

**SYSTEM-LEVEL MODELING OF ENDOTHELIAL
PERMEABILITY PATHWAY AND HIGH-THROUGHPUT
DATA ANALYSIS FOR DISEASE BIOMARKER
SELECTION**

WEI XIAONA

(M.Sc., Nankai University)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTATION AND SYSTEMS BIOLOGY (CSB)
SINGAPORE-MIT ALLIANCE
NATIONAL UNIVERSITY OF SINGAPORE**

2012

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A handwritten signature in black ink, appearing to read 'Wei Xiaona', written over a horizontal line.

WEI XIAONA

1st August 2012

ACKNOWLEDGEMENTS

First and foremost, my heartfelt appreciation and thanks go to my supervisor and mentor, Professor Chen Yu Zong, for his excellent supervision, invaluable advices and constructive suggestions throughout my whole research progress. I have tremendously benefited from his profound knowledge, expertise in scientific research, as well as his enormous support, which will inspire and motivate me to go further in my future professional career. My many thanks also go to my co-supervisor Professor Bruce Tidor and Associate Professor Low Boon Chuan. Thank you for their good suggestion for my project and invaluable encouragement.

I would like to dedicate my thesis to my parents, my husband, and my lovely son. The beautiful time and memories we have in Singapore are definitely great treasures in my life, I cherish it very much. And I am eternally grateful for everything you do for me, I appreciate it very much.

Special thanks go to our present and previous BIDD Group members. Without their help and group effort, this work could not be properly finished. I thank them for their valuable support and encouragement in my work.

Finally, I am very grateful to the Singapore-MIT Alliance, National University of Singapore for awarding me the Research Scholarship.

TABLE OF CONTENTS

DECLARATION.....	I
ACKNOWLEDGEMENTS.....	II
TABLE OF CONTENTS.....	III
SUMMARY.....	VIII
LIST OF ABBREVIATIONS.....	XV
Chapter 1 Introduction.....	1
1.1 Introduction to endothelial permeability and related disease.....	2
1.1.1 Overview of endothelial permeability.....	2
1.1.2 Molecular mechanism of endothelial permeability.....	3
1.1.3 Endothelia permeability related disease - Sepsis.....	8
1.2 Overview of mathematical modelling of signalling pathways.....	10
1.3 Introduction to high-throughput biomarker selection.....	13
1.3. 1 Introduction to microarray experiments.....	13
1.3.2 Statistical analysis of microarray data.....	15
1.3.3 Brief introduction to the Copy Number Variation.....	19
1.3.3 Overview of disease marker selection.....	24
1.4 Objective and outline of this thesis.....	29
Chapter 2 Methodology.....	32
2.1 Methods for mathematics model of signalling pathway.....	32
2.1.1 ODE for model development.....	32
2.1.2 Parameter estimation.....	36
2.1.3 Sensitivity analysis.....	41

2.2 Processing of microarray data.....	43
2.2.1 Missing data estimation	43
2.2.2 Normalization of microarray data.....	45
2.3 Processing Copy Number Variations	46
2.3.1 Overview of CNV calling calculation.....	46
2.3.2 HMM modelling strategy.....	47
2.3.3 Inference of log R Ratio (LRR) and B Allele Frequency (BAF)	48
2.4 Support Vector Machines	50
2.4.1 Theory and algorithm.....	50
2.4.2 Performance evaluation	58
2.5 Methodology for gene selection.....	59
2.5.1 Overview of the gene selection procedure.....	59
2.5.2 Recursive feature elimination	62
2.5.3 Sampling, feature elimination and consistency evaluation.....	63
Chapter 3 Mathematical Model of Thrombin-, Histamine-and VEGF-Mediated Signalling in Endothelial Permeability	66
3.1 Introduction.....	66
3.2 Thrombin-, Histamine-and VEGF-Mediated Signaling Cascades in endothelial permeability mediators.....	70
3.2.1 Thrombin mediated GPCR activation.....	70
3.2.2 Role of MAP Kinase in Cell Migration	73
3.2.3 VEGF mediated ERK activation.....	74
3.2.4 Thrombin, VEGF and Histamine mediated Ca ²⁺ release, PKC activation MLC activation	75
3.2.5 Thrombin, VEGF and Histamine mediated MLC activation....	76

3.3 Methods.....	77
3.3.1 Model Development.....	77
3.3.3 Model Optimization, Validation and Parameter Sensitivity Analysis.....	88
3.3.4 Estimation of kinetic parameters	90
3.4 Results and discussion	92
3.4.1 Model validation with experimental studies of the regulation of MLC activation, calcium release, and Rho activation by thrombin ..	92
3.4.2 Model validation with experimental studies of MLC activation and ERK activation by VEGF.....	98
3.4.3 Model validation with experimental studies of MLC activation by histamine	101
3.4.4 Comparison of the simulated thrombin-mediated IP3 and Ca ²⁺ release with that of an existing model.....	103
3.4.5 Simulation of the effects of thrombin receptor PAR-1 over-expression on thrombin-mediated MLC activation.....	105
3.4.6 Simulation of the effects of Rho GTPase and ROCK over-expression on thrombin-mediated MLC activation.....	106
3.4.7 Simulation of effects of VEGF and VEGFR2 over-expression on VEGF-mediated MLC activation.....	108
3.4.8 Simulation of synergistic activation of MLC by thrombin and histamine	110
3.4.9 Prediction of the collective regulation of MLC activation by thrombin and VEGF	118
3.4.10 Prediction of the effect of CPI-17 over-expression on MLC activation in the presence of lower concentration of thrombin, histamine and VEGF	122
3.5 Conclusion remarks	123
Chapter 4 Sepsis Biomarker selection	125
4.1 Introduction.....	125

4.2 Materials and methods	127
4.2.1 Sepsis microarray datasets	127
4.2.2 Gene selection procedure	129
4.2.3 Performance evaluation of signatures	130
4.3 Results and discussion	131
4.3.1 System of the disease marker selection	131
4.3.2 Consistency analysis of the identified disease markers	132
4.3.3 The function of the identified sepsis markers	144
4.3.4 The predictive performance of identified signatures in disease differentiation	146
Chapter 5 Breast cancer biomarker selection based on Copy number variation	149
5.1 Introduction	149
5.2 Materials and methods	152
5.2.1 Breast cancer and normal people CNV datasets	152
5.2.2 CNV calling calculation	153
5.2.3 CNV annotation	162
5.2.4 Breast cancer gene selection procedure	163
5.2.5 Performance evaluation of signatures	164
5.3 Results and discussion	165
5.3.1 CNV calls	165
5.3.2 Statistics of the selected predictor genes from Breast cancer dataset	166
5.3.3 The function of the identified breast cancer markers	167
5.3.4 Hierarchical clustering analysis of samples	170
Chapter 6 Concluding Remarks	193

TABLE OF CONTENTS

6.1 Finding and merits	193
6.2 Limitations and suggestions for future study.....	195
BIBLIOGRAPHY	198
List of Publication.....	232

SUMMARY

Understanding the behavior of biological systems is a challenging task. Computational models can assist us to understand biological systems by providing a framework within which their behavior can be explored. Constructing the models of these systems enables their behavior to be simulated, observed and quantified on a scale.

We constructed a model of endothelial permeability signaling pathway which is involved in injury, inflammation, diabetes and cancer. Detailed molecular interactions are specific and ordinary differential equations (ODEs) were used in our model to capture the time-dependent dynamic behavior of the concentration of proteins. All equations for molecular interactions in this study were derived based on laws of Mass Action. Our model was validated against a number of experimental findings and the observed synergistic effects of low concentrations of thrombin and histamine in mediating the activation of MLC. It can be used to predict the effects of altered pathway components, collective actions of multiple mediators and the potential impact to various diseases.

Another perspective for deciphering the mechanism of endothelial permeability and related disease is identifying the gene markers responsible for disease initiation. Current microarray data analysis tools provided good predictive performance. However, the signatures produced by those tools have

been found to be highly unstable with the variation of patient sample size and combination. To solve this problem, we developed a novel gene selection method based on Support Vector Machines, recursive feature elimination, multiple random sampling strategies and multi-step evaluation of gene-ranking consistency.

After program implementation, we first use microarray datasets to test. The dataset is endothelia permeability related disease - sepsis microarray. The expression levels of 18 control and 22 patient samples were used for sepsis marker discovery. 20 sets of sepsis gene signatures were generated. 41 gene signatures are fairly stable with 69%~93% of all predictor-genes shared by all 20 signatures sets. The predictive ability of the selected signature shared by all of the 20 sets is evaluated by SVM models on an independent dataset collected from GEO Database. Unsupervised hierarchical clustering analysis provides additional indication of the predictive ability of selected signatures.

Then the other type of high-throughput dataset used for signature selection system is breast cancer copy number variation based dataset. Total of 373 breast cancer samples and 517 normal people samples were used. We first calculated the breast cancer and normal people CNV calling by hidden Markov model. In this case, the derived 91 breast cancer signatures are found to be fairly stable with 80% of the top 50 ranked genes and 65% to 85% of all genes in each signature were shared by 20 signature sets.

LIST OF FIGURES

Figure 1-1: Signal transduction in endothelial permeability.....	4
Figure 1-2: GPCR activation and Ca ²⁺ release.	6
Figure 1-3: The Rho GTPase cycle.....	7
Figure 1-4: Procedure of microarray experiment.....	16
Figure 1-5 : The patterns of Copy-number variation (CNV).....	21
Figure 1-6: The procedure of comparative genomic hybridization (CGH)	22
Figure 1-7: Affymetrix Human Genome-Wide 6.0 SNP Arrays.....	23
Figure 1-8 : Filter method versus wrapper method for feature selection.....	26
Figure 2-1: Unidentifiable model parameters.	39
Figure 2-2: Affymetrix CNV calling overview	47
Figure 2-3: An illustration of log R Ratio (LRR) and B Allele Freq (BAF) values for the chromosome 15 q-arm of an individual.	50
Figure 2-4: Margins and hyperplanes	52
Figure 2-5 : Architecture of support vector machines	57
Figure 2-6: Overview of the gene selection procedure.....	61
Figure 3-1: The detailed pathway map of the thrombin-mediated signalling component of our integrated pathway simulation model. ROCK (f) and ROCK (o) refer to ROCK in folded and open conformation respectively.	71
Figure 3-2: The detailed pathway map of the VEGF-mediated signalling component of our integrated pathway simulation model.....	72
Figure 3-3: The detailed pathway map of the histamine-mediated signalling component of our integrated pathway simulation model.....	73
Figure 3-4: Framework of integrated pathway simulation model of thrombin-, histamine-, and VEGF-mediated MLC activation.	78
Figure 3-5: Fit to experimental data for Ras activation.	87

Figure 3-6: Parameter sensitivity analysis	90
Figure 3-7: Simulated time course and experimental data of thrombin-mediated MLC activation (left) and calcium release (right).	93
Figure 3-8: Simulated time course and experimental data of thrombin-mediated MLC activation in the first 20 min.	94
Figure 3-9: Simulated time course of thrombin-mediated MLC activation in terms of different components.	95
Figure 3-10: Simulated time course and experimental results of thrombin-mediated Rho GTPase activation in units of percentage of initial Rho concentration.	97
Figure 3-11: Simulated time course of thrombin-mediated MLC activation in terms of different components.	98
Figure 3-12: Simulated time course and experimental result of VEGF-mediated MLC activation (left) and ERK activation (right)	100
Figure 3-13: Simulated time course of VEGF-mediated MLC activation in terms of different components.	101
Figure 3-14: Simulated time course and experimental result of Histamine-mediated MLC activation in units of percentage of initial MLC concentration with thrombin and VEGF level set at zero values. The shaded area indicates the time range in which histamine has been experimentally found to induce a transient endothelial permeability. The histamine concentrations were taken as 0.005 μ M.	102
Figure 3-15: Simulated time course of Histamine-mediated MLC activation in terms of different components.	103
Figure 3-16: Comparison of simulation result of Ca ²⁺ and IP3 in our model and Maeda's model	104
Figure 3-17: ppMLC activation at different PAR-1 concentrations.....	106
Figure 3-18: MLC activation at different Rho GTPase (A) and ROCK (B) concentrations.	108
Figure 3-19: MLC activation at different VEGF(V) and VEGFR2 (VR) concentrations	110
Figure 3-20: MLC activation induced by combination of thrombin and histamine stimuli.....	111

Figure 3-21: The contribution of Ca²⁺- dependent, ROCK-dependent and CPI-17-dependent signaling cascade to thrombin-mediated MLC activation at low concentration of thrombin (0.0015 μM)..... 115

Figure 3-22: The contribution of Ca²⁺- dependent, NO-dependent and CPI-17-dependent signaling cascade to histamine-mediated MLC activation at low concentration of histamine (0.005 μM)..... 115

Figure 3-23: The contribution of Ca²⁺- dependent, NO-dependent and CPI-17-dependent cascade to thrombin + histamine mediated MLC activation at low concentration of thrombin (0.0015 μM) and histamine (0.005 μM)..... 116

Figure 3-24 : MLC activation induced by the combination of thrombin and VEGF stimuli..... 120

Figure 3-25: Prediction of the effect of CPI-17 over-expression on MLC activation at low concentration of stimuli..... 123

Figure 4-1: The system of sepsis genes derivation and sepsis differentiation 132

Figure 5-1: A flowchart outlining the procedure for CNV calling from genotyping data..... 156

Figure 5-2: Classes of genes involved in oncogenic transformation 169

Figure 5-3: Hierarchical clustering analysis of copy number enrichment patterns of 91 genes in breast cancer samples and normal samples. (Red for higher relative enrichment level, blue for lower relative enrichment level and white for medium enrichment level) 171

LIST OF TABLES

Table 2-1: Hidden states, copy numbers and their descriptions.....	48
Table 2-2 : List of some popular used support vector machines software	57
Table 2-3: Relationships among terms of performance evaluation.	59
Table 3-1: List of chemical reactions and related kinetic parameters in model.....	79
Table 3-2: Comparison of the areas with respect to different time ranges in	113
Table 3-3: Comparison of the areas with respect to different time ranges in Figure 9	121
Table 4-1 : List of sepsis biomarkers shared by all 20 groups, 15groups and 10 groups.....	134
Table 4-2: Statistics of the selected sepsis genes from sepsis microarray dataset by class-differentiation systems constructed from 20 different sampling-sets each composed of 500 training-testing sets generated by random sampling.....	143
Table 4-3: Overall accuracies of 500 training-test sets on the optimal SVM parameters	143
Table 4-4: Average sepsis prediction accuracy and standard deviation of 500 SVM class-differentiation systems constructed by 30 samples from GSE28750 dataset. The results were obtained from the overall accuracies of 500 test sets TP: True positive, FN: False negative, SE: Sensitivity.....	147
Table 5-1: Breast cancer and normal people CNV dataset used in biomarker selection.....	153
Table 5-2: Format of CNV calls.....	166
Table 5-3: Statistics of the selected predictor genes from Breast cancer dataset	168
Table 5-4: List of predictor genes of breast cancer data set shared by all 20 signatures	172
Table 5-5: Distribution of the selected predictor gene on chromosome (gene number >10).....	172

Table 5-6 : List of function of breast cancer signatures..... 173

LIST OF ABBREVIATIONS

MLC	Myosin Light Chain
MLCK	Myosin light chain kinase
MYPT	Myosin Light chain phosphatase
Arp2/3	Actin-related protein 2/3
PLC	Phospholipase C
PIP2	Phosphatidylinositol 4,5-bisphosphate
DAG	Diacyl glycerol
CPI-17	Protein kinase C-potentiated inhibitor protein of 17 kDa
SNP	Single-nucleotide polymorphism
PAR	Protease-activated receptor
cdc42	Cell division control protein 42 homolog
cAMP	cyclic adenosine monophosphate
DNA	deoxyribonucleic acid
EST	expressed sequence tag
ER	endoplasmic reticulum
FN	false negative
FP	false positive
KEGG	Kyoto encyclopedia of genes and genomes database
KNN	k-nearest neighbors
LS	least square method
ML	machine learning
NCBI	national center for biotechnology information

Q	overall accuracy
RFE	recursive feature elimination
RNA	ribonucleic acid
SE	sensitivity
SMO	sequential minimal optimization
SP	specificity
SNP	single nucleotide polymorphism
SBML	systems Biology Markup Language
STDEV	standard deviation
SVM	support vector machines
TN	true negative
TP	true positive
PDEs	partial differential equations
ODEs	ordinal differential equations
SDEs	stochastic differential equations

Chapter 1 Introduction

Endothelial permeability is involved in injury, inflammation, diabetes and cancer. Computational models can assist us to understand the mechanism by providing a framework within which their behavior can be explored. Besides, computational model can be used to predict the effects of altered pathway components, collective actions of multiple mediators and the potential impact to various diseases. Computational model also can potentially be used to identify important disease genes through sensitivity analysis of signaling components. Another perspective for deciphering the mechanism of endothelial permeability and related disease is identifying the gene markers. Thanks to the rapid progress on the research of genomics and genetics, more and more high-throughput data is available. The first section (Section 1.1) of this chapter gives an overview of endothelial permeability and related disease. The second section introduces mathematical modeling of signaling pathways (Section 1.2). The following sections of this chapter introduce the disease biomarker selection using high throughput data, includes microarray and copy number variation datasets (Section 1.3). The motivation of this work and outline of the structure of this document are presented in Section 1.4.

1.1 Introduction to endothelial permeability and related disease

1.1.1 Overview of endothelial permeability

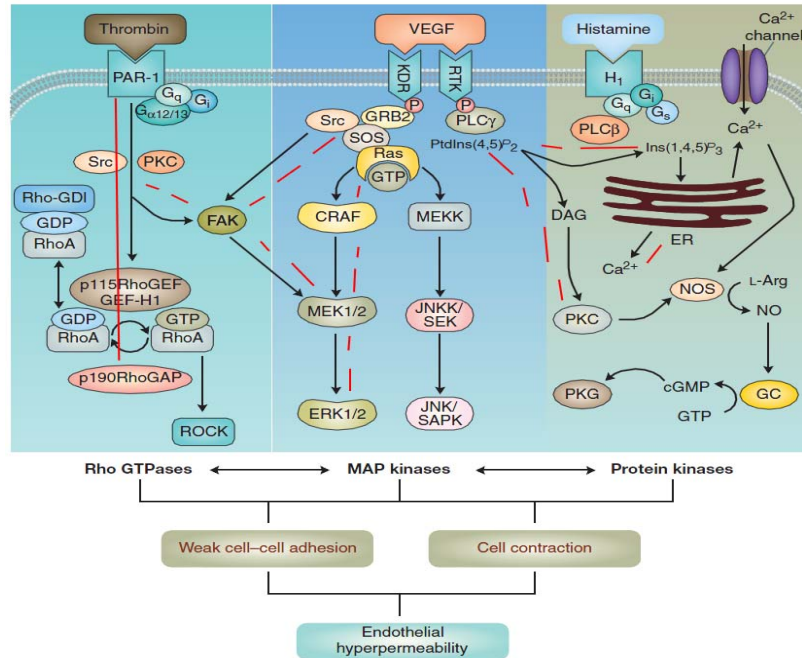
Endothelial permeability is a significant problem in vascular inflammation associated with trauma, ischaemia–reperfusion injury, sepsis, adult respiratory distress syndrome, diabetes, thrombosis and cancer [1]. The mechanism underlying this process is increased paracellular leakage of plasma fluid and protein [2]. Inflammatory stimuli such as histamine, thrombin, vascular endothelial growth factor (VEGF) and activated neutrophils can cause dissociation of cell–cell junctions between endothelial cells as well as cytoskeleton contraction, leading to a widened intercellular space that facilitates transendothelial flux [3, 4]. Such structural changes initiate with agonist-receptor binding, followed by activation of intracellular signalling molecules including calcium, protein kinase C (PKC), tyrosine kinases, myosin light chain kinase (MLCK), and small Rho-GTPases; these kinases and GTPases then phosphorylate or alter the conformation of different subcellular components that control cell–cell adhesion, resulting in endothelial hypermeability [5]. Targeting key signaling molecules that mediate endothelial junction - cytoskeleton dissociation demonstrates a therapeutic potential to improve vascular barrier function during inflammatory injury [1].

1.1.2 Molecular mechanism of endothelial permeability

Endothelial cells lining the inner surface of microvessels form a semipermeable barrier that actively participates in blood–tissue exchange of plasma fluid, proteins and cells [6] [7]. The maintenance by the endothelium of a semi-permeable barrier is particularly important in controlling the passage of macromolecules and fluid between the blood and interstitial space [7, 8] .

Many inflammatory mediators are capable of disrupting the interendothelial junction assembly, thereby causing endothelial permeability [9-12]. More in-depth molecular analyses suggest that the mechanism underlying inflammation-induced endothelial hyperpermeability involves phosphorylation, internalisation or degradation of the junctional molecules [13, 14] [15]. In addition, the junction - cytoskeleton complex participates in other cellular processes including molecular scaffolding, intracellular trafficking, transcription and apoptosis that may directly or indirectly alter vascular barrier function [16][17].

Regardless of the molecular details, however, essentially all permeability responses in the vascular endothelium are initiated with receptor occupancy followed by a series of intracellular signalling cascades (**Figure 1-1**) [1], some of which are described below.

Figure 1-1: Signal transduction in endothelial permeability

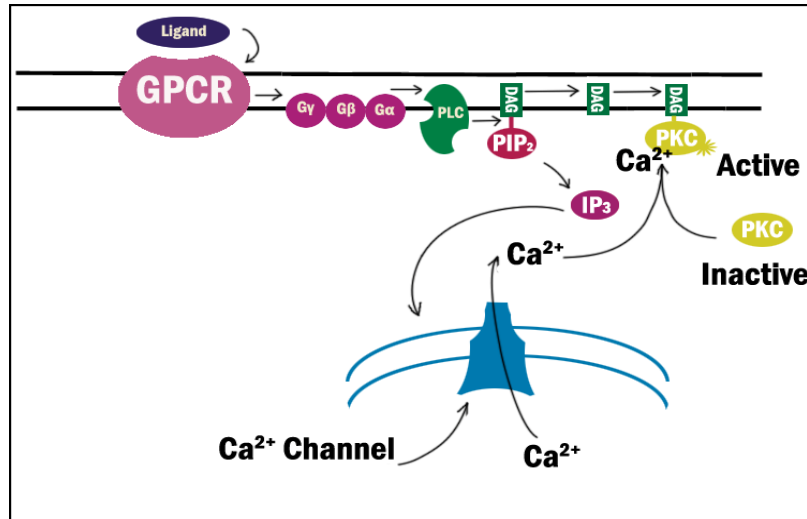
1.1.2.1 Ca²⁺ release

In endothelial cells, binding to GPCRs by agonists causes G α _q to switch from a GDP-bound to a GTP-bound state, allowing the release of G α _q from the G β γ dimer. The GTP bound G α _q subunit subsequently activates phosphoinositide phospholipase PLC- β , which then hydrolyses the lipid precursor phosphatidylinositol-4, 5-bisphosphate (PIP₂) to yield IP₃ and diacylglycerol [18-21]. IP₃ receptors constitute the most clearly identified Ca²⁺ channels that pump Ca²⁺ from the ER [22-27]. Most cells have at least one form of IP₃ receptor, and many express all three. Structurally, the IP₃ receptor channels are tetramers composed of four subunits, IP₃-mediated Ca²⁺ release responses are co-operative, indicating that several and perhaps all subunits are required to

bind IP3 for the channel to open [28] (**Figure1-2**). A characteristic feature of IP3 receptors is that they are regulated by both IP3 and Ca^{2+} .

1.1.2.2 PKC activation

PKC activation occurs when plasma membrane receptors coupled to phospholipase C, releasing diacylglycerol. The conventional isoforms, α , βI , βII , and γ , are activated by phosphatidylserine, diacylglycerol and Ca^{2+} [29-33]. The unconventional isoforms, δ , ϵ , η , and θ , require phosphatidylserine and diacylglycerol but do not require Ca^{2+} . The ζ and λ isoforms are called atypical and require only phosphatidylserine for activation. The G-protein activates phospholipase C (PLC), which cleaves phosphoinositol-4, 5-bisphosphate (PIP2) into 1, 2-diacylglycerol and inositol-1, 4, 5-trisphosphate (IP3). The IP3 interacts with a calcium channel in the endoplasmic reticulum (ER), releasing Ca^{2+} into the cytoplasm. The increase in Ca^{2+} levels activates PKC [34, 35], which translocates to the membrane, anchoring to diacylglycerol (DAG) and phosphatidylserine.

Figure 1-2: GPCR activation and Ca²⁺ release.

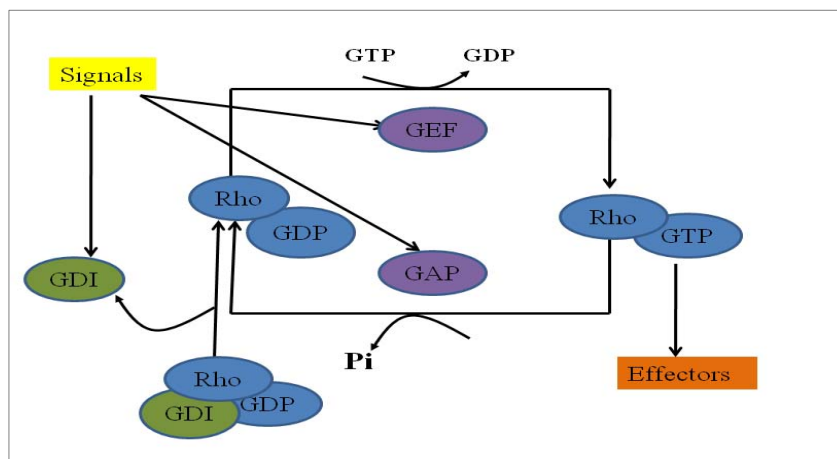
External stimulus activates a G-Protein-Coupled Receptor (GPCR), which activates a stimulating G-protein. The G-protein activates phospholipase C (PLC), which cleaves phosphoinositol-4, 5-bisphosphate (PIP₂) into 1, 2-diacylglycerol and inositol-1, 4, 5-trisphosphate (IP₃). The IP₃ interacts with a calcium channel in the endoplasmic reticulum (ER), releasing Ca²⁺ into the cytoplasm. The increase in Ca²⁺ levels activates PKC, which translocates to the membrane, anchoring to diacylglycerol (DAG) and phosphatidylserine [36] (From Promega signal transduction, Source: Signal Transduction Resource, www.promega.com).

1.1.2.3 Rho GTPase activation

The Rho GTPase cycle is tightly regulated by three groups of proteins. Guanine nucleotide exchange factors (GEFs) promote the exchange of GDP for GTP to activate the GTPase, GTPase-activating proteins (GAPs) negatively regulate

the switch by enhancing its intrinsic GTPase activity and guanine nucleotide dissociation inhibitors (GDIs) are thought to block the GTPase cycle by sequestering and solubilizing the GDP-bound form [37]. Extracellular signals could regulate the switch by modifying any of these proteins, but so far at least, they appear to act predominantly through GEFs. Once activated, Rho GTPases interact with cellular target proteins (effectors) to generate a downstream response (**Figure 1-3**) [38].

Figure 1-3: The Rho GTPase cycle.



Rho GTPases cycle between an inactive GDP-bound form and an active GTP-bound form. The cycle is tightly regulated mainly by guanine exchange factors (GEFs), GTPase activating proteins (GAPs) and guanine dissociation inhibitors (GDIs) [39-44]. In their active form, Rho GTPases can bind to effector molecules such as kinases and scaffold proteins [44-49].

1.1.2.4 NO activation

Cytosolic Ca^{2+} elevation is a typical initial response of endothelial cells to hormonal and chemical signal and to changes in physical parameters, and many endothelial functions are dependent on changes in Ca^{2+} concentration [37]. For instance, the activity of endothelial nitric oxide synthase (eNOS) in producing nitric oxide in endothelial cells absolutely requires CaM [50] and it appears to also require Ca^{2+} to sustain elevated level of activity [37].

Nitric oxide plays a critical role in the endothelial cell proliferation, migration, and tube formation, as well as increased vascular permeability, hypotension, and angiogenesis in vivo [37]. VEGF- and histamine - induced microvascular hyperpermeability are both mediated by a signalling cascade triggered by receptor binding and transduced by a serial activation of intracellular enzymes, including PLC, eNOS, soluble guanylate cyclase (sGC), and protein kinase G (PKG). Subsequently, the VEGF-activated NO-PKG pathway was linked to ERK1/2-mediated proliferation of cultured endothelial cells via phosphorylation and activation of the upstream p42/44 MAPK cascade component RAF by PKG [37].

1.1.3 Endothelia permeability related disease - Sepsis

The precise regulation of endothelial permeability is essential for maintaining circulatory homeostasis and the physiological function of different organs. As a result, microvascular barrier dysfunction and endothelial permeability represent

crucial events in the development of a variety of disease processes, such as adult respiratory distress syndrome (ARDS), ischemia–reperfusion (I–R) injury, diabetic vascular complications, and tumor metastasis. Better insight into the molecular mechanisms underlying pathogenic conditions related to microvascular permeability is required for developing effective therapeutic strategies [51-66].

Sepsis is one of the major causes of mortality in critically ill patients and develops as a result of the host response to infection. The endothelium is a major target of sepsis-induced events and endothelial cell damage accounts for much of the pathology of septic shock [67]. Vascular endothelial cells are among the first cells in the body that come into contact with circulating bacterial molecules. Endothelial cells possess mechanisms that recognize structural patterns of bacterial pathogens and subsequently initiate the expression of inflammatory mediators [68-72].

The cellular response to bacterial toxins normally provides protection against microorganism - induced infection critical injury. Under normal conditions, the biological activity of sepsis-involved mediators is under the stringent control of specific inhibitors. In sepsis this balance is disrupted and the disturbance is manifested by profound changes in the relative production of different mediators. Therefore the pathogenesis of sepsis can be described as a pro- and anti-inflammatory disequilibrium syndrome [73].

If a person has sepsis, they often will have fever. Sometimes, though, the body temperature may be normal or even low. Sepsis symptoms and signs are as followings: The individual may also have chills and severe shaking; The heart may be beating very fast, and breathing may be rapid; Low blood pressure is often observed in septic patients; Confusion, disorientation, and agitation may be seen as well as dizziness and decreased urination; Some patients who have sepsis develop a rash on their skin; The rash may be a reddish discoloration or small dark red dots seen throughout the body; Those with sepsis may also develop pain in the joints of the wrists, elbows, back, hips, knees, and ankles.

The prognosis of sepsis depends on age, previous health history, overall health status, how quickly the diagnosis is made, and the type of organism causing the sepsis. For elderly people with many illnesses or for those whose immune system is not working well because of illness or certain medications and sepsis is advanced, the death rate may be as high as 80%. On the other hand, for healthy people with no prior illness, the death rate may be low, at around 5%. The overall death rate from sepsis is approximately 40%. It is important to remember that the prognosis also depends on any delay in diagnosis and treatment. The earlier the treatment is started, the better the outcome will be.

1.2 Overview of mathematical modelling of signalling pathways

Biological pathways are the most common pathways which include metabolic

pathways and signaling networks of the cell. The metabolic pathways constitute the enzymatic reactions where a certain product is formed from a combination of substrates under particular kinetic parameters and specific concentration of the substrate(s) [74-81]. Signaling networks comprises of the cellular processes under different intracellular conditions and in responses to various external stimuli. These pathways are studied in greater detail for disease related process such as cancer, diabetes, etc [82-88]. The pathways are known in detail because of the knowledge obtained from the wide number of interactions between various components of the pathway.

The different interacting proteins trigger the cellular process such as that of signal transduction where the signals get transduce from extracellular surface to the nucleus to activate gene transcription. Specific cells carry out the signaling process based on tissue and cell specific gene expression. Hence it is difficult to quantify the biological pathways in terms of their biological function in the cell. Although much has been known about biological pathways and databases have been constructed to store the pathway information, there is always a gap to be filled to gain more knowledge about the already known pathways in highly detailed manner and expanding their horizons. Using systems biology approach scientists have tried to reconstruct pathways using pathway models from the information as known from the existing pathways [17]. Reconstruction of pathways are carried out using the well known pathways, the different components and the interacting partners, the kinetic parameters, the inhibitors and the activating factors. Most of the information is obtained either from the

pathway database such as KEGG [67] and literature. Pathways maps can be described in mathematical terms [89-95]. By describing these pathways in mathematical models, it is then possible to perform computer simulations of the changes in the responses to changing input. This procedure of predicting biological responses through mathematical modeling and simulation is known as pathway simulation. Pathway simulation is therefore a quantitative prediction of complex biological pathways [96-101]. Pathway simulation will allow us to predict or explain complex biological process outcomes that cannot be easily foresee or explained with fundamental principles [102-110]. For example, Li *et al* [111] described a model of ERK activity with a crosstalk between MEKK1-mediated and EGFR-ERK pathways. The simulation of the ERK activity under various conditions, such as differing RhoA and Ras levels, displayed ERK activity that are not directly observed. Subsequently new hypothesis about the potential drug targets can be generated [112-119].

Unknown information such as kinetic parameters are obtained using manual estimation and prediction using similar proteins such as sequence similarity to proteins which share homology with the proteins under study. Parameter estimation is significant because it determines how the pathway acts in terms of the substrate concentration and the product formation from its respective substrates [120-122]. Reconstruction of pathways is followed by pathway

simulation wherein the different kinetic parameters along with the difference components of the pathway are input into the simulation software which helps

us to understand better about the pathway components and gives us ideas about the behavior of the various interactions involved in the pathway. Pathway simulation has been an important topic in Systems Biology [90, 123-129]. It gives us an overall idea of how the pathways act inside the cell in a quantitative manner and this is facilitated by the kinetic parameters used for each reaction of the pathway reconstructed.

The complexity of the pathway interactions makes it a hard task to understand the behavior of cellular networks. Also as the *in vivo* experiments are time consuming process with minimum time are desirable [130]. Mathematical modeling and computer simulation techniques have played an important role in understanding the topology and dynamics of such complex networks. Pathway simulations have an edge over conventional experimental biology in terms of cost, ease and speed.

A pathway simulation can be defined mathematically by differential equations defining the law of mass action or Michaelis-Menton Kinetics with formats like systems biology mark-up language (SBML), a standard for representing models of biochemical and gene-regulatory networks [131, 132].

1.3 Introduction to high-throughput biomarker selection

1.3. 1 Introduction to microarray experiments

Microarray technology, also known as DNA chip, gene ship or biochip, is one

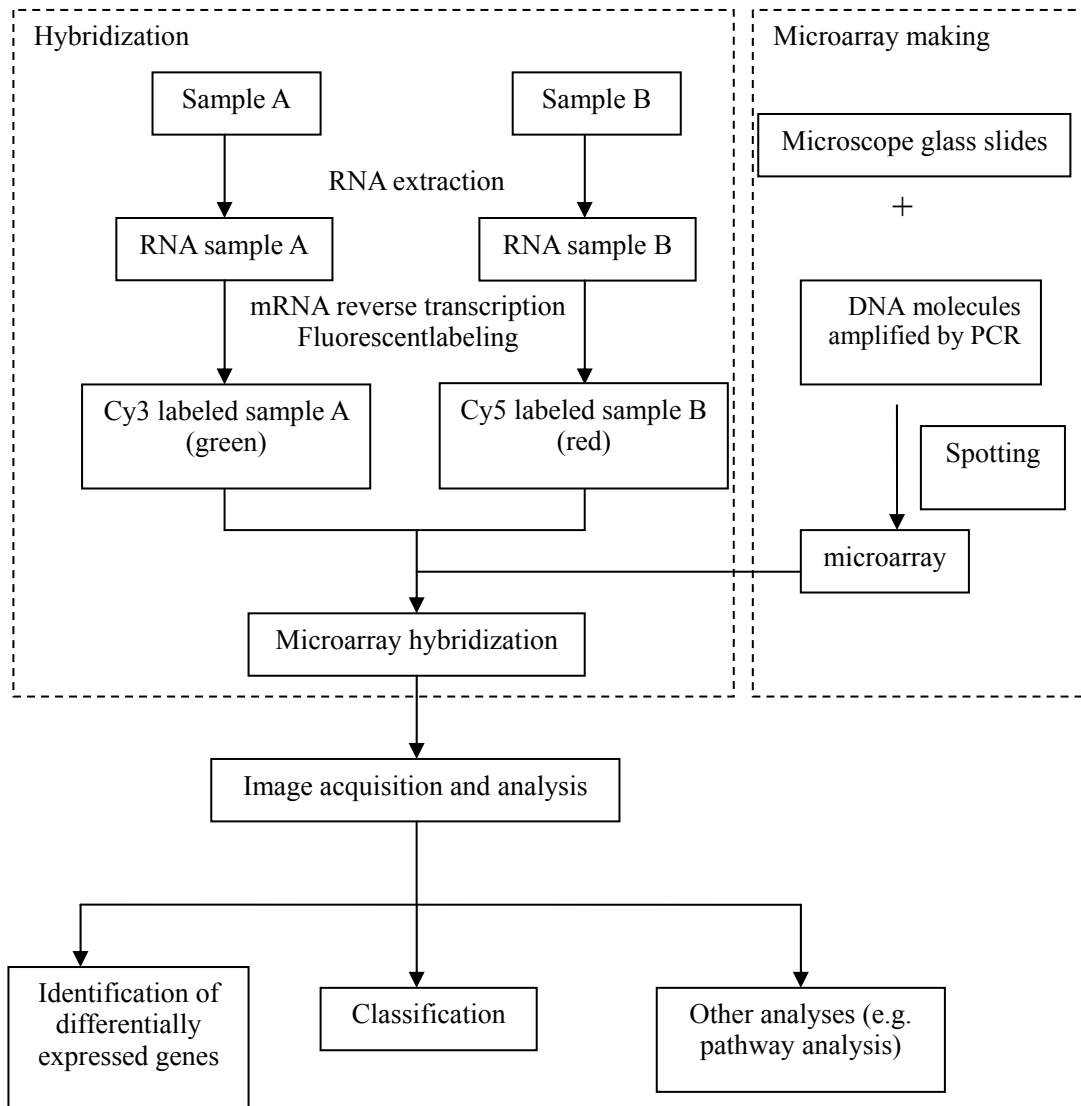
of the indispensable tools in monitoring genome wide expression levels of genes in a given organism [133, 134]. Microarrays measure gene expression in many ways, one of which is to compare expression of a set of genes from cells maintained in a particular condition A (such as disease status) with the same set of genes from reference cells maintained under conditions B (such as normal status).

Figure 1-4 shows a typical procedure of microarray experiments [135, 136]. A microarray is a glass substrate surface on which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots, and each spot may contain a few million copies of identical DNA molecules (probes) that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA [137], or synthesized oligo-nucleotide strands that correspond to a gene [138-140]. This microarray can be made by the experimenters themselves (such as cDNA array) or purchased from some suppliers (such as Affymetrix GeneChip). The actual microarray experiment starts from the RNA extraction from cells. These RNA molecules are reverse transcribed into cDNA, labeled with fluorescent reporter molecules, and hybridized to the probes formatted on the microarray slides. At this step, any cDNA sequence in the sample will hybridize to specific spots on the glass slide containing its complementary sequence. The amount of cDNA bound to a spot will be directly proportional to the initial number of RNA molecules present for that gene in both samples. Following, an instrument is used to read the reporter molecules and create microarray image. In this image,

each spot, which corresponds to a gene, has an associated fluorescence value, representing the relative expression level of that gene. Then the obtained image is processed, transformed and normalized. And the analysis, such as differentially expressed gene identification, classification of disease/normal status, and pathway analysis, can be conducted.

1.3.2 Statistical analysis of microarray data

Since microarray contains the expression level of several thousands of genes, it requires sophisticated statistical analysis to extract useful information such as gene selection. Theoretically, one would compare a group of samples of different conditions and identify good candidate genes by analysis of the gene expression pattern. However, microarray data contain some noises arising from measurement variability and biological differences [73, 141]. The gene-gene interaction also affects the gene-expression level. Furthermore, the high dimensional microarray data can lead to some mathematical problems such as the curse of dimensionality and singularity problems in matrix computations, causing data analysis difficult. Therefore choosing a suitable statistical method for gene selection is very important.

Figure 1-4: Procedure of microarray experiment

The statistical methods in microarray data analysis can be classified into two groups: unsupervised learning methods and supervised learning methods. Unsupervised analysis of microarray data aims to group relative genes without knowledge of the clinical features of each sample [142]. A commonly-used unsupervised method is hierarchical clustering method. This method groups

genes together on the basis of shared expression similarity across different conditions, under the assumption that genes are likely to share the same function if they exhibit similar expression profiles [143-146]. Hierarchical clustering creates phylogenetics trees to reflect higher-order relationship between genes with similar expression patterns by either merging smaller clusters into larger ones, or by splitting larger clusters into smaller ones. A dendrogram is constructed, in which the branch lengths among genes also reflect the degree of similarity of expression [147, 148]. By cutting the dendrogram at a desired level, a clustering of the data items into the disjoint groups can be obtained. Hierarchical clustering of gene expression profiles in rheumatoid synovium identified 121 genes associated with Rheumatoid arthritis I and 39 genes associated with Rheumatoid arthritis II [149]. Unsupervised methods have some merits such as good implementations available online and the possibility of obtaining biological meaningful results, but they also possess some limitations. First, unsupervised methods require no prior knowledge and are based on the understanding of the whole data set, making the clusters difficult to be maintained and analyzed. Second, genes are grouped based on the similarity which can be affected by input data with poor similarity measures. Third, some of the unsupervised methods require the predefinition of one or more user-defined parameters that are hard to be estimated (e.g. the number of clusters). Changing these parameters often have a strong impact on the final results [150].

In contrast to the unsupervised methods, supervised methods require a priori

knowledge of the samples. Supervised methods generate a signature which contains genes associated with the clinical response variable. The number of significant genes is determined by the choice of significance level. Support vector machines (SVM) [151] and artificial neural networks (ANN) [152] are two important supervised methods. Both methods can be trained to recognize and characterize complex pattern by adjusting the parameters of the models fitting the data by a process of error (for example, mis-classification) minimization through learning from experience (using training samples). SVM separates one class from the other in a set of binary training data with the hyperplane that is maximally distant from the training examples. This method has been used to rank the genes according to their contribution to defining the decision hyperplane, which is according to their importance in classifying the samples. Ramaswamy *et al.* used this method to identify genes related to multiple common adult malignancies [153]. ANN consists of a set of layers of perceptrons to model the structure and behavior of neurons in the human brain. ANN ranks the genes according to how sensitive the output is with respect to each gene's expression level. Khan *et al* identified genes expressed in rhabdomyosarcoma from such strategy [154].

In classification of microarray datasets, it has been found that supervised machine learning methods generally yield better results [155], particularly for smaller sample sizes [73]. In particular, SVM consistently shows outstanding performance, is less penalized by sample redundancy, and has lower risk for over-fitting [156, 157]. Furthermore, some studies demonstrated that

SVM-based prediction system was consistently superior to other supervised learning methods in microarray data analysis [158-160]. SVM for microarray data analysis are used in this study.

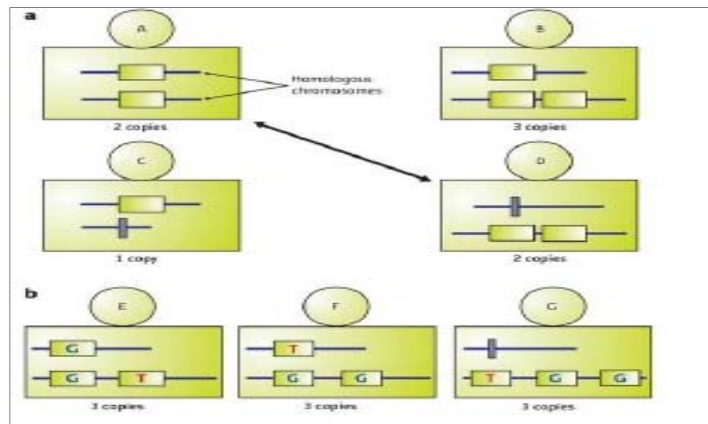
1.3.3 Brief introduction to the Copy Number Variation

1.3.3.1 Copy Number Variation

Human populations show extensive polymorphism — both additions and deletions — in the number of copies of chromosomal segments, and the number of genes in those segments[161]. This is known as copy number variation (CNV). A high proportion of the genome, currently estimated at up to 12%, is subject to copy number variation [162]. Copy number variants (CNVs) can arise both meiotically and somatically, as shown by the finding that identical twins can have different CNVs and that repeated sequences in different organs and tissues from the same individual can vary in copy number [163]. Copy number variation seems to be at least as important as SNPs in determining the differences between individual humans [164] and seems to be a major driving force in evolution, especially in the rapid evolution that has occurred, and continues to occur, within the human and great ape lineage [165]. Changes in copy number might change the expression levels of genes included in the regions of variable copy number, allowing transcription levels to be higher or lower than those that can be achieved by control of transcription of a single copy per haploid genome. The patterns of CNV are in **Figure 1-5.**

Additional copies of genes also provide redundancy that allows some copies to evolve new or modified functions or expression patterns while other copies maintain the original function. The nonhomologous recombination events that underlie changes in copy number also allow generation of new combination of exons between different genes by translocation, insertion or deletion [166], so that proteins might acquire new domains, and hence new or modified activities.

However, much of the variation in copy number is disadvantageous. Change in copy number is involved in cancer formation and progression [166, 167], and contributes to cancer proneness. In many situations, a change in copy number of any one of many specific genes is not well tolerated, and leads to a group of pathological conditions known as genomic disorders. Because particular gene imbalances are associated with specific clinical syndromes, data on rare clinical cases of change in copy number are available and have facilitated the study of the chromosomal changes underlying copy number variation [168-173].

Figure 1-5 : The patterns of Copy-number variation (CNV)

(a) Individuals in a population may have different copy numbers on homologous chromosomes at CNV loci. (b) Individuals may also have CNVs that contain SNPs.

1.3.3.2 Copy number analysis techniques

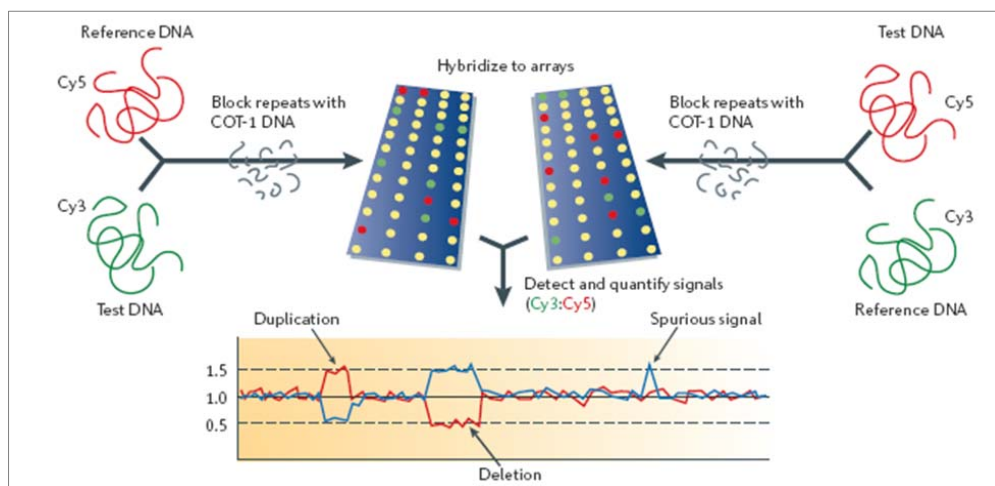
The study of chromosomal copy number analysis techniques is important in biology primarily because presence of copy number aberrations is known to be associated with the development of cancerous tumors.

1.3.3.2.1 Comparative genomic hybridization

Traditionally, the method of comparative genomic hybridization (CGH) [137, 174] has been used to identify chromosomal copy number aberrations (**Figure 1-6**). In CGH, cancerous test chromosomes and normal reference chromosomes are each chemically labeled with different colors, and then hybridized to a

genome of metaphase chromosomes. By quantifying the relative fluorescence intensity, the copy number can be deduced. However, the known disadvantage of using this cytogenetic technique for copy number analysis is its limited resolution: usually about 10Mb, and 2Mb at best.

Figure 1-6: The procedure of comparative genomic hybridization (CGH)

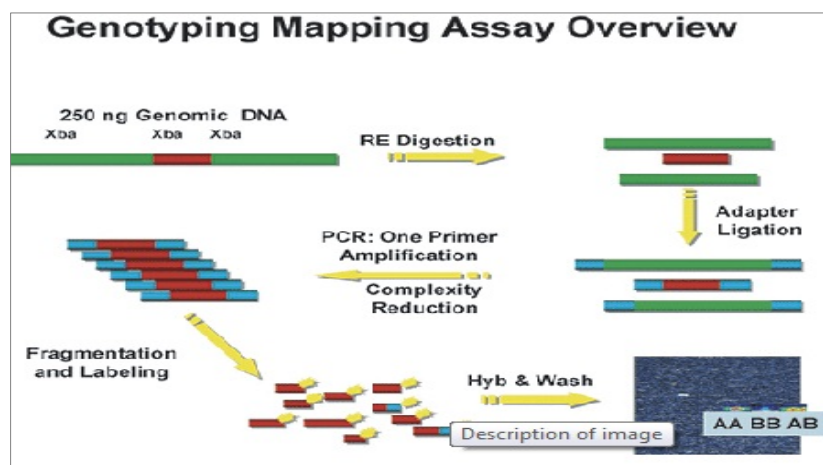


1.3.2.2.2 Copy number analysis with SNP microarrays

Despite developments in CGH microarray technology and methodology, the use of SNP microarrays for determining chromosomal copy number is of interest for three principal reasons (**Figure 1-7**). First, since SNP microarrays are already commonly used for SNP genotyping, an effective method of copy number analysis for SNP microarrays would enable the microarray assay to elucidate chromosomal copy number in addition to SNP genotypes. Second, the copy number call resolution in the genome is potentially much greater for SNP microarrays than for CGH microarrays since SNP microarrays have so many

probes. Third, since SNP microarrays are fundamentally similar to CGH microarrays, existing copy number analysis methods for CGH microarrays can be adapted for use with SNP microarrays. Thus, SNP microarrays have the potential to be useful tools for copy number analysis.

