-8, MAP2K3, GZMA, MYC/c-myc, NR3C1/GR, and BCL2L11/Bim) of which were found regulated more than two-fold in the Schmidt et al. (2006) study. Our study found two out of seven known GC-regulated genes met our new criteria for T-ALL: NR3C1/GR and BCL2L11/Bim. Only three out of 31 genes (FKBP51, SOCS1, and DDIT4/Dig2) were found in the Schmidt et al. (2006) study. Altogether, our study found that six genes in the subtypes overlapped the genes listed in Schmidt et al. (2004). B-ALL had one unique gene, FKBP51, and four genes (BCL2L11/Bim, NR3C1/GR, TUBB, and HES1) were unique found in T-ALL. SOCS1 is the only one found in both patients.
- Finally, the Schmidt et al. (2006) study proposed 22 novel GC-regulated genes as common to B-ALL and T-ALL. We confirmed that only 8/22 candidate genes belonged to both B-ALL and T-ALL, and 14/22 genes were found only in B-ALL or T-ALL, as shown in Table 3.5.

Table 3.3: Differentially expressed genes

Differentially expressed genes six hours after treatment, between six hours and 24 hours, and 24 hours after treatment (before and after deleting cell cycle genes) at ± 1 fold change

	0-6 hours				6-24 hours				0-24 hours			
	Before		After		Before		After		Before		After	
	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed
T-ALL	19	10	19	9	59	51	56	49	58	40	56	33
B-ALL	24	23	24	9	16	13	16	9	73	108	71	61
Common	1	1	1	1	1	5	1	4	5	13	5	9

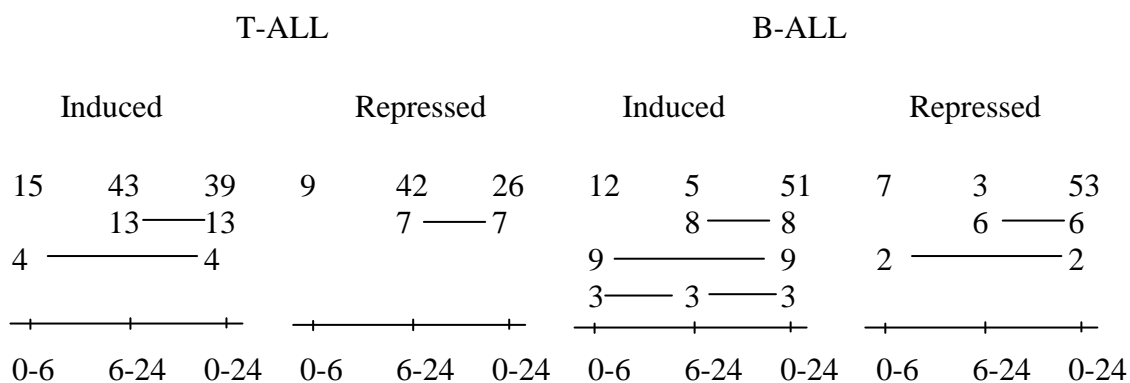


Figure 3.4: Candidate differentially expressed GC-induced apoptosis genes in Table 3.4 broken down into the four patterns of expression.

Numbers indicate the number of genes that are turned on or off (induced and repressed **after deleting cell cycle genes**-same numbers, as indicated in Table 3.4) at three time points for both T-ALL (left hand side) and B-ALL (right hand side). The numbers connected with a solid line indicate that they are all the same genes active at different time points. For example, the four genes found differentially up-regulated at 0-6 were also found at 0-24.

Table 3.4: GC-regulated genes and their activity patterns

No.	Patterns	Genes
1	Genes differentially expressed (turned on) only at six hours or 24 hours	EGR1, SOCS1, IGHM, LYZ, PIK3IP1, HES1, S100A12, GNG11, BCL2L11, ZNF24, and 1552230_AT
2	Genes turned on from six hours to 24 hours	S100A8, ZBTB16, and P2RY14
3	Genes turned on at six hours and stay at same expression level at 24 hours	GIMAP7, SLA, FKBP5, SNF1LK, PFKFB2, EPPK1, WFS1, C6ORF85, BTNL9, and TFP1
4	Genes turned off at six hours but turned on at 24 hours	STAB1, TNFSF8, KIAA0101, DTL, TYMS, RRM2, PPBP, FEN1, TMSL8, TUBB1, PF4, NRGN, and HBG2

Table 3.5: Gene list found by original authors using the combined dataset and the presence or absence of these genes in the two sub-types of ALL from separate analysis of the subtypes

Gene symbol	Description	T-ALL	B-ALL
PFKFB2	PFK2.2	√	√
BTNL9	Butyrophilin-like 9		√
SNFILK	SNF1-like kinase	√	√
FKBP5	FK506 binding protein 51	√	√
ZBTB16	ZFand BTB domain 16		√
KIF26A	Kinesin family member 26A		√
SLA	Src-like-adaptor	√	√
SOCS1	SOCS-1	√	
DDIT4	DNA-damage-ind.transcript 4	√	√
GBP4	Guanylate binding protein 4		√
MGC17330	HGFL gene	√	
ZFP36L2	Zinc finger protein 36	√	
Unknown	Unknown	√	√
EPPK1	Epiplakin 1		√
P2RY14	Purinergic receptor P2Y		√
FGR	Gardner-Rasheed v-fgr		√
WFS1	Wolframin		√
ARPP-21	cAMP-regulated PP21	√	√
SERPINA1	Proteinase inhibitor, clade A		√
GIMAP7	GTPase, IMAP family M7	√	√
MYCPBP	c-myc promoter BP		√
LGALS3	Galectin 3		√

Many previous hypotheses and studies revealed possible GC-regulated genes based on other systems, such as cell-lines and mouse tissues, while the dataset used in this study was retrieved from patients. The only concern with this set is the small number of samples. Another issue that needs to be included is the ambiguity surrounding the differences in drugs that may alter treatment results and side effects. All studies reviewed by Schmidt et al. (2004) used dexamethasone while we have used datasets from prednisolone. The difference between prednisolone and dexamethasone is still inconclusive and not clear. Some researchers have also compared side effects, taste and efficacy of these two drugs in several diseases, i.e., asthma and ALL.

In addition, some studies mentioned no difference (or advantage) between dexamethasone and prednisolone (Igarashi et al., 2005; van Beek et al., 2006) while some do mention their toxicity and treatment results (Juruena et al., 2006; Kaspers et al., 1996; Mitchell et al., 2005). For example, dexamethasone has higher potency (16-fold antileukaemic activity) than prednisolone (Kaspers et al., 1996). Furthermore, prednisone is slightly different from prednisolone. Prednisone is rapidly converted into prednisolone by the liver.

We then compared our new gene list with previous studies. As can be seen in Table 3.1, none of the previous studies, using the same drugs and the same times at which data were collected, showed similar results. Our main feature here being the use of experimental data from clinical samples, the most similar study is that by Tissing et al. (2007), they used same drugs but had a different normalisation process, the tissues were retrieved from childhood leukaemia patients, not cell cultures. They compared the primary childhood ALL cells treated with *in vivo* prednisolone and leukaemic cells of childhood ALL exposed to *in vitro* prednisolone, for finding early apoptosis response genes at 6/8 hours after treatment. Our differentially expressed 29 probe sets from T-ALL and 47 probe sets from B-ALL (Table 3.4 first two column) were compared with the 39 up-regulated and 12 down-regulated genes from Tissing et al. (2007). We found these common induced genes: BTG1 (T-ALL and Tissing), and FKBP5, ZBTB16, SNF1LK (B-ALL and Tissing). No common repressed gene between T- or B-ALL and Tissing was found.

We selected another study from Table 3.1, Thompson and Johnson's (2003), based on using different chemotherapeutic drugs and different time points and different tissues but with similar gene selection criteria. Thompson and Johnson (2003) identified 39 up-regulated genes and 21 down-regulated genes in CEM (a cell line derived from human lymphoid cells). In addition, they proposed the time frame for apoptosis gene regulation after CEM-C7 were exposed to dexamethasone (Thompson & Johnson, 2003). In comparing the gene sets reported by Thompson and Johnson (2003) with our differentially expressed genes from T-ALL and B-ALL patients, we found a few overlapping genes, but more than those found in comparison with Tissing et al. (2007). T-ALL had five overlapping genes (BCL2L1, SOCS1, BTG1, CD69 and NR3C1) and B-ALL has four overlapping genes (SOCS1, DFNA5, WFS1 and SLA). Of these two sets, only BTG1 overlapped with the common genes between our T-ALL and the study of Tissing et al. (2007) for T-ALL. There was no gene common to B-ALL and Tissing et al.'s (2007) study.

In summary, the selected dataset is robust and reproducible. We reproduced the original authors differentially expressed genes using the same normalisation process and criteria. Different platforms (software) for normalisation process behaved similarly with minor differences. In some cases, there was only a \pm one patient difference dividing whether a gene was selected or not. Twenty two differentially expressed genes were proposed by the original authors as GC-regulated genes and common to T and B-ALL (Schmidt et al., 2006); however, our study, that separated T- and B-ALL, revealed that only some of these genes were found in each subtype. We also proposed 327 unique probe sets (304 genes) responsive to GCs. Furthermore, we extended the analysis to find differentially expressed genes between 6 and 24 hours. This can add a temporal picture because previous studies only considered 0 and 6 hours and 0 and 24 hours. We then identified the temporal activation pattern of the proposed GCs-regulated genes. Finally, we compared this gene set with two previous studies and found only a few common genes, possibly indicating that different chemotherapeutic agents and tissues may produce different results for the target gene set.

3.3.2 Extraction of intrinsic biological patterns with four emergent clustering methods applied to gene clustering

After identifying GC-induced apoptosis genes, further analysis focused on finding similar gene expression patterns, which may indicate genes with similar function or co-regulation. First, we identified gene function using The Database for Annotation, Visualisation and Integrated Discovery (DAVID). The results from DAVID after processing our 327 probe sets are as follows:

For **functional annotation clustering**, the top three annotation clusters were defence response (p-value $5.5e^{-3}$), cell cycle ($7.8e^{-4}$) and apoptosis ($3.3e^{-3}$). Specifically, DAVID reported 30 probe sets from our list were involved in apoptosis annotation. This means 30 probe sets from our list are already known as apoptosis related genes. Those 30 probe sets (highlighted gene names) are shown in Table 3.6. Even though we had already deleted cell cycle genes, we still retrieved cell cycle function genes from DAVID. This is an ambiguous issue; different sources/databases can define different set of cell cycle genes.

For **gene functional classification**, the results indicated that there were nine gene functional clusters (170 genes were not clustered), as presented in Table 3.7. Each cluster had many key biological functions, and we selected one function to represent each group. As we were interested in finding similar gene function from gene expression clustering methods, the nine gene functional clusters were used for further comparison with gene clusters from the four selected clustering methods: SOM, ESOM, STEM and FLAME.

In this study, a temporal (time series) gene expression dataset was used. This gene set was collected at three time points, making it a short time series gene expression dataset. This data has not been used with any clustering tools. We aimed to explore the behaviour of our final gene set from the perspective of each of the four selected clustering methods. The results may shed light on to possible consistent functional grouping of genes in childhood leukaemia.

Table 3.6: Results on 30 probe sets found relevant to the apoptosis process from Functional annotation clustering function by DAVID.

Probe sets	Gene name	Gene symbol
1557257_at	B-cell CLL/lymphoma 10	BCL 10
236439_at	B-cell CLL/lymphoma 6:	BCL 6
1559975_at	B-cell translocation gene 1, anti-proliferative	BTG 1
205780_at	BCL2-interacting killer (apoptosis-inducing)	BIK
206665_s_at	BCL2-like 1	Bcl-x/BCL2L1
1555372_at and 1558143_a_at	BCL2-like 11 (apoptosis facilitator)	BCL2L11/Bim
205681_at	BCL2-related protein A1	BCL2A1
226530_at	Bcl2 modifying factor	BMF/FLJ00065
1565752_at, 1553906_s_at and 215602_at	FYVE, RhoGEF and PH domain containing 2	FGD2/FLJ00048/ZFYVE4
244447_at	Kruppel-like factor 10	EGRA/KIF10
211341_at	POU class 4 homeobox 1	POU4F1/FLJ13449
232344_at	RAS p21 protein activator (GTPase activating protein) 1	RASGAP
209936_at	RNA binding motif protein 5	FLJ39876/RBM5
235412_at	Rho guanine nucleotide exchange factor (GEF) 7	ARHGEF7/p50
211899_s_at	TNF receptor-associated factor 4	TRAF4
210314_x_at	TNFSF12-TNFSF13 read-through transcript; tumour necrosis factor (ligand) superfamily, member 12; tumour necrosis factor (ligand) superfamily, member 13	TNFSF12/Tnfsf12-Tnfsf13
202908_at	Wolfram syndrome 1 (wolframin)	WFS1
229958_at	ceroid-lipofuscinosis, neuronal 8 (epilepsy, progressive with mental retardation)	C8orf61/FLJ39417
223556_at and 220085_at	helicase, lymphoid-specific	FLJ10339/SMARCA6
207826_s_at	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein	HEIR-1
203949_at	myeloperoxidase	MPO
32431_at	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	GR/NR3C1
204285_s_at	phorbol-12-myristate-13-acetate-induced protein 1	PMAIP1
206390_x_at	platelet factor 4	MGC138298/Pf4
1552742_at	potassium voltage-gated channel, subfamily H (eag-related), member 8	ELK1
213093_at	protein kinase C, alpha	PRKCA
239504_at	similar to Bcl-2-associated transcription factor 1 (Btf); BCL2-associated transcription factor 1	KIAA0164/LOC731605
241819_at and 235735_at	tumor necrosis factor (ligand) superfamily, member 8	TNFSF8
202643_s_at	tumour necrosis factor, alpha-induced protein 3	TNFAIP3
205883_at, 227762_at, 228854_at and 244697_at	zinc finger and BTB domain containing 16	ZBTB16/PLZF

Table 3.7: Gene functional classification results from DAVID

Functional Group 1 (12) Mitosis	Functional Group 2 (15) Microtubule	Functional Group 3 (14) ATP binding	Functional Group 4 (10) Oxygen transport
218662_s_at, 218663_at 233223_at 213599_at 218542_at 218039_at, 219978_s_at 242787_at 210052_s_at 204026_s_at 223381_at 212949_at	232069_at 212364_at, 212365_at 230690_at 218355_at 220585_at, 237324_s_at 219306_at 218039_at, 219978_s_at 242787_at 221258_s_at 210052_s_at 204825_at 241403_at	232069_at 212912_at 220585_at, 237324_s_at 206828_at 222740_at, 218782_s_at 209642_at 225207_at 208078_s_at 1552921_a_at 204825_at 241403_at 208438_s_at	209458_x_at, 211745_x_at, 204018_x_at, 217414_x_at, 211699_x_at 226632_at 240336_at 204848_x_at, 204419_x_at, 213515_x_at
Functional Group 5 (4) Transcription regulatory activity	Functional Group 6 (5) Positive regulation of transcription, DNA-dependent	Functional Group 7 (11) Negative regulation of transcription	Functional Group 8 (5) Ubiquitin-protein ligase activity
228033_at 203574_at 219990_at 241926_s_at	201694_s_at 210971_s_at 1561973_at 232231_at 244414_at	201694_s_at 244447_at 228964_at 239504_at 226677_at 227762_at, 244697_at, 205883_at, 228854_at 203543_s_at 242210_at	243649_at 242829_x_at 236528_at 223229_at 225328_at
Functional Group 9 (41) Integral to membrane/ Immune response			
206637_at 210146_x_at, 207697_x_at 212771_at 237009_at 206749_at 230175_s_at 235735_at, 241819_at 219230_at	1564424_at 217478_s_at 205898_at 231437_at 1552925_at 221756_at, 221757_at 1565602_at 244523_at 223194_s_at, 242055_at	205857_at 210982_s_at 227405_s_at, 224325_at 215894_at 216510_x_at, 211430_s_at 230597_at 212282_at, 212281_s_at, 212279_at	219607_s_at 212998_x_at 212235_at, 38671_at 205099_s_at 228434_at 206618_at 242551_at 235568_at

Clustering with SOM

The first clustering tool is self organising maps (SOM), which can produce a two dimensional picture of high dimensional gene expression data. SOM can reveal clusters genes with similar expression pattern vectors across the time points. We present here some results for SOM clustering of GC-induced apoptosis genes for two typical patients: patient 2 (T-ALL) and patient 13 (B-ALL), as shown in Figure 3.5. The lighter areas in the U-Matrix map indicate neurons that are clustered closer to each other and reveal some cluster patterns picked by the algorithm on the leftmost panels for the two patients. The cluster maplet reveals two clusters for T-ALL and B-ALL. The last three maplets indicate the expression of genes at 0, 6 h and 24 h. The expression of a gene at the three time points can be traversed by following the expression at the same location of the three maplets. The maplets for the three time points indicate that cluster one (red colour) mainly contains low gene expression values (less than about 6.0) and vice versa for cluster two. The red areas show highly up-regulated genes, whereas the blue areas show highly down-regulated genes.

These patterns indicate that, overall, the majority of genes have similar activity patterns at the three time points (i.e. either high or low). However, although activity pattern across the three time points showed similarity, there was a general decrease in activity in T-ALL (enlarged blue region with increasing time) and general increase in expression in B-ALL (enlarged red/yellow region). But SOM works well with high dimensional inputs. Therefore, we further investigated SOM with the whole set of gene expression at 24 hours after treatment of 13 patients (input dimension is 13). The results are shown in Figure 3.6, where last three panels indicate the last three T-ALL patients. The activity patterns of three T-ALL show differences compared to the eight B-ALL patients (first and second rows). The two gene clusters shown on the cluster maplet indicate the highly expressed and weakly expressed genes, respectively. This cluster pattern is dominated by the 10 B-ALL patients; however, some general similarities between the B-ALL and T-ALL can be found in individual maplets informing the two cluster structure. As the same gene is shown at the same spot in the maplets, SOM can give an overview of how interested genes behave across all 13 patients.

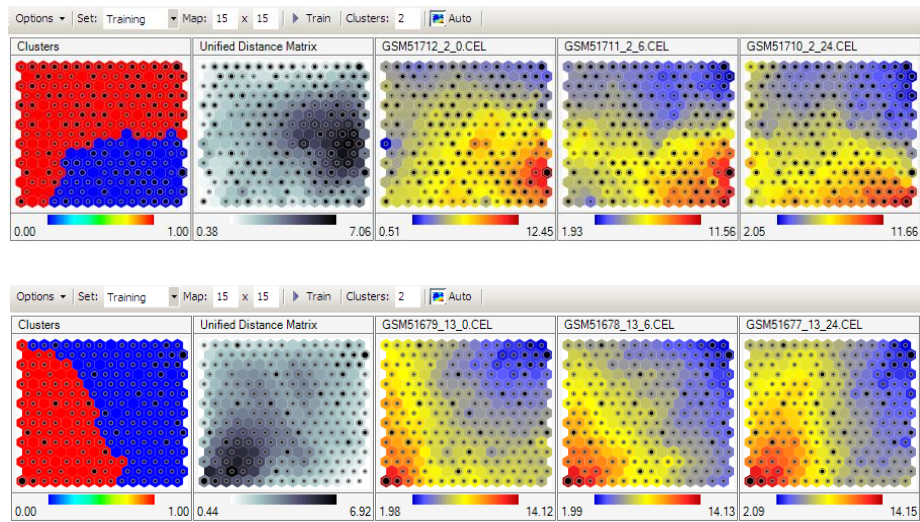
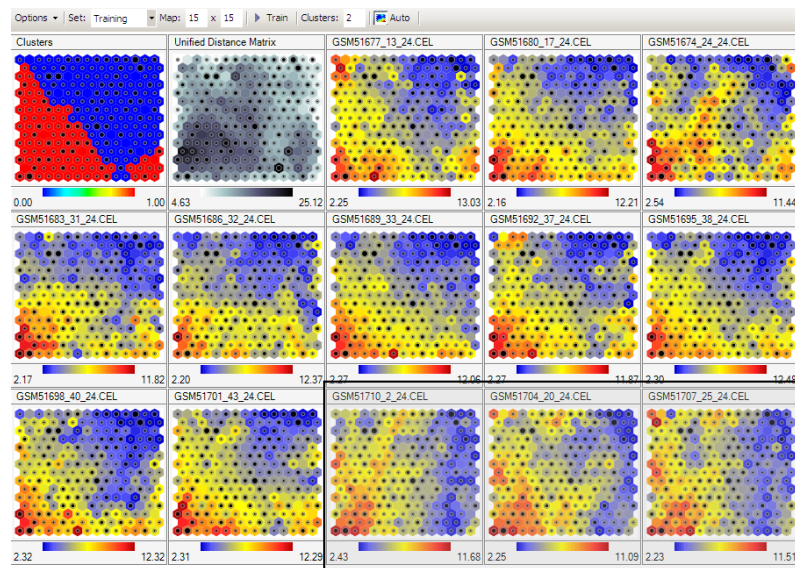


Figure 3.5: Self-organising map of selected T- and B-ALL patients depicting gene expression across the three time points

Self-organising map (maplets from left to right: clusters, U-Matrix, time 0 h, 6 h and 24 h for two selected patients) (Top T-ALL, Bottom B-ALL). Both show that with time, more genes tend to attain high expression values.



T-ALL

Figure 3.6: Self-organising map of ALL patients at 24 hours

Self-organising map (maplets from left to right: clusters, U-Matrix, eight B-ALL patients (patient numbers 13, 17, 24, 31, 32, 33, 37, 38, 40 and 43) and three T-ALL patients (patient numbers 2, 20 and 25))

The advantage of using SOM is we can visualise the macro picture of the three scenarios (time 0, 6 and 24h) for each patient simultaneously. SOM are good to use when data analysis starts because they give an overview of data. However, we found a drawback in using SOM for clustering in that clusters from SOM may not be consistent due to the instability of neural gas, the method used to cluster neurons on the trained map, as it produced unclear cluster boundaries.

Clustering with ESOM

The second clustering method used in this study was Emergent Self organising maps (ESOM). ESOM was created based on two concepts: emergence and border less map. P-Matrix (density-based) visualisation can be used to enhance the visibility of overlapping clusters and the map on the right-hand side continues with the left-hand side, and bottom and top are continued as a torous space. The darker colour represents high density, while the lighter colour represents low density. In Figure 3.7, ESOM provides a better visualisation of separated clusters when compared with Unified dimension matrix (U-Matrix) from SOM in Figure 3.5. However, in ESOM the number of clusters are user-defined, so different users can define different numbers of clusters. In Figure 3.7, we defined at least four gene clusters for both the selected T- and B-ALL patients. We further tested whether the four temporal patterns in Table 3.5 can be separated by density clustering.

We selected ten genes scattered across the four temporals from Table 3.5 and located them on the P-Matrix map. For patient number two (T-ALL): only one group (group 1) from Table 3.4 was represented by a cluster (cluster 3) on the map, the three genes in this cluster were S100A8, ZBTB16 and P2RY14. The other selected gene groups from Table 3.5 did not fall into the same clusters on the map. Likewise, only one gene group (group 4) from Table 3.5-correlated with one cluster (cluster 2) on the ESOM map for patient number 13 (B-ALL). There are many studies claiming the success of using ESOM for clustering (Drachen, Canossa, & Yannakakis, 2009; Haddad et al., 2009; Lehwark, Risi, & Ultsch, 2007; J. Poelmans, P. Elzinga, S. Viaene, M. Van Hulle, & G. Dedene, 2009a; Poelmans et al., 2009b).

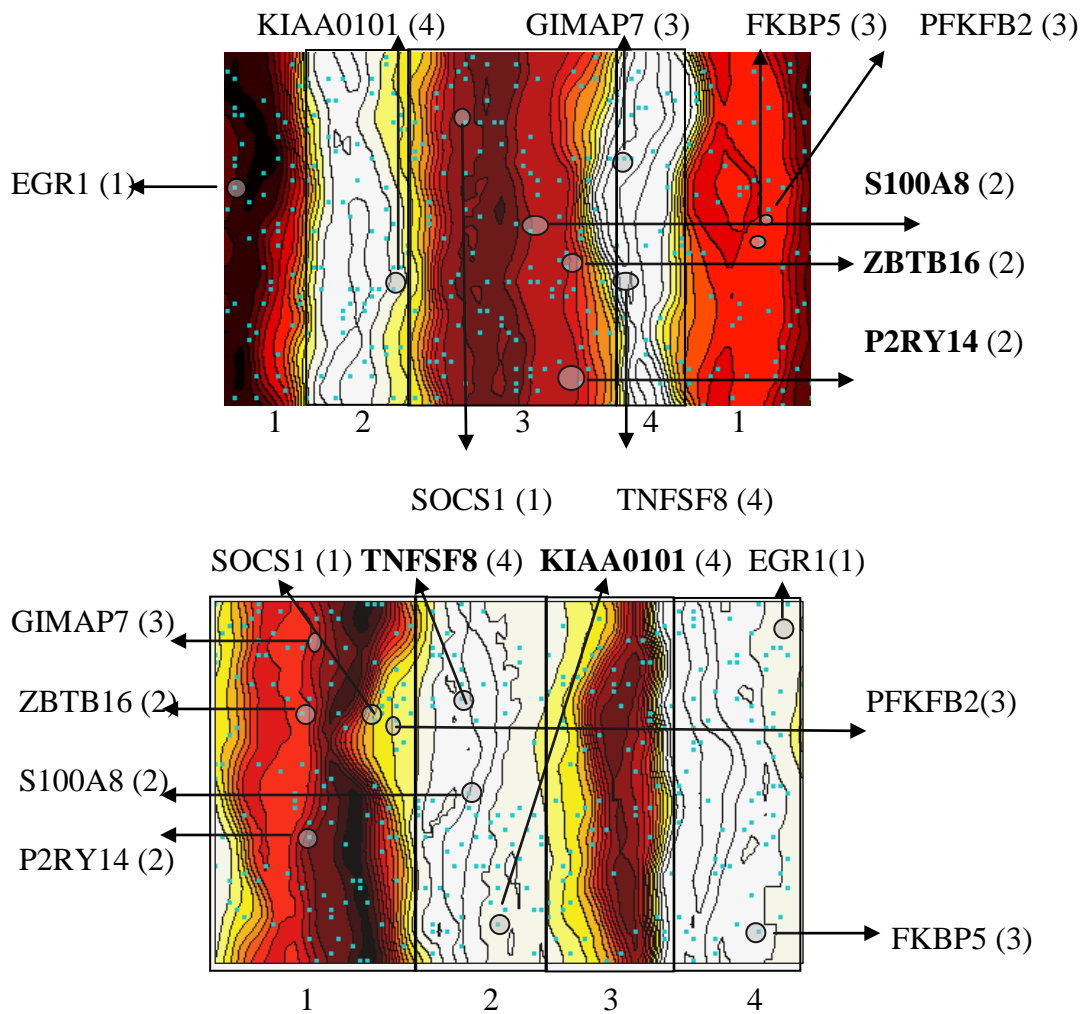


Figure 3.7: ESOM: P-Matrix (density-based) plot for the two selected patients (T-ALL top panel and B-ALL bottom panel)

Dots represent best match (closest) vector to each data point. Number indicates possible cluster.

Gene	Original	ESOM	
		T-ALL	B-ALL
EGR1	1	1	4
SOC1	1	3	1
P2RY14	2	3	1
S100A8	2	3	2
ZBTB16	2	3	1
FKBP5	3	1	4
GIAMP7	3	4	1
PFKFB2	3	1	1
KIAA0101	4	2	2
TNFSF8	4	4	2

Even though ESOM has been used for unsupervised clustering, the previous studies mentioned above have used it for subtype clustering; for example, leukaemia subtypes: ALL or AML, or music types: hiphop, jazz, metal or punk and so on. For such studies, it is easy to evaluate the effectiveness of using ESOM for clustering. In our study, we used it for visualisation of genes and possible gene clustering. ESOM is good for visualising how data are organised in terms of density. ESOM drawbacks are that clusters need to be defined by the user and it produces unclear cluster boundaries as in SOM. Therefore, clusters from both SOM and ESOM methods were not used in further comparisons.

Clustering with STEM

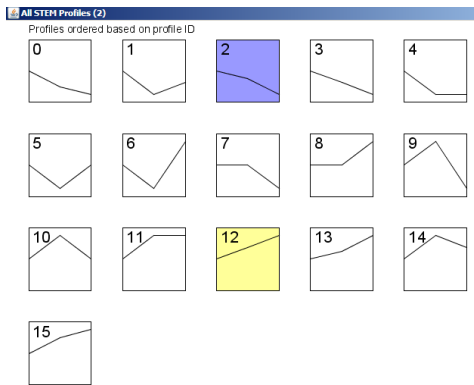
We analysed the 327 probes using Short Time series Expression Miner (STEM) for each patient separately. STEM process starts from converting raw expression data for 6 h and 24 h into log-ratios with respect to the first time point (0 hour); then using reference profiles as the standard permutations to identify the significant model profiles that do not happen by chance. Finally, significant profiles are clustered. Results for three selected patients from T- and B-ALL are shown in Figure 3.8: T-ALL (patients 2, 20 and 25) on the left-hand side and B-ALL (patients 13, 24 and 31) on the right-hand side. Colours indicate significant profiles. Significant profiles for all B-ALL and T-ALL patients were:

- **B-ALL:** patient 13: profiles 4, 13 and 15; patient 24: profiles 8 and 13; patient 31: profiles 2, 3, 11 and 15 (significant profiles for the rest of the B-ALL patients were: patient 17: profiles 2 and 13, and 15; patient 32: profiles 2, 10 and 14; patient 33: profiles 14 and 15; patient 37: profiles 11 and 12; patient 38: profile 3, patient 40: profiles 2 and 8 and patient 43: profiles 2, 12, 13 and 15).
- **T-ALL:** patient 2: profiles 2 and 12; patient 20: profiles 0, 2, 3 and 11 and patient 25: profiles 8, 12 and 13.

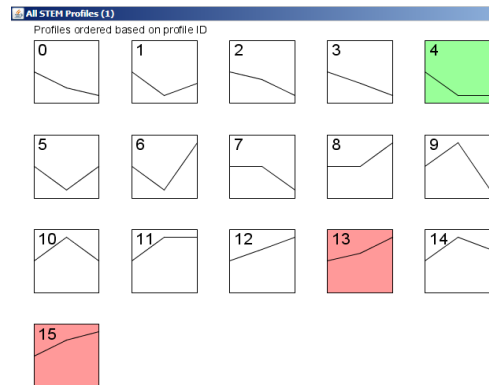
After clustering genes for each patient, group results for T-ALL and B-ALL were assessed. The result across all B-ALL patients showed ten significant cluster profiles: 2, 3, 4, 8, 10, 11, 12, 13, 14, and 15. T-ALL had seven clusters profiles: 0, 2, 3, 8, 11, 12 and 13. After clustering similar significant profiles, each B-ALL and T-ALL patient showed either one or two clusters, as shown in Table 3.8.

T-ALL

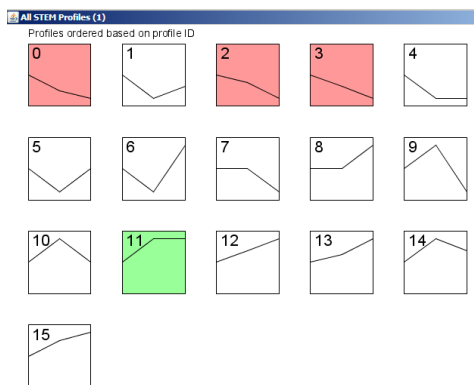
B-ALL



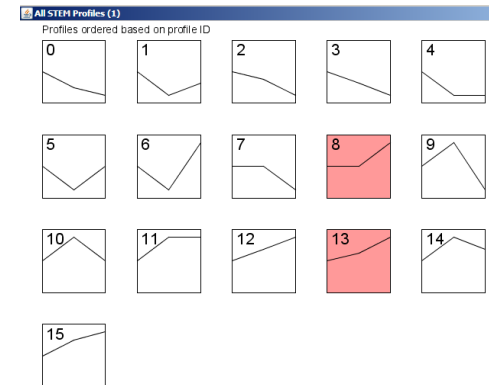
Patient 2



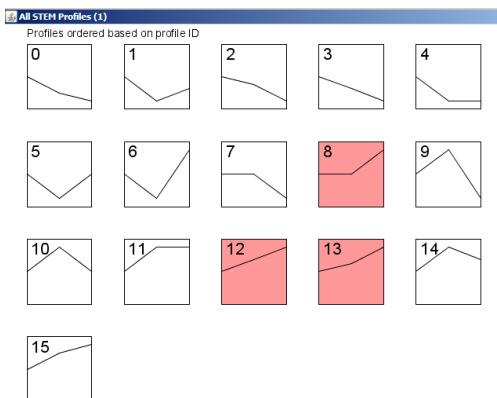
Patient 13



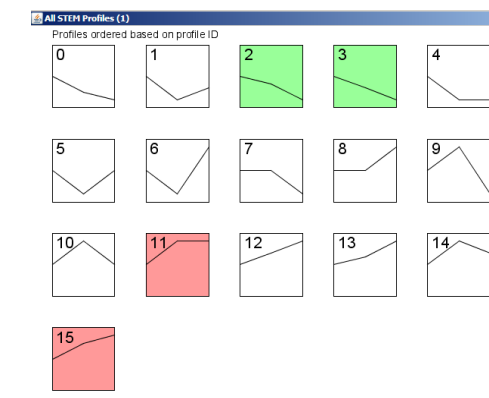
Patient 20



Patient 24



Patient 25



Patient 31

Figure 3.8: Results from STEM for T- and B-ALL

T-ALL (left-hand panel) for three patients (patient numbers 2, 20 and 25). B-ALL (right-hand panel) for patient numbers 13, 24 and 31. Colours denote significant clusters and the same colour belongs to the same cluster.