Figure 1-7: Affymetrix Human Genome-Wide 6.0 SNP Arrays.



These arrays contain over 900,000 SNP and over 900,000 Copy Number Variation (CNV) probes to allow researchers to conduct whole genome scans on a single array. Probes were chosen from restriction digestion fragments that had been size selected as shown in order to reduce target complexity before labeling and hybridization. As a result, the probe distribution should not be expected to be completely uniform across the entire genome. This is currently the highest density genotyping array commercially available.

1.3.3 Overview of disease marker selection

1.3.3.1 Overview of Feature selection

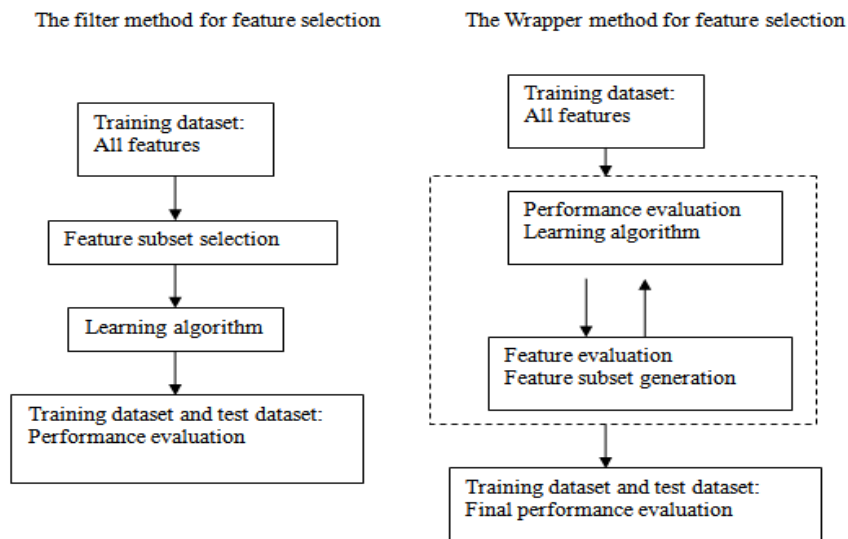
No matter whether the supervised or unsupervised methods are used, one critical problem encountered in both methods is feature selection, which has become a crucial challenge of microarray data analysis. The challenge comes from the presence of thousands of genes and only a few dozens of samples in currently available data. From the mathematical view, thousands of genes are thousands of dimensions. Such a large number of dimensions leads microarray data analysis to problems such as the curse of dimensionality [175, 176] and singularity problems in matrix computations. Therefore, there is a need of robust techniques capable of selecting the subsets of genes relevant to a particular problem from the entire set of microarray data both for the disease classification and for the disease target discovery.

Gene selection from microarray data is to search through the space of gene subsets in order to identify the optimal or near-optimal one with respect to the performance measure of the classifier. Many gene selection methods have been developed, and generally fall into two categories: filter method and wrapper method [177]. **Figure 1-8** shows how these two methods work.

In brief, the filter method selects genes independent of the learning algorithms [178-180]. It evaluates the goodness of the genes from simple statistics computed from the empirical distribution with the class label [181]. Filter

method has some pre-defined criteria. Mutual information and statistical testing (e.g. T-test and F-test) are two typical examples of filter method [178, 182-187]. Filter method can be easily understood and implemented, and needs little computational time. But the pitfall of this method is that it is based on the assumption that genes are not connected to each other, which is not true in real biological process.

Wrapper method generates genes from the evaluation of a learning algorithm. It is conducted in the space of genes, evaluating the goodness of each gene or gene subsets by such criteria as cross-validation error rate or accuracy from the validation dataset [188]. The wrapper method is very popular among machine learning methods for gene discovery [177, 189, 190]. Although the wrapper method needs extensive computational resources and time, it considers the gene-gene interaction and its accuracy is normally higher than the filter method [177, 189, 190]. Recursive feature elimination (RFE) is a good example of the wrapper method for disease gene discovery. The RFE method uses the prediction accuracy from SVM to determine the goodness of a selected subset. This thesis will employ RFE for disease gene discovery from microarray data.

Figure 1-8 : Filter method versus wrapper method for feature selection

1.3.3.2 The problems of current marker selection methods

The methodology of SVM and RFE will be discussed in Chapter 2 in details. Here, some problems encountered in current marker discovery from microarray data analysis are discussed. One problem is to specify the number of genes for differentiating disease. The number of derived breast cancer genes and leukemia genes ranges from 1 to 200 [183, 191-196]. 50 genes were arbitrarily chosen for differentiating acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) by Golub *et al*, since they supposed that 50 genes might reflect the difference between AML and ALL [183]. In most cases, the gene number was decided by the classification performance of different gene combinations. The gene combination which produced the highest classification accuracy constituted the gene signature. This strategy might

produce small sets of genes (one or two genes) that formed accurate classifier [194-196]. For example, Slonim *et al* reported that the classifier consisting of one gene (HOXA9) outperformed all of other classifiers consisting of other gene combinations for recurrence prediction in AML patients [196]. Li and Yang showed that one gene (Zyxin) constituted the best classifiers for AML/ALL differentiation [194]. Nevertheless these results were only obtained and tested on one dataset. Considering that the number of genes should correlate with the disease situation, the selected genes should be large enough to be robust against noise and small enough to be readily applied in clinical settings. Therefore, it is not appropriate to use the arbitrary gene number. Similarly, to use just one dataset to decide the optimal gene number may not be satisfactory, because the optimal gene number varies with the different sample sizes and sample combinations [141, 197, 198].

Another problem in gene discovery is the gene signatures were highly unstable and strongly depended on the selection of patients in the training sets [73, 141, 154, 183, 199-202] [141, 197, 198], despite the use of sophisticated class differentiation and gene selection methods by various groups. The unstable signatures were observed in most microarray datasets including breast cancer, lung adenocarcinoma, non-Hodgkin lymphoma, acute lymphocytic leukemia, acute myeloid leukemia, breast cancer, medulloblastoma, and hepatocellular carcinoma [141, 147, 158, 177, 180, 199, 203-206]. While these signatures display high predictive accuracies, the highly unstable and patient-dependent nature of these signatures diminishes their application potential for diagnosis

and prognosis [141]. Moreover, the complex and heterogenic nature of disease such as cancer may not be adequately described by the few cancer-related genes in some of these signatures. The unstable nature of these signatures and their lack of disease-relevant genes also limit their potential for target discovery. The instability of derived signatures is likely caused by the noises in the microarray data arising from such factors as the precision of measured absolute expression levels, capability for detecting low abundance genes, quality of design and probes, annotation accuracy and coverage, and biological differences of expression profiles [73, 207]. Apart from enhancing the quality of measurement and annotation, strategies for improving signature selection have also been proposed. These strategies include the use of multiple random validation [141], large sample size [208], known mechanisms [209], and robust signature-selection methods which is insensitive to noises [73, 210, 211].

This thesis will explore a new signature selection method aiming at reducing the chances of erroneous elimination of predictor-genes due to the noises contained in microarray dataset. Multiple random sampling and gene-ranking consistency evaluation procedures will be incorporated into RFE signature selection method. The consistent genes obtained from the multiple random sampling method may give us a better understanding to the disease initiation and progress, and may provide potential disease targets.

1.4 Objective and outline of this thesis

The ultimate goal of this thesis is to get the molecular mechanism of endothelial permeability and related disease using computational method. In order to meet this, this thesis has been divided into three sections, each of which deals with one sub-objective.

The first objective is to develop a mathematical model of endothelial permeability. Thrombin, VEGF, and histamine are hallmarks of endothelial hyper-permeability, which perform their regulatory roles individually and collectively under different disease conditions, and with different dynamic profiles. Thrombin and VEGF can increase microvascular permeability ~50,000 times more potently than histamine [212]. Thrombin, VEGF, and histamine induce prolonged (1-1.5 hr), intermediate (15-20 min) and transient (~5 min) increases of endothelial permeability, respectively. Using the model, we can interpret temporal effects and the dynamics of multi-mediator regulation.

The second objective is to design bioinformatics tools for endothelial permeability disease marker discovery using high-throughput dataset. A disease marker discovery system is developed by using gene selection strategies from microarray data. This system aims to provide gene signatures which should produce good prediction performance for disease differentiation, and show a certain level of stability with the variation of sampling method.

The strategies include the incorporation of multiple random sampling methods and the evaluation of gene-consistency into RFE gene selection procedure. The stable gene signatures may help us understand the mechanism of disease initiation and process, and may provide an insight for diagnosing disease, predicting disease types, prognosis of the outcome of a specific therapeutic strategy, and drug resistance before drug treatment.

The complete outline of this thesis is as follows:

In **chapter 1** an overview of endothelial permeability, related disease and molecular mechanism is described and an introduction to mathematics model of signaling pathway. Then we have give background to microarray, copy number variation and disease biomarker selection.

Chapter 2 methods used in this work are described. In particular, methods for pathway simulation, Parameter estimation, Sensitivity analysis, Processing of microarray data and copy number variation calling calculation, Support Vector Machines, Performance evaluation, Recursive feature elimination, Sampling, feature elimination and consistency evaluation are presented in more detail.

Chapter 3, Mathematical Model of Thrombin-, Histamine-and VEGF-Mediated signaling in Endothelial Permeability were demonstrated. In particular, the simulated effects of PAR-1, Rho GTPase, ROCK, VEGF and VEGFR2 over-expression on MLC activation, and the collective modulation by thrombin and histamine, enhanced MLC activation by CPI-17

over-expression and by synergistic action of thrombin and VEGF at low mediator levels was the focus of the study.

Chapter 4, Endothelial permeability related disease-Sepsis biomarker selection method from microarray data was described. The new method of Consensus Scoring of Multiple Random Sampling and Gene-Ranking Consistency Evaluation method used for identifying of stable disease-differentiating signatures was presented. The predictive ability of the selected signature shared is evaluated by independent dataset.

Chapter 5, The other type of high-throughput dataset for signature selection system – Breast cancer copy number variation based signature selection were provided. The procedure of CNV calling calculation was presented. Hierarchical clustering analysis and literature search are used to evaluate the expression pattern of the identified markers.

Finally, in the last chapter, **Chapter 6**, major findings and contributions of current work to endothelial permeability were discussed. Limitations and suggestions for further studies were also rationalized in this chapter.

Chapter 2 Methodology

This chapter introduces the methodologies for (1) mathematics model of signaling pathway, (2) disease biomarker selection. In section 2.1, Methods for mathematics model of signaling pathway was described, includes how to develop the model, how to do the parameter estimation and how to do the sensitivity analysis. In Section 2.2, 2.3, Processing for microarray data and copy number variation were described. Section, 2.4 2.5 present the method and strategies used for marker selection from microarray data.

2.1 Methods for mathematics model of signalling pathway

2.1.1 ODE for model development

Biological process can be described in mathematical terms in many ways for pathway simulations. Many different methods have been utilized to describe the various biological processes. For example, ordinary differential equations (ODE) [213-217] were used to describe the glycolytic oscillations, difference equations were used to model population growth, stochastic equations were used for signaling pathways and Boolean networks were used for gene expressions [218]. Each method has got its own strength and limitation and choosing the method to do pathway simulation often depends on the modeler's familiarity to the method and the availability and limitations of the computational power.

ODEs and algebraic closed form equations are the most popular methods of describing the biological systems. This is because despite the various mathematical methods available for describing biological processes, such as partial differential equations (PDEs) [219-224], there is a lack of efficient solvers for these problems. On the other hand, there are well established algorithms solvers to efficiently solve the ODEs, given the limited computational power.

ODEs are describes the change in state variable (eg, protein amount or concentrations) with respect to one dependent variable (eg, time). The ODEs are appropriate to describe the temporal changes of state variables under the assumption that other variables, such as space, are constant. The ODEs are appropriate in biological processes where the number of molecules involved in the reaction is large; the reaction follows well-defined kinetic laws that are often zero- or first-order reactions. In most situations, this assumption is valid as in most enzymatic reactions, the concentration of the reactants are in large excess to the enzymes that catalyze the reaction.

In other situations when other variables are also changing with time, partial differential equations are more appropriate. PDEs can account for state variable that change temporally and spatially. These equations may be appropriate for describing the structural changes in the self-organization biological systems, such as in embryonic development. However, there is a lack of efficient PDE solvers and many biological processes are adequately described by ODEs since

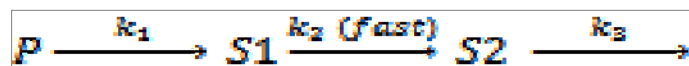
many of the cellular processes reacts within well-defined cellular spaces.

In another situation, when the number of interacting molecules is small; the kinetics of the chemical reaction cannot be assumed to follow a zero- or first order reaction. Instead the interaction follows a random process and stochastic differential equations (SDEs) [225-228] that accounts for random noise may be more suitable. SDEs incorporate the random “white” noise into the differential equation and can model the random fluctuation. Many of the numerical solvers that were developed to solve ODEs can be modified and used to solve the SDEs. Because of the stochastic process, the emphasis is placed on estimating the mean and variance of the distribution of the process. There are other more time-consuming simulation methods, such as Monte-Carlo sampling, to solve ODEs and their uses are limited by the computational power required for large simulations.

ODEs are usually appropriate to describe the biological systems and it is the most popular approach because of the following advantages: One, ODEs can be used to describe many non-linear systems that otherwise have no closed form equations. Two, ODEs employs the continuous timescale unlike in the difference equations which employs discrete timescale. This allows for simulations of biological responses at any time point. Three, there are many ODEs solvers readily available to give reasonably good estimates of the parameters with various algorithm. Non-linear differential equations cannot be solved to give an exact solution and approximate methods through numerical

analysis are almost always utilized to estimate the system parameters. Having an efficient and accurate solver is important in solving these problems.

Under specific condition such as a rapid reaction compared to the observation timescale, it may be appropriate to approximate the intermediate state variable to have reached steady-state or equilibrium. The corresponding differential equation can be simplified to a closed form algebraic expression to reduce the extensive computational requirement to perform numerical analysis of ODEs. In many cases, such simplifications may also be justified because the underlying biological processes occurs rapidly or have no major impact on the temporal relationship with other state variable. There are many methods for transforming the ODES into close formed algebraic expression. One example is the quasi-steady state as described in [218]



The rate of the change of the $S1$ as described by ODE is

$$\frac{dS1}{dt} = P \cdot k_1 - S1 \cdot k_2$$

If the reaction of k_2 is fast, the rate of change of $S1$ can be considered constant and at steady state.

$$\frac{dS1}{dt} = 0, \therefore S1 = \frac{Pk_1}{k_2}$$

A closed form equation of $S1$ may be used to increase the efficiency of analyzing a complex pathway by reducing the number ODEs required to be analyzed numerically.

2.1.2 Parameter estimation

After a model with correct components and connectivity has been constructed, the next critical step is determining the values of the parameters such as the rate constants and initial conditions. Parameters can be obtained directly, or from literature. Even when there is considerable experimental data available, it is common for many parameters in a pathway model to remain unmeasured and require estimation.

The pathway simulation model often contains many parameters that are usually hard to determine accurately *in vivo*. In fact, some of the model parameters are kinetic constants that may not have physiological interpretation and are difficult to quantify in *in vitro* experiments. Nonetheless, there are several methods that were routinely utilized to approximate model parameters for pathway simulations.

There are, however, many of the kinetic parameters [91, 229] previously studied *in vitro* by careful laboratory experiments. Reported rate constants have been published in the scientific literature for many different chemical and biological reactions can be used to approximate the model kinetic parameters. Initial concentrations of various biological molecule concentrations can also be set to the normal ranges reported for various cellular, animal or human species. It can be assumed that the biological pathway kinetics follows the same kinetics law and the model parameters can be estimated to approximate these reported values. Hatakeyama *et al* [230] and colleagues developed a systems biology

model of the mitogen-activated protein kinase (MAPK) and Akt pathways in ErbB signaling. The model contains a total of 33 differential equations describing the temporal change in the concentrations of the molecules involved in the signaling pathways. The model is parameterized by 68 kinetic constants, 38 of these kinetic constants were estimated and taken directly from the published literature.

There are other parameters where the values are not readily available in the scientific database. For these parameters, other estimation methods will need to be considered. One such method to estimate parameters is to using simulation and evaluation algorithm. In the model described above, Hatakeyama *et al* [230] utilized the genetic algorithm to estimate the parameters. The genetic algorithm randomly generates a large number of parameter estimates and then randomly chooses a set of parameter estimates that has the best fitness. The process is repeated with mutations and crossover for a large number of cycles and the parameter estimates with the best calculated fitness is then taken as the parameter estimates for the pathway simulation. The genetic algorithm allows the modeler to estimate the parameter when there is no other information available to base the estimation upon. Caution must be observed when using such algorithm to estimate parameters. Because these estimates are purely mathematical random generation, when experimental or other information are available, the parameter estimates needs to be reviewed again.

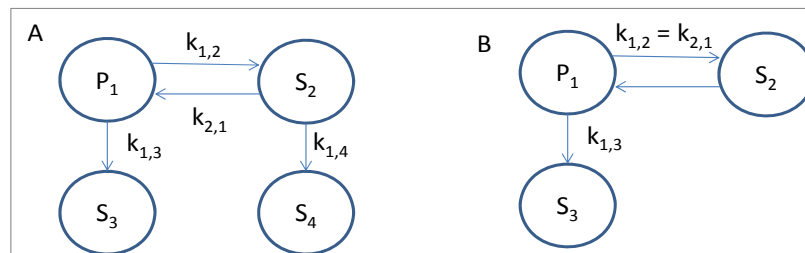
When experimental data are available, models can be fitted to the data and the

parameter can be estimated using regression methods [231]. An important note before estimating parameter through regression is that structural identifiability is a concern. For large complex pathways, it is possible that some parameters are unidentifiable given the incompleteness and uncertainty in the experimental data. **Figure 2-1** illustrates the identifiability issue by comparing two models (models A and B). When the observations are limited and are constrained within certain compartments, there are indistinguishable parameters such as $k_{1, 2}$ and $K_{2, 1}$ in Model A of **Figure 2-1** for example. The structural identifiability of a parameter can be tested by the sensitivity (discussed in the next section) of the parameter to the output. The model is structurally identifiable when the parameter has a large sensitivity and the effects of parameters on the output are uncorrelated [231].

For the structurally identifiable model, many are often described using the ODEs and an appropriate ODE solver is required. However, because biological pathways are often non-linear systems, there are multiple solutions to the differential equation and the traditional methods of identifying the zero gradient may cause the computer to be stuck at a local minimum [232] (Moles *et al*, 2003). Moles *et al* [232] discussed several global optimization methods that are both deterministic (ie. Parameter determined through estimation methods) or stochastic (ie. Parameter determined through random simulations). Some methods mentioned included: Adaptive stochastic methods, Clustering methods, Evolutionary computation and simulated annealing *etc*. The optimization methods used for estimating the model parameters are often determined by the

quality and quantity of the experimental data, the familiarity of the methods by the modeler and the availability of the computer software and computational power. When data is rich and the model is fairly stable, the deterministic methods are appropriate and will the results can usually be determined in a fairly reasonable computer runtime. When data are sparse and little, the stochastic methods might be suitable, although computer runtime may be significantly longer.

Figure 2-1: Unidentifiable model parameters.



Other methods to estimate parameter is to approximate it to known parameters of related molecules. Some proteins share sequence homology and have similar structural conformation. The binding and interaction of these analog proteins can be assumed to be similar. Indeed, Li *et al* (2001) [233] reported that proteins that shares 90% sequence homology exhibits high affinity to a specific epitope of the study protein. The rate constants of binding to the study protein by the different analog proteins differ by only several folds, assuming that these proteins behave similarly. This method of approximating kinetic constants to proteins with sequence homology can be applied to protein-protein interaction

[233], protein-protein interaction with different substrates or protein-protein interaction with the same substrates [234]. However, as always, this method of comparing sequence homology and estimating the parameter must be used with caution. There are instances where the proteins with sequence similarity do not display the same kinetic constant for interaction. For example, Brown and colleagues [235] investigated the kinetics of four different species of Parasite lactate dehydrogenase (pLDH) with one inhibitor. The group reported that although the four species of pLDH shares 90% sequence homology, have similar catalytic residues and have similar crystal structural; the kinetic interactions of the these enzymes with the inhibitor display significant different characteristics.

There could be instances when the above mentioned methods are not suitable for estimating the parameters. For example when there is very little published literature for the molecule or when the other methods for approximating the parameters are difficult. The parameters can be estimated to an arbitrary value which is within reasonable biological ranges. This method is prone to criticism but may be a reasonable step to take to enable the simulation which would otherwise be unattainable. The parameters could be varied within a range to find the best fit of the model to the data. Again, caution must be exercised and not make too many assumptions to the too many of the parameters which would otherwise invalidate the model.

Whichever method the researcher chooses to estimate the parameter, there will

be uncertainties in the parameter values because of it is difficult to accurately measure the in vivo biochemical activities. Gutenkuns *et al* (2007) [236] reported that the systems biology model have a “sloppy spectrum of parameter sensitivities”. Even with the most ideal experimental data, the optimization of parameter estimates can be poor. The authors suggested that when building pathways model and estimating parameters, it is essential to check the predictive power of the model rather than the accuracy of the parameter estimates.

2.1.3 Sensitivity analysis

The change of output with respect to the change in the parameter is described as the sensitivity of the systems. For example when the amount input of the pathway changed, the output can be changed in a corresponding amount. Mathematically, if the output equals y and the parameter of the interest is x_1 , the sensitivity is the simple derivative of y with respect to x_1 , dy/dx_1 . A parameter with large sensitivity is capable of substantially change the system output given a small change in the input. There are several usages of sensitivity analyses. Sensitivity analysis can be used to estimate model parameters, assess the robustness of the model, make predictions of critical points in a pathway and devise experimental condition through optimal design [237].

There are many methods to perform sensitivity analysis. The more common method is a sampling based strategy where the pathway is randomly simulated repeatedly with varying input. The input-output relationship can then be further

investigated with graphical or statistical analysis. There is also less computationally intensive procedure which does not require stochastic simulation. This is done by approximating the information matrix through mathematical algorithm such as the linear noise approximation [237]. Stochastic simulation remains the popular approach to sensitivity analysis and the other deterministic approach maybe suitable when significant computation limitation becomes an issue.

The time profiles of the output with varying inputs can be graphed to observe the trend and magnitude of the output changes with changing inputs. When a parameter estimate cannot be determine through literature database search and has to set arbitrarily, the sensitivity analysis can be used as a parameter estimation method. The parameter input can be varied over a range of values and the simulated output can be graphed against experimental observations. Visual predictive check can be used to determine which parameter estimate best fits the model to the observations. The selection of model parameter can be based on the goodness-of-fit of the model predictions to the experimental data. This allows the modeler to select the parameter estimates that will otherwise be inestimable.

Sensitivity analysis can also be used to test the robustness of a model. There are several ways of assessing the robustness of the model, depending on the knowledge on the pathway. If a particular interaction is known to be a critical control point in the pathway, the sensitivities of the associated model

parameters will be large comparatively. Conversely, if a known interaction does not influence the pathway significantly, the sensitivities of the parameters will be smaller. Otherwise, when a parameter is estimated using sensitivity analysis, the sensitivities of the parameter should ideally be small to increase the robustness of the model. The robustness of the model is defined as the stability of the model over multiple parameters variation.

Additionally, simulations can be performed to determine which parameter has significant impact on the output of the system. Different what-if scenarios can be simulated by altering different model parameters to investigate the critical control points in the pathway. This is a powerful predictive tool for identifying potential drug targets and therapeutic interventions to disease pathways.

2.2 Processing of microarray data

2.2.1 Missing data estimation

Missing values is a common issue existing in microarray data. The missing values arise from experimental errors due to spotting problems (cDNA), dust, poor hybridization, inadequate resolution, fabrication errors (e.g. scratch) and image corruption [238, 239]. They could also come from the suspicious data with low expression (e.g. background is stronger than signal) or censored data [240]. Repeating experiments could be a solution but often not be a realistic option because of economic reasons or limitations in biological material [160, 241]. However, many microarray data analysis methods, such as classification,

clustering and gene selection methods, require complete data matrix. Therefore in many microarray projects, one needs to determine how to estimate missing values. Proper missing value estimation can significantly improve performance of the analysis methods [242-244]. The simplest way is to remove all genes and arrays with missing values, or to replace missing values with an arbitrary constant (usually zero), row (gene) average or column (array) average. The better approaches had also been proposed such as k-nearest neighbors method (KNN) [244], least square methods (LS) [241, 245], and principal component analysis (PCA) [246, 247]. Among these estimation methods, KNN is the most widely used and is also a standard method for missing value estimation currently [242, 244, 248].

The KNN-based method for missing value estimation involves selecting k neighbor genes with similar expression profiles to the target gene (the gene with missing values in one or more arrays), and estimating the missing value of the target gene in specific array as the weighted mean of the expression levels of the k neighbor genes in this array. A popular KNN-based method is KNNimpute [244], which is the only imputation method available in many microarray data analysis tools for missing value estimation [249-251]. KNNimpute can be downloaded from Stanford Microarray Database [248, 252]. In this thesis, KNNimpute is employed if the microarray data contains missing values.

2.2.2 Normalization of microarray data

The purpose of normalization is to remove systematic variations from the expression values, so that biological difference can be easily distinguished and the comparison of expression levels across samples can be performed. In microarray experiments, all the values are fluorescent intensities, which are directly comparable. Therefore the normalization among genes and arrays [159] are both possible.

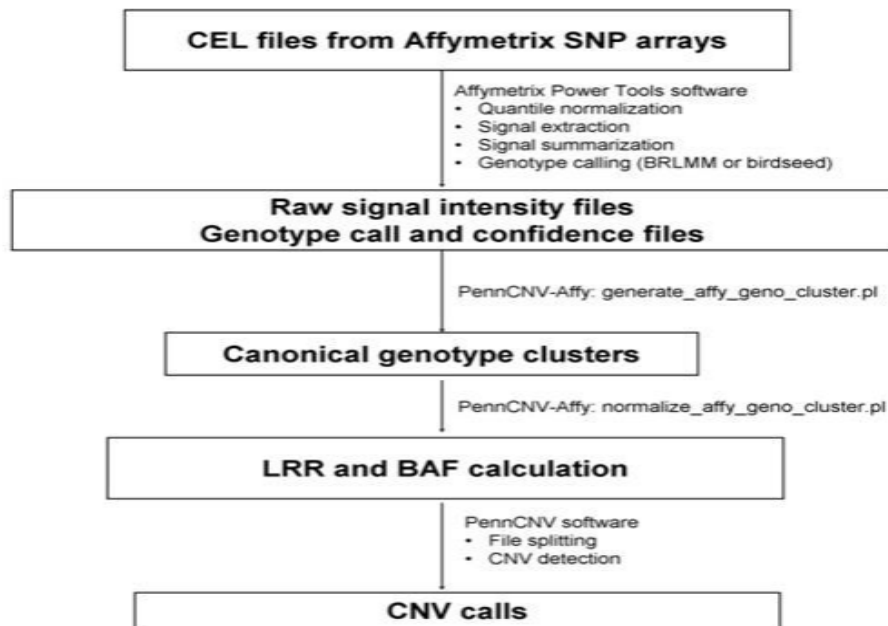
The popular normalization methods for microarray experiments include global normalization using all genes on the array, and housekeeping genes normalization using constantly expressed housekeeping/invariant genes [136]. Since Housekeeping genes are not as constantly expressed as assumed previously [253], using housekeeping genes normalization might introduce extra potential sources of error. It was further approved that normalization using a reduced subset of genes was less statistically robust than the normalization using the entire gene set [254]. Currently, a typical normalization procedure is (1) normalizing the expression levels of each sample to zero-mean and unit variance, and then (2) normalizing the expression levels of each gene to zero-mean and unit variance over all the samples. This normalization method have been shown to perform well [255, 256] and is applied in this thesis.

2.3 Processing Copy Number Variations

2.3.1 Overview of CNV calling calculation

We mainly use integrated hidden Markov model (HMM) algorithm, called “PennCNV,” to detect CNVs with high resolution using the Illumina Infinium assay [257]. To better reflect the distribution of the intensity data, log R Ratio and B Allele Frequency are developed for state transition between different copy number states. In addition, PennCNV incorporates the population allele frequency for each SNP and the distance between adjacent SNPs. Several studies have demonstrated the heritability of CNVs, suggesting that using information from related family members can improve the sensitivity for CNV detection and accuracy of boundary mapping. The application of PennCNV to a large group of individuals demonstrates the feasibility of whole-genome fine-mapping of CNVs through high-density SNP genotyping.

The procedure below (**Figure 2-2**) [257] outlines how to process raw CEL files and generates canonical genotype clusters, then convert signal intensity for each sample to LRR/BAF values, then generates CNV calls. For this protocol to work, one needs to use at least 100 CEL files to generate a reasonably good clustering file. If the user has only a few CEL files, then it is necessary to use the default canonical clustering file in the PennCNV-Affy package, but in this case the CNV calls may not be reliable.

Figure 2-2: Affymetrix CNV calling overview

2.3.2 HMM modelling strategy

Six-state definition [258] for more precise modeling of CNV events is considered (**Table 2-1**). To exploit all available information for each SNP to its full potential, PennCNV incorporates several components together into a hidden Markov model (HMM), including the LRR (The log R Ratio), the BAF (B Allele Frequency), the distance between neighboring SNPs, and the population frequency of the B allele. The LRR is a measure of normalized total signal intensity, and the BAF is a measure of normalized allelic intensity ratio (**Figure 2-3**). Both the LRR and BAF values can be displayed and exported from BeadStudio given that there is an appropriate clustering file with canonical cluster positions for each SNP. The distance between neighboring SNPs determines the probability of having a copy number state

change between them. Each SNP has two alleles referred to as the A and B alleles, thus we use the term “population frequency of B allele” to differentiate it from the BAF term that measures allelic intensity ratio. The values for population frequency of B allele for all SNPs are compiled from a large set of individuals with mixed ethnic backgrounds and of normal phenotypes; the likelihood of the copy number genotypes for each copy number state is then determined.

Table 2-1: Hidden states, copy numbers and their descriptions

Copy no. state	Total copy no.	Description (for autosome)	CNV genotypes
1	0	Deletion of two copies	Null
2	1	Deletion of one copy	A, B
3	2	Normal state	AA, AB, BB
4	2	Copy-neutral with LOH	AA, BB
5	3	Single copy duplication	AAA, AAB, ABB, BBB
6	4	Double copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

Each state has a different distribution of CNV genotypes.

2.3.3 Inference of log R Ratio (LRR) and B Allele Frequency (BAF)

For each SNP, its two alleles are referred to as the A and B alleles using a set of specific naming rules (see http://www.illumina.com/downloads/TopBot_TechNote.pdf). The raw signal intensity values measured for the A and B alleles are then subject to a five-step normalization procedure using the signal intensity of all SNPs (see Illumina white paper at <https://icom.illumina.com/icom/software.ilmn>). This procedure produces the X and Y values for each SNP, representing the experiment-wide normalized signal intensity on the A and B alleles, respectively. Two additional measures

are then calculated for each SNP, where $\mathbf{R} = \mathbf{X} + \mathbf{Y}$ refers to the total signal intensity, and $\theta = \arctan(\mathbf{Y}/\mathbf{X}) / (\pi/2)$ refers to the relative allelic signal intensity ratio.

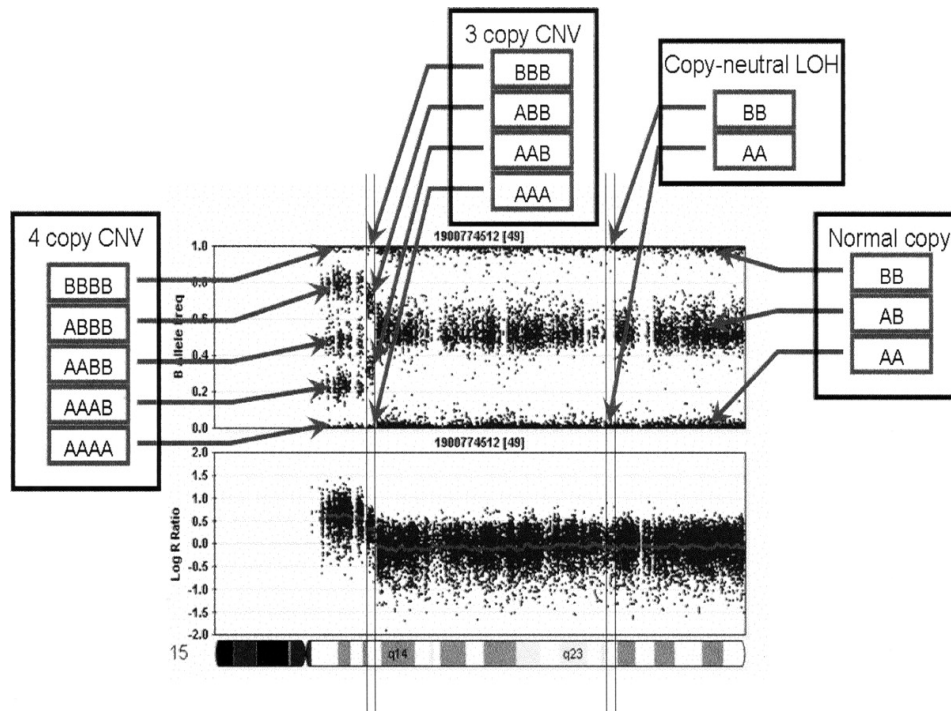
As a normalized measure of total signal intensity, the log R Ratio (LRR) value for each SNP is then calculated as $LRR = \log_2(\mathbf{R}_{\text{observed}}/\mathbf{R}_{\text{expected}})$, where $\mathbf{R}_{\text{expected}}$ is computed from linear interpolation of canonical genotype clusters. The B Allele Frequency (BAF) is a somewhat confusing term that actually refers to a normalized measure of relative signal intensity ratio of the

$$BAF = \begin{cases} 0, & \text{if } \theta < \theta_{AA} \\ 0.5(\theta - \theta_{AA})/(\theta_{AB} - \theta_{AA}), & \text{if } \theta \leq \theta < \theta_{AB} \\ 0.5 + 0.5(\theta - \theta_{AB})/(\theta_{BB} - \theta_{AB}), & \text{if } \theta_{AB} \leq \theta < \theta_{BB} \\ 1, & \text{if } \theta \geq \theta_{BB} \end{cases} \quad (1)$$

B and A alleles:

where θ_{AA} , θ_{AB} , and θ_{BB} are the θ values for three canonical genotype clusters generated from a large set of reference samples. The transformation from θ to BAF values adjusts for different chemical characteristics of each SNP so that values for different SNPs are more comparable to each other.

Figure 2-3: An illustration of log R Ratio (LRR) and B Allele Freq (BAF) values for the chromosome 15 q-arm of an individual.



2.4 Support Vector Machines

2.4.1 Theory and algorithm

Support vector machines (SVM) is a relatively new machine learning method proposed by Vapnik [151, 259, 260]. It defines a mapping, or a decision function, from feature vector space to the class label space. Over the past decades, SVM has become a popular supervised learning method in variety applications including image classification and object detection [261, 262], text categorization [263], prediction of protein solvent accessibility [264], microarray data analysis [159, 160, 192, 205], protein fold recognition [265], protein secondary structure prediction [266], prediction of protein-protein

interaction [267] and protein functional class classification [268].

SVM can be divided into linear and non-linear SVM. Linear SVM directly constructs a hyperplane in the feature space to separate positive examples from negative examples. On the other hand, non-linear SVM projects both positive and negative examples into a higher-dimensional feature space and then separates them in that space.

Linear SVM is the simplest form of SVM, in which the data represented as a p-dimensional vector (a list of p numbers) can be separated by a p-1 dimensional hyperplane. On each side of this p-1 hyperplane, two parallel hyperplanes can be constructed (**Figure 2-4**). The separating hyperplane is the one that maximizes the distance between these two parallel hyperplanes. Many linear hyperplanes (also called classifiers) can separate the data. However, only one achieves maximum separation. Under the assumption that the larger the margin or distance between these two parallel hyperplanes the better the generalization error of the classifier will be [269], the maximum separating hyperplane (also known as maximum-margin hyperplane) is clearly of interest (**Figure 2-5**).

Mathematically, supposed the training set is composed of n examples with two classes, it could represent as

$$\mathcal{X} = \{ (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \}, \quad i=1,$$

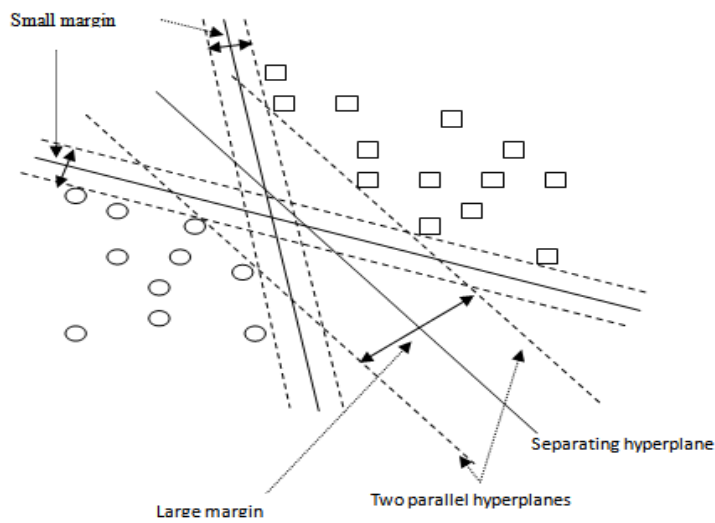
$$2, \dots, n, \quad (1)$$

where $x_i \in R^N$ is an N-dimensional real vector and $y_i \in \{-1, +1\}$ indicates class label. The separating hyperplane can be described by equation:

$$w \bullet x + b = 0$$

(2) where $w=(w_1, w_2, \dots, w_n)^T$ is a unit vector of n elements and b is a constant, and the relative two parallel hyperplanes can be described by equations

Figure 2-4: Margins and hyperplanes



$$w \bullet x + b = +1 \quad \text{for} \quad y_i = +1 \quad (3)$$

$$w \bullet x + b = -1 \quad \text{for} \quad y_i = -1 \quad (4)$$

If the training data are linearly separable, we can select those two parallel

hyperplanes with no data point between them and try to maximize their distance. By using geometry, we find the distance between the two parallel hyperplanes is $2/|w|$. Therefore, to obtain the solution of SVM, $|w|$ should be minimized.

To exclude data points between the two parallel hyperplanes, we need to ensure that for all i either

$$w \cdot x + b \geq +1 \quad \text{for } y_i = +1 \quad \text{or} \quad (5)$$

$$w \cdot x + b \leq -1 \quad \text{for } y_i = -1 \quad (6)$$

It can be rewritten as

$$y_i (w \cdot x_i + b) \geq 1, \quad 1 \leq i \leq n \quad (7)$$

The problem now is to minimize $|w|$ subject to the above constraint. More clearly,

$$\text{Minimize } \frac{1}{2} \|w\|^2 \quad (8)$$

$$\text{Subject to } y_i (w \cdot x_i + b) \geq 1, \quad 1 \leq i \leq n$$

This is a quadratic programming (QP) optimization problem.

Such optimization problem could be efficiently solved with the introduction of lagrangian multiplier α_i ,

$$L_p(w, b, a) = \frac{1}{2} \|w\|^2 - \sum \alpha_i (y_i \cdot ((x_i \cdot w) + b) - 1) \quad (9)$$

where $\alpha_i \geq 0$.

The solution to this QP optimization problem requires that the gradient of

$L(w,b,\alpha)$ with respect to w and b vanishes,

$$\boxed{\frac{\partial}{\partial w} L_p(w, b, a) = 0} \quad \text{and} \quad (10)$$

$$\boxed{\frac{\partial}{\partial b} L_p(w, b, a) = 0} \quad (11)$$

resulting in the following conditions:

$$\boxed{w = \sum_{i=1}^n a_i y_i x_i} \quad \text{and} \quad (12)$$

$$\boxed{\sum_{i=1}^n a_i y_i = 0} \quad (13)$$

By substituting Equations (12) and (13) into Equation (9), the QP problem becomes the maximization of the following expression:

$$\boxed{L_p(w, b, a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \bullet x_j)} \quad (14)$$

under the constraints $\boxed{\sum_{i=1}^n a_i y_i = 0}$, $0 \leq \alpha_i \leq C$, $i=1, 2, \dots, n$. C is a penalty for training errors for soft-margin SVM and is equal to infinity for hard-margin SVM.

The points located on the two optimal margins will have nonzero coefficients α_i among the solutions to Equation (14), and are called Support Vectors (SV). The

bias b_0 can be calculated as follows:

$$b_0 = -\frac{1}{2} \left\{ \min_{\{x_i|y_i=+1\}} (w_0 \cdot x_i) + \max_{\{x_i|y_i=-1\}} (w_0 \cdot x_i) \right\} \quad (15)$$

After determination of support vectors and bias, the decision function that separates two classes can be written as:

$$f(x) = \text{sign} \left[\sum_{i=1}^n a_i y_i \mathbf{x}_i \cdot \mathbf{x} + b_0 \right] = \text{sign} \left[\sum_{SV} a_i y_i \mathbf{x}_i \cdot \mathbf{x} + b_0 \right] \quad (16)$$

When the examples are inseparable by linear SVM, nonlinear SVM can be applied, which projects the input data to a higher dimensional feature space by using a kernel function $K(x,y)$. The linear SVM procedure is then applied to the feature vectors in this feature space. After the determination of w and b , a given vector x can be classified by using

$$f(x) = \text{sign} \left[\sum_{SV} a_i y_i K(x \cdot x_i) + b_0 \right] \quad (17)$$

A positive or negative value indicates that the vector x belongs to the positive or negative class respectively.

In equation (17), kernel function $K(x,y)$ represents a legitimate inner product in the input space:

$$K(x, y) = \varphi(x) \cdot \varphi(y) \quad (18)$$

A number of kernel functions have been used in SVM. Examples of the most popular ones are:

Polynomial

kernel

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^p \quad (19)$$

Sigmoid kernel
$$K(x_i, x_j) = \tanh(\kappa x_i x_j + c) \quad (20)$$

Radial basis function (RBF)
$$K(x_i, x_j) = e^{-\|x_j - x_i\|^2 / 2\sigma^2} \quad (21)$$

In practice, RBF kernel is the most widely used kernel function due to three reasons. First, linear kernel and sigmoid kernel can be treated as special cases of RBF kernel since RBF kernel in certain parameters has the same performance as the linear kernel [270] or sigmoid kernel [271]. Second, comparing with polynomial kernel, RBF kernel has few parameters which influence the complexity of model selection. Third, RBF function has less computational cost compared with polynomial kernels in which kernel values may go to infinity or zero while the degree is large. Based on these reasons, this thesis mainly used RBF kernel.

Several specialized algorithms can be used to solve the QP problem of SVM by heuristically breaking the problem down into smaller, more-manageable chunks. **Table 2-2** listed some popular SVM software tools. In our case, we modified the source code of libSVM to fit our program requirements. libSVM is a sequential minimal optimization (SMO) algorithm[272], which breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating

the need for a numerical optimization algorithm such as conjugate gradient methods.

Figure 2-5 : Architecture of support vector machines

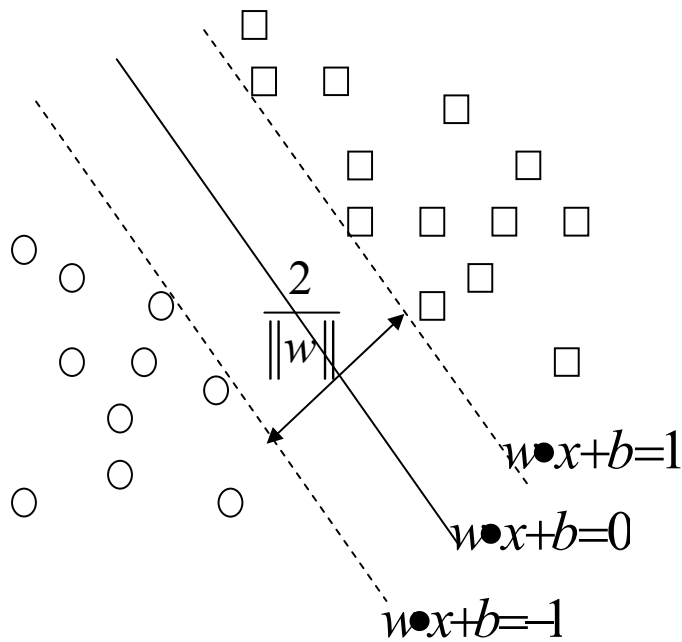


Table 2-2 : List of some popular used support vector machines software

Software	URL
SVM-Light	http://svmlight.joachims.org/
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/
WinSVM	http://www.cs.ucl.ac.uk/staff/M.Sewell/winsvm/
LS-SVMlab	http://www.esat.kuleuven.ac.be/sista/lssvmlab/
GIST SVM Server	http://svm.sdsc.edu/svm-intro.html

2.4.2 Performance evaluation

The performance of SVM can be measured as true positive TP (the number of positive examples which are correctly predicted as positive), false negative FN (the number of positive examples which are incorrectly predicted as negative), true negative TN (the number of negative examples which are correctly predicted as negative) and false positive FP (the number of negative examples which are incorrectly predicted as positive) (**Table 2-3**).

The simplest way to evaluate the performance of a classification is overall accuracy (Q), which measures the proportion of the total number of the correctly predicted examples.

$$Q = \frac{TP + TN}{TP + FN + TN + FP} \quad (22)$$

Another two concepts, sensitivity (SE) and specificity (SP), which measure the positive and negative prediction performance respectively, are also frequently used in classification.

$$SE = \frac{TP}{TP + FN} \quad (23)$$

$$SP = \frac{TN}{TN + FP} \quad (24)$$

In some cases such as epidemiology and the evaluation of diagnostic tests [273], positive predictive value (PPV, also called precision rate) and negative predictive value (NPV) are very important.

$$PPV = \frac{TP}{TP + FP} \quad (25)$$

$$NPV = \frac{TN}{TN + FN} \quad (26)$$

Table 2-3: Relationships among terms of performance evaluation.

		Condition		
		True	false	
Test outcome	Positive	True positive (TP)	False positive (FP)	→Positive predictive value (PPV)
	Negative	False negative (FN)	True negative (TN)	→Negative predictive value (NPV)
		↓ Sensitivity (SE)	↓ Specificity (SP)	

2.5 Methodology for gene selection

2.5.1 Overview of the gene selection procedure

A novel gene selection procedure method based on Support Vector Machines classifier, recursive feature elimination, multiple random sampling strategies and multi-step evaluation of gene-ranking consistency was established (**Figure 2-6**):

(1) After preprocessing the original data, by using random sampling method, a large number of training-test sample combinations are generated from the original data set.

(2) The large number of sample combinations is divided into n groups, and each group contains 500 sample combinations.

(3) In each training-test sample combination of each group, SVM and RFE are used to classify the samples (SVM classifiers) and rank the genes (RFE gene rank criteria). Therefore 500 gene ranking sequences are formed.

(4) The consistency evaluation can be performed based on the 500 sequences and a certain number of genes (for example, k genes) can be eliminated.