Table 3.8: Significant clusters and profiles from STEM for T and B-ALL patients

	Patient Number	Cluster 1	Cluster 2	Total number of probe sets in profiles
B-ALL	13	4	13,15	325
	17	2	13	95
	24	-	8,13	163
	31	2,3	11,12	125
	32	2	10,14	128
	33	-	14,15	76
	37	-	11,12	144
	38	3	-	119
	40	2	8	117
	43	2	12,13,15	184
T-ALL	2	2	12	149
	20	0,2,3	11	119
	25	-	8,12,13	170

Although significant profiles in the two clusters were not identical, Table 3.8 appears to separate increasing and decreasing temporal activity patterns into two clusters. The total statistically significant probe sets from STEM clusters were 141 probe sets (129 genes) for T-ALL and 221 probe sets (208 genes) for B-ALL, with 67 probe sets (63 genes) in common. For some selected significant profiles, Table 3.9 shows all the probe sets represented by them.

Generally, clustering methods always exhibit patterns or clusters for all input data. It is very difficult to validate the significance or robustness of the results. We verified our cluster results by comparing the significant clusters for T- and B-ALL with gene functional groups from Table 3.7, with the results shown in Table 3.10.

From Table 3.10 most probe sets in significant profiles from B-ALL were found scattered in functional groups; for example, functional group 1 contains probe sets from profiles 2, 3, 4 and 13; functional group 2 contains probe sets from profiles 2, 3, 4, 8 and 10 and functional group 3 contains probe sets from profiles 2, 3, 4, 10, 11, 12, 13, 14 and 15. Similarly, for T-ALL most probe sets in significant profiles were also found scattered in functional groups; for example, functional group 1 contains probe sets from profiles 0, 3 and 13 and functional group 2 contains probe sets from profile 0, 2 and 13. We can conclude from this analysis that some genes from the same cluster have similar gene function (as they were found in the same functional cluster). But there was more than one function for each gene cluster. How to assign genes to the right function can be a future research topic.

Table 3.9: List of probe sets in some examples of significant clusters from STEM for B-ALL patients

3	4	8	10	11	12
ProbeIDs	ProbeIDs	ProbeIDs	ProbeIDs	ProbeIDs	ProbeIDs
1554733_AT	1552921_A_AT	1560706_AT	1556472_S_AT	1556682_S_AT	1555372_AT
202503_S_AT	1554696_S_AT	1564424_AT	202908_AT	1564424_AT	1555745_A_AT
204026_S_AT	1557910_AT	1569225_A_AT	203395_S_AT	1565599_AT	1556472_S_AT
204127_AT	1565602_AT	203543_S_AT	204018_X_AT	1569225_A_AT	1569225_A_AT
204146_AT	201013_S_AT	204018_X_AT	205883_AT	202833_S_AT	202833_S_AT
204825_AT	201014_S_AT	204150_AT	206637_AT	202908_AT	203760_S_AT
205024_S_AT	201761_AT	204560_AT	208949_S_AT	202975_S_AT	203761_AT
206102_AT	202345_S_AT	205033_S_AT	209301_AT	203543_S_AT	204018_X_AT
209936_AT	203708_AT	205099_S_AT	209458_X_AT	203760_S_AT	205681_AT
210052_S_AT	204127_AT	205883_AT	209992_AT	203761_AT	205883_AT
212949_AT	204146_AT	206618_AT	210001_S_AT	204560_AT	205950_S_AT
218039_AT	204439_AT	209458_X_AT	211429_S_AT	205883_AT	209458_X_AT
218355_AT	204700_X_AT	211430_S_AT	211430_S_AT	208078_S_AT	211429_S_AT
218542_AT	204836_AT	211699_X_AT	211699_X_AT	208949_S_AT	211699_X_AT
218663_AT	209642_AT	211745_X_AT	211745_X_AT	210146_X_AT	211745_X_AT
219306_AT	210948_S_AT	212771_AT	213975_S_AT	210448_S_AT	212912_AT
219493_AT	211302_S_AT	213515_X_AT	217022_S_AT	212195_AT	213515_X_AT
219918_S_AT	212281_S_AT	217022_S_AT	224325_AT	212771_AT	215602_AT
219978_S_AT	213599_AT	217414_X_AT	224840_AT	213817_AT	217414_X_AT
220448_AT	214452_AT	218638_S_AT	227265_AT	215528_AT	219607_S_AT
220651_S_AT	215117_AT	219230_AT	227405_S_AT	215602_AT	222062_AT
221521_S_AT	218039_AT	221756_AT	227611_AT	218638_S_AT	224856_AT
221591_S_AT	218355_AT	223027_AT	227762_AT	219230_AT	225207_AT
222680_S_AT	218663_AT	223028_S_AT	228854_AT	219607_S_AT	225239_AT
223229_AT	219306_AT	224325_AT	232069_AT	221756_AT	225685_AT
223381_AT	219978_S_AT	224840_AT		221757_AT	226530_AT
227921_AT	219990_AT	224856_AT		222303_AT	226733_AT
228273_AT	220651_S_AT	226530_AT		223027_AT	227265_AT
229490_S_AT	223062_S_AT	226733_AT		224840_AT	227611_AT
	223229_AT	226982_AT		224856_AT	227762_AT
	223381_AT	227611_AT		225207_AT	228434_AT
	224797_AT	227762_AT		225949_AT	228697_AT
	226980_AT	228697_AT		226530_AT	228854_AT
	228071_AT	228854_AT		226982_AT	231332_AT
	235088_AT	228964_AT		227062_AT	232583_AT
	241926_S_AT	232344_AT		227265_AT	236512_AT
		235735_AT		227611_AT	238999_AT
		235735_AT		228697_AT	240019_AT
		236450_AT		228854_AT	240038_AT
		240019_AT		229958_AT	240665_AT
		240665_AT		232344_AT	241819_AT
		242551_AT		232431_AT	
		244357_AT		232583_AT	
		244447_AT		235735_AT	
		244697_AT		236450_AT	
				236512_AT	
				240890_AT	
				244026_AT	
				244357_AT	

Table 3.10: Comparison between gene functional groups (9 groups) with STEM significant profiles (16 profiles) for T- and B-ALL patients

Numbers in blanket indicate number of total probe sets. Total probe sets from STEM are 159 probe sets for T-ALL and 450 probe sets for B-ALL and gene functional groups are comprised of 117 probe sets.

	STEM significant profiles																
	T-ALL							B-ALL									
	0 (21)	2 (33)	3 (13)	8 (26)	11 (18)	12 (29)	13 (19)	2 (65)	3 (29)	4 (36)	8 (45)	10 (25)	11 (49)	12 (41)	13 (70)	14 (32)	15 (58)
1(12)	5		1				1	5	6	4					1		
2(15)	5	1					1	5	6	4	1	1					
3(14)	1	3					1	4	1	2		1	2	2	4	2	1
4(10)				3			1	1			6	4		6	1		3
5(4)		1				1		1		2							1
6(5)	1				1	1	1								1	1	
7(11)	1	1		11							5	3	3	3	1	1	1
8(5)	1		2			1		1	1	1							
9(41)	2	2	1	3	1	6	1	5		2	8	4	6	2	8	3	9

Clustering with FLAME

The last clustering method used in this study is Fuzzy clustering by Local Approximation of MEMbership (FLAME). FLAME identified clusters without the need for pre-defined numbers of clusters while providing fuzzy clustering characteristics. However, the number of clusters for each patient varied, depending on the distance measure, with Pearson's correlation producing a larger number of clusters than Euclidean distance, as shown below:

- B-ALL:

Euclidean distance: patient 13 (14 clusters), patient 17 (12), patient 24 (10), patient 31 (12), patient 32 (13), patient 33 (12), patient 37(11), patient 38(11), patient 40 (13) and patient 43 (13).

Pearson correlation coefficient: patient 13 (20 clusters), patient 17 (16), patient 24 (16), patient 31 (21), patient 32 (18), patient 33 (18), patient 37 (18), patient 38 (18), patient 40 (21) and patient 43 (19).

- T-ALL:

Euclidean distance: patient 2 (9 clusters), patient 20 (11) and patient 25 (10).

Pearson correlation coefficient: patient 2 (14 clusters), patient 20 (17) and patient 25 (20).

Figure 3.9 shows 14 clusters for patient 2 (T-ALL) and 20 clusters for patient 13 (B-ALL) from FLAME with Pearson's correlation coefficient. The mean number of clusters for T-ALL was 17 clusters and 18 clusters for B-ALL with Pearson's correlation coefficient and 10 clusters for T-ALL and 12 clusters for B-ALL when using Euclidean distance. The choice of distance measure can influence the final numbers of output cluster. The selection of the distance measure should be based on prior knowledge for each application. In this study, we selected two distance measures to show their effect. FLAME and STEM are different in how they presented final results. FLAME considers that all clusters are possible while STEM only selects significant profiles. We aimed for consistent gene groups. Therefore, results from both clustering methods were compared. FLAME clusters, based on Pearson's correlation coefficient, were compared with STEM because STEM also found similarities in shape between two gene expression patterns (i.e. correlation). We compared STEM and FLAME for each patient; specifically, probe sets in all 16 profiles of STEM were compared with probe sets in all clusters in FLAME for each patient.

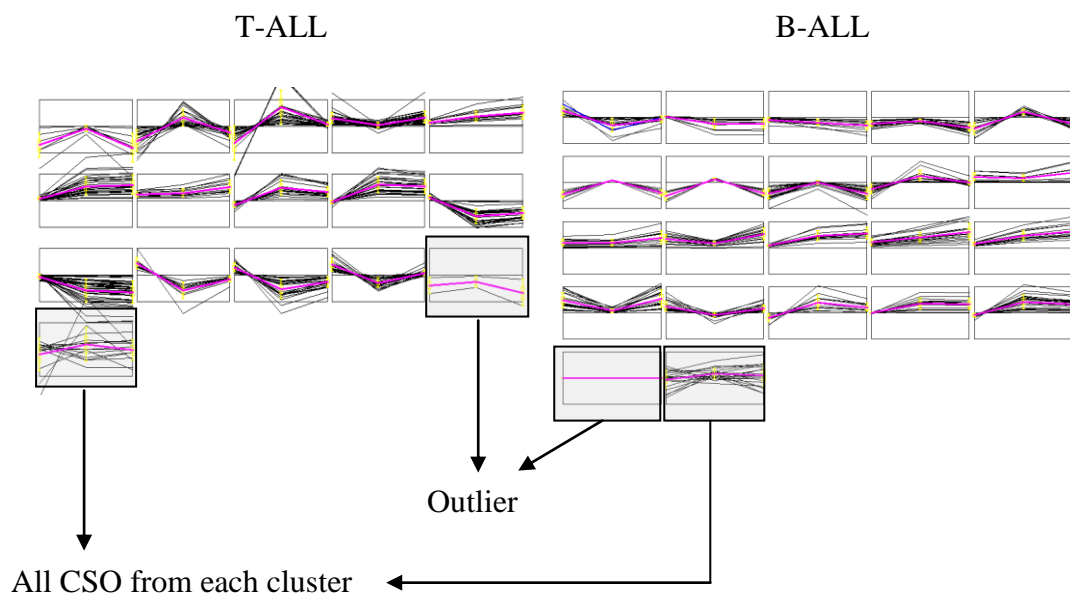


Figure 3.9: Clustering results from FLAME based on Pearson correlation for one selected patient from each subtype

FLAME: for T-ALL (patient 2) - line plots of gene expression for each Cluster, Outlier and Cluster Supporting Object (CSO) clusters: there are 14 clusters (0 to 13) for patient 2 (T-ALL) and 20 clusters for patient 13 (B-ALL).

The patterns in each profile from STEM and clusters from FLAME showed similar characteristics and number of genes in each profile/cluster for the cluster from patient with high number of probe sets. For example, 16 profiles (325 probe sets) from STEM for patient number 13 (B-ALL) compared with 20 clusters (327 probe sets) from FLAME: profile number one in STEM had 15 probe sets of which nine probe sets were found in cluster number nine in FLAME (there were nine probe sets in cluster number nine). Similarly, profile number 13(10/15) from STEM and cluster number four from FLAME (10/18) had ten genes in common. However, the patterns in each profile from STEM and cluster in FLAME were not correlated or were scattered for the lower number of probe sets in the cluster, for example, patient number 2 (T-ALL). The results of the comparison between 16 profiles (149 probe sets) from STEM with 13 cluster and one outlier (327 probe sets) are shown in Table 3.11.

The results for each patient from FLAME cannot be combined like the results from STEM. For STEM, we selected only significant profiles from each patient and then combined all probe sets of all patients under each significant profile to represent all T-and B-ALL patients (see more detail in Table 3.9). Nevertheless, for FLAME, each patient had a different number of genes in each cluster. To address this issue, we prepared a new input dataset for FLAME analysis. In this dataset, each row represented expression of one gene at the three time points for all patients. FLAME analysis was conducted for the three T-ALL, ten B-ALL, and 13 ALL. FLAME identified six clusters, eight clusters, and six clusters for T-ALL, B-ALL and ALL, respectively. The probe sets under these clusters were used for further comparison with Table 3.7.

The results from FLAME were similar to the results from STEM, when comparing probe sets of clusters with the members in each functional group (comparison between FLAME results and gene functional groups are shown in Table 3.12). For example, for B-ALL, functional group 1 can be found in clusters 2, 3, 4, 7 and 8; functional group 2 can be found in clusters 2, 3, 4, 7 and 8 and so on. One cluster contains genes from several functional groups.

Table 3.11: Comparison of probe sets from STEM profiles (16 profiles) and FLAME cluster (15 clusters: 14 clusters and one outlier) for patient number 2 (T-ALL)

Numbers in brackets indicate number of total genes. Total probe sets from STEM are 149 probe sets and 327 probe sets from FLAME.

	STEM significant profiles															
	0 (2)	1 (9)	2 (15)	3 (6)	4 (7)	5 (12)	6 (10)	7 (15)	8 (13)	9 (5)	10 (2)	11 (17)	12 (13)	13 (8)	14 (9)	15 (6)
0(7)			2					3		1						
1(23)			6	1				5		3		1			2	
2(25)			6	3	1			6		1	1				1	
3(39)		1				4	4		6			2	6	3		1
4(14)													1			
5(33)					1											1
6(13)						1			1							1
7(19)	1		1								1	2			2	
8(28)	1			1	2							1				
9(25)					2								2			2
10(41)				1								6			4	
11(12)		2				3	3						2	1		
12(24)		4				1	2		1			4	2	1		
13(22)		2			1	3	1		5			1		3		1
14(2)								1								

Table 3.12: Comparison between gene functional groups (nine groups) with FLAME clusters for T- ALL (six clusters and one outlier cluster) and B-ALL (eight clusters and one outlier cluster) patients

Numbers in brackets indicate number of total probe sets. Total probe sets from FLAME are 327 probe sets and 114 probe sets from gene functional groups.

	FLAME clusters															
	T-ALL							B-ALL								
	0 (32)	1 (31)	2 (23)	3 (26)	4 (67)	5 (136)	6 (12)	0 (14)	1 (15)	2 (28)	3 (42)	4 (41)	5 (15)	6 (18)	7 (149)	8 (5)
1(12)	1			1	9	1				1	1	7			2	1
2(15)	2			1	10	1	1			2	1	7			3	2
3(14)		1			9	3	1	2		1	2	3			4	2
4(10)			9				1			1					9	
5(4)			1		3			1				2			1	
6(5)	1	1		1		2				1	1			1	2	
7(11)		1		1	1	8					1		1		9	
8(5)	1				1	3						1	1		3	
9(41)	1	7	2	6	4	21	2	1	2	8	4	1	3	1	22	

Clearly, clustering tools can give possible groups that might have similar function or co-regulated genes. However, clustering cannot describe interactions between genes. Therefore, prior knowledge can help validate the output from clustering. Consequently, the next step of this study, which is presented in Chapter 4, used literature-based network tools to infer the gene networks of candidate GC-induced apoptosis genes.

3.6 Conclusion

The overall theme of the study in this chapter addressed the identification of GC-regulated genes. Schmidt et al (2006) claimed that they identified a novel set of glucocorticoid-response genes in children with acute lymphoblastic leukaemia. Their study was carried out using whole-genome expression profiling and then compared gene regulation with additional systems, such as mice, and those from previous studies, as reviewed by Schmidt et al (2004). We validated the original paper's results starting from the normalisation process. Studies by many researchers have shown different results from different normalisation methods. We compared the same pre-processing methods for gene expression analysis on different software platforms (Matlab (Bioinformatics toolbox), R, and RMAExpress) which resulted in slightly different outputs. However, R and RMAExpress produced identical gene lists. If a threshold of \pm two patients was used (i.e. 6 ± 2 out of 13 instead of 6 out of 13), all the software provided almost the same gene lists. Analysis of the processed data continued with the focus on finding novel GC-regulated genes. A discrepancy between T-ALL and B-ALL was shown to exist; this led to the proposal of new and separate criteria for the two subtypes. Altogether, we found 327 probe sets (304 genes) differentially expressed at the three time points. These genes can be classified into four different gene activities occurring at the three time points: for example, some genes are active at a particular time point while some other genes are active at all times.

The comparison between the GC-regulated genes reported by Schmidt et al. (2004) and our new set showed six out of 31 common/overlapping genes. We confirmed that only eight out of twenty two novel GC-regulated genes proposed by Schmidt et al. (2006) were common to both T- and B-ALL subtypes, while 14 out of 22 genes were only found in either T- or B-ALL. Furthermore, considering the two previous studies [Tissing et al. (2007), and Thompson

& Johnson (2003)], which were compared with our final gene lists, only a few common genes were found. All investigations in this study lead to the conclusion that different chemotherapeutic drugs, sample tissues, period of data collection and gene selection criteria may affect the final discovery of differentially expressed gene sets.

We then further analysed the data with four emergent clustering tools. Each computational method provided different insights into our short time series data. SOM and ESOM are visualisation methods for high dimensional data projected onto a map and give an overview of how data are organised in terms of distance and density; the drawback is that clusters from SOM are not consistent and ESOM require users to define the number of clusters. The SOM identified two clusters of expression profiles: genes that were highly expressed throughout the time points and genes weakly expressed throughout the time points. The more consistent and detailed clusters were obtained from STEM and FLAME. STEM was used to find expression patterns that were statistically significant and have only a very little probability of happening by chance. FLAME can be used to find clusters without predefined groups and was useful to verify clusters from STEM. But FLAME considers all clusters as possible. Each patient had different cluster results from both STEM and FLAME. Since we needed to analyse groups of patients, the challenge was to develop data analysis tools which can analyse multiple samples (patients) and multiple time points at the same time. We also found the different clusters when applying different distance measures. Prior knowledge needs to be incorporated with the selection of distance measure used in clustering tools.

The next stage of our study, in Chapter 4, will look into GC responsive gene networks that may control the gene expression patterns behind the scenes which, in turn, will help identify target genes for better treatment procedures.

Chapter 4

Identification of GC-induced apoptosis gene network

4.1 Introduction

To understand a complex apoptosis process, identification of differentially expressed genes alone is not enough. There is a need to extend the study level from the identified genes to a gene network. Gaining a better understanding of gene networks is currently a challenge for scientists. Previous work on inferring gene regulatory networks (GRN) used statistical and mathematical modelling. Inferring gene networks from rapidly growing microarray gene expression using network/pathway databases has been shown to be a very promising approach in cancer research. Prior biological knowledge plays an important role in inferring genes in the network. Taking this into account in this chapter, we present how to infer gene networks using time series microarray data from selected pathway databases based on prior knowledge. In this study, we used publicly available childhood leukaemia treated with prednisolone (Schmidt et al., 2006) short time series data, and this data have not been maximally leveraged or integrated with existing knowledge pathway databases. Therefore, in this chapter we used computational analysis with identified GC-regulated genes to identify their gene networks. We used GC-regulated genes before and after deleting cell cycle genes as an input gene list.

Usually, inferred networks are the final step of gene expression analysis. The number of inferred networks from time series varies according to the number of time points in the input data. However, the scientist/biologist or clinician needs a small but essential group of genes for further experiment. This leads to the question of how inferred networks, specifically, from time series data, can be combined in order to minimise the relevant genes or identify novel genes for further study.

The specific objective in this chapter is to infer gene networks at three time intervals and find the common genes which play roles in these three networks. This inferred network was then used to answer the questions of how networks for different time points can reduce the number of interested genes to a few possible novel genes for further clinical study.

4.2 Methods

4.2.1 Dataset

Raw data were collected from 13 patients (three T-ALL patients and ten B-ALL patients). The data used in this part was the same data used in the previous chapter. Specifically, there were two main datasets used in this chapter, these datasets were the results from Chapter 3. (i) Candidate GC-regulated genes it uses, as shown in Table 4.1. These differentially expressed genes from three time intervals: 6 hours after treatment (0-6 hours), between 6 hours and 24 hours (6-24 hours), and 24 hours after treatment (0-24 hours) after deleting cell cycle genes at ± 1 fold change. (ii) Statistically significant probe sets from STEM clustering method.

Table 4.1: GC-regulated differentially expressed genes

	0-6 hours		6-24 hours		0-24 hours	
	Induced	Repressed	Induced	Repressed	Induced	Repressed
T-ALL	19	9	56	49	56	33
B-ALL	24	9	16	9	71	61

4.2.2 Computational Methods

Gene network analysis and recovery of key biological pathways in this study were conducted using Ingenuity Pathway Analysis software (IPA) (Ingenuity® Systems, Redwood City, CA, USA, <http://www.ingenuity.com>). IPA is a web-based application that integrates a systems biology approach to solve various biological problems. The knowledge base of IPA comes from journal articles, textbooks and other data sources. This software has many applications; only the functional analysis of genes and their networks have been used in this study. The p-value defines the significance of gene function in a network as well as gene to gene relationships, and a p-value less than 0.05 signifies a statistically significant and non-random association. The right-tailed Fisher Exact Test was used to calculate the p-value.

4.3 Results and Discussion

4.3.1 Inferring GR gene networks from GC-induced apoptosis genes

Datasets containing expression values of GC-regulated genes were uploaded and analysed through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com). Each gene list was an input to Ingenuity Pathway Analysis software (IPA) which maps the genes to pathways generating networks using an algorithm based on gene connectivity with a cut-off of 35 molecules per network and produces a table containing molecules in each network. In each table, ‘the focus molecules’ means input genes (presented in bold letters) that are overlaid onto a global molecular network and networks to create algorithmically generated networks based on their connectivity. In each network, up-regulated genes are in red and the intensity indicating the degree of up-regulation while green denotes down-regulated genes. An uncoloured node was not found in the uploaded dataset but computationally generated by the IPA on the basis of stored knowledge. Network scores and p-values indicate the significance of each network, process or pathway. The score tells the possibility of the Network Eligible Molecules happening by chance. The ‘score’ is calculated by the negative exponent of the p-value. The interpretation of the score number is that if the score is 53, this implies that p-value was $10e^{-53}$ that is, the higher the score the lower the p-value. The ‘top functions’ in each table show the three most significant functions of each network. Dashed lines indicate indirect interactions while solid lines indicate direct interaction. There are 20 possible edge labels: A (Activation), B (Binding), C (Causes/Leads to), CC (Chemical-Chemical interaction), CP (Chemical- Protein interaction), E (Expression includes metabolism/synthesis for chemicals), EC (Enzyme Catalysis), I (Inhibition), L (ProteoLysis includes degradation for Chemicals), LO (Localisation), M (Biochemical Modification), MB (Group/complex Membership), P (Phosphorylation/ Dephosphorylation), PD (Protein-DNA binding), PP (Protein-Protein binding), PR (Protein-RNA binding), RB (Regulation of Binding), RE (Reaction), T (Transcription), and TR (Translocation).

The differentially expressed gene list from Table 4.1 was used for inferring the gene networks through the IPA software. We used the gene list after deleting cell cycle genes. For T-ALL, there were 28 probe sets at 0-6 hours, 105 probe sets at 6-24 hours and 89 probe sets at 0-24 hours. Similarly, for B-ALL there were 33 probe sets at 0-6 hours, 25 probe sets at 6-24 hours and 132 probe sets at 0-24 hours. Results from IPA (data not shown) typically were a number

of networks ranking from one to seven for each time point with a maximum of 35 molecules in each network consisting of those genes from our list plus those given by IPA. Processing these networks was fairly time consuming but we manually identified common genes active throughout the period (at least between two time points) which can be referred to as predominant genes to T- and B-ALL, separately and common to both.

For T-ALL patients, 48 unique genes were found for the three different time points from IPA: Akt, ARHGEF7, **ASPM**, ATAD2, BCL2L11, BMF, CKSCR1, DTL, ERK*, **E2f**, FEN1, HBG2, hCG, hemin, HES1, Histone h3, **HNF4A**, IL6, JNK, **KIAA0101**, LDL, LYZ, MCM10, MAPK, MS4A1, **MYC***, **NFκB (complex)***, PF4, PMAIP1, PKMYT1, PPBP, PTGDR, p38 MARK, RNF4, SEPT5, SLC2A4, STOM, **S100A8**, TEPI, TMSB15A, TNF*, TNFSF13, TRIP6, TUBB1, UHRF1, Vegf, and ZNF24. Out of these, the seven genes were common to both T- and B-ALL patients. The asterisks indicate four genes found active throughout the three time intervals. Twenty three genes were found in our gene lists and 25 genes were added by IPA. Only four genes were found to be active at all three time points: ERK, MYC, NFκB (complex) and TNF. None of these genes was found in our analysis or that of the original author. These four genes were strongly related to cancer. ERK (Extracellular-signal-Regulated Kinase) is an essential component and part of the mitogen-activated protein kinase MARK/ERK signalling pathway. This pathway plays a major role in cancer therapies. MYC or cMyc gene is found in many cancers. NFκB (Nuclear Factor Kappa-light-chain-enhancer of activated B cells) is activated in the development of many types of cancer and is a therapeutic target for cancer treatment. TNF (Tumour Necrosis Factors) is also involved in cancer therapy because this protein can cause cell death.

For B-ALL patients, the 47 unique genes are ANAPC5, **ASPM**, Beta-estradiol, BUB1, CCNG1, CDC2, CDC20, CDC42EP3, CDC45L, CENPA, CENPF, CEP55, Cyclin A, DLGAP5, **E2f**, FKBP5, GBP1, **HNF4A**, IFNG, IL13, IL4R, **KIAA0101**, KIF14, KIF23, KIF20A, LY6A, MKI67, **MYC**, **NFκB (complex)**, NUF2, RPL3B, PFKFB2, PRC1, PYH1N1, P2RY14, RIPK2, Rb, SLK1, SPARC, STAB1, **S100A8**, TGFB1, TOP2A, TP53, TXNIP, UBE2C, and WFS1. Again, the seven highlighted genes were common to both B- and T-ALL. From the total of 47 genes, 21 genes were from our list and 26 genes were added by IPA. Five genes were found to be active at all three time intervals: HNF4A, P2RY14, Rb, S100A8 and TOP2A. Three were from our list: P2RY14, S100A8 and TOP2A, and the rest

were added by IPA. HNF4A or Hepatocyte nuclear factor 4 alpha is involved in development of the liver, kidney, and intestines. P2RY14 is a protein encoded by P2Y purinoceptor 14 and is a member of the family of G-protein coupled receptors. P2RY14 is involved in the immune system and the regulation of stem cell compartments. Rb or retinoblastoma protein is found to be dysfunctional in several types of cancer. S100A8 (S100 calcium binding protein A8) plays an important role in inflammation-associated cancer. TOP2A or Topoisomerase 2-alpha is used as the target for anticancer agents.

The above analysis was based on GC-induced apoptosis genes at three time points separately, but involved a large amount of processing of several networks for each time point. Therefore, in the next step, we only selected genes that were found in at least in two of three time intervals from our list to create gene networks for T- and B-ALL patients using the IPA program. For B-ALL patients, the selected genes are described as follows (the asterisk denotes genes that were found multiple times in the input dataset):

- 0-6 hours, 6-24 hours and 0-24 hours (3): P2RY14, S100A8 and TOP2A.
- 0-6 hours and 0-24 hours (12): BUB1, CENPF, DLGAP5, EPPK1, FKBP5, GBP4, HHMR*, NUF2, PFKFB2, SIK1, SLA and UBE2C.
- 6-24 hours and 0-24 hours (8): CDC42EP3, CEP55, DTL, KIAA0101, MCM10, RRM2, STAB1 and TYMS.

For T-ALL patients, the selected genes are described as follows:

- 0-6 hours and 6-24 hours (5): BCL2L11, GNG11, HES1, S100A12, and ZNF24.
- 0-6 hours and 0-24 hours (1): TFPI.
- 6-24 hours and 0-24 hours (13): DEFA3, DTL, FEN1, HBG2, KIAA0101, MCM10, NRG1, PPBB, PMAIP1, RHOBTB3, STOM, TMSB15A and TUBB1.

There were more probe sets belonging to the above time categories but since they had unmapped gene id or gene symbol, they were not included in the sets. With the selected, reduced gene list, IPA was used again to construct a network for each time point and the results are shown in Tables 4.2 and 4.3 where genes from our list are in bold. Others were added by IPA. The selected corresponding networks are shown in Figure 4.1.

Table 4.2: B-ALL gene network based on the reduced gene set for 0-6 hours, 6-24 hours and 0-24 hours

ID	Molecules in Network	Score	Focus Molecules	Top Functions
0-6 hours				
1	AP4B1, BRAP, BUB1 , CCNB1, CDK1 , CENPE, CENPF , DLGAP5 , EGFR, EPPK1* , FBXO7, FKBP5 , GBP4 (includes EG:115361), HMMR , HNF4A, IFNG, IL5, MIS12, NDE1, norepinephrine, NUF2 , P2RY14 , PDE4B, PFKFB2 , PPME1, RBMS3, S100A8 , SIK1 , SLA , SMOC2, SPC25, TGFB1, TOP2A* , UBE2C , YWHAZ	51	16	Cell Cycle , Cellular Assembly and Organisation, DNA Replication, Recombination, and Repair
6-24 hours				
1	ASB3, CDC7, CDC2B, CDC42EP3 , CDK1, CDK1/2, CEP55 , CKAP2, DLGAP5 , DTL , E2f, E2F1, FEN1, GRWD1, Histone h4, HNF4A, Immunoglobulin, KIAA0101 , LATS2, MCM8, MCM10 , ORC3L, P2RY14 , PYHIN1 (includes EG:149628), Rb, RRM2* , RRM2B, S100A8 , SHOX, SMOC2, STAB1 , TOP2A* , TP53, TYMS , UMPS	35	11	Cell Cycle , Genetic Disorder, Metabolic Disease
0-24 hours				
1	beta-estradiol, BUB1 , CCNG1, CEP55 , E2F8, FEN1, FKBP5 , GBP4 (includes EG:115361), HELLS, HMMR* , IFNG, Immunoglobulin, KIAA0101 , KLC2, MKI67, NAP1L1, NDE1, PFKFB2 , POLA1, POLD1, POLD3, RAD51AP1, RFC3, RFC4, RPA, RPRM, S100A8 , SIK1 , SLA , STAB1 , TIMM50, TOP2A* , TP53, TPX2, YWHAZ	38	12	DNA Replication, Recombination, and Repair, Cell Cycle, Cancer
2	AURKB, BUB1 , catechol, CDC6, CDC7, CDC45L, CDK1, CDK1/2, CDT1, CENPF , DLGAP5 , DTL , E2f, E2F2, FZR1, hCG, Histone h4, hydroquinone, leucovorin, MCM8, MCM10 , NUSAP1, ORC2L, ORC3L, ORC6L, PFKFB3, POLA1, Rb, RRM2* , SIRT2, SMOC2, TOP2A* , TYMS , UBE2A, UBE2C	30	9	Cell Cycle, DNA Replication, Recombination, and Repair, Cancer
3	CCDC53, CDC42EP2, CDC42EP3 , CEBPB, CETN3, DSN1, E2F4, EPPK1* , ERBB2, FOXO1, GSTK1, HNF1A, HNF4A, MAP3K3, MIS12, NDC80, NSL1, NUF2 , P2RY14 , PHB2, PLDN, RB1, SMAD4, SMC1A, SPC24, SPC25, TGFB1, TP53, TRAF6, UMPS, ZW10, ZWINT (includes EG:11130)	9	4	Cell Cycle , Gene Expression, Cellular Assembly and Organisation

Table 4.3: List of T-ALL genes in networks based on the reduced gene set for 0-6 hours,6-24 hours and 0-24 hours

ID	Molecules in Network	Score	Focus Molecules	Top Functions
0-6 hours				
1	2-methoxyestradiol, AIFM1, BAK1, BCL2A1, BCL2L2, BCL2L11 , Cbp, CDC6, CFD, DHFR, ENDOG, F9, FBXO32, FOXG1, GCLC, GSR, heparin, HES1 , HUWE1, LTBP1, LY6A, NKX2-2, NOTCH4, NOV, PTMA, S100A12 , SPHK2, SPN, stearic acid, TFPI , TNF, TNFSF13, TNFSF13B, TP53, ZNF24 (includes EG:7572)	13	5	Cell Death , Cellular Growth and Proliferation, Haematological System Development and Function
2	AMOTL2, G protein beta gamma, G-protein gamma, GNAI1, GNAI2, GNAI3, GNB1, GNB2, GNB3, GNB4, GNB5, GNG2, GNG3, GNG4, GNG5, GNG7, GNG10, GNG11 , GNG12, GNG13, SMURF1, ZHX1	2	1	Cell-To-Cell Signalling and Interaction, Cellular Assembly and Organisation, Cell Signalling
6-24 hours				
1	BCL2L2, BCL2L11 , BNC1, BOK, CFBF, COL18A1, CTSG, D-sphingosine, DEFA3 (includes EG:1668) , DTL , DYNC111, Dynein, DYNLL2, EDN1, ENDOG, FEN1 , hCG, IL6, KIAA0101 , LPA, MCM10 , MYB (includes EG:293405), NR3C1, NRGN , PFDN1, PMAIP1 , POLD1, PPBP , retinoic acid, RNASE2, S100A12 , stearic acid, STOM , TCEB2, TUBB1	31	12	Cell Signalling, Molecular Transport, Vitamin and Mineral Metabolism
2	AMOTL2, Ca ²⁺ , Cbp, CCND3, CTNNB1, ERK, FANCA, GNG11 , HDAC5, HES1 , ID3, JAG1, Jnk, KAT5, MAPK8, MMP2, MMP9, NOTCH3, NR3C1, NUMB, PCNA, PRKCA, PTGDS, RBL1, RUNX2, SMAD2, SMAD3, SMARCB1, SMURF1, TMSB15A (includes EG:11013) , TP53, TP63, Vegf, ZHX1, ZNF24 (includes EG:7572)	8	4	Cellular Development, Organ Development, Cellular Growth and Proliferation
3	HGS, MIR302A (includes EG:407028), RHOBTB3	3	1	Cellular Assembly and Organisation, Cell Morphology, Organ Development
4	HBA1, HBA2, HBB (includes EG:3043), HBE1, HBG1, HBG2 , HBZ, hemin, SMAD5, UBQLN4	2	1	Protein Synthesis, Genetic Disorder, Haematological Disease
0-24 hours				
1	APOA2, beta-estradiol, CDC7, CDC45L, CDT1, COX17, CUL4A, DEFA3 (includes EG:1668) , DTL , EDN1, FEN1 , hCG, HUS1, KIAA0101 , LPA, MCM5, MCM6, MCM10 , MMP7, MMP1 (includes EG:4312), NRGN , PFDN1, PMAIP1 , POLA1, POLD1, PPBP , retinoic acid, RFC4, SLC2A1, STOM , TFPI , TMSB15A (includes EG:11013) , TP53, TUBB1 , VLDLR	33	12	DNA Replication, Recombination, and Repair, Cell Cycle , Cancer
2	HGS, MIR302A (includes EG:407028), RHOBTB3	3	1	Cellular Assembly and Organisation, Cell Morphology, Organ Development
3	HBA1, HBA2, HBB (includes EG:3043), HBE1, HBG1, HBG2 , HBZ, hemin, SMAD5, UBQLN4	2	1	Protein Synthesis, Genetic Disorder, Haematological Disease

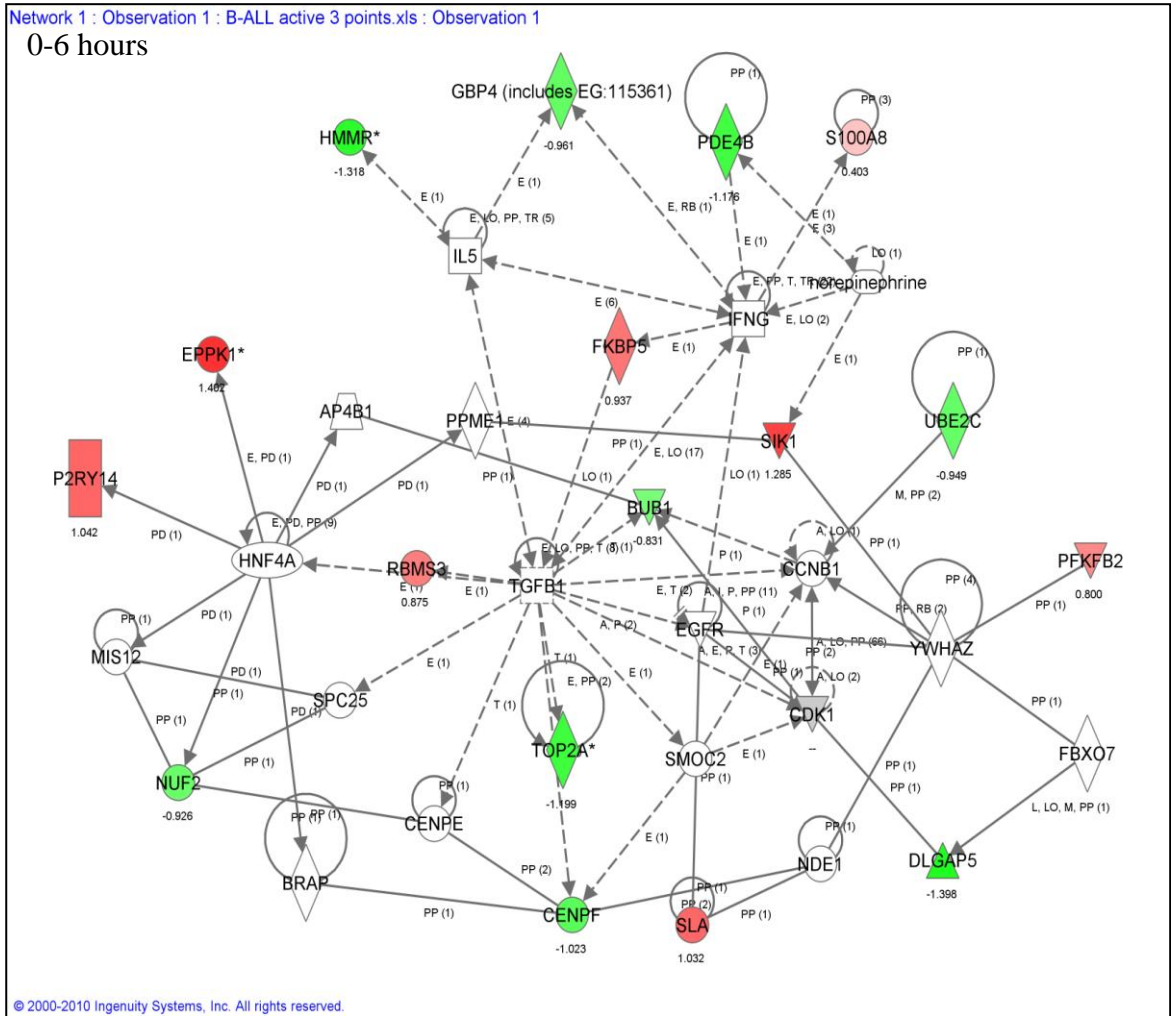


Figure 4.1: B-ALL gene networks at 0-6 hours

Coloured nodes are from our input list; red indicates over expression and green indicates under expression.

Table 4.4 presents the biological functions and pathways of the same networks from Tables 4.2 and 4.3. They indicated that T- and B-ALL patients were quite different in the Molecular and cellular functions, Canonical pathways and Functions, for example, molecular and cellular functions of T-ALL were involved more in cell death while B-ALL were more involved with cell cycling. This finding may imply that (i) the apoptosis process in T-ALL may occur before B-ALL during the same period of treatment (ii) there are many cells progressing through the cell cycle (cycling cells) in B-ALL while many non-cycling cells are in T-ALL. Many more pathways/steps are involved in cycling cells than in non-cycling cells in the apoptosis process after glucocorticoid treatment (King & Cidlowski, 1998). Genes found in both T-and B-ALL were involved with cancers functions.

We hypothesised that genes found active throughout the time period were dominant genes that may be needed in the GC-induced apoptosis process. In the next step, we identified how these genes were connected and in what form of relationship. We manually extracted relationships/connections from B- and T-ALL gene networks at three time intervals. We tried to minimise the gene network size by minimising the number of connected genes, for example, B-ALL patients, we first started from the three common genes for the three time intervals, and then we checked what other genes were connected to these genes; for example, TP53 interacts with HNF4A, and HNF4A interacts with DTL and P2RY14, and so on. We continued to add more connections between these genes.

Finally, we drew the possible gene relationships with the selected genes, as shown in Figures 4.2 and 4.3. For B-ALL patients, we started with the three common genes and extended to the others (Figure 4.3). For T-ALL patients, common genes were found for only two sets of time intervals (0-6 h and 6-24 h, and 6-24 h and 0-24 h); therefore, we started with these common genes separately and continued adding connections in the same way as for B-ALL; the results are shown in Figure 4.3 (a) and (b). For the two resulting networks, we then found four hub genes linking the two networks. These hub genes were BCL2L11, S100A12, TEPI and TP53. Obviously, from both curated B-and T-ALL gene networks, TP53 is the most intensive node which has many connections with other genes. The glucocorticoid receptor (GR/NR3C1) was found in the T-ALL gene network but not in the B-ALL gene network.

Table 4.4: List of top three activities of the selected genes from Tables 4.2 and 4.3 in molecular and cellular functions, canonical pathways and functions for T- and B-ALL patients at three time intervals by IPA

B-ALL			T-ALL		
0-6	6-24	0-24	0-6	6-24	0-24
Molecular and Cellular Functions (p-value)					
Cell Cycle (8.49E-06-3.17E-02)	Cell Cycle (9.16E-06-4.24E-02)	Cell Cycle (1.86E-11-4.89E-02)	Carbohydrate Metabolism (4.08E-04-4.08E-04)	Cellular Assembly and Organisation (4.50E-05-4.85E-02)	Cell-To-Cell Signalling and Interaction (1.697E-04-4.41E-02)
DNA Replication, Recombination, and Repair (1.02E-04-4.29E-02)	DNA Replication, Recombination, and Repair (6.25E-05-4.38E-02)	Cellular Assembly and Organisation (1.85E-11-4.65E-02)	Cell Death (4.08E-04-4.83E-02)	Cell Morphology (5.72E-04-4.85E-02)	Carbohydrate Metabolism (8.83E-04-3.22E-02)
Cellular Assembly and Organisation (4.67E-04-2.80E-02)	Cellular Movement (3.25E-04-2.43E-02)	DNA Replication, Recombination, and Repair (1.85E-05-4.89E-02)	Cellular Compromise (4.08E-04-1.42E-02)	Cell Death (1.15E-03-4.55E-02)	Cell Death (8.83E-04-4.67E-02)
Canonical Pathways (-log (p-value)/ Ratio)					
Cell Cycle: G2/M DNA Damage Checkpoint Regulation (1.32/0.023)	Cell Cycle: G2/M DNA Damage Checkpoint Regulation (3.281/0.045)	Role of CHK Proteins in Cell Cycle Checkpoint Control (2.516/0.057)	G Protein Signalling Mediated by Tubby (1.874/0.024)	Breast Cancer Regulation by Stathmin 1 (1.733/0.01)	Coagulation System (1.492/0.027)
Fructose and Mannose Metabolism (1.212/0.007)	Pyrimidine Metabolism (2.116/0.009)	Cell Cycle: G2/M DNA Damage Checkpoint Regulation (2.395/0.045)	Coagulation System (1.824/0.027)	G Protein Signalling Mediated by Tubby (1.426/0.024)	TR/RXR Activation (1.134/0.01)
Glioma Invasiveness Signalling (1.171/0.018)	One Carbon Pool by Folate (1.703/0.026)	Pyrimidine Metabolism (2.229/0.013)	Notch Signalling (1.813/ 0.023)	Notch Signalling (1.366/ 0.023)	P53 Signalling (1.115/0.011)
Functions					
Cell Cycle (8.49E-06-3.17E-02)	Cancer (8.15E-08-4.98E-02)	Cell Cycle (1.86E-11-4.89E-02)	Lymphoid Tissue Structure and Development (7.27E-05-1.70E-02)	Cancer (1.84E-05-4.96E-02)	Cancer (5.46E-05-4.96E-02)
DNA Replication, Recombination, and Repair (1.02E-04-4.29E-02)	Genetic Disorder (8.15E-08-4.80E-02)	Cancer (8.65E-06-4.89E-02)	Organ Development (7.27E-05-4.72E-02)	Gastrointestinal Disease (1.84E-05-1.15E-03)	Gastrointestinal Disease (5.46E-05-3.80E-03)
Cancer (1.62E-04-4.79E-02)	Skeletal and Muscular Disorders (3.25E-06-3.58E-02)	Gastrointestinal Disease (8.65E-06-4.89E-02)	Carbohydrate Metabolism (4.08E-04-4.08E-04)	Cellular Assembly and Organisation (4.50E-05-4.85E-02)	Cell-To-Cell Signalling and Interaction (1.697E-04-4.41E-02)

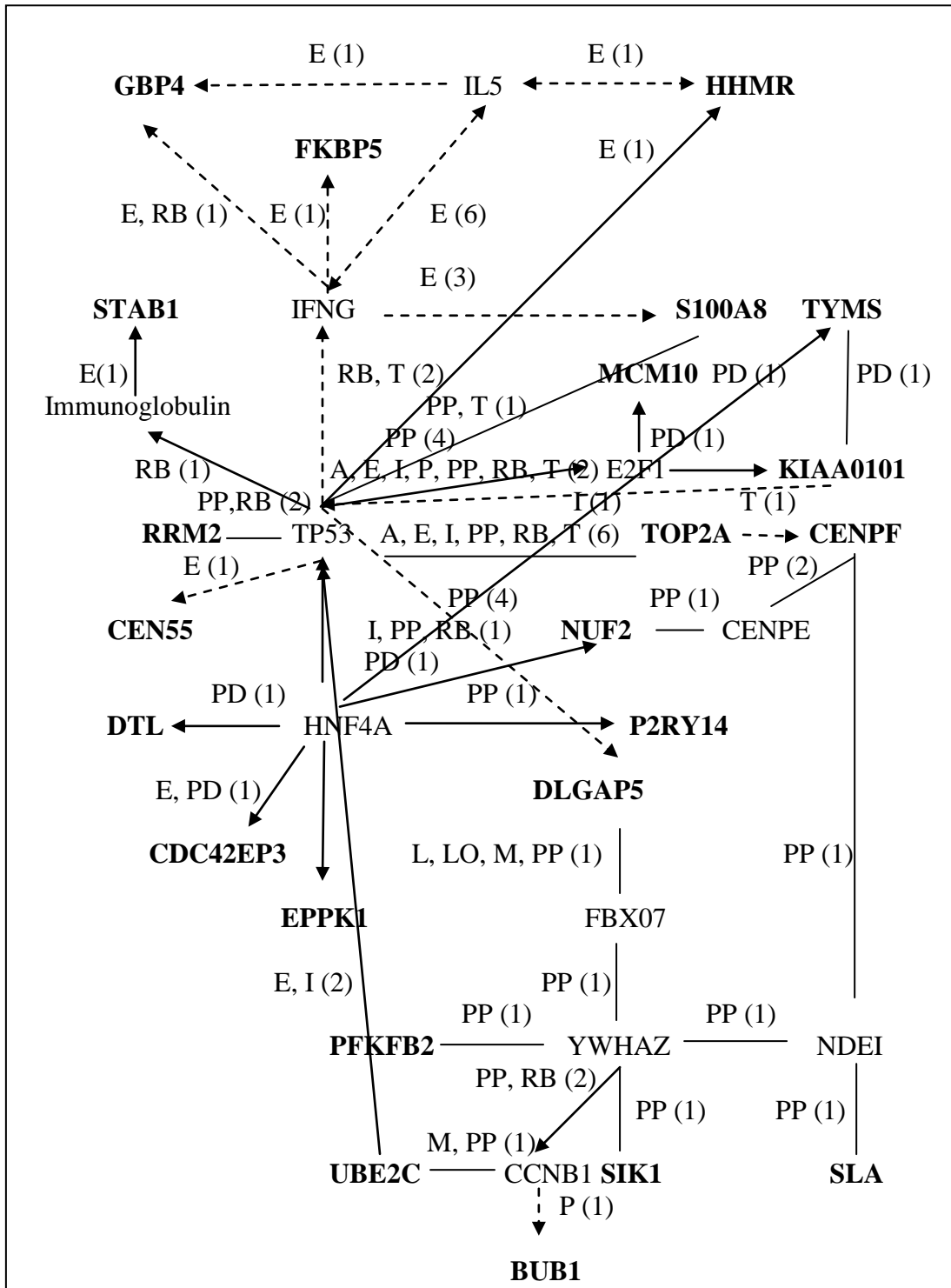


Figure 4.2: Proposed B-ALL GC-induced apoptosis gene network for all genes active in at least one time interval (0-6 hours, 6-24 hours and 0-24 hours)

Gene found in our study are highlighted in bold.

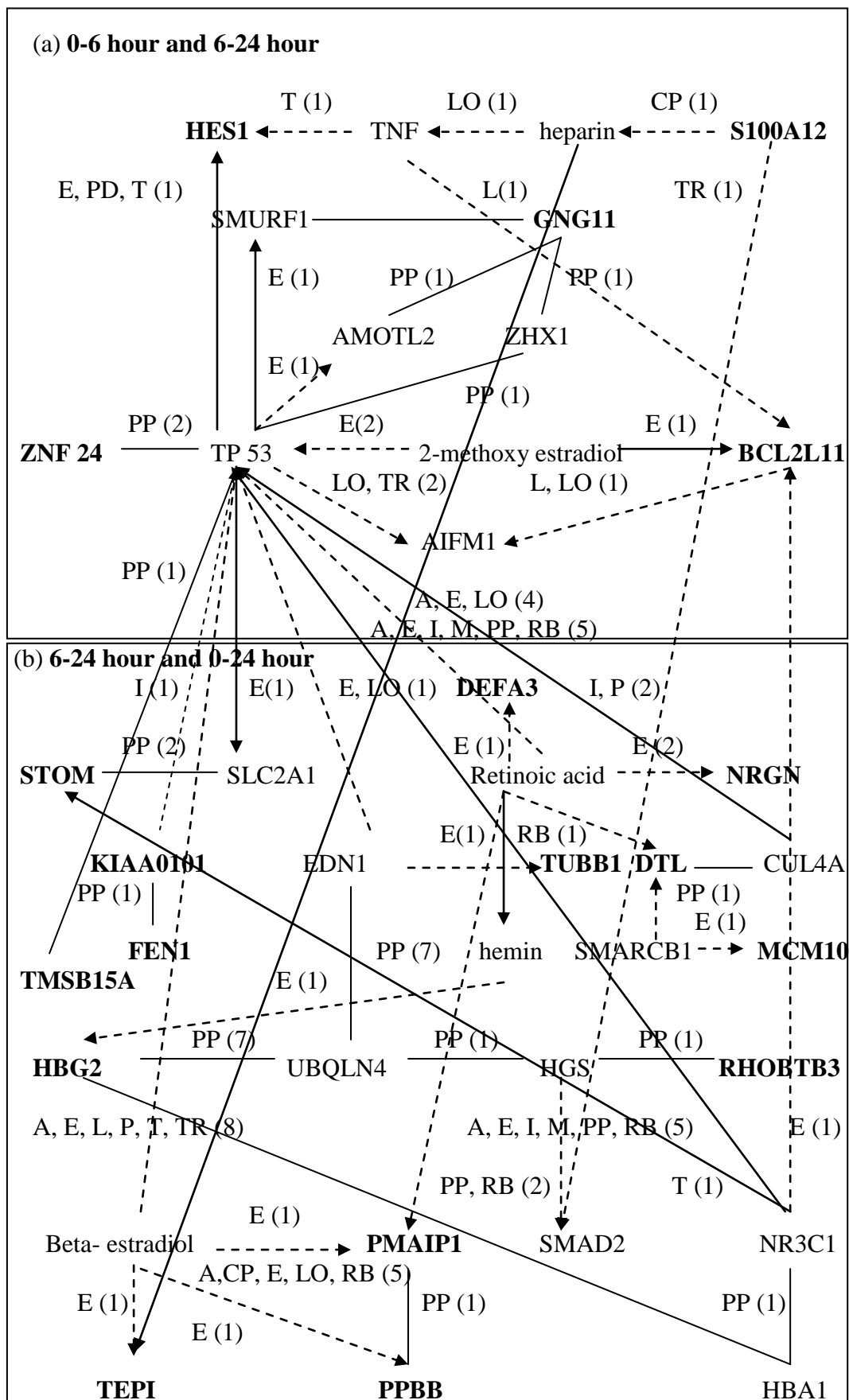


Figure 4.3: Proposed T-ALL GC-induced apoptosis gene network for all genes active in at least one time interval (0-6 hours, 6-24 hours and 0-24 hours)

Gene found in our study are highlighted in bold.

In summary, we elucidated gene networks for the three intervals using candidate GC-induced apoptosis genes. The number of inferred networks varied according to number of input genes. We used two input sizes in this study: (i) the whole gene set found differentially expressed at each time interval and (ii) genes found in at least two out of three time intervals. Different input gene sets produce many possible inferred networks; therefore, many genes appearing in these networks need further study. The question we try to answer after inferring gene networks from time series data is how to reduce the number of genes that need further study. We emphasised common genes (strong genes or prominent genes) because if they appeared more than once in a network, there was a high chance that this common gene needs to be activated in our input gene list.

We identified common genes by manually extracting connections from inferred gene networks for each time interval. In addition, results from IPA showed the different Molecule and cellular functions, Canonical pathways and Functions between T- and B-ALL. T-ALL is more involved with cell death while B-ALL is more involved with cell cycle. We proposed GC-induced apoptosis gene networks for T- and B-ALL separately. This network still needs further study because the networks were retrieved from relationships already mentioned in the literature from studies on different cells, tissues or diseases.

In the next section, we focus on elucidating the relationship between GC-induced apoptosis genes and GR/NR3C1 using genes found statistically significant probe sets from our STEM clustering that was presented in Chapter 3.

4.3.2 Inferring GR gene networks from selected genes from STEM

As shown in the previous section, genes extracted for the time intervals resulted in several possible inferred gene networks. Therefore, using Short Time series Expression Miner (STEM) to cluster genes before inferring networks can help reduce the number of networks as the attention here is on the temporal characteristics of data as a time series. The hypothesis behind clustering methods is that they cluster gene groups with similar gene expression patterns; these genes may have similar gene function or co-regulation. Therefore, we further analysed gene networks from the set of GC-regulated genes that were identified by the STEM clustering method and reported in Chapter 3 (327 probes were analysed for each patient separately). The 141 probe sets for T-ALL and 221 probe sets for B-ALL were used to construct gene networks. The networks were generated through the use of Ingenuity Pathway Analysis (Ingenuity System®, www.ingenuity.com). The list of genes above did not include the glucocorticoid receptor (GR) or NR3C1 gene for B-ALL even ± 2 patients would not have made NR3C1 pass the criteria. However, according to the literature, GCs mediated their function by binding with their receptor (glucocorticoid receptor: GR/NR3C1). Therefore, we aimed to understand the relationships between GC-regulated genes and GR/NR3C1 in the apoptosis process. Thus, we added gene expression of NR3C1 into the list of B-ALL patients as input into IPA.

Tables 4.5 and 4.6 which report all genes that form the gene networks and the selected gene networks are shown in Figures 4.4 and 4.5 for B-ALL and T-ALL, respectively. In Tables 4.5 and 4.6, the top three functions from B- and T-ALL still show that gene listed from B-ALL is involved in cell cycles and cancer whereas those from T-ALL are involved in cell death functions. Of all the networks for B-ALL in Table 4.5, only network 1 had NR3C1 and, for T-ALL in Table 4.6, only networks 2 and 5 had NR3C1. Figures 4.4 and 4.5 show networks for each time interval with focus on NR3C1 or GR gene from network number one of B-ALL (Table 4.5) and network number two of T-ALL (Table 4.6), respectively.

Table 4.5: List of B-ALL gene networks based on gene list including GR from STEM clustering method

ID	Molecules in Network	Score	Focus Molecules	Top Functions
1	BCL2A1, CDK6 , Cyclin A, E2f, E2F7, E2F8, ELL2, FCER1G, FKBP5 , Histone h4, Hsp90, IGHM, IL12 (complex), IL18R1, IL27RA , Interferon alpha, KIF15, LEF1, MCM10 , NFkB (complex), NR3C1, NUSAP1, OIP5 , Pias, PRDM1, RAG2 , Rb, RBMS3, RRM2, SOCS1, STAT5a/b, TPX2, TYMS , tyrosine kinase, ZBTB16	53	24	Haematological System Development and Function, Haematopoiesis, Cancer
2	ACP5, ACPP, AGTR2, AP1B1, ASPM , beta-estradiol, BUB1, BUB1B, BYSL, CCR1, CCR3, CDC42EP3, CTSH, DFNA5, ERG, FABP5, ICAM2, IGF2R, KIF4A, KLF9 , Mhc ii, MKI67, NCAPD2, NCAPG (includes EG:64151), NCAPH, PRC1, SMC2, SMC4, SPARC, STAB1, TGFB1, TRO, TXNIP, WFS1, ZWINT (includes EG:11130)	29	15	DNA Replication, Recombination, and Repair, Cell Cycle , Cellular Assembly and Organisation
3	BCAT1, CEP55 , CTNN β -LEF1, CTNNB1, DEPDC1, DGKE, DPM1, DSN1, FZD8, GINS2, GSTM4, HELLS, HNF4A, HSPH1, LY6A, MIR124-1, MIS12, MYC, NUF2, P2RY14, PAICS, PDE4B, PFKFB2, PHB2, PP2A, PSAT1, SNX9 (includes EG:51429), SOX17, SPN, Tcf/lef, TCF7L2 (includes EG:6934), TP53, TPX2, UBE2T, VPS37C	24	14	Cell Cycle , Cellular Assembly and Organisation, DNA Replication, Recombination, and Repair
4	ANAPC5, ATAD2, ATP, BUB3, CDKN2A, CDKN2D, DEFA3 (includes EG:1668), DTL, EPPK1, GFRA1, GINS1, ICAM3, IL6, KIAA0101, KIF11, MELK, Na+,K+ -ATPase, Oas, OAS1, OAS2, ORC4L, P2RX1, P2RX2, P2RX3, P2RX4, P2RX5, P2RX6, PHB2, retinoic acid, SHC1, SHCBP1, TARSL2, TMEM97, TRAF6, UBA1	21	12	Protein Synthesis, Cardiovascular System Development and Function, Cell Morphology
5	ARPP-21, CASP3, CCL6, CCL23, CD24, CD8A, CSF1R, CTSH, CTSL2, DARC, FAM171A1, FEZ1, FGD2, FGL2, GRIN1, HLA-F, HLA-G, HTT, IFNG, IFNGR1, IL13, KIF18A, LILRB2, MS4A4A , neuroprotectin D1, NFKBIB, PAPP, PDE4B, PRTN3, RAB27A, RPS6KA2, S100A8, SERPINF1, SHANK1, SIK1	19	11	Cellular Movement, Immune Cell Trafficking, Organismal Injury and Abnormalities
6	Akt, ANXA7, Ap1, CD3E, CFLAR, CYP8B1, ERK, ERK1/2, FCGR1A/2A/3A, FGR, FSH, HSP90AA2, HSP90AB1 , Ige, IGHE, IL31, IRS1/2, LGALS3, MGST1, MMP12, P38 MAPK, PDE4B, PDGF BB, PI3K, PIK3IP1 , Proteasome, PRSS3 (includes EG:5646), RAD51AP1 , RNA polymerase II, SERPINA1, SHB, SLA, SPHK1, SRC, TCR	13	8	Cellular Movement, Haematological System Development and Function, Immune Cell Trafficking

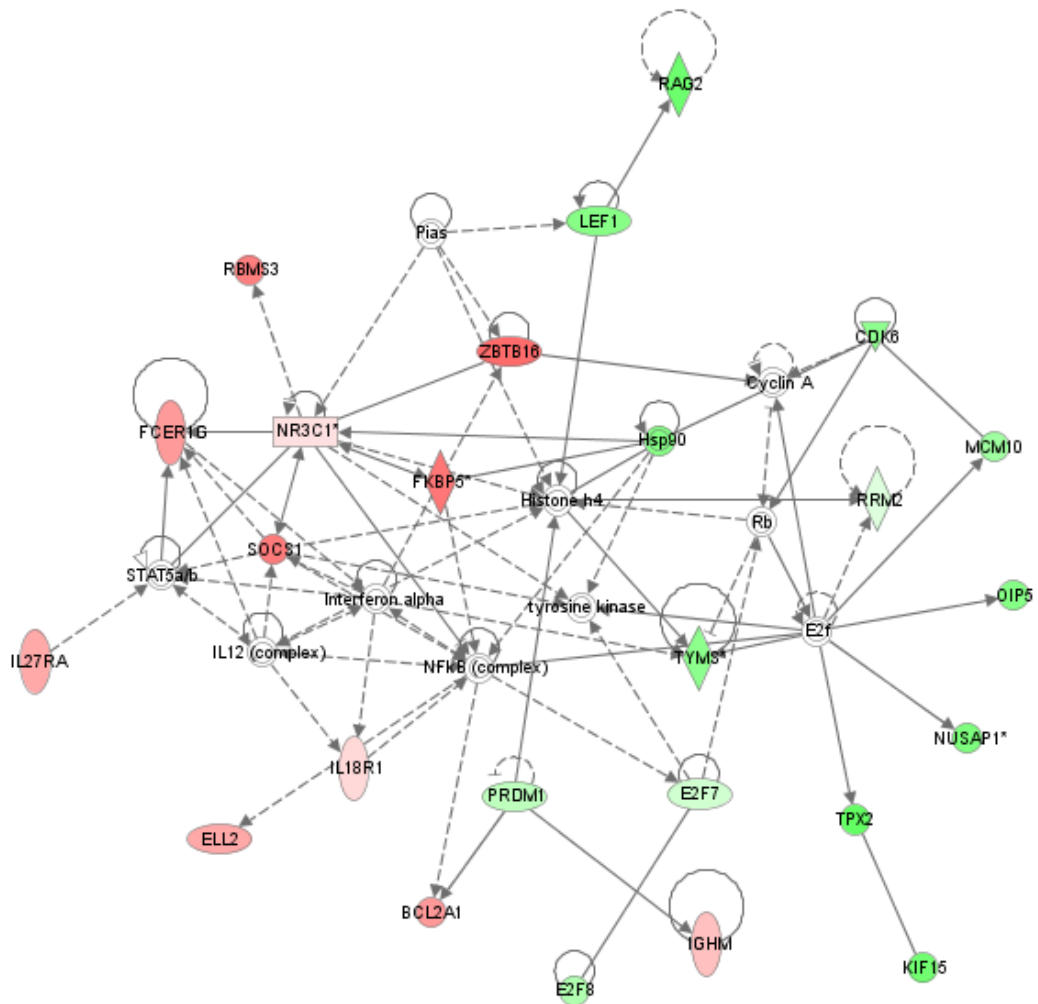


Figure 4.4: Glucocorticoid receptor gene networks for B-ALL at 0-6 hours

Gene network curated from network 1 from Table 4.4. Coloured genes are from our list.