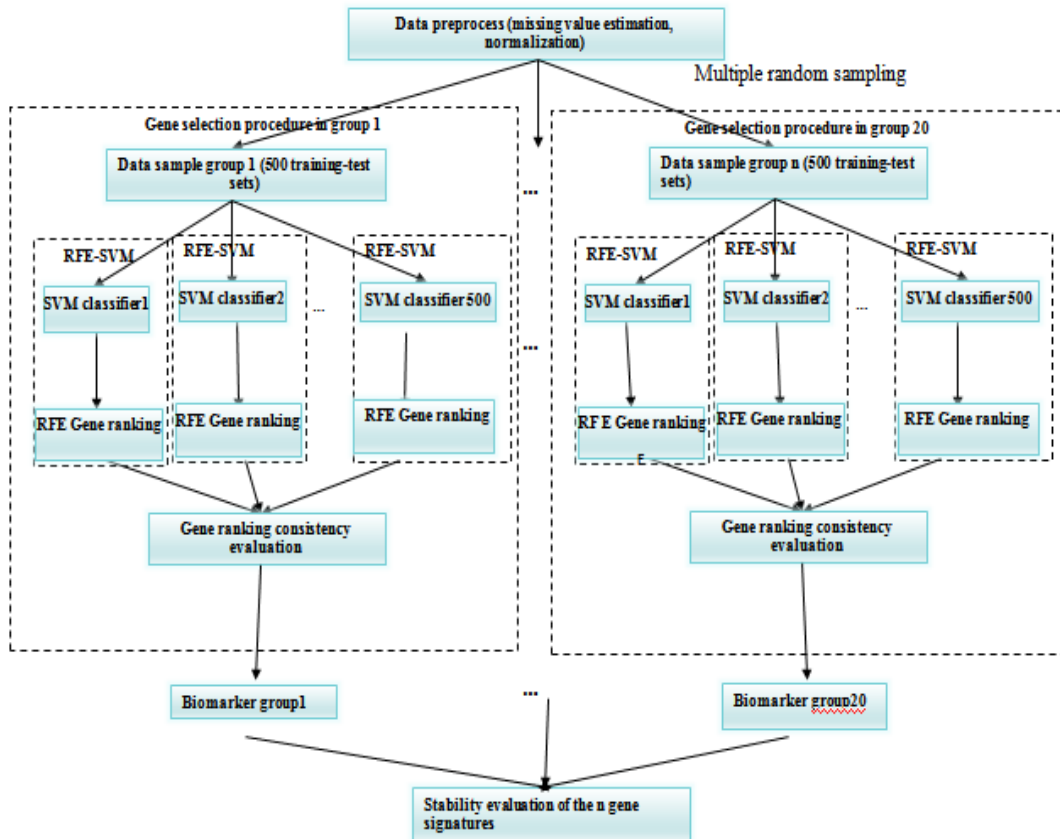
(5) Step (3) and (4) can be iteratively done until no gene can be eliminated.

(6) The gene subset which gives us the highest overall accuracies of the 500 test sample sets can be selected as gene signatures of this group. By this way, we can obtain n gene signatures.

(7) The stability evaluation of the gene signatures can be performed by looking into the overlap gene rate of the n gene signatures.

Below Recursive feature elimination is introduced first and followed by a detailed introduction of the whole feature selection procedure.

Figure 2-6: Overview of the gene selection procedure



2.5.2 Recursive feature elimination

During gene selection procedure, the genes ranked according to their contribution to the SVM classifiers. The contributions of genes are calculated by Recursive feature elimination (RFE) procedure, which sort genes according to a gene-ranking function generated from SVM classifier. As described in Section 2.1, SVM training process needs to find the solution for the optimum problem (also known as objective function or cost function) shown in equation (14), which can be rewritten as

$$J = \frac{1}{2} \alpha^T H \alpha - \alpha^T 1 \quad (27)$$

Under the constraints $\sum_{i=1}^n a_i y_i = 0$ and $\alpha_i \geq 0$, $i=1,2,\dots,n$.

Where $H(i, j) = y_i y_j K(x_i, x_j)$, K is the kernel function.

The gene-ranking function of RFE can be defined as the change in the objective function J upon removing a certain gene. When a given feature is removed or its weight w_k is reduced to zero, the change in the cost function $J(k)$ is

$$DJ(k) = \frac{1}{2} \frac{\partial^2 J}{\partial w_k^2} (Dw_k)^2 \quad (28)$$

where the change in weight $Dw_k = w_k - 0$ corresponds to the removal of feature k .

Under the assumption that the removal of one feature will not significantly influence the values of α_s , the change of cost function can be estimated as

$$DJ(k) = \frac{1}{2} \alpha^T H \alpha - \frac{1}{2} \alpha^T H(-k) \alpha \quad (29)$$

Where H is the matrix with elements $y_i y_j K(x_i, x_j)$, and H(-k) is the matrix computed by using the same method as that of matrix H but with its kth component removed.

The change in the cost function indicates the contribution of the feature to the decision function, and serves as an indicator of gene ranking position [274].

2.5.3 Sampling, feature elimination and consistency evaluation

In order to present statistical meaning, gene selection is conducted based on multiple random sampling. Each random sampling divide all microarray samples into a training set which contains half number of samples and an associates test set which contains another half number of samples. This sampling method can be treated as a special case of the bootstrap technique. Many researchers showed that bootstrap-related techniques present more accurate estimation than cross-validation on small sample sets [275, 276]. By using this random sampling, thousands of training-test sets, each containing a unique combination of samples, are generated. These thousands of randomly generated training-test sets are randomly divided into several sampling groups, with equal number of training-test sets (such as 500 training-test sets) in each group. Every sampling group is then used to derive a signature by RFE-SVM.

In every training-test sampling group generated by multiple random sampling, each training-set (totally 500 training-test sets) is used to train a SVM

class-differentiation system and the genes are ranked by using Recursive feature elimination (RFE), according to the contribution of genes to the SVM classifier. In order to derive a gene-ranking criterion consistent for all iterations and all the 500 training-test sets in this group, a SVM class-differentiation system with a universal set of globally optimized parameters, which give the best average class-differentiation accuracy over all of the 500 test sets in this group, is applied by RFE gene-ranking function at every iteration step and for every training-test set.

To further reduce the chance of erroneous elimination of predictor-genes, additional gene-ranking consistency evaluation steps are implemented on top of the normal RFE procedures in each group:

(1) For every training-set, subsets of genes ranked in the bottom (which give least contribution to the SVM classification procedure) with combined score lower than the first few top-ranked genes (which give highest contribution to the SVM classification procedure) are selected such that collective contribution of these genes less likely outweigh top-ranked ones.

(2) For every training-set, the step (1) selected genes are further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes are consistently ranked lower.

(3) A consensus scoring scheme is applied to step (2) selected genes such that

only those appearing in most of the 500 testing-sets were eliminated.

For each sampling group, different SVM parameters are scanned, various RFE iteration steps are evaluated to identify the globally optimal SVM parameters and RFE iteration steps that give the highest average class-differentiation accuracy for the 500 testing-sets.

The several signatures derived from these sampling-groups are then applied to evaluate the stability and performance.

Chapter 3 Mathematical Model of Thrombin-, Histamine-and VEGF-Mediated Signalling in Endothelial Permeability

This chapter describes a mathematical model of thrombin-, histamine-and VEGF-Mediated signaling in endothelial permeability. The model was validated against experimental data for calcium release and thrombin-, histamine-, and VEGF-mediated MLC activation. The simulated effects of PAR-1, Rho GTPase, ROCK, VEGF and VEGFR2 over-expression on MLC activation, and the collective modulation by thrombin and histamine are consistent with experimental findings. Our model was used to predict enhanced MLC activation by CPI-17 over-expression and by synergistic action of thrombin and VEGF at low mediator levels. These may have impact in endothelial permeability and metastasis in cancer patients with blood coagulation. The model also can be used to predict the effects of altered pathway components, collective actions of multiple mediators and the potential impact to various diseases. Similar to the published models of other pathways, our model can potentially be used to identify important disease genes through sensitivity analysis of signaling components.

3.1 Introduction

The endothelium is a semi-permeable barrier that regulates the flux of liquid and solutes between the blood and surrounding tissues. Endothelial

permeability increases paracellular leakage of plasma fluid and proteins to surrounding tissues, and intravasation of tissue-released contents to the blood in the development of multiple diseases related to injury (such as edema, trauma, ischaemia-reperfusion injury, respiratory distress syndrome, and thrombosis), inflammation (such as atherosclerosis and sepsis), diabetes, and cancer [1, 277-279]. The level of endothelial permeability is regulated individually or in combination by multiple mediators, particularly thrombin, histamine, and vascular endothelial growth factor (VEGF), under various disease conditions [1].

The proinflammatory and vasoactive factors thrombin, generated in thrombosis and inflammatory diseases, and histamine, produced in acute inflammatory responses to trauma, burns, allergy, and infection, induce transient endothelial permeability to link inflammation, tissue injury and vascular leakage to cellular responses and symptoms [280-282]. VEGF, released in diabetic retinopathy, I-R injury, vasculogenesis, angiogenesis, and tumour development and metastasis, causes endothelial permeability to enable extravasation of fluids and solutes and intravasation of tumor cells [283-285]. These three key mediators stimulate their respective receptors on endothelial cells to individually and collectively activate Ca^{2+} , Rho GTPase/ROCK, and Myosin light chain kinase (MLCK) signalling pathways that subsequently activate myosin light chain (MLC) to induce cytoskeleton contraction in endothelial cells and dissociation of cell-cell junctions, resulting in endothelial hyper-permeability [1, 286].

Significant progress has been made in understanding the molecular mechanism and dynamics of the relevant signalling events [1, 282, 284, 286, 287] and the roles of different regulators [288, 289]. Nonetheless, some puzzles still remain to be elucidated. For instance, it is unclear what contributes to the different temporal effects and permeability recovery rates by histamine, thrombin, and VEGF mediated signalling, given that they share similar signalling cascades in triggering endothelial permeability. Another question is how multiple mediators under certain complicated inflammatory conditions collectively reduce the effectiveness of antagonizing agents directed at individual mediator-mediated signalling [1].

As part of the efforts for solving these puzzles and for quantitative and mechanistic study of the relevant signalling events, mathematical models have been developed for analyzing the relevant signalling and regulation processes [290-295]. In particular, ordinary differential equation (ODE) based mathematical models of thrombin, Ca^{2+} -calmodulin (CaM), and Rho activation have been developed for investigating the thrombin-mediated activation of MLC [293], and Ca^{2+} -CaM, MLCK and Myosin Light chain phosphatase (MYCP) on MLC activation [290, 291, 296]. To enable more comprehensive analysis of signaling in endothelial permeability, there is a need to develop an expanded ODE model that covers the signaling mediated by multiple mediators, particularly thrombin, histamine and VEGF.

In this work, we developed a mathematical model that integrates thrombin,

histamine, and VEGF mediated signalling in endothelial permeability by extending the published ODE models of the thrombin-mediated pathway and Ca^{2+} -CaM and MLCK activation of MLC [290, 291, 293, 296]. The framework of our integrated mathematical model is illustrated in **Figure 3-1** and the detailed pathway maps of all three signalling components and thrombin-, histamine- and VEGF-mediated signalling cascades are given in **Figure 3-1**, **Figure 3-2**, and **Figure 3-3** respectively. Detailed molecular interactions and the corresponding kinetic data were obtained from the literature, including published simulation models [290, 291, 293, 296], which are summarized in **Table 3-1**. Our model was validated by evaluating whether the time course of MLC activation by each individual mediator (thrombin, histamine, and VEGF) is in agreement with published experimental and computational findings. The sensitivity of our model with respect to parameters was analyzed to evaluate its robustness. The validated model was then used to study the modulation of other pathway components by each individual mediator (thrombin, histamine, and VEGF) [1, 286] and the modulation of MLC activation by combination of a pair of key mediators thrombin and histamine [297, 298]. Our model was further used to predict the regulation of MLC activation by PKC-potentiated inhibitory protein of 17 kDa (CPI-17) over-expression and by combination of thrombin and VEGF at low mediator levels. The effects of the protein variation of key signalling components protease-activated receptor-1 (PAR-1), VEGF, VEGFR2, Rho GTPase, and ROCK on MLC activation were also studied. Some of these components are significantly elevated in different diseases and have been

explored as therapeutic targets for pharmacological intervention of endothelial permeability and barrier function in these diseases [295].

3.2 Thrombin-, Histamine-and VEGF-Mediated Signaling

Cascades in endothelial permeability mediators

3.2.1 Thrombin mediated GPCR activation

Thrombin regulates endothelial permeability, inflammation and other events via activation of thrombin receptors such as PAR-1 by proteolytically cleaving the N-terminus of these receptors [299]. PAR-1 is the main receptor in the regulation of endothelial permeability (**Figure 3-1**). It interacts with Gq to increase the concentration of Ca^{2+} and activate protein kinase C, inositol 1, 4, 5-triphosphate and diacylglycerol [37]. It is also linked to G12/13 [300] to activate the small G-protein Rho [301].

Figure 3-2: The detailed pathway map of the thrombin-mediated signalling component of our integrated pathway simulation model. ROCK (f) and ROCK (o) refer to ROCK in folded and open conformation respectively.

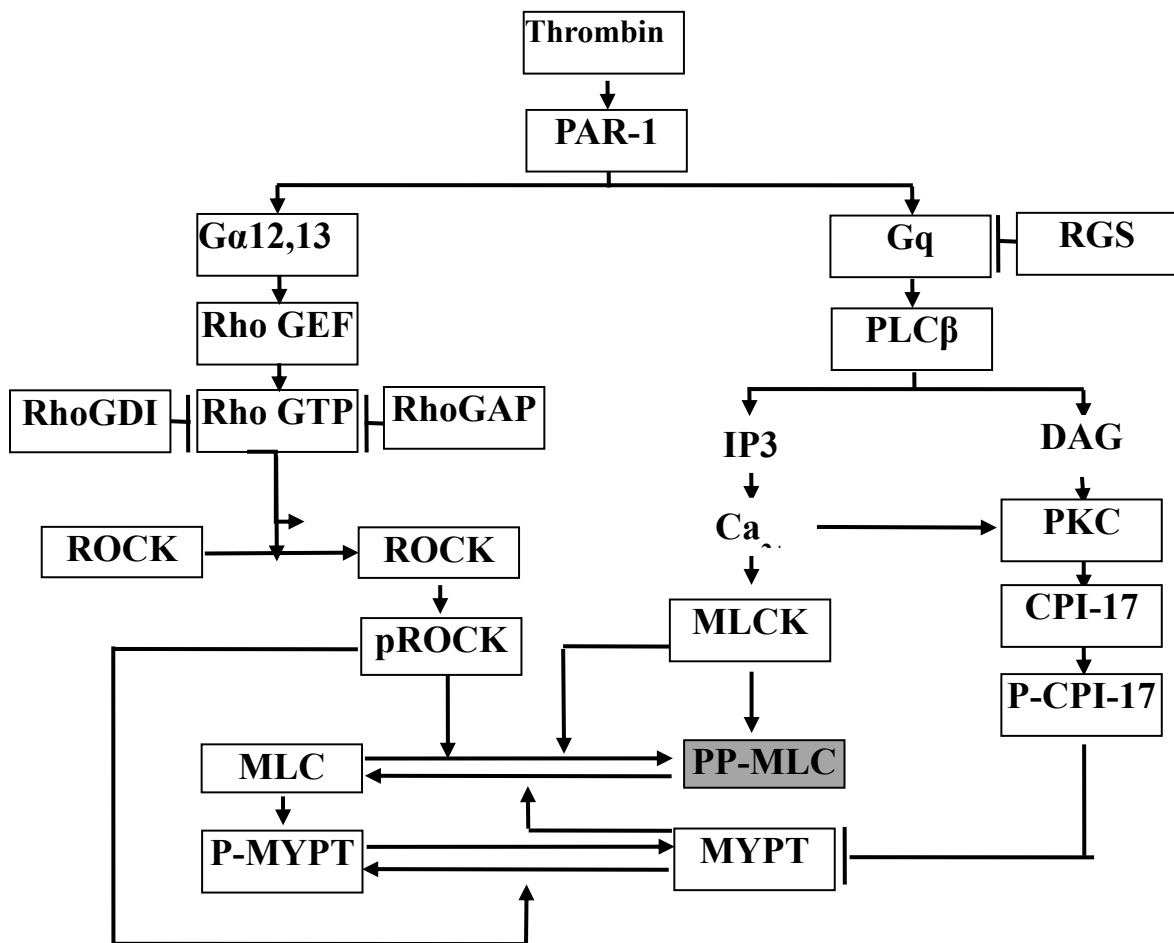


Figure 3-3: The detailed pathway map of the VEGF-mediated signalling component of our integrated pathway simulation model.

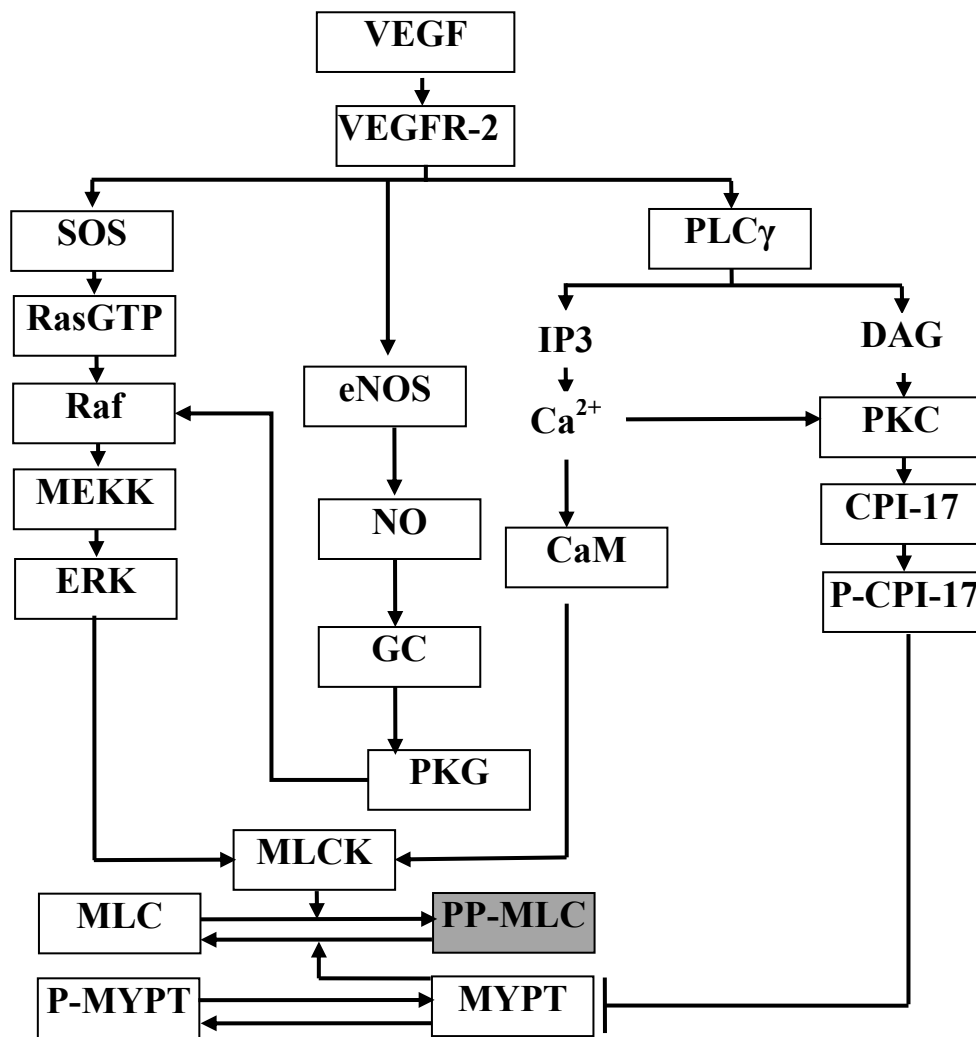
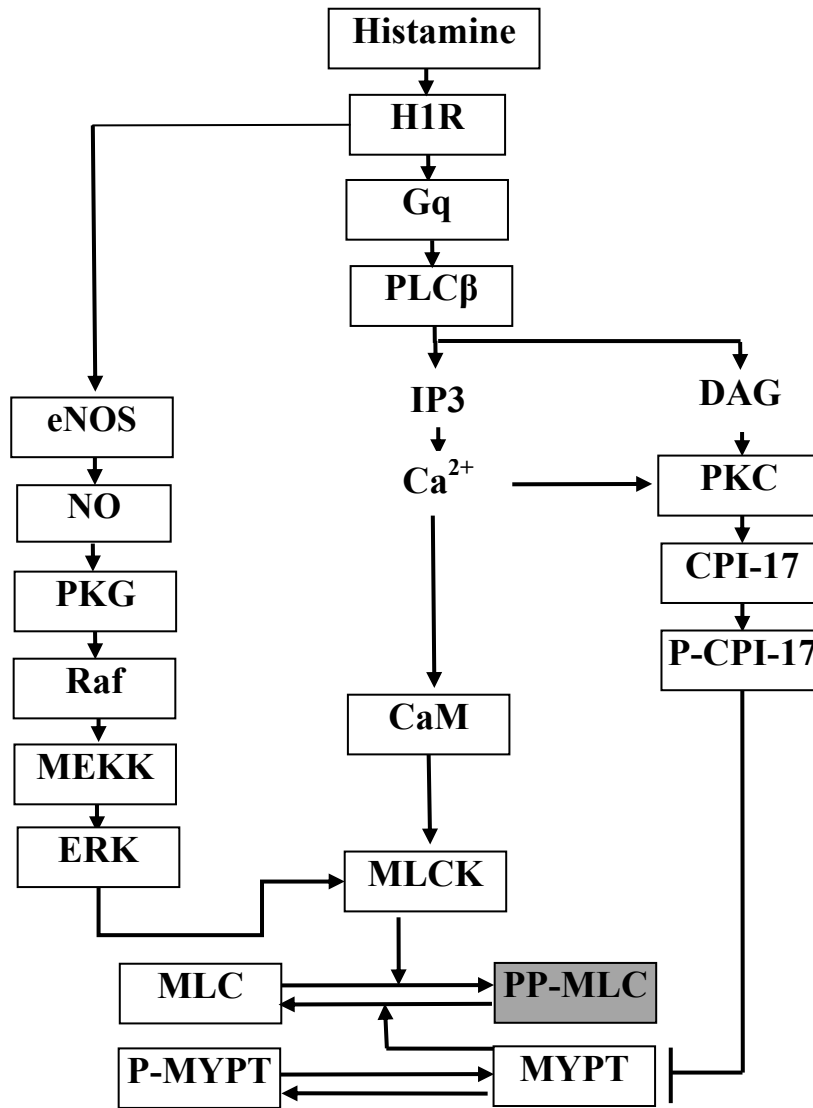


Figure 3-4: The detailed pathway map of the histamine-mediated signalling component of our integrated pathway simulation model.



3.2.2 Role of MAP Kinase in Cell Migration

VEGF can activate ERK-1/2 through the typical signaling of Raf-MEK-ERK pathway [302]. For the subsequent signaling of MLC activation, there are

accumulating evidences that ERK-MLCK-mediated cytoskeletal responses contribute to VEGF-elicited microvascular hyperpermeability. Shoemaker *et al.* has examined that MLCK contains multiple MAP kinase consensus phosphorylation sites (P-x-S[37]-P) and it can be directly phosphorylated by MAP kinase [37]. Evidence presented in by Richard's experiment demonstrates that MLCK, a key regulator of cell motility and contraction, is a substrate for MAP kinase [303].

3.2.3 VEGF mediated ERK activation

VEGF regulates angiogenesis, cancer and microvascular permeability under various physiological and pathological conditions by activating transmembrane tyrosine kinase receptors VEGFR-2 and Flt-1, which promotes mitogenic, chemotactic, and prosurvival signalling and activates phospholipase C (PLC), intracellular Ca^{2+} , and various protein kinase C (PKC) isoforms. In particular, VEGF activates ERK-1/2 via the Raf-MEK-ERK cascade [302]. Accumulative evidences suggest that ERK-MLCK-mediated cytoskeletal responses contribute to VEGF-elicited microvascular hyperpermeability. For instance, MLCK has been found to contain multiple MAP kinase consensus phosphorylation sites (P-x-S[37]-P) that can be directly phosphorylated by MAP kinase [37], which is supported by additional experimental evidence indicating MLCK as a substrate for MAP kinase [303].

3.2.4 Thrombin, VEGF and Histamine mediated Ca²⁺ release, PKC

activation MLC activation

Phosphorylation of regulatory light chain (MLC) of myosin II plays a critical role in controlling actomyosin contractility in both smooth muscle and nonmuscle cells [304]. MLC phosphorylation is regulated by the balance of two enzymatic activities, i.e., Myosin light chain kinase (MLCK) and myosin phosphatase (MYCP). MLCK is regulated by Ca²⁺ /calmodulin and is believed to be a major kinase in both smooth muscle and nonmuscle cells. In addition, Rho-kinase can directly phosphorylate MLC in vitro [304]. MYCP is a holoenzyme composed of three subunits: a catalytic subunit of 38 kDa that was identified as protein phosphatase 1 (PP1) catalytic subunit δ -isoform (PP1C δ) [305] and two noncatalytic subunits of 21 and 110–130 kDa [37]. The larger one, called myosin phosphatase targeting subunit 1 (MYPT1), binds to the catalytic subunit and targets it to MLC, providing substrate specificity [37]. Rho-kinase (RhoK) and protein kinase C (PKC) have been proposed to mediate the inhibition of smooth muscle MYCP, leading to increased MLC phosphorylation in response to various agonists. Phosphorylation of the MYPT1 regulatory site (Thr695 in chicken MYPT1) by RhoK induces inhibition of MYCP activity [37]. A number of experimental facts suggest that CPI-17 (for PKC-potentiated inhibitory protein of 17 kDa) is involved in PKC-dependent inhibition of MYCP. CPI-17 is a soluble globular protein described as a specific inhibitor for MYCP [37].

3.2.5 Thrombin, VEGF and Histamine mediated MLC activation

MLC of myosin II plays a critical role in controlling actomyosin contractility in both smooth muscle and nonmuscle cells [306-308]. MLC phosphorylation is regulated by the balance of two enzymatic activities, i.e., MLCK and myosin phosphatase (MYCP). MLCK is regulated by Ca^{2+} /calmodulin and is believed to be a major kinase in both smooth muscle and nonmuscle cells. In addition, Rho-kinase (ROCK) can directly phosphorylate MLC in vitro [304]. MYCP is a holoenzyme composed of three subunits: a catalytic subunit of 38 kDa that was identified as protein phosphatase 1 (PP1) catalytic subunit δ -isoform (PP1C δ) [305] and two noncatalytic subunits of 21 and 110–130 kDa [309], [310]. The larger one, called myosin phosphatase targeting subunit 1 (MYPT1), binds to the catalytic subunit and targets it to MLC, providing substrate specificity [311]. ROCK and PKC have been proposed to mediate the inhibition of smooth muscle MYCP, leading to increased MLC phosphorylation in response to various agonists. Phosphorylation of the MYPT1 regulatory site (Thr695 in chicken MYPT1) by ROCK induces inhibition of MYCP activity [312]. Some experimental findings suggest that CPI-17, a soluble globular protein, is involved in PKC-dependent inhibition of MYCP and it has thus been considered as a specific inhibitor for MYCP [313].

3.3 Methods

3.3.1 Model Development

One of the most commonly used approaches to model biological systems is that of ODEs. In general, a differential equation can be used to describe the chemical reaction rate that depends on the change of participating species over time. The temporal dynamic behavior of molecular species in the biological signalling pathway network can be captured by a set of coupled ODEs. Our pathway model is illustrated in **Figure 3-4**. Thrombin, VEGF and histamine induced MLC activation, as well as Ca^{2+} -dependent and ROCK-dependent activation of MLC, were included in the model. The constituent molecular interactions, their kinetic constants and molecular concentrations are described in detail in **Table 3-1**. The ODEs for these interactions were derived based on mass-action laws with interaction rate constants defined by the forward and reverse rate constants K_f and K_b or turnover number K_{cat} for enzymatic reactions derived from published models [37] and other literature. Our simulation model contains 200 equations and interactions and 185 distinct molecular species, characterized by 319 kinetic parameters and 48 initial molecular concentrations. These ODEs were then solved by using the Dormand-Prince pair based Ode45 solver of Matlab with the absolute tolerance of $1.0\text{E-}6$ and relative tolerance 0.0010. A Systems Biology Markup Language (SBML) version of our model is provided at <http://bidd.nus.edu.sg/group/Supplement.htm>, and uploaded into the BioModels [314] and KDBI [315] databases.

Figure 3-5: Framework of integrated pathway simulation model of thrombin-, histamine-, and VEGF-mediated MLC activation.

The components in existing models are highlighted by red, blue, and red + blue + green background color for models from reference 15, 16 and 18 respectively. The protein in the gray box represents output signaling. ROCK (f) and ROCK (o) refer to ROCK in folded and open conformation respectively.

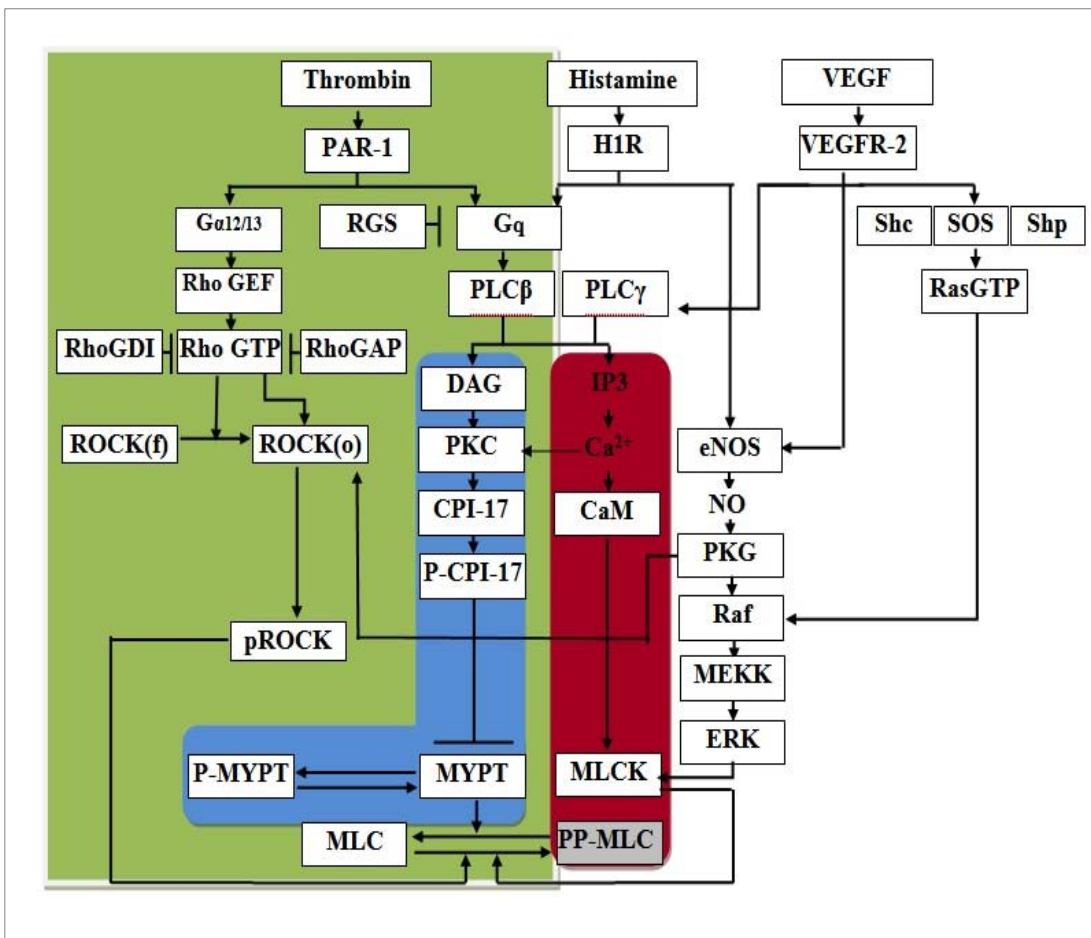


Table 3-1: List of chemical reactions and related kinetic parameters in model.

The relevant references from which the parameters obtained are given. Some of the kinetics values used in this study are not necessary exactly the same as the values given in the cited references but were scaled and optimized in 10-fold ranges according to the performance and kinetics of current model (See Model Optimization and Validation under Materials and Methods section in main text for detailed description).

For those kinetics parameters that are not readily available, parameter values from their homologous partners were taken and were subsequently scaled and optimized in 10-fold ranges (denoted as “Estimated” in the Table).

“=” reversible reaction

“->” enzyme catalytic reaction

Reaction Number	Chemical Reactions and Description	Kf (uM.s)-1	Kb (s)-1	Kcat (s)-1	References
	<i>Activation of GPCR by thrombin</i>				
1	Thrombin + Pro_thrombinR (PAR-1) = Thrombin-Pro_thrombinR	3	0.01		[300]
2	Thrombin-Pro_thrombinR -> ThrombinR-active + Thrombin			0.8	[300]
3	ThrombinR-active + G12α_Gβγ_GDP = ThrombinR-active-G12α_Gβγ_GDP	0.6	0.0006		[300]
4	ThrombinR-active-G12α_Gβγ_GDP + GTP -> G12α_GTP+ Gβγ + GDP + ThrombinR			0.1	[300]
5	ThrombinR-active + Gqα_Gβγ_GDP = ThrombinR-active-Gqα_Gβγ_GDP	0.68	0.0060		[300]
6	ThrombinR-active-Gqα_Gβγ_GDP + GTP -> Gqα_GTP + Gβγ + GDP + ThrombinR			0.08	[300]
7	RGS + Gqα_GTP = RGS-Gqα_GTP	0.8	1.0E-5		[37]
8	RGS-Gqα_GTP -> RGS + Gqα_GDP			0.88	[37]
9	ThrombinR-active -> degradation			0.002	[37]
10	G12α_GTP -> G12α_GDP			0.005	[37]
11	Gqα_GTP ->Gqα_GDP			0.0133	[37]
12	Gβγ+ G12α_GDP = G12α_Gβγ_GDP	0.01	0		[37]
13	Gβγ+ Gqα_GDP = Gqα_Gβγ_GDP	0.01	0		[37]

	<i>Rho activation</i>				
14	$\text{RhoGEF} + \text{G12}\alpha_GTP = \text{RhoGEF-G12}\alpha_GTP$	1.0	0.1		[37]
15	$\text{RhoGEF-G12}\alpha_GTP = \text{G12}\alpha_GDP + \text{RhoGEF}$	0.0117	0		[37]
16	$\text{Rho_GDP} + \text{RhoGEF-G12}\alpha_GTP = \text{RhoGEF-G12}\alpha_GTP\text{-Rho_GDP}$	7.6	0.0010		[37]
17	$\text{RhoGEF-G12}\alpha_GTP\text{-Rho_GDP} \rightarrow \text{Rho_GTP} + \text{RhoGEF-G12}\alpha_GTP$			0.07	[37]
18	$\text{Rho_GTP} + \text{RhoGAP} = \text{Rho_GTP-RhoGAP}$	8.0E-4	1.0E-4		Estimated
19	$\text{Rho_GTP-RhoGAP} \rightarrow \text{Rho_GDP} + \text{RhoGAP}$			0.08	Estimated
20	$\text{Rho-GTP} \rightarrow \text{Rho-GDP}$			3.6E-4	[37]
21	$\text{Rho-GDP} + \text{RhoGDI} = \text{Rho-GDP-RhoGDI}$	5.0	0.05		[38]
22	pROCK \rightarrow degradation			8.0E-4	[37]
	<i>Ca²⁺ release</i>				
23	$\text{PLC}\beta + \text{Ca}^{2+} = \text{PLC}\beta\text{-Ca}^{2+}$	0.8	0.001		[37]
24	$\text{PLC}\beta\text{-Ca}^{2+} + \text{PIP2} = \text{PLC}\beta\text{-Ca}^{2+}\text{-PIP2}$	0.8	0.0001		[37]
25	$\text{PLC}\beta\text{-Ca}^{2+}\text{-PIP2} \rightarrow \text{IP3} + \text{PLC}\beta\text{-Ca}^{2+} + \text{DAG}$			0.8	[37]
26	$\text{PLC}\beta\text{-Ca}^{2+} + \text{Gq}\alpha_GTP = \text{PLC}\beta\text{-Ca}^{2+}\text{-Gq}\alpha_GTP$	1	1.0E-5		[37]
27	$\text{PLC}\beta\text{-Ca}^{2+}\text{-Gq}\alpha_GTP + \text{PIP2} = \text{PLC}\beta\text{-Ca}^{2+}\text{-Gq}\alpha_GTP\text{-PIP2}$	0.7	4.0E-4		[37]
28	$\text{PLC}\beta\text{-Ca}^{2+}\text{-Gq}\alpha_GTP\text{-PIP2} \rightarrow \text{PLC}\beta\text{-Ca}^{2+}\text{-Gq}\alpha_GTP + \text{IP3} + \text{DAG}$			0.68	[37]
29	$\text{PLC}\beta + \text{Gq}\alpha_GTP = \text{PLC}\beta\text{-Gq}\alpha_GTP$	1.5201	0.5		Estimated
30	$\text{PLC}\beta\text{-Gq}\alpha_GTP + \text{Ca}^{2+} = \text{PLC}\beta\text{-Ca}^{2+}\text{-Gq}\alpha_GTP$	1.6	0.1		[37]
31	$\text{PLC}\beta\text{-Ca}^{2+}\text{-Gq}\alpha_GTP = \text{Gq}\alpha_GDP + \text{PLC}\beta\text{-Ca}^{2+}$	0.0133	0		[37]
32	$\text{DAG} = \text{PC}$	0.15	0		[37]
33	$\text{IP3} = \text{Inositol}$	0.01	0		[37]
34	$\text{IP3R} + \text{IP3} + \text{IP3} + \text{IP3} = 3\text{IP3-IP3R}$	0.5	10		[37]
35	$\text{ER.Ca}^{2+}_store + \text{IP3-IP3R} = \text{Ca}^{2+}\text{-IP3-IP3R}$	17.0	0.0010		Estimated
36	$\text{Ca}^{2+}\text{-IP3-IP3R} \rightarrow \text{IP3-IP3R} + \text{Ca}^{2+}$			3	[28]
37	$\text{Ca}^{2+}_extleak + \text{Ca}^{2+}_ext = \text{Ca}^{2+}_ext1$	0.0080	1.0E-4		[28]
38	$\text{Ca}^{2+}_ext1 \rightarrow \text{Ca}^{2+} + \text{Ca}^{2+}_extleak$			0.5	[28]
39	$\text{ER.Ca}^{2+}_store + \text{Ca}^{2+}_intleak = \text{Ca}^{2+}_int1$	5.0	1.0E-4		[28]
40	$\text{Ca}^{2+}_int1 \rightarrow \text{Ca}^{2+}_intleak + \text{Ca}^{2+}$			0.5	[28]
41	$\text{Ca}^{2+}_pump + \text{Ca}^{2+} = \text{Ca}^{2+}_pump\text{-Ca}^{2+}$	6.0	1.0E-4		[28]
42	$\text{Ca}^{2+}_pump\text{-Ca}^{2+} \rightarrow \text{Ca}^{2+}_pump + \text{Ca}^{2+}_ext$			10.0	[28]

Chapter 3 Mathematical Model of Endothelial Permeability Signalling

43	$2Ca^{2+} + Ca^{2+}_{trunsp} = Ca^{2+}_{trunsp} - 2Ca^{2+}$	10.0	0.3		[28]
44	$Ca^{2+}_{trunsp} - 2Ca^{2+} + Ca^{2+}_{trunsp} = ER.Ca^{2+}_{store}$	5.0	1.0		[28]
45	$CaM + Ca^{2+} = Ca^{2+} - CaM$	10.0	45.0		[37]
46	$Ca^{2+} - CaM + Ca^{2+} = CaM - 2Ca^{2+}$	8.0	40.0		Estimated
47	$CaM - 2Ca^{2+} + Ca^{2+} = CaM - 3Ca^{2+}$	10.0	170.0		[37]
48	$CaM - 3Ca^{2+} + Ca^{2+} = CaM - 4Ca^{2+}$	10.0	500		[37]
49	$Ca^{2+} - CaM + MLCK = MLCK - Ca^{2+} - CaM$	0.03	0.08		[37]
50	$CaM - 2Ca^{2+} + MLCK = MLCK - 2Ca^{2+} - CaM$	0.04	0.15		[37]
51	$CaM - 3Ca^{2+} + MLCK = MLCK - 3Ca^{2+} - CaM$	1.0	7.0E-4		[37]
52	$CaM - 4Ca^{2+} + MLCK = MLCK - 4Ca^{2+} - CaM$	10.0	0.01		[37]
	PKC activation				
53	$PKC + Ca^{2+} = PKC - Ca^{2+}$	0.3	0.01		Estimated
54	$PKC - Ca^{2+} + DAG = PKC - Ca^{2+} - DAG$	0.3	0.0010		Estimated
55	$PKC - Ca^{2+} - DAG = PKC_{active}$	1.78	0.01		Estimated
56	$PKC_{active} = degradation$			4.6E-4	[316]
57	$PKC_{active} + CPI-17 = PKC_{active} - CPI-17$	3.2	0.01		[37]
58	$PKC_{active} - CPI-17 \rightarrow PKC_{active} + pCPI-17$			1.68	[37]
59	$pCPI-17 = CPI-17$	0.5	0.0010		[37]
60	$MYPT1_{PPase} + pCPI-17 = pCPI-17 - MYPT1_{PPase}$	7.2	0.62		[37]
61	$MYPT1_{PPase} + CPI-17 = CPI-17 - MYPT1_{PPase}$	0.01	0.0010		[37]
62	$pCPI-17 - MYPT1_{PPase} = CPI-17 - MYPT1_{PPase}$	0.0050	0.0010		[37]
	MLC activation				
63	$MLC + pROCK = pROCK - MLC$	2.02	0.0010		[37]
64	$pROCK - MLC \rightarrow pMLC + pROCK$			1.35	[37]
65	$pMLC + pROCK = pROCK - pMLC$	1.38	0.0010		[37]
66	$pROCK - pMLC \rightarrow pROCK + ppMLC$			0.07	[37]
67	$MLC + Rho-GTP-ROCK = Rho-GTP-ROCK - MLC$	0.05	1.0E-4		[37]
68	$Rho-GTP-ROCK - MLC \rightarrow pMLC + Rho-GTP-ROCK$			0.06	[37]
69	$pMLC + Rho-GTP-ROCK = Rho-GTP-ROCK - pMLC$	1.36	2.0E-4		[37]
70	$Rho-GTP-ROCK - pMLC \rightarrow ppMLC + Rho-GTP-ROCK$			1.6	[37]
71	$MLCK - Ca^{2+} - CaM + MLC = MLCK - Ca^{2+} - CaM - MLC$	0.8	1.0E-4		[37]

Chapter 3 Mathematical Model of Endothelial Permeability Signalling

72	$\text{MLCK-Ca}^{2+}\text{-CaM-MLC} \rightarrow \text{MLCK-Ca}^{2+}\text{-CaM} + \text{pMLC}$			0.35	[37]
73	$\text{MLCK-Ca}^{2+}\text{-CaM} + \text{pMLC} = \text{MLCK-Ca}^{2+}\text{-CaM-pMLC}$	0.082	$1.0\text{E-}4$		[37]
74	$\text{MLCK-Ca}^{2+}\text{-CaM-pMLC} \rightarrow \text{ppMLC} + \text{MLCK-Ca}^{2+}\text{-CaM}$			0.25	[37]
75	$\text{MLCK-2Ca}^{2+}\text{-CaM} + \text{MLC} = \text{MLCK-2Ca}^{2+}\text{-CaM-MLC}$	0.5	0.0060		[37]
76	$\text{MLCK-2Ca}^{2+}\text{-CaM-MLC} \rightarrow \text{MLCK-2Ca}^{2+}\text{-CaM} + \text{pMLC}$			2.0	[37]
77	$\text{MLCK-2Ca}^{2+}\text{-CaM} + \text{pMLC} = \text{MLCK-2Ca}^{2+}\text{-CaM-pMLC}$	1.0	0.1		[37]
78	$\text{MLCK-2Ca}^{2+}\text{-CaM-pMLC} \rightarrow \text{MLCK-2Ca}^{2+}\text{-CaM} + \text{ppMLC}$			1.6	[37]
79	$\text{MLCK-3Ca}^{2+}\text{-CaM} + \text{MLC} = \text{MLCK-3Ca}^{2+}\text{-CaM-MLC}$	0.8	$1.0\text{E-}5$		[37]
80	$\text{MLCK-3Ca}^{2+}\text{-CaM-MLC} \rightarrow \text{pMLC} + \text{MLCK-3Ca}^{2+}\text{-CaM}$			0.3	[37]
81	$\text{MLCK-3Ca}^{2+}\text{-CaM} + \text{pMLC} = \text{MLCK-3Ca}^{2+}\text{-CaM-pMLC}$	0.4	$1.0\text{E-}4$		[37]
82	$\text{MLCK-3Ca}^{2+}\text{-CaM-pMLC} \rightarrow \text{MLCK-3Ca}^{2+}\text{-CaM} + \text{ppMLC}$			1.2	[37]
83	$\text{MLCK-4Ca}^{2+}\text{-CaM} + \text{MLC} = \text{MLCK-4Ca}^{2+}\text{-CaM-MLC}$	0.5	0.05		[37]
84	$\text{MLCK-4Ca}^{2+}\text{-CaM-MLC} \rightarrow \text{pMLC} + \text{MLCK-4Ca}^{2+}\text{-CaM}$			1.5	[37]
85	$\text{MLCK-4Ca}^{2+}\text{-CaM} + \text{pMLC} = \text{MLCK-4Ca}^{2+}\text{-CaM-pMLC}$	1.0	0.02		[37]
86	$\text{MLCK-4Ca}^{2+}\text{-CaM-pMLC} \rightarrow \text{ppMLC} + \text{MLCK-4Ca}^{2+}\text{-CaM}$			1.35	[37]
87	$\text{MLCK} + \text{MLC} = \text{MLCK-MLC}$	0.16	0.01		Estimated
88	$\text{MLCK-MLC} \rightarrow \text{pMLC} + \text{MLCK}$			0.6	Estimated
89	$\text{MLCK} + \text{pMLC} = \text{MLCK-pMLC}$	0.55	$1.0\text{E-}5$		[37]
90	$\text{MLCK-pMLC} \rightarrow \text{ppMLC} + \text{MLCK}$			0.2	[37]
91	$\text{MYPT1_PPase} + \text{ppMLC} = \text{MYPT1_PPase-ppMLC}$	0.15	$1.0\text{E-}4$		[37]
92	$\text{MYPT1_PPase-ppMLC} \rightarrow \text{MYPT1_PPase} + \text{pMLC}$			0.2	[37]
93	$\text{MYPT1_PPase} + \text{pMLC} = \text{MYPT1_PPase-pMLC}$	4.5	0.0010		[37]
94	$\text{MYPT1_PPase-pMLC} \rightarrow \text{MLC} + \text{MYPT1_PPase}$			1.68	[37]
95	$\text{pMYPT1_PPase} + \text{ppMLC} = \text{pMYPT1_PPase-ppMLC}$	0.08	$1.0\text{E-}4$		[37]
96	$\text{pMYPT1_PPase-ppMLC} \rightarrow \text{pMLC} + \text{pMYPT1_PPase}$			0.3	[37]
97	$\text{pMYPT1_PPase} + \text{pMLC} = \text{pMYPT1_PPase-pMLC}$	0.25	$1.0\text{E-}4$		[37]
98	$\text{pMYPT1_PPase-pMLC} \rightarrow \text{MLC} + \text{pMYPT1_PPase}$			0.7	[37]
99	$\text{MYPT1_PPase} + \text{Rho-GTP-ROCK} = \text{Rho-GTP-ROCK-MYPT1_PPase}$	1.0	0.0050		Estimated
100	$\text{Rho-GTP-ROCK-MYPT1_PPase} \rightarrow \text{pMYPT1_PPase} + \text{Rho-GTP-ROCK}$			2.5	Estimated
101	$\text{MYPT1_PPase} + \text{pROCK} = \text{pROCK-MYPT1_PPase}$	1.43	$1.0\text{E-}4$		[37]
102	$\text{pROCK-MYPT1_PPase} \rightarrow \text{pMYPT1_PPase} + \text{pROCK}$			0.72	[37]

164	ppMLCK -> degradation			8.0E-4	[303]
	<i>Activation of Ca²⁺ and PKC by VEGF</i>				
165	PLC_γ + Ca ²⁺ = PLC_γ-Ca ²⁺	0.8	0.0010		[37]
166	PIP2 + PLC_γ-Ca ²⁺ = PLC_γ-Ca ²⁺ -PIP2	0.78	1.0E-4		[37]
167	PLC_γ-Ca ²⁺ -PIP2 -> IP3 + DAG + PLC_gamma-Ca ²⁺			0.8	[37]
168	PLC_γ-Ca ²⁺ + VEGF-pVEGFR2-2 = PLC_γ-Ca ²⁺ -VEGF-pVEGFR2-2	1.5	0.01		[37]
169	Ca ²⁺ + VEGF-pVEGFR2-2- PLC_γ = VEGF-pVEGFR2-2-PLC_γ-Ca ²⁺				[37]
170	PIP2 + PLC_γ-Ca ²⁺ -VEGF-pVEGFR2-2 = PLC_γ-Ca ²⁺ -VEGF-pVEGFR2-2-PIP2	0.78	4.0E-4		[37]
171	VEGF-pVEGFR2-2-PLC_gamma-Ca ²⁺ -PIP2 -> VEGF-pVEGFR2-2-PLC_gamma-Ca ²⁺ + IP3 + DAG			0.5	[37]
172	PLC_γ + VEGF-pVEGFR2-2 = VEGF-pVEGFR2-2- PLC_γ	1.52	0.1		Estimated
	<i>NO-PKG-activation</i>				
173	VEGF-pVEGFR2-2 + eNOS = VEGF-pVEGFR2-2-eNOS	55.0	1.0E-4		[321]
174	VEGF-pVEGFR2-2-eNOS -> eNOS-active + VEGF-pVEGFR2-2			0.25	[321]
175	H1R-active + eNOS = H1R-active-eNOS	55.0	1.0E-4		[321]
176	H1R-active-eNOS -> H1R-active + eNOS-active			0.6	[321]
177	eNOS-active + L-Arg = eNOS-active-L-Arg	20.0	0.01		Estimated
178	eNOS-active-L-Arg -> L-Cit + eNOS-active + NO			1.2	[321]
179	NO + GC = NO-GC	10.0	0.1		[321]
180	NO-GC + GTP = NO-GC-GTP	2.11	0.05		[321]
181	NO-GC-GTP -> NO-GC + cGMP			0.8	[321]
182	cGMP + PKG = cGMP-PKG	0.68	0.0010		[321]
183	cGMP-PKG -> PKG-active + cGMP			0.06	[321]
184	PKG-active + Raf = PKG-Raf	0.22	0.0010		Estimated
185	PKG-Raf -> PKG-active + pRaf			0.21	[37]
186	NO -> NO1			0.15	[321]
187	NO -> NO2			0.25	[321]
188	NO -> NO3			0.25	[321]
189	PKG-active -> degradation			0.0010	[37]
	<i>Activation of Gqa and Ca²⁺ by histamine</i>				
190	Histamine + H1R = Histamine-H1R	3	0.01		[282]
191	Histamine-H1R -> H1R-active + Histamine			0.8	[282]

192	H1R -> degradation			0.0022	[282]
193	H1R-active -> degradation			0.0133	[282]
194	H1R-active + Gqα_Gβγ_GDP = Gqα_Gβγ_GDP -H1R-active	0.68	0.0006		[282]
195	GTP + Gqα_Gβγ_GDP -H1R-active -> GDP + Gqα_GTP + Gβγ_GDP			0.08	[282]
196	Rho-GTP + ROCK-fold = ROCK-open + Rho-GTP	25	0		[322]
197	ROCK-open + Rho-GTP = Rho-GTP-ROCK-open	0.05	0.0005		[322]
198	Rho-GTP-ROCK-open -> pROCK + Rho-GTP			0.085	[37]
199	PKG-active + ROCK-open = PKG-ROCK-open	0.35	0.0001		[37]
200	PKG-ROCK-open -> pROCK + PKG			1.85	[37]
Initial Concentration					
ID	Component	Concentration		References (PubMed ID)	
1	Pro_thrombinR (PAR-1)	0.05		[130]	
2	RGS	0.2		[130]	
3	GTP	50		[130]	
4	GDP	5		[130]	
5	IP3R	0.33		Estimated	
6	RhoGEF	0.25		Estimated	
7	RhoGAP	0.15		[320]	
8	PIP2	10.0		[37]	
9	Rho-GDP	0.1		[320]	
10	MLCK	0.69		[320]	
11	PKC	0.2		[37]	
12	CPI-17	0.08		[37]	
13	MLC	4.2		[37]	
14	pMLC	0.6		[37]	
15	Ca ²⁺	0.0083		[37]	
16	CaM	20		[37]	
17	PLCβ	0.57		[37]	
18	Ca ²⁺ _trunsp	20.0		Estimated	
19	Ca ²⁺ _pump	0.08		Estimated	
20	ppMLC	0.8		[37]	

21	MYPT1_PPase	0.4	[37]
22	pMYPT1_PPase	0.01	Estimated
23	G12 α _G β γ _GDP	0.4	[130]
24	Gq α _G β γ _GDP	0.5	[130]
25	Ca ²⁺ _extleak	4.0	Estimated
26	Ca ²⁺ _intleak	0.8	Estimated
27	RhoGDI	0.05	[37]
28	ROCK	0.16	[37]
29	VEGF	0.02	Estimated
30	VEGFR2	0.02	Estimated
31	SHP	0.25	[320]
32	Grb2	0.15	[320]
33	Sos	0.12	[320]
34	RasGDP	0.5	[320]
35	Shc	0.5	[320]
36	Raf	0.5	[320]
37	MEK	0.05	[320]
38	ERK	0.05	[320]
39	Pase	0.5	[320]
40	PP2A	0.02	[320]
41	MKP3	0.01	[320]
42	RasGAP	0.8	[320]
43	PLC_ γ	0.57	[37]
44	eNOS	0.33	[321]
45	GC	0.05	[321]
46	PKG	0.015	[321]
47	L-Arg	0.55	[321]
48	H1R	0.35	[282]

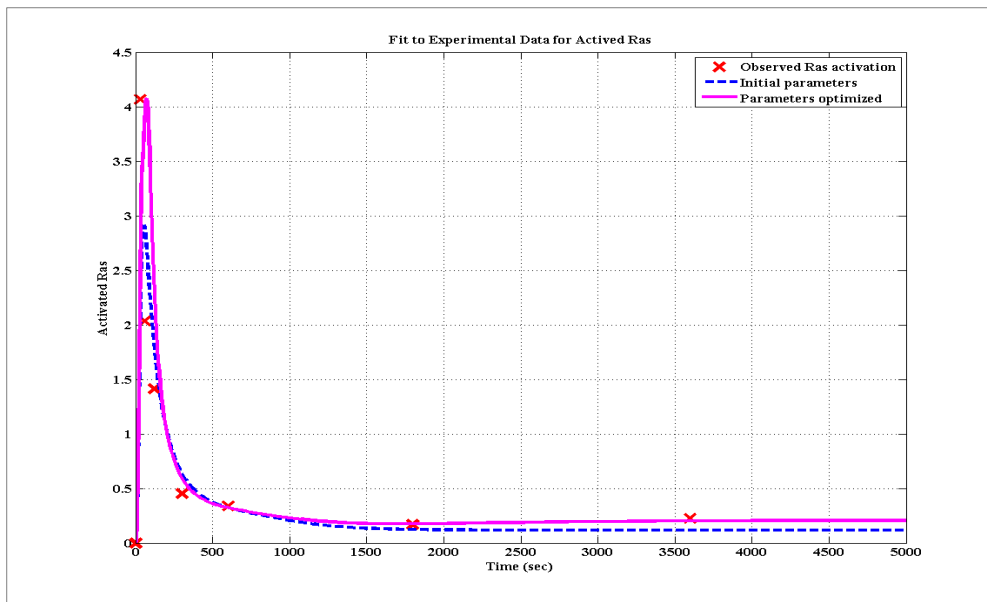
3.3.2 Collection and estimation of kinetic parameters

The types of parameters used in our model are parameters governing protein–protein interactions and catalytic activities. The published simulation studies have found that most parameters are robust and moderate changes do not significantly alter the overall pathway behavior [290, 291, 293]. Apart from the use of the parameters of the published simulation models, additional parameters were obtained from the literature based on the widely used assumption that parameters measured in vitro and in some cell lines are generally applicable in most cases. For those protein-protein interactions without available parameters, their parameters were putatively estimated from the known parameters of the relevant interacting domain profile pairs [323, 324] or other interacting protein pairs of similar sequences. As a biological network is believed to be robust, and protein-protein binding interactions for proteins in similar families that mediate similar types of biochemical reactions (such as Ras and Rho) often differ within a 10-fold range, the values of kinetic parameters obtained from previous models were optimized within this range.