Table 4.6: List of T-ALL gene networks based on gene list including GR from STEM clustering method

ID	Molecules in Network	Score	Focus Molecules	Top Functions
1	ARHGEF7, BCL2L11, DTL, E2f, ERAF, hCG, HES1, Histone h3, Histone h4, HLA-DQB1, HLA-DR, HLA-DRA, ID3, Ifn gamma, ITPKB, MCM10, MHC Class II, NFkB (complex), NRG, PF4, PMAIP1, PPBP, Proteasome, PTGDR, Rb, RNA polymerase II, RUNX2, SMARCC2 (includes EG:6601), SOCS1, TPX2, TYMS, tyrosine kinase, UHRF1, Vegf, ZNF24 (includes EG:7572)	49	22	Cell Death, Haematological System Development and Function, Cell-To-Cell Signalling and Interaction
2	Akt, Ap1, ARNTL, ASB3, BMF, CAMP, CCR7, Collagen(s), CTF1, CTSG, DEFA3 (includes EG:1668), ERK, GLP2R, IL1, IL1F9, IRS1/2, Jnk, LRRK1, LYZ, Mapk, MPO, NR3C1, P38 MAPK, PDGF BB, PI3K, PIK3IP1, Pkc(s), PRKCA, PSMC3IP, S100A8, S100A9, SPARC, STOM, Tgf beta, TNFSF8	30	15	Cell Death, Antigen Presentation, Cell-mediated Immune Response
3	AHSG, ANKZF1, ARPC5, ATN1, C4A, CA2, CA1 (includes EG:759), CLTCL1, CREB1, CYP11A1, DBT, EDN1, ELL, EP300, FEN1, Fgfr, FRMD6, GHRHR, HNF4A, IFNAR1, ING4, KIAA0101, L-triiodothyronine, LPIN1, MEGF11, Na⁺,K⁺ -ATPase, PHKB, SCAND1, SIK1, SLC18A2, SNX5, SRC, TUBB1, WASL (includes EG:8976), ZYX	20	11	Cellular Development, Connective Tissue Development and Function, Cell Morphology
4	ASPM, ATAD2, BRE, Ca²⁺, CCR7, CD27, CSGALNACT1, CTF1, CXCL9, DAD1, EIF5A, ETHE1, GNB1, GNG11, GSR, ING4, IRAK3, IRF2, IRF9, NOTCH3, PIM2 (includes EG:11040), PMAIP1, PPP1R13L, PRMT2, PRNP, RAD51AP1, RELA, RPL5 (includes EG:6125), S100A12, S100P, SLC2A4, SPI1, TMSB15A (includes EG:11013), TNFSF9, TP53	20	11	Cellular Development, Cellular Growth and Proliferation, Cell Death
5	AKT1, BCL2, BIRC2, CBL, CCNB1, CCND1, CCND2, CCND3, CDC34 (includes EG:997), CDKN1A, CDKN1B, CHEK1, CTNNB1, CUL1, DLGAP5, EGFR, ELF4, FBXL3, FBXO7, HNRNPU, HSPD1, MLF1, NR3C1, PFKFB2, RBL1, RBX1 (includes EG:9978), SKP1, SMAD3, SMAD4, TNF, YWHAG, YWHAH, YWHAQ (includes EG:10971), YWHAZ, ZFP36	5	4	Cell Cycle, Cancer, Cellular Development

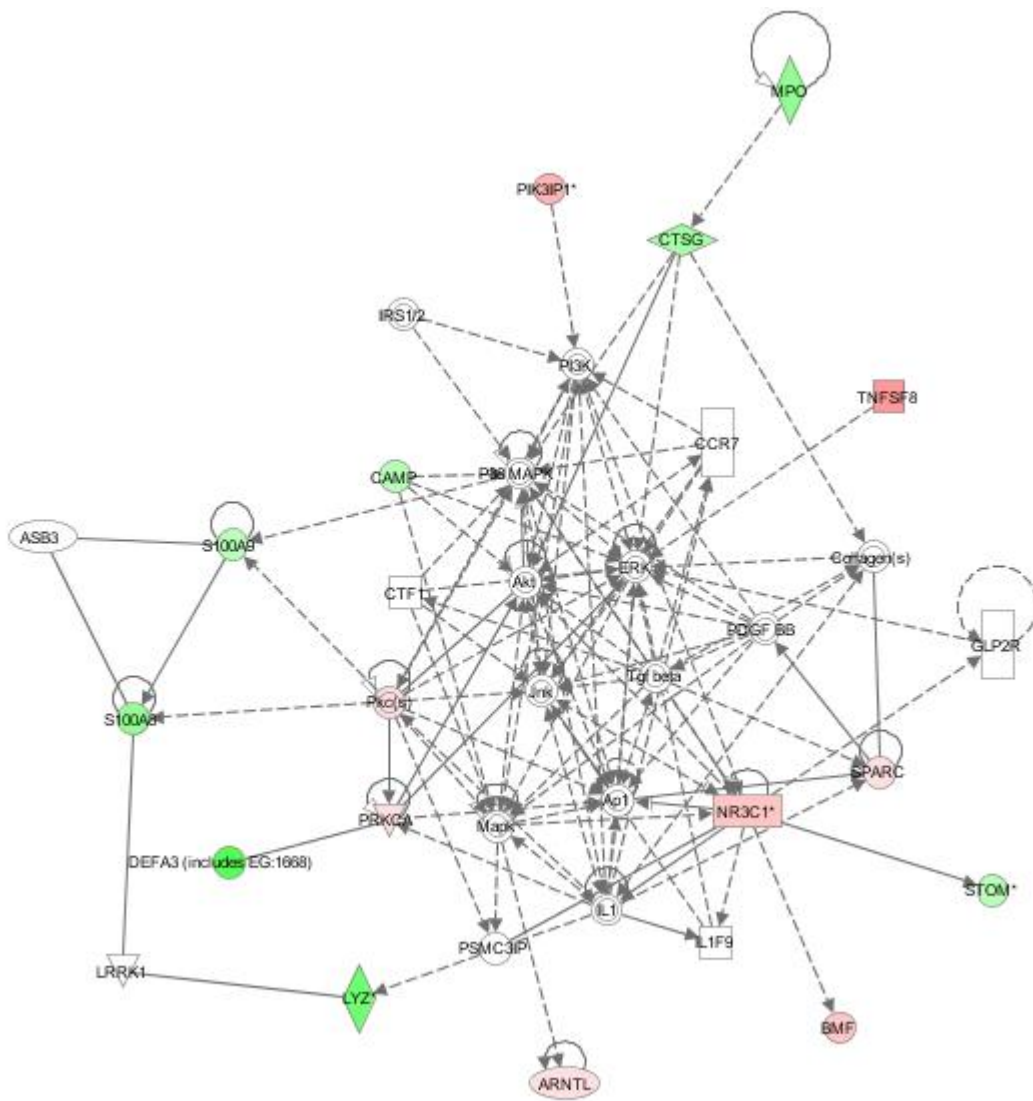


Figure 4.5: Glucocorticoid receptor gene networks for T-ALL at 0-6 hours

Gene network curated from network 2 from Table 4.6. Coloured genes are from our list.

In Figure 4.4, the B-ALL gene network containing NR3C1 found connections with FCER1G, FKBP5, Histone h4, Hsp90, NFkB, Pias, RBMS3, STAT5a/b, SOCS1, tyrosine kinase and ZBTB16. From Figure 4.5, the T-ALL gene network containing NR3C1 found connections with Akt, BMF, ERK, GLP2R, IL1, Jnk, LIF9, Mapk, PSMC3P, STOM and Tgf beta. These two GR gene networks from B- and T-ALL will be used to combine with the relevant pathway networks and existing GR gene networks selected from literature which will be used to expand the existing GR gene network, in Chapter 6.

4.4 Conclusions

After identifying the set GC-regulated genes in Chapter 3, in this chapter, we focused on finding the gene networks of GC-regulated genes in childhood leukaemia. First, the GC-regulated genes from Chapter 3 were used to infer gene networks at three interval time points: before treatment and six hours after treatment (0-6 hours), six and 24 hours after treatment (6-24 hours) and before treatment and 24 hours after treatment (0-24 hours)). We inferred networks from two different sizes of input dataset with the focus on finding prominent or predominant genes through common genes. We manually extracted gene connections from the inferred networks at three time intervals and proposed GC-induced apoptosis gene networks for T- and B-ALL (Figures 4.2 and 4.3). The major finding of this study from inferring gene networks using IPA software was that we found a set of three genes (P2RY14, S100A8 and TOP2A) that remained active after treatment for the whole 24 hour period for B-ALL but none for T-ALL. Seven more common genes (ASPM, E2f, HNF4A, KIAA0101, MYC, NFkB (complex) and S100A8) were added by IPA to both T- and B-ALL networks with other two extra genes (HNF4A and Rb) added to B-ALL.

Next, GCs activate by binding to the glucocorticoid receptor (GR). To understand the relationship between GR and gene network of GC-induced apoptosis genes, we then focused on finding the connection between gene networks of GC-induced apoptosis genes with the GR gene network. The GR gene network was created from a gene set selected by the STEM clustering method from Chapter 3. Inferred gene networks from gene clusters by STEM added more connected genes to NR3C1. For B-ALL (Figure 4.4), NR3C1 found connections with FCER1G, FKBP5, Histone h4, Hsp90, NFkB, Pias, RBMS3, STAT5a/b, SOCS1, tyrosine

kinase and ZBTB16. In addition, for T-ALL (Figure 4.5), NR3C1 found connections with Akt, BMF, ERK, GLP2R, IL1, Jnk , LIF9, Mapk, PSMC3P, STOM and Tgf beta.

The next step of our study, in Chapter 5 will investigate GR gene networks that may control the GC-induced apoptosis mechanism which, in turn, will help to better understand the GCs-induced apoptosis process.

Chapter 5

Inferring Gene Networks from Microarray data and Pathway Databases

5.1 Introduction

Systems Biology, the concept of multi-disciplinary, systematic and holistic study of biological systems, is now becoming rooted in the biomedical field. The integration of available information (from genes to gene networks/pathways) leads to better understanding of complex biological systems including human response to treatment. The biological system of gene expression and their products can be classified as transcriptional regulatory, protein interaction, metabolic and signal transduction networks. These represent connections and relationships between tissues, cells, genes, gene products (protein), networks and pathways. This information provides a higher level of understanding of phenomena and cellular processes in an organism. To date, gene expression profiling is the most widely used approach to construct gene networks from gene expression data. Gene expression data mainly come from publicly available databases. Due to increasing availability of high throughput data, especially, microarray data, a database has been developed to store and standardise the information for further analysis. Lists of existing public databases, internet-based platforms and software for networks/pathways analysis (construction, data mining, and visualisation) can be found in various sources such as Babu, (2008) and Tsui, Chari, Buys, & Lam, (2007). The knowledge of unknown gene function and transcriptional regulatory networks are not retrieved when using a pathway-based method. Therefore, integration with clustering methods may help to identify function of unknown genes assigned to the same function or pathway of known genes in the cluster (Cavalieri & De Filippo, 2005).

Gene expression data from microarray technology is used extensively in genome wide analysis of cancer, for example, gene clustering and inferring gene networks. A number of methods, including machine learning and mathematical approaches have been used to construct gene networks from synthetic and experimental data. In addition, microarray data can be used to infer gene networks by using forward and reverse engineering methods. Reverse engineering or inferring gene networks is the process of reconstructing networks from experimental systems. A genetic network is an interaction between genes and its products which indicate the regulation between genes. Gene networks are often known as

gene regulatory networks and they are large-scale interactions and connections of the cell at mRNA level. However, these networks do not give the whole picture of regulatory networks. For example, networks inferred from time series gene expression data only describe a phenomenon in the form of: every time that gene A is downregulated (underexpressed), gene B is upregulated (overexpressed). Thus, an inferred network gives only part of the whole actual regulatory process; however, this is an important step that will lead to further investigation (Ferrazzi & Bellazzi, 2007).

The aim of gene network inference is to retrieve interactive or dynamic networks from given data. Usually, the network can be represented as a graph where nodes represent genes and edges represent the relationship or interaction between connected genes. The relationship may indicate coexpression or coregulation of genes, which may share regulatory inputs, common pathways, biological function, location or process. Many studies on network inference have been carried out using artificial intelligence, statistical methods and mathematical methods, including Boolean networks, Differential equations, Neural Networks, and Stochastic models. A network/pathway database is another method currently used after microarray data analysis. Many studies start with the experimental process, then pre-process and analyse data to obtain differentially expressed genes and, finally, IPA or similar program is used to infer gene networks (Panetta, Evans, & Cheok, 2005; Phillip et al., 2005; Raponi et al., 2004; Winter et al., 2007).

The main focus of this study is glucocorticoids-induced apoptosis in childhood leukaemia treated with prednisolone. Generally, the apoptosis mechanism is mediated through two major pathways: the intrinsic (mitochondrial) pathway and extrinsic (cell death receptor) pathway. Many researchers have reported that the apoptosis mechanism initiated by conventional anticancer drugs is via the intrinsic pathway (R. Kim et al., 2002). In addition, glucocorticoid-induced apoptosis involves other pathways. Knowledge of the relationship and crosstalk between the apoptotic pathways and other pathways can increase the understanding of the apoptosis process. Defects in apoptotic pathways cause resistance in cancer cells during chemotherapy. The response to chemotherapy may be mediated by the involvement of pathways including p53-dependent and independent mechanisms (Min et al., 2006). Figure 5.1 shows the relationship of chemotherapeutic drugs and other factors that together affect the intrinsic apoptosis process. After treatment, chemotherapeutic drugs induce apoptosis through

microtubule damage or the DNA damage pathway and, subsequently, these pathways connect with the intrinsic and extrinsic apoptosis pathways. The chemotherapeutic drug referred to in this study is prednisolone, which is a synthetic glucocorticoid intensively used to treat childhood leukaemia. Generally, it is commonly known that the effect of glucocorticoids is mediated by the glucocorticoid receptor (Bachmann et al., 2007; Costlow, Pui, & Dahl, 1982; Kofler, 2000; Tissing et al., 2003; Tonko et al., 2001). Thus, knowledge of GR gene networks can enhance the understanding of GC-induced apoptosis mechanisms. Only few studies have been conducted on GR gene networks and these were not based on time series microarray data (Donn et al., 2007; Miller, Komak, Webb, Leiter, & Thompson, 2007; Phillip et al., 2005; Webb et al., 2003). As mentioned previously, in Chapter 3, we found that ALL subtypes: T- and B-ALL have some common and distinct genes. The inferred networks are proposed as GC-induced apoptosis networks for T- and B-ALL separately. Nevertheless, many studies have focused on finding common networks and pathways for childhood ALL. Therefore, we combined all possible networks from our data analysis and included previously proposed networks from the literature. We used three different gene sets to infer gene networks: (i) genes selected in this study by Short Time series Expression Miner (STEM) clustering method (ii) gene selected from three pathways of relevance to the apoptosis process and (iii) genes reported in the previously mentioned study by Phillip et al. (2005). These gene sets were analysed mainly by using Ingenuity Pathway Analysis software (IPA), only the last gene set was further analysed with BiblioSphere Pathway Edition (BSPE) and Oncomine. Then, a comparison and combination among these networks was carried out and proposed as a GR gene network.

The specific objective in this chapter is to understand behaviour of known genes from the apoptosis, p53 and NF κ B pathways. In addition, to study the effect of different tissues and chemotherapeutics drugs on the inferred GR gene network. Finally, to maximise use of publicly available pathway databases to identify GR gene networks. The outcome of this chapter is a comprehensive inferred GR gene network that might be used for further clinical study.

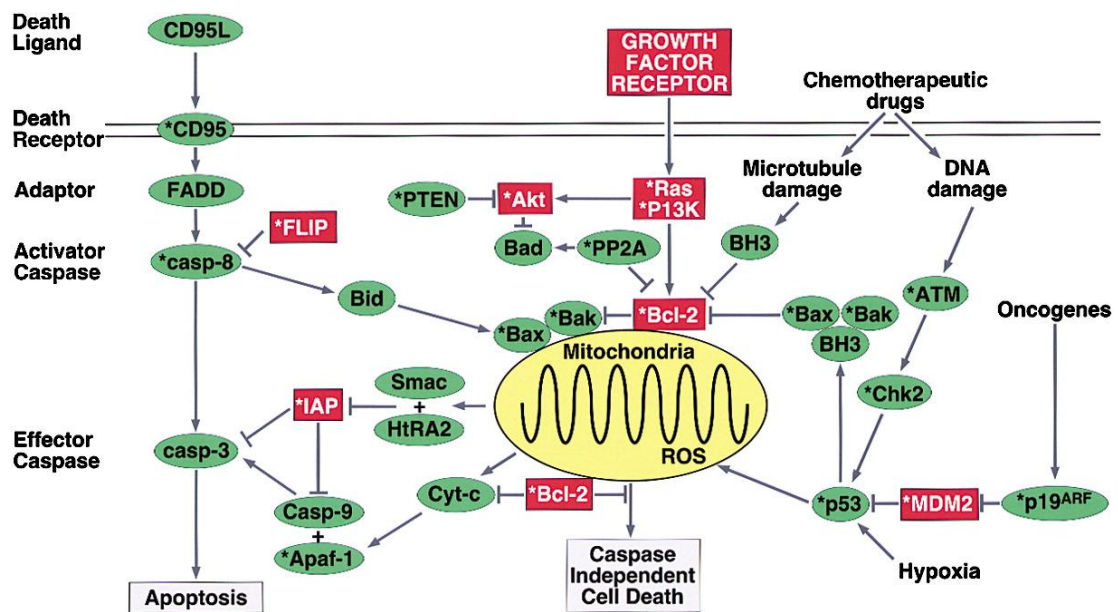


Figure 5.1: The Integrated Apoptotic Pathways

A schematic diagram showing some of the known components of the intrinsic and death receptor apoptotic programs that may modulate tumour development and therapy. An asterisk denotes components that are frequently mutated or aberrantly expressed in human cancers. Components in red inhibit apoptosis while those in green promote apoptosis. Abbreviations used: casp, caspase; cyt, cytochrome (Johnstone, Ruefli, & Lowe, 2002).

5.2 Methods

5.2.1 Dataset

The data used in this part is the same as that used in the previous chapter. Specifically, the secondary microarray dataset by Schmidt et al. (2006) is used. Raw data in the format of CEL files and normalised microarray data were obtained online from the Gene Expression Omnibus (GEO). Raw data were collected from 13 patients (three T-ALL patients and ten B-ALL patients) at three time points: 0 hour, 6/8 hours, and 24 hours.

5.2.2 Computational Methods

Gene network analysis and recovery of key biological pathways in this study were conducted using three network/pathway tools, summarised as follows:

(1) Ingenuity Pathway Analysis software (IPA) (Ingenuity® Systems, Redwood City, CA, USA, <http://www.ingenuity.com>). IPA is a web-based application that integrates a systems biology approach to solve various biological problems. The knowledge base of IPA comes from journal articles, textbooks and other data sources. This software has many applications; only functional analysis of genes and their networks have been used in this study. The p-value defines the significance of gene function in a network as well as gene to gene relation, and a p-value less than 0.05 signifies a statistically significant and non-random association. The right-tailed Fisher Exact Test is used to calculate the p-value.

(2) The BiblioSphere Pathway Edition (BSPE) (Genomatix Software, Munich, Germany, <http://www.genomatix.de>). This software can be used to analyse gene relationship networks, which combines literature analysis (from Pubmed), gene annotation and promoter analysis. Statistics are performed to check over- or under-represented groups of genes by using Z-score. The intensity of node (gene) varies from red to blue, denoting overexpression to underexpression, respectively.

(3) The Oncomine (<http://www.oncomine.org>) is a knowledge-based database curated from existing literature of human cancer gene expression profiles and integrated data-mining platforms. The differentially expressed genes are analysed using t-statistics and corrected measure of significance by using false discovery rate. As of 22 June, 2009, there were 41

cancer types with 392 studies and 28,880 microarray experiments available for further analysis with integration of other 18 bioinformatics resources.

5.3 Results and Discussion

The aim of our study was to better understand the GC-induced apoptosis mechanism via two major GCs issues: GC-induced apoptosis genes and GR gene network. After identifying candidate GC-induced apoptosis genes in Chapter 3 and their network in Chapter 4, we focused on GR gene networks for childhood leukaemia in this chapter.

5.3.1 Inferring GR gene networks from selected genes from three pathways (apoptosis, p53 and NF κ B)

Generally, there are two main apoptosis signalling pathways: the extrinsic and the intrinsic Apoptosis is regulated by various death inducing signals and interplay of several initiator, regulator and executioner genes. After receiving the apoptotic stimulus, the biochemical reactions and signalling pathways lead to apoptosis through several molecules, for example, Bax (the prototypic pro-apoptotic protein), Apaf-1 (apoptotic protease-activating factor 1), and caspases 9. In addition, the molecular mechanism of apoptosis signalling pathways is activated through anti- and pro-apoptotic molecules: the Bcl-2 family. The Bcl-2 family of proteins is essential to induce the apoptosis process; it can be either pro-apoptotic such as BH3-only, BAD, Bax and Bim (Bcl-2L11) or anti-apoptotic such as Bcl-2, Bcl-x_L and Bcl-w.

We first selected five vital genes commonly referred to in the literature about extrinsic and intrinsic pathway to investigate their behaviour before and after treatment. From the **extrinsic pathway**, the five selected genes were: **caspase 8** (cysteine-aspartic acid protease (caspase)), **caspase 10**, **FADD** (Fas-Associated protein with Death Domain), **FAS** (tumor necrosis factor receptor superfamily, member 6) and **TRADD** (tumour necrosis factor receptor type-1-associated DEATH domain protein). The five selected genes from the **intrinsic pathway** were: **Apaf-1** (apoptotic protease activating factor 1), **Bad** (bcl-2-associated death promoter), **Bcl2** (B-cell CLL/lymphoma 2), **caspase 3** and **caspase 7**.

None of these genes passed our criteria for finding differentially expressed genes. Their log ratios for the three time intervals (0-6, 6-24, and 0-24 hours) were relatively low, varying from 0.005 (mostly) up to 0.7. Only few genes with ratios above 0.9 were found in one out three patients for T-ALL and one or two out of ten patients (B-ALL). We can speculate from these results about why known apoptosis genes were not differentially expressed: (i) the time frame for apoptosis process can vary, taking up to 96 hours; Thomson and Johnson (2003) studied the apoptosis time frame for gene regulation after CEM-C7 were exposed to dexamethasone. They indicated that 24 hours after treatment was still pre-apoptotic and a reversible process (Thompson & Johnson, 2003). Therefore, the selected data only represented the early stage of the GC-induced apoptosis mechanism. (ii) The selected threshold (two fold) may be too high, but the question still remains as to how low a threshold could be and still provide biologically statistically and meaningful differentially expressed genes. Overall, it was possible to conclude that ten of known vital genes from **extrinsic and intrinsic apoptosis pathways** were not yet activated within 24 hours after the treatment.

Next, we looked into the inferred networks from the three related pathways (apoptosis, p53 and NFκB). We considered these three pathways because apoptosis is involved with multiple processes and pathways. Many studies reported that the GC-induced apoptosis process in ALL was mainly involved with the intrinsic pathway (Laane et al., 2007; Ploner et al., 2005; Schmidt et al., 2004). Thus, we selected only known genes from the intrinsic pathway; in addition, we focused on the Bcl2 family instead of caspases. The reason why we added two pathways (p53 and NFκB) into our network analysis was because previous studies have shown their strong relationship with apoptosis pathways. In addition, p53 and NFκB play an important role in cancer research. First, p53 is a transcription factor and a tumour suppressor protein which plays an important role in the control between apoptosis and survival. There are many positive and negative feedback loops involved with p53, as reviewed by Harris and Levine (Harris & Levine, 2005) and a recent review of p53 can be found in Batchelor, Loewer, and Lahav (2009). A recent study on crosstalk between p53 and apoptosis process can be found in Sun et al., (2009). Second, a model of GC-induced apoptosis in leukemic cells showed a connection between NFκB and the GC-GR complex when it was translocated to the nucleus (Tissing et al., 2003).

Fifteen key genes from the three pathways (apoptosis (intrinsic pathways) and five each from p53 and NFκB) were selected and highlighted in bold (Table 5.1) and detailed, as follows:

- Intrinsic pathway: **Apaf-1*** (Apoptotic Peptidase Activating Factor), **Bcl2** (B-cell CLL/lymphoma2), **BAD** (Bcl2-antagonist of cell death), **BAX** (Bcl2-associated x protein), and **BCL2L1** (Bcl-2-like1). *Apaf-1 involved in both intrinsic and p53 pathways (Soengas et al., 1999).
- p53 pathway: **Apaf-1***, **CASP9** (Caspase 9, apoptosis-related cysteine peptidase), **MDM2** (Transformed mouse 3T3 cell double minute 2, p53 binding protein), **MYC**, and **TP53** (Tumour protein p53).
- NFκB pathway: NFκB family (Nuclear Factor Kappa-light-chain-enhancer of activated B cells) consist of **NFκB1**, **NFκB2**, **REL**, **RELA**, and **RELB**.

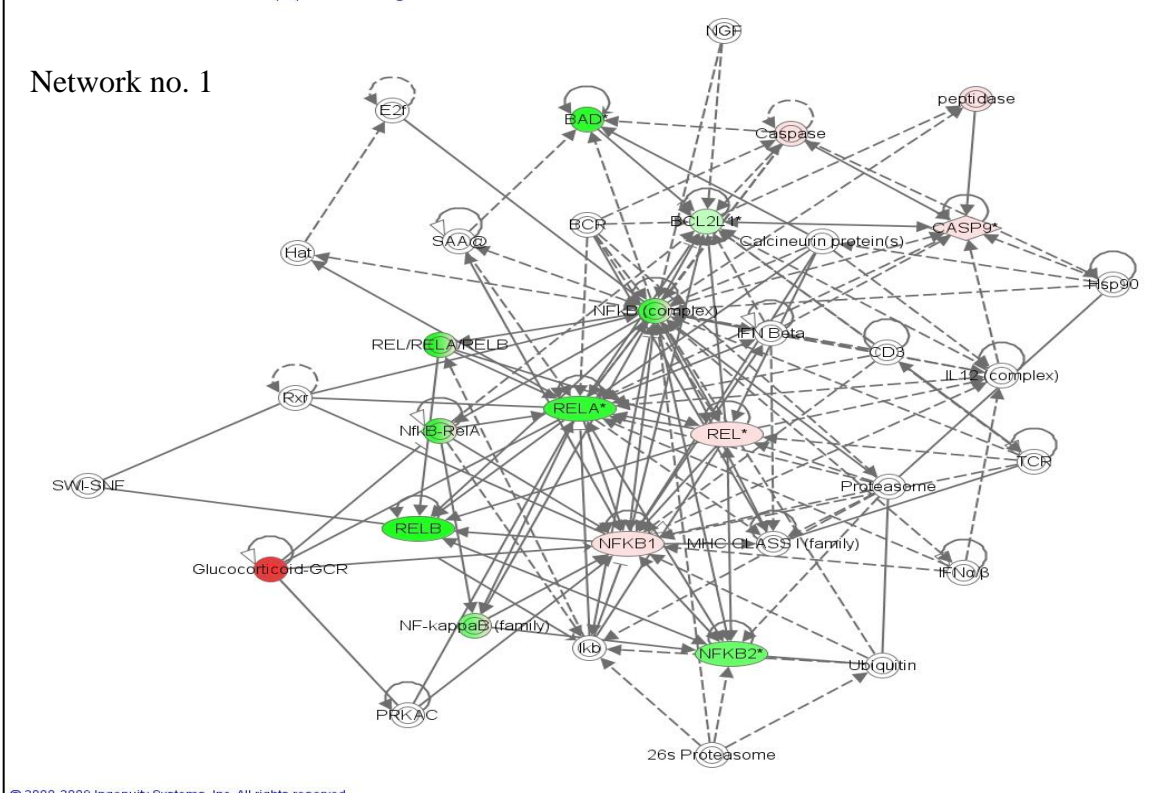
The total 15 genes mentioned above and two known genes (**NR3C1** and AP-1) involved with the GC-induced apoptosis process (Tissing et al., 2003). AP1 (Activator Protein 1) consisted of proteins belonging to the c-Fos and c-Jun families. In this study we selected: **FOS**, **FOSB**, **JUN**, **JUNB**, and **JUND**. Therefore, a total of 20 genes was input into IPA. It revealed gene network (s) common to the three pathways, as shown in Table 5.1, indicating three possible gene networks. In the table, the genes from the list of 20 genes are highlighted in bold and others are added by IPA. We further analysed the three networks with expression values for 20 genes from T-ALL data and the networks are shown in Figure 5.2. The generic network structures will be the same with B-ALL (data not shown). Only network number no.3 from IPA showed a gene connection with NR3C1: Akt, MDM2, MYC and TP53 or p53. The apoptosome (Apaf1, CASP9 and Cytochrome c) were found scattered in different networks (network no.1 and 2).

Table 5.1: Selected gene lists from Apoptosis, p53 and NFκB pathway

ID	Molecules in Network	Score	Focus Molecules	Top Functions
1	26s Proteasome, BAD , BCL2L1 , BCR, Calcineurin protein(s), CASP9 , Caspase, CD3, E2f, Glucocorticoid-GCR, Hat, Hsp90, IFN Beta, IFNα/β, Ikb, IL12 (complex), MHC CLASS I (family), NF-kappaB (family), NFKB1 , NFKB2 , NFkB (complex), NfkB-RelA, NGF, peptidase, PRKAC, Proteasome, REL , REL/RELA/RELB, RELA , RELB , Rxr, SAA@, SWI-SNF, TCR, Ubiquitin	17	8	Lymphoid Tissue Structure and Development, Organ Morphology, Gene Expression
2	Ant, APAF1 , BAX , BCL2 , Cdc2, Creb, Cyclin A, Cyclin D, Cytochrome c, ERK, FOSB , Growth hormone, Gsk3, hCG, Hexokinase, HISTONE, Ige, IgG, Il12 (family), Il8r, JUN , JUN/JUNB/JUND, JUNB , JUND , LDL, MAP2K1/2, Mek, Pdgf, PDGF BB, Pias, Pkg, PP2A, Rb, STAT5a/b, Top2	15	7	Behaviour, Nervous System Development and Function, Cancer
3	14-3-3, Akt, ALP, Ap1, Arf, Calmodulin, Calpain, Cbp/p300, Ck2, ERK1/2, Fgf, FOS , Hsp70, IL1, Insulin, Interferon alpha, Jnk, Mapk, MDM2 , MYC , NR3C1 , P38 MAPK, PI3K, Pkc(s), PLA2, Ras, RNA polymerase II, Sapk, Shc, SNCA, STAT, Tgf beta, Thyroid hormone receptor, TP53 , Vegf	12	5	Inflammatory Disease, Renal Nephritis, Renal and Urological Disease

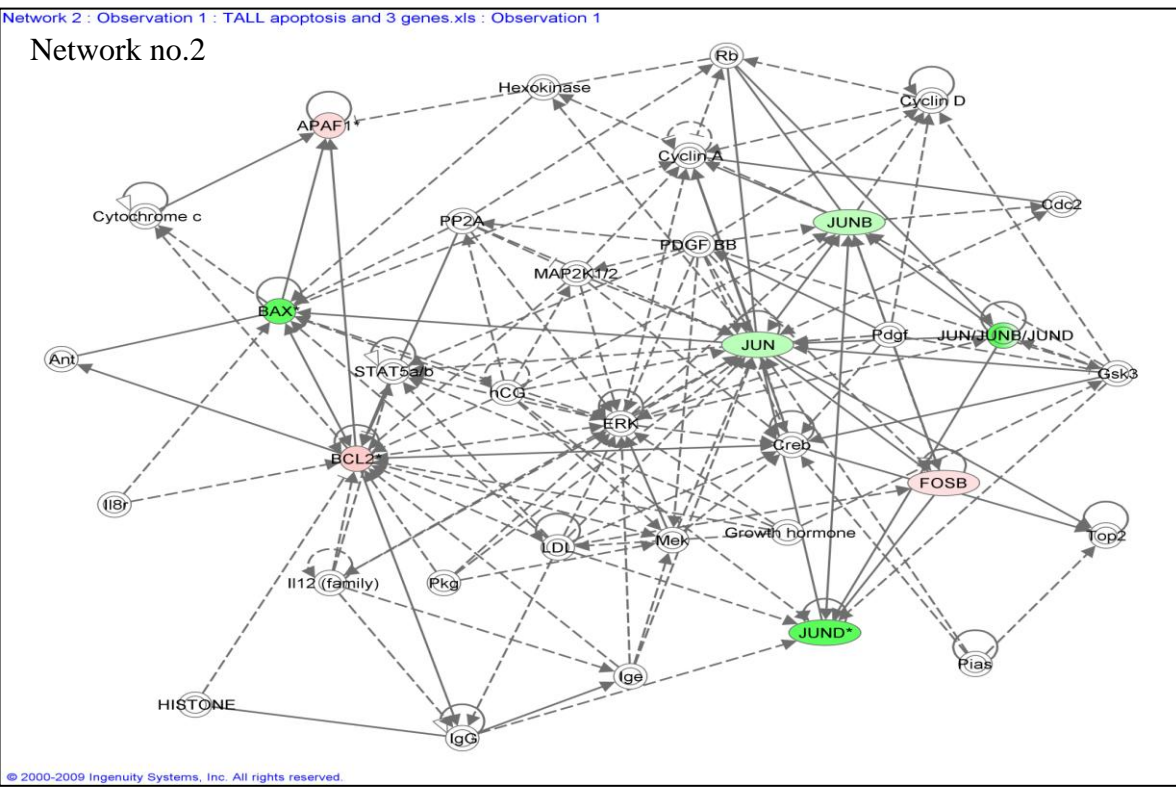
Network 1 : Observation 1 : TALL apoptosis and 3 genes.xls : Observation 1

Network no. 1



Network 2 : Observation 1 : TALL apoptosis and 3 genes.xls : Observation 1

Network no.2



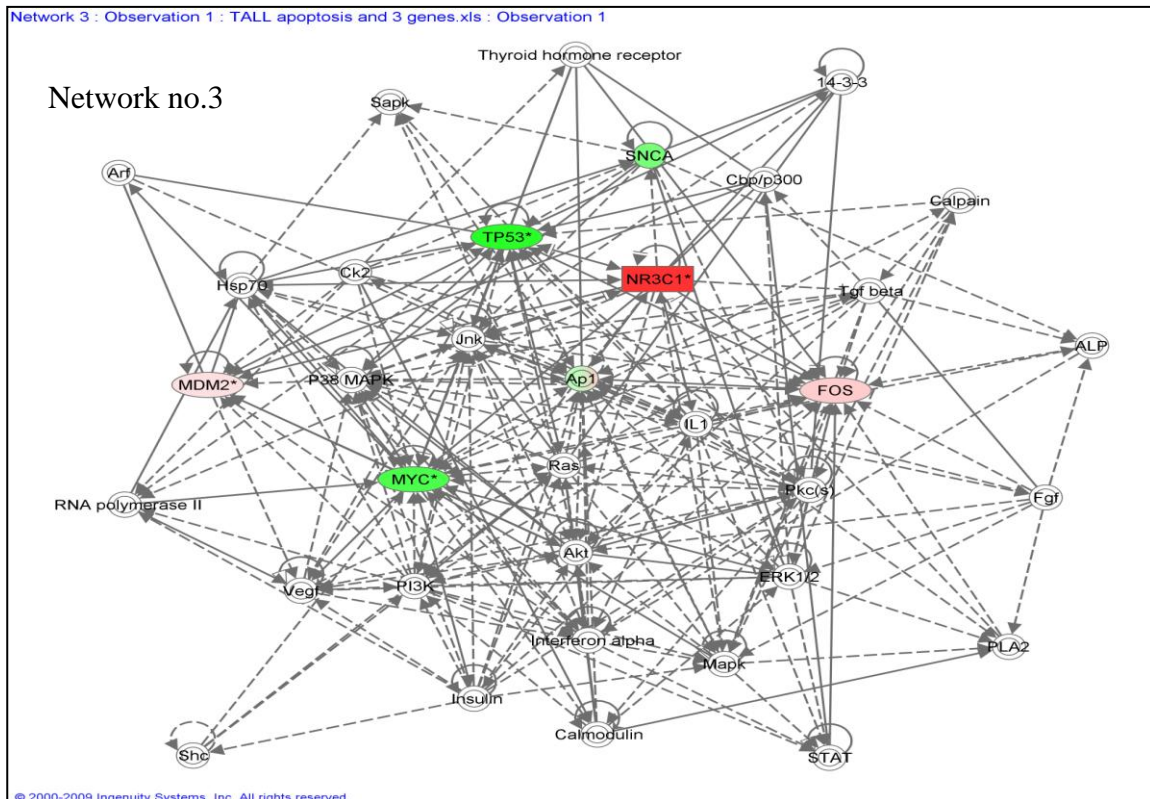


Figure 5.2: The three gene networks from Table 5.1

Figures illustrated with expression values for T-ALL patients (colours indicate intensity spectrum from highly upregulated (red) to highly down regulated (green)).

In Figure 5.2, there was no direct connection between NR3C1 and NFκB but after merging networks 1 to 3 (data not shown), three connections between NR3C1 and NFκB were found, details of their connections follow:

- NR3C1 → NFκB (i) activation: GR protein is involved in activation of NFκB (ii) inhibition: GR increases inhibition of NFκB
- NR3C1 → NFκB1 (i) expression: in U2-OS cells, human GR alpha A protein is involved in expression of human NFκB1 mRNA (ii) protein-protein interactions: binding of human p50 (NFκB-p50) protein and human GR (NR3C1) protein occurs in cell extracts from COS cells.
- NR3C1 → NFκB2 (i) expression: in U2-OS cells, human GR alpha A protein is involved in expression of human NFκB2 mRNA (ii) protein-protein interactions: binding of human NFκB2 protein and human GR (NR3C1) protein occurs.

The details of NFκB connections given above, as well as overall results, showed that the inferred network from IPA only gave genes (from the literature) connected to the uploaded gene set without being limited to any specific cells/tissues/processes. Furthermore, we used input genes that we thought could possibly be involved with the apoptosis process, the final inferred network gave a gene network based on literature from other studies that may or may be not involved with the apoptosis process. Nevertheless, the inferred network may give information for clinicians and scientists to use for further investigation.

Although the IPA networks highlight generic gene network structures, gene connectivities in IPA have been rigorously validated from literature, therefore, the networks generated in this study may be useful for constructing a possible GR network, as discussed in section 5.3.3. Prior to final section, we study one more aspect: GR gene networks from previous studies.

In the following section, three selected web-based knowledge network/pathway tools were used to construct a glucocorticoid receptor gene network based on prior studies.

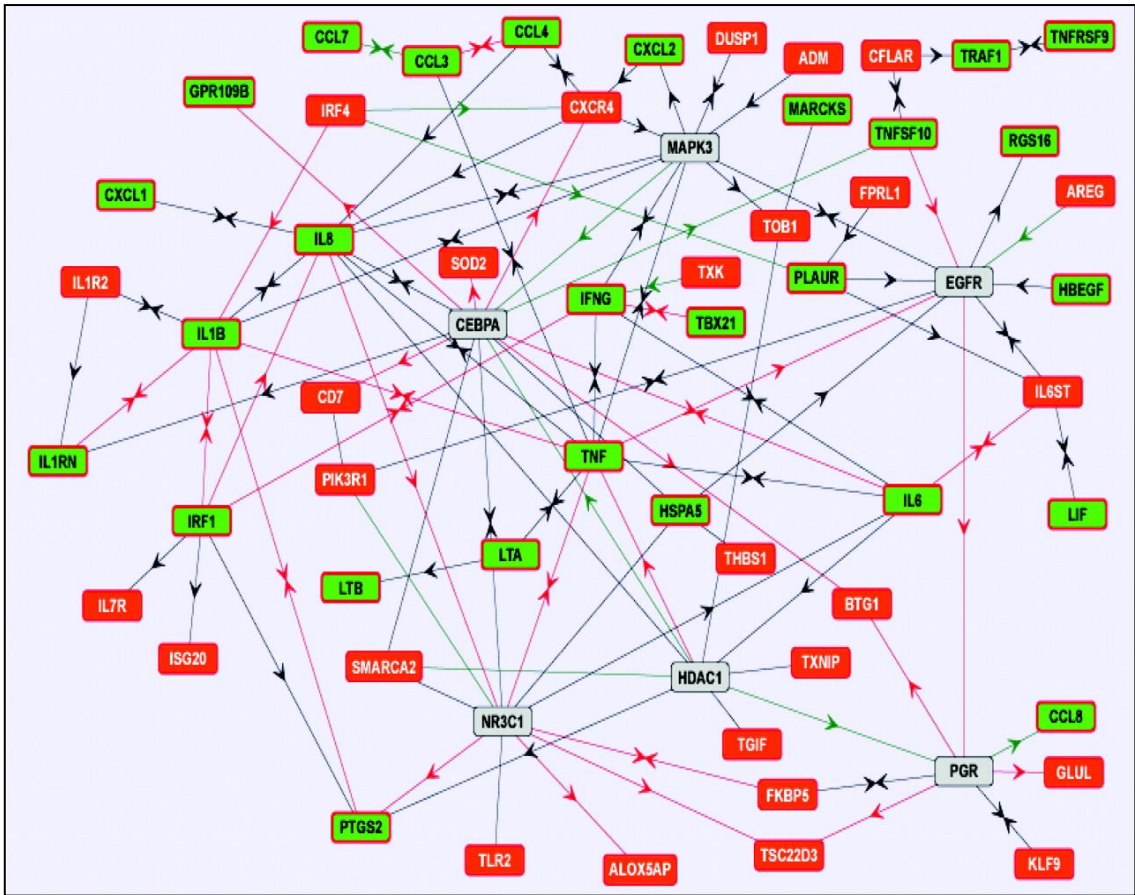
5.3.2 Inferring GR gene networks using genes found vital in previous studies

In this section we analysed gene networks based on genes from previous studies. Additionally, we investigated whether there was an effect from drugs or tissues on gene networks. We reviewed previous studies on GR network and found three studies: (i) Donn et al. (2007) (Figure 5.3a) and (ii) Miller et al. (2007) (Figure 5.3b) and (iii) Phillip et al. (2005) network (Figure 5.3c).

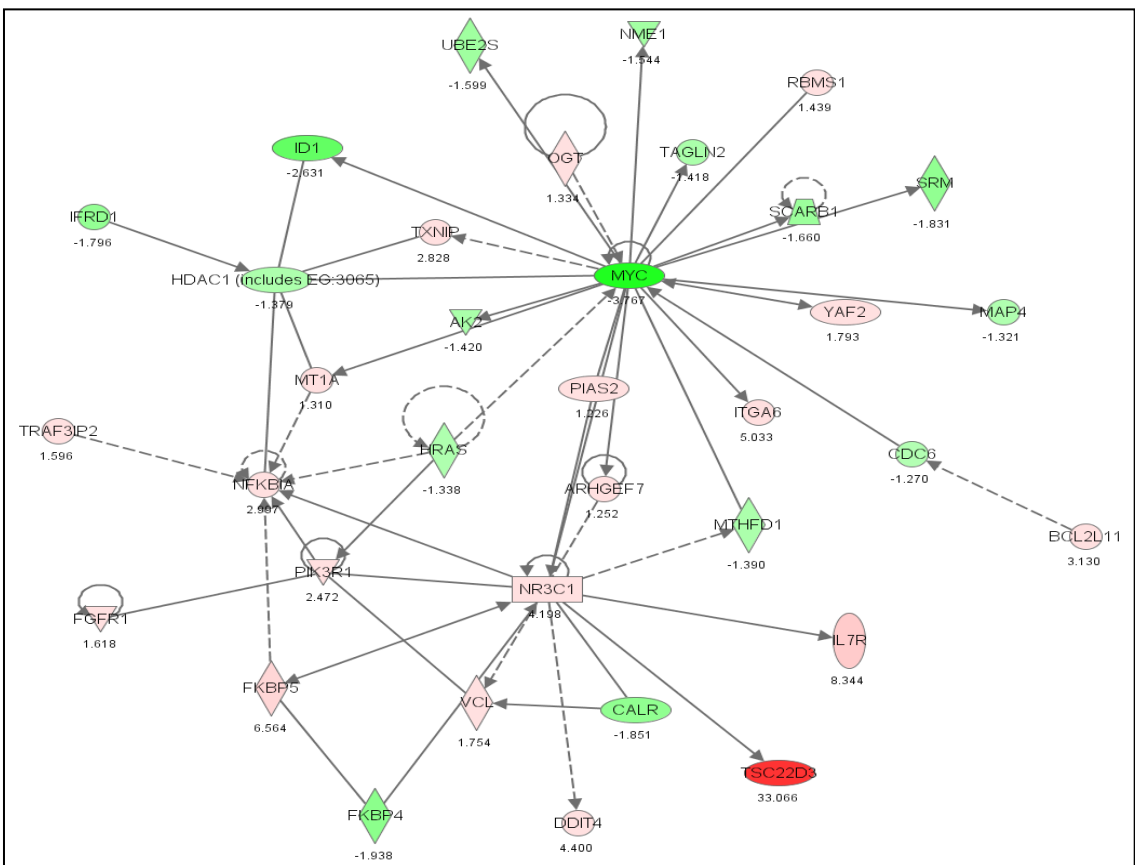
We selected Phillip et al.'s (2005) study for further investigation because our study was based on humans treated with prednisolone whereas the Phillip et al. (2005) study was in mouse livers treated with dexamethasone three hours before being sacrificed. Phillip et al. (2005) illustrated glucocorticoid receptor gene networks using data from two high-throughput technologies (microarray and genome wide location analysis (ChIP-on-Chip)). The 53 overlapping probe sets (23 genes) between the two methods were used to create a GR gene network (in Figure 5.3c) using Ingenuity Pathway Analysis (Ingenuity Systems, <http://www.ingenuity.com>) (Phillip et al., 2005). We constructed a GR gene network based on Schmidt et al.'s (2006) microarray data and pathway analysis (two selected network/pathway software (IPA and BSPE)) by using the 23 differentially expressed genes from for T-ALL and B-ALL. Our GR gene network was then compared with the three glucocorticoid networks mentioned above.

The gene expression level of the selected 23 genes is shown in Table 5.2. We extracted the gene expression level of the 23 genes from Phillip et al. (2005), as these data were retrieved three hours after treatment; therefore, we extracted gene expression levels six or eight hours after treatment from the Schmidt data. One out of 23 genes (SERPINA1) was found differentially expressed in B-ALL only under our new criteria, as discussed in Chapter 3. Also, NR3C1 was found differentially expressed only in T-ALL. The Table 5.2 shows the average gene expression level for T- and B-ALL for the 23 genes. Eight out of twenty three genes have the same expression pattern (up-regulation or down-regulation) between Phillip et al. (2005) and our study: CKS1B, FNTA, HSPCB, IGFBP1, MKNK2, NR3C1, and TXN (noted with * in Table 5.2).

a)



b)



c)

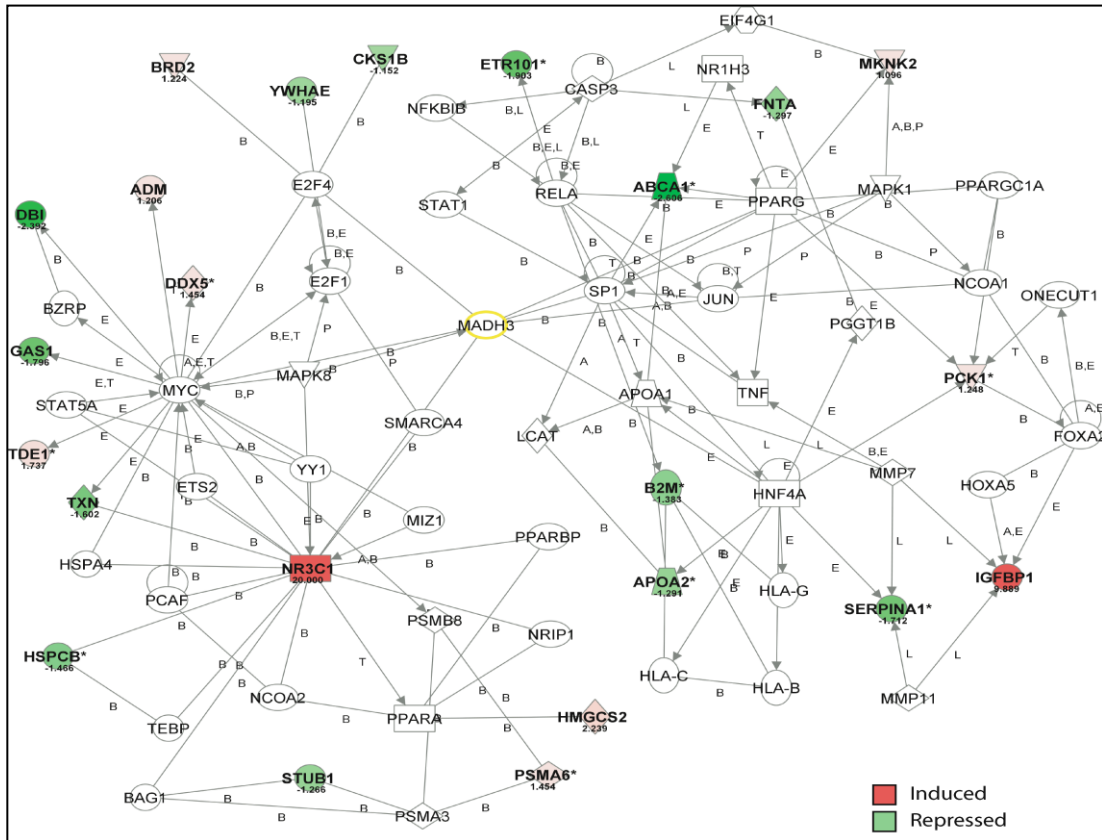


Figure 5.3: Existing GR gene networks from three studies (Donn et al., 2007; Miller, Komak et al., 2007; Phillip et al., 2005)

a) **Network analysis of the transcriptome response to glucocorticoid in human, primary T lymphoblasts.** Induced genes are shown in red and repressed genes in green. Linking nodes not contained in the query set are shown in gray. Relationships involving changes in expression levels are coloured red (increase) and green (decrease). All other relationships are represented by a black line. Arrows indicate the orientation of the relationship (Donn et al., 2007). b) **A signalling network links genes regulated by GCs in CEM cells.** Ingenuity® bioinformatics pathway analysis tool was used to connect a subset of 35 genes from the CEM signatory list based upon a database of published observations. Symbols for genes representing specific categories of cellular molecules as well as interactive relationships are depicted in the legend. Colour gradations are based upon gene regulation at the fold-change level. Red: induced gene; green: repressed gene. Fold-change data from CEM-C7–14 cells treated with Dex are presented as representative of the CEM signature (Miller, Komak et al., 2007). c) **A Regulatory Network for the GR.** Pathway analysis was seeded with the 53 differentially expressed and GR-bound genes, plus the GR itself, as described in Materials and Methods. Genes in coloured, bold text were in the seed set, while all others were brought into the network by the pathway analysis program based on their known relationships to the genes in the seed set. Colour indicates induction (red) or repression (green) of expression (Phillip et al., 2005).

Table 5.2: Comparison of gene expression between the original article by Phillippe et al. (2005) and our data for T-ALL and B-ALL patients

Gene	Article	T-ALL	B-ALL	Gene	Article	T-ALL	B-ALL
ABCA1	-2.606	0.349	0.151	HSPCB*	-1.466	-0.091	-0.313
ADM	1.206	-0.046	0.583	IGFBP1*	9.689	0.019	0.058
APOA2	1.291	-0.259	-0.089	MKNK2*	1.096	0.035	0.279
BRD2	1.224	-0.132	0.107	NR3C1*	20	0.594	0.178
B2M	-1.383	0.427	-0.267	PCK1	1.248	-0.074	0.075
CKS1B*	-1.152	-0.242	-0.485	PSMA6	1.454	-0.158	-0.073
DB1	-2.392	0.310	0.120	STUB1	-1.266	-0.041	0.132
DDX5	1.454	0.038	-0.105	SERPINA1	-1.712	-0.564	1.098
ETR101	-1.963	-0.191	0.451	TDE1	1.737	-0.020	-0.096
FNTA*	-1.297	-0.219	-0.097	TXN*	-1.602	-0.316	-0.161
GAS1	-1.796	0.218	-0.653	YWHAE*	-1.195	-0.275	-0.372
HMGCS2	2.239	-0.091	0.090				

The asterisk (*) indicates the same expression pattern (up or down regulation) in all.

Networks were constructed with both- BSPE and IPA and results are shown in Figure 5.4 (a) and (b) for BSPE and Figure 5.4 (c) and (d) for IPA. IPA networks included gene, protein, enzyme, transcription factor, nuclear receptor, kinase and peptidase, while BSPE networks can be viewed as depicting gene-gene relations or gene-transcription factor relationships.


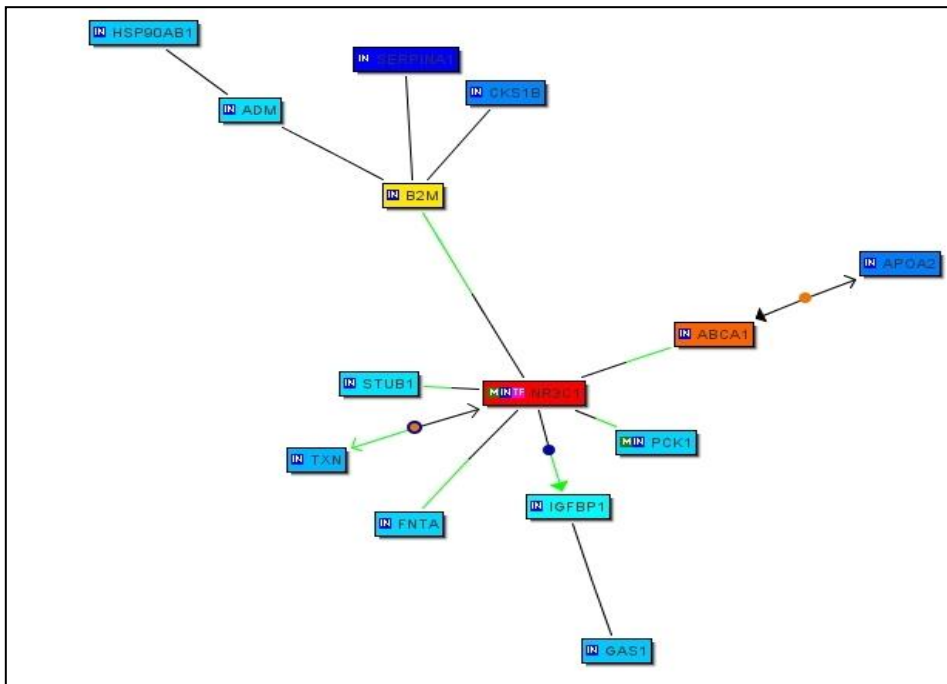
Figures 5.4 a and b show gene networks of NR3C1 or GR from BSPE - Bibliographic relationships for differentially expressed gene profiles analyzed with the Genomatrix Bibliosphere software tool. Arrow heads indicate the type of functional relationship between the connected genes. The half line green () connection means a gene encoding for a transcription factor is connected to a gene with the binding site for this transcription factor in its promoter, for example, NR3C1 and B2M or FNTA in Figure 5.4 a. A Genomatrix expert verified that gene-gene relationships are indicated by a blue circle in the centre of the connection line, e.g., NR3C1 and IGFBP1 in Figure 5.4 a. Red indicates up-regulated genes and blue indicates down-regulated genes during 0 to 6 hours of T-ALL and B-ALL patient. The colour intensity indicates the level of up- or down-regulation.

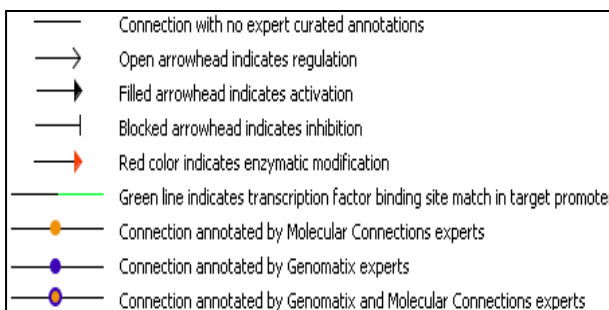
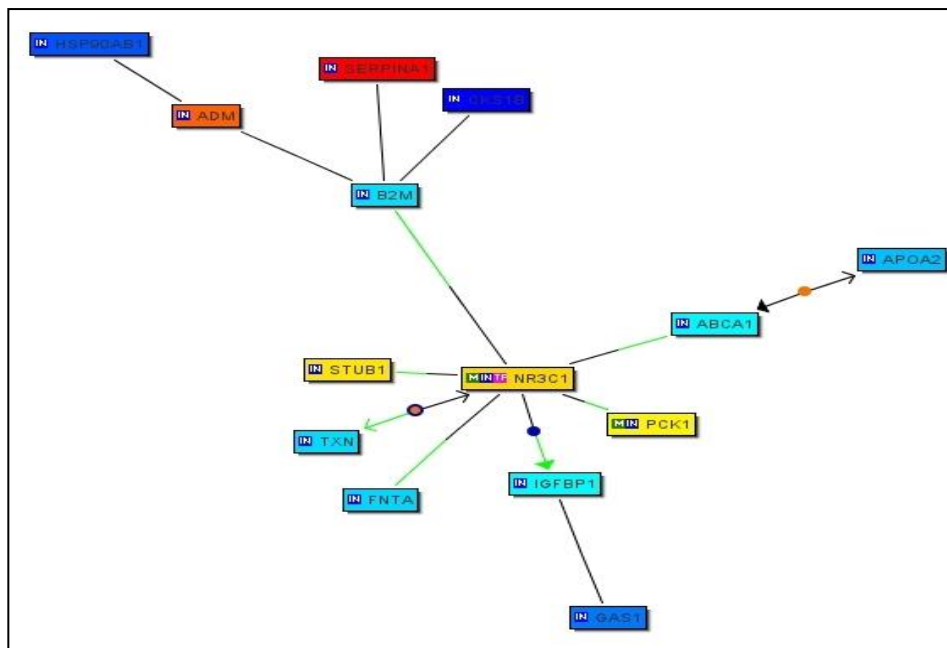
Figure 5.4 c and d show gene networks of NR3C1 or GR from IPA. Even though the input probe sets are the same as those originally used in Phillip et al. (2005), we found different output gene names due to different species (human and mouse). Three genes from the original study were changed to different gene names: ETR101 to IER2, HSPCB to HSP90AB1, and TDE1 to SERINC3.

Comparing results from BSPE and IPA, IPA found connections for all 23 genes but BSPE only found 14 connections. IPA added other relevant genes (12 genes) from databases to the network. IPA confirmed six connections found in BSPE: NR3C1 connects with ABCA1, B2M, FNTA, IGFBP1, STUB1, and TXN. In addition, the connections between and ABCA1 and APOA2 from BSPE were also found in IPA. However, connections between B2M with ADM, CKS1B, SERPINA1 and ADM with HSP90AB1 found in BSPE were not found in IPA. In fact, these genes (ADM, CKS1B, SERPINA1, and HSP90AB1) and APOA2 were found to have direct connections with NR3C1 in IPA.

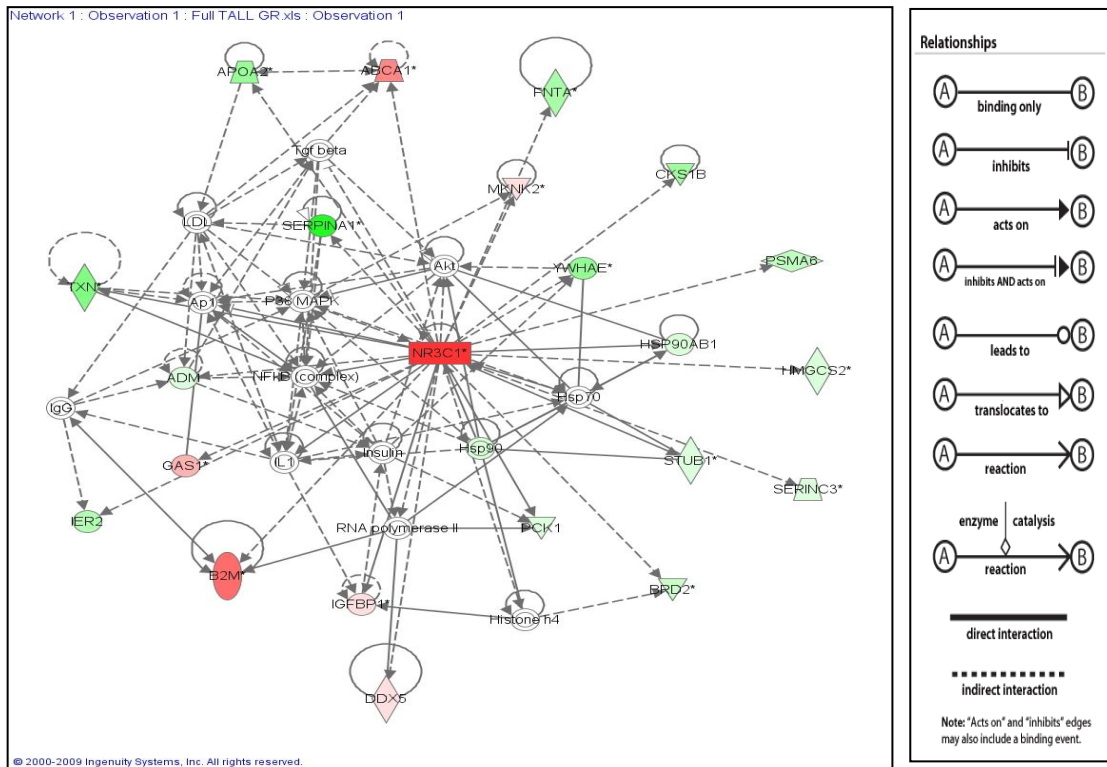
a)



b)



c)



d)

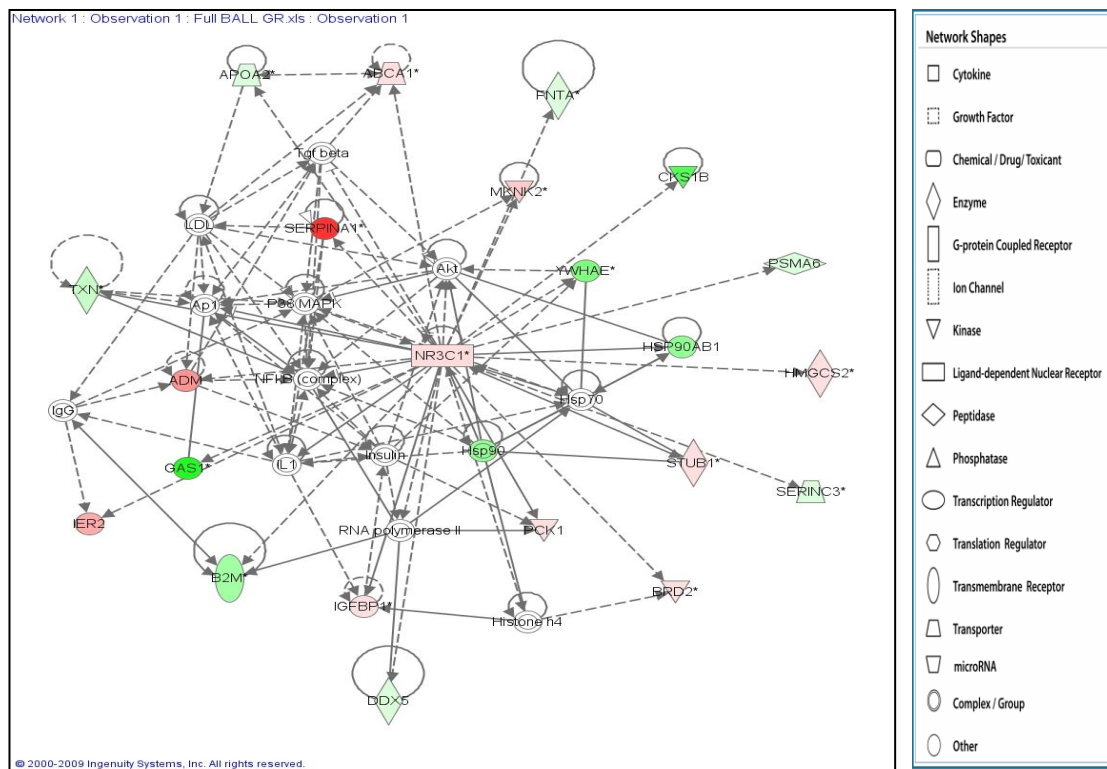


Figure 5.4: GR gene network from; a) BSPE for T-ALL; b) BSPE for B-ALL; c) IPA for T-ALL; and d) IPA for B-ALL

We further investigated an effect of drugs and tissues on gene networks by identifying common genes. The GR network used in this study was inferred from blood sample data from patients treated with prednisolone and collected at, before treatment, and at 6/8 and 24 hours after treatment; whereas, Phillip et al.'s (2005) samples were mice liver treated with dexamethasone and measured three hours before being sacrificed. In addition, Miller et al. (2007) used paediatric patient cell lines treated with dexamethasone and sampled at 20-24 hours, while Donn et al. (2007) collected data from healthy adults treated with dexamethasone.

- **Same tissues and same drug (human and dexamethasone)**

Next, we directly compared the previously mentioned two existing glucocorticoid networks from (i) Donn et al. (2007) and (ii) Miller et al. (2007). Four genes were found common between (i) and (ii): IL7R, FKBP5, PIK3R1 and TXNIP. NR3C1 and HDAC1 were added to network (i) by IPA but found in network (ii).

- **Different tissues and same drug (human vs mouse and dexamethasone)**

Finally, we compared network (i) Donn et al. (2007) and (ii) Miller et al. (2007) with network (iii) Phillip et al. (2005). There were two common genes between networks (ii) and (iii): NR3C1 was from both original sets while MYC was added to the network (ii) by IPA software and found in network (iii). ADM was the only common gene between network (i) and (iii).

- **Different tissues and drugs (human vs mouse and prednisolone vs dexamethasone)**

First, we compared the gene list from STEM (Table 4.4 and 4.5) and network (iii) Phillip et al. (2005). For B-ALL, there were six common genes (MYC, STAT5A, HLA-G, CASP3, NFKB1B, and HNF4A) and for T-ALL only three common genes were found (HNF4A, RELA, and TNF).

Some existing literature confirmed a relationship between common genes and the apoptosis process, specifically, in leukaemia, as follows: CASP3 (T. Liu et al., 2002), FKBP5 (Kajiyama et al., 2009; Schmidt et al., 2006), HDAC (Rosato, Almenara, Dai, & Grant, 2003; Tsapis et al., 2007), HLA-G (Gros et al., 2006), IL7R (Karawajew et al., 2000), MYC (Ceballos et al., 2005), NFKBIB (Zhuang et al., 2004), PIK3R1 (Kharas et al., 2008), RELA (Dai, Rahmani, Dent, & Grant, 2005), STAT5A (Nosaka et al., 1999), TNF (Wen et al., 2000; Wuchter et al., 2001), and TXNIP (Z. Wang et al., 2005). No literature was found for HNF4A from a PubMed search using the key words “HNF4A apoptosis leukaemia” (<http://www.ncbi.nlm.nih.gov/sites/pubmed>).