The parameters of the protein-pairs not available from previous models were obtained by the following procedure: The first step in finding the parameters of a specific protein-pair is to search protein-pairs that are both with available parameters and with each individual protein similar in sequence with the respective protein of the studied protein-pair. If one or more such protein-pairs are found, then the average values of the parameters of these protein-pairs are used as the initial parameters of the studied protein-pair, which are further

optimized in ± 10 -fold range with respect to experimental data. For instance, the parameters for the Rho activation cycle were obtained from the Ras activation cycle and were further optimized within a 10-fold range. The cycle of optimization and validation was repeated in order to obtain simulated results that agreed well with known experimental trends. If no such protein-pair is found, we proceed to the second step to search protein-pairs that are both with available parameters and with each individual protein belonging to the same domain family of the respective protein of the studied protein-pair. If one or more such protein-pairs are found, then the average values of the parameters of these protein-pairs are used as the initial parameters of the studied protein-pair, which are further optimized in ± 10 -fold range with respect to experimental data.

Figure 3-6: Fit to experimental data for Ras activation.



The parameters of RhoGAP and PKC related protein-pairs were determined by

the first and second step. The parameters of 14 protein-pairs could not be determined by the first and second step due to lack of experimental data and relevant protein-pairs with known parameters. These parameters were determined by using the trust regions algorithm [325] to fit the simulation results to the experimental data of RAS, ERK, MYPT and CPI-17 activation curves [37]. **Figure 3-5** shows the fitting curve against experimental RAS activation data. The level of fitting is based on the least-squares method and the fitting process proceeds in iterations until the R-square value is >0.6 [326]. In each iteration, the parameter values derived from the previous iteration were used as the starting parameters for further optimization.

3.3.3 Model Optimization, Validation and Parameter Sensitivity Analysis

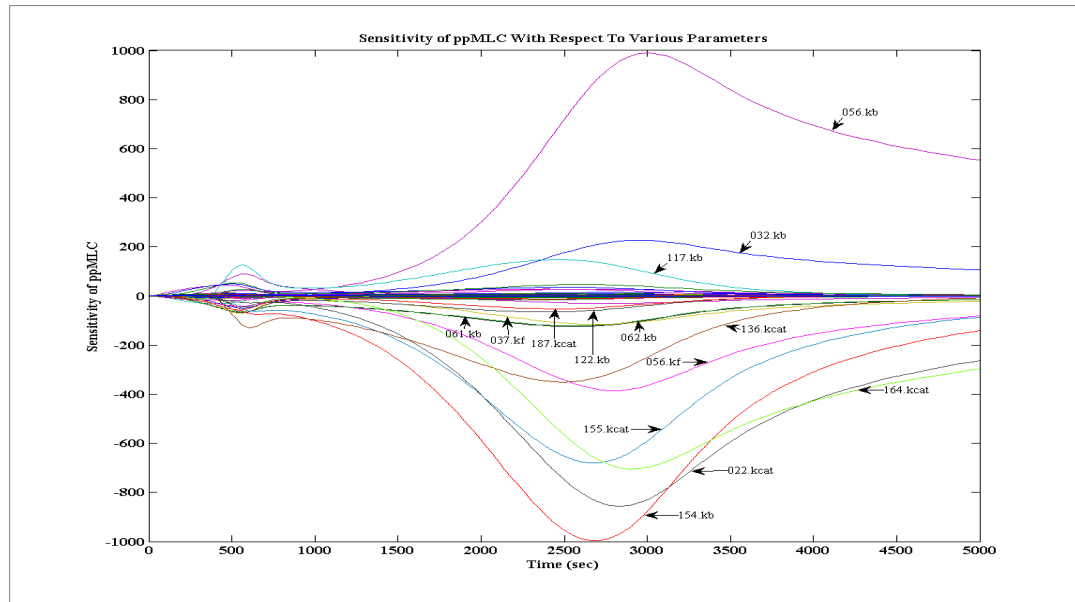
Mathematical models developed by ‘one-set-fits-all’ generic parameters need not reproduce quantitative behavior in all systems, but may be able to reproduce the behavior or trend for specific systems. For instance, a mathematical model developed for a biological pathway from parameters obtained experimentally from one cell type can behave slightly different in another cell types. Differences in model behavior between cell types can be due to the presence or absence of crosstalk (i.e., differences in model topology) and variation in effective values of kinetic parameters. Hence, in this study, we developed a generic model of the thrombin-, VEGF-, and histamine-MLC signaling pathway to investigate the role of these three main mediators in regulating MLC activation. The simulated results are represented as

trajectories of concentrations of chemical species with respect to time that are validated against available experimental data. If the trend or dynamic behavior of a particular reactant or product behaves in such a way that is consistent with the experimental data, then the model is said to be reasonable and can be used to analyze and predict unknown biological phenomena within some difficulty to define range of conditions. If the simulation results are not in agreement with known experimental facts, then the model has to be revisited to examine possible errors, such as incorrect interaction kinetics or values of kinetic parameters. Optimized parameters obtained from previous mathematical models are not necessarily optimized in a new study as the scope of these models can be different. The cycle of optimization and validation are repeated in order to obtain simulated results that agree well with known experimental trends.

The sensitivity of the simulation results with respect to the optimized and other parameters need to be systematically analyzed to determine if the model is sufficiently robust to be able to analyze and predict the true dynamic behavior of biological networks without the artifact of parameters. Differential analysis of parameter sensitivity, also referred to as the direct method, was utilized to compute the time-dependent sensitivities of all the species states with respect to each parameter values in the model [327]. Complex-step derivative approximation [328] was used to calculate numerical derivatives of the reactions in the model to achieve near analytical accuracy, robustness and easy implementation. We used sensitivity analysis function of Matlab to

conduct sensitivity analysis. The sensitivity value of ppMLC with respect to all parameters in the model was provided in Additional File and **Figure 3-6**. As shown in Additional File, **Figure 3-6**, only 14 (4%) kinetic parameters including CPI-17, PKC and ROCK related reactions showed some sensitivity in affecting the output. The majority of the parameters are insensitive in affecting the output. Thus, our model can be considered as sufficiently robust.

Figure 3-7: Parameter sensitivity analysis



3.3.4 Estimation of kinetic parameters

We estimated the values of unknown parameters in the model when some parameters cannot be determined from direct experiments or the literatures. Unknown or only roughly known parameters were estimated by minimizing the discrepancy between the experiment data and model simulation. These parameters were determined by using the trust regions algorithm [325] to fit

the simulated to the experimental data of RAS, ERK, MYPT and CPI-17 activation curves [37]. The level of fitting is based on the least-squares method and the fitting process proceeds in iterations until the R-square value is >0.6 .

The procedure can be summarized as follows:

1. Import target experiment data
2. Simulate the model with the rough parameter values
3. Compute R-Square value for the simulation and experiment data before parameter estimation
4. Set up the parameters to estimate and the state to match.
5. Use the current values of parameters in the model as the starting point for optimization
6. Simulating the model with the new estimate parameters and computing R-Square value for comparison.
7. Plot the results

For example, we estimated parameters by fit our simulation results to experiment data of Activated Ras [320]. The experiment data, simulation results before and after parameter estimation were shown as **Figure 3-6**. The fitness has been improved with R-Square value from 0.4508 to 0.6088.

3.4 Results and discussion

3.4.1 Model validation with experimental studies of the regulation of MLC activation, calcium release, and Rho activation by thrombin

Our simulation model was first validated by determining whether the simulation results were consistent with experimental observations of MLC activation and calcium release by the single mediator thrombin. Thrombin-mediated processes were investigated computationally by zeroing out the initial concentration of VEGF and histamine. It has been observed that MLC activation increases from low initial levels to $39\% \pm 2\%$, $66\% \pm 10\%$, $68\% \pm 13\%$, $64\% \pm 13\%$, and $67 \pm 9\%$ of the MLC population at 30s, 60s, 2.5 min, 15 min, and 30 min after thrombin stimulation, respectively, which subsequently drops to 48% at 60 min [329]. The amplitude of MLC activation has been found to correlate linearly with the strength of endothelial cell contraction [330, 331]. As illustrated in **Figure 3-7** (Left), our simulated time-dependent MLC activation levels are in fair agreement with this observation (the simulation results for the first 20 min are also shown in **Figure 3-8**). Our simulations showed that the amplitude of MLC activation reaches two peaks, the first at ~ 2.5 min and the main peak at ~ 30 min, which is compared to the observation that the levels of active MLC levels at 2.5 min and 30 min are higher than those at 15 min and 60 min [329]. Our analysis suggested that these two peaks arise primarily from the Ca^{2+} -dependent and Rho GTPase/ROCK-dependent mechanisms, respectively, as described below.

Figure 3-8: Simulated time course and experimental data of thrombin-mediated MLC activation (left) and calcium release (right).

⊞ denotes experimentally measured MLC activation at 30s ($39\% \pm 2$), 60s ($66\% \pm 10\%$), 2.5min ($68\% \pm 13\%$).

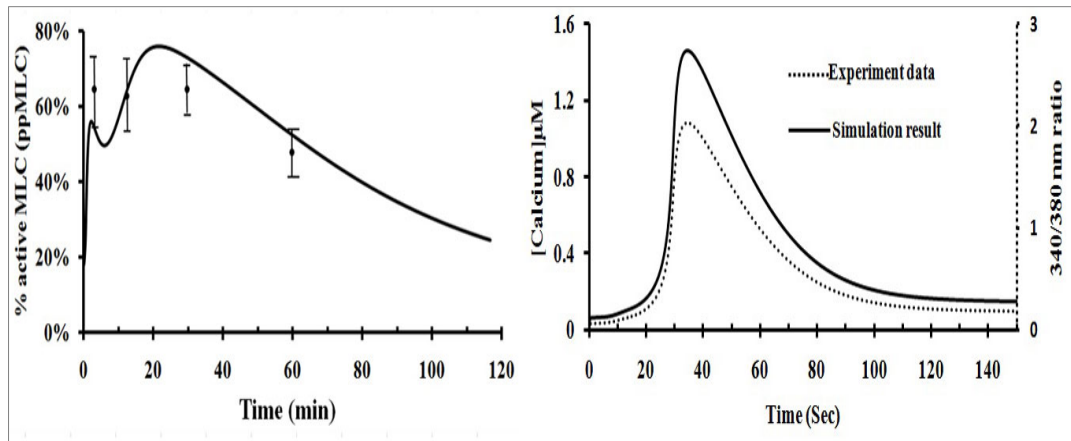
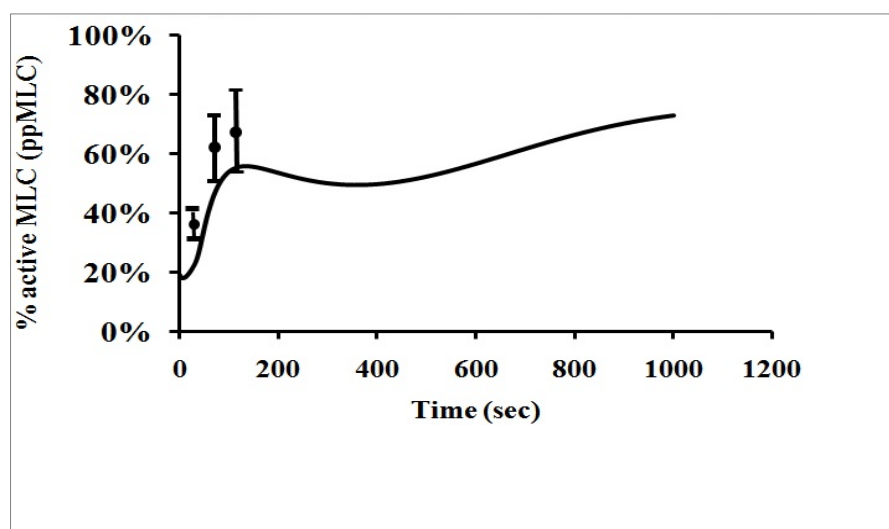


Figure 3-9: Simulated time course and experimental data of thrombin-mediated MLC activation in the first 20 min.




┆ denotes experimentally measured MLC activation at 30s (39%±2), 60s (66%±10%), 2.5min (68%±13%).

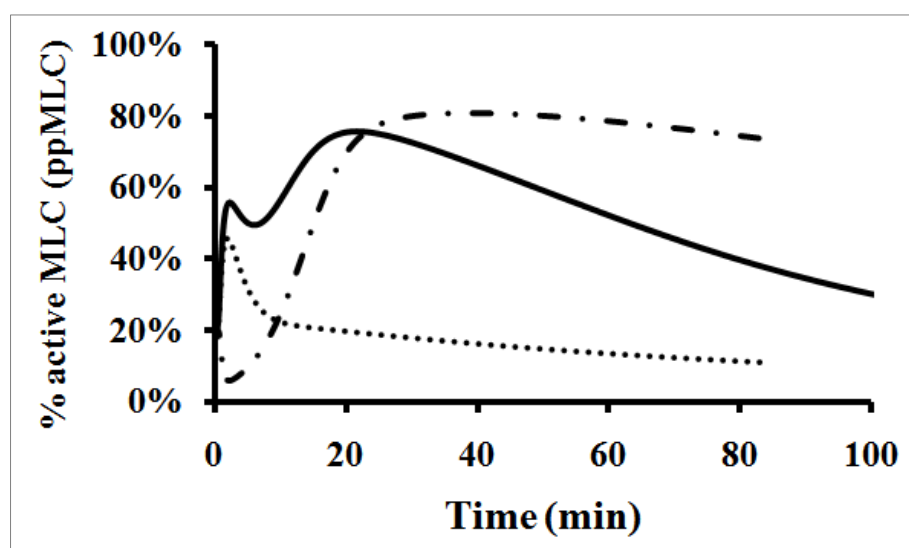


Elevation in cytosolic Ca^{2+} concentration ([332]) is a common initial response of endothelial cells to various changes such as the exposure to hormonal and inflammatory stimuli and variation of physical conditions [333]. Jeng *et al.* [334] have shown that the binding of thrombin and PAR-1 induces rapid calcium mobilization and increase of $[Ca^{2+}]_i$, with $[Ca^{2+}]_i$ peaking at 30-40 s followed by a rapid drop. The simulated calcium release profile in **Figure 3-7** (Right) exhibits a peak concentration at 38 s followed by a rapid decay, consistent with Jeng *et al.*'s experiment results. The increased intracellular Ca^{2+} influx is expected to enhance the binding of Ca^{2+} to CaM, which subsequently activates MLCK to phosphorylate the MLC of myosin II. To evaluate which

signaling event is primarily responsible for the large transient increase in the level of MLC activation (the first peak at ~2.5 min in the left **Figure 3-7**), we systematically varied the strength of protein-protein interactions upstream of MLC. As shown in **Figure 3-9**, the first peak disappears when the Ca^{2+} -dependent MLC activation (Reaction 73-86) was switched off, while that peak remains largely intact when the ROCK-dependent MLC activation and CPI-17-MYPT interactions were switched off (Reactions 57-58, 63-70, 99-102), Therefore, our analysis suggests that this Ca^{2+} -dependent mechanism was primarily responsible for the large transient increase of the levels of MLC activation.

Figure 3-10: Simulated time course of thrombin-mediated MLC activation in terms of different components.

The curve  ,  and  represents the signaling from the complete pathway (Control), the Ca^{2+} -dependent component (with ROCK-dependent MLC activation and P-CPI-17-MYPT interaction switched off, Reactions 57-58, 63-70, 99-102), and the non- Ca^{2+} -dependent component (with Ca^{2+} -dependent MLC activation switched off, Reactions 71-86) respectively.



Thrombin induces a prolonged increase of endothelial permeability lasting for 1-1.5 hr. This prolonged elevated permeability is attributed to the activation of the small Rho GTPase and Rho kinase [335, 336]. It has been found that Rho GTPase activation can be observed after 2 min and the elevated activation is maintained up to 60 min after thrombin stimulation, and the time course of Rho GTPase activation correlates well with the time course of MLC activation increase by **Figure 3-10** is consistent with this observation, which shows that the simulated Rho GTPase activation was maintained for 60 min. Rho GTPase activation induces MLC activation via both direct and indirect routes. Rho GTPase and ROCK directly activate MLC to subsequently induce the contraction of the non-muscle cell systems [304, 337]. In the indirect route, ROCK inhibits myosin phosphatase activity by phosphorylating the myosin binding subunit (MBS) of myosin phosphatase [312], which increases the activation level of MLC, actomyosin interaction, stress fiber formation, and subsequent endothelial permeability. We studied whether these direct and

indirect Rho GTPase -dependent mechanisms are primarily responsible for the sustained levels of MLC activation (the main peak at ~30 min in the left **Figure 3-7**) by systematically varying the protein-protein interactions upstream of MLC. As shown in **Figure 3-11**, this peak remains largely intact when the Ca^{2+} -dependent MLC activation and P-CPI-17-MYPT interaction (Reaction 57-58, 71-76) were switched off, but disappeared when the ROCK-dependent MLC activation (Reactions 63-70, 99-102) were switched off. Therefore, our analysis suggests that both the direct and indirect Rho GTPase -dependent mechanisms play an important role for the sustained levels of MLC activation.

Figure 3-11: Simulated time course and experimental results of thrombin-mediated Rho GTPase activation in units of percentage of initial Rho concentration.

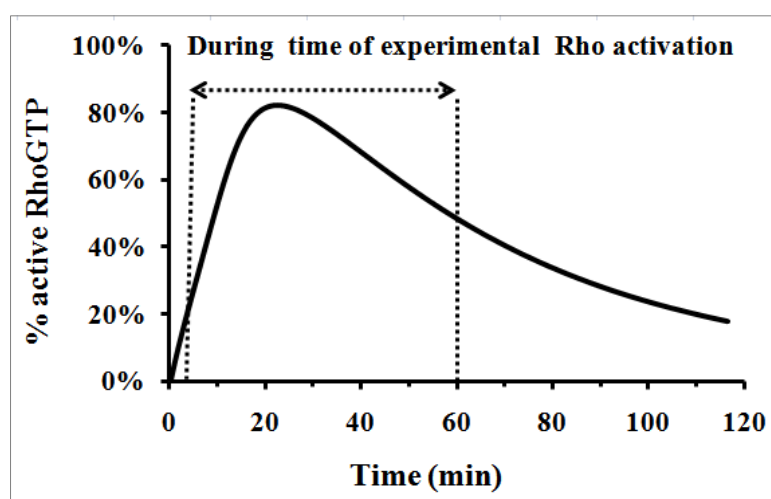
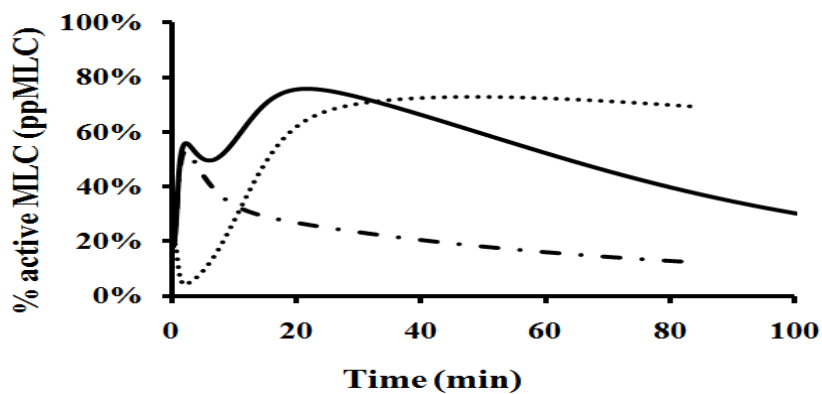


Figure 3-12: Simulated time course of thrombin-mediated MLC activation in terms of different components.

The curve **————**, **.....** and **- . -** represents the signalling from the complete pathway (Control), ROCK-dependent component (with Ca²⁺-dependent MLC activation and P-CPI-17-MYPT interaction switched off, Reactions 57-58, 71-76), and the non- ROCK-dependent component (with ROCK-dependent MLC activation switched off, Reactions 63-70, 99-102) respectively.



3.4.2 Model validation with experimental studies of MLC activation and ERK activation by VEGF

Our simulation model was also validated by determining whether the simulation results are consistent with experimentally observed regulation of MLC activation as well as ERK and MLCK activation by another mediator VEGF. These VEGF-mediated processes were simulated by using our model with thrombin and histamine switched off by setting their initial concentrations to zero values. It has been reported that injection of VEGF

induces vascular leakage in 5 min, and the leakage peaks in 15-20 min and then diminishes after 30 min [338]. As shown in **Figure 3-12** (Left), the simulated duration of MLC activation is about 30 min with the first peak at 2.5 min and the main peak at 15 min. The 15 min time range of the main peak of MLC activation is consistent with the reported 15-20 min time range for VEGF-induced vascular leakage to reach its peak [338]. While we have not found an experimental finding to support the true existence of the first peak exhibited by our simulation, it is noted that the time of the first peak matches the experimentally determined on-set time of VEGF-induced vascular leakage [338]. As described in the previous section, the first peak of MLC activation at ~2.5 min in **Figure 3-12** (Left) was induced mainly by Ca^{2+} -dependent mechanism. We further investigate which signalling event is primarily responsible for the main peak at ~15 min. As shown in **Figure 3-13**. We found that this peak remained when NO-dependent MLC activation was switch off (Reactions 179-185) but disappeared when Ras-Raf-ERK-dependent MLC activation was switch off (Reactions 152-163). This suggests that the main peak is induced by Ras-dependent ERK activation. As shown in **Figure 3-12** (Right), the simulated ERK activation peaks at about 7 min and decays within 25 min, which is consistent with the observation that the amount of phosphorylated ERK-1/2 reaches maximum value at 5-10 min after administration of VEGF and decreases back to the control level 30 min afterward [37].

Figure 3-13: Simulated time course and experimental result of VEGF-mediated MLC activation (left) and ERK activation (right)

Thrombin and histamine level set at zero values. The shaded area in the left figure indicates the time range in which VEGF-induced vascular leakage reaches its peak in experimental studies (Ref 57). The shaded area in the right figure indicates the time range in which the amount of ERK-1/2 activation reaches maximum value after VEGF administration (Ref 58). The VEGF concentrations were set as 0.02 μM .

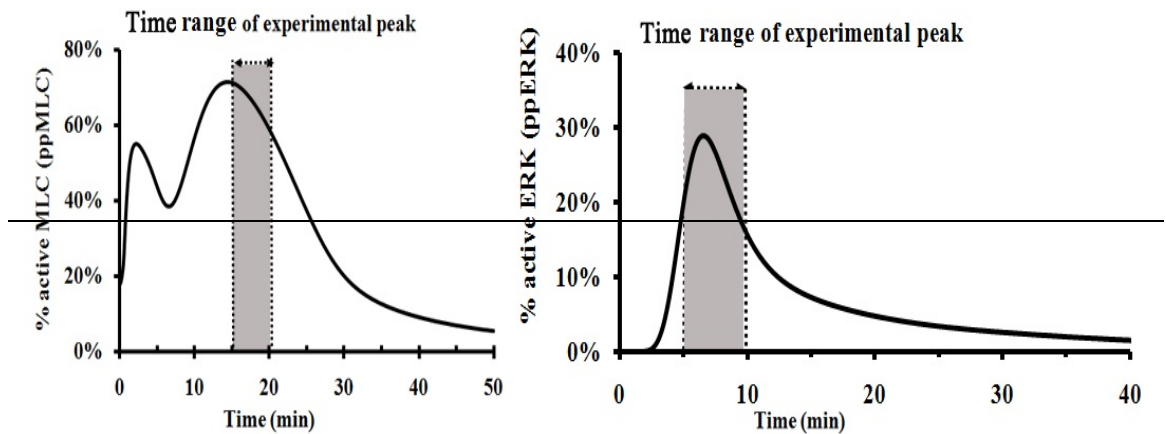
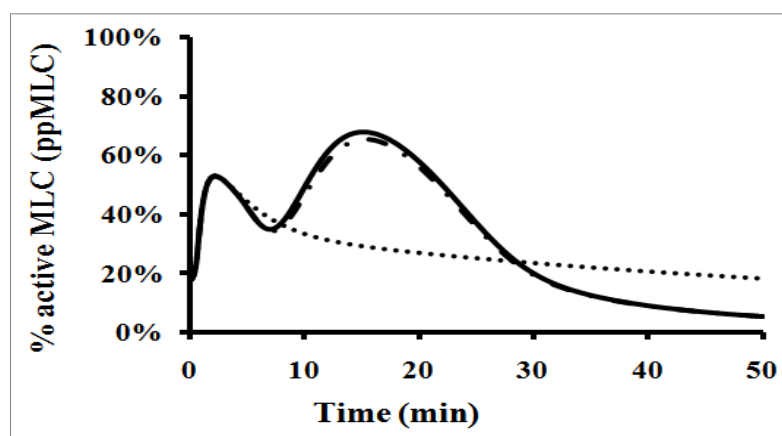


Figure 3-14: Simulated time course of VEGF-mediated MLC activation in terms of different components.

The curve **————**, **.....** and **- . -** represents the signaling from the complete pathway (Control), non-ERK-dependent component (with Ras-Raf-ERK-dependent MLC activation switched off, Reactions 152-163), and the non- NO-dependent component (with NO- dependent MLC activation switched off Reactions 179-185) respectively.

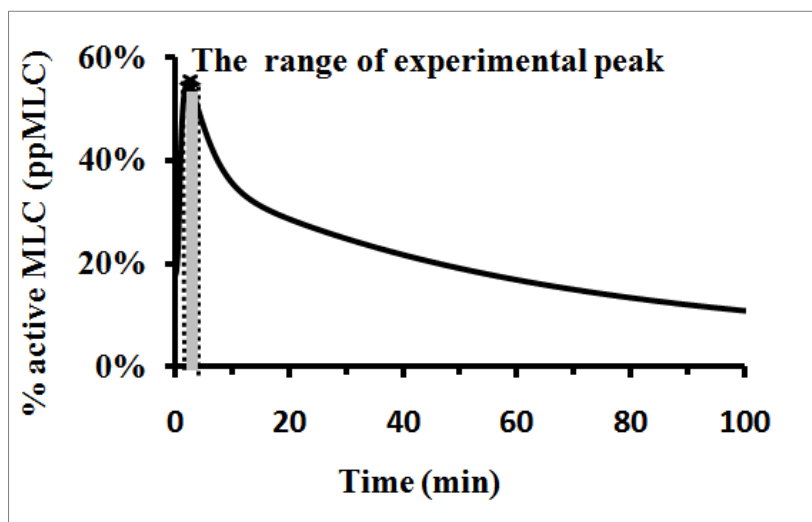


3.4.3 Model validation with experimental studies of MLC activation by histamine

The model was further validated by determining whether the simulation results are consistent with experimentally observed regulation of MLC activation by the third individual mediator histamine. This histamine-mediated process was simulated by using our model with thrombin and VEGF switched off by setting their initial concentrations to zero values. The simulation results in **Figure 3-14** indicated that histamine causes a transient increase of MLC activation that peaked at 2.5 min, which is consistent with the experimental

finding that histamine induces a transient endothelial permeability peaked at 2-5 min [282]. Further investigation showed that this peak is primarily induced by Ca²⁺-dependent mechanism and the contribution from the NO-dependent ERK activation path is very small, as shown in **Figure 3-15** by switching off each individual path. Moreover, the contribution from the NO-dependent ERK activation path is much weaker compared with Ras-dependent ERK activation and MLC activation by the individual mediator VEGF.

Figure 3-15: Simulated time course and experimental result of Histamine-mediated MLC activation in units of percentage of initial MLC concentration with thrombin and VEGF level set at zero values. The shaded area indicates the time range in which histamine has been experimentally found to induce a transient endothelial permeability. The histamine concentrations were taken as 0.005 μ M.



3.4.4 Comparison of the simulated thrombin-mediated IP3 and Ca²⁺ release with that of an existing model

The thrombin signalling cascade of our model is very similar to that of Maeda et al. that has been developed a computational model of thrombin-regulated ROCK pathway [293]. Hence, it is appropriate to compare the simulation results of our model with Maeda's model. In their studies, they measured and simulated thrombin-mediated IP3 and Ca²⁺ release. We therefore compared our simulated IP3 and Ca²⁺ release with their results. As shown in **Figure 3-16**, our simulation showed essentially the same transit IP3 release and Ca²⁺ release patterns as those presented in Maeda's studies.

Figure 3-16: Simulated time course of Histamine-mediated MLC activation in terms of different components.

The curve **————**, **.....** and **- . -** represents the signaling from the complete pathway (Control), non- Ca²⁺-dependent component (with Ca²⁺-dependent MLC activation switched off, Reaction 71-86), and the non-NO-dependent component (with NO- dependent MLC activation switched off, Reaction 179-185) respectively.

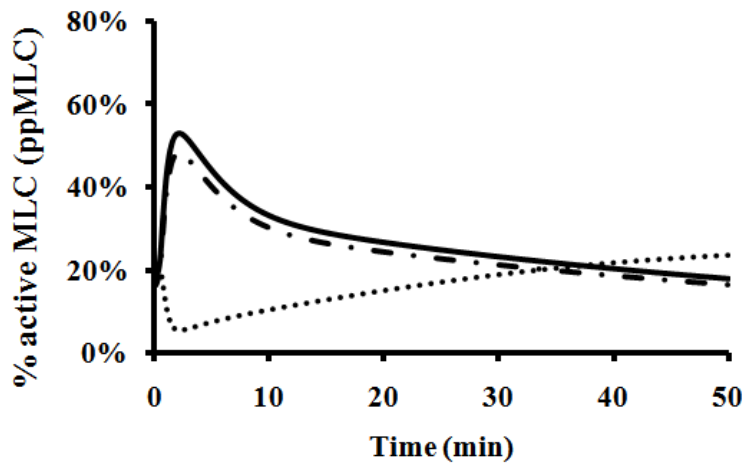
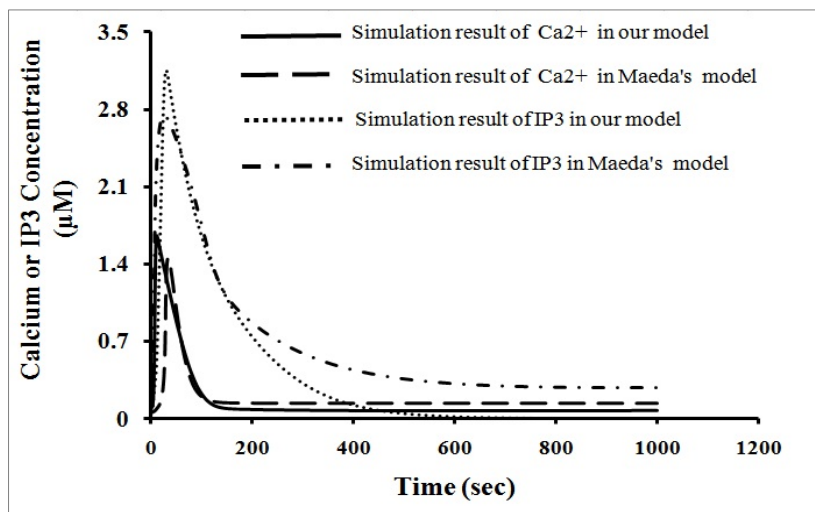
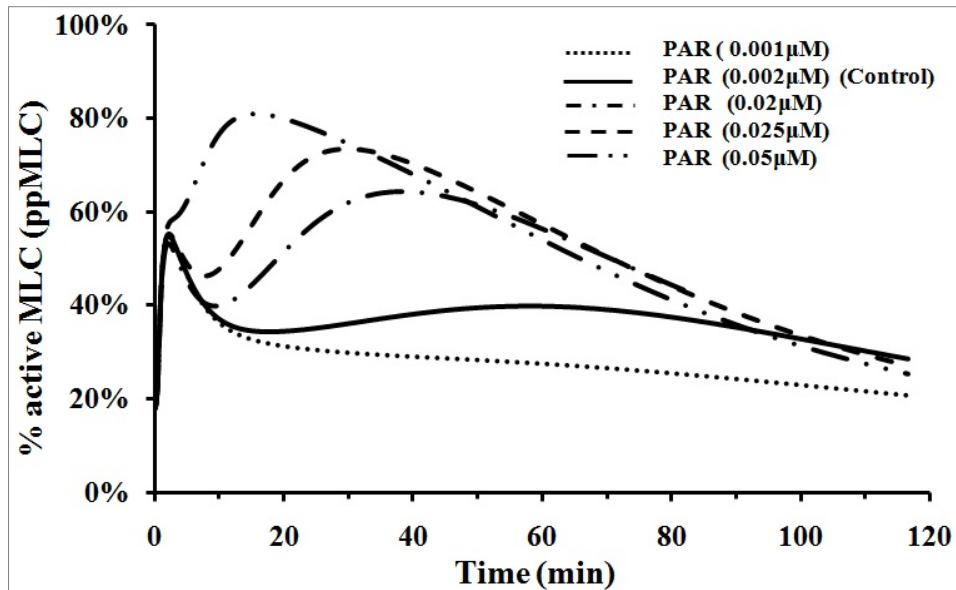


Figure 3-17: Comparison of simulation result of Ca^{2+} and IP3 in our model and Maeda's model



3.4.5 Simulation of the effects of thrombin receptor PAR-1 over-expression on thrombin-mediated MLC activation

PAR-1 is the major thrombin-activated receptor involved in platelet aggregation, endothelial permeability, and tumor cell migration. Activated PAR-1 is coupled via several members of the heterotrimeric G-proteins, $G_{\alpha 12/13}$ and $G_{\alpha q}$, to transduce a substantial network of signalling pathways [300]. It has been reported that during atherogenesis, PAR-1 expression is enhanced in regions of inflammation associated with macrophage influx, smooth muscle cell proliferation, and an increase in mesenchymal-like intimal cells [339]. It is of interest to quantitatively evaluate the effects of PAR-1 elevation on thrombin-mediated MLC activation. We further used our model to simulate thrombin mediated MLC activation at different PAR-1 levels with VEGF and histamine switched off [340]. Our simulation results, in **Figure 3-17**, showed that PAR-1 at elevated levels significantly increases the amplitude of MLC activation and reduces the time for MLC activation to reach the main peak. There is a direct correlation between the level of PAR-1 expression and the degree of invasiveness of breast carcinoma cell lines [37], in which endothelial permeability is one of the prerequisites for cancer invasiveness as it facilitates cell transmigration and plasma accumulation in the matrix to support new vessel formation [340]. Therefore, this experiment indicated that PAR-1 over-expression leads to enhanced endothelial hyper-permeability, and our simulation results are in good agreement with this experimental finding.

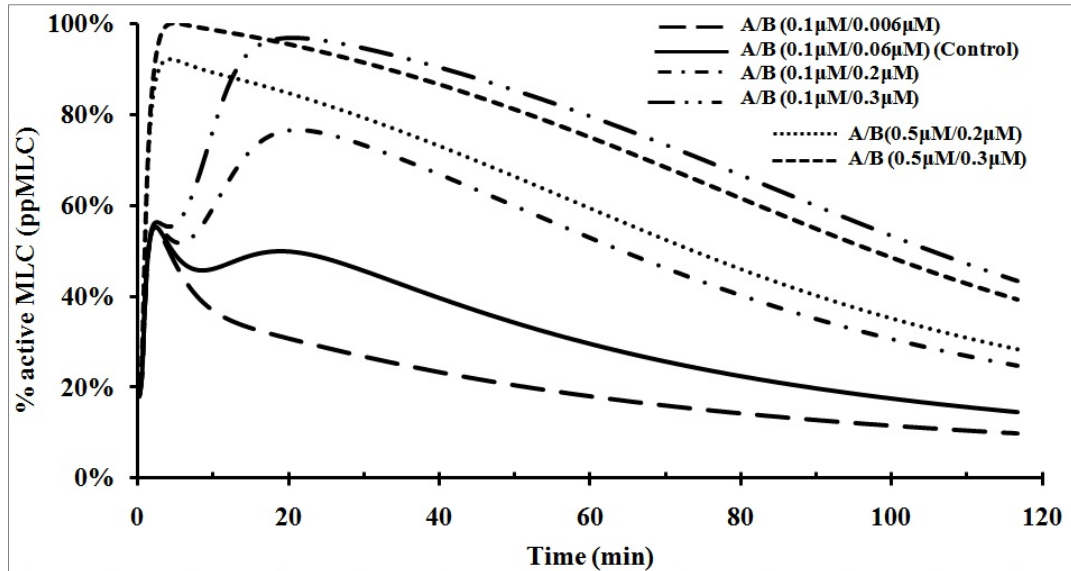
Figure 3-18: ppMLC activation at different PAR-1 concentrations

3.4.6 Simulation of the effects of Rho GTPase and ROCK over-expression on thrombin-mediated MLC activation

Rho GTPase and ROCK in endothelial cells have been found to be elevated in hypoxia [341]. Over-expression of dominant activated Rho GTPase/ROCK in NIH3T3 cells results in an increase of MLC activation [312]. Over-expressed ROCK in human brain microvascular endothelial cells has been found to induce endothelial permeability and to significantly increase the transmigration rate of NCI-H209 cells through the human brain microvascular endothelial cells [342]. The effects of elevated Rho GTPase and ROCK on thrombin-mediated MLC activation were quantitatively evaluated by using our model with VEGF and histamine switched off. As shown in **Figure 3-18**, an increased ROCK level with Rho GTPase at control level significantly

enhanced the amplitude of activation of MLC in a dose-dependent manner. When ROCK and Rho GTPase levels were simultaneously elevated, the amplitude of MLC activation was significantly increased and the time to reach the activation peak was reduced. Rho GTPase and ROCK are abundant in lymph nodes with metastasis, and the ability to enter either blood or lymphatic vasculature is necessary for tumor cells to metastasize to distant sites [343]. Furthermore, Rho GTPase and ROCK reportedly are required in both endothelial and migrating cells for them to cross the vascular endothelium [344, 345]. Thus, by quantifying the effect of Rho GTPase /ROCK, we can gain more insight into the mechanism of sustained MLC activation, which may aid the search for and evaluation of new therapeutic strategies for the prevention and treatment of endothelial hyper-permeability and cancer metastasis-related diseases.

Figure 3-19: MLC activation at different Rho GTPase (A) and ROCK (B) concentrations.



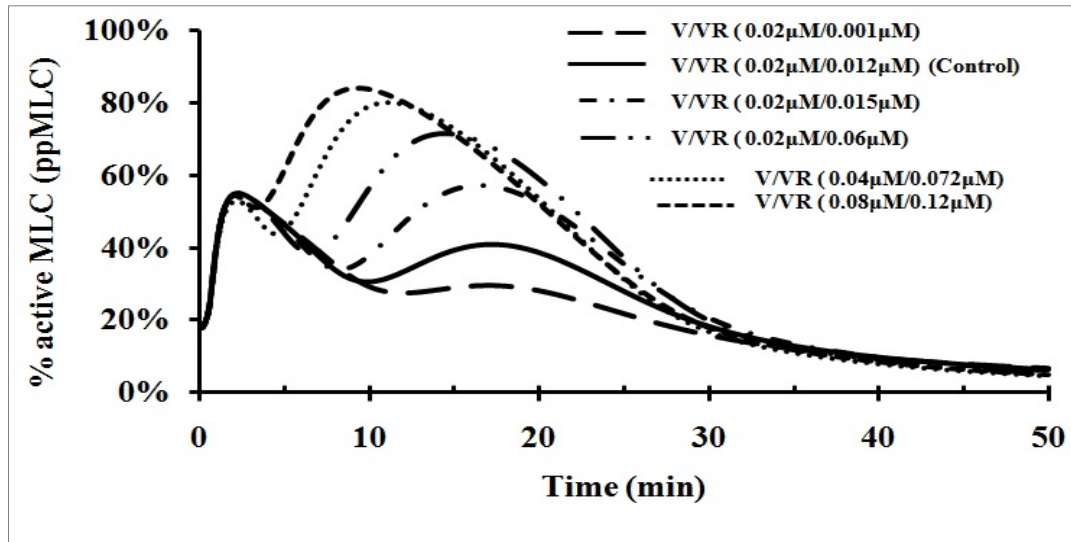
3.4.7 Simulation of effects of VEGF and VEGFR2 over-expression on VEGF-mediated MLC activation

VEGFR2 is recognized as the principal mediator of physiological and pathological effects of VEGF on endothelial cells, which include proliferation, migration, survival, and permeability [346]. The expression of VEGF and VEGFR2 in endothelial cells has been found to be elevated in oxidative stress [347], type 1 leprosy reaction [348], and during diabetes to induce microvascular complications, especially diabetic retinopathy [349]. Over-expression of VEGF and VEGFR2 has been shown to correlate with increased risk of metastatic disease and overall poor prognosis in different

carcinomas [350]. Apart from their primary functions in angiogenesis, the roles of VEGF and VEGFR2 in metastasis likely involve the regulation of endothelial permeability to facilitate cell transmigration and plasma accumulation in the matrix in support of new vessel formation [351]. The effects of VEGF and VEGFR2 over-expression on VEGF-mediated MLC activation were quantitatively evaluated by using our model with thrombin and histamine switched off.

As shown in **Figure 3-19**, the increased amount of VEGFR2 with VEGF at control level significantly enhanced MLC activation. For instance, the small increase of VEGFR2 concentration from 0.010 to 0.012 μM increased the amplitude of the main peak of MLC activation by 15%, suggesting that MLC activation was very sensitive to VEGFR2 concentration. When VEGF and VEGFR2 levels were simultaneously increased, the amplitude of MLC activation was further increased by a significant amount with respect to that when only VEGFR2 was over-expressed. This is consistent with the observed correlation of VEGF and VEGFR2 over-expression with increased risk of metastatic disease and overall poor prognosis in different carcinomas [350].