It is possible to conclude from our results that the glucocorticoid response depends on the type of tissue, chemotherapeutic drugs and elapsed time after treatment. Another factor that can affect the final gene network result is the software selected. In our study, both software (BSPE and IPA) were based on curated connections from existing literature, however, we found differences in the inferred networks from the two methods.

Since NR3C1 was likely to be the most common gene, this gene was further investigated using Oncomine - a cancer microarray database. This tool gives a visualisation of gene expression from many different studies at the same time. However, analysis of the network using Oncomine can only be done using the available datasets on the database because a user cannot upload their dataset. NR3C1 has been found differentially expressed in many types of cancer; for example, ovarian, prostate, breast and lymphoma. In leukaemia, NR3C1 has been found up- or down-regulated. Existing human leukaemia cancer datasets on Oncomine were analysed in relation to NR3C1 and the level of gene expression between normal cells, T-ALL and B-ALL is shown in Figure 5.5. There were 15 independent experiments/studies. Details of each experiment/study, including p-values, are shown in the table under Figure 5.5. Colours indicate different classes: class one (blue), class two (red), class three (green) and class four (yellow). Analysis no.1 was between normal bone marrow (left-hand side) and B-ALL (right-hand side) while analysis no.7 showed expression between normal bone marrow (left-hand side) and T-ALL (right-hand side). Analysis no. 10 shows the differences expression of NR3C1 between males and females and analysis no. 15 shows the resistance and sensitivity to GCs.



Schmidt et al. (2006) (current study)

Figure 5.5: Box-plot distribution of NR3C1 across 15 independent experiments/studies created by OncoPrint

Analysis No.	Study	P-value
1	Anderson leukaemia class 1: Normal bone marrow (6) class 2: B-ALL (6)	$5.7e^{-11}$
2	Maser leukaemia T-ALL (18)	$1e^{-7}$
3	Anderson leukaemia class 1: B-ALL (87) class 2: T-ALL (11)	$1.1e^{-5}$
4	Holleman leukaemia class 1: B-ALL (146) class 2: T-ALL (27)	$2.2e^{-5}$
5	Raetz leukaemia class 1: T-ALL (9) class 2: B-ALL (10)	$7.3e^{-5}$
6	Bhojwani leukaemia class 1: Pre B-ALL (103) class 2: T-ALL (10)	$3.1e^{-4}$
7	Anderson leukaemia class 1: Normal (6) class 2: T-ALL (11)	0.002
8	Schmidt leukaemia class 1: B-ALL (30) class 2: T-ALL (9)	0.005
9	Raetz leukaemia class 1: B-ALL (10) class 2: T-ALL (10)	0.022
10	Heuser leukaemia class 1: Female (18) class 2: Male (17)	0.035
11	Choi leukaemia class 1: healthy (6) class 2: Adult T-ALL (41)	0.067
12	Cario leukaemia class 1: Common (32) class 2: Pre B-ALL (19)	0.073
13	Schmidt leukaemia ALL class 1: prior to treatment (14) class 2: after 8 and 24 hours	0.188
14	Schmidt leukaemia ALL class 1: prior to treatment (13) class 2: after 6 hours (9) class 3: after 8 hours (4) class 4: after 24 hours (13)	0.245
15	Holleman leukaemia class 1: Resistant (28) class 2: sensitive (94)	4.68

For B-ALL, there was strong evidence that NR3C1 is up-regulated and in T-ALL it appeared to be mainly down-regulated. In our study, we did not find NR3C1 differentially expressed in B-ALL (only 2/10 patients show that NR3C1 was differentially expressed: up-regulated). For T-ALL, NR3C1 was found differentially expressed (up-regulated) in 2/3 patients. We compared this with analysis no.14 which showed the expression from NR3C1 of Schmidt et al.'s (2006) study whose data were used in our study. (Schmidt et al. (2006) combined B- and T-ALL subtypes into one group but we analysed subtypes separately). The level of expression of NR3C1 when compared before treatment (0 hour- blue box plot) indicated up-regulation after treatment (6 or 8 hours and 24 hours- red, green and yellow box plots). The results from Oncomine indicated that NR3C1 was differentially expressed with a p-value of only 0.245; this was confirmed by our study; altogether only four out of thirteen patients' expression value passed the criteria. Thus, the average NR3C1 gene expression value was definitely low and statistically insignificant.

The next section is the final section in this chapter and focuses on combining all possible GR gene networks from the previous section and proposed the most intensive GR gene network.

5.3.3 Proposed GR gene network

We come to the last part, the most important finding, and the core of this study which is the possible GR gene network. In addition, we also combine previously known GR gene networks and GR gene network curated genes from the most relevant pathways. This gene network is a comprehensive network and it may provide a good base for scientists to select target genes for future research.

Information from previous studies were combined and presented as a GR gene network that extended the existing network proposed by Phillip et al. (2005). These studies are: Miller et al. (2007), Donn et al. (2007), Phillip et al. (2005), and our selected gene networks of GC-induced apoptosis genes based on NR3C1 gene, as shown in Figures 4.4, 4.5, 5.2 and 5.4 (Figures 4.4 and 4.5 contain our GR gene networks based on gene clusters from STEM for B-ALL and T-ALL, respectively; Figure 5.2 contains the GR gene networks based on apoptosis, p53, NFκB genes; and Figure 5.4 shows the GR gene network we curated from IPA, based on

genes from the Phillip et al. (2005) study). Gene lists from these six gene network were used to develop the gene network on the CellDesignerTM program (Funahashi et al., 2008) which can be freely downloaded from <http://www.celldesigner.org/>. The result is shown in Figure 5.6. Most figures generated by IPA program have nodes presented in several shapes, for example, cytokine (□), growth factor (⊖), enzyme (◇), and kinase (▽). These network shapes not exist in CellDesignerTM. Therefore, we replaced most of the components (nodes) from IPA as a gene node (□) in CellDesignerTM, only transmembrane receptor (⊖) and complex or group (⊙) nodes retained their representation as receptor and complex node, respectively.

Common genes were found in more than two networks, for example, **ABAC1, ADM, B2M, DDX5, FKBP5, HDAC, MYC, PIK3R1, SERPINA1, TNF, TXN and YWHAE**. This possible GC network in childhood leukaemia was curated from literature-based knowledge from many experiments on different tissues/organism and conditions. A future research topic is to verify the gene connections; specifically, for childhood leukaemia.

All the methods in our study used hand/manual literature curated networks from existing databases as a part of network construction based on co-occurring terms. Therefore, the proposed network may be used with caution as the networks created from databases may have some incorrectly incorporated gene pairs (no relationship or incorrect relationship). Jessen et al. (2001) pointed out possible errors from gene symbols and short names; for example, different symbols have been used for different species and cell lines and short names refer to something else other than gene names (Jessen, Lægroid, Komorowski, & Hovig, 2001)

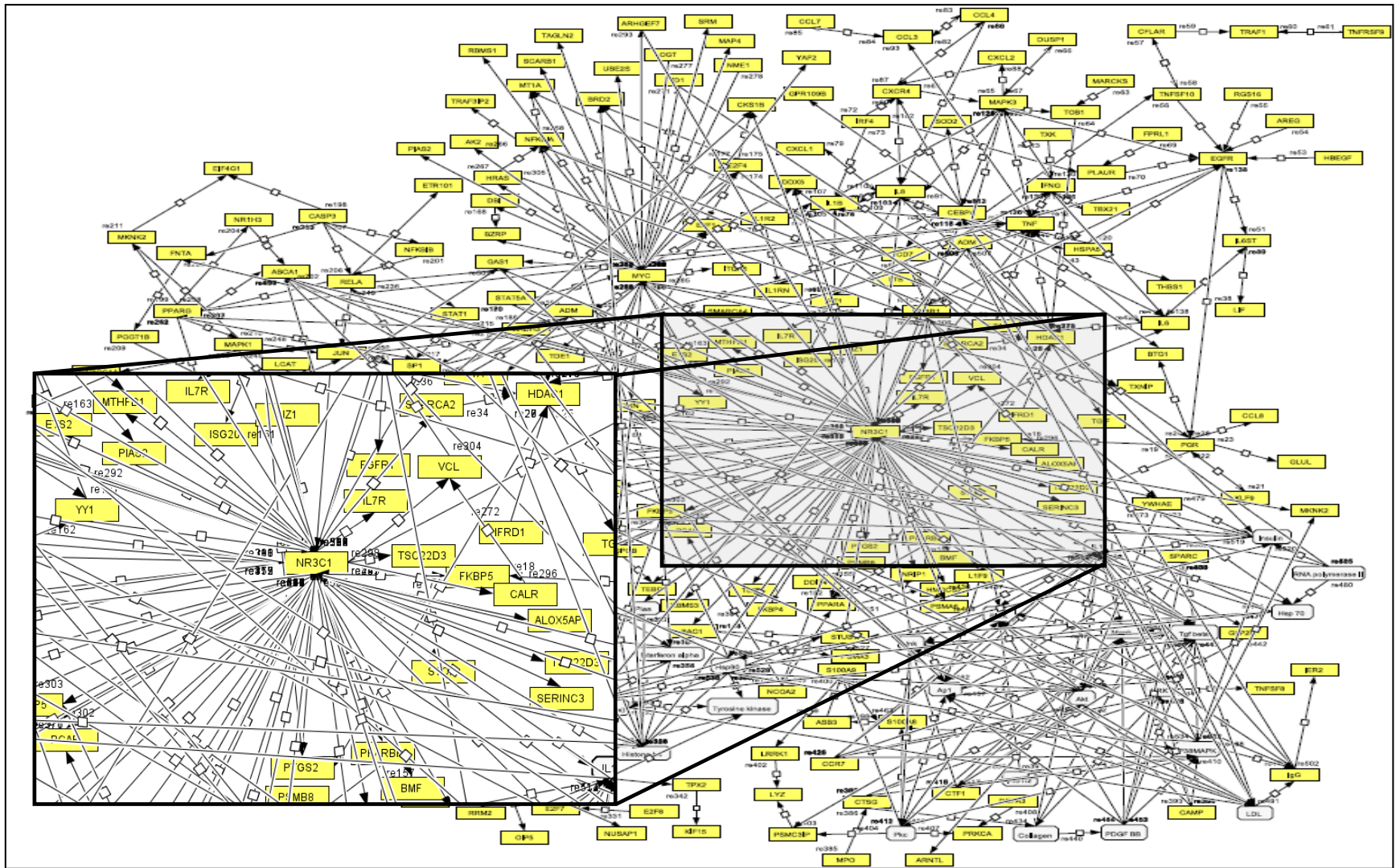


Figure 5.6: Proposed GC-induced apoptosis based on NR3C1 gene network

5.4 Conclusions

After identifying the set GC-regulated genes in Chapter 3, in this chapter, we focused on finding GR gene networks in childhood leukaemia. We constructed the gene network based on three aspects: GR gene networks from GC-induced apoptosis genes after using the STEM clustering method (illustrated in Chapter 4), GR gene networks from relevant pathways (apoptosis, p53 and NFκB) and GR gene networks using genes from previous study (Phillip et al. (2005)).

We investigated how known genes in known pathways involved in the apoptosis process behaved in our selected dataset. All the selected genes had expression levels under the threshold (\pm two fold) and relatively small changes in expression. This led to our conclusion that these known genes may not yet be active 24 hours after treatment or small subtle changes in gene expression should be considered in the process of selecting differentially expressed genes. When we further investigated gene connections from the three pathways given by the IPA program, we found that even the selected gene (NFκB), known in the literature to be related to GC-induced apoptosis process, did not show up in one of the three networks from three pathways. Tissing et al. (2003) showed a direct connection between GR (NR3C1) and NFκB in leukaemia. These two genes are also vital genes in the three selected pathways. However, no direct connection was found between these two in any of the three gene networks based on the 20 genes from the three pathways. When the three networks were combined, however, connection between GR and NFκB was reinstated by IPA, possibly indicating that the connection come from the combined activity of several functions represented by these networks. Another possibility is that IPA database is from a variety of cells/tissues/diseases/processes, therefore, may not highlight connections specific to a particular disease situation. Therefore, inferred networks need further study in relation to a specific biology process. However, the inferred gene networks from IPA still a good base to add to the final gene network because their connections have been rigorously validated from the literature.

Network no. 3 (Figure 5.2) was used to add to the final gene network with the connection between NR3C1 with other 34 genes including: 14-3-3, Akt, ALP, Ap1, Arf, Calmodulin, Calpain, Cbp/p300, Ck2, ERK1/2, Fgf, FOS, Hsp70, IL1, Insulin, Interferon alpha, Jnk, Mapk, MDM2, MYC, P38 MAPK, PI3K, Pkc(s), PLA2, Ras, RNA polymerase II, Sapk, Shc, SNCA, STAT, Tgf beta, Thyroid hormone receptor, TP53, Vegf.

The GR gene network was elucidated from IPA based on the gene list of Phillip et al. (2005) (Figure 5.3c) but this study used a different chemotherapeutic agent (dexamethasone) and mouse liver tissues. The comparison of networks between the study of Phillip et al. (2005) and our study showed only a few overlapping genes. Few overlapping genes were also found when our network was compared with other two recently proposed GR gene networks. This leads to the conclusion that GR networks may depend on chemotherapeutic agents and tissue type. In this study, we utilised the strengths of three existing network/pathway tools to improve the understanding of GC-induced apoptosis through GC-regulated genes and GR gene networks. IPA is considered to be the best tool to visualise and create pathways and view networks while BSPE can add more details on transcription factor levels. Oncomine was used because it is a cancer database, so it systematically compares and links with other cancer databases. The only limitation is that it does not allow users to upload their own datasets. As there are many available tools for scientists to choose from, using different software to create gene networks can cause variation in the proposed gene networks. The key question for further research is how to verify gene networks from differing sources.

Finally, we proposed a GC-induced apoptosis network with the main focus on GR or NR3C1 genes. We combined three existing GR gene networks from previous studies (Donn et al. (2007), Miller et al. (2007) and Phillip et al. (2005)) with our three inferred gene networks from the selected genes (Figures 4.4, 4.5, 5.2 and 5.4) using CellDesigner™ program. This network is a starting point for scientists to conduct further investigations.

Chapter 6

Summary, Conclusions and Future Directions

In this thesis, glucocorticoid-induced apoptosis genes and glucocorticoid receptor (GR) gene networks were identified in order to increase an understanding of the underlying biological mechanisms of GC-induced apoptosis in childhood leukaemia. The dataset used in this study was retrieved from the original study by Schmidt et al. (2006). Differentially expressed genes were identified according to the following criteria: log-ratio and number of patients. Thereafter, these genes were used to find gene clusters using four emergent clustering methods: Self organising map (SOM), Emergent self organising maps (ESOM), Short time series expression miner (STEM) and Fuzzy clustering by local approximations of memberships (FLAME). Finally, these genes were used for gene network construction through selected network/pathway knowledge-based databases.

We extended the investigation further from the original Schmidt et al. (2006) study. Specifically, new criteria were proposed to select novel genes in B-ALL and T-ALL subtypes separately and more genes were found than in the original research that combined the two subgroups in the analysis. The relationship between GC-induced apoptosis and GR gene network was illustrated. From short time series data, we defined novel genes and elucidated the gene network which we believe expands the current knowledge about GC-induced apoptosis in childhood leukaemia. A previously proposed GR gene network from liver tissue was updated in this study for clinical tissues. Finally, we proposed GC-induced apoptosis based on the NR3C1 gene network which lead to an understanding of the underlying mechanism and will lead to better clinical treatment. The following section presents a summary of what we discovered and contributions of this research, as well as future research that may provide more insights into the GC-induced apoptosis mechanism.

6.1 Summary

We started with an extensive literature review, which was published and presented in Chapter 1. This review focused on the existing research on the use of machine learning approaches to investigate childhood leukaemia. Even though childhood therapy achieves a highly successful survival rate, there are still many children who face severe side effects and failure from treatments. Therefore, chemotherapeutic treatment for childhood acute lymphoblastic leukaemia was the main focus in this study. The first aim of this study was to identify GC-induced apoptosis genes in childhood leukaemia, starting from a set of genes proposed in a previous study. The previous study showed little agreement with findings from other studies (a few or no common genes) (Schmidt et al., 2006). Different gene sets have been reported from previous studies that used different drugs and tissues, as mentioned in Chapter 3. Experiments based on a clinical setting are extremely rare, difficult and expensive to carry out, so most studies used *in vitro* experiments and cell lines instead. We used a dataset from Schmidt et al. (2006); the most prominent dataset because it used gene expression data collected from childhood leukaemia patients at two time points after treatment. We analysed this dataset with specific objectives and the findings were reported as follows.

The first specific objective of this study is to identify GC-regulated candidate genes.

Objectives:

- (1) How reproducible or robust are the original authors' results?
- (2) Do different platforms (software) available to normalise data have an effect on the final gene sets?
- (3) Do leukaemia subtypes: T and B-ALL produce similar differentially expressed gene sets?

Findings:

We found that the selected dataset is reproducible and robust. In addition to the R software used by Schmidt et al. (2006), we applied Matlab and RMEExpress to normalise raw data. Different software platforms produced slightly different sets of differentially expressed genes. Among the differentially expressed genes, there are some genes that are common between subtypes and some genes are found only in each subtype?

The analysis process started using robust-multi array average (RMA) to normalise the raw dataset (from B- and T-ALL patients). Next, we identified differentially expressed genes using the same criteria ((i) log ratio of ± 0.7 or higher (ii) log ratio of ± 1.0 or higher, for at least six out of thirteen patients) as in the original research; we found more differentially expressed genes than the ones reported in the original research. Then, we investigated the proposed novel gene set (22 genes) in the original research. Most genes were found for B-ALL only (11/22 genes), while only three were found in T-ALL, and the remaining eight genes were found in both subtypes. The original authors (Schmidt et al. (2006)) assumed the commonality of the two subtypes (T-ALL and B-ALL), whereas there is a discrepancy between them. Therefore, we proposed new criteria for the two subtypes separately (log ratio of ± 1.0 for five out of ten B-ALL patients and two out of three T-ALL patients and new gene sets was reported. For T-ALL, there were 237 probe sets (203 unique probe sets after removing repeats) and for B-ALL there were a total of 257 probe sets (207 unique probe sets) for three time intervals. We combined these two gene sets from T- and B-ALL. This set contained 380 unique probe sets (only 24 probe sets were common to T-ALL and B-ALL, of which three probes were not found in the original paper). In the next step, we deleted cell cycle genes from this list by using known cell cycle genes from KEGG, Cell cycle database, and the original article. After deleting cell cycle genes, T-ALL contained 222 probe sets (172 unique probe sets) and B-ALL contained 190 probe sets (155 unique probe sets) for the three time intervals. After combining the gene lists from the two subtypes, the final set had 327 unique probe sets (304 genes) responsive to GCs (19 probe sets were common to T-ALL and B-ALL). Thereafter, we compared the new gene list with GC responsive gene reported in other previous research, even so, only few overlapping genes were found.

The second specific objective is to identify group of GCs-induced apoptosis genes that may have similar functions.

(1) Do gene clusters differ when analysed by general clustering methods as opposed to using clustering methods specifically designed for short time series data?

The second aim was achieved by processing the new gene set through four different emergent clustering methods to identify genes with similar expression. SOM and ESOM are good at visualising how genes are organised in terms of distance and density, respectively. Both clustering methods provided a clear overview of the expression patterns of genes. However, SOM clusters may not be consistent due to the instability of neural gas, the method used in

this study to cluster neurons on the trained map as it produces unclear cluster boundaries. For ESOM, similar issues are known found true in this study. In addition, its number of clusters needs to be pre-defined by the user. Therefore, clusters from both SOM and ESOM methods were not used in further comparisons.

STEM is a clustering approach that combines temporal characteristics with statistical significance analysis into cluster analysis to reduce possible networks that can happen by chance. Gene clusters from STEM were selected and used for gene network analysis in the second part of this research. The FLAME clustering method can be used to find clusters without predefined groups as it considers all possible clusters. The result from FLAME was useful to verify clusters from STEM. Clusters had to be compared manually for each patient. STEM and FLAME identified similar gene clusters. As more and more data are collected from many patients, the challenge is to develop data analysis tools which can analyse multiple samples (patients) at a time with respect to genes and their temporal patterns. We also showed the effect of distance measure on the final gene cluster. Therefore, prior knowledge needs to be incorporated with the selection of distance measure that is appropriate for a specific dataset.

Next, we compare gene clusters from both methods with known gene functional clusters from the Database for Annotation, Visualisation and Integrated Discovery (DAVID). The results lead to the finding that only some genes from the same cluster have similar gene function (as they were found in the same gene functional cluster in DAVID). But there was more than one function for each cluster. How to assign genes to their right function can be a future research topic.

The third specific objective is to construct gene networks of proposed GC-induced apoptosis genes.

Findings

In regard to objective one, we inferred gene networks of GC-induced apoptosis gene sets using IPA software to find the common (dominant) genes between the three time intervals. The connection between these gene networks with GR was displayed. We proposed GC-induced apoptosis gene networks for T- and B-ALL separately. It was found from IPA that

the main function of the input genes in T-ALL was involved with cell death while the function of B-ALL was involved with cell cycle processes. We found a different time-lag (time-delay) response to treatment between these two subtypes.

The fourth specific objective is to elucidate glucocorticoid receptor (GR) gene network.

The objective and specific findings were defined as follows:

- (1) To investigate the behaviour of selected known genes from the two main apoptosis pathways (extrinsic and intrinsic).
- (2) To illustrate GR gene networks based on selected genes from previously proposed glucocorticoid receptor gene networks by Phillip et al. (2005).
- (3) To elucidate GC-induced apoptosis network with emphasis on GR or NR3C1 gene.

Findings

To answer the first sub-objective, we indicated that the gene expression level of selected known genes from the two main apoptosis pathways (extrinsic and intrinsic). Thereafter, we constructed an interplay network between three selected pathways (apoptosis, p53 and NFκB). We selected a limited number of genes which have previously been defined in relevant literature as known genes in apoptosis, p53 and NFκB pathways. The inferred network showed possible gene connections that were created by IPA, which is curated from journal articles, on different cell/tissue/process/disease studies may be not highlight the specific process. For example, the relationship between NR3C1 and NFκB identified by IPA was not included in the Tissing et al. (2003) study which also mentioned the relationship between these two genes in the GC-induced apoptosis process. However, the inferred gene networks from IPA and their gene connection have been rigorously validated from the literature. Therefore, the networks generated by IPA may be useful for constructing a possible GC gene network with focus on GR.

For the second sub-objective, we extracted the same genes that were used in the Phillip et al. (2005) study from Schmidt et al.'s (2006) dataset. This led to the finding that different chemotherapeutic agents (prednisolone and dexamethasone), tissues (blood and liver), species (human and mouse) and time (6/8 hours and three hours after treatment) may have an effect

on the final candidate genes. In the final objective, we manually combined gene networks: three existing networks (Donn et al. (2007), Miller et al. (2007) and Phillip et al. (2005)) and our inferred network from STEM, three relevant pathways and Phillip et al. (2005) gene set to retrieve the possible GC-induced apoptosis network.

We used three databases: Ingenuity Pathway Analysis software (IPA), the BiblioSphere Pathway Edition (BSPE) and the Oncomine. The main pathway analysis tool used with all input gene sets was IPA, whereas, BSPE was only used with the gene set from previous study (Phillip et al. (2005)).

For the final sub-objective, we manually curated the GC-induced apoptosis network with emphasis on GR or NR3C1 gene from selected studies and our own investigations. This proposed GR gene network is far from being complete but it may be a starting point for further investigation and may be added to future networks modelling.

We concluded, after gene expression data analysis, the following:

- (i) ALL subtypes share some common molecular profiling: however, they have distinct patterns; although this was based on a small sample size, this issue should be taken into account.
- (ii) While different sample types: cell lines, clinical samples and mice treated with different chemotherapeutic drugs may share common response, there still have unique patterns and the final genes discovered can vary.

6.2 Conclusion

Glucocorticoids are used intensively in the treatment of childhood acute lymphoblastic leukaemia. GCs induce apoptosis in immature lymphoid cells. However, molecular mechanism of GC-induced apoptosis has not been clearly defined. This study focused on extending the understanding of the underlying mechanism of GC-induced apoptosis process in childhood leukaemia. This goal was accomplished by using short time series clustering methods and web-based knowledge network/pathway tools. In this study, glucocorticoid treated childhood leukaemia short time series gene expression profiles were used (i) to identify GC-induced apoptosis genes and (ii) to infer GR gene networks. Even though there were many issues to extend this study in future research, we hope the gene lists and gene networks identified in this study will add new knowledge to the field and will lead to further experiments and clinical trials in order to increase the survival rate and reduce side effects in treatment of childhood leukaemia.

6.3 Contributions

In this thesis, we addressed the problem of understanding the glucocorticoid-induced apoptosis mechanism in childhood leukaemia. The following summarises the contributions of the thesis.

- The original study from which we retrieved the dataset identified only novel genes involved with GC-induced apoptosis process. The expensive costs and difficult processes involved with data collection still exist in this research field. The selected dataset is invaluable as an extremely rare time series dataset from clinical samples. My contribution is extending the data analysis of this dataset to further analyse gene networks which may shed light on understanding the GC-induced apoptosis mechanism in childhood leukaemia.

- We have used short time series clustering to discover relationships among selected differentially expressed genes in order to find similar gene functions based on emergent clustering methods. The results have helped with understanding how four emergent clustering methods work with short time series data from childhood ALL and produced novel and consistent GC-regulated gene sets that will be extremely useful in clinical settings, in understanding GC gene networks and in identifying new drug targets.
- The work focused on using identified gene sets to elucidate networks based on prior knowledge by incorporating some manually curated networks/pathways from the relevant literature. This has produced the gene network of GC-induced apoptosis in relation to the most important GC-regulated genes.
- GC-induced apoptosis genes which have been reported in this study are from various tissues treated with different drugs, resulting in only small numbers of overlapping genes. My contribution to this work is in the investigation and confirmation of the effect of subtypes, samples sources, and chemotherapeutic drugs to the final novel genes or selected differentially expressed genes.

6.4 Future Research

Biomedical informatics research is an emerging field, still in its infancy, with much on-going research. There are various directions in which to consider further research. We indicate the specific points that are of prime importance.

6.4.1 Computational methods

- **Integration of ‘omics’ data.**

This research field requires systematic analysis in order to uncover the whole picture of the GC-induced apoptosis mechanism. Gene expression profiles may not provide a complete whole insight into GC-induced apoptosis mechanisms because the changes in gene expression may not always refer to a simultaneous change in protein expression (Carroll et al., 2005). Therefore, integrating other high throughput proteomic, splicing, or other newer techniques with gene expression profiles may provide true insight into GC-induced apoptosis process. In

addition, gene networks are constructed based on different sources; in this research glucocorticoid receptor gene networks were constructed based on microarray data. In future, a combination of various data sources including transcription factor binding information, protein-protein interaction, ChIP (Chromatin Immunoprecipitation) on chip and other high throughput technology data will increase the reliability, provide more understanding and insight, and unravel the sophisticated and complex nature of gene networks.

- **Selection of normalisation process**

The normalisation process has an effect on inferring gene networks, and more details can be found in Lim et al. (2007). In addition, normalisation affects array-to-array precision and accuracy (Stafford & Tak, 2008). As there is no definitive conclusion on the best normalisation process, using different normalisation methods to analyse the dataset, as used in our study, then comparing results might be another way to verify differentially expressed genes. In addition, it is interesting to study further what cause the different results when using different software but same normalisation method.

- **Criteria for selection of differentially expressed genes**

Differentially expressed genes are selected from a comparison between a target sample and control sample (e.g. no treatment/after treatment) as log ratio (base two) of target to control gene expression, which is called a fold change. The selection of statistical methods affects the selection of novel gene lists. Differentially expressed genes in this study were selected based on the fold change method which considered genes above a fold change cut-off as significant genes. Fold change has been criticised for its propensity to variation or unreliability. This method does not take into account the variability of inter-experiment noise and outliers. Genes with large fold ratios may probably come from high variability; consequently, genes with more than two fold changes may not always be significant genes. Several statistical methods have been developed to improve the outcome and reduce system variability. Further analysis should apply these statistical methods and compare the results in order to verify the novel genes. The decision on which method to use should be based on the nature of the target biological system; for example, the modified t-test is suitable for gene expression that changes according to the underlying noise, and fold-change is suitable for gene expressions that have large absolute changes (Tibshirani & Witten, 2007). A comparison of ten gene selection methods with several cancer data can be found in Jeffery et al., (2006). Apart from the existing methods, there are new methods under development or recently published in relevant

bioinformatics journals for public use; for example, FCPC methods, based on gene-to-gene correlation and principal component analyses (Qin, Feng, Harding, Tsai, & Zhang, 2008).

- **Selection of short time series clustering method**

Although gene clustering is the most common, important and essential method in microarray data analysis, the issue still remains as to which of the available methods should be selected and the choice of the corresponding parameters to generate clusters for selected data in order to reveal data structure and characteristics. For analysis of short time series data, STEM is one of a possible number of clustering methods. This field is still at an early stage; suitable approaches are being developed in order to capture the nature of short time series data, such as TA-clustering (Temporal Abstractions) (Sacchi et al., 2005). A comparison of three existing methods for analysing time series gene expression data can be found in Di Camillo, Toffolo, Nair, Greenlund, and Cobelli (2007). There is still no one method for all data; therefore, further study should consider cluster validity in order to evaluate/validate clusters. Currently, there are a number of cluster validation techniques available (Bolshakova & Azuaje, 2003; Bolshakova, Zamolotskikh, & Cunningham, 2006).

- **Selection of network/Pathway analysis tools**

Networks/Pathways analysis is still in its infancy; new tools are under development or are currently being released. Most new tools are free and user-friendly with web access; furthermore, these tools increase the capability and stability in handling the large datasets from existing databases, generating heterogeneous and complex cellular networks, and analysing noisy and incomplete gene expression data. Examples of these new tools are: Gene Network Generator (GeNGe) (Hache, Wierling, Lehrach, & Herwig, 2009), GraphWeb (Reimand, Tooming, Peterson, Adler, & Vilo, 2008), Network Analysis Tool (NeAT) (Brohee et al., 2008) and VisANT (Hu et al., 2009; Hu, Snitkin, & DeLisi, 2008). In addition, there are software/tools designed specifically for time series data including JCell- a java-based tool to reconstruct gene networks from time series gene expression data (Spieth, Supper, Streichert, Speer, & Zell, 2006). A more comprehensive list of network visualisation and analysis tools can be found in S. Zhang, Jin, Zhang, and Chen (2007). In addition, reverse engineering using network/pathway tools or other methods (such as neural networks and genetic algorithm) along with mathematical modelling will help to complete the whole picture of selected study process. Each method has different approaches and these differences will provide different perspectives that complement each other which may lead to better understanding of complex processes.

6.4.2 Biology of Childhood Leukaemia

- **Subtypes of T-and B-ALL**

Childhood leukaemia can be divided into at least six prognostic subtypes under T- and B-cell precursor: T-ALL, TEL-AML1, E2A-PBX1, BCR-ABL, MLL gene arrangement and hyperdiploidy > 50 chromosomes (Pui, 2004; Ross et al., 2003; Yeoh et al., 2002). Furthermore, they can be divided into good-, standard-, high- or very high-risks groups (Moos et al., 2002). These issues should be taken into account in order to understand the underlying GC-induced apoptosis in each subtype and risk groups.

- **Different chemotherapeutic agents**

Gene expression profiles from patients (clinical samples) were shown to be different from their cell line samples in the case of chronic myeloid leukaemia and acute myeloid leukaemia (Leupin et al., 2006). In GC-induced apoptosis gene studies on acute lymphoblastic leukaemia, there has been an extrapolation from leukemic cell lines to clinical samples with various chemotherapeutics. As a consequence, two main points are of concern: (i) there are many synthetic glucocorticoids as well as other antileukaemic drugs that have been used; for example, L-asparaginase (L-asp), daunorubicin, and some that have been tested in clinical trials, such as clofarabine, nelarabine and forodesine (Pui, Robison, & Look, 2008) (ii) there is still no conclusion on how chemotherapy drugs may vary in effect on the treatment and the GC-induced apoptosis process. Cheok et al. (2003) suggested that the changes in gene expression are treatment-specific, in addition, different leukaemia subtypes share common responses to antileukaemic agents. However, this conclusion is for specific antileukemic agents (mercaptopurine and methotrexate). The degree of sharing is still in question; from previous studies, there is only a small number of overlapping genes between cell lines and clinical samples. Moreover, samples from clinical settings with different drug treatments give distinct gene sets.

- **Relevant Pathways**

Unravelling the underlying mechanism of the glucocorticoid-induced apoptosis signalling pathway for specific cell types is still in its infancy. GCs are involved with several signalling pathways, which are not all included in this study. For example, the cAMP/protein kinase A (PKA) and the mitogen-activated protein kinase MAPKs pathways, which include extra-cellular signal-regulated kinase (ERK), c-Jun N-terminal kinase (JNK) and p58 (Miller, Garza, Johnson, & Thompson, 2007). Twelve molecules and pathways associated with GCs

have been reported by Herr et al. (2007) including mitochondria, death receptor signalling, Bcl-2-family, Caspases, c-myc, I κ B, Granzyme A, TDAG8, Lysosomes, Proteasomal degradation, Stress pathway and other modulators, such as IL-6 and T-cell receptor. New discoveries in biology for each individual pathway may be combined to increase the completeness of pathways/networks; for example, the ubiquitin-proteasome pathway, which is the main pathway involving degradation of intracellular protein in eukaryotes and controls the regulation of apoptosis process. Manual collection is a painful task to cover all information in the literature; therefore existing databases will help to gather most of the information. However, the final networks/pathways still need to be verified and validated. The combination of forward modelling using mathematical techniques with reverse engineering schemes would fulfil this validation task.

Appendix A

A.1 Comparing genes from original author with our data analysis

The original author proposed 128 probe sets (104 genes) as follows:

Subset	# of probe sets	# of genes
Table S2A Induced (early response: 6/8 hours)	25	19
Table S2B Repressed (early response: 6/8 hours)	37	30
Total	62	49
Table S2C Induced (late response: 24 hours)	28	24
Table S2D Repressed (late response: 24 hours)	38	31
Total	66	55

We reanalysed the dataset from original authors using the same method and criteria. Then we compared number of patients that passed the criteria in the original authors' data analysis and our data analysis. In Tables A.1.1-1.4, we highlighted in grey colour when the number of patients was different and in bold letters when they were quite different.

A.1.1 Comparison of the number of patients who passed selection criteria in Table S2A from original article (left) and reproduced (right) in our study

ID	Gene symbol	Induced 6 h	Induced 24 h	ID	Induced 6 h	Induced 24 h
208078_s_at	SNF1LK	9	10	208078_s_at	9	10
226733_at	PFKFB2	10	9	226733_at	10	9
204560_at	FKBP5	8	9	204560_at	8	10
224856_at	FKBP5	7	9	224856_at	7	10
228854_at	ZBTB16	8	9	228854_at	7	9
224840_at	FKBP5	7	8	224840_at	7	8
203761_at	SLA	7	11	203761_at	7	11
202887_s_at	DDIT4	8	11	202887_s_at	8	11
203760_s_at	SLA	8	10	203760_s_at	8	10
210001_s_at	SOCS1	6	9	210001_s_at	6	9
221756_at	MGC17330	6	9	221756_at	6	9
228434_at	BTNL	9	8	228434_at	9	8
209992_at	PFKFB2	6	8	209992_at	7	8
232069_at		7	7	232069_at	7	7
201369_s_at	ZFP36L2	6	9	201369_s_at	6	9
236450_at		6	9	236450_at	7	10
232164_s_at	EPPK1	7	7	232164_s_at	7	7
202833_s_at	SERPINA1	7	5	202833_s_at	7	5
232165_at	EPPK1	7	5	232165_at	7	4
206637_at	P2RY14	7	7	206637_at	7	7
208438_s_at	FGR	6	5	208438_s_at	6	5
208949_s_at	LGALS3	7	5	208949_s_at	6	5
211429_s_at	MYCPBP	6	3	211429_s_at	6	4
229985_at	BTNL9	7	8	229985_at	6	9
202908_at	WFS1	7	7	202908_at	8	7

A.1.2 Comparison of the number of patients who passed selection criteria in Table S2B from original article (left) and reproduced (right) in our study

ID	Gene symbol	Repressed 6 h	Repressed 24 h	ID	Repressed 6 h	Repressed 24 h
207165_at	HMMR	9	11	207165_at	9	11
207828_s_at	CENPF	9	11	207828_s_at	8	11
201291_s_at	TOP2A	10	10	201291_s_at	10	10
219918_s_at	ASPM	9	11	219918_s_at	9	11
203764_at	DLG7	10	10	203764_at	9	10
202870_s_at	CDC20	9	11	202870_s_at	10	11
201292_at	TOP2A	9	11	201292_at	9	11
204962_s_at	CENPA	9	10	204962_s_at	10	9
202954_at	UBE2C	7	10	202954_at	7	10
228273_at	FLJ11029	6	11	228273_at	6	11
209709_s_at	HMMR	9	10	209709_s_at	9	10
225834_at		11	9	225834_at	10	9
223381_at	CDCA1	7	11	223381_at	8	10
202705_at	CCNB2	8	10	202705_at	7	10
235574_at	GBP4	8	9	235574_at	8	11
204444_at	KIF11	6	10	204444_at	6	10
214710_s_at	CCNB1	8	11	214710_s_at	8	11
218542_at	C10orf3	8	10	218542_at	8	10
204709_s_at	KIF23	7	7	204709_s_at	7	8
209714_s_at	CDKN3	8	8	209714_s_at	7	8
218755_at	KIF20A	8	9	218755_at	8	10
1555758_a_at	CDKN3	9	9	1555758_a_at	9	10
219148_at	TOPK	6	9	219148_at	6	9
212022_s_at	MKI67	7	8	212022_s_at	7	8
212023_s_at	MKI67	8	9	212023_s_at	9	9
206364_at	KIF14	7	9	206364_at	7	9
202095_s_at	BIRC5	8	11	202095_s_at	5	11
212020_s_at	MKI67	6	8	212020_s_at	7	7
222958_s_at	DEPDC1	8	7	222958_s_at	8	7
205046_at	CENPE	7	7	205046_at	7	6
228071_at	GIMAP7	7	3	228071_at	7	3
204822_at	TTK	7	9	204822_at	7	10
204641_at	NEK2	6	8	204641_at	8	7
220359_s_at	ARPP-21	7	5	220359_s_at	7	6
228729_at	CCNB1	6	8	228729_at	7	10
203554_x_at	PTTG1	7	10	203554_x_at	7	10
212021_s_at	MKI67	7	9	212021_s_at	6	10

A.1.3 Comparison of the number of patients who passed selection criteria in Table S2C from original article (left) and reproduced (right) in our study

ID	Gene symbol	Induced 6 h	Induced 24 h	ID	Induced 6 h	Induced 24 h
224325_at	FZD8	4	8	224325_at	5	7
235735_at	TNFSF8	5	12	235735_at	5	11
233921_s_at	MAD1L1	5	8	233921_s_at	5	7
236512_at	SESN1	3	9	236512_at	3	10
202917_s_at	S100A8	5	6	202917_s_at	5	6
204150_at	STAB1	5	8	204150_at	5	8
234989_at	TncRNA	4	8	234989_at	4	8
244357_at	LOC64744	3	9	244357_at	3	9
205786_s_at	ITGAM	3	8	205786_s_at	3	8
225685_at	CDC42EP3	3	10	225685_at	3	10
209286_at	CDC42EP3	4	9	209286_at	4	9
201012_at	ANXA1	5	4	201012_at	5	4
204232_at	FCER1G	5	6	204232_at	5	5
219230_at	FLJ10970	4	8	219230_at	4	8
209288_s_at	CDC42EP3	4	10	209288_s_at	4	10
38487_at	STAB1	5	8	38487_at	6	8
1556472_s_at	SCML4	5	7	1556472_s_at	5	7
213975_s_at	LYZ	4	4	213975_s_at	4	4
242551_at		3	9	242551_at	3	8
210538_s_at	BIRC3	4	6	210538_s_at	4	5
222281_s_at		5	9	222281_s_at	5	9
227405_s_at	FZD8	3	8	227405_s_at	3	7
203535_at	S100A9	4	5	203535_at	4	5
214414_x_at	HBA1	5	5	214414_x_at	4	5
244358_at		4	7	244358_at	5	7
1569477_at	FOXO3A	5	8	1569477_at	4	8
241893_at		1	8	241893_at	1	8
240665_at	CUGBP2	3	8	240665_at	3	8

A.1.4 Comparison of the number of patients who passed selection criteria in Table S2D from original article (left) and reproduced (right) in our study

ID	Gene symbol	Repressed 6 h	Repressed 24 h	ID	Repressed 6 h	Repressed 24 h
210052_s_at	TPX2	5	9	210052_s_at	4	9
1554696_s_at	TYMS	5	10	1554696_s_at	5	9
222680_s_at	RAMP	4	10	222680_s_at	5	9
209642_at	BUB1	5	8	209642_at	5	9
219306_at	KNSL7	5	9	219306_at	6	9
218663_at	HCAP-G	5	10	218663_at	5	9
205394_at	CHEK1	2	10	205394_at	3	9
212141_at	MCM4	4	9	212141_at	3	9
220651_s_at	MCM10	3	11	220651_s_at	2	11
202503_s_at	KIAA0101	1	10	202503_s_at	1	10
1554037_a_at	ZBTB24	4	6	1554037_a_at	4	5
242870_at	KIAA1238	5	4	242870_at	4	4
1560610_at	FLJ37673	5	7	1560610_at	5	7
203213_at	CDC2	4	10	203213_at	6	11
203214_x_at	CDC2	5	9	203214_x_at	6	9
219978_s_at	NUSAP1	4	9	219978_s_at	5	7
1554768_a_at	MAD2L1	5	9	1554768_a_at	5	10
204146_at	RAD51AP1	4	10	204146_at	4	9
210559_s_at	CDC2	4	9	210559_s_at	3	10
212281_s_at	MAC30	3	8	212281_s_at	3	9
212949_at	BRRN1	3	9	212949_at	3	9
228033_at	E2F7	3	10	228033_at	3	9
213599_at	OIP5	4	11	213599_at	4	9
218039_at	NUSAP1	3	10	218039_at	5	10
204768_s_at	FEN1	2	8	204768_s_at	2	8
204825_at	MELK	4	9	204825_at	4	9
212142_at	MCM4	3	7	212142_at	3	7
212282_at	MAC30	2	9	212282_at	1	9
218782_s_at	ATAD2	1	8	218782_s_at	1	9
219990_at	FLJ23311	3	9	219990_at	4	9
235609_at	BRIP1	3	9	235609_at	2	9
204127_at	RFC3	3	7	204127_at	1	7
222740_at	ATAD2	1	10	222740_at	2	10
204126_s_at	CDC45L	2	10	204126_s_at	2	10
209773_s_at	RRM2	0	8	209773_s_at	1	8
218585_s_at	RAMP	0	9	218585_s_at	0	10
221521_s_at	Pfs2	3	9	221521_s_at	4	9
227350_at		3	7	227350_at	2	6

A.2 Extra probe sets from our data analysis

After repeating the analysis of original authors' microarray data, we found that the two subtypes should be separated in the data analysis. We separated patients (T- and B-ALL) and proposed new criteria: log ratio of ± 1 or higher for at least five out of ten B-ALL patients and two out of three T-ALL patients. We also extended the analysis to response between 6 and 24 hours. The results are shown in the following tables with details of probe sets at each time point:

	0-6 hours		6-24 hours		0-24 hours	
	Induced	Repressed	Induced	Repressed	Induced	Repressed
T-ALL	19	10	59	51	58	40
B-ALL	24	23	16	13	73	108

A.2.1 Extra probe sets found in this study for B-ALL at each time point

B-ALL			
0-6 hours		6-24 hours	
Induced	Repressed	Induced	Repressed
1556472_s_at	201291_s_at	1564077_at	1554696_s_at
202833_s_at	201292_at	1565599_at	201291_s_at
202908_at	202870_s_at	202917_s_at	201292_at
202917_s_at	202954_at	204150_at	201890_at
203761_at	203708_at	206637_at	202503_s_at
204232_at	203764_at	209286_at	209773_s_at
204560_at	204709_s_at	228854_at	210559_s_at
205681_at	204962_s_at	228964_at	211430_s_at
205883_at	205046_at	233223_at	217022_s_at
206637_at	206364_at	235735_at	218542_at
208078_s_at	207165_at	236450_at	218585_s_at
208438_s_at	207828_s_at	240665_at	220651_s_at
208949_s_at	209642_at	241893_at	235609_at
211429_s_at	209709_s_at	242551_at	
219607_s_at	211302_s_at	244357_at	
223194_s_at	217373_x_at	244414_at	
224856_at	218755_at		
226733_at	219918_s_at		
227265_at	222326_at		
228854_at	222958_s_at		
232069_at	223381_at		
232164_s_at	228071_at		
232165_at	235574_at		
238447_at			

B-ALL				
0-24 hours				
Induced	Induced	Repressed	Repressed	Repressed
1553906_s_at	227405_s_at	1554696_s_at	209709_s_at	227921_at
1560706_at	227611_at	1554733_at	209773_s_at	228033_at
1564424_at	227762_at	1554768_a_at	210052_s_at	228273_at
1564525_at	228434_at	1555758_a_at	210334_x_at	228729_at
1565752_at	228697_at	1556598_at	210559_s_at	229490_s_at
1569225_a_at	228854_at	1557910_at	210948_s_at	235088_at
202908_at	229958_at	1560610_at	211341_at	235287_at
202917_s_at	229985_at	1565602_at	212020_s_at	235574_at
203543_s_at	232164_s_at	201013_s_at	212021_s_at	235609_at
203695_s_at	232583_at	201014_s_at	212022_s_at	236641_at
203760_s_at	233921_s_at	201291_s_at	212023_s_at	241926_s_at
203761_at	234989_at	201292_at	212141_at	242787_at
204150_at	235735_at	201577_at	212142_at	
204560_at	236450_at	201890_at	212279_at	
205033_s_at	236512_at	202095_s_at	212281_s_at	
205099_s_at	236931_at	202345_s_at	212282_at	
205786_s_at	240019_at	202503_s_at	212949_at	
205883_at	240038_at	202589_at	213599_at	
206618_at	240665_at	202705_at	214452_at	
206637_at	241893_at	202870_s_at	214710_s_at	
207697_x_at	242551_at	202954_at	215117_at	
208078_s_at	244026_at	203213_at	218039_at	
209286_at	244357_at	203214_x_at	218355_at	
209288_s_at	244697_at	203362_s_at	218542_at	
210001_s_at	38487_at	203554_x_at	218585_s_at	
210146_x_at		203612_at	218662_s_at	
210448_s_at		203755_at	218663_at	
212771_at		203764_at	218755_at	
212912_at		203968_s_at	219148_at	
215528_at		204026_s_at	219306_at	
215602_at		204126_s_at	219493_at	
218638_s_at		204128_s_at	219918_s_at	
219230_at		204146_at	219978_s_at	
221756_at		204444_at	219990_at	
222062_at		204641_at	220085_at	
222281_s_at		204709_s_at	220448_at	
223027_at		204822_at	220651_s_at	
223028_s_at		204825_at	221258_s_at	
223194_s_at		204962_s_at	221521_s_at	
224325_at		205393_s_at	221591_s_at	
224840_at		205394_at	222037_at	
224856_at		206102_at	222680_s_at	
225207_at		206364_at	222740_at	
225239_at		207165_at	223229_at	
225685_at		207828_s_at	223381_at	
225949_at		209172_s_at	223556_at	
226733_at		209642_at	225834_at	
226982_at		209714_s_at	226980_at	

A.2.2 Extra probe sets found in this study for T-ALL at each time point

T-ALL					
0-6 hours		6-24 hours			
Induced	Repressed	Induced	Induced	Repressed	Repressed
1555372_at	1555745_a_at	1552439_s_at	226530_at	1552742_at	242829_x_at
1559975_at	203395_s_at	1553338_at	226632_at	1552921_a_at	242894_at
1562230_at	204115_at	1556682_s_at	226811_at	1552925_at	243649_at
209992_at	204700_x_at	1557961_s_at	230690_at	1555372_at	244523_at
215447_at	205094_at	201061_s_at	231005_at	1557257_at	
226733_at	205863_at	201694_s_at	235384_at	1558143_a_at	
227062_at	210279_at	202917_s_at	238999_at	1559119_at	
228434_at	212998_x_at	202975_s_at	240336_at	1561973_at	
230175_s_at	219918_s_at	202976_s_at	243398_at	1562230_at	
231437_at	225834_at	203395_s_at	243904_at	1569600_at	
232231_at		203502_at	244447_at	201761_at	
232744_x_at		203535_at	38671_at	201890_at	
235213_at		203574_at		202503_s_at	
235412_at		203936_s_at		202643_s_at	
235735_at		204018_x_at		204146_at	
237009_at		204081_at		204285_s_at	
241819_at		204115_at		204444_at	
242210_at		204419_x_at		204768_s_at	
242248_at		204848_x_at		205024_s_at	
		205033_s_at		205347_s_at	
		205262_at		205559_s_at	
		205653_at		209522_s_at	
		205780_at		209773_s_at	
		205863_at		210001_s_at	
		205950_s_at		210034_s_at	
		206390_x_at		210356_x_at	
		206655_s_at		212195_at	
		209301_at		213459_at	
		209458_x_at		214805_at	
		210314_x_at		215330_at	
		210384_at		216834_at	
		210982_s_at		218585_s_at	
		211560_s_at		218782_s_at	
		211699_x_at		220651_s_at	
		211745_x_at		222018_at	
		211899_s_at		222303_at	
		212235_at		222680_s_at	
		213515_x_at		223570_at	
		213817_at		226287_at	
		213975_s_at		229882_at	
		214146_s_at		232344_at	
		215894_at		234150_at	
		217414_x_at		236439_at	
		217478_s_at		236528_at	
		219672_at		239504_at	
		222164_at		241403_at	
		225481_at		242210_at	

T-ALL			
0-24 hours			
Induced	Induced	Repressed	Repressed
1553338_at	215150_at	1554696_s_at	222740_at
1557961_s_at	215447_at	201291_s_at	223062_s_at
1569153_at	215679_at	202503_s_at	223475_at
200665_s_at	215894_at	202870_s_at	223570_at
201060_x_at	216067_at	204127_at	223666_at
201061_s_at	219049_at	204285_s_at	224797_at
202975_s_at	220585_at	204439_at	225285_at
202976_s_at	221756_at	204695_at	225655_at
203949_at	221757_at	204768_s_at	226013_at
204081_at	225202_at	204836_at	226677_at
204419_x_at	225328_at	205347_s_at	228190_at
205033_s_at	226530_at	205552_s_at	228273_at
205857_at	227309_at	205898_at	
206390_x_at	228854_at	206102_at	
206655_s_at	230690_at	206749_at	
206828_at	231332_at	207826_s_at	
208078_s_at	232431_at	208782_at	
209771_x_at	232744_x_at	210052_s_at	
210244_at	235343_at	212141_at	
210971_s_at	235412_at	214710_s_at	
212364_at	235568_at	216510_x_at	
212365_at	235735_at	218585_s_at	
213093_at	237324_s_at	219918_s_at	
213515_x_at	238342_at	222036_s_at	
214146_s_at		222680_s_at	

References

- Aburatani, S., Goto, K., Saito, S., Toh, H., & Horimoto, K. (2005). ASIAN: a web server for inferring a regulatory network framework from gene expression profiles. *Nucleic Acids Research*, 33(Web Server Issue), W659-W664.
- Adams, J. M., & Cory, S. (2007). The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene*, 26(9), 1324-1337.
- Affymetrix. (2002). *Statistical algorithms description document* (Technical report).
- Allison, D. B., Page, G. P., Beasley, T. M., & Edwards, J. W. (2006). *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments*: CRC Press.
- Amaratunga, D., & Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*: Wiley-Interscience.
- Amato, R., Ciaramella, A., Deniskina, N., Mondo, C. D., di Bernardo, D., Donalek, C. (2006). A multi-step approach to time series analysis and gene expression clustering. *Bioinformatics*, 22(5), 589-596.
- Ananko, E. A., Podkolodny, N. L., Stepanenko, I. L., Podkolodnaya, O. A., Rasskazov, D. A., Miginsky, D. S., et al. (2005). GeneNet in 2005. *Nucleic Acids Research*, 33(Database Issue), D425-D427.
- Andersson, A., Edén, P., Lindgren, D., Nilsson, J., Lassen, C., Heldrup, J. (2005). Gene expression profiling of leukemic cell lines reveals conserved molecular signatures among subtypes with specific genetic aberrations. *Leukaemia*, 19(6), 1042-1050.
- Andersson, A., Olofsson, T., Lindgren, D., Nilsson, B., Ritz, C., Edén, P. (2005). Molecular signatures in childhood acute leukaemia and their correlations to expression patterns in normal hematopoietic subpopulations. *Proceedings of the National Academy of Sciences*, 102(52), 19069-19074.
- Andersson, A., Ritz, C., Lindgren, D., Edén, P., Lassen, C., Heldrup, J. (2007). Microarray-based classification of a consecutive series of 121 childhood acute leukaemias: prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukaemia*, 21, 1198-1203.
- Andrecut, M., & Kauffman, S. A. (2006). A simple method for reverse engineering causal networks. *Journal of Physics A: Mathematical and General*, 39(46), L647-L655.
- Androulakis, I. P., Yang, E., & Almon, R. R. (2007). Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities. *Annual Review of Biomedical Engineering*, 9, 205-228.
- Ao, S. I., & Ng, M. K. (2006). Gene expression time series modelling with principal component and neural network. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 10(4), 351-358.
- Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia. *Nature Genetics*, 30(1), 41-47.
- Babu, M. (2008). Computational approaches to study transcriptional regulation. *Biochemical Society Transactions*, 36, 758-765.
- Bachmann, P. S., Gorman, R., Papa, R. A., Bardell, J. E., Ford, J., Kees, U. R. (2007). Divergent mechanisms of glucocorticoid resistance in experimental models of pediatric acute lymphoblastic leukaemia. *Cancer Research*, 67(9), 4482-4490.
- Bandyopadhyay, S., Maulik, U., & Wang, J. T. L. (2007). *Analysis of Biological Data: A Soft Computing Approach*: World Scientific Publishing Company.
- Bandyopadhyay, S., Mukhopadhyay, A., & Maulik, U. (2007). An improved algorithm for clustering gene expression data. *Bioinformatics*, 23(21), 2859-2865.

- Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3, 78.
- Bar-Joseph, Z. (2004). Analyzing time series gene expression data. *Bioinformatics*, 20(16), 2493-2503.
- Barnes, M. R. (2007). *Bioinformatics for geneticists: a bioinformatics primer for the analysis of genetic data*: Wiley.
- Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Research*, 35(Database issue), D760-D765.
- Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C. (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37(Database issue), D885-D890.
- Batchelor, E., Loewer, A., & Lahav, G. (2009). The ups and downs of p53: understanding protein dynamics in single cells. *Nature Reviews Cancer*, 9(5), 371-377.
- Benson, M., & Breitling, R. (2006). Network theory to understand microarray studies of complex diseases. *Current Molecular Medicine*, 6(6), 695-701.
- Berkhin, P. (2006). Survey of clustering data mining techniques. In *Grouping Multidimensional Data* (pp. 25-71): Springer Berlin Heidelberg.
- Bhojwani, D., Kang, H., Moskowitz, N. P., Min, D. J., Lee, H., Potter, J. W. (2006). Biologic pathways associated with relapse in childhood acute lymphoblastic leukaemia: a Children's Oncology Group study. *Blood*, 108(2), 711-717.
- Bhojwani, D., Moskowitz, N., Raetz, E. A., & Carroll, W. L. (2007). Potential of Gene Expression Profiling in the Management of Childhood Acute Lymphoblastic Leukaemia. *Pediatric Drugs*, 9(3), 149-156.
- Bolshakova, N., & Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Processing*, 83(4), 825-833.
- Bolshakova, N., Zamolotskikh, A., & Cunningham, P. (2006, 22-23 June). *Comparison of the Data-based and Gene Ontology-Based Approaches to Cluster Validation Methods for Gene Microarrays*. Paper presented at the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS 2006), Salt Lake City, Utah, USA.
- Brohee, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Research*, 36(Web Server), W444-W451.
- Cario, G., Fetz, A., Bretscher, C., Mörnicke, A., Schrauder, A., Stanulla, M. (2008). Initial leukemic gene expression profiles of patients with poor in vivo prednisone response are similar to those of blasts persisting under prednisone treatment in childhood acute lymphoblastic leukaemia. *Annals of Hematology*, 87(9), 709-716.
- Carroll, W. L., Bhojwani, D., Min, D. J., Moskowitz, N., & Raetz, E. A. (2005). Childhood acute lymphoblastic leukaemia in the age of genomics. *Pediatric Blood Cancer*, 46, 560-578.
- Carroll, W. L., Bhojwani, D., Min, D. J., Raetz, E., Relling, M., Davies, S. (2003). Pediatric Acute Lymphoblastic Leukaemia. *ASH Education Program Book*, 2003(1), 102-131.
- Cary, M. P., Bader, G. D., & Sander, C. (2005). Pathway information for systems biology. *Federation of European Biochemical Societies (FEBS) Letters*, 579(8), 1815-1820.
- Cavaliere, D., & De Filippo, C. (2005). Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug discovery today: Biosilico*, 10(10), 727-734.
- Ceballos, E., Muñoz-Alonso, M. J., Berwanger, B., Acosta, J. C., Hernández, R., Krause, M. (2005). Inhibitory effect of c-Myc on p53-induced apoptosis in leukaemia cells. Microarray analysis reveals defective induction of p53 target genes and upregulation of chaperone genes. *Oncogene*, 24(28), 4559-4571.

- Chaiboonchoe, A., Samarasinghe, S., & Kulasiri, D. (2010). Machine Learning for Childhood Acute Lymphoblastic Leukaemia Gene Expression Data Analysis: A Review. *Current Bioinformatics*, 5(2), 118-133.
- Chan, Z. S. H., Havukkala, I., Jain, V., Hu, Y., & Kasabov, N. (2008). Soft computing methods to predict gene regulatory networks: An integrative approach on time-series gene expression data. *Applied Soft Computing Journal*, 8(3), 1189-1199.
- Chen, K. C., Wang, T. Y., Tseng, H. H., Huang, C. Y. F., & Kao, C. Y. (2005). A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics*, 21(12), 2883-2890.
- Cheok, M. H., Yang, W., Pui, C. H., Downing, J. R., Cheng, C., Naeve, C. W. (2003). Treatment-specific changes in gene expression discriminate in vivo drug response in human leukaemia cells. *Nature Genetics*, 34(1), 85-90.
- Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F. (2004). Gene expression profile of adult T-cell acute lymphocytic leukaemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7), 2771-2778.
- Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Wang, K. S., Mandelli, F. (2005). Gene Expression Profiles of B-lineage Adult Acute Lymphocytic Leukaemia Reveal Genetic Patterns that Identify Lineage Derivation and Distinct Mechanisms of Transformation. *Clinical Cancer Research*, 11, 7209-7219.
- Cho, K.-H., Choo, S.-M., Jung, S. H., Kim, J.-R., Choi, H.-S., & Kim, J. (2007). Reverse engineering of gene regulatory networks. *IET Systems Biology* 1(3), 149-163.
- Cho, S. B., & Won, H. H. (2003, 4-7 February). *Machine learning in DNA microarray analysis for cancer classification*. Paper presented at the the First Asia-Pacific bioinformatics conference on Bioinformatics (APBC 2003), Adelaide, Australia
- Choi, Y. L., Tsukasaki, K., O'Neill, M. C., Yamada, Y., Onimaru, Y., Matsumoto, K., et al. (2007). A genomic analysis of adult T-cell leukaemia. *Oncogene*, 26, 1245-1255.
- Chou, J. W., Zhou, T., Kaufmann, W. K., Paules, R. S., & Bushel, P. R. (2007). Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes. *BMC Bioinformatics*, 8(1), 427.
- Christensen, C., Thakar, J., & Albert, R. (2007). Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. *IET System Biology*, 1(2), 61-77.
- Chu, G., Narasimhan, B., Tibshirani, R., & Tusher, V. (2002). *SAM (Significance Analysis of Microarrays)-Users guide and technical document*: Stanford University.
- Conesa, A., Nueda, M. J., Ferrer, A., & Talon, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9), 1096-1102.
- Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., & Speed, T. P. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20(3), 323-331.
- Corduas, M., & Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis*, 52(4), 1860-1872.
- Costlow, M. E., Pui, C. H., & Dahl, G. V. (1982). Glucocorticoid receptors in childhood acute lymphocytic leukaemia. *Cancer Research*, 42(11), 4801-4806.
- D'haeseleer, P., Liang, S., & Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8), 707-726.
- Dai, Y., Rahmani, M., Dent, P., & Grant, S. (2005). Blockade of histone deacetylase inhibitor-induced RelA/p65 acetylation and NF- κ B activation potentiates apoptosis in leukaemia cells through a process mediated by oxidative damage, XIAP downregulation, and c-Jun N-terminal kinase 1 activation. *Molecular and cellular biology*, 25(13), 5429-5444.

- Dam, H. H., Abbass, H. A., Lokan, C., & Yao, X. (2007). Neural-Based Learning Classifier Systems. *IEEE Transactions on Knowledge and Data Engineering*, 20(1), 26-39.
- Das, R., Kalita, J., & Bhattacharyya, D. K. (2009). A new approach for clustering gene expression time series data. *International Journal of Bioinformatics Research and Applications*, 5(3), 310-328.
- De Bosscher, K., Vanden Berghe, W., & Haegeman, G. (2003). The interplay between the glucocorticoid receptor and nuclear factor- B or activator protein-1: molecular mechanisms for gene repression. *Endocrine reviews*, 24(4), 488-522.
- de Jong, H. (2002). Modelling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of computational biology*, 9(1), 67-103.
- De Pitta, C., Tombolan, L., Campo Dell'Orto, M., Accordi, B., te Kronnie, G., Romualdi, C. (2005). A leukaemia-enriched cDNA microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukaemia. *Haematologica*, 90(7), 890-898.
- DeAngelo, D. J. (2005). The Treatment of Adolescents and Young Adults with Acute Lymphoblastic Leukaemia. *ASH Education Program Book*, 2005(1), 123-130.
- Den Boer, M. L., van Slegtenhorst, M., De Menezes, R. X., Cheok, M. H., Buijs-Gladdines, J. G., Peters, S. T. (2009). A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *The lancet oncology*, 10(2), 125-134.
- Di Camillo, B., Toffolo, G., Nair, S., Greenlund, L., & Cobelli, C. (2007). Significance analysis of microarray transcript levels in time series experiments. *BMC Bioinformatics*, 8(Suppl 1), S10.
- Donn, R., Berry, A., Stevens, A., Farrow, S., Betts, J., Stevens, R. (2007). Use of gene expression profiling to identify a novel glucocorticoid sensitivity determining gene, BMPRII. *The Federation of American Societies for Experimental Biology (FASEB) Journal*, 21(2), 402-414.
- Douzal-Chouakria, A., Diallo, A., & Giroud, F. (2009). Adaptive clustering for time series: Application for identifying cell cycle expressed genes. *Computational Statistics and Data Analysis*, 53(4), 1414-1426.
- Drachen, A., Canossa, A., & Yannakakis, G. N. (2009, 7-10 September). *Player Modelling using Self-Organization in Tomb Raider: Underworld*. Paper presented at the IEEE Symposium on Computational Intelligence and Games, Milano, Italy.
- Dubitzky, W., Granzow, M., Downes, C., & Berrar, D. (2003). Introduction to microarray data analysis. In D. P. Berrar, W. Dubitzky & M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis* (pp. 1–46): Kluwer Academic Publishers.
- Dunphy, C. H. (2006). Gene Expression Profiling Data in Lymphoma and Leukaemia: Review of the Literature and Extrapolation of Pertinent Clinical Applications. *Archives of Pathology and Laboratory Medicine*, 130(4), 483-520.
- Dzeroski, S., Gjorgjioski, V., Slavkov, I., & Struyf, J. (2007). Analysis of time series data with predictive clustering trees. In S. Džeroski & J. Struyf (Eds.), *Knowledge Discovery in Inductive Databases* (Vol. 4747, pp. 63-80): Springer Berlin/ Heidelberg.
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences (PNAS)*, 95(25), 14863-14868.
- Ernst, J., & Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7(1), 191.
- Ernst, J., Nau, G. J., & Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, 21(Suppl 1), S159-S168.
- Ertel, A., Verghese, A., Byers, S. W., Ochs, M., & Tozeren, A. (2006). Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Molecular Cancer*, 5, 55.