Figure 3-20: MLC activation at different VEGF(V) and VEGFR2 (VR) concentrations

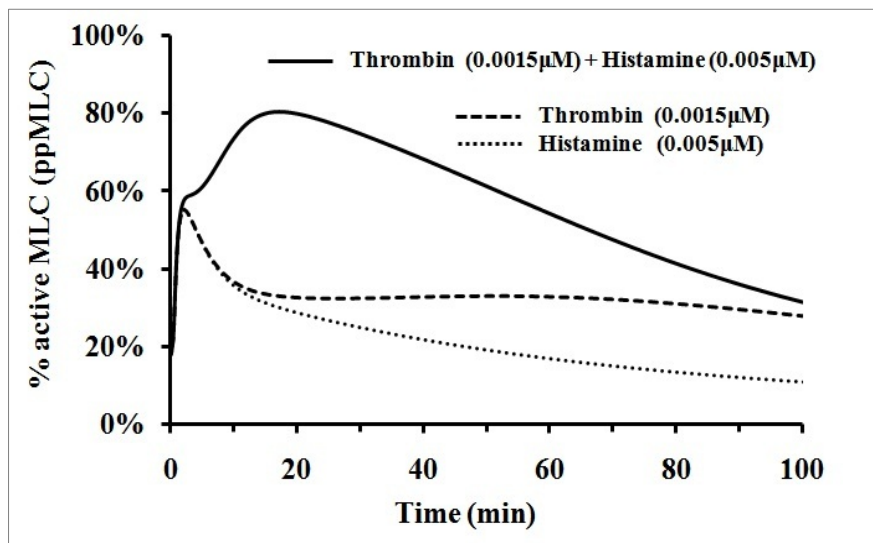


3.4.8 Simulation of synergistic activation of MLC by thrombin and histamine

It has been reported that combination of low concentrations of stimuli of thrombin and histamine induces more significantly enhanced endothelial permeability than the simple sum of the permeability change induced by each mediator alone [352]. The effect of the combination of low concentrations of thrombin and histamine on MLC activation was explored by using our model with the third mediator VEGF switched off. As illustrated in **Figure 3-20**, from 10 min to 50 min after stimulation with combination of 0.0015 μM thrombin and 0.0050 μM histamine, the amplitude of MLC activation reached levels of >65%, which is greater than the simple sum of <35% and <22% when only one individual mediator, thrombin and histamine, respectively, was switched

on. Therefore, our simulation results indicated a synergistic effect of histamine and thrombin, in agreement with observations [352]. Moreover, the levels of MLC activation induced by these low concentrations of thrombin and histamine are comparable to (higher than) those induced by individual mediator thrombin and histamine at concentrations of $0.0500 \mu\text{M}$ and $0.005 \mu\text{M}$, respectively, which suggests that the synergistic effect is at a substantial level.

Figure 3-21: MLC activation induced by combination of thrombin and histamine stimuli.



The level of synergistic effect can be more clearly revealed by the comparison of the areas under the thrombin and histamine induced MLC activation curve with those of thrombin-induced and histamine-induced MLC activation curves at different 10 min time intervals in Figure 8, which are provided in **Table 3-2**.

In particular, the level of synergistic effect can be reflected by the difference between the area under the thrombin and histamine induced curve and the simple sum of the areas under the thrombin-induced and histamine-induced curves, with positive values corresponding to synergistic effect (better than simple sum of thrombin-induced and histamine-induced activation). From Table 1, the largest synergistic effect occurs in the 10-20 min, 20-30 min and 30-40 min time ranges with net area gain of 1.3, 1.8 and 1.5 (corresponding to an average of 13%, 18% and 15% more amount of activated MLC with respect to that of simple sum of thrombin-induced and histamine-induced activation).

Table 3-2: Comparison of the areas with respect to different time ranges in Figure 3-20

Comparison of the areas under the thrombin and histamine induced MLC activation curve with those of thrombin-induced and histamine-induced MLC activation curves with respect to different time ranges in **Figure 3-20**

MLC activation curve	Area under MLC activation curve with respect to different time range							
	0-10 min	10-20 min	20-30 min	30-40 min	40-50 min	50-60 min	60-70 min	70-80 min
Curve 1: Histamine + Thrombin induced activation	5.9	7.8	7.6	6.9	6.1	5.2	4.7	4.0
Curve 2: Histamine-mediated activation	4.3	3.1	2.6	2.3	2.0	1.6	1.4	1.3
Curve 3: Thrombin-mediated activation	4.4	3.4	3.2	3.1	3.1	3.0	3.0	2.9
Simple sum of curve 2 and 3	8.7	6.5	5.8	5.4	5.0	4.7	4.5	4.3
Area difference between curve 1 and simple sum of curve 2 and 3	-2.8	1.3	1.8	1.5	1.1	0.6	0.2	-0.3

As shown in **Figure 3-20**, the synergistic effect at low concentrations of thrombin and histamine only occur during the time range from 10 min to 50 min. Before and after this time range, the level of MLC activation by thrombin

+ histamine is less than the simple sum of that by thrombin and histamine alone. The less than additive effect during the first 10 min is primarily due to the time-dependent behavior of MLC activation by the Ca^{2+} -dependent signalling cascade. The transient MLC activation curve by the Ca^{2+} -dependent cascade is largely the same for the thrombin, histamine, and thrombin + histamine mediated processes (**Figure 3-21**, **Figure 3-22**, **Figure 3-23** solid line). It is thus not difficult to understand that the simple sum of the level of MLC activation by thrombin and histamine alone is superficially larger than that by thrombin + histamine. The less than additive effect after 50 min is primarily due to the variation of time-dependent behavior of MLC activation by the ROCK-dependent signalling cascade. The level of MLC activation slowly rises to significant levels without decay in the presence of thrombin alone for up to 100 min (**Figure 3-21**, dotted and dash-dotted line). On the other hand, the MLC activation level rises slowly to moderate levels without decay in the presence of histamine alone (**Figure 3-22**). In contrast, the MLC activation level quickly rises to high levels and rapidly decays to low levels after 50 min in the presence of thrombin + histamine, the signalling strength thus becomes less than additive after 50 min (**Figure 3-23**, dash-dotted line).

Figure 3-22: The contribution of Ca²⁺- dependent, ROCK-dependent and CPI-17-dependent signaling cascade to thrombin-mediated MLC activation at low concentration of thrombin (0.0015 μM).

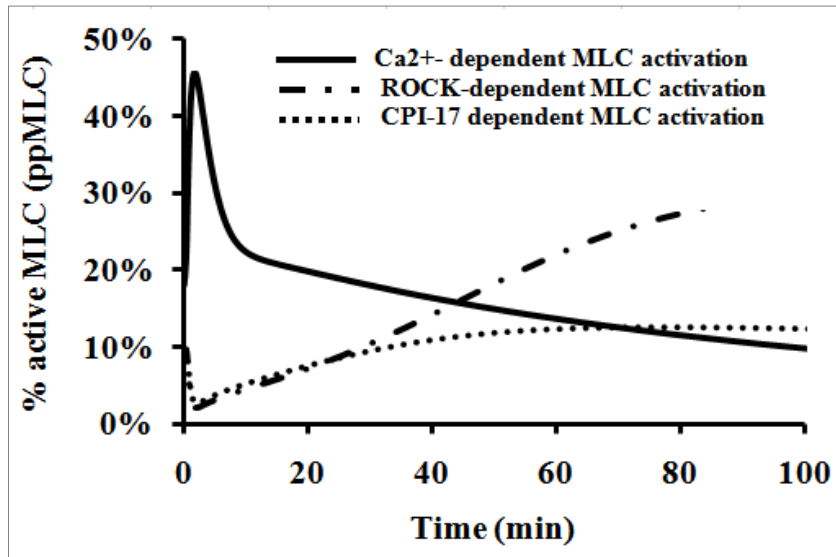


Figure 3-23: The contribution of Ca²⁺- dependent, NO-dependent and CPI-17-dependent signaling cascade to histamine-mediated MLC activation at low concentration of histamine (0.005 μM).

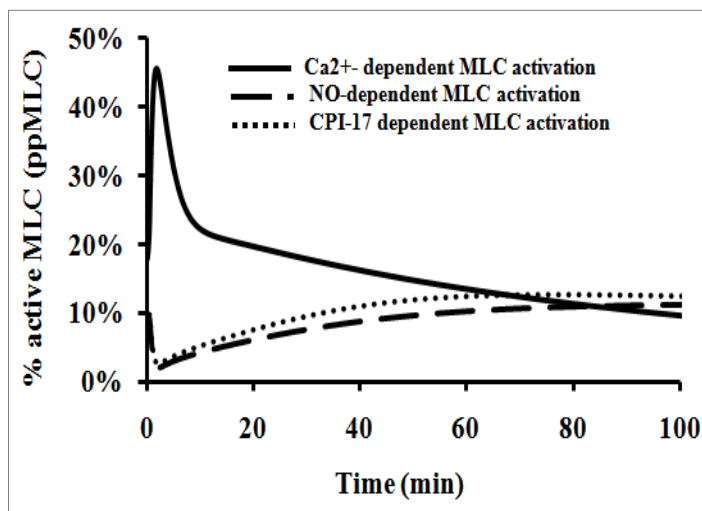
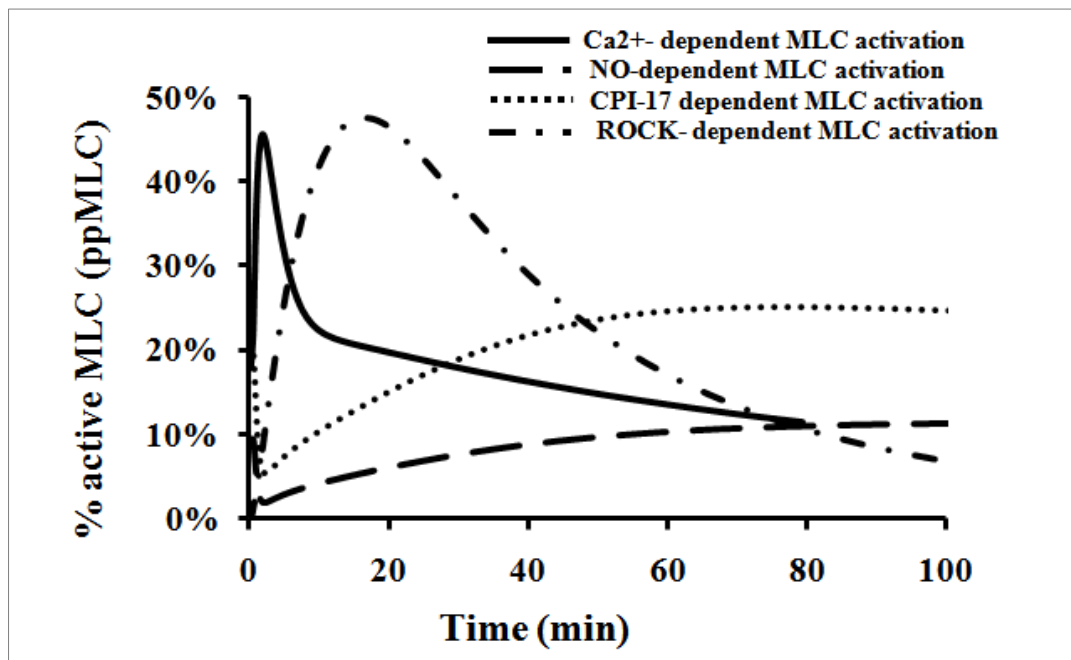


Figure 3-24: The contribution of Ca^{2+} - dependent, NO-dependent and CPI-17-dependent cascade to thrombin + histamine mediated MLC activation at low concentration of thrombin ($0.0015 \mu\text{M}$) and histamine ($0.005 \mu\text{M}$).



The underlying mechanism of the significant synergistic effect during 10-50min time period can be elucidated from the perspective of network regulation. MLC activation is regulated by at least four signalling cascades Ca^{2+} -dependent, CPI-17-dependent, NO-dependent, and ROCK-dependent cascades. As shown in **Figure 3-21**, **Figure 3-22**, **Figure 3-23**, The MLC activation curve induced by the Ca^{2+} -dependent cascade is roughly the same for the thrombin, histamine and thrombin + histamine mediated processes. The level of MLC activation induced by the CPI-17-dependent, NO-dependent cascade in the presence of thrombin + histamine is close to the simple sum of that in the presence of thrombin and histamine alone. While the MLC activation induced by the

ROCK-dependent cascade in the presence of thrombin + histamine is at significantly higher levels and shows more transient pattern than that in the presence of thrombin and histamine alone. These differences in signaling behavior lead to synergistic effect within 10 min to 50 min time range.

The different signaling behavior of the Rho-ROCK signaling stimulated by different mediators or mediator combinations primarily arises from the dynamics of ROCK activation [322]. The kinase activity of ROCK is off when ROCK is intra-molecularly folded. ROCK can be activated only after it is unfolded by the binding of Rho GTPase to its Rho-binding domain to disrupt the auto-inhibitory interaction, which subsequently allows such proteins as Rho GTPase and PKG to activate ROCK at phosphorylation site. Hence, in the presence of thrombin + histamine, thrombin-activated Rho GTPase unfolds ROCK to allow histamine-activated PKG to activate ROCK thereby enhancing the level of ROCK activation in combination with thrombin-mediated Rho GTPase activation of ROCK. When stimulated by histamine or VEGF alone, ROCK is in the inactive state and does not contribute to MLC activation. When stimulated by thrombin alone, ROCK is activated only by Rho GTPase without the contribution from PKG, leading to a slower increase and lower peak strength of MLC activation than that in the presence of thrombin + histamine. Such an integrated communication network is expected to enable fine tuning of the strength and duration of MLC activation, thereby enabling fine regulation of physiological responses, including synergistic or more complex effects. Network models have suggested that partial inhibition of a surprisingly small

number of targets can be more efficient than the complete inhibition of a single target [353]. Experimental and simulation studies of synergistic effects of thrombin and histamine on endothelial monolayer permeability may provide useful information for developing multi-target drugs against endothelial permeability and related diseases [353].

3.4.9 Prediction of the collective regulation of MLC activation by thrombin and VEGF

Our simulation model was further used to study the collective regulation of MLC activation by thrombin and VEGF, with a particular focus on whether or not thrombin and VEGF synergistically activate MLC in certain time ranges. Systemic activation of blood coagulation is often present in cancer patients, and thrombin generated during thrombosis can augment malignant phenotypes by activating tumor cell adhesion to platelets and endothelial cells, enhancing tumor cell growth and metastasis, and stimulating tumor cell angiogenesis [277]. Moreover, thrombin promotes VEGF secretion and platelet activation, thus causing a mutual stimulation between endothelial cells and cancer cells [354, 355]. Therefore, the collective effect of thrombin and VEGF on MLC activation and subsequently endothelial hyperpermeability may have substantial influence on the tumor growth and metastasis process in cancer patients with blood coagulation near and at the tumor sites [356].

As shown in **Figure 3-24**, from 15 min to 30 min after stimulation with

combination of 0.002 μM thrombin and 0.010 μM VEGF, the amplitude of MLC activation reached levels of $>62\%$, which is greater than the simple sum of $<32\%$ and $<28\%$ when only one individual mediator, thrombin and VEGF, respectively, was switched on. Therefore, our simulation results indicated a synergistic effect of histamine and VEGF on MLC activation. The level of synergistic effect can be reflected by the difference between the area under the thrombin and VEGF induced curve and the simple sum of the areas under the thrombin-induced and VEGF-induced curves in **Figure 3-24**, which are shown in **Table 3-3**. From **Table 3-3**, the largest synergistic effect occurs in the 20-30 min time range with net area gain of 1.8 corresponding to an average of 18% more amount of activated MLC with respect to simple sum of thrombin-induced and VEGF-induced activation. The high level MLC activation by thrombin and VEGF likely has significant impact on the promotion of cancer metastasis in the cancer patients with blood coagulation near and at the tumor sites. These patients may be more effectively treated by combinations of drugs targeting the VEGF and thrombin signalling pathways [356].

Figure 3-25 : MLC activation induced by the combination of thrombin and VEGF stimuli.

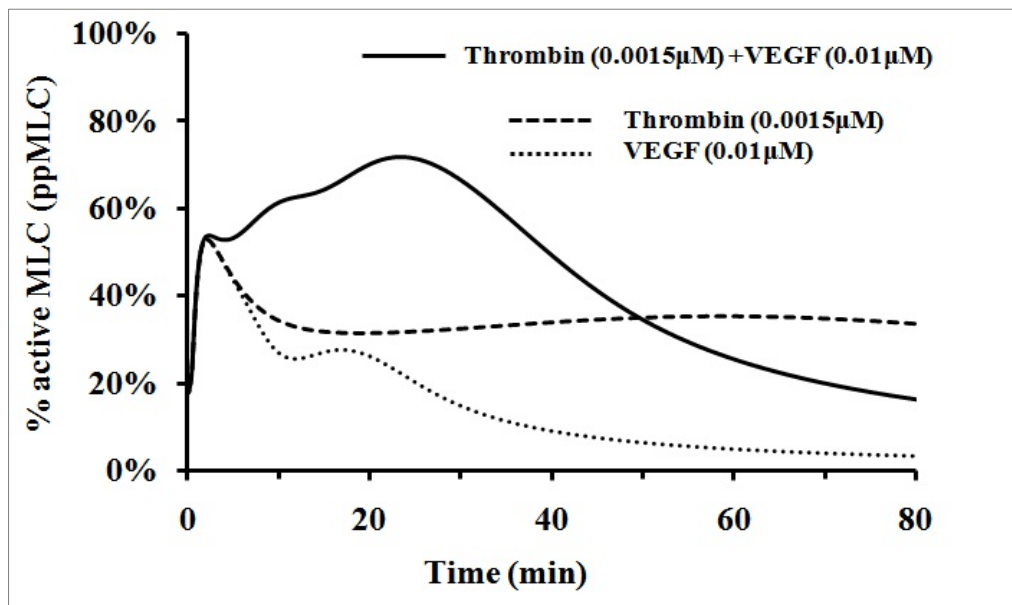


Table 3-3: Comparison of the areas with respect to different time ranges in Figure 9

Comparison of the areas under the thrombin and VEGF induced MLC activation curve with those of thrombin-induced and VEGF-induced MLC activation curves with respect to different time ranges in Figure 9

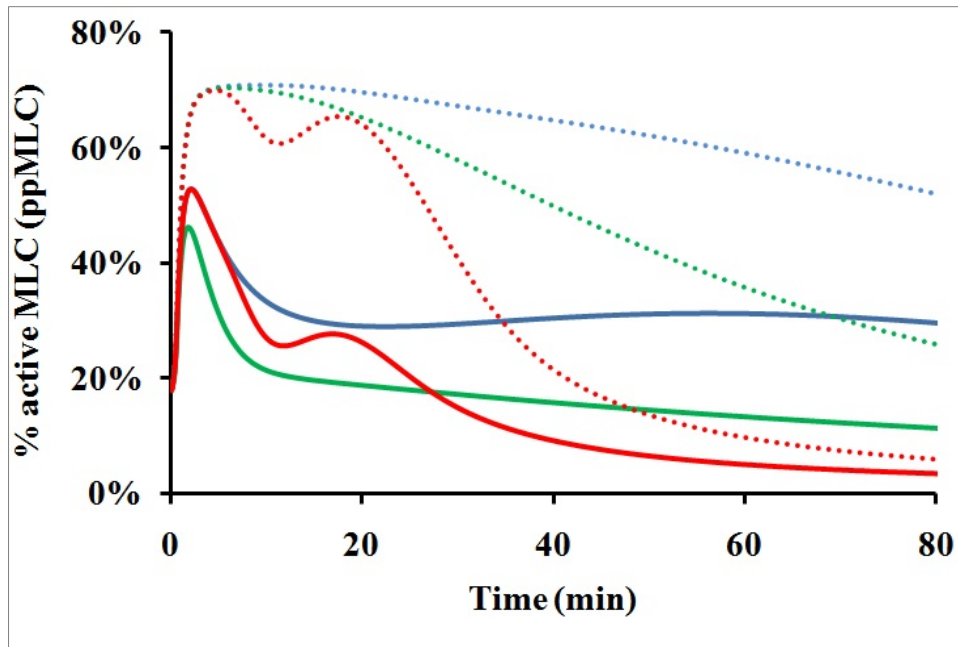
MLC activation curve	Area under MLC activation curve with respect to different time range							
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
	min	min	min	min	min	min	min	min
Curve 1: Thrombin + VEGF induced activation	5.0	6.2	6.7	5.1	3.4	2.4	1.9	1.4
Curve 2: VEGF-mediated activation	4.1	3.5	2.9	3.2	3.3	3.4	3.4	3.1
Curve 3: Thrombin-mediated activation	4.0	2.6	2.0	1.1	0.05	0.5	0.4	0.3
Simple sum of curve 2 and 3	8.1	6.1	4.9	4.3	3.35	3.9	3.8	3.4
Area difference between curve 1 and simple sum of curve 2	-3.1	0.1	1.8	0.7	0.05	-1.5	-1.9	-2.0

3.4.10 Prediction of the effect of CPI-17 over-expression on MLC activation in the presence of lower concentration of thrombin, histamine and VEGF

CPI-17 inhibits MYCP to hinder its dephosphorylation of MLC, leading to increased MLC activation [357]. Altered CPI-17 level is associated with smooth muscle-related diseases, such as intestinal bowel disease [358], asthma [359], pulmonary hypertension [360] and diabetic dysfunction of smooth muscle [361]. It is of interest to evaluate the effect of CPI-17 over-expression on MLC activation, particularly at lower level of thrombin, histamine and VEGF. In this work, CPI-17 over-expression was simulated by 5-fold increase of its level from $0.08\mu\text{M}$ to $0.4\mu\text{M}$ [362]. Each of the thrombin-, histamine- and VEGF- mediated processes was simulated by setting the concentration of thrombin, histamine and VEGF set at lower value of $0.0015\mu\text{M}$, $0.005\mu\text{M}$ and $0.01\mu\text{M}$ respectively with the other two mediators switched off by setting their initial concentrations to zero values. As shown in **Figure 3-24**, CPI-17 over-expression significantly strengthened and prolonged MLC activation to the levels higher than those at normal CPI-17 level and normal concentration of thrombin, histamine and VEGF respectively [363].

Figure 3-26: Prediction of the effect of CPI-17 over-expression on MLC activation at low concentration of stimuli.

Solid and dotted lines correspond to the activation by default (CPI-17 = 0.08 μM) and elevated (CPI-17 = 0.4 μM) concentration of CPI-17. Blue line refers to thrombin stimuli, green line refers to histamine stimuli, and red line refers to VEGF stimuli respectively.



3.5 Conclusion remarks

Thrombin, VEGF, and histamine are hallmarks of endothelial hyper-permeability, which perform their regulatory roles individually and collectively under different disease conditions, and with different dynamic profiles. Thrombin and VEGF can increase microvascular permeability $\sim 50,000$ times more potently than histamine [212]. Thrombin, VEGF, and histamine

induce prolonged (1-1.5 hr), intermediate (15-20 min) and transient (~5 min) increases of endothelial permeability, respectively. An integrated simulation model that includes the signalling of all these hallmark mediators enables more comprehensive analysis of the signalling processes involved in different disease processes and regulated by different combinations of these mediators.

Based on published models of relevant signalling, we developed an integrated mathematical model including the signalling pathways of all three of these mediators. Simulation results from our model were consistent with available experimental data of signalling mediated by both individual mediators and combinations of two mediators, and could be used to interpret the sustained and transient phases of MLC activation. Our model was able to predict the effects of altered pathway components and synergistic combination of multiple mediators, some of which are consistent with experimental findings [352]. Similar to the published models of other pathways, our model can potentially be used to identify important disease genes through sensitivity analysis of signalling components [37]. Our model may also be extended to emphasize other components to facilitate further investigation of the effects of different mediators, cascades, and cross-talk on endothelial permeability and related diseases.

Chapter 4 Sepsis Biomarker selection

This chapter describes endothelial permeability related disease-Sepsis biomarker selection method using microarray data. The Consensus Scoring of Multiple Random Sampling and Gene-Ranking Consistency Evaluation method were used for identifying of stable disease-differentiating signatures. 20 sets of sepsis gene signatures were generated. 41 gene signatures are fairly stable with 69%~93% of all predictor-genes shared by all 20 signatures sets. The predictive ability of the selected signature shared by all of the 20 sets is evaluated by SVM models on an independent dataset collected from GEO Database (GSE28750). The overall accuracy for the 41 predictor-genes was 93.26%. The accuracies for all predictor-genes were in the range of 92.97~94.57%. These results suggest that the selected signatures using our system can perform well in classification of drug sensitivity.

4.1 Introduction

In complex biological systems, in order to understand and explain the mechanisms of various biological processes and their various disease states, it becomes increasingly important to model the various markers for both the healthy state and the disease state in order to enable prompt, accurate and timely diagnosis as well as an appropriate intervention to treat the disease state if it is present.

Sepsis is a leading cause of death in critically ill patients despite the use of

modern antibiotics and resuscitation therapies [364]. The septic response is an extremely complex chain of events involving inflammatory and anti-inflammatory processes, humoral and cellular reactions and circulatory abnormalities [365]. The diagnosis of sepsis and evaluation of its severity is complicated by the highly variable and non-specific nature of the signs and symptoms of sepsis [366]. However, the early diagnosis and stratification of the severity of sepsis is very important, increasing the possibility of starting timely and specific treatment [367].

Disease signatures can have an important place in this process because they can indicate the presence or absence or severity of sepsis [368] and can differentiate bacterial from viral and fungal infection, and systemic sepsis from local infection. Other potential uses of biomarkers include roles in prognostication, guiding antibiotic therapy, evaluating the response to therapy and recovery from sepsis. The simple and direct way to identify disease signatures is through analyzing the change of expression level across a series of samples. There are around 25,000 genes in human genome [369]. Therefore, Microarray becomes a very important tool for disease gene discovery because microarray can measure the gene expression profiles of tens of thousands of genes at one time. By discovering the differences in gene expression between normal and disease tissues, we can focus on the genes with different expressions and those genes that might be activated or inactivated in association with a particular disease.

In this chapter, we explored a new gene selection method aiming at reducing the chances of erroneous elimination of predictor-genes. We employed the recursive feature selection method based on a model built from support vector

machines to identify novel molecular signatures with respect to the interactions among genes. Derived from the consensus scoring of multiple random sampling and the evaluation of gene-ranking consistency embedded in the recursive feature selection system, totally 41 genes were selected after 20 times of experiments. The gene signatures are fairly stable with 69%~93% of all predictor-genes shared by all 20 signatures. To test the prediction of selected signature, the derived signatures were evaluated by independent dataset (GSE28750) which contains 20 sepsis samples and 10 normal people. The differential expression and function analysis of the identified marker genes implies that the selected genes should play important roles in sepsis initiation and progress. For accurate disease diagnosis and proper treatment selection, it is very important to identify the gene markers responsible for disease initiation. Moreover, the discovery of the markers responsible for disease progress is critical because such markers can be used to identify disease stages, subtypes and prognosis effect in an accurate manner. As such, proper treatment regime can be applied and the survivability of the patients can be ultimately extended [370].

4.2 Materials and methods

4.2.1 Sepsis microarray datasets

Two independent data sets of sepsis (GSE13904 and GSE28750) were used for sepsis gene discovery and for validating the effect of our selected genes.

The dataset of GSE13904 contained the expression levels of 18 control and 22 patients. This dataset was obtained by using the Affymetrix Human Genome

U133 Plus 2.0 Array [140]. This array completely covered of the Human Genome U133 Set plus 6,500 additional genes for analysis of over 47,000 transcripts. All probe sets represented on the GeneChip Human Genome U133 Set are identically replicated on the GeneChip Human Genome U133 Plus 2.0 Array. The sequences from which these probe sets were derived were selected from GenBank®, dbEST, and RefSeq. The sequence clusters were created from the UniGene database (Build 133, April 20, 2001) and then refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database (April 2001 release).

In addition, there are 9,921 new probe sets representing approximately 6,500 new genes. These gene sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from the UniGene database (Build 159, January 25, 2003) and refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the NCBI human genome assembly (Build 31).

In order to evaluate the performance of selected genes, the other dataset GSE28750 contains 10 normal people and 20 sepsis patients were used. The platform for this dataset is Affymetrix Human Genome U133 Plus 2.0 Array. Sepsis patients were recruited if they met the 1992 Consensus Statement criteria and had clinical evidence of systemic infection based on microbiology diagnoses (n=27). Participants in the post-surgical (PS) group were recruited pre-operatively and blood samples collected within 24 hours following surgery

(n=38). Healthy controls (HC) included hospital staff with no known concurrent illnesses (n=20). Each participant had minimally 5ml of PAXgene blood collected for leucocyte RNA isolation and gene expression analyses.

4.2.2 Gene selection procedure

By using repeated random sampling [141], 10,000 training-testing sets were generated. These 10,000 randomly generated training-testing sets were randomly placed into 20 sampling groups, and each group contains 500 training-testing sets.

Each of the 20 sampling groups was used to derive a signature. In the 500 training-testing sets in every sampling group, each training-set was used to select genes by RFE based on SVM system. For all iterations and testing-sets, SVM system employed a set of globally modified parameters which gave the best average class-differentiation accuracy over the 500 testing-sets.

On every sampling group, three gene-ranking consistency evaluation steps were implemented on top of the normal RFE procedures in all sampling groups:

For every training-set, subsets of genes ranked in the bottom 10% (if no gene was selected in current iteration, this percentage was gradually increased to the bottom 40%) with combined score lower than the first top-ranked gene were selected such that collective contribution of these genes less likely outweighed higher-ranked ones.

For every training-set, the step (1) selected genes was further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes were consistently ranked lower.

A consensus scoring scheme was applied to step (2) selected genes such that only those appearing in >90% (if no gene was selected in current iteration, this percentage was gradually reduced to 60%) of the 500 training-sets were eliminated.

4.2.3 Performance evaluation of signatures

The predictive capability and robustness of gene signatures was evaluated by using several microarray data analysis methods on independent microarray datasets. The microarray data analysis methods included hierarchical clustering and SVM.

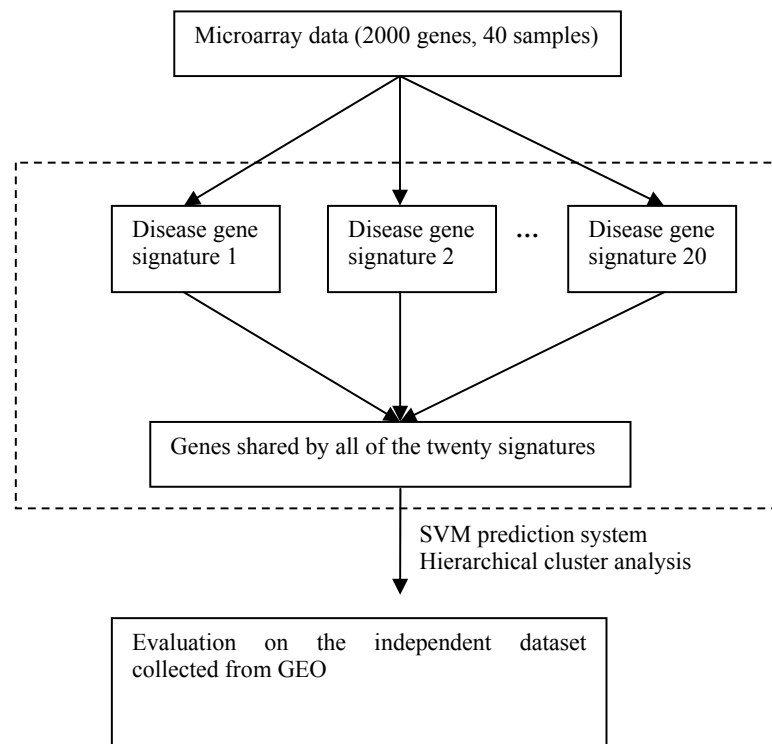
By using hierarchical clustering analysis, the performance of gene signatures was analyzed. As a popular unsupervised method, hierarchical clustering analysis groups genes and samples which have similar expression in the microarray data. Typically, the analysis begins with each gene/sample considered as a separate cluster. They are successively merged until one large cluster comprising the whole dataset is achieved. Later, these clusters are displayed in the form of a branching tree diagram, which can be broken into distinct clusters by cutting across the tree at a particular height. Hierarchical cluster analysis was carried out using the selected signatures by the software from Eisen *et al* [148, 371]. The results from hierarchical clustering were displayed by TreeView, which was also provided by Eisen *et al* [148, 371].

In SVM evaluation system, 500 random-generated training-test set were generated. The performance of the gene signatures was evaluated by overall accuracies Q (Equation 2-22) from the associated 500 test sets of the 500 SVM classification systems.

4.3 Results and discussion

4.3.1 System of the disease marker selection

The aim of this study was to identify the important gene signatures with regard to the intrinsic complex interactions of genes in sepsis disease initiation. Moreover, considering the noise in the microarray data arising from measurement variability and biological differences, the selected important gene signatures should be stable with regarding to such kind of variations. Based on the above concerns, recursive feature elimination method based on SVM was used to identify the different signatures from the multiple random combinations of samples. 20 sets of survival marker signatures were obtained by using RFE-SVM from 500 training-testing datasets with random sampling methods. SVM classifiers and hierarchical cluster analysis were used to evaluate the prediction system constructed from selected signatures (**Figure 4-1**).

Figure 4-1: The system of sepsis genes derivation and sepsis differentiation

4.3.2 Consistency analysis of the identified disease markers

The consistency level of the 20 derived signatures was estimated from the percentage of predictor-genes shared by them. 41 genes were shared by all of 20 signatures (**Table 4-1**) in which the number of disease genes ranged from 45 to 65 indicating that 63%~91% of all genes in each signature were shared by 20 signatures (**Table 4-2**). Comparing to 20 sets of signatures derived from the same dataset of 40 samples by other groups, our selected signatures are stable.

There are two aspects explaining why our selected gene signatures possess better stability. First, a SVM class-differentiation system with a universal set of

globally optimized parameters, which gave the best average class-differentiation accuracy over the 500 testing-sets, was used to derive RFE gene-ranking function at every iteration step and for every testing-set. As such, the effect from the parameter-dependence of conventional gene selection can be reduced dramatically. In earlier studies using RFE or other wrapper methods for selecting signatures, non-predictor-genes have been eliminated in multiple iterations, and at every iteration step a different class-differentiation system, characterized by a different set of optimized parameters, has been constructed [156, 274]. As gene-elimination is parameter-dependent, these selected predictor-genes are likely path-dependent and heavily influenced by sampling method, composition, order of gene evaluation, computational algorithm and parameters. These characteristics partly explain the highly unstable and patient-dependent characteristics of the previously-derived signatures [147]. Second, an additional gene-ranking consistency evaluation was performed on top of the normal RFE procedure to reduce the variations of erroneous eliminations of predictor-genes.

The optimal SVM parameters for the 20 sample-sets were in a narrow range of 17~18 and the highest average accuracies were 89.72%~93.63% for sepsis patients and 99.30%~91.78% for normal people respectively (**Table 4-3**). At these parameters, the accuracies for the individual testing-sets ranged from 89.79~93.31% for sepsis patients and 99.30%~99.78% for normal people. Further deviation from these optimal parameters had relatively small effect on prediction accuracy and composition of predictor-genes. The relatively small variations of optimal SVM parameters and prediction accuracies across the 20 sampling-sets suggests that the performance of the SVM class-differentiation

systems constructed by using globally optimized parameters and RFE iteration steps are fairly stable across different sampling combinations.

Table 4-1 : List of sepsis biomarkers shared by all 20 groups, 15 groups and 10 groups.

Official Name	Full Name	Other name	Summary
List of sepsis biomarkers shared by all 20 groups			
ACE	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	DCP; ICH; ACE1; DCP1; CD143; MVCD3	This gene encodes an enzyme involved in catalyzing the conversion of angiotensin I into a physiologically active peptide angiotensin II. Angiotensin II is a potent vasopressor and aldosterone-stimulating peptide that controls blood pressure and fluid-electrolyte balance. This enzyme plays a key role in the renin-angiotensin system. Many studies have associated the presence or absence of a 287 bp Alu repeat element in this gene with the levels of circulating enzyme or cardiovascular pathophysiologies.
ADAMTS13	ADAM metalloproteinase with thrombospondin type 1 motif	13; TTP; VWFCP; C9orf8; vWF-CP; ADAM-TS; ADAM-TS13; ADAMTS-13	This gene encodes a member of the ADAMTS (a disintegrin and metalloproteinase with thrombospondin motif) protein family. Members of the family share several distinct protein modules, including a propeptide region, a metalloproteinase domain, a disintegrin-like domain, and a thrombospondin type 1 (TS) motif. Individual members of this family differ in the number of C-terminal TS motifs, and some have unique C-terminal domains
CALCA	calcitonin-related polypeptide alpha	CT; KC; CGRP; CALC1; CGRP1; CGRP-I	This gene encodes the peptide hormones calcitonin, calcitonin gene-related peptide and katecalcine by tissue-specific alternative RNA splicing of the gene transcripts and cleavage of inactive precursor proteins. Calcitonin is involved in calcium regulation and acts to regulate phosphorus metabolism.

CCR 3	chemokine (C-C motif) receptor 3	chemokine (C-C motif) receptor	The protein encoded by this gene is a receptor for C-C type chemokines. It belongs to family 1 of the G protein-coupled receptors. This receptor binds and responds to a variety of chemokines, including eotaxin (CCL11), eotaxin-3 (CCL26), MCP-3 (CCL7), MCP-4 (CCL13), and RANTES (CCL5).
CD10	membrane metallo-endopeptidase	NEP; SFE; CALLA	This gene encodes a common acute lymphocytic leukemia antigen that is an important cell surface marker in the diagnosis of human acute lymphocytic leukemia (ALL). This protein is present on leukemic cells of pre-B phenotype, which represent 85% of cases of ALL. This protein is not restricted to leukemic cells, however, and is found on a variety of normal tissues. It is a glycoprotein that is particularly abundant in kidney, where it is present on the brush border of proximal tubules and on glomerular epithelium.
CD14	CD14 molecule		The protein encoded by this gene is a surface antigen that is preferentially expressed on monocytes/macrophages. It cooperates with other proteins to mediate the innate immune response to bacterial lipopolysaccharide. Alternative splicing results in multiple transcript variants encoding the same protein
CNP	2',3'-cyclic nucleotide 3' phosphodiesterase	CNP1	
CRP	C-reactive protein, pentraxin-related	PTX1	The protein encoded by this gene belongs to the pentaxin family. It is involved in several host defense related functions based on its ability to recognize foreign pathogens and damaged cells of the host and to initiate their elimination by interacting with humoral and cellular effector systems in the blood. Consequently, the level of this protein in plasma increases greatly during acute phase response to tissue injury, infection, or other inflammatory stimuli
CRTh2	chemoattractant receptor-homologous molecule		

	expressed on TH2 cells		
ESR	esterase 5 regulator		
FLT1	fms-related tyrosine kinase 1	FLT; FLT-1; VEGFR1; VEGFR-1	This gene encodes a member of the vascular endothelial growth factor receptor (VEGFR) family. VEGFR family members are receptor tyrosine kinases (RTKs) which contain an extracellular ligand-binding region with seven immunoglobulin (Ig)-like domains, a transmembrane segment, and a tyrosine kinase (TK) domain within the cytoplasmic domain.
flt-1	Protein FLT-1		
GAS6	growth arrest-specific 6	AXSF; AXLLG	This gene product is a gamma-carboxyglutamic acid (Gla)-containing protein thought to be involved in the stimulation of cell proliferation, and may play a role in thrombosis. Alternatively spliced transcript variants encoding different isoforms have been found for this gene
GFAP	glial fibrillary acidic protein		This gene encodes one of the major intermediate filament proteins of mature astrocytes. It is used as a marker to distinguish astrocytes from other glial cells during development. Mutations in this gene cause Alexander disease, a rare disorder of astrocytes in the central nervous system. Alternative splicing results in multiple transcript variants encoding distinct isoforms
HMGB-1	high mobility group box 1	HMG1; HMG3; SBP-1	
ICAM1	intercellular adhesion molecule 1	BB2; CD54; P3.58	This gene encodes a cell surface glycoprotein which is typically expressed on endothelial cells and cells of the immune system. It binds to integrins of type CD11a / CD18, or CD11b / CD18 and is also exploited by Rhinovirus as a receptor
IL-1	interleukin 1 complex		
IL-2	interleukin 2		cytokine produced by T-cells in response to antigen or mitogen stimulation
IL-4	interleukin 4	interleukin 4	Th2-type cytokine; may be involved in inflammatory response in eosinophils

IL6	interleukin 6	IL6; Ifnb2	a cytokine involved in development and possibly in neurodegenerative processes
IL-6	interleukin 6	IL6; Ifnb2	a cytokine involved in development and possibly in neurodegenerative processes
IL6ST	interleukin 6 signal transducer	CD130; GP130; CDW130; IL-6RB	The protein encoded by this gene is a signal transducer shared by many cytokines, including interleukin 6 (IL6), ciliary neurotrophic factor (CNTF), leukemia inhibitory factor (LIF), and oncostatin M (OSM). This protein functions as a part of the cytokine receptor complex. The activation of this protein is dependent upon the binding of cytokines to their receptors.
IL-8	interleukin 8		
LAMB1	laminin, beta 1	CLM	Laminins, a family of extracellular matrix glycoproteins, are the major noncollagenous constituent of basement membranes. They have been implicated in a wide variety of biological processes including cell adhesion, differentiation, migration, signaling, neurite outgrowth and metastasis. Laminins are composed of 3 non identical chains: laminin alpha, beta and gamma (formerly A, B1, and B2, respectively) and they form a cruciform structure consisting of 3 short arms, each formed by a different chain, and a long arm composed of all 3 chains
LBP	lipopolysaccharide binding protein	BPIFD2	The protein encoded by this gene is involved in the acute-phase immunologic response to gram-negative bacterial infections. Gram-negative bacteria contain a glycolipid, lipopolysaccharide (LPS), on their outer cell wall. Together with bactericidal permeability-increasing protein (BPI), the encoded protein binds LPS and interacts with the CD14 receptor, probably playing a role in regulating LPS-dependent monocyte responses. Studies in mice suggest that the encoded protein is necessary for the rapid acute-phase response to LPS but not for the clearance of LPS from circulation
LBP	lipopolysaccharide	BPIFD2	The protein encoded by this gene is

	e binding protein		involved in the acute-phase immunologic response to gram-negative bacterial infections. Gram-negative bacteria contain a glycolipid, lipopolysaccharide (LPS), on their outer cell wall. Together with bactericidal permeability-increasing protein (BPI), the encoded protein binds LPS and interacts with the CD14 receptor, probably playing a role in regulating LPS-dependent monocyte responses
MIF	macrophage migration inhibitory factor (glycosylation-inhibiting factor)	GIF; GLIF; MMIF	This gene encodes a lymphokine involved in cell-mediated immunity, immunoregulation, and inflammation. It plays a role in the regulation of macrophage function in host defense through the suppression of anti-inflammatory effects of glucocorticoids.
NAMPT	nicotinamide phosphoribosyltransferase	VF; PBEF; PBEF1; VISFATIN; 1110035O14Rik	This gene encodes a protein that catalyzes the condensation of nicotinamide with 5-phosphoribosyl-1-pyrophosphate to yield nicotinamide mononucleotide, one step in the biosynthesis of nicotinamide adenine dinucleotide. The protein belongs to the nicotinic acid phosphoribosyltransferase (NAPRTase) family and is thought to be involved in many important biological processes, including metabolism, stress response and aging. This gene has a pseudogene on chromosome 10
NPPB	natriuretic peptide B	BNP	This gene is a member of the natriuretic peptide family and encodes a secreted protein which functions as a cardiac hormone. The protein undergoes two cleavage events, one within the cell and a second after secretion into the blood. The protein's biological actions include natriuresis, diuresis, vasorelaxation, inhibition of renin and aldosterone secretion, and a key role in cardiovascular homeostasis. A high concentration of this protein in the bloodstream is indicative of heart failure. Mutations in this gene have been associated with postmenopausal osteoporosis
PDGFRA	platelet-derived	platelet-derived	This gene encodes a cell surface

	growth factor receptor, alpha polypeptide	growth factor receptor, alpha polypeptide	tyrosine kinase receptor for members of the platelet-derived growth factor family. These growth factors are mitogens for cells of mesenchymal origin. The identity of the growth factor bound to a receptor monomer determines whether the functional receptor is a homodimer or a heterodimer, composed of both platelet-derived growth factor receptor alpha and beta polypeptides. Studies suggest that this gene plays a role in organ development, wound healing, and tumor progression.
PF4	platelet factor 4	PF-4; CXCL4; SCYB4	This gene encodes a member of the CXC chemokine family. This chemokine is released from the alpha granules of activated platelets in the form of a homotetramer which has high affinity for heparin and is involved in platelet aggregation. This protein is chemotactic for numerous other cell type and also functions as an inhibitor of hematopoiesis, angiogenesis and T-cell function.
SAA1	serum amyloid A1	SAA; PIG4; SAA2; TP53I4	This gene encodes a member of the serum amyloid A family of apolipoproteins. The encoded protein is a major acute phase protein that is highly expressed in response to inflammation and tissue injury. This protein also plays an important role in HDL metabolism and cholesterol homeostasis. High levels of this protein are associated with chronic inflammatory diseases including atherosclerosis, rheumatoid arthritis, Alzheimer's disease and Crohn's disease
SPP1	secreted phosphoprotein 1	OPN; BNSP; BSPI; ETA-1	The protein encoded by this gene is involved in the attachment of osteoclasts to the mineralized bone matrix. The encoded protein is secreted and binds hydroxyapatite with high affinity. The osteoclast vitronectin receptor is found in the cell membrane and may be involved in the binding to this protein. This protein is also a cytokine that upregulates expression of interferon-gamma and interleukin-12. Several transcript variants encoding different isoforms

			have been found for this gene.
TLR2	toll-like receptor 2	TIL4; CD282	The protein encoded by this gene is a member of the Toll-like receptor (TLR) family which plays a fundamental role in pathogen recognition and activation of innate immunity. TLRs are highly conserved from Drosophila to humans and share structural and functional similarities. They recognize pathogen-associated molecular patterns (PAMPs) that are expressed on infectious agents, and mediate the production of cytokines necessary for the development of effective immunity.
TLR4	toll-like receptor 4	TOLL; CD284; TLR-4; ARMD10	The protein encoded by this gene is a member of the Toll-like receptor (TLR) family which plays a fundamental role in pathogen recognition and activation of innate immunity. TLRs are highly conserved from Drosophila to humans and share structural and functional similarities. They recognize pathogen-associated molecular patterns that are expressed on infectious agents, and mediate the production of cytokines necessary for the development of effective immunity.
TNF-α	tumour necrosis factor alpha-like		
TREM-1	triggering receptor expressed on myeloid cells 1		
TRPV1	transient receptor potential cation channel, subfamily V, member 1	VR1	Capsaicin, the main pungent ingredient in hot chili peppers, elicits a sensation of burning pain by selectively activating sensory neurons that convey information about noxious stimuli to the central nervous system. The protein encoded by this gene is a receptor for capsaicin and is a non-selective cation channel that is structurally related to members of the TRP family of ion channels.
VCAM-1	vascular cell adhesion molecule 1	CD106; INCAM-100	This gene is a member of the Ig superfamily and encodes a cell surface sialoglycoprotein expressed by cytokine-activated endothelium. This type I membrane protein

			mediates leukocyte-endothelial cell adhesion and signal transduction, and may play a role in the development of arteriosclerosis and rheumatoid arthritis.
VEGFA	vascular endothelial growth factor A	VPF; VEGF; MVCD1	This gene is a member of the PDGF/VEGF growth factor family and encodes a protein that is often found as a disulfide linked homodimer. This protein is a glycosylated mitogen that specifically acts on endothelial cells and has various effects, including mediating increased vascular permeability, inducing angiogenesis, vasculogenesis and endothelial cell growth, promoting cell migration, and inhibiting apoptosis. Elevated levels of this protein is linked to POEMS syndrome, also known as Crow-Fukase syndrome.
VIP	vasoactive intestinal peptide	PHM27	The protein encoded by this gene belongs to the glucagon family. It stimulates myocardial contractility, causes vasodilation, increases glycogenolysis, lowers arterial blood pressure and relaxes the smooth muscle of trachea, stomach and gall bladder. Alternative splicing occurs at this locus and two transcript variants encoding distinct isoforms have been identified
List of other 7 genes shared in 15 groups			
PRKACA	Protein kinase, cAMP-dependent, catalytic, alpha	MGC102831, MGC48865, PKACA	This protein is a signaling molecule important for a variety of cellular functions
CCL14	chemokine (C-C motif) ligand 14	CC-1, CC-3, CKb1, HCC-1, HCC-3, MCIF, NCC-2, NCC2, SCYA14, SCYL2, SY14	Chemokines play an important role in leukocyte mobilization, hematopoiesis, and angiogenesis. Tissue-specific expression of particular chemokines also influences tumor growth and metastasis [372]
PECAM1	platelet/endothelial cell adhesion molecule	CD31, PECAM-1	This protein participates in adhesive and/or signaling phenomena required for the motility and organization of endothelial cells [373]
KRT8	Keratin 8, cytokeratin 8; keratin, type II cytoskeletal 8	CARD2, CK8, CYK8, K2C8, K8, KO	This protein alters the epidermal cell differentiation, favors the neoplastic transformation of cells, and is ultimately responsible of the invasive behavior of transformed epidermal cells leading of conversion of benign to malignant tumors [374]

S100A11	S100 calcium binding protein A11 (calgizzarin)	1	MLN70, S100C	
MUC2	Mucin 2		MLP	MUC2 expression was regulated in human colon cancer cells at the level of transcription via AP-1 activation [375]. Reduction of MUC2 expression may be associated with the occurrence and progression of colorectal carcinomas [376]
IL1R2	Interleukin receptor, type II	1	CD121b, IL1RB, MGC47725	.
List of other 4 genes shared in 15 groups				
GSTM4	glutathione S-transferase M4		GSTM4-4, GTM4, MGC131945, MGC9247	A T2517C polymorphism in the GSTM4 gene is associated with risk of developing lung cancer [377]
GUCA2B	Guanylate cyclase activator 2B (uroguanylin)		GCAP-II, UGN	This protein synthesizes cGMP, which concentration of human colon tumors was higher than that of the surrounding mucosa [378]
HNRPA1	heterogeneous nuclear ribonucleoprotein A1		HNRNPA1, MGC102835	This protein may contribute to maintenance of telomere repeats in cancer cells with enhanced cell proliferation and the quantitative alteration of this protein could facilitate colon epithelial cell transformation through transcriptional and translational perturbation [379]
HSP90AB1	Heat shock protein 90kDa alpha (cytosolic), class B member 1		D6S182, FLJ26984, HSP90-BETA, HSP90B, HSPC2, HSPCB	This protein is important for signaling by types I and II interferons [380]

Table 4-2: Statistics of the selected sepsis genes from sepsis microarray dataset by class-differentiation systems constructed from 20 different sampling-sets each composed of 500 training-testing sets generated by random sampling.

Sampling Set	No of selected predictor genes in signature	No of predictor-genes also included in N other signatures derived by using different sampling-set																			
		19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	60	41	1	0	2	1	3	0	0	1	0	1	1	2	1	1	2	1	1	1	0
2	58	41	3	2	0	0	0	2	1	2	1	0	1	2	1	0	1	1	0	0	0
3	62	41	3	2	1	3	0	3	0	0	3	1	0	1	0	0	0	0	0	1	3
4	63	41	3	2	1	3	1	0	0	0	0	1	2	2	1	2	2	1	0	0	1
5	48	41	0	1	2	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0
6	45	41	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0
7	56	41	2	2	3	1	3	0	3	0	0	1	0	0	0	0	0	0	0	0	0
8	56	41	0	2	0	3	2	1	2	1	1	1	1	0	0	1	0	0	0	0	0
9	58	41	0	2	0	2	2	1	1	1	1	1	0	2	1	2	1	0	0	0	1
10	57	41	3	2	0	1	0	3	0	0	0	0	0	2	1	2	0	0	0	1	1
11	55	41	1	2	0	3	0	0	3	0	2	1	0	1	1	0	0	0	0	0	0
12	50	41	2	2	1	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
13	59	41	0	2	1	1	1	3	0	2	1	1	2	2	1	0	0	1	0	0	0
14	58	41	1	2	0	0	0	3	0	0	0	0	1	2	2	1	2	1	0	0	2
15	65	41	0	2	0	3	1	3	3	0	0	0	3	2	0	3	1	0	0	0	3
16	62	41	2	2	0	3	2	3	0	2	0	0	3	2	1	0	0	0	1	0	0
17	61	41	3	2	1	0	0	2	3	1	1	0	2	1	1	0	1	1	1	0	0
18	58	41	0	2	1	3	2	3	1	0	1	0	2	0	0	1	1	0	0	0	0
19	61	41	1	2	1	2	1	1	1	0	2	1	0	2	1	1	0	1	1	1	1
20	57	41	1	1	2	2	2	1	1	2	1	1	0	0	1	1	0	0	0	0	0

Table 4-3: Overall accuracies of 500 training-test sets on the optimal SVM parameters

Sampling set	Overall performance in 500 training dataset						Q
	Sepsis			Normal people			
	TP	FN	SE	TN	FP	SP	
1	2182	205	91.41%	8959	54	99.40%	97.73%
2	2127	233	90.13%	9001	39	99.57%	97.61%
3	2129	242	89.79%	8984	45	99.50%	97.48%
4	2138	211	91.02%	9007	44	99.51%	97.76%
5	2160	181	92.27%	9039	20	99.78%	98.24%
6	2165	165	92.92%	9021	49	99.46%	98.12%
7	2130	185	92.01%	9037	48	99.47%	97.96%
8	2150	231	90.30%	8993	26	99.71%	97.75%
9	2159	168	92.78%	9040	33	99.64%	98.24%
10	2120	243	89.72%	9004	33	99.63%	97.58%
11	2158	219	90.79%	8987	36	99.60%	97.76%
12	2176	148	93.63%	9042	34	99.63%	98.40%
13	2189	179	92.44%	8993	39	99.57%	98.09%
14	2147	198	91.56%	8992	63	99.30%	97.71%
15	2135	200	91.43%	9010	55	99.39%	97.76%
16	2158	206	91.29%	8997	39	99.57%	97.85%
17	2219	165	93.08%	8961	55	99.39%	98.07%
18	2147	192	91.79%	9024	37	99.59%	97.99%
19	2163	155	93.31%	9054	28	99.69%	98.39%
20	2158	180	92.30%	8999	63	99.30%	97.87%

4.3.3 The function of the identified sepsis markers

Sepsis can be defined as a general- cellular and humoral pathways with the generation of proized inflammatory response of the entire organism and and anti-inflammatory mediators. These mediators include often manifests itself as the systemic inflammatory re- cytokines, coagulation factors, adhesion molecules, sponse syndrome (SIRS). The pathogenesis of sepsis is a result of a complex network of events.

Coagulation, complement, contact system activation, inflammation, and apoptosis are all involved in the sepsis process. The detail functions of these biomarkers in the **Table 4-1**. C-reactive protein (CRP) is widely used as a marker of acute inflammation and one of the more studied sepsis biomarkers. It is produced by the liver in response to interleukin (IL)-6 generated during the inflammatory response to cellular injury. CRP is thought to assist in complement binding to foreign and damaged cells. The interleukins have been logical targets of sepsis biomarker investigations related to their role in inflammation and sepsis. Interleukin-6 (IL-6) is a pro-inflammatory cytokine that is produced in response to infection and other conditions of inflammation [8, 11]. IL-6 is an integral part of the cytokine activation cascade [10]. IL-6 is found to inhibit tumor necrosis factor-alpha and interleukin-1 but activate interleukin-1 receptor antagonist and interleukin-10. Interleukin-8 (IL-8) is an inflammatory cytokine that is released from monocytes, endothelial cells, and neutrophils in response to IL-1 and TNF- α . IL-8 responds by activating T cells, neutrophils and basophils [29]. Increases in circulating IL-8 are seen early in the infectious course. Interleukin-8 (IL-8) is an inflammatory cytokine that is released from monocytes, endothelial cells, and neutrophils in response to IL-1. IL-8 responds by activating T cells, neutrophils and basophils [29]. Increases in circulating IL-8 are seen early in the infectious course. tumor necrosis factor-alpha (TNF- α), a pro-inflammatory cytokine that is known to mediate inflammatory conditions including sepsis [35]. It is produced by dendritic cells, activated T cells and monocytes, macrophages, Langerhans cells, keratinocytes, fibroblasts, and astrocytes in response to cellular insult [35, 36]. Acting as one of the primary agents in initiating the cellular response to sepsis, TNF- α regulates the body's immune response by influencing production of a

variety of cells including prothymocytes and thymocytes. In addition, TNF- α activates macrophages and NK cells in the cytotoxic cascade [27, 35].

Because of biological differences and complex nature of cancers, a signature applicable for many patients is expected to include a substantial percentage of these sepsis-related genes, together with some of their interacting-partners and consequence-genes [381]. Moreover, because of measurement variability, a certain number of irrelevant genes may be inevitably included in a signature. Therefore, it is not surprising that the number of selected predictor-genes in our signatures ranged from 45 to 65. Moreover, it is probably unrealistic to assume that only a few genes stand out from the thousands of gene with sufficient clarity allowing target selection [382], which is a very important application of gene selection from microarray analysis.

4.3.4 The predictive performance of identified signatures in disease differentiation

To further evaluate the predictive capability of our selected signatures sets, we collected the gene expression profiles from GSE28750 dataset (from GEO database) which contains 10 normal people and 20 sepsis patients were used. Sepsis patients were recruited if they met the 1992 Consensus Statement criteria and had clinical evidence of systemic infection based on microbiology diagnoses (n=27). Participants in the post-surgical (PS) group were recruited pre-operatively and blood samples collected within 24 hours following surgery (n=38). Healthy controls (HC) included hospital staff with no known concurrent illnesses (n=20). The predictive capability of our selected signatures was evaluated by using the SVM classification system and 500

randomly-generated training-test sets generated from this dataset. The performance was evaluated using the associated test set and are shown in **Table 4-4**. The overall accuracy for the 83 predictor-genes was 93.26%. The accuracies for all predictor-genes were in the range of 92.97~94.57%. These results suggest that the selected signatures using our system can perform well in classification of drug sensitivity.

Table 4-4: Average sepsis prediction accuracy and standard deviation of 500 SVM class-differentiation systems constructed by 30 samples from GSE28750 dataset. The results were obtained from the overall accuracies of 500 test sets
TP: True positive, FN: False negative, SE: Sensitivity.

Signature (method)	No of selected predictor genes in signature	No of selected predictor-genes in signature	Sepsis Patient			Normal people			Q
			TP	FN	SE	TN	FP	SP	
1	60	155	1322	294	81.81%	4777	107	97.81%	93.83%
2	58	135	1317	305	81.20%	4728	150	96.93%	93.00%
3	62	156	1359	260	83.94%	4775	106	97.83%	94.37%
4	63	146	1331	294	81.91%	4791	84	98.28%	94.18%
5	48	116	1402	240	85.38%	4744	114	97.65%	94.55%
6	45	112	1344	254	84.11%	4780	122	97.51%	94.22%
7	56	119	1354	263	83.74%	4771	112	97.71%	94.23%
8	56	127	1334	285	82.40%	4767	114	97.66%	93.86%
9	58	133	1306	324	80.12%	4751	119	97.56%	93.18%
10	57	156	1345	281	82.72%	4770	104	97.87%	94.08%
11	55	139	1320	316	80.68%	4758	106	97.82%	93.51%
12	50	115	1327	292	81.96%	4747	134	97.25%	93.45%
13	59	144	1348	266	83.52%	4771	115	97.65%	94.14%
14	58	157	1329	338	79.72%	4717	116	97.60%	93.02%
15	65	149	1390	266	83.94%	4757	87	98.20%	94.57%
16	62	136	1279	334	79.29%	4764	123	97.48%	92.97%
17	61	136	1344	269	83.32%	4752	135	97.24%	93.78%
18	58	127	1385	268	83.79%	4745	102	97.90%	94.31%
19	61	146	1375	240	85.14%	4771	114	97.67%	94.55%
20	57	122	1350	290	82.32%	4748	112	97.70%	93.82%

4.4 Concluding Remarks

This chapter described a system for marker discovery. The system was designed to overcome the unstable signatures from different combination of samples and different classification method. Multiple random sampling method and consistency evaluation strategy were incorporated into the normal RFE gene selection procedure. The system was tested on colon cancer marker discovery. The results show that our selected markers could present both better stability and higher predictive performance on different microarray datasets than other signatures. 41 gene signatures are fairly stable with 69%~93% of all predictor-genes shared by all 20 signatures sets. These gene signatures includes inflammation factor such as CRP, cytokine/chemokine biomarkers, IL-6, IL-8 and tumor necrosis factor-alpha (TNF- α). The predictive ability of the selected signature shared by all of the 20 sets is evaluated by SVM models on an independent dataset collected from GEO Database. Unsupervised hierarchical clustering analysis provides additional indication of the predictive ability of selected signatures.

Chapter 5 Breast cancer biomarker selection based on Copy number variation

This chapter introduces signature selection with other source of high-throughput dataset – Breast cancer copy number variation (CNV) based signature selection. Total of 373 breast cancer samples and 517 normal people samples were used. We first calculated the breast cancer and normal people CNV calling by hidden Markov model. Then the strategies include the incorporation of multiple random sampling methods and the evaluation of gene-consistency into RFE gene selection procedure was used for gene signature selection. The predictive ability of these signatures are evaluated by SVM models, and unsupervised hierarchical clustering analysis. Hierarchical clustering analysis and literature search are used to evaluate the pattern of the identified markers.

5.1 Introduction

Human populations show extensive polymorphism — both additions and deletions — in the number of copies of chromosomal segments, and the number of genes in those segments [161] [383]. This is known as copy number variation (CNV). A high proportion of the genome, currently estimated at up to 12%, is subject to copy number variation [384]. We defined a CNV as a DNA segment that is 1 kb or larger and present at variable copy number in

comparison with a reference genome [385]. A CNV can be simple in structure, such as tandem duplication, or may involve complex gains or losses of homologous sequences at multiple sites in the genome.

CNV arise from genomic rearrangements, primarily owing to deletion, duplication, insertion, and unbalanced translocation events. CNV seems to be a major driving force in evolution, especially in the rapid evolution that has occurred, and continues to occur, within the human and great ape lineage [386-388]. However, much of the variation in copy number is disadvantageous. Change in copy number is involved in cancer formation and progression [166] and contributes to cancer proneness. In many situations, a change in copy number of any one of many specific genes is not well tolerated, and leads to a group of pathological conditions known as genomic disorders [389].

Breast cancer is the most common cause of cancer-related death among women worldwide ([390]. It is the leading cause of cancer mortality with around 411,000 annual deaths worldwide [391]. Like other solid cancers, breast cancer presents genomic instability. The current concept is that frequently occurring regions of DNA amplification commonly harbor oncogenes, whereas regions of recurrent deletion harbor tumor suppressor genes. Classical cytogenetic methods have been used to detect such copy number changes in tumors [174], which have deepened our understanding of the genomic hallmarks of breast cancer [392].

Genome-wide CNV in breast cancer have been profiled using a number of BAC clone- or cDNA-based array comparative genomic hybridization (CGH)

technologies [393]. These surveys revealed genomic regions with copy number gains, such as 8q11, 1q21, 17q11, and 11q13, as well as genomic segments that are frequently deleted, including regions that harbor known tumor suppressors such as CDKN2A and PTEN. Often, the amplified genomic regions were examined. To identify the “driver genes,” those genes that, when amplified, provide selective growth advantage for cancer cells. However, the resolutions of the previous array platforms, noise effect, and wrong algorithm were usually inadequate in defining the fine boundaries of DNA CNV or in pinpointing the genes under the most selective pressure. Because of these limitations, prior biological knowledge was heavily used to infer causal genes in amplified regions. Consequently, many known or putative oncogenes were credited as the driver genes, while some potentially novel cancer-driving genes may have been overlooked.

In this chapter, we explored gene selection method based on breast cancer copy number variation. Total of 373 breast cancer samples and 517 normal people samples were used. We first calculated the breast cancer and normal people CNV calling by hidden Markov model. Then the strategies include the incorporation of multiple random sampling methods and the evaluation of gene-consistency into RFE gene selection procedure was used for gene signature selection. 91 genes were selected after 20 times of experiments. The gene signatures are fairly stable with 80% of top-50 and 69%~93% of all predictor-genes shared by all 20 signatures. These shared predictor-genes include 48 cancer-related and 16 cancer-implicated genes, as well as 50% of

the previously-derived predictor-genes.

5.2 Materials and methods

5.2.1 Breast cancer and normal people CNV datasets

We mainly used SNP array for calculating CNV calling. For breast cancer datasets, we choose GSE 16619, GSE9154, GSE7545, GSE11977 and GSE11976 datasets (**Table 5-1**). The total number sample is 373. The main platform for these datasets is Affymetrix Mapping 250K Nsp SNP Array, Affymetrix Mapping 250K Sty2 SNP Array, Affymetrix Genome-Wide Human SNP 5.0 Array. The GeneChip® Human Mapping 500K Array Set provides consistently high coverage across different populations. It is comprised of two arrays, each capable of genotyping on average 250,000 SNPs (approximately 262,000 for Nsp arrays and 238,000 for Sty arrays). The SNP Array 5.0 is a single microarray featuring all single nucleotide polymorphisms (SNPs) from the original two-chip Mapping 500K Array Set, as well as 420,000 additional non-polymorphic probes that can measure other genetic differences, such as copy number variation. SNPs on the array are present on 200 to 1,100 base pair (bp) Nsp I or Sty I digested fragments in the human genome, and are amplified using the fifth generation of the Whole-genome Sampling Assay (WGSA).

The normal people datasets is from GSE16904, GSE17094, GSE16985, GSE16896 and GSE16894. The total number of normal people sample is 517 (**Table 5-1**). Two independent data sets of sepsis (GSE13904 and GSE28750)

were used for sepsis gene discovery and for validating the effect of our selected genes. The main platform is Illumina Human1M-Duov3 DNA Analysis BeadChip (Human1M-Duov3_B). Over one million markers to interrogate human genetic variation using single nucleotide polymorphisms (SNPs) and copy number variation (CNV) probes.

Table 5-1: Breast cancer and normal people CNV dataset used in biomarker selection

Type	Datasets	No. Sample	Platform
Breast Cancer	GSE 16619	203	Affymetrix Mapping 250K Nsp SNP Array
			Affymetrix Mapping 250K Sty2 SNP Array
			Affymetrix Genome-Wide Human SNP 5.0
	GSE 9154	42	Affymetrix Mapping 250K Sty
	GSE 7545	102	Affymetrix Mapping 250K Nsp
			Affymetrix Mapping 250K Sty
GSE 11977	6	Illumina Human Hsp 550v3	
GSE 11976	20	Illumina Human CNV 370	
Total		373	
Healthy people	GSE 16904	125	Illumina Human 1 Mv1_C
	GSE 17094	123	Illumina Human CNV 370v1
	GSE 16895	90	Illumina Human 1M-Duov3
	GSE 16896	90	Illumina Human 1M-Duov3
	GSE 16894	89	Illumina Human 1M-Duov3
Total		517	

5.2.2 CNV calling calculation

Step 1. Generate the signal intensity data based on raw CEL files

The goal of the first step is to generate the cross-marker normalized signal intensity data from an Affymetrix genotyping project to a text file, so that it can be analyzed subsequently by the PennCNV software [394]. This step has 3 substeps. A flowchart outlining the procedure for CNV calling from genotyping data is shown in **Figure 5-1**.

Suppose all the files from a genotyping project is stored in a directory called gw6/. Under this directory, there are several sub-directories: a CEL/ directory that stores the raw CEL files for each genotyped sample, a lib/ directory that stores library and annotation files provided by Affymetrix and by PennCNV-Affy,. We will try to write output files to the apt/ directory.

We need to download the PennCNV software from <http://www.openbioinformatics.org/penncnv/download/penncnv.latest.tar.gz> and uncompress the file.

Next download the PennCNV-Affy programs and library files from <http://www.openbioinformatics.org/penncnv/download/gw6.tar.gz> and uncompress the file. These files are required for signal pre-processing and also for CNV calling. There will be a lib/ directory that contains some PennCNV-specific library files for genome-wide 6.0 array; in addition, the library files for the genome-wide 5.0 arrays and Mapping 500K arrays are in the libgw5/ and gw6/lib500k/ directories, respectively.

Next download the Affymetrix Power Tools (APT) software package from <http://www.affymetrix.com/support/developer/powertools/index.affx>. We need to log into the website to download the software.

Substep 1.1 Generate genotyping calls from CEL files

This step uses the apt-probeset-genotype program in Affymetrix Power Tools (APT) to generate genotyping calls from the raw CEL files using the Birdseed algorithm (for genome-wide 6.0 array) or BRLMM-P (for genome-wide 5.0 array) algorithm. Note that the genotyping calling requires lots of CEL files.

(a) Genome-wide 6.0 array

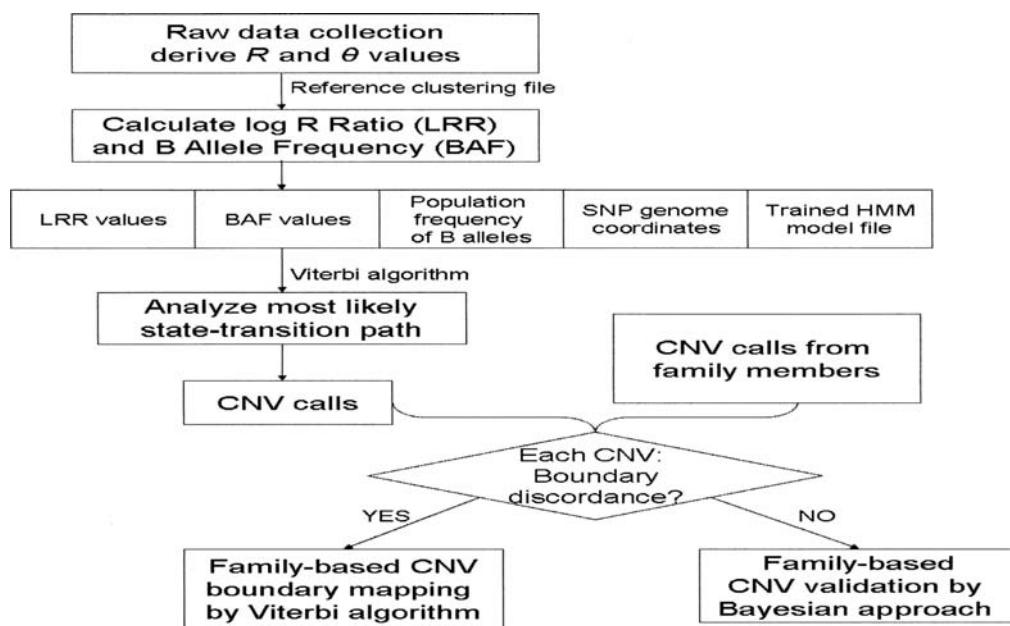
Before performing this step, we need to download the library files for the genome-wide 6.0 array from http://www.affymetrix.com/Auth/support/downloads/library_files/genomewidesnp6_libraryfile.zip, and save the decompressed files to the lib/ directory. Several files in this directory, including a CDF file and a Birdseed model file, will be used in the genotype calling step.

```
$ apt-probeset-genotype -c lib/GenomeWideSNP_6.cdf -a birdseed
--read-models-birdseed lib/GenomeWideSNP_6.birdseed.models
--special-snps lib/GenomeWideSNP_6.specialSNPs --out-dir apt --cel-files
listfile
```

The above command generates genotyping calls using all CEL files specified in the listfile, and generates several output files in the apt/ directory. The listfile contains a list of CEL file names, with one name per line, and with the first line being “cel_files”. The output files for this command include

birdseed.confidences.txt, birdseed.report.txt and birdseed.calls.txt. In addition, a birdseed.report.txt file is generated, that will be helpful to infer sample gender to generate a sexfile (see Substep 1.3 below).

Figure 5-1: A flowchart outlining the procedure for CNV calling from genotyping data.



(b) Genome-wide 5.0 array

For genome-wide 5.0 arrays, the command line is slightly different. First download the CDF and model files for GW5 array from http://www.affymetrix.com/Auth/support/downloads/library_files/genomewid_esnp5_libraryfile_rev1.zip and

<http://www.affymetrix.com/Auth/support/downloads/library>

_files/GenomeWideSNP_5.r2.zip. Then save decompressed files to the lib/ directory. There are several CDF files but we will need to use the GenomeWideSNP_5.Full.r2.cdf file. The genotype calling can be done using a command like this:

```
$ apt-probeset-genotype -c lib/GenomeWideSNP_5.Full.r2.cdf --chrX-snps
lib/GenomeWideSNP_5.Full.chrx --read-models-brlmm-p
lib/GenomeWideSNP_5.models -a brlmm-p --out-dir apt --cel-files listfile
```

(c) Mapping 500K array

For Mapping 500K array set with Nsp and Sty arrays, the genotype calling and signal extraction need to be done separately for each array. The command for genotype calling should use brlmm (instead of brlmm-p) as the algorithm (this is the default algorithm). In addition, there is no need to specify the --read-models-brlmm-p argument as shown above for Affy 5.0 arrays.

```
$apt-probeset-genotype -c
lib/CD_Mapping250K_Nsp_rev4/Full/Mapping250K_Nsp/LibFiles/Mapping250K_
Nsp.cdf --chrX-snps
lib/affy500k/CD_Mapping250K_Nsp_rev4/Full/Mapping250K_Nsp/LibFiles/Map
ping250K_Nsp.chrx --out-dir apt_nsp/ --cel-files list.nsp
```

As mentioned in the note above, if the program takes forever to run (during "computing prior" step), try to analyze only 500 samples and write the prior to a file (via --write-prior argument), then reanalyze the entire sample using --read-priors-brlmm argument to expedite the process.

Subsetp 1.2 Allele-specific signal extraction from CEL files

This step uses the Affymetrix Power Tools software to extract allele-specific signal values from the raw CEL files. Here “allele-specific” refers to the fact that for each SNP, we have a signal measure for the A allele and a separate signal measure for the B allele.

(a) Genome-wide 6.0 array

```
$ apt-probeset-summarize --cdf-file lib/GenomeWideSNP_6.cdf --analysis
quant-norm.sketch=50000,pm-only,med-polish,expr.genotype=true
--target-sketch lib/hapmap.quant-norm.normalization-target.txt
--out-dir apt --cel-files listfile
```

The above command read signal intensity values for PM probes in all the CEL files specified in listfile, apply quantile normalization to the values, apply median polish on the data, then generates signal intensity values for A and B allele for each SNP. The file hapmap.quant-norm.normalization-target.txt is provided in the PennCNV-Affy package: it is generated using all HapMap samples, as a reference quantile distribution to use in the normalization process, so that the quantile normalization procedures for different genotyping projects are more comparable to each other.

(b) Genome-wide 5.0 array

For genome-wide 5.0 arrays, the target-sketch can be found in the libgw5/ directory. An example command is given below:

```
$ apt-probeset-summarize --cdf-file lib/GenomeWideSNP_5.Full.r2.cdf
--analysis
quant-norm.sketch=50000,pm-only,med-polish,expr.genotype=true
--target-sketch libgw5/agre.quant-norm.normalization-target.txt
--out-dir apt --cel-files listfile
```

(c) Mapping 500K array

The signal extraction needs to be done for each array type separately. The pm-only option need to be used in --analysis argument since Mapping 500K array contains both PM and MM probes for each probe set.

```
$apt-probeset-summarize --cdf-file  
lib/affy500k/CD_Mapping250K_Nsp_rev4/Full/Mapping250K_Nsp/LibFiles/Map  
ping250K_Nsp.cdf --out-dir apt_nsp/ --cel-files list.nsp -a  
quant-norm.sketch=50000,pm-only,med-polish,expr.genotype=true  
--target-sketch lib/hapmap.nsp.quant-norm.normalization-target.txt
```

Substep 1.3 Generate canonical genotype clustering file

This step generates a file that contains the parameters for the canonical clustering information for each SNP or CNV marker, such that this file can be used later on to calculate LRR and BAF values.

If the user has only a few dozen CEL files, then it is unlikely that a clustering file can be generated successfully and accurately.

(a) Genome-wide 6.0 array

To generate canonical genotype clusters, use the generate_affy_genocuster.pl program in the downloaded PennCNV-Affy package (see gw6/bin/ directory).

```
$ generate_affy_genocuster.pl birdseed.calls.txt  
birdseed.confidences.txt  
quant-norm.pm-only.med-polish.expr.summary.txt  
-locfile ../lib/affygw6.hg18.pfb -sexfile file_sex -out gw6.genocuster
```

The affygw6.hg18.pfb file is provided in PennCNV-Affy package, which contains the annotated marker positions in hg18 (NCBI 36) human genome

assembly. The `file_sex` file is a two-column file that annotates the sex information for each CEL file, one file per line, and each line contains the file name and the sex separated by tab. The `file_sex` file is important for chrX markers and chrY markers, such that only females are used for constructing canonical clusters for chrX markers and that only males are used for constructing canonical clusters for chrY markers.

For example, the first 10 lines of a `file_sex` file are below:

```
10918.CEL male
10924.CEL male
11321_2.CEL female
10998.CEL female
11039.CEL female
11345.CEL female
10909.CEL female
11035.CEL female
11569_2.CEL female
```

Alternatively, one can use 1 to specify male and 2 to specify female in the sexfile. If the sex information for some CEL file is not known, you do not need to include them in the sexfile.

If the `--sexfile` argument is not provided, then chrX and chrY markers will not be processed and the resulting cluster file is only suitable for autosome CNV detection!

For a typical modern computer, the command should take several hours to process files generated from 1000-2000 CEL files.

(b) Genome-wide 5.0 array

An example command is given below:

```
$ generate_affy_genoclust.pl brlmm-p.calls.txt
brlmm-p.confidences.txt quant-norm.pm-only.med-polish.expr.summary.txt
-locfile ../libgw5/affygw5.hg18.pfb -sexfile file_sex -out
gw5.genoclust
```

(c) Mapping 500K array

Similar command as genome-wide arrays should be used for Nsp and Sty array separately.

```
$ generate_affy_genoclust.pl ../apt_nsp/brlmm.calls.txt ../apt_nsp/
brlmm.confidences.txt ../apt_nsp/quant-norm.pm-only.med-polish.expr.su
mmmary.txt -locfile lib/affy500k.hg18.pfb -sexfile file_sex -out
nsp.genoclust
```

Substep 1.4 LRR and BAF calculation

This step use the allele-specific signal intensity measures generated from the last step to calculate the Log R Ratio (LRR) values and the B Allele Frequency (BAF)

values for each marker in each individual. The `normalize_affy_genoclust.pl` program in the downloaded PennCNV-Affy package (see `gw6/bin/` directory) is used:

```
$ normalize_affy_genoclust.pl gw6.genoclust
quant-norm.pm-only.med-polish.expr.summary.txt
-locfile ../lib/affygw6.hg18.pfb -out gw6.lrr_baf.txt
```

The above command generates LRR and BAF values using the summary file generated in last step, and using a cluster file called `gw6.genoclust` generated in the last step. The location file specifies the chromosome position of each

SNP or CN probe, and this information is printed in the output files as well to facilitate future data processing.

For a typical modern computer, the command should take several hours to process files generated from 1000-2000 CEL files. A new tab-delimited file called gw6.lrr_baf.txt will be generated that contains one SNP per line and one sample per two columns (LRR column and BAF column).

If we do not have sufficient number of CEL files for the above substep 1.1 and 1.3, then you can alternatively use the default canonical clustering file provided in the PennCNV-Affy package. Right now several files are provided: hapmap.genocluster for GW6 arrays, agre.genocluster for GW5 arrays, and affy500k.nsp.genocluster/affy500k.sty.genocluster for Mapping 500K arrays. The results won't be optimal and are probably highly unreliable (the QC measures during PennCNV calling can give some clue on the signal-to-noise ratio of the resulting signal intensity files)

5.2.3 CNV annotation

5.2.3.1 Scan genomic regions against annotated genes

Functionality of the scan_region.pl program is to scan regions against annotated genes or exons. For example, suppose there is a list of copy number variation (CNV) regions, and we can use the program to specifically pick out those gene-disrupting CNVs as well as exonic CNVs. Another example is to map SNPs in a given array to either an overlapping gene (if the SNP is located

within the gene) or its closest gene (if the SNP is located in intergenic regions).

For this analysis, we need to download gene annotations from UCSC Genome Browser. Two types of annotations are widely used: RefGene and UCSC Gene. The former is well annotated but may miss some genuine transcripts, yet the later consists of many computationally predicted genes and is more comprehensive. Note that despite the names, both annotations are "transcript" annotations, rather than real "gene" annotations. (In contrast, the Ensembl does provide gene annotations, in addition to transcript annotations.)

5.2.4 Breast cancer gene selection procedure

By using repeated random sampling [141], 10,000 training-testing sets were generated, each constituted a training set which contains 373 samples and an associates test set which contains the other 517 samples from normal people dataset [147]. These 10,000 randomly generated training-testing sets were randomly placed into 20 sampling groups, and each group contains 500 training-testing sets.

Each of the 20 sampling groups was used to derive a signature. In the 500 training-testing sets in every sampling group, each training-set was used to select genes by RFE based on SVM system. For all iterations and testing-sets, SVM system employed a set of globally modified parameters which gave the best average class-differentiation accuracy over the 500 testing-sets.

On every sampling group, three gene-ranking consistency evaluation steps were implemented on top of the normal RFE procedures in all sampling groups:

- (1) For every training-set, subsets of genes ranked in the bottom 10% (if no gene was selected in current iteration, this percentage was gradually increased to the bottom 40%) with combined score lower than the first top-ranked gene were selected such that collective contribution of these genes less likely outweighed higher-ranked ones.
- (2) For every training-set, the step (1) selected genes was further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes were consistently ranked lower.
- (3) A consensus scoring scheme was applied to step (2) selected genes such that only those appearing in $>90\%$ (if no gene was selected in current iteration, this percentage was gradually reduced to 60%) of the 500 training-sets were eliminated.

5.2.5 Performance evaluation of signatures

By using hierarchical clustering analysis, the performance of gene signatures was analyzed. As a popular unsupervised method, hierarchical clustering analysis groups genes and samples which have similar expression in the microarray data. Typically, the analysis begins with each gene/sample considered as a separate cluster. They are successively merged until one large cluster comprising the whole dataset is achieved. Later, these clusters are

displayed in the form of a branching tree diagram, which can be broken into distinct clusters by cutting across the tree at a particular height. Hierarchical cluster analysis was carried out using the selected signatures by the software from Eisen *et al* [148, 371]. The results from hierarchical clustering were displayed by TreeView, which was also provided by Eisen *et al* [148, 371].

In SVM evaluation system, 500 random-generated training-test set were generated. The performance of the gene signatures was evaluated by overall accuracies Q (Equation 2-22) from the associated 500 test sets of the 500 SVM classification systems.

5.3 Results and discussion

5.3.1 CNV calls

The raw CEL files were used to generate canonical genotype clusters. The result converts signal intensity for each sample to LRR/BAF values, and then generates CNV calls. For the annotation of CNV callings, Human RefGene annotations (NCBI36 build) were used. The format of CNV calls is below (**Table 5-2**).

Table 5-2: Format of CNV calls

A1BG	cnv	-3	chr19:62877863-63731511
A2LD1	cnv	-3	chr13:99973265-100369439
ASGALT2	cnv	-3	chr1:3331266183-1344004812
A4GNT	cnv	-3	chr3:13886728834-13993734409
AACS	cnv	-3	chr12:123390784-124445406
AACSL	cnv	-3	chr5:1780988466-1789984791
AADAC	cnv	-3	chr9:152882773-153269184
AADACL3	cnv	-3	chr1:126224838-128527223
AADAT	cnv	-3	chr4:170276130-177178854
AAMP	cnv	-3	chr2:218729153-219045203
AANAT	cnv	-3	chr17:71348737-72843573
AATF	cnv	-3	chr17:131923810-132418471
ABCA1	cnv	-3	chr9:106391930-106917047
ABCA12	cnv	-3	chr2:215210139-221607247
ABCB1	cnv	-3	chr7:85256767-87228266
ABCC4	cnv	-3	chr13:993081765-944855400
ABCC5	cnv	-3	chr3:183484162239-1838666007
ABCG1	cnv	-3	chr1:4241382-43022001
ABCG5	cnv	-3	chr2:438743993-444008771
ABHD10	cnv	-3	chr3:112881899-1137855789
ABHD5	cnv	-3	chr3:422196226-440882403
ABIN3	cnv	-3	chr17:141595492-145883422
ABLIM1	cnv	-3	chr10:115993043-116886259
ABRA	cnv	-3	chr5:107222273-1079960877
ABT1	cnv	-3	chr6:26220872-30048241
ACAA1	cnv	-3	chr3:38076682-395255540
ACAA2	cnv	-3	chr18:45270216-458896441
ACAD11	cnv	-3	chr3:132267796-1335916176
ACAD9	cnv	-3	chr3:129849416-1300756400
ACADS	cnv	-3	chr12:11196449998-11199919947
ACADS8	cnv	-3	chr10:124477828-1248417024
ACAP2	cnv	-3	chr3:194881982-1977261133
ACCN1	cnv	-3	chr7:111111111-111111111
ACER2	cnv	-3	chr9:18710977-201833051
ACLY	cnv	-3	chr17:3660104995-3738581181
ACMSD	cnv	-3	chr2:13496010-13500000
ACN9	cnv	-3	chr7:96230424-96784635
ACOT1	cnv	-3	chr14:172944097-173652259
ACOT12	cnv	-3	chr5:806668962-811819112
ACOT13	cnv	-3	chr6:222594643-225566022
ACOX3	cnv	-3	chr4:8306593-8306593
ACOXL1	cnv	-3	chr2:1111142809-1111376131
ACEL2	cnv	-3	chr12:141802786-142622463
ACEP1	cnv	-3	chr18:1556652888-1565919725
ACRV1	cnv	-3	chr11:1224114987-1252268210
ACSF3	cnv	-3	chr16:87678937-87817464
ACSL1	cnv	-3	chr4:185379249-1855952970
ACSL3	cnv	-3	chr10:111347487-111444483
ACSL4	cnv	-3	chr1:13134868-13204483
ACSM4	cnv	-3	chr12:7193249-80083336
ACSS2	cnv	-3	chr20:3235211-33256820
ACSS3	cnv	-3	chr12:79735442-81178043
ACTL6A	cnv	-3	chr3:179147615-183997330

5.3.2 Statistics of the selected predictor genes from Breast cancer dataset

The stability levels of the 20 derived signatures can be estimated from the percentage of predictor genes shared by all 20 signatures. From **Table 5-3**, 80% of the top 50 ranked genes and 65% to 85% of all genes in each signature were shared by 20 signatures. This suggests that our selected signatures are fairly stable. One reason is that a SVM class differentiation system with a universal set of globally optimized variables, which gave the best average class differentiation accuracy over the 500 test sets, was used to derive RFE gene ranking function at every iteration step and for every test set. In earlier studies using RFE or other wrapper methods for selecting signatures, non-predictor genes have been eliminated in multiple iterations, and at every iteration step a different class differentiation system, characterized by a different set of optimized variables, has been constructed. As gene elimination

is variable dependent, these selected predictor genes are likely path dependent and heavily influenced by sampling method, composition, order of gene evaluation, computational algorithm, and variables. These characteristics partly explain the highly unstable and patient-dependent characteristics of the previously derived signatures. Another reason is that an additional gene-ranking consistency evaluation is done on top of the normal RFE procedure to reduce the change of erroneous elimination of predictor genes.

5.3.3 The function of the identified breast cancer markers

It is now well known that cancer is caused and driven by mutations in DNA that change the signal pathways which normally operate to regulate proliferation and death in normal cell. The activation of oncogene (drive excessive proliferation of cells) and inactivation of tumor suppressor genes (lose the inhibitory effect which is crucial to prevent inappropriate growth) change the normal signal pathway and hence leads to various well-defined phenotypic traits of cancer in **Figure 5-4** [395, 396]. These traits include proliferation, inappropriate survival, immortalization, invasion, angiogenesis and metastasis [395, 396]. Considering such complexity of tumorigenesis, the number of cancer genes in the signatures should not be very few. It was reported that there are 291 known cancer genes [381], 15 cancer-associated pathways [397], and 34 angiogenesis genes [398, 399]. Because of biological differences and complex nature of cancers, a signature applicable for many patients is expected to include a substantial percentage of these cancer-related

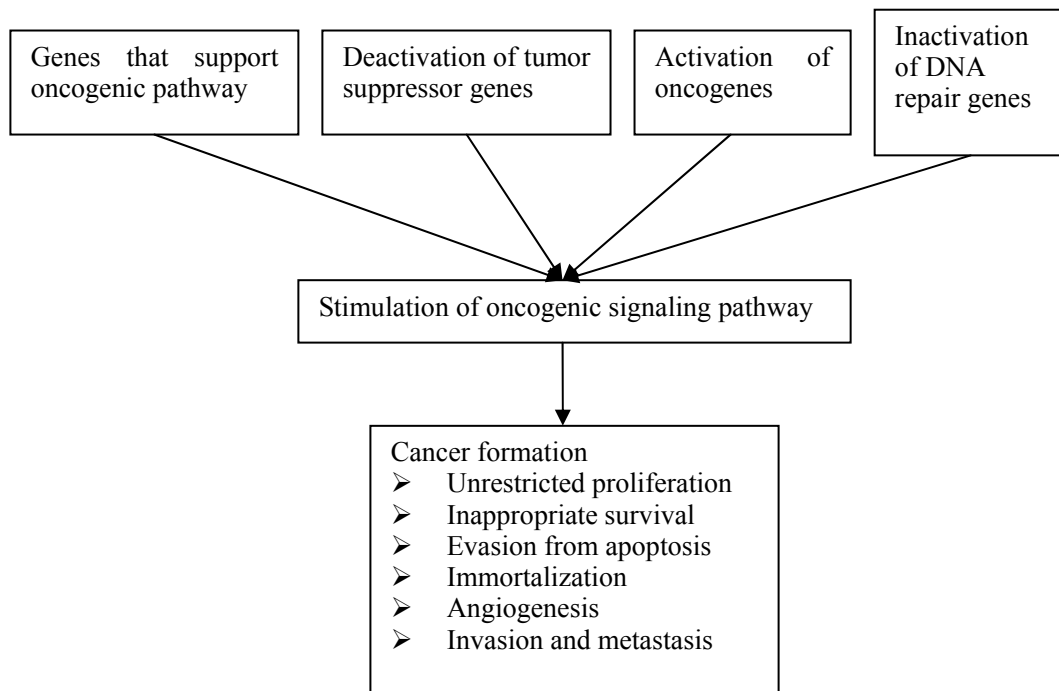
genes, together with some of their interacting-partners and consequence-genes [381]. Moreover, because of measurement variability, a certain number of irrelevant genes may be inevitably included in a signature. Therefore, it is not surprising that the number of selected predictor-genes in our signatures ranged from 112 to 157. Moreover, it is probably unrealistic to assume that only a few genes stand out from the thousands of gene with sufficient clarity allowing target selection [382], which is a very important application of gene selection from copy number variation analysis.

Table 5-3: Statistics of the selected predictor genes from Breast cancer dataset

Statistics of the selected predictor genes for predicting cancer outcome from a breast cancer data set by class differentiation systems constructed from 20 different sampling sets each composed of 500 training test sets generated by random sampling																					
Sampling set	No. selected predictor genes in signature	No. predictor genes also included in n other signatures derived by using different sampling set																			
		19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	130	91	5	3	3	4	4	2	2	3	5	2	0	0	2	2	0	0	2	0	0
2	136	91	4	4	4	1	4	3	5	2	3	0	3	0	3	2	0	2	2	0	3
3	124	91	4	2	3	2	2	0	3	3	2	2	3	4	0	0	0	0	0	2	1
4	129	91	3	1	2	3	3	2	2	1	3	3	4	0	0	0	0	0	3	3	2
5	109	91	0	3	2	3	1	0	2	0	1	2	2	1	0	0	0	0	0	0	0
6	117	91	3	2	1	2	2	3	3	3	0	3	0	0	1	2	1	0	0	0	0
7	135	91	4	5	4	3	5	4	4	3	2	2	0	3	4	0	0	1	0	0	0
8	119	91	3	2	2	4	2	1	2	0	2	0	0	2	3	0	2	1	0	0	2
9	125	91	5	5	3	1	3	2	2	4	3	2	1	0	0	0	3	0	0	0	0
10	133	91	3	5	4	3	2	3	3	5	4	2	0	0	1	1	3	1	0	0	2
11	120	91	4	2	3	1	1	5	4	2	2	2	0	0	1	0	0	2	0	0	0
12	114	91	5	3	1	2	1	1	3	1	2	0	0	1	1	2	0	0	0	0	0
13	112	91	3	2	2	0	0	2	2	2	3	2	0	3	0	0	0	0	0	0	0
14	126	91	4	3	2	2	3	4	3	2	4	3	1	0	0	2	0	0	2	0	0
15	138	91	3	4	3	0	5	4	4	4	4	2	3	0	0	2	1	2	0	1	1
16	124	91	2	2	2	4	2	2	3	2	3	3	0	0	2	0	2	0	1	0	3
17	128	91	2	4	2	2	0	2	2	3	2	5	0	3	0	3	1	2	2	0	2
18	113	91	0	2	0	3	3	1	2	2	2	2	0	3	0	1	1	0	0	0	0
19	122	91	1	0	1	2	2	2	1	1	4	5	3	0	2	2	0	2	1	1	1
20	110	91	3	1	3	1	2	0	2	4	2	1	0	0	0	0	0	0	0	0	0

The selected 91 predictor-genes shared by all 20 signatures shown in **Table 5-4; Table 5-6**. The cancer-related genes and cancer-pathways were taken from recent publications [381, 397-401] . Breast cancer is the most common cancer in women, comprising 23% of all female cancers, and it ranks second in overall cancer incidence when both sexes are considered. There were an estimated 1.15 million patients diagnosed with breast cancer worldwide in 2002[390]. Like other solid cancers, breast cancer presents genomic instability. The current concept is that frequently occurring regions of DNA amplification commonly harbor oncogenes, whereas regions of recurrent deletion harbor tumor suppressor genes.

Figure 5-2: Classes of genes involved in oncogenic transformation



The frequent aberration regions were as follows: gains in 2p25.3 – q37.3, 3q11.2 – 13.13, 3q21.1 – 29, 4p16.2 – q35.1, and 8q11.21 – q24.3, whereas losses in 1p36.31

- 33, 3p21.31 - 21.1, 9q33.3 - q34.3, 14q23.2 - 32.33, 15q11.2 - 26.3, 16p11.2 - q12.1, 17p13.3 - q21.32, 17q25.1 - 25.3, 19p13.3 - q13.43, 22q11.23 - 13.33, and Xp22.2 - q21.1 [392].

Amplifications involving chromosomes 8p (RAB11FIP1, FGFR1), 11q (CCND1) and 17q (ERBB2) are among the most common high level copy number aberrations in breast tumors. Amplification of 8p and 11q are most often observed in estrogen receptor positive tumors, while amplification of 17q (ERBB2) occurs in both estrogen receptor positive and negative tumors. Poor prognosis is associated with the presence of these amplicons in breast cancer. Thus, overexpressed genes within amplicons are attractive targets for therapy, as exemplified by the targeted use of herceptin to treat patients with tumors with ERBB2 amplification. Co-amplification of 8p12 and 11q13 is frequent. Amplification of 8p12 and 11q13 frequently occur together suggesting possible interactions between the genes in these two amplicons (**Table 5-5**). Indeed, it has been reported previously that FGFR1 (at 8p12) is up-regulated by increased expression of CCND1 (at 11q13) in fibroblasts, and occurs via CCND1 mediated activation of the pRB/E2F pathway.

5.3.4 Hierarchical clustering analysis of samples

The screened 91 predictor-genes show differential copy number enrichment pattern between normal samples (517 samples) and tumor samples (373 samples). We performed unsupervised 2D hierarchical clustering analysis of the 91 genes with 11 well annotated samples (6 breast cancer samples and 5

normal samples) to visualize the expression patterns of these 91 genes. The clustering shows clear differential enrichment pattern between the sample groups (**Figure 5-3**).