- Estes, D. A., Lovato, D. M., Khawaja, H. M., Winter, S. S., & Larson, R. S. (2007). Genetic alterations determine chemotherapy resistance in childhood T-ALL: modelling in stage-specific cell lines and correlation with diagnostic patient samples. *British journal of haematology*, 139(1), 20-30.
- Famili, A. F., Liu, Z., Ouyang, J., Walker, P. R., Smith, B., O'Connor, M. (2004, 23-27 August). *A novel data mining technique for gene identification in time-series gene expression data*. Paper presented at the The 16th European Conference on Artificial Intelligence (ECAI 2004), Valencia, Spain.
- Fernandez, E. A., & Balzarini, M. (2007). Improving cluster visualization in self-organizing maps: Application in gene expression data analysis. *Computers in Biology and Medicine*, 37(12), 1677-1689.
- Ferrazzi, F., & Bellazzi, R. (2007). Control and Systems Fundamentals. In G. Alterovitz & M. F. Ramoni (Eds.), *Systems bioinformatics: an engineering case-based approach* (pp. 127-147). Norwood, MA, USA: Artech House, Inc.
- Flotho, C., Coustan-Smith, E., Pei, D., Cheng, C., Song, G., Pui, C. H. (2007). A set of genes that regulate cell proliferation predicts treatment outcome in childhood acute lymphoblastic leukaemia. *Blood*, 110(4), 1271-1277.
- Folarin, A., & Bioinformatics, M. R. (2003). *Partitioner, ClusterBrowser And ClusterMapper Modules Of MicroCore Provide Gene Ontology And Pathway Network Driven Paradigms For The Exploration Syn-expression Groups in Microarray Data*. University of London., Birkbeck College.
- Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4), 601-620.
- Fu, L., & Medico, E. (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, 8, 3.
- Fulci, V., Colombo, T., Chiaretti, S., Messina, M., Citarella, F., Tavoraro, S. (2009). Characterization of B-and T-lineage acute lymphoblastic leukaemia by integrated analysis of MicroRNA and mRNA expression profiles. *Genes, Chromosomes and Cancer*, 48(12), 1069-1082.
- Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., & Kitano, H. (2008). CellDesigner 3.5: a versatile modelling tool for biochemical networks. *Proceedings of the IEEE*, 96(8), 1254-1265.
- Gardiner-Garden, M., & Littlejohn, T. (2001). A comparison of microarray databases. *Briefings in Bioinformatics*, 2(2), 143-158.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439), 531-537.
- Greenstein, S., Ghias, K., Krett, N. L., & Rosen, S. T. (2002). Mechanisms of glucocorticoid-mediated apoptosis in hematological malignancies. *Clinical Cancer Research*, 8(6), 1681-1694.
- Gros, F., Sebti, Y., de Guibert, S., Branger, B., Bernard, M., Fauchet, R. (2006). Soluble HLA-G molecules are increased during acute leukaemia, especially in subtypes affecting monocytic and lymphoid lineages. *Neoplasia (New York, NY)*, 8(3), 223-230.
- Guicciardi, M. E., Leist, M., & Gores, G. J. (2004). Lysosomes in cell death. *Oncogene*, 23, 2881-2890.
- Hache, H., Wierling, C., Lehrach, H., & Herwig, R. (2009). GeNGe: systematic generation of gene regulatory networks. *Bioinformatics*, 25(9), 1205-1207.
- Haddad, I., Hiller, K., Frimmersdorf, E., Benkert, B., Schomburg, D., & Jahn, D. (2009). An emergent self-organizing map based analysis pipeline for comparative metabolome studies. *In Silico Biology*, 9, 0014.

- Hammer, B., Micheli, A., Neubauer, N., Sperduti, A., & Strickert, M. (2005, 5-8 September). *Self-Organizing Maps for Time Series*. Paper presented at the 5th Workshop On Self-Organizing Maps (WSOM 05) University Paris, Panthéon-Sorbonne, Paris, France.
- Harris, S. L., & Levine, A. J. (2005). The p53 pathway: positive and negative feedback loops. *Oncogene*, *24*(17), 2899-2908.
- Heiner, M., Koch, I., & Will, J. (2004). Model validation of biological pathways using Petri nets—demonstrated for apoptosis. *Biosystems*, *75*(1-3), 15-28.
- Herold, R., von Stackelberg, A., Hartmann, R., Eisenreich, B., & Henze, G. (2004). Acute Lymphoblastic Leukaemia-Relapse Study of the Berlin-Frankfurt-Munster Group (ALL-REZ BFM) Experience: Early Treatment Intensity Makes the Difference. *Journal of Clinical Oncology*, *22*(3), 569-570.
- Herr, I., Gassler, N., Friess, H., & Büchler, M. W. (2007). Regulation of differential pro-and anti-apoptotic signaling by glucocorticoids. *Apoptosis*, *12*(2), 271-291.
- Herrero, J., Valencia, A., & Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, *17*(2), 126-136.
- Holleman, A., Cheok, M. H., den Boer, M. L., Yang, W., Veerman, A. J. P., Kazemier, K. M. (2004). Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukaemia Cells and Response to Treatment. *New England Journal of Medicine*, *351*(6), 533-542.
- Holleman, A., den Boer, M. L., de Menezes, R. X., Cheok, M. H., Cheng, C., Kazemier, K. M. (2006). The expression of 70 apoptosis genes in relation to lineage, genetic subtype, cellular drug resistance, and outcome in childhood acute lymphoblastic leukaemia. *Blood*, *107*(2), 769-776.
- Howell, D. L., Ward, K. C., Austin, H. D., Young, J. L., & Woods, W. G. (2007). Access to Pediatric Cancer Care by Age, Race, and Diagnosis, and Outcomes of Cancer Treatment in Pediatric and Adolescent Patients in the State of Georgia. *Journal of Clinical Oncology*, *25*(29), 4610-4615.
- Hu, Z., Hung, J. H., Wang, Y., Chang, Y. C., Huang, C. L., Huyck, M. (2009). VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Research*, *37*(Web Server issue), W115-W121.
- Hu, Z., Snitkin, E. S., & DeLisi, C. (2008). VisANT: an integrative framework for networks in systems biology. *Briefings in Bioinformatics*, *9*(4), 317-325.
- Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, *19*(17), 2271-2282.
- Igarashi, S., Manabe, A., Ohara, A., Kumagai, M., Saito, T., Okimoto, Y. (2005). No advantage of dexamethasone over prednisolone for the outcome of standard-and intermediate-risk childhood acute lymphoblastic leukaemia in the Tokyo Children's Cancer Study Group L95-14 protocol. *Journal of Clinical Oncology*, *23*(27), 6489.
- Igney, F. H., & Krammer, P. H. (2002). Death and anti-death: tumour resistance to apoptosis. *Nat Rev Cancer*, *2*(4), 277-288.
- Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, *31*(4), e15.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. W. E. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*(2), 249.
- Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C. (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods*, *2*, 345-350.
- Irizarry, R. A., Wu, Z., & Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, *22*(7), 789-794.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys*, *31*(3), 265-323.

- Jeffery, I. B., Higgins, D. G., & Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7(1), 359.
- Jenssen, T. K., Lægreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1), 21-28.
- Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 1370-1386.
- Jiang, N., Leach, L. J., Hu, X., Potokina, E., Jia, T., Druka, A. (2008). Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*, 9(1), 284.
- Johnstone, R. W., Ruefli, A. A., & Lowe, S. W. (2002). Apoptosis a Link between Cancer Genetics and Chemotherapy. *Cell*, 108(2), 153-164.
- Juruena, M. F., Cleare, A. J., Papadopoulos, A. S., Poon, L., Lightman, S., & Pariante, C. M. (2006). Different responses to dexamethasone and prednisolone in the same depressed patients. *Psychopharmacology*, 189(2), 225-235.
- Kajiyama, Y., Iijima, Y., Chiba, S., Furuta, M., Ninomiya, M., Izumi, A. (2009). Prednisolone causes anxiety-and depression-like behaviors and altered expression of apoptotic genes in mice hippocampus. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 34(1), 159-165.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27-30.
- Karawajew, L., Ruppert, V., Wuchter, C., Kosser, A., Schrappe, M., Dorken, B. (2000). Inhibition of in vitro spontaneous apoptosis by IL-7 correlates with bcl-2 up-regulation, cortical/mature immunophenotype, and better early cytoreduction of childhood T-cell acute lymphoblastic leukaemia. *Blood*, 96(1), 297-306.
- Kaspers, G. J. L., Veerman, A. J. P., Popp-Snijders, C., Lomecky, M., Van Zantwijk, C. H., Swinkels, L. (1996). Comparison of the antileukemic activity in vitro of dexamethasone and prednisolone in childhood acute lymphoblastic leukaemia. *MEDICAL AND PEDIATRIC ONCOLOGY*, 27(2), 114-121.
- Kauffman, S., Peterson, C., Samuelsson, B., & Troein, C. (2003). Random Boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences (PNAS)*, 100(25), 14796-14799.
- Kerr, G., Ruskin, H. J., Crane, M., & Doolan, P. (2007). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, 38, 283-293.
- Kharas, M. G., Janes, M. R., Scarfone, V. M., Lilly, M. B., Knight, Z. A., Shokat, K. M. (2008). Ablation of PI3K blocks BCR-ABL leukemogenesis in mice, and a dual PI3K/mTOR inhibitor prevents expansion of human BCR-ABL+ leukaemia cells. *The Journal of Clinical Investigation*, 118(9), 3038-3050.
- Kim, J., & Kim, J. H. (2007). Difference-based clustering of short time-course microarray data with replicates. *BMC Bioinformatics*, 8(1), 253.
- Kim, R., Tanabe, K., Uchida, Y., Emi, M., Inoue, H., & Toge, T. (2002). Current status of the molecular mechanisms of anticancer drug-induced apoptosis. *Cancer chemotherapy and pharmacology*, 50(5), 343-352.
- Kimura, S., Ide, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R. (2005). Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics*, 21(7), 1154-1163.
- King, K. L., & Cidlowski, J. A. (1998). Cell cycle regulation and apoptois. *Annual review of physiology*, 60(1), 601-617.
- Kirschner-Schwabe, R., Lottaz, C., Todling, J., Rhein, P., Karawajew, L., Eckert, C. (2006). Expression of Late Cell Cycle Genes and an Increased Proliferative Capacity Characterize Very Early Relapse of Childhood Acute Lymphoblastic Leukaemia. *Clinical Cancer Research*, 12(15), 4553-4561.

- Knudsen, S. (2006). *Cancer Diagnostics with DNA Microarrays*. Hoboken, New Jersey: John Wiley&Sons,Inc.
- Kofler, R. (2000). The molecular basis of glucocorticoid-induced apoptosis of lymphoblastic leukaemia cells. *Histochemistry and Cell Biology*, 114(1), 1-7.
- Kohonen, T. (1997, 9-12 June). *Exploration of very large databases by self-organizing maps*. Paper presented at the International Conference on Neural Networks, Houston, TX, USA.
- Korenberg, M. J. (2007). *Microarray Data Analysis: Methods and Applications*: Humana Press.
- Kriegel, H. P., Kroger, P., Pryakhin, A., Renz, M., & Zherdin, A. (2008). Approximate Clustering of Time Series Using Compact Model-based Descriptions. In J. R. Haritsa, R. Kotagiri & V. Pudi (Eds.), *Database Systems for Advanced Applications* (Vol. 4947, pp. 364-379): Springer Berlin/ Heidelberg.
- Krishnamurthy, L., Nadeau, J., Ozsoyoglu, G., Ozsoyoglu, M., Schaeffer, G., Tasan, M. (2003). Pathways database system: an integrated system for biological pathways. *Bioinformatics*, 19(8), 930-937.
- Kuiper, R. P., Schoenmakers, E., van Reijmersdal, S. V., Hehir-Kwa, J. Y., van Kessel, A. G., van Leeuwen, F. N. (2007). High-resolution genomic profiling of childhood ALL reveals novel recurrent genetic lesions affecting pathways involved in lymphocyte differentiation and cell cycle progression. *Leukaemia*, 21(6), 1258-1266.
- Kustanovich, A. M., Savitskaja, T. V., Bydanov, O. I., Belevtsev, M. V., & Potapnev, M. P. (2005). Aberrant expression of tumor suppressor genes and their association with chimeric oncogenes in pediatric acute lymphoblastic leukaemia. *Leukaemia Research*, 29(11), 1271-1276.
- Laane, E., Panaretakis, T., Pokrovskaja, K., Buentke, E., Corcoran, M., Soderhall, S. (2007). Dexamethasone-induced apoptosis in acute lymphoblastic leukaemia involves differential regulation of Bcl-2 family members. *Haematologica*, 92(11), 1460-1469.
- Lau, M., & Schultz, M. (2002). *A Feature Selection Method for Gene Expression Data with Thousands of Features*: Technical Report, CS-490, Yale University.
- Lee, W. P., & Tzou, W. S. (2009). Computational methods for discovering gene networks from expression data. *Briefings in Bioinformatics*, 10(4), 408-423.
- Lee, W. P., & Yang, K. C. (2008). A clustering-based approach for inferring recurrent neural networks as gene regulatory networks. *Neurocomputing*, 71(4-6), 600-610.
- Lehwark, P., Risi, S., & Ultsch, A. (2007). Visualization and clustering of tagged music data. In *Data Analysis, Machine Learning and Applications* (pp. 673-680): Springer Berlin Heidelberg.
- Leung, Y. F., & Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *TRENDS in Genetics*, 19(11), 649-659.
- Leupin, N., Kuhn, A., Hügli, B., Grob, T. J., Jaggi, R., Tobler, A. (2006). Gene expression profiling reveals consistent differences between clinical samples of human leukaemias and their model cell lines. *British journal of haematology*, 135(4), 520-523.
- Lewin, B. (2008). *Genes IX*. Sudbury, MA: Jones and Barlett Publishers.
- Lim, W. K., Wang, K., Lefebvre, C., & Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, 23(13), i282-i288.
- Liu, T., Raetz, E., Moos, P. J., Perkins, S. L., Bruggers, C. S., Smith, F. (2002). Diversity of the apoptotic response to chemotherapy in childhood leukaemia. *Leukaemia*, 16(2), 223-232.
- Liu, X., & Kellam, P. (2003). Mining gene expression data. In C. Orengo, D. Jones & J. Thornton (Eds.), *Bioinformatics: genes, proteins and computers* (pp. 229-244): BIOS Scientific Oxford, UK.

- Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A. (2003). *Molecular cell biology* (6 ed.). New York: W.H. Freeman and company.
- Lugthart, S., Cheok, M. H., den Boer, M. L., Yang, W., Holleman, A., Cheng, C. (2005). Identification of genes associated with chemotherapy crossresistance and treatment response in childhood acute lymphoblastic leukaemia. *Cancer Cell*, 7(4), 375-386.
- Ma, P., Castillo-Davis, C. I., Zhong, W., & Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34(4), 1261-1269.
- Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, 1(1), 24-45.
- Magni, P., Ferrazzi, F., Sacchi, L., & Bellazzi, R. (2008). TimeClust: a clustering tool for gene expression time series. *Bioinformatics*, 24(3), 430-432.
- Markowitz, F., & Spang, R. (2007). Inferring cellular networks—a review. *BMC Bioinformatics*, 8(Suppl 6), S5.
- Miller, A. L., Garza, A. S., Johnson, B. H., & Thompson, E. B. (2007). Pathway interactions between MAPKs, mTOR, PKA, and the glucocorticoid receptor in lymphoid cells. *Cancer Cell International*, 7(1), 3.
- Miller, A. L., Komak, S., Webb, M. S., Leiter, E. H., & Thompson, E. B. (2007). Gene expression profiling of leukemic cells and primary thymocytes predicts a signature for apoptotic sensitivity to glucocorticoids. *Cancer Cell International*, 7, 18.
- Min, D. J., Moskowitz, N. P., Brownstein, C., Lee, H., Horton, T. M., & Carroll, W. L. (2006). Diverse pathways mediate chemotherapy-induced cell death in acute lymphoblastic leukaemia cell lines. *Apoptosis*, 11(11), 1977-1986.
- Mingoti, S. A., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, 174(3), 1742-1759.
- Mitchell, C. D., Richards, S. M., Kinsey, S. E., Lilleyman, J., Vora, A., & Eden, T. O. B. (2005). Benefit of dexamethasone compared with prednisolone for childhood acute lymphoblastic leukaemia: results of the UK Medical Research Council ALL97 randomized trial. *British journal of haematology*, 129(6), 734-745.
- Moos, P. J., Raetz, E. A., Carlson, M. A., Szabo, A., Smith, F. E., Willman, C. (2002). Identification of Gene Expression Profiles That Segregate Patients with Childhood Leukaemia 1. *Clinical Cancer Research*, 8, 3118-3130.
- Mullighan, C. G., Goorha, S., Radtke, I., Miller, C. B., Coustan-Smith, E., Dalton, J. D. (2007). Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*, 446(7137), 758-764.
- Nikkilä, J., Törönen, P., Kaski, S., Venna, J., Castrén, E., & Wong, G. (2002). Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, 15(8-9), 953-966.
- Nosaka, T., Kawashima, T., Misawa, K., Ikuta, K., Mui, A. L. F., & Kitamura, T. (1999). STAT5 as a molecular regulator of proliferation, differentiation and apoptosis in hematopoietic cells. *The EMBO Journal*, 18(17), 4754-4765.
- Pal, N. R., Aguan, K., Sharma, A., & Amari, S. (2007). Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. *BMC Bioinformatics*, 8(1), 5.
- Panetta, J. C., Evans, W. E., & Cheok, M. H. (2005). Mechanistic mathematical modelling of mercaptopurine effects on cell cycle of human acute lymphoblastic leukaemia cells. *British journal of cancer*, 94(1), 93-100.
- Phan, S., Famili, F., Tang, Z., Pan, Y., Liu, Z., Ouyang, J. (2007). A novel pattern based clustering methodology for time-series microarray data. *International Journal of Computer Mathematics*, 84(5), 585-597.

- Phillip, P. L., R., F. J., Schug, J., Brestelli, J. E., Parker, J. B., Bochkis, I. M. (2005). Glucocorticoid receptor-dependent gene regulatory networks. *PLoS Genetic*, *1*(2), e16.
- Ploner, C., Schmidt, S., Presul, E., Renner, K., Schrocksnadel, K., Rainer, J. (2005). Glucocorticoid-induced apoptosis and glucocorticoid resistance in acute lymphoblastic leukaemia. *Journal of Steroid Biochemistry and Molecular Biology*, *93*(2-5), 153-160.
- Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M., & Dedene, G. (2009a, 31 March-2 April 2009). *How Emergent Self Organizing Maps can help counter domestic violence*. Paper presented at the 2009 WRI World Congress on Computer Science and Information Engineering, California, USA.
- Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M. M., & Dedene, G. (2009b). Gaining insight in domestic violence with Emergent Self Organizing Maps. *Expert Systems With Applications*, *36*(9), 11864-11874
- Pui, C. H. (2004). Recent advances in childhood acute lymphoblastic leukaemia. *Journal Formosan Medical Association*, *103*(2), 85-95.
- Pui, C. H., Robison, L. L., & Look, A. T. (2008). Acute lymphoblastic leukaemia. *The Lancet*, *371*(9617), 1030-1043.
- Pui, C. H., Schrappe, M., Ribeiro, R. C., & Niemeyer, C. M. (2004). Childhood and Adolescent Lymphoid and Myeloid Leukaemia. *ASH Education Program Book*, *2004*(1), 118-145.
- Qin, H., Feng, T., Harding, S. A., Tsai, C. J., & Zhang, S. (2008). An efficient method to identify differentially expressed genes in microarray experiments. *Bioinformatics*, *24*(14), 1583-1589.
- Ramanujachar, R., Richards, S., Hann, I., Goldstone, A., Mitchell, C., Vora, A. (2007). Adolescents with acute lymphoblastic leukaemia: Outcome on UK national paediatric (ALL97) and adult (UKALLXII/E2993) trials. *Pediatric Blood Cancer*, *48*(3), 254-261.
- Raponi, M., Belly, R. T., Karp, J. E., Lancet, J. E., Atkins, D., & Wang, Y. (2004). Microarray analysis reveals genetic pathways modulated by tipifarnib in acute myeloid leukaemia. *BMC cancer*, *4*(1), 56.
- Reimand, J., Tooming, L., Peterson, H., Adler, P., & Vilo, J. (2008). GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Research*, *36*(Web Server issue), W452-459.
- Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B. (2007). Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia*, *9*(2), 166-180.
- Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D. (2004). ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. *Neoplasia*, *6*(1), 1-6.
- Rogers, P. C., Broemeling, A., Pritchard, S. L., Goddard, K., Xie, L., Poole, B. (2007). Research, policy and practice related to survivors of childhood, adolescent and young adult cancers in British Columbia (BC), Canada: A population-based approach. *Journal of Clinical Oncology*, *25*(18s), 9539.
- Rosato, R. R., Almenara, J. A., Dai, Y., & Grant, S. (2003). Simultaneous activation of the intrinsic and extrinsic pathways by histone deacetylase (HDAC) inhibitors and tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) synergistically induces mitochondrial damage and apoptosis in human leukaemia cells. *Molecular Cancer Therapeutics*, *2*(12), 1273-1284.
- Ross, M. E., Zhou, X., Song, G., Shurtleff, S. A., Girtman, K., Williams, W. K. (2003). Classification of pediatric acute lymphoblastic leukaemia by gene expression profiling. *Blood*, *102*(8), 2951-2959.
- Russo, G., Zegar, C., & Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. *Oncogene*, *22*(42), 6497-6507.

- Sacchi, L., Bellazzi, R., Larizza, C., Magni, P., Curk, T., Petrovic, U. (2005). TA-clustering: Cluster analysis of gene expression profiles through Temporal Abstractions. *International Journal of Medical Informatics*, 74(7-8), 505-517.
- Samarasinghe, S. (2006). *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition*: Auerbach Publications.
- Santos, C. C. d., & Liu, M. (2007). Gene expression profiling by microarray. In S. Q. Ye (Ed.), *Bioinformatics: a practical approach* (pp. 131-187): Chapman & Hall/CRC.
- Savvides, A., Promponas, V. J., & Fokianos, K. (2008). Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognition*, 41(7), 2398-2412.
- Schäfer, J., & Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6), 754-764.
- Schlitt, T., & Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8(Suppl 6), S9.
- Schmidt, S., Rainer, J., Ploner, C., Presul, E., Riml, S., & Kofler, R. (2004). Glucocorticoid-induced apoptosis and glucocorticoid resistance: molecular mechanisms and clinical relevance. *Cell death and differentiation*, 11(Suppl 1), S45-S55.
- Schmidt, S., Rainer, J., Riml, S., Ploner, C., Jesacher, S., Achmuller, C. (2006). Identification of glucocorticoid-response genes in children with acute lymphoblastic leukaemia. *Blood*, 107(5), 2061-2069.
- Schrøder, H., Kjeldahl, M., Boesen, A. M., Nielsen, O. J., Schmidt, K., Johnsen, H. E. (2006). Acute lymphoblastic leukaemia in adolescents between 10 and 19 years of age in Denmark--secondary publication. *Danish Medical Bulletin*, 53, 76-79.
- Shamir, R., & Sharan, R. (2001). Algorithmic approaches to clustering gene expression data. *Current Topics in Computational Biology*, 269-299.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498-2504.
- Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., & Zhao, Y. (2003). *Design and analysis of DNA microarray investigations*: Springer Verlag.
- Sinnett, D., Labuda, D., & Krajcinovic, M. (2006). Challenges Identifying Genetic Determinants of Pediatric Cancers--the Childhood Leukaemia Experience. *Familial Cancer*, 5(1), 35-47.
- Smedmyr, B., & Heyman, M. (2006). Treatment Outcome in Young Adults and Children > 10 Years of Age with Acute Lymphoblastic Leukaemia in Sweden. *Cancer*, 107(7), 1551-1561.
- Soengas, M. S., Alarcon, R. M., Yoshida, H., Giaccia, A. J., Hakem, R., Mak, T. W. (1999). Apaf-1 and caspase-9 in p53-dependent apoptosis and tumor inhibition. *Science*, 284, 156-159.
- Spieth, C., Supper, J., Streichert, F., Speer, N., & Zell, A. (2006). JCell-a Java-based framework for inferring regulatory networks from time series data. *Bioinformatics*, 22(16), 2051-2052.
- Stafford, P., & Tak, Y. (2008). Biological Interpretation for Microarray Normalization Selection. In P. Stafford (Ed.), *Methods in microarray normalization*. Boca Raton, FL: CRC Press.
- Steinhoff, C., & Vingron, M. (2006). Normalization and quantification of differential expression in gene expression microarrays. *Briefings in Bioinformatics*, 7(2), 166-177.
- Strickert, M., & Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, 64, 39-71.
- Styczynski, M. P., & Stephanopoulos, G. (2005). Overview of computational methods for the inference of gene regulatory networks. *Computers & Chemical Engineering*, 29(3), 519-534.

- Summa, M. G., Steyaert, J. M., Vautrain, F., & Weitkunat, R. (2007). A New Clustering Method for Time Series to Discover Geographical Cancer Trends from 1960 to 2000. *Annals of Epidemiology*, *17*(9), 744.
- Sun, T., Chen, C., Wu, Y., Zhang, S., Cui, J., & Shen, P. (2009). Modelling the role of p 53 pulses in DNA damage- induced cell death decision. *BMC Bioinformatics*, *10*(1), 190.
- Swain, M., Hunniford, T., Dubitzky, W., Mandel, J., & Palfreyman, N. (2005). Reverse-Engineering Gene-Regulatory Networks using Evolutionary Algorithms and Grid Computing. *Journal of Clinical Monitoring and Computing*, *19*(4), 329-337.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences (PNAS)*, *96*(6), 2907-2912.
- Thalamuthu, A., Mukhopadhyay, I., Zheng, X., & Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, *22*(19), 2405-2412.
- Thompson, E. B., & Johnson, B. H. (2003). Regulation of a distinctive set of genes in glucocorticoid-evoked apoptosis in CEM human lymphoid cells. *Recent Progress in Hormone Research*, *58*(1), 175-197.
- Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., & Brown, P. (1999). *Clustering methods for the analysis of DNA microarray data*. Stanford, CA: Department Statistics, Stanford University.
- Tibshirani, R., & Witten, D. M. (2007). *A comparison of fold-change and the t-statistic for microarray data analysis*: Standford University.
- Tissing, W. J. E., den Boer, M. L., Meijerink, J. P. P., Menezes, R. X., Swagemakers, S., van der Spek, P. J., et al. (2007). Genomewide identification of prednisolone-responsive genes in acute lymphoblastic leukaemia cells. *Blood*, *109*(9), 3929-3935.
- Tissing, W. J. E., Meijerink, J. P. P., den Boer, M. L., & Pieters, R. (2003). Molecular determinants of glucocorticoid sensitivity and resistance in acute lymphoblastic leukaemia. *Leukaemia*, *17*(1), 17-25.
- Toh, H., & Horimoto, K. (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modelling. *Bioinformatics*, *18*(2), 287-297.
- Tonko, M., Ausserlechner, M. J., Bernhard, D., Helmberg, A., & Kofler, R. (2001). Gene expression profiles of proliferating vs. G1/G0 arrested human leukaemia cells suggest a mechanism for glucocorticoid-induced apoptosis. *The FASEB Journal*, *15*(3), 693-699.
- Toronen, P., Kolehmainen, M., Wong, G., & Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *Federation of European Biochemical Societies (FEBS) Letter*, *451*(2), 142-146.
- Tsapis, M., Lieb, M., Manzo, F., Shankaranarayanan, P., Herbrecht, R., Lutz, P. (2007). HDAC inhibitors induce apoptosis in glucocorticoid-resistant acute lymphatic leukaemia cells despite a switch from the extrinsic to the intrinsic death pathway. *The international journal of biochemistry & cell biology*, *39*(7-8), 1500-1509.
- Tseng, G. C., & Wong, W. H. (2005). Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, *61*(1), 10-16.
- Tsui, I. F. L., Chari, R., Buys, T. P. H., & Lam, W. L. (2007). Public databases and software for the pathway analysis of cancer genomes. *Cancer Informatics*, *3*, 389-407.
- Tukey, J. W. (1977). Exploratory data analysis. *Applied Psychological Measurement*, *2*(1), 151-155.
- Ultsch, A. (2003a, 11-14 September). *Maps for the visualization of high-dimensional data spaces*. Paper presented at the Workshop on Self-Organizing Maps (WSOM 03), Fukuoka, Japan.

- Ultsch, A. (2003b, 12-14 March). *Pareto density estimation: probability density estimation for knowledge discovery*. Paper presented at the the 27th Annual Conference of the German Classification Society (GfKI), University of Technology, Cottbus, Germany.
- Ultsch, A., & Morchen, F. (2005). *ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM*: Data Bionics Research Group, University of Marburg.
- van Beek, R. D., de Muinck Keizer-Schrama, S., Hakvoort-Cammel, F. G., van der Sluis, I. M., Krenning, E. P., Pieters, R. (2006). No difference between prednisolone and dexamethasone treatment in bone mineral density and growth in long term survivors of childhood acute lymphoblastic leukaemia. *Pediatric blood & cancer*, 46(1).
- van Someren, E. P., Wessels, L. F., Backer, E., & Reinders, M. J. (2002). Genetic network modelling. *Pharmacogenomics*, 3(4), 507-525.
- Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (1999, 16-17 November). *Self-organizing map in Matlab: the SOM toolbox*. Paper presented at the Matlab DSP Conference, Espoo, Finland.
- Wang, J., Li, J., & Ruan, X. (2005, 13-15 October). *Mining Leukaemia Gene Association Structure with DNA Microarray*. Paper presented at the International Conference on Neural Networks and Brain (ICNN&B 05), Beijing
- Wang, L., Chu, F., & Xie, W. (2007, 10-12 June). *Accurate Cancer Classification Using Expressions of Very Few Genes*. Paper presented at the IEEE International Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland
- Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F. X. (2005). Gene selection from microarray data for cancer classification-a machine learning approach. *Computational Biology and Chemistry*, 29(1), 37-46.
- Wang, Y. P., Gunampally, M., Chen, J., Bittel, D., Butler, M. G., & Cai, W. W. (2008). A Comparison of Fuzzy Clustering Approaches for Quantification of Microarray Gene Expression. *Journal of Signal Processing Systems*, 50(3), 305-320.
- Wang, Z., Rong, Y. P., Malone, M. H., Davis, M. C., Zhong, F., & Distelhorst, C. W. (2005). Thioredoxin-interacting protein (txnip) is a glucocorticoid-regulated primary response gene involved in mediating glucocorticoid-induced apoptosis. *Oncogene*, 25(13), 1903-1913.
- Warren Liao, T. (2005). Clustering of time series data-a survey. *Pattern Recognition*, 38(11), 1857-1874.
- Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids A Structure for deoxyribose nucleic acid. *American Journal of Psychiatry*, 160(4), 623-624.
- Webb, M. S., Miller, A. L., Johnson, B. H., Fofanov, Y., Li, T., Wood, T. G. (2003). Gene networks in glucocorticoid-evoked apoptosis of leukemic cells. *Journal of Steroid Biochemistry and Molecular Biology*, 85(2-5), 183-193.
- Wen, J., Ramadevi, N., Nguyen, D., Perkins, C., Worthington, E., & Bhalla, K. (2000). Antileukemic drugs increase death receptor 5 levels and enhance Apo-2L-induced apoptosis of human acute leukaemia cells. *Blood*, 96(12), 3900-3906.
- Willenbrock, H., Juncker, A. S., Schmiegelow, K., Knudsen, S., & Ryder, L. P. (2004). Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukaemia using DNA microarrays. *Leukaemia*, 18, 1270-1277.
- Willman, C. L. (2004). Discovery of novel molecular classification schemes and genes predictive of outcome in leukaemia. *The Hematology Journal*, 5(Suppl 3), S138-S143.
- Winter, S. S., Jiang, Z., Khawaja, H. M., Griffin, T., Devidas, M., Asselin, B. L. (2007). Identification of genomic classifiers that distinguish induction failure in T-lineage acute lymphoblastic leukaemia: a report from the Children's Oncology Group. *Blood*, 110(5), 1429-1438.

- Wu, C. C., Huang, H. C., Juan, H. F., & Chen, S. T. (2004). GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data. *Bioinformatics*, 20(18), 3691-3693.
- Wu, Z., & Irizarry, R. A. (2004). Preprocessing of oligonucleotide array data. *Nature Biotechnology*, 22(6), 656-658.
- Wuchter, C., Krappmann, D., Cai, Z., Ruppert, V., Scheidereit, C., Dörken, B. (2001). In vitro susceptibility to TRAIL-induced apoptosis of acute leukaemia cells in the context of TRAIL receptor gene expression and constitutive NF- B activity. *Leukaemia*, 15(6), 921-928.
- Xiao, X., Dow, E. R., Eberhart, R., Miled, Z. B., & Oppelt, R. J. (2003, 22-26 April). *Gene clustering using self-organizing maps and particle swarm optimization*. Paper presented at the 17th International Symposium on Parallel and Distributed Processing , Nice, France.
- Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645-678.
- Xuwei, W., Ming, W., Zheng, L., & Chan, C. (2008). Short time-series microarray analysis: Methods and challenges. *BMC Systems Biology*, 2(58).
- Yeloglu, O., Heywood, A., & Malcolm, I. (2007, 7-10 Oct. 2007). *Growing recurrent self organizing map*. Paper presented at the IEEE International Conference on Systems, Man and Cybernetics, 2007. ISIC. , Montreal, Canada.
- Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukaemia by gene expression profiling. *Cancer Cell*, 1(2), 133-143.
- Yin, L., Huang, C. H., & Ni, J. (2007). Clustering of gene expression data: performance and similarity analysis. *BMC Bioinformatics*, 7(Suppl 4), 1-11.
- Youle, R. J., & Strasser, A. (2008). The BCL-2 protein family: opposing activities that mediate cell death. *Nature Reviews Molecular Cell Biology*, 9(1), 47-59.
- Zhang, L., Miles, M. F., & Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology*, 21, 818-821.
- Zhang, S., Jin, G., Zhang, X. S., & Chen, L. (2007). Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics*, 7(16), 2856-2869.
- Zhuang, W. J., Fong, C. C., Cao, J., Ao, L., Leung, C. H., Cheung, H. Y. (2004). Involvement of NF-[kappa] B and c-myc signaling pathways in the apoptosis of HL-60 cells induced by alkaloids of *Tripterygium hypoglaucom* (levl.) Hutch. *Phytomedicine*, 11(4), 295-302.
- Zou, M., & Conzen, S. D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1), 71-79.