Figure 5-3: Hierarchical clustering analysis of copy number enrichment patterns of 91 genes in breast cancer samples and normal samples. (Red for higher relative enrichment level, blue for lower relative enrichment level and white for medium enrichment level)

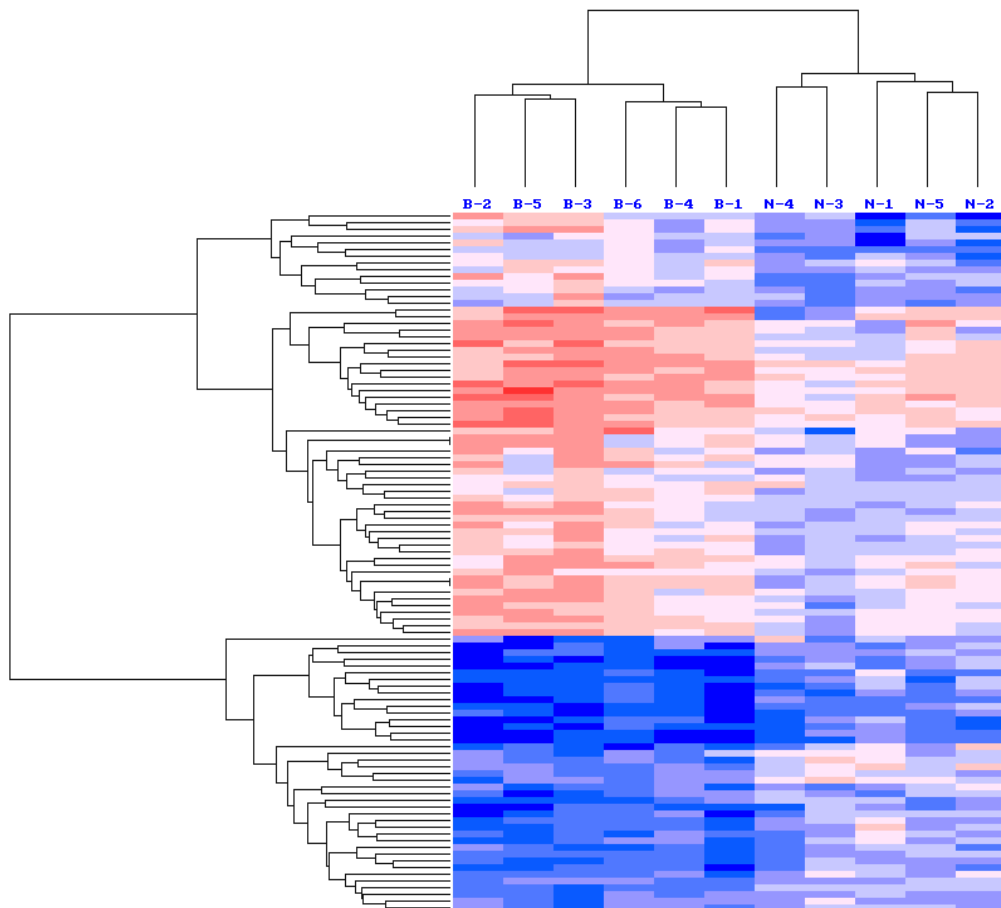


Table 5-4: List of predictor genes of breast cancer data set shared by all 20 signatures

List of predictor genes by other studies and our work		
Gene group	Predictor genes selected by both this work and other studies (no. studies)	Predictor genes selected by this work only
Cancer genes (Oncogene)	EGFR (3), ERBB2(7), AKT1(3), KRAS (2), RASA1(2), MLH1(1), RASA2(1),MYC	AKT2
Cancer genes (tumor suppressors)	TP53(3), PTEN(2)	BRCA1, BRCA2,
Cancer cell proliferation	MAPK10(1), CCND1(4), CCNA2(1), CDKN2A(1), CCNE1(1),	CCND2,PRKCI, PIK3CA, PTPRK
Angiogenesis Genes	FGFR2(2), FGFR1(2), FGF19(1)	FGF4, FGF3,
Cancer metastasis Genes	CTTN(3), PPF1A1(1), CDH26(1), TBL1XR1(1), PAK1(1), ASPH(1)	CDC42EP4, FREM2, ANKRD11, MICAL2
Tumor immune tolerance genes		CCR3, NFRKB, IDO1, CR2
Apoptosis Genes	PDCD5(1), BCL2L2(1), IGF1R(1), BAG4(1), FADD(2), TAF3(1)	
Interacting partner of cancer genes	GRB7(3), MDM2(1), RAB11FIP1(2)	BRIP1
Cancer pathway affiliated genes	C8orf76 (1), C17orf37(2), CDH4(1), CLNS1A(1), PPP1R3D(1), PROSC(3), PPM1D(3)	GPR116, C5orf22
DNA-synthesis/damage/mismatch repair	MDC1(1), MYST2(3), MSH3(1), RAD51(1), MLH3(1), TLK2(1),	MYST3, HIST1H2AH, HIST1H2BH, PFAS, DNASE1
Transcription Genes	E2F3(1), EIF4EBP2(1), INTS2(1), INTS4(2), GATA3(1), TAF4(1)	MRPL10
Others	ALG8(1), NGFR(1), JUND(1)	STARD3, CHRDL2, SERPINA4, ERI1, ACSL6, IRX4, KBTBD11, PRMT7, CACNB1, GPRIN1

Table 5-5 : Distribution of the selected predictor gene on chromosome (gene number >10)

Distribution of the selected predictor gene on chromosome (gene number >10)	
Chromosome 8 (12)	FGFR1, MYC, ASPH, IDO1, BAG4, C8orf76, PROSC, MYST3, RAB11FIP1, ERI1, KBTBD11, SLA
Chromosome (14)	CCND1, FGF3, FGF4, FGF19, CTTN, PPF1A1, PAK1, NFRKB, FADD, CLNS1A, INTS4, ALG8, MICAL2, CHRDL2
Chromosome (16)	ERBB2, TP53, BRCA1, CDC42EP4, GRB7, BRIP1, C17orf37, PPM1D, MYST2, PFAS, TLK2, INTS2, NGFR, STARD3, MRPL10, CACNB1
Amplifications involving chromosomes 8p, 11q, and 17q are among the most common high level copy number aberrations in breast cancer tumors. Poor prognosis is associated with the presence of these amplicons in breast cancer.	
Co-amplification of 8p12 and 11q13 is frequent. Amplification of 8p12 and 11q13 frequently occur together suggesting possible interactions between the genes in these two amplicons. For example, it has been reported previously that FGFR1 (at 8p12) is up-regulated by increased expression of CCND1 (at 11q13) in fibroblasts, and occurs via CCND1 mediated activation of the pRB/E2F pathway.	

Table 5-6 : List of function of breast cancer signatures.

Gene Name	Gene description	Gene aliases	
MAPK10	mitogen-activated protein kinase 10	JNK3; JNK3A; PRKM10; SAPK1b; p493F12; p54bSAPK	The protein encoded by this gene is a member of the MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development.
EGFR	epidermal growth factor receptor	ERBB; HER1; mENA; ERBB1; PIG61	The protein encoded by this gene is a transmembrane glycoprotein that is a member of the protein kinase superfamily. This protein is a receptor for members of the epidermal growth factor family.
FGF19	fibroblast growth factor 19		The protein encoded by this gene is a member of the fibroblast growth factor (FGF) family. FGF family members possess broad mitogenic and cell survival activities, and are involved in a variety of biological processes including embryonic development cell growth, morphogenesis, tissue repair, tumor growth and invasion.
FGF4	fibroblast growth factor 4	HST; KFGF; HST-1; HSTF1; K-FGF; HBGF-4	The protein encoded by this gene is a member of the fibroblast growth factor (FGF) family. FGF family members possess broad mitogenic and cell survival activities and are involved in a variety of biological processes including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion.
FGFR2	fibroblast growth factor receptor 2	BEK; JWS; BBDS; CEK3; CFD1; ECT1; KGFR; TK14; TK25; BFR-1; CD332; K-SAM	The protein encoded by this gene is a member of the fibroblast growth factor receptor family, where amino acid sequence is highly conserved between members and throughout evolution. FGFR family members differ

			from one another in their ligand affinities and tissue distribution.
FGFR1	fibroblast growth factor receptor 1	CEK; FLG; OGD; FLT2; KAL2; BFGFR; CD331; FGFR; FLT-2; HBGFR; N-SAM; FGFR-1; bFGF-R-1	The protein encoded by this gene is a member of the fibroblast growth factor receptor (FGFR) family, where amino acid sequence is highly conserved between members and throughout evolution. FGFR family members differ from one another in their ligand affinities and tissue distribution.
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2	NEU; NGL; HER2; TKR1; CD340; HER-2; MLN 19; HER-2/neu	This gene encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases. This protein has no ligand binding domain of its own and therefore cannot bind growth factors. However, it does bind tightly to other ligand-bound EGF receptor family members to form a heterodimer, stabilizing ligand binding and enhancing kinase-mediated activation of downstream signalling pathways, such as those involving mitogen-activated protein kinase and phosphatidylinositol-3 kinase.
AKT2	v-akt murine thymoma viral oncogene homolog 2	PKBB; PRKBB; HIHGHH; PKBBETA; RAC-BETA	This gene is a putative oncogene encoding a protein belonging to a subfamily of serine/threonine kinases containing SH2-like (Src homology 2-like) domains. The gene was shown to be amplified and overexpressed in 2 of 8 ovarian carcinoma cell lines and 2 of 15 primary ovarian tumors.
CCND2	cyclin D2	KIAK0002	The protein encoded by this gene belongs to the highly conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance through the cell cycle. Cyclins function as regulators of CDK kinases.
CCND1	cyclin D1	BCL1;PRAD1; U21B31; D11S287E	The protein encoded by this gene belongs to the highly

			conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance throughout the cell cycle. Cyclins function as regulators of CDK kinases.
CDKN2A	cyclin-dependent kinase inhibitor 2A	ARF; MLM; P14; P16; P19; CMM2; INK4; MTS1; TP16; CDK4I; CDKN2; INK4A; MTS-1; P14ARF; P19ARF; P16INK4; P16INK4A;	This gene generates several transcript variants which differ in their first exons. At least three alternatively spliced variants encoding distinct proteins have been reported, two of which encode structurally related isoforms known to function as inhibitors of CDK4 kinase.
PIK3CA	phosphoinositide-3-kinase	PI3K; p110-alpha	Phosphatidylinositol 3-kinase is composed of an 85 kDa regulatory subunit and a 110 kDa catalytic subunit. The protein encoded by this gene represents the catalytic subunit, which uses ATP to phosphorylate PtdIns, PtdIns4P and PtdIns(4,5)P2. This gene has been found to be oncogenic and has been implicated in cervical cancers.
AKT1	v-akt murine thymoma viral oncogene homolog 1	AKT; PKB; RAC; PRKBA; PKB-ALPHA; RAC-ALPHA	The serine-threonine protein kinase encoded by the AKT1 gene is catalytically inactive in serum-starved primary and immortalized fibroblasts. AKT1 and the related AKT2 are activated by platelet-derived growth factor. The activation is rapid and specific, and it is abrogated by mutations in the pleckstrin homology domain of AKT1.
CCNE1	cyclin E1	CCNE	The protein encoded by this gene belongs to the highly conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance through the cell cycle. Cyclins function as regulators of CDK kinases. Different cyclins exhibit distinct expression and degradation patterns which contribute to the temporal coordination of each mitotic event.

TP53	tumor protein p53	P53; LFS1; TRP53	This gene encodes tumor protein p53, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. p53 protein is expressed at low level in normal cells and at a high level in a variety of transformed cell lines, where it's believed to contribute to transformation and malignancy
PTEN	phosphatase and tensin homolog	BZS; DEC; GLM2; MHAM; TEPI; MMAC1; PTEN1; 10q23del	This gene was identified as a tumor suppressor that is mutated in a large number of cancers at high frequency. The protein encoded this gene is a phosphatidylinositol-3,4,5-tris phosphate 3-phosphatase.
BRCA1	breast cancer 1	IRIS; PSCP; BRCA1; BRCC1; PNCA4; RNF53; BROVCA1; PPP1R53	This gene encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. The encoded protein combines with other tumor suppressors, DNA damage sensors, and signal transducers to form a large multi-subunit protein complex known as the BRCA1-associated genome surveillance complex (BASC).
BRCA2	breast cancer 2	FAD; FACD; FAD1; GLM3; BRCC2; FANCB; FANCD; PNCA2; FANCD1; BROVCA2	Inherited mutations in BRCA1 and this gene, BRCA2, confer increased lifetime risk of developing breast or ovarian cancer. Both BRCA1 and BRCA2 are involved in maintenance of genome stability, specifically the homologous recombination pathway for double-strand DNA repair
TFAR19	programmed cell death 5	TFAR19	This gene encodes a protein that is upregulated during apoptosis where it translocates rapidly from the cytoplasm to the nucleus. The encoded protein may be an important regulator of K(lysine) acetyltransferase 5 (a protein involved in transcription, DNA damage response and cell cycle control) by

				inhibiting its proteasome-dependent degradation.
BCL2L2	BCL2-like 2	BCLW; BCL-W; PPP1R51; BCL2-L-2		This gene encodes a member of the BCL-2 protein family. The proteins of this family form hetero- or homodimers and act as anti- and pro-apoptotic regulators. Expression of this gene in cells has been shown to contribute to reduced cell apoptosis under cytotoxic conditions. Studies of the related gene in mice indicated a role in the survival of NGF- and BDNF-dependent neurons
IGF1R	ulin-like growth factor 1 receptor	IGFR; CD221; IGFIR; JTK13		This receptor binds insulin-like growth factor with a high affinity. It has tyrosine kinase activity. The insulin-like growth factor I receptor plays a critical role in transformation events. Cleavage of the precursor generates alpha and beta subunits.
BAG4	BCL2-associated athanogene 4	SODD; BAG-4		The protein encoded by this gene is a member of the BAG1-related protein family. BAG1 is an anti-apoptotic protein that functions through interactions with a variety of cell apoptosis and growth related proteins including BCL-2, Raf-protein kinase, steroid hormone receptors, growth factor receptors and members of the heat shock protein 70 kDa family
FADD	Fas (TNFRSF6)-associated via death domain	GIG3; MORT1		The protein encoded by this gene is an adaptor molecule that interacts with various cell surface receptors and mediates cell apoptotic signals. Through its C-terminal death domain, this protein can be recruited by TNFRSF6/Fas-receptor, tumor necrosis factor receptor, TNFRSF25, and TNFSF10/TRAIL-receptor, and thus it participates in the death signaling initiated by these receptors
TAF3	TAF3 RNA polymerase II,	TAF140; TAFII140; TAFII-140		The highly conserved RNA polymerase II transcription

	TATA box binding protein (TBP)-associated factor				factor TFIID (see TAF1; MIM 313650) comprises the TATA box-binding protein (TBP; MIM 600075) and a set of TBP-associated factors (TAFs), including TAF3. TAFs contribute to promoter recognition and selectivity and act as antiapoptotic factors
MDM2	Mdm2, ubiquitin ligase (mouse)	p53 E3 protein homolog	HDMX; ACTFS	hdm2;	This gene is a target gene of the transcription factor tumor protein p53. The encoded protein is a nuclear phosphoprotein that binds and inhibits transactivation by tumor protein p53, as part of an autoregulatory negative feedback loop. Overexpression of this gene can result in excessive inactivation of tumor protein p53, diminishing its tumor suppressor function
CTHRC1	collagen helix containing1	triple repeat			This locus encodes a protein that may play a role in the cellular response to arterial injury through involvement in vascular remodeling. Mutations at this locus have been associated with Barrett esophagus and esophageal adenocarcinoma
ASPH	aspartate beta-hydroxylase		AAH; BAH; HAAH; JCTN; CASQ2BP1	AAH; BAH; HAAH; junctin;	This gene is thought to play an important role in calcium homeostasis. The gene is expressed from two promoters and undergoes extensive alternative splicing. The encoded set of proteins share varying amounts of overlap near their N-termini but have substantial variations in their C-terminal domains resulting in distinct functional properties.
CTTN	cortactin		EMS1		This gene is overexpressed in breast cancer and squamous cell carcinomas of the head and neck. The encoded protein is localized in the cytoplasm and in areas of the cell-substratum contacts
PPFIA1	protein tyrosine phosphatase		LIP1; LIP.1; LIPRIN		The protein encoded by this gene is a member of the LAR protein-tyrosine phosphatase-interacting

			protein (liprin) family. Liprins interact with members of LAR family of transmembrane protein tyrosine phosphatases, which are known to be important for axon guidance and mammary gland development
CDH26	cadherin 26	VR20	Cadherins are a family of adhesion molecules that mediate Ca ²⁺ -dependent cell-cell adhesion in all solid tissues and modulate a wide variety of processes, including cell polarization and migration. Cadherin domains occur as repeats in the extracellular region and are thought to contribute to the sorting of heterogeneous cell types and the maintenance of orderly structures such as epithelium
TBL1XR1	transducin (beta)-like X-linked receptor 1	C21; DC42; IRA1; TBLR1	The protein encoded by this gene has sequence similarity with members of the WD40 repeat-containing protein family. The WD40 group is a large family of proteins, which appear to have a regulatory function
PAK1	protein (Cdc42/Rac)-activated kinase 1	PAKalpha	This gene encodes a family member of serine/threonine p21-activating kinases, known as PAK proteins. These proteins are critical effectors that link RhoGTPases to cytoskeleton reorganization and nuclear signaling, and they serve as targets for the small GTP binding proteins Cdc42 and Rac
C8orf76	chromosome 8 open reading frame 76		
C17ORF37	migration and invasion enhancer 1	C35; ORB3; XTP4; RDX12; C17orf37	
PPM1D	protein phosphatase, Mg ²⁺ /Mn ²⁺ dependent	WIP1; PP2C-DELTA	The protein encoded by this gene is a member of the PP2C family of Ser/Thr protein phosphatases. PP2C family members are known to be negative regulators of cell stress response pathways. The expression of this gene is

			induced in a p53-dependent manner in response to various environmental stresses
CDH4	cadherin 4, type 1, R-cadherin	CAD4; RCAD	This gene is a classical cadherin from the cadherin superfamily. The encoded protein is a calcium-dependent cell-cell adhesion glycoprotein comprised of five extracellular cadherin repeats, a transmembrane region and a highly conserved cytoplasmic tail.
GPR116	G protein-coupled receptor 116	KPG_001	
CLNS1A	chloride channel, nucleotide-sensitive, 1A	CLCI; ICln; CLNS1B	This gene encodes a protein that functions in multiple regulatory pathways. The encoded protein complexes with numerous cytosolic proteins and performs diverse functions including regulation of small nuclear ribonucleoprotein biosynthesis, platelet activation and cytoskeletal organization
PPP1R3D	protein phosphatase 1, regulatory subunit 3D	PPP1R6	Phosphorylation of serine and threonine residues in proteins is a crucial step in the regulation of many cellular functions ranging from hormonal regulation to cell division and even short-term memory. The level of phosphorylation is controlled by the opposing actions of protein kinases and protein phosphatases. Protein phosphatase 1 (PP1) is 1 of 4 major serine/threonine-specific protein phosphatases which have been identified in eukaryotic cells.
RASA2	GAP1M	GAP1M	The protein encoded by this gene is member of the GAP1 family of GTPase-activating proteins. The gene product stimulates the GTPase activity of normal RAS p21 but not its oncogenic counterpart. Acting as a suppressor of RAS function, the protein enhances the weak intrinsic GTPase activity of RAS proteins

				resulting in the inactive GDP-bound form of RAS, thereby allowing control of cellular proliferation and differentiation. This particular family member has a perinuclear localization and is an inositol 1,3,4,5-tetrakisphosphate-binding protein
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	NS; NS3; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-RAS2B; K-RAS4A; K-RAS4B		This gene, a Kirsten ras oncogene homolog from the mammalian ras gene family, encodes a protein that is a member of the small GTPase superfamily. A single amino acid substitution is responsible for an activating mutation. The transforming protein that results is implicated in various malignancies, including lung adenocarcinoma, mucinous adenoma, ductal carcinoma of the pancreas and colorectal carcinoma. Alternative splicing leads to variants encoding two isoforms that differ in the C-terminal region
RAB11FIP1	RAB11 family interacting protein 1 (class I)	RCP; NOEL1A; rab11-FIP1		Proteins of the large Rab GTPase family (see RAB1A; MIM 179508) have regulatory roles in the formation, targeting, and fusion of intracellular transport vesicles. RAB11FIP1 is one of many proteins that interact with and regulate Rab GTPases
MDC1	mediator of DNA-damage checkpoint 1	NFBD1		The protein encoded by this gene contains an N-terminal forkhead domain, two BRCA1 C-terminal (BRCT) motifs and a central domain with 13 repetitions of an approximately 41-amino acid sequence. The encoded protein is required to activate the intra-S phase and G2/M phase cell cycle checkpoints in response to DNA damage
MYST3	MYST histone acetyltransferase	MOZ; KAT6A; Zfp220; 1500036M03; 9930021N24Rik		
MSH3	DUP; MRP1	DUP; MRP1		The protein encoded by this gene forms a heterodimer with MSH2 to form MutS beta, part

				of the post-replicative DNA mismatch repair system. MutS beta initiates mismatch repair by binding to a mismatch and then forming a complex with MutL alpha heterodimer. This gene contains a polymorphic 9 bp tandem repeat sequence in the first exon.
MLH1	mutL homolog 1, colon cancer, nonpolyposis type 2	FCC2; HNPCC; HNPCC2	COCA2; hMLH1;	This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog of the E. coli DNA mismatch repair gene mutL, consistent with the characteristic alterations in microsatellite sequences (RER+phenotype) found in HNPCC. Alternative splicing results in multiple transcript variants encoding distinct isoforms. Additional transcript variants have been described, but their full-length natures have not been determined
HDAC2	histone deacetylase 2	HD2; RPD3; YAF1		This gene product belongs to the histone deacetylase family. Histone deacetylases act via the formation of large multiprotein complexes, and are responsible for the deacetylation of lysine residues at the N-terminal regions of core histones (H2A, H2B, H3 and H4). This protein forms transcriptional repressor complexes by associating with many different proteins, including YY1, a mammalian zinc-finger transcription factor. Thus, it plays an important role in transcriptional regulation, cell cycle progression and developmental events
RAD51	RAD51 homolog (S. cerevisiae)	RECA; MRMV2; RAD51A; HsT16930	BRCC5; HRAD51; HsRad51;	The protein encoded by this gene is a member of the RAD51 protein family. RAD51 family members are highly similar to bacterial RecA and Saccharomyces cerevisiae Rad51, and are known to be involved in the homologous recombination and repair of DNA. This

				protein can interact with the ssDNA-binding protein RPA and RAD52, and it is thought to play roles in homologous pairing and strand transfer of DNA. This protein is also found to interact with BRCA1 and BRCA2, which may be important for the cellular response to DNA damage
MLH3	mutL homolog 3 (E. coli)	3	HNPCC7	This gene is a member of the MutL-homolog (MLH) family of DNA mismatch repair (MMR) genes. MLH genes are implicated in maintaining genomic integrity during DNA replication and after meiotic recombination. The protein encoded by this gene functions as a heterodimer with other family members.
HIST1H2BH	histone cluster 1, H2bh	H2B/j;	H2BFJ	Histones are basic nuclear proteins that are responsible for the nucleosome structure of the chromosomal fiber in eukaryotes. Two molecules of each of the four core histones (H2A, H2B, H3, and H4) form an octamer, around which approximately 146 bp of DNA is wrapped in repeating units, called nucleosomes. The linker histone, H1, interacts with linker DNA between nucleosomes and functions in the compaction of chromatin into higher order structures
PFAS	phosphoribosylformylglycinamide synthase	PURL;	FGAMS; FGARAT	Purines are necessary for many cellular processes, including DNA replication, transcription, and energy metabolism. Ten enzymatic steps are required to synthesize inosine monophosphate (IMP) in the de novo pathway of purine biosynthesis. The enzyme encoded by this gene catalyzes the fourth step of IMP biosynthesis
TLK2	tousled-like kinase 2	PKU-ALPHA		The Tousled-like kinases, first described in Arabidopsis, are nuclear serine/threonine kinases that are potentially involved in the regulation of chromatin assembly

DNASE1	deoxyribonuclease I	DNL1; DRNI	This gene encodes a member of the DNase family. This protein is stored in the zymogen granules of the nuclear envelope and functions by cleaving DNA in an endonucleolytic manner. At least six autosomal codominant alleles have been characterized, DNASE1*1 through DNASE1*6, and the sequence of DNASE1*2 represented in this record. Mutations in this gene have been associated with systemic lupus erythematosus (SLE), an autoimmune disease
E2F3	E2F transcription factor 3	E2F-3	The protein encoded by this gene is a member of the E2F family of transcription factors. The E2F family plays a crucial role in the control of cell cycle and action of tumor suppressor proteins and is also a target of the transforming proteins of small DNA tumor viruses. The E2F proteins contain several evolutionally conserved domains found in most members of the family. These domains include a DNA binding domain, a dimerization domain which determines interaction with the differentiation regulated transcription factor proteins (DP), a transactivation domain enriched in acidic amino acids, and a tumor suppressor protein association domain which is embedded within the transactivation domain.
INTS2	integrator complex subunit 2	INT2; KIAA1287	INTS2 is a subunit of the Integrator complex, which associates with the C-terminal domain of RNA polymerase II large subunit (POLR2A; MIM 180660) and mediates 3-prime end processing of small nuclear RNAs U1
INTS4	integrator complex subunit 4	INT4; MST093	INTS4 is a subunit of the Integrator complex, which associates with the C-terminal domain of RNA polymerase II large subunit (POLR2A; MIM 180660) and mediates 3-prime end processing of small

			nuclear RNAs U1 (RNU1; MIM 180680) and U2 (RNU2; MIM 180690)
PROSC	proline synthetase co-transcribed homolog (bacterial)		
RPL19	ribosomal protein L19 L19		Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. This gene encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L19E family of ribosomal proteins. It is located in the cytoplasm
MRPL9	mitochondrial ribosomal protein L9	L9mt	Mammalian mitochondrial ribosomal proteins are encoded by nuclear genes and help in protein synthesis within the mitochondrion. Mitochondrial ribosomes (mitoribosomes) consist of a small 28S subunit and a large 39S subunit. They have an estimated 75% protein to rRNA composition compared to prokaryotic ribosomes, where this ratio is reversed. Another difference between mammalian mitoribosomes and prokaryotic ribosomes is that the latter contain a 5S rRNA.
BRD4	bromodomain containing 4	CAP; MCAP; HUNK1; HUNKI	The protein encoded by this gene is homologous to the murine protein MCAP, which associates with chromosomes during mitosis, and to the human RING3 protein, a serine/threonine kinase. Each of these proteins contains two bromodomains, a conserved sequence motif which may be involved in chromatin targeting
FOXA1	forkhead box A1	HNF3A; TCF3A	This gene encodes a member of the forkhead class of DNA-binding proteins. These hepatocyte nuclear factors are transcriptional activators for

					liver-specific transcripts such as albumin and transthyretin, and they also interact with chromatin. Similar family members in mice have roles in the regulation of metabolism and in the differentiation of the pancreas and liver
SPFH2	ER associated 2	lipid raft		NET32;SPG18; C8orf2; Erlin-2	This gene encodes a member of the SPFH domain-containing family of lipid raft-associated proteins. The encoded protein is localized to lipid rafts of the endoplasmic reticulum and plays a critical role in inositol 1,4,5-trisphosphate (IP3) signaling by mediating ER-associated degradation of activated IP3 receptors. Mutations in this gene are a cause of spastic paraplegia-18 (SPG18). Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene.
GATA3	GATA protein 3	binding		HDR; HDRS	This gene encodes a protein which belongs to the GATA family of transcription factors. The protein contains two GATA-type zinc fingers and is an important regulator of T-cell development and plays an important role in endothelial cell biology. Defects in this gene are the cause of hypoparathyroidism with sensorineural deafness and renal dysplasia
ALG8	asparagine-linked glycosylation 8, alpha-1,3-glucosyl transferase homolog			CDG1H	This gene encodes a member of the ALG6/ALG8 glucosyltransferase family. The encoded protein catalyzes the addition of the second glucose residue to the lipid-linked oligosaccharide precursor for N-linked glycosylation of proteins. Mutations in this gene have been associated with congenital disorder of glycosylation type 1h (CDG-1h). Alternatively spliced transcript variants encoding different isoforms have been identified
NGFR	nerve growth			CD271; p75NTR;	Nerve growth factor receptor

	factor receptor	TNFRSF16; p75(NTR); Gp80-LNGFR	contains an extracellular domain containing four 40-amino acid repeats with 6 cysteine residues at conserved positions followed by a serine/threonine-rich region, a single transmembrane domain, and a 155-amino acid cytoplasmic domain. The cysteine-rich region contains the nerve growth factor binding domain
TAF4	TAF4 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 135kDa	TAF2C; TAF4A; TAF2C1; TAFII130; TAFII135	Initiation of transcription by RNA polymerase II requires the activities of more than 70 polypeptides. The protein that coordinates these activities is transcription factor IID (TFIID), which binds to the core promoter to position the polymerase properly, serves as the scaffold for assembly of the remainder of the transcription complex, and acts as a channel for regulatory signals. TFIID is composed of the TATA-binding protein (TBP) and a group of evolutionarily conserved proteins known as TBP-associated factors or TAFs.
KIF26B	kinesin family member 26B		
MNX1	motor neuron and pancreas homeobox1	HB9; HLXB9; SCRA1; HOXHB9	This gene encodes a nuclear protein, which contains a homeobox domain and is a transcription factor. Mutations in this gene result in Currarino syndrome, an autosomic dominant congenital malformation. Alternatively spliced transcript variants encoding different isoforms have been found for this gene
CAPN9	calpain 9	GC36; nCL-4	Calpains are ubiquitous, well-conserved family of calcium-dependent, cysteine proteases. The calpain proteins are heterodimers consisting of an invariant small subunit and variable large subunits. The large subunit possesses a cysteine protease domain, and both subunits possess calcium-binding domains. Calpains have been implicated

			in neurodegenerative processes, as their activation can be triggered by calcium influx and oxidative stress.
CDR2L	cerebellar degeneration-related protein 2-like		
FANCA	Fanconi anemia, complementation group A	FA; FA1; FAA; FAH; FA-H; FACH	The Fanconi anemia complementation group (FANC) currently includes FANCA, FANCB, FANCC, FANCD1 (also called BRCA2), FANCD2, FANCE, FANCF, FANCG, FANCI, FANCI (also called BRIP1), FANCL, FANCM and FANCN (also called PALB2). The previously defined group FANCI is the same as FANCA. Fanconi anemia is a genetically heterogeneous recessive disorder characterized by cytogenetic instability, hypersensitivity to DNA crosslinking agents, increased chromosomal breakage, and defective DNA repair. The members of the Fanconi anemia complementation group do not share sequence similarity; they are related by their assembly into a common nuclear protein complex.
OR8G1	olfactory receptor, family 8, subfamily G, member 1	OR8G1P; TPCR25; HSTPCR25	Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell. The olfactory receptor proteins are members of a large family of G-protein-coupled receptors (GPCR) arising from single coding-exon genes. Olfactory receptors share a 7-transmembrane domain structure with many neurotransmitter and hormone receptors and are responsible for the recognition and G protein-mediated transduction of odorant signals
ARFGAP3	ADP-ribosylation factor GTPase activating protein 3	ARFGAP1	The protein encoded by this gene is a GTPase-activating protein (GAP) that associates with the Golgi apparatus and regulates the early secretory

		<p>pathway of proteins. The encoded protein promotes hydrolysis of ADP-ribosylation factor 1 (ARF1)-bound GTP, which is required for the dissociation of coat proteins from Golgi-derived membranes and vesicles. Dissociation of the coat proteins is a prerequisite for the fusion of these vesicles with target compartments.</p>
PERLD1	per1-like domain containing 1	
STARD3	CAB1; es64; CAB1; es64; MLN64 MLN64	<p>This gene encodes a member of a subfamily of lipid trafficking proteins that are characterized by a C-terminal steroidogenic acute regulatory domain and an N-terminal metastatic lymph node 64 domain. The encoded protein localizes to the membranes of late endosomes and may be involved in exporting cholesterol. Alternative splicing results in multiple transcript variants</p>
GRB7	growth factor receptor-bound protein 7	<p>The product of this gene belongs to a small family of adapter proteins that are known to interact with a number of receptor tyrosine kinases and signaling molecules. This gene encodes a growth factor receptor-binding protein that interacts with epidermal growth factor receptor (EGFR) and ephrin receptors. The protein plays a role in the integrin signaling pathway and cell migration by binding with focal adhesion kinase (FAK). Several transcript variants encoding two different isoforms have been found for this gene</p>
JUND	jun D AP-1 proto-oncogene	<p>The protein encoded by this intronless gene is a member of the JUN family, and a functional component of the AP1 transcription factor complex. It has been proposed to protect cells from p53-dependent senescence and apoptosis. Alternate</p>

			translation initiation site usage results in the production of different isoforms
CRP	C-reactive protein, pentraxin-related	PTX1	The protein encoded by this gene belongs to the pentaxin family. It is involved in several host defense related functions based on its ability to recognize foreign pathogens and damaged cells of the host and to initiate their elimination by interacting with humoral and cellular effector systems in the blood. Consequently, the level of this protein in plasma increases greatly during acute phase response to tissue injury, infection, or other inflammatory stimuli
HTR3B	5-hydroxytryptamine (serotonin) receptor 3B, ionotropic	5-HT3B	The product of this gene belongs to the ligand-gated ion channel receptor superfamily. This gene encodes subunit B of the type 3 receptor for 5-hydroxytryptamine (serotonin), a biogenic hormone that functions as a neurotransmitter, a hormone, and a mitogen. This receptor causes fast, depolarizing responses in neurons after activation. It is not functional as a homomeric complex, but a pentaheteromeric complex with subunit A (HTR3A) displays the full functional features of this receptor
FBXO43	F-box protein 43	EMI2; ERP1; FBX43	Members of the F-box protein family, such as FBXO43, are characterized by an approximately 40-amino acid F-box motif. SCF complexes, formed by SKP1 (MIM 601434), cullin (see CUL1; MIM 603134), and F-box proteins, act as protein-ubiquitin ligases.
ARFGAP3	ADP-ribosylation factor GTPase activating protein 3	ARFGAP1	The protein encoded by this gene is a GTPase-activating protein (GAP) that associates with the Golgi apparatus and regulates the early secretory pathway of proteins. The encoded protein promotes hydrolysis of

ADP-ribosylation factor 1 (ARF1)-bound GTP, which is required for the dissociation of coat proteins from Golgi-derived membranes and vesicles. Dissociation of the coat proteins is a prerequisite for the fusion of these vesicles with target compartments.

5.4 Concluding Remarks

In this chapter, the comprehensive gene selection system was further evaluated on the selection of cancer biomarker based on Copy number variation. By way of multiple random sampling, 91 genes were selected by all of twenty sets of breast cancer marker signatures. The derived 91 breast cancer signatures are found to be fairly stable with 80% of the top 50 ranked genes and 65% to 85% of all genes in each signature were shared by 20 signature sets. One reason is that a SVM class differentiation system with a universal set of globally optimized variables, which gave the best average class differentiation accuracy over the 500 test sets, was used to derive RFE gene ranking function at every iteration step and for every test set. The biomarker contains cell proliferation genes, like MAPK10, EGFR, tumor suppressor gene, apoptosis gene, tumorigenesis gene, cytoskeleton gene. Amplifications involving chromosomes 8p, 11q, and 17q are among the most common high level copy number aberrations in breast cancer tumors. Poor prognosis is associated with the presence of these amplicons in breast cancer. Co-amplification of 8p12 and 11q13 is frequent. Amplification of 8p12 and 11q13 frequently occur together

suggesting possible interactions between the genes in these two amplicons. For example, it has been reported previously that FGFR1 (at 8p12) is up-regulated by increased expression of CCND1 (at 11q13) in fibroblasts, and occurs via CCND1 mediated activation of the pRB/E2F pathway. These 21 markers were then used to develop PNN and SVM prediction models to predict prognosis for lung adenocarcinoma patients from different datasets. The survivability analysis by hierarchical clustering analysis and Kaplan-Meier survival analysis further suggested that the derived signatures from our system could provide better performance when comparing with other signatures. Most of the selected genes have been experimentally proved that high expression of the genes is relevant to adverse survivability of patients. 12 markers, including 5 known targets and 7 novel targets, were successfully predicted as therapeutic targets by using a therapeutic target prediction system.

Chapter 6 Concluding Remarks

6.1 Finding and merits

Thrombin, VEGF, and histamine are hallmarks of endothelial hyper-permeability, which perform their regulatory roles individually and collectively under different disease conditions, and with different dynamic profiles. Thrombin and VEGF can increase microvascular permeability ~50,000 times more potently than histamine [212]. Thrombin, VEGF, and histamine induce prolonged (1-1.5 hr), intermediate (15-20 min) and transient (~5 min) increases of endothelial permeability, respectively. An integrated simulation model that includes the signalling of all these hallmark mediators enables more comprehensive analysis of the signalling processes involved in different disease processes and regulated by different combinations of these mediators. Based on published models of relevant signaling, we developed an integrated mathematical model including the signaling pathways of all three of these mediators. Simulation results from our model were consistent with available experimental data of signaling mediated by both individual mediators and combinations of two mediators, and could be used to interpret the sustained and transient phases of MLC activation.

After building the mathematical model of endothelial permeability, to further explore the molecular mechanism of endothelial permeability, we developed a

robust computational system for gene signature using high-throughput datasets. A popular and accurate machine learning method, support vector machines, was applied to classify the samples. Recursive feature selection incorporating with multiple random sampling method and gene consistency evaluation strategies was used in gene selection procedure. The use of consensus scoring for multiple random sampling and evaluation of gene-ranking consistency seem to have impressive capability in avoiding erroneous elimination of predictor-genes due to such noise as measurement variability and biological differences. This system was used to select sepsis markers based on gene expression level. Then the system was expanded to the other type high-throughput data, breast cancer copy number variations dataset. For both cases, the markers were consistent with the variation of the samples, and present good predictive performances. The first case (endothelia permeability related disease sepsis) which contained the expression levels of 18 control and 22 patients were used for sepsis marker discovery. 20 sets of sepsis gene signatures were generated. 41 gene signatures are fairly stable with 69%~93% of all predictor-genes shared by all 20 signatures sets. For the second case (breast cancer copy number variation datasets), total of 373 breast cancer samples and 517 normal people samples were used, the derived 91 breast cancer marker signature are found to be fairly stable with 80% of the top 50 ranked genes and 65% to 85% of all genes in each signature were shared by 20 signatures.

In summary, the integrated endothelial permeability model was able to predict

the effects of altered pathway components and synergistic combination of multiple mediators, some of which are consistent with experimental findings [352]. Similar to the published models of other pathways, our model can potentially be used to identify important disease genes through sensitivity analysis of signalling components [402]. Our model may also be extended to emphasize other components to facilitate further investigation of the effects of different mediators, cascades, and cross-talk on endothelial permeability and related diseases. For both cases (sepsis biomarker selection and copy number variation based biomarker selection), the biomarker results suggest that our system can derive stable and good predictive marker signatures. Since the cost for high-throughput experiments is very high, the sample size is much smaller than what is required for a satisfactory diagnosis and prognosis of a certain disease such as cancer. In such situations, our system is particular useful to get real important markers for disease initiation, diagnosis, patient survival prediction and therapeutic target discovery. The use of consensus scoring for multiple random sampling and evaluation of gene-ranking consistency seem to have impressive capability in avoiding erroneous elimination of predictor-genes due to such noise as measurement variability and biological differences.

6.2 Limitations and suggestions for future study

A major function of the endothelial cell (EC) is to serve as a barrier to fluid and solute flux across the blood vessel wall. Breakdown of this barrier leads to

increased permeability and the development of edema. This process has been implicated in cancer metastasis, angiogenesis, ischaemic heart disease, inflammation, trauma, sepsis, and many other pathological conditions [1]. Due to the data limitation, we just construct the model of endothelial permeability model. Further, based on the model of endothelial permeability, we will construct the model of inflammation, sepsis and other disease (depended on the data available). Then we can observe how the endothelial permeability leads to these diseases.

Another aspect of this work was a robust computational system for gene signature derivation was developed. We mainly used the microarray and copy number variation (SNP data) to test the implementation. For the microarray dataset, further improvement in measurement quality, annotation accuracy and coverage, and signature-selection will enable the derivation of more accurate signatures for facilitating biomarker and target discovery. The currently available platforms for microarray data are different. Therefore if we could synchronize the platform and provide more samples, we could further improve the accuracy of our system and reduce the computational time. The gene ontology information also could be integrated into the system and the selected genes would be given a biological meaning directly. While for the platform of copy number variation, we mainly used the Penncnv software which based on hidden Markova model to calculate CNV calling. For further work, other algorithm of CNV calculation should be used as a comparison.

Furthermore, combined analysis of DNA copy number and gene expression microarrays and mutation of the same or similar tumor samples has revealed a major and direct effect of allelic imbalance on gene expression. We will continue to identify the biomarker by integration of gene amplification, expression and mutation. We wish to find which individual copy number changes affect gene expression levels and mutation within the same chromosomal region. Integrated analysis of both copy number variation, gene expression and mutation data could give additional information about the role of copy number alterations in the development of cancer.

BIBLIOGRAPHY

1. Kumar, P., et al., *Molecular mechanisms of endothelial hyperpermeability: implications in inflammation*. Expert Rev Mol Med, 2009. **11**: p. e19.
2. Vestweber, D., *VE-cadherin: the major endothelial adhesion molecule controlling cellular junctions and blood vessel formation*. Arterioscler Thromb Vasc Biol, 2008. **28**(2): p. 223-32.
3. Mehta, D. and A.B. Malik, *Signaling mechanisms regulating endothelial permeability*. Physiol Rev, 2006. **86**(1): p. 279-367.
4. Xiao, K., et al., *p120-Catenin regulates clathrin-dependent endocytosis of VE-cadherin*. Mol Biol Cell, 2005. **16**(11): p. 5141-51.
5. Clarke, H., A.P. Soler, and J.M. Mullin, *Protein kinase C activation leads to dephosphorylation of occludin and tight junction permeability increase in LLC-PK1 epithelial cell sheets*. J Cell Sci, 2000. **113** (Pt **18**): p. 3187-96.
6. Tiruppathi, C., et al., *Ca²⁺ signaling, TRP channels, and endothelial permeability*. Microcirculation, 2006. **13**(8): p. 693-708.
7. Wong, B.W., et al., *Vascular endothelial growth factor increases human cardiac microvascular endothelial cell permeability to low-density lipoproteins*. J Heart Lung Transplant, 2009. **28**(9): p. 950-7.
8. Bates, D.O. and F.E. Curry, *Vascular endothelial growth factor increases microvascular permeability via a Ca(2+)-dependent pathway*. Am J Physiol, 1997. **273**(2 Pt 2): p. H687-94.
9. Tharakan, B., et al., *beta-Catenin dynamics in the regulation of microvascular endothelial cell hyperpermeability*. Shock, 2012. **37**(3): p. 306-11.
10. Sawant, D.A., et al., *Role of beta-catenin in regulating microvascular endothelial cell hyperpermeability*. J Trauma, 2011. **70**(2): p. 481-7; discussion 487-8.
11. You, Q.H., et al., *Role of src-suppressed C kinase substrate in rat pulmonary microvascular endothelial hyperpermeability stimulated by inflammatory cytokines*. Inflamm Res, 2010. **59**(11): p. 949-58.

12. Xiong, C., et al., *The lectin-like domain of TNF protects from listeriolysin-induced hyperpermeability in human pulmonary microvascular endothelial cells - a crucial role for protein kinase C-alpha inhibition*. *Vascul Pharmacol*, 2010. **52**(5-6): p. 207-13.
13. Esser, S., et al., *Vascular endothelial growth factor induces VE-cadherin tyrosine phosphorylation in endothelial cells*. *J Cell Sci*, 1998. **111** (Pt 13): p. 1853-65.
14. Sandoval, R., et al., *Ca(2+) signalling and PKCalpha activate increased endothelial permeability by disassembly of VE-cadherin junctions*. *J Physiol*, 2001. **533**(Pt 2): p. 433-45.
15. Kowalczyk, A.P., et al., *VE-cadherin and desmoplakin are assembled into dermal microvascular endothelial intercellular junctions: a pivotal role for plakoglobin in the recruitment of desmoplakin to intercellular junctions*. *J Cell Sci*, 1998. **111** (Pt 20): p. 3045-57.
16. Lampugnani, M.G., et al., *The molecular organization of endothelial cell to cell junctions: differential association of plakoglobin, beta-catenin, and alpha-catenin with vascular endothelial cadherin (VE-cadherin)*. *J Cell Biol*, 1995. **129**(1): p. 203-17.
17. Dejana, E., *Endothelial cell-cell junctions: happy together*. *Nat Rev Mol Cell Biol*, 2004. **5**(4): p. 261-70.
18. Tiburu, E.K., et al., *Human cannabinoid 1 GPCR C-terminal domain interacts with bilayer phospholipids to modulate the structure of its membrane environment*. *AAPS J*, 2011. **13**(1): p. 92-8.
19. Peeters, M.C., et al., *GPCR structure and activation: an essential role for the first extracellular loop in activating the adenosine A2B receptor*. *FASEB J*, 2011. **25**(2): p. 632-43.
20. Xu, X., et al., *Coupling mechanism of a GPCR and a heterotrimeric G protein during chemoattractant gradient sensing in Dictyostelium*. *Sci Signal*, 2010. **3**(141): p. ra71.
21. Swanson, R. and J.R. Beasley, *Pathway-specific, species, and sub-type counterscreening for better GPCR hits in high throughput screening*. *Curr Pharm Biotechnol*, 2010. **11**(7): p. 757-63.
22. Eno, C.O., et al., *Distinct roles of mitochondria- and ER-localized Bcl-xL in apoptosis resistance and Ca2+ homeostasis*. *Mol Biol Cell*, 2012. **23**(13): p. 2605-18.
23. Huang, G., et al., *ER stress disrupts Ca2+-signaling complexes and*

- Ca²⁺ regulation in secretory and muscle cells from PERK-knockout mice.* J Cell Sci, 2006. **119**(Pt 1): p. 153-61.
24. Kuang, E., et al., *ER Ca²⁺ depletion triggers apoptotic signals for endoplasmic reticulum (ER) overload response induced by overexpressed reticulon 3 (RTN3/HAP).* J Cell Physiol, 2005. **204**(2): p. 549-59.
25. Li, Y.X., et al., *Ca²⁺ excitability of the ER membrane: an explanation for IP₃-induced Ca²⁺ oscillations.* Am J Physiol, 1995. **269**(5 Pt 1): p. C1079-92.
26. Villa, A., et al., *Intracellular Ca²⁺ stores in chicken Purkinje neurons: differential distribution of the low affinity-high capacity Ca²⁺ binding protein, calsequestrin, of Ca²⁺ ATPase and of the ER luminal protein, Bip.* J Cell Biol, 1991. **113**(4): p. 779-91.
27. Walz, B., *Association between cytoskeletal microtubules and Ca²⁺-sequestering smooth ER in Semper cells of fly ommatidia.* Eur J Cell Biol, 1983. **32**(1): p. 92-8.
28. Meyer, T. and L. Stryer, *Molecular model for receptor-stimulated calcium spiking.* Proc Natl Acad Sci U S A, 1988. **85**(14): p. 5051-5.
29. Xia, J., et al., *The role of PKC isoforms in the inhibition of NF-kappaB activation by vitamin K2 in human hepatocellular carcinoma cells.* J Nutr Biochem, 2012.
30. Chen, T., et al., *Protective effects of mGluR5 positive modulators against traumatic neuronal injury through PKC-dependent activation of MEK/ERK pathway.* Neurochem Res, 2012. **37**(5): p. 983-90.
31. Pinton, P., C. Pavan, and B. Zavan, *PKC-beta activation and pharmacologically induced weight gain during antipsychotic treatment.* Pharmacogenomics, 2011. **12**(4): p. 453-5.
32. Aktan, I., B. Dunkel, and F.M. Cunningham, *PKC isoenzymes in equine platelets and stimulus induced activation.* Vet Immunol Immunopathol, 2011. **141**(3-4): p. 276-82.
33. Buitrago, C., M. Costabel, and R. Boland, *PKC and PTPalpha participate in Src activation by 1alpha,25OH₂ vitamin D₃ in C2C12 skeletal muscle cells.* Mol Cell Endocrinol, 2011. **339**(1-2): p. 81-9.
34. Sheldahl, L.C., et al., *Dishevelled activates Ca²⁺ flux, PKC, and CamKII in vertebrate embryos.* J Cell Biol, 2003. **161**(4): p. 769-77.

35. Bissonnette, M., et al., *1,25(OH)₂ vitamin D₃ activates PKC- α in Caco-2 cells: a mechanism to limit secosteroid-induced rise in [Ca²⁺]_i*. *Am J Physiol*, 1994. **267**(3 Pt 1): p. G465-75.
36. Cheng, Z., et al., *Luciferase Reporter Assay System for Deciphering GPCR Pathways*. *Curr Chem Genomics*, 2010. **4**: p. 84-91.
37. Maeda A, Ozaki Y, Sivakumaran S, Akiyama T, Urakubo H, Usami A, Sato M, Kaibuchi K, Kuroda S: *Ca²⁺ -independent phospholipase A₂-dependent sustained Rho-kinase activation exhibits all-or-none response*. *Genes Cells* 2006, **11**:1071-1083.
38. Fujita, H., et al., *Molecular decipherment of Rho effector pathways regulating tight-junction permeability*. *Biochem J*, 2000. **346 Pt 3**: p. 617-22.
39. Jung, C.H., et al., *The role of Rho/Rho-kinase pathway in the expression of ICAM-1 by linoleic acid in human aortic endothelial cells*. *Inflammation*, 2012. **35**(3): p. 1041-8.
40. Schwenke, D.O., et al., *Role of Rho-kinase signaling and endothelial dysfunction in modulating blood flow distribution in pulmonary hypertension*. *J Appl Physiol*, 2011. **110**(4): p. 901-8.
41. Dunoyer-Geindre, S., R.J. Fish, and E.K. Kruithof, *Regulation of the endothelial plasminogen activator system by fluvastatin. Role of Rho family proteins, actin polymerisation and p38 MAP kinase*. *Thromb Haemost*, 2011. **105**(3): p. 461-72.
42. Harvey, K.A., et al., *Role of Rho kinase in sphingosine 1-phosphate-mediated endothelial and smooth muscle cell migration and differentiation*. *Mol Cell Biochem*, 2010. **342**(1-2): p. 7-19.
43. van Nieuw Amerongen, G.P., et al., *Thrombin-induced endothelial barrier disruption in intact microvessels: role of RhoA/Rho kinase-myosin phosphatase axis*. *Am J Physiol Cell Physiol*, 2008. **294**(5): p. C1234-41.
44. van Nieuw Amerongen, G.P., M.A. Vermeer, and V.W. van Hinsbergh, *Role of RhoA and Rho kinase in lysophosphatidic acid-induced endothelial barrier dysfunction*. *Arterioscler Thromb Vasc Biol*, 2000. **20**(12): p. E127-33.
45. Williams, J., J. Bogwu, and A. Oyekan, *The role of the RhoA/Rho-kinase signaling pathway in renal vascular reactivity in endothelial nitric oxide synthase null mice*. *J Hypertens*, 2006. **24**(7): p. 1429-36.

46. Zheng, H.Z., K.S. Zhao, and Q.B. Huang, [*Role of Rho kinase in reorganization of the vascular endothelial cytoskeleton induced by rat burn serum*]. *Zhonghua Shao Shang Za Zhi*, 2005. **21**(3): p. 181-4.
47. Stamatovic, S.M., et al., *Potential role of MCP-1 in endothelial cell tight junction 'opening': signaling via Rho and Rho kinase*. *J Cell Sci*, 2003. **116**(Pt 22): p. 4615-28.
48. Eto, M., et al., *Thrombin suppresses endothelial nitric oxide synthase and upregulates endothelin-converting enzyme-1 expression by distinct pathways: role of Rho/ROCK and mitogen-activated protein kinase*. *Circ Res*, 2001. **89**(7): p. 583-90.
49. van Nieuw Amerongen, G.P., et al., *Activation of RhoA by thrombin in endothelial hyperpermeability: role of Rho kinase and protein tyrosine kinases*. *Circ Res*, 2000. **87**(4): p. 335-40.
50. Bredt, D.S. and S.H. Snyder, *Isolation of nitric oxide synthetase, a calmodulin-requiring enzyme*. *Proc Natl Acad Sci U S A*, 1990. **87**(2): p. 682-5.
51. Funyu, J., et al., *VEGF can act as vascular permeability factor in the hepatic sinusoids through upregulation of porosity of endothelial cells*. *Biochem Biophys Res Commun*, 2001. **280**(2): p. 481-5.
52. Pocock, T.M., et al., *VEGF and ATP act by different mechanisms to increase microvascular permeability and endothelial [Ca(2+)](i)*. *Am J Physiol Heart Circ Physiol*, 2000. **279**(4): p. H1625-34.
53. Miyamoto, K., et al., *Vascular endothelial growth factor (VEGF)-induced retinal vascular permeability is mediated by intercellular adhesion molecule-1 (ICAM-1)*. *Am J Pathol*, 2000. **156**(5): p. 1733-9.
54. Levitas, E., et al., *Periovarian and interleukin-1 beta-dependent up-regulation of intraovarian vascular endothelial growth factor (VEGF) in the rat: potential role for VEGF in the promotion of periovarian angiogenesis and vascular permeability*. *J Soc Gynecol Investig*, 2000. **7**(1): p. 51-60.
55. Feng, D., et al., *Ultrastructural localization of the vascular permeability factor/vascular endothelial growth factor (VPF/VEGF) receptor-2 (FLK-1, KDR) in normal mouse kidney and in the hyperpermeable vessels induced by VPF/VEGF-expressing tumors and adenoviral vectors*. *J Histochem Cytochem*, 2000. **48**(4): p. 545-56.
56. Stacker, S.A., et al., *A mutant form of vascular endothelial growth*

- factor (VEGF) that lacks VEGF receptor-2 activation retains the ability to induce vascular permeability.* J Biol Chem, 1999. **274**(49): p. 34884-92.
57. Christov, C., et al., *Vascular permeability factor/vascular endothelial growth factor (VPF/VEGF) and its receptor flt-1 in microcystic meningiomas.* Acta Neuropathol, 1999. **98**(4): p. 414-20.
 58. Proescholdt, M.A., et al., *Vascular endothelial growth factor (VEGF) modulates vascular permeability and inflammation in rat brain.* J Neuropathol Exp Neurol, 1999. **58**(6): p. 613-27.
 59. Fischer, S., et al., *Hypoxia induces permeability in brain microvessel endothelial cells via VEGF and NO.* Am J Physiol, 1999. **276**(4 Pt 1): p. C812-20.
 60. Isaji, M., et al., *Inhibition by tranilast of vascular endothelial growth factor (VEGF)/vascular permeability factor (VPF)-induced increase in vascular permeability in rats.* Life Sci, 1998. **63**(4): p. PL71-4.
 61. Ruckman, J., et al., *2'-Fluoropyrimidine RNA-based aptamers to the 165-amino acid form of vascular endothelial growth factor (VEGF165). Inhibition of receptor binding and VEGF-induced vascular permeability through interactions requiring the exon 7-encoded domain.* J Biol Chem, 1998. **273**(32): p. 20556-67.
 62. Yan, Z., et al., *Vascular endothelial growth factor (VEGF)/vascular permeability factor (VPF) production by luteinized human granulosa cells in vitro; a paracrine signal in corpus luteum formation.* Gynecol Endocrinol, 1998. **12**(3): p. 149-53.
 63. Simon, M., et al., *Receptors of vascular endothelial growth factor/vascular permeability factor (VEGF/VPF) in fetal and adult human kidney: localization and [125I]VEGF binding sites.* J Am Soc Nephrol, 1998. **9**(6): p. 1032-44.
 64. Grutzkau, A., et al., *Synthesis, storage, and release of vascular endothelial growth factor/vascular permeability factor (VEGF/VPF) by human mast cells: implications for the biological significance of VEGF206.* Mol Biol Cell, 1998. **9**(4): p. 875-84.
 65. Klanke, B., et al., *Effects of vascular endothelial growth factor (VEGF)/vascular permeability factor (VPF) on haemodynamics and permselectivity of the isolated perfused rat kidney.* Nephrol Dial Transplant, 1998. **13**(4): p. 875-85.
 66. Watanabe, Y., et al., *Vascular permeability factor/vascular endothelial*

- growth factor (VPF/VEGF) delays and induces escape from senescence in human dermal microvascular endothelial cells.* Oncogene, 1997. **14**(17): p. 2025-32.
67. Aach, J., W. Rindone, and G.M. Church, *Systematic management and analysis of yeast gene expression data.* Genome Res, 2000. **10**(4): p. 431-45.
 68. Sun, J., et al., *Cytotoxicity, permeability, and inflammation of metal oxide nanoparticles in human cardiac microvascular endothelial cells: cytotoxicity, permeability, and inflammation of metal oxide nanoparticles.* Cell Biol Toxicol, 2011. **27**(5): p. 333-42.
 69. Sheikpranbabu, S., et al., *Silver nanoparticles inhibit VEGF-and IL-1beta-induced vascular permeability via Src dependent pathway in porcine retinal endothelial cells.* J Nanobiotechnology, 2009. **7**: p. 8.
 70. Sahni, A., et al., *The VE-cadherin binding domain of fibrinogen induces endothelial barrier permeability and enhances transendothelial migration of malignant breast epithelial cells.* Int J Cancer, 2009. **125**(3): p. 577-84.
 71. Alghisi, G.C., L. Ponsonnet, and C. Ruegg, *The integrin antagonist cilengitide activates alphaVbeta3, disrupts VE-cadherin localization at cell junctions and enhances permeability in endothelial cells.* PLoS One, 2009. **4**(2): p. e4449.
 72. Li, Q., et al., *[Effect of tumor necrosis factor-alpha on the permeability of strial capillary endothelial cells in guinea pig cochlea].* Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi, 2008. **43**(9): p. 691-5.
 73. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 2006. **7**(1): p. 55-65.
 74. Blair, R.H., D.J. Kliebenstein, and G.A. Churchill, *What can causal networks tell us about metabolic pathways?* PLoS Comput Biol, 2012. **8**(4): p. e1002458.
 75. Arita, M., *From metabolic reactions to networks and pathways.* Methods Mol Biol, 2012. **804**: p. 93-106.
 76. Faust, K., D. Croes, and J. van Helden, *Prediction of metabolic pathways from genome-scale metabolic networks.* Biosystems, 2011. **105**(2): p. 109-21.
 77. Pang, T.Y. and S. Maslov, *A toolbox model of evolution of metabolic pathways on networks of arbitrary topology.* PLoS Comput Biol, 2011.

- 7(5): p. e1001137.
78. Rezola, A., et al., *Exploring metabolic pathways in genome-scale networks via generating flux modes*. Bioinformatics, 2011. **27**(4): p. 534-40.
 79. Schuster, S., L.F. de Figueiredo, and C. Kaleta, *Predicting novel pathways in genome-scale metabolic networks*. Biochem Soc Trans, 2010. **38**(5): p. 1202-5.
 80. Jevremovic, D., et al., *On algebraic properties of extreme pathways in metabolic networks*. J Comput Biol, 2010. **17**(2): p. 107-19.
 81. Lavoie, H., H. Hogues, and M. Whiteway, *Rearrangements of the transcriptional regulatory networks of metabolic pathways in fungi*. Curr Opin Microbiol, 2009. **12**(6): p. 655-63.
 82. Moutselos, K., et al., *KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database*. BMC Bioinformatics, 2009. **10**: p. 324.
 83. Yang, X., et al., *Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks*. Nat Genet, 2009. **41**(4): p. 415-23.
 84. Yeung, M., I. Thiele, and B.O. Palsson, *Estimation of the number of extreme pathways for metabolic networks*. BMC Bioinformatics, 2007. **8**(1): p. 363.
 85. Rahman, S.A. and D. Schomburg, *Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks*. Bioinformatics, 2006. **22**(14): p. 1767-74.
 86. Tun, K., et al., *Metabolic pathways variability and sequence/networks comparisons*. BMC Bioinformatics, 2006. **7**: p. 24.
 87. Croes, D., et al., *Inferring meaningful pathways in weighted metabolic networks*. J Mol Biol, 2006. **356**(1): p. 222-36.
 88. Croes, D., et al., *Metabolic PathFinding: inferring relevant pathways in biochemical networks*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W326-30.
 89. Pescini, D., et al., *Simulation of the Ras/cAMP/PKA pathway in budding yeast highlights the establishment of stable oscillatory states*. Biotechnol Adv, 2012. **30**(1): p. 99-107.

90. Nagasaki, M., et al., *Systems biology model repository for macrophage pathway simulation*. Bioinformatics, 2011. **27**(11): p. 1591-3.
91. Matsubara, K., et al., *Sensitivity of kinetic macro parameters to changes in dopamine synthesis, storage, and metabolism: a simulation study for [(1)(8)F]FDOPA PET by a model with detailed dopamine pathway*. Synapse, 2011. **65**(8): p. 751-62.
92. Li, L., L. Yu, and Q. Huang, *Molecular trigger for pre-transfer editing pathway in Valyl-tRNA synthetase: a molecular dynamics simulation study*. J Mol Model, 2011. **17**(3): p. 555-64.
93. Guisasola, A., et al., *Modelling and simulation revealing mechanisms likely responsible for achieving the nitrite pathway through aeration control*. Water Sci Technol, 2010. **61**(6): p. 1459-65.
94. Hawari, A.H. and Z.A. Mohamed-Hussein, *Simulation of a Petri net-based model of the terpenoid biosynthesis pathway*. BMC Bioinformatics, 2010. **11**: p. 83.
95. Enkavi, G. and E. Tajkhorshid, *Simulation of spontaneous substrate binding revealing the binding pathway and mechanism and initial conformational response of GlpT*. Biochemistry, 2010. **49**(6): p. 1105-14.
96. Das, A. and C. Mukhopadhyay, *Mechanical unfolding pathway and origin of mechanical stability of proteins of ubiquitin family: an investigation by steered molecular dynamics simulation*. Proteins, 2009. **75**(4): p. 1024-34.
97. Li, H., et al., *Simulation of crosstalk between small GTPase RhoA and EGFR-ERK signaling pathway via MEKK1*. Bioinformatics, 2009. **25**(3): p. 358-64.
98. Raychaudhuri, S., et al., *Monte Carlo simulation of cell death signaling predicts large cell-to-cell stochastic fluctuations through the type 2 pathway of apoptosis*. Biophys J, 2008. **95**(8): p. 3559-62.
99. Li, Y., et al., *Neither replication nor simulation supports a role for the axon guidance pathway in the genetics of Parkinson's disease*. PLoS One, 2008. **3**(7): p. e2707.
100. Mukherjee, A., et al., *On the molecular mechanism of drug intercalation into DNA: a simulation study of the intercalation pathway, free energy, and DNA structural changes*. J Am Chem Soc, 2008. **130**(30): p. 9747-55.

101. Arikuma, T., et al., *Drug interaction prediction using ontology-driven hypothetical assertion framework for pathway generation followed by numerical simulation*. BMC Bioinformatics, 2008. 9 Suppl 6: p. S11.
102. Li, W., et al., *Possible pathway(s) of metyrapone egress from the active site of cytochrome P450 3A4: a molecular dynamics simulation*. Drug Metab Dispos, 2007. 35(4): p. 689-96.
103. Suresh Babu, C.V., et al., *Simulation and sensitivity analysis of phosphorylation of EGFR signal transduction pathway in PC12 cell model*. Syst Biol (Stevenage), 2004. 1(2): p. 213-21.
104. Xu, J. and G.A. Voth, *Computer simulation of explicit proton translocation in cytochrome c oxidase: the D-pathway*. Proc Natl Acad Sci U S A, 2005. 102(19): p. 6795-800.
105. Li, W., et al., *Possible pathway(s) of testosterone egress from the active site of cytochrome P450 2B1: a steered molecular dynamics simulation*. Drug Metab Dispos, 2005. 33(7): p. 910-9.
106. Xu, C., et al., *Simulation of a mathematical model of the role of the TFPI in the extrinsic pathway of coagulation*. Comput Biol Med, 2005. 35(5): p. 435-45.
107. Matrai, J., et al., *Simulation of the activation of alpha-chymotrypsin: analysis of the pathway and role of the propeptide*. Protein Sci, 2004. 13(12): p. 3139-50.
108. Sung, M.H. and R. Simon, *In silico simulation of inhibitor drug effects on nuclear factor-kappaB pathway dynamics*. Mol Pharmacol, 2004. 66(1): p. 70-5.
109. Visser, D. and J.J. Heijnen, *Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics*. Metab Eng, 2003. 5(3): p. 164-76.
110. Chassagnole, C., et al., *Dynamic simulation of pollutant effects on the threonine pathway in Escherichia coli*. C R Biol, 2003. 326(5): p. 501-8.
111. Ung, C.Y., et al., *Simulation of the regulation of EGFR endocytosis and EGFR-ERK signaling by endophilin-mediated RhoA-EGFR crosstalk*. FEBS Lett, 2008. 582(15): p. 2283-90.
112. Wei, D. and S. Mashima, *Prediction of accessory pathway locations in Wolff-Parkinson-White syndrome with body surface potential Laplacian maps . A simulation study*. Jpn Heart J, 1999. 40(4): p.

- 451-9.
113. Bond, C.J., et al., *Characterization of residual structure in the thermally denatured state of barnase by simulation and experiment: description of the folding pathway*. Proc Natl Acad Sci U S A, 1997. **94**(25): p. 13409-13.
 114. Darji, A., et al., *In vitro simulation of immunosuppression caused by Trypanosoma brucei: active involvement of gamma interferon and tumor necrosis factor in the pathway of suppression*. Infect Immun, 1996. **64**(6): p. 1937-43.
 115. Cachau, R.E., et al., *Computer simulation and analysis of the reaction pathway for the decomposition of the hydrated peptide bond in aspartic proteases*. Adv Exp Med Biol, 1995. **362**: p. 461-5.
 116. Jones, K.C. and K.G. Mann, *A model for the tissue factor pathway to thrombin. II. A mathematical simulation*. J Biol Chem, 1994. **269**(37): p. 23367-73.
 117. Bash, P.A., et al., *Computer simulation and analysis of the reaction pathway of triosephosphate isomerase*. Biochemistry, 1991. **30**(24): p. 5826-32.
 118. Jackson, R.C., *Kinetic simulation of anticancer drug effects on metabolic pathway fluxes: two case studies*. Bull Math Biol, 1986. **48**(3-4): p. 337-51.
 119. Duggleby, R.G. and R.I. Christopherson, *Metabolic resistance to tight-binding inhibitors of enzymes involved in the de novo pyrimidine pathway. Simulation of time-dependent effects*. Eur J Biochem, 1984. **143**(1): p. 221-6.
 120. Koh, G., et al., *A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk*. Bioinformatics, 2006. **22**(14): p. e271-80.
 121. Kutalik, Z., K.H. Cho, and O. Wolkenhauer, *Optimal sampling time selection for parameter estimation in dynamic pathway modeling*. Biosystems, 2004. **75**(1-3): p. 43-55.
 122. Sorribas, A., et al., *Metabolic pathway characterization from transient response data obtained in situ: parameter estimation in S-system models*. J Theor Biol, 1993. **162**(1): p. 81-102.
 123. Lambrou, G.I., et al., *Pathway simulations in common oncogenic drivers of leukemic and rhabdomyosarcoma cells: a systems biology*

- approach*. Int J Oncol, 2012. **40**(5): p. 1365-90.
124. Mosca, E., et al., *Systems biology of the metabolic network regulated by the Akt pathway*. Biotechnol Adv, 2012. **30**(1): p. 131-41.
 125. Jegga, A.G., et al., *Systems biology of the autophagy-lysosomal pathway*. Autophagy, 2011. **7**(5): p. 477-89.
 126. Zhang, F. and J.Y. Chen, *Discovery of pathway biomarkers from coupled proteomics and systems biology methods*. BMC Genomics, 2010. **11 Suppl 2**: p. S12.
 127. Navas-Delgado, I., et al., *Social pathway annotation: extensions of the systems biology metabolic modelling assistant*. Brief Bioinform, 2011. **12**(6): p. 576-87.
 128. Ozbayraktar, F.B. and K.O. Ulgen, *Drug target identification in sphingolipid metabolism by computational systems biology tools: metabolic control analysis and metabolic pathway analysis*. J Biomed Inform, 2010. **43**(4): p. 537-49.
 129. Li, S., Q. Lu, and Y. Cui, *A systems biology approach for identifying novel pathway regulators in eQTL mapping*. J Biopharm Stat, 2010. **20**(2): p. 373-400.
 130. Tiruppathi, C., et al., *Thrombin receptor 14-amino acid peptide binds to endothelial cells and stimulates calcium transients*. Am J Physiol, 1992. **263**(5 Pt 1): p. L595-601.
 131. Hucka, M., et al., *Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project*. Syst Biol (Stevenage), 2004. **1**(1): p. 41-53.
 132. Hucka, M., et al., *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models*. Bioinformatics, 2003. **19**(4): p. 524-31.
 133. Ma, Q., et al., *[Differential expression of colon cancer microRNA in microarray study]*. Sichuan Da Xue Xue Bao Yi Xue Ban, 2011. **42**(3): p. 344-8.
 134. Sundaresh, S., et al., *How noisy and replicable are DNA microarray data?* Int J Bioinform Res Appl, 2005. **1**(1): p. 31-50.
 135. Babu, M.M., *An Introduction to Microarray Data Analysis*. Computational Genomics: Theory and Application, ed. C. Richard P.

- Grant Laboratory of Molecular Biology, UK. 2004: Horizon Bioscience.
136. Leung, Y.F. and D. Cavalieri, *Fundamentals of cDNA microarray data analysis*. Trends Genet, 2003. **19**(11): p. 649-59.
137. Pinkel, D., et al., *High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays*. Nat Genet, 1998. **20**(2): p. 207-11.
138. Hughes, T.R., et al., *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer*. Nat Biotechnol, 2001. **19**(4): p. 342-7.
139. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays*. Nat Biotechnol, 1996. **14**(13): p. 1675-80.
140. Dalma-Weiszhausz, D.D., et al., *The affymetrix GeneChip platform: an overview*. Methods Enzymol, 2006. **410**: p. 3-28.
141. Michiels, S., S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy*. Lancet, 2005. **365**(9458): p. 488-92.
142. Schoch C, D.M., Kern W, Kohlmann A, Schnittger S, Haferlach T, *"Deep insight" into microarray technology*. Atlas Genet Cytogenet Oncol Haematol, 2004.
143. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science, 1997. **278**(5338): p. 680-6.
144. Tavazoie, S., et al., *Systematic determination of genetic network architecture*. Nat Genet, 1999. **22**(3): p. 281-5.
145. Jansen, R., D. Greenbaum, and M. Gerstein, *Relating whole-genome expression data with protein-protein interactions*. Genome Res, 2002. **12**(1): p. 37-46.
146. Ramirez-Benitez Mdel, C., G. Moreno-Hagelsieb, and J.C. Almagro, *VIR.II: a new interface with the antibody sequences in the Kabat database*. Biosystems, 2001. **61**(2-3): p. 125-31.
147. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. Proc Natl Acad Sci U S A, 1999. **96**(12): p.

- 6745-50.
148. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns*. Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
 149. van der Pouw Kraan, T.C., et al., *Discovery of distinctive gene expression profiles in rheumatoid synovium using cDNA microarray technology: evidence for the existence of multiple pathways of tissue destruction and repair*. Genes Immun, 2003. **4**(3): p. 187-96.
 150. Sherlock, G., *Analysis of large-scale gene expression data*. Curr Opin Immunol, 2000. **12**(2): p. 201-5.
 151. Vapnik, V., *Statistical Learning Theory*. 1998.
 152. Bishop, C., *neural networks for pattern recognition*. 1995.
 153. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures*. Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15149-54.
 154. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*. Nat Med, 2001. **7**(6): p. 673-9.
 155. Qiu, P., Z.J. Wang, and K.J. Liu, *Ensemble dependence model for classification and prediction of cancer and normal gene expression data*. Bioinformatics, 2005. **21**(14): p. 3114-21.
 156. Li, F. and Y. Yang, *Analysis of recursive gene selection approaches from microarray data*. Bioinformatics, 2005. **21**(19): p. 3741-7.
 157. Pochet, N., et al., *Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction*. Bioinformatics, 2004. **20**(17): p. 3185-95.
 158. Isabelle Guyon, J.W., Stephen Barnhill, Vladimir Vapnik, *Gene Selection for Cancer Classification using Support Vector Machines*. Machine Learning, 2002. **46**(1-3): p. 389-422.
 159. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data*. Bioinformatics, 2000. **16**(10): p. 906-14.
 160. Brown, M.P., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines*. Proc Natl Acad Sci

- U S A, 2000. **97**(1): p. 262-7.
161. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome*. Nat Genet, 2004. **36**(9): p. 949-51.
162. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome*. Science, 2004. **305**(5683): p. 525-8.
163. Piotrowski, A., et al., *Somatic mosaicism for copy number variation in differentiated human tissues*. Hum Mutat, 2008. **29**(9): p. 1118-24.
164. Beckmann, J.S., X. Estivill, and S.E. Antonarakis, *Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability*. Nat Rev Genet, 2007. **8**(8): p. 639-46.
165. Marques-Bonet, T., et al., *A burst of segmental duplications in the genome of the African great ape ancestor*. Nature, 2009. **457**(7231): p. 877-81.
166. Volik, S., et al., *Decoding the fine-scale structure of a breast cancer genome and transcriptome*. Genome Res, 2006. **16**(3): p. 394-404.
167. Debies, M.T. and D.R. Welch, *Genetic basis of human breast cancer metastasis*. J Mammary Gland Biol Neoplasia, 2001. **6**(4): p. 441-51.
168. Hong, S.M., et al., *Genome-Wide Somatic Copy Number Alterations in Low-Grade PanINs and IPMNs from Individuals with a Family History of Pancreatic Cancer*. Clin Cancer Res, 2012.
169. *Copy number aberrations define breast cancer subgroups*. Cancer Discov, 2012. **2**(6): p. OF6.
170. Krepischi, A.C., et al., *Germline DNA copy number variation in familial and early-onset breast cancer*. Breast Cancer Res, 2012. **14**(1): p. R24.
171. Ueno, T., et al., *Genome-wide copy number analysis in primary breast cancer*. Expert Opin Ther Targets, 2012. **16 Suppl 1**: p. S31-5.
172. Nielsen, K.V., et al., *Lack of independent prognostic and predictive value of centromere 17 copy number changes in breast cancer patients with known HER2 and TOP2A status*. Mol Oncol, 2012. **6**(1): p. 88-97.
173. Lonigro, R.J., et al., *Detection of somatic copy number alterations in cancer using targeted exome capture sequencing*. Neoplasia, 2011. **13**(11): p. 1019-25.

174. Struski, S., M. Doco-Fenzy, and P. Cornillet-Lefebvre, *Compilation of published comparative genomic hybridization studies*. Cancer Genet Cytogenet, 2002. **135**(1): p. 63-90.
175. Bellman., R.E., *Adaptive Control Processes*. 1961.
176. Koeppen, M., *The Curse of Dimensionality*. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), 2000.
177. Inza, I., et al., *Filter versus wrapper gene selection approaches in DNA microarray domains*. Artif Intell Med, 2004. **31**(2): p. 91-103.
178. Model, F., et al., *Feature selection for DNA methylation based cancer classification*. Bioinformatics, 2001. **17 Suppl 1**: p. S157-64.
179. Robnik-Šikonja, M. and I. Kononenko, *Theoretical and Empirical Analysis of ReliefF and RReliefF*. Machine Learning, 2003. **53**(1-2): p. 23-69.
180. Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data*. J Bioinform Comput Biol, 2005. **3**(2): p. 185-205.
181. Ben-Bassat, M., *Pattern recognition and reduction of dimensionality*. Handbook of statistics II, 1982: p. p. 773—91.
182. Cheng, J. and R. Greiner, *Comparing Bayesian Network Classifiers*. Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99), 1999: p. 101-10.
183. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science, 1999. **286**(5439): p. 531-7.
184. Aris V, R.M., *A method to improve detection of disease using selectively expressed genes in microarray data*. Methods of Microarray Data Analysis. Proceedings of the First Conference on Critical Assessment of Microarray Data Analysis, CAMDA'00., 2002. **p. 69—80**.
185. Beibel, M., *Selection of informative genes in gene expression based diagnosis: a nonparametric approach*. Lecture Notes in Computer Sciences. Proceedings of the First International Symposium in Medical Data Analysis, ISMDA'00, 2000. **1933**: p. p. 300-7.
186. Ding, C., *Analysis of gene expression profiles: class discovery and leaf*

- ordering*. Proceedings of the Sixth International Conference on Research in Computational Molecular Biology, 2002: p. p. 127-36.
187. Baker, S.G. and B.S. Kramer, *Identifying genes that contribute most to good classification in microarrays*. BMC Bioinformatics, 2006. **7**: p. 407.
188. Kohavi, R. and G.H. John, *Wrappers for feature subset selection*. Artificial Intelligence, 97 **Special issue on relevance**(1-2): p. 273 - 324
189. Xiong, M., X. Fang, and J. Zhao, *Biomarker identification by feature wrappers*. Genome Res, 2001. **11**(11): p. 1878-87.
190. Kohavi, R. and G.H. John, *Wrappers for feature subset selection*. Artificial Intelligence, 1997 **Special issue on relevance**(1-2): p. 273 - 324
191. Talvinen, K., et al., *Biochemical and clinical approaches in evaluating the prognosis of colon cancer*. Anticancer Res, 2006. **26**(6C): p. 4745-51.
192. Ancona, N., et al., *On the statistical assessment of classifiers using DNA microarray data*. BMC Bioinformatics, 2006. **7**: p. 387.
193. Zhang, X.W., et al., *Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis*. Eur J Hum Genet, 2005. **13**(12): p. 1303-11.
194. Li, W. and Y. Yang, *How Many Genes Are Needed for a Discriminant Microarray Data Analysis ?* Methods of Microarray Data Analysis. Proceedings of the First Conference on Critical Assessment of Microarray Data Analysis, CAMDA'00., 2002: p. 137-150.
195. Grate, L.R., *Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery*. BMC Bioinformatics, 2005. **6**: p. 97.
196. Slonim, D.K., et al., *Class prediction and discovery using gene expression data*. Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB), 2000.
197. Ahmed, A.A. and J.D. Brenton, *Microarrays and breast cancer clinical studies: forgetting what we have not yet learnt*. Breast Cancer Res, 2005. **7**(3): p. 96-9.
198. Brenton, J.D., et al., *Molecular classification and molecular*

- forecasting of breast cancer: ready for clinical application?* J Clin Oncol, 2005. **23**(29): p. 7350-60.
199. Bullinger, L. and P.J. Valk, *Gene expression profiling in acute myeloid leukemia*. J Clin Oncol, 2005. **23**(26): p. 6296-305.
200. Conzelmann, H., et al., *Reduction of mathematical models of signal transduction networks: simulation-based approach applied to EGF receptor signalling*. Syst Biol (Stevenage), 2004. **1**(1): p. 159-69.
201. Valk, P.J.M., et al., *Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia*. N Engl J Med, 2004. **350**(16): p. 1617-1628.
202. Ntzani, E.E. and J.P. Ioannidis, *Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment*. Lancet, 2003. **362**(9394): p. 1439-44.
203. Zhou, X. and K.Z. Mao, *LS Bound based gene selection for DNA microarray data*. Bioinformatics, 2005. **21**(8): p. 1559-64.
204. Bo, T. and I. Jonassen, *New feature subset selection procedures for classification of expression profiles*. Genome Biol, 2002. **3**(4): p. RESEARCH0017.
205. Huang, T.M. and V. Kecman, *Gene extraction for cancer diagnosis by support vector machines--an improvement*. Artif Intell Med, 2005. **35**(1-2): p. 185-94.
206. Liu, X., A. Krishnan, and A. Mondry, *An entropy-based gene selection method for cancer classification using microarray data*. BMC Bioinformatics, 2005. **6**(1): p. 76.
207. Draghici, S., et al., *Reliability and reproducibility issues in DNA microarray measurements*. Trends Genet, 2006. **22**(2): p. 101-9.
208. Ioannidis, J.P., *Microarrays and molecular research: noise discovery?* Lancet, 2005. **365**(9458): p. 454-5.
209. Gardner, S.N. and M. Fernandes, *Prediction of cancer outcome with microarrays*. Lancet, 2005. **365**(9472): p. 1685.
210. Biganzoli, E., et al., *Prediction of cancer outcome with microarrays*. Lancet, 2005. **365**(9472): p. 1683; author reply 1684-5.
211. Winegarden, N., *Microarrays in cancer: moving from hype to clinical reality*. Lancet, 2003. **362**(9394): p. 1428.

212. Wang, W., M.J. Merrill, and R.T. Borchardt, *Vascular endothelial growth factor affects permeability of brain microvessel endothelial cells in vitro*. Am J Physiol, 1996. **271**(6 Pt 1): p. C1973-80.
213. Cao, J., X. Qi, and H. Zhao, *Modeling gene regulation networks using ordinary differential equations*. Methods Mol Biol, 2012. **802**: p. 185-97.
214. Cheng, J., et al., *Real-time vector quantization and clustering based on ordinary differential equations*. IEEE Trans Neural Netw, 2011. **22**(12): p. 2143-8.
215. Berglund, M., et al., *Investigations of a compartmental model for leucine kinetics using non-linear mixed effects models with ordinary and stochastic differential equations*. Math Med Biol, 2011.
216. Soliman, S. and M. Heiner, *A unique transformation from ordinary differential equations to reaction networks*. PLoS One, 2010. **5**(12): p. e14284.
217. Filici, C., *Error estimation in the neural network solution of ordinary differential equations*. Neural Netw, 2010. **23**(5): p. 614-7.
218. Klipp, E., et al., *Integrative model of the response of yeast to osmotic shock*. Nat Biotechnol, 2005. **23**(8): p. 975-82.
219. Ham, B., D. Min, and K. Sohn, *A Robust Scale-Space Filter using Second Order Partial Differential Equations*. IEEE Trans Image Process, 2012.
220. Zhang, K., G. Achari, and H. Li, *A comparison of numerical solutions of partial differential equations with probabilistic and possibilistic parameters for the quantification of uncertainty in subsurface solute transport*. J Contam Hydrol, 2009. **110**(1-2): p. 45-59.
221. Cho, K.H., et al., *A new methodology for determining dispersion coefficient using ordinary and partial differential transport equations*. Water Sci Technol, 2009. **59**(11): p. 2197-203.
222. Tang, C., et al., *The oriented-couple partial differential equations for filtering in wrapped phase patterns*. Opt Express, 2009. **17**(7): p. 5606-17.
223. Caselles, V. and J. Morel, *Introduction to the special issue on partial differential equations and geometry-driven diffusion in image processing and analysis*. IEEE Trans Image Process, 1998. **7**(3): p. 269-73.

-
224. Lagaris, I.E., A. Likas, and D.I. Fotiadis, *Artificial neural networks for solving ordinary and partial differential equations*. IEEE Trans Neural Netw, 1998. **9**(5): p. 987-1000.
225. Traulsen, A., J.C. Claussen, and C. Hauert, *Stochastic differential equations for evolutionary dynamics with demographic noise and mutations*. Phys Rev E Stat Nonlin Soft Matter Phys, 2012. **85**(4-1): p. 041901.
226. Cuenod, C.A., et al., *Parameter estimation and change-point detection from Dynamic Contrast Enhanced MRI data using stochastic differential equations*. Math Biosci, 2011. **233**(1): p. 68-76.
227. Atalla, A. and A. Jeremic, *Modeling bacterial clearance using stochastic-differential equations*. Conf Proc IEEE Eng Med Biol Soc, 2010. **2010**: p. 746-51.
228. Brown, S.D., R. Ratcliff, and P.L. Smith, *Evaluating methods for approximating stochastic differential equations*. J Math Psychol, 2006. **50**(4): p. 402-410.
229. Kopoplia, E.F., et al., *[Kinetic parameters of androgen receptor complexes and the activities of the glycolysis and oxidative pentose phosphate pathway key enzymes in rat testis cytosol after whole body 60-min exposure to high frequency electromagnetic field (39.5 Ghz)]*. Radiats Biol Radioecol, 2003. **43**(5): p. 535-7.
230. Hatakeyama, M., et al., *A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signalling*. Biochem J, 2003. **373**(Pt 2): p. 451-63.
231. Jaqaman, K. and G. Danuser, *Linking data to models: data regression*. Nat Rev Mol Cell Biol, 2006. **7**(11): p. 813-9.
232. Moles, C.G., P. Mendes, and J.R. Banga, *Parameter estimation in biochemical pathways: a comparison of global optimization methods*. Genome Res, 2003. **13**(11): p. 2467-74.
233. Li, W., et al., *Highly discriminating protein-protein interaction specificities in the context of a conserved binding energy hotspot*. J Mol Biol, 2004. **337**(3): p. 743-59.
234. Gabdoulline, R.R., M. Stein, and R.C. Wade, *qPIPSA: relating enzymatic kinetic parameters and interaction fields*. BMC Bioinformatics, 2007. **8**: p. 373.

235. Aldridge, B.B., et al., *Physicochemical modelling of cell signalling pathways*. Nat Cell Biol, 2006. **8**(11): p. 1195-203.
236. Gutenkunst, R.N., et al., *Universally sloppy parameter sensitivities in systems biology models*. PLoS Comput Biol, 2007. **3**(10): p. 1871-78.
237. Komorowski, M., et al., *Sensitivity, robustness, and identifiability in stochastic chemical kinetics models*. Proc Natl Acad Sci U S A, 2011. **108**(21): p. 8645-50.
238. Schuchhardt, J., et al., *Normalization strategies for cDNA microarrays*. Nucleic Acids Res, 2000. **28**(10): p. E47.
239. Tu, Y., G. Stolovitzky, and U. Klein, *Quantitative noise analysis for gene expression microarray experiments*. Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14031-6.
240. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*. Nature, 2000. **403**(6769): p. 503-11.
241. Bo, T.H., B. Dysvik, and I. Jonassen, *LSimpute: accurate estimation of missing values in microarray data with least squares methods*. Nucleic Acids Res, 2004. **32**(3): p. e34.
242. de Brevern, A.G., S. Hazout, and A. Malpertuy, *Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering*. BMC Bioinformatics, 2004. **5**: p. 114.
243. Hu, J., et al., *Integrative missing value estimation for microarray data*. BMC Bioinformatics, 2006. **7**: p. 449.
244. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays*. Bioinformatics, 2001. **17**(6): p. 520-5.
245. Kim, H., G.H. Golub, and H. Park, *Missing value estimation for DNA microarray gene expression data: local least squares imputation*. Bioinformatics, 2005. **21**(2): p. 187-98.
246. Oba, S., et al., *A Bayesian missing value estimation method for gene expression profile data*. Bioinformatics, 2003. **19**(16): p. 2088-96.
247. Scholz, M., et al., *Non-linear PCA: a missing data approach*. Bioinformatics, 2005. **21**(20): p. 3887-95.
248. Demeter, J., et al., *The Stanford Microarray Database: implementation of new analysis tools and open source release of software*. Nucleic

- Acids Res, 2007. **35**(Database issue): p. D766-70.
249. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*. Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
250. Bair, E. and R. Tibshirani, *Semi-supervised methods to predict patient survival from gene expression data*. PLoS Biol, 2004. **2**(4): p. E108.
251. Scheel, I., et al., *The influence of missing value imputation on detection of differentially expressed genes from microarray data*. Bioinformatics, 2005. **21**(23): p. 4272-9.
252. <http://helix-web.stanford.edu/pubs/impute/>. Available from: <http://helix-web.stanford.edu/pubs/impute/>.
253. Lee, P.D., et al., *Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies*. Genome Res, 2002. **12**(2): p. 292-7.
254. Norman Morrison, M.R., Martin Brutsche, Stephen G. Oliver, Andrew Hayes, Nianshu Zhang, Chris Penkett, Jacqui Lockey, Sudha Rao, Ian Hayes, Ray Jupp, Andy Brass, *Robust normalization of microarray data over multiple experiments*. Nature Genetics, 1999. **23**: p. 64.
255. Chu, W., et al., *Biomarker discovery in microarray gene expression data with Gaussian processes*. Bioinformatics, 2005. **21**(16): p. 3385-93.
256. Michael E. Wall, A.R., Luis M. Rocha, *Microarray analysis techniques: Singular value decomposition and principal component analysis*. Understanding and Using Microarray Analysis Techniques: A Practical Guide, ed. W.D. D.P. Berrar, M. Granzow. 2002: Kluwer Academic Press.
257. Wang, K., et al., *PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data*. Genome Res, 2007. **17**(11): p. 1665-74.
258. Colella, S., et al., *QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data*. Nucleic Acids Res, 2007. **35**(6): p. 2013-25.
259. Vapnik, V., *Estimation of dependences based on empirical data [in Russian]*. [English translation: Springer Verlag, New York, 1982]. 1979.

-
260. Vapnik, V.N., *The nature of statistical learning theory*. 1995, New York: Springer.
261. Souheil Ben-Yacoub, Y.A., and Eddy Mayoraz, *Fusion of Face and Speech Data for Person Identity Verification*. IEEE transactions on neural networks, 1999. **10**: p. 1065-1074.
262. Karlsen, R.E.G, David J.; Gerhart, Grant R., *Target classification via support vector machines*. Optical Engineering, 2000. **39**(3): p. 704-711.
263. Shin, C.S.K., K.I. Park, M.H. Kim, H.J. , *Support vector machine-based text detection in digital video*. Pattern recognition, 2001. **34**: p. 527-529.
264. Yuan, Z., K. Burrage, and J.S. Mattick, *Prediction of protein solvent accessibility using support vector machines*. Proteins, 2002. **48**(3): p. 566-70.
265. Ding, C.H. and I. Dubchak, *Multi-class protein fold recognition using support vector machines and neural networks*. Bioinformatics, 2001. **17**(4): p. 349-58.
266. Hua, S. and Z. Sun, *A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach*. J Mol Biol, 2001. **308**(2): p. 397-407.
267. Bock, J.R. and D.A. Gough, *Predicting protein--protein interactions from primary structure*. Bioinformatics, 2001. **17**(5): p. 455-60.
268. Cai, C.Z., et al., *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence*. Nucleic Acids Res, 2003. **31**(13): p. 3692-7.
269. Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery 1998. **2**: p. 121-167.
270. Keerthi, S.S. and C.J. Lin, *Asymptotic behaviors of support vector machines with Gaussian kernel*. Neural Comput, 2003. **15**(7): p. 1667-89.
271. Lin, H.-T., C.-J. Lin, *A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods*. Technical report, Department of Computer Science, National Taiwan University. 2003.
272. Platt, J., *Fast Training of Support Vector Machines using Sequential*

- Minimal Optimization*. Advances in Kernel Methods - Support Vector Learning, ed. C.B. B. Schölkopf, and A. Smola, eds. 1998: MIT Press.
273. Gunnarsson, R.K. and J. Lanke, *The predictive value of microbiologic diagnostic tests if asymptomatic carriers are present*. Stat Med, 2002. **21**(12): p. 1773-85.
274. Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines*. Machine Learning, 2002. **46**(1-3): p. 389-422.
275. Sima, C., U. Braga-Neto, and E.R. Dougherty, *Superior feature-set ranking for small samples using bolstered error estimation*. Bioinformatics, 2005. **21**(7): p. 1046-54.
276. Fu, W.J., R.J. Carroll, and S. Wang, *Estimating misclassification error with small samples via bootstrap cross-validation*. Bioinformatics, 2005. **21**(9): p. 1979-86.
277. Nierodzik, M.L. and S. Karpatkin, *Thrombin induces tumor growth, metastasis, and angiogenesis: Evidence for a thrombin-regulated dormant tumor phenotype*. Cancer Cell, 2006. **10**(5): p. 355-62.
278. Martorell, L., et al., *Thrombin and protease-activated receptors (PARs) in atherothrombosis*. Thromb Haemost, 2008. **99**(2): p. 305-15.
279. Finigan, J.H., *The coagulation system and pulmonary endothelial function in acute lung injury*. Microvasc Res, 2009. **77**(1): p. 35-8.
280. Jutel, M., K. Blaser, and C.A. Akdis, *The role of histamine in regulation of immune responses*. Chem Immunol Allergy, 2006. **91**: p. 174-87.
281. Coughlin, S.R., *Thrombin signalling and protease-activated receptors*. Nature, 2000. **407**(6801): p. 258-64.
282. van Nieuw Amerongen, G.P., et al., *Transient and prolonged increase in endothelial permeability induced by histamine and thrombin: role of protein kinases, calcium, and RhoA*. Circ Res, 1998. **83**(11): p. 1115-23.
283. Keck, P.J., et al., *Vascular permeability factor, an endothelial cell mitogen related to PDGF*. Science, 1989. **246**(4935): p. 1309-12.
284. Sun, H., et al., *Rho and ROCK signaling in VEGF-induced microvascular endothelial hyperpermeability*. Microcirculation, 2006. **13**(3): p. 237-47.

-
285. Langley, R.R. and I.J. Fidler, *Tumor cell-organ microenvironment interactions in the pathogenesis of cancer metastasis*. *Endocr Rev*, 2007. **28**(3): p. 297-321.
286. Vandenbroucke, E., et al., *Regulation of endothelial junctional permeability*. *Ann N Y Acad Sci*, 2008. **1123**: p. 134-45.
287. Hirano, K., et al., *Protein kinase network in the regulation of phosphorylation and dephosphorylation of smooth muscle myosin light chain*. *Mol Cell Biochem*, 2003. **248**(1-2): p. 105-14.
288. Wang, L. and S.M. Dudek, *Regulation of vascular permeability by sphingosine 1-phosphate*. *Microvasc Res*, 2009. **77**(1): p. 39-45.
289. Hu, G. and R.D. Minshall, *Regulation of transendothelial permeability by Src kinase*. *Microvasc Res*, 2009. **77**(1): p. 21-5.
290. Lukas, T.J., *A signal transduction pathway model prototype I: From agonist to cellular endpoint*. *Biophys J*, 2004. **87**(3): p. 1406-16.
291. Lukas, T.J., *A signal transduction pathway model prototype II: Application to Ca²⁺-calmodulin signaling and myosin light chain phosphorylation*. *Biophys J*, 2004. **87**(3): p. 1417-25.
292. Moraru, II and L.M. Loew, *Intracellular signaling: spatial and temporal control*. *Physiology (Bethesda)*, 2005. **20**: p. 169-79.
293. Maeda, A., et al., *Ca²⁺ -independent phospholipase A₂-dependent sustained Rho-kinase activation exhibits all-or-none response*. *Genes Cells*, 2006. **11**(9): p. 1071-83.
294. Viswanathan, G.A., et al., *Getting started in biological pathway construction and analysis*. *PLoS Comput Biol*, 2008. **4**(2): p. e16.
295. van Nieuw Amerongen, G.P. and V.W. van Hinsbergh, *Targets for pharmacological intervention of endothelial hyperpermeability and barrier function*. *Vascul Pharmacol*, 2002. **39**(4-5): p. 257-72.
296. Fajmut, A., A. Dobovisek, and M. Brumen, *Mathematical modeling of the relation between myosin phosphorylation and stress development in smooth muscles*. *J Chem Inf Model*, 2005. **45**(6): p. 1610-5.
297. Caunt, M., et al., *Growth-regulated oncogene is pivotal in thrombin-induced angiogenesis*. *Cancer Res*, 2006. **66**(8): p. 4125-32.
298. Zania, P., et al., *Thrombin mediates mitogenesis and survival of human endothelial cells through distinct mechanisms*. *Am J Physiol Cell*

- Physiol, 2008. **294**(5): p. C1215-26.
299. Grand, R.J., A.S. Turnell, and P.W. Grabham, *Cellular consequences of thrombin-receptor activation*. Biochem J, 1996. **313** (Pt 2): p. 353-68.
300. Parry, M.A., et al., *Cleavage of the thrombin receptor: identification of potential activators and inactivators*. Biochem J, 1996. **320** (Pt 1): p. 335-41.
301. Buhl, A.M., et al., *G alpha 12 and G alpha 13 stimulate Rho-dependent stress fiber formation and focal adhesion assembly*. J Biol Chem, 1995. **270**(42): p. 24631-4.
302. Cobb, M.H., *MAP kinase pathways*. Prog Biophys Mol Biol, 1999. **71**(3-4): p. 479-500.
303. Klemke, R.L., et al., *Regulation of cell motility by mitogen-activated protein kinase*. J Cell Biol, 1997. **137**(2): p. 481-92.
304. Amano, M., et al., *Phosphorylation and activation of myosin by Rho-associated kinase (Rho-kinase)*. J Biol Chem, 1996. **271**(34): p. 20246-9.
305. Hartshorne, D.J., M. Ito, and F. Erdodi, *Myosin light chain phosphatase: subunit composition, interactions and regulation*. J Muscle Res Cell Motil, 1998. **19**(4): p. 325-41.
306. Kamm, K.E. and J.T. Stull, *The function of myosin and myosin light chain kinase phosphorylation in smooth muscle*. Annu Rev Pharmacol Toxicol, 1985. **25**: p. 593-620.
307. Moussavi, R.S., C.A. Kelley, and R.S. Adelstein, *Phosphorylation of vertebrate nonmuscle and smooth muscle myosin heavy chains and light chains*. Mol Cell Biochem, 1993. **127-128**: p. 219-27.
308. Somlyo, A.P. and A.V. Somlyo, *Signal transduction and regulation in smooth muscle*. Nature, 1994. **372**(6503): p. 231-6.
309. Alessi, D., et al., *The control of protein phosphatase-1 by targeting subunits. The major myosin phosphatase in avian smooth muscle is a novel form of protein phosphatase-1*. Eur J Biochem, 1992. **210**(3): p. 1023-35.
310. Shirazi, A., et al., *Purification and characterization of the mammalian myosin light chain phosphatase holoenzyme. The differential effects of the holoenzyme and its subunits on smooth muscle*. J Biol Chem, 1994. **269**(50): p. 31598-606.

311. Johnson, D., et al., *Identification of the regions on the M110 subunit of protein phosphatase 1M that interact with the M21 subunit and with myosin*. Eur J Biochem, 1997. **244**(3): p. 931-9.
312. Kimura, K., et al., *Regulation of myosin phosphatase by Rho and Rho-associated kinase (Rho-kinase)*. Science, 1996. **273**(5272): p. 245-8.
313. Eto, M., et al., *A novel protein phosphatase-1 inhibitory protein potentiated by protein kinase C. Isolation from porcine aorta media and characterization*. J Biochem, 1995. **118**(6): p. 1104-7.
314. Li, C., et al., *BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models*. BMC Syst Biol, 2010. **4**: p. 92.
315. Kumar, P., et al., *Update of KDBI: Kinetic Data of Bio-molecular Interaction database*. Nucleic Acids Res, 2009. **37**(Database issue): p. D636-41.
316. Newton, A.C., *Protein kinase C: structure, function, and regulation*. J Biol Chem, 1995. **270**(48): p. 28495-8.
317. Matsumoto, T. and H. Mugishima, *Signal transduction via vascular endothelial growth factor (VEGF) receptors and their roles in atherogenesis*. J Atheroscler Thromb, 2006. **13**(3): p. 130-5.
318. Yamada, S., T. Taketomi, and A. Yoshimura, *Model analysis of difference between EGF pathway and FGF pathway*. Biochem Biophys Res Commun, 2004. **314**(4): p. 1113-20.
319. Kholodenko, B.N., et al., *Quantification of short term signaling by the epidermal growth factor receptor*. J Biol Chem, 1999. **274**(42): p. 30169-81.
320. Sasagawa, S., et al., *Prediction and validation of the distinct dynamics of transient and sustained ERK activation*. Nat Cell Biol, 2005. **7**(4): p. 365-73.
321. Roy, B. and J. Garthwaite, *Nitric oxide activation of guanylyl cyclase in cells revisited*. Proc Natl Acad Sci U S A, 2006. **103**(32): p. 12185-90.
322. Riento, K. and A.J. Ridley, *Rocks: multifunctional kinases in cell behaviour*. Nat Rev Mol Cell Biol, 2003. **4**(6): p. 446-56.
323. Singhal, M. and H. Resat, *A domain-based approach to predict*

- protein-protein interactions*. BMC Bioinformatics, 2007. **8**: p. 199.
324. Wojcik, J. and V. Schachter, *Protein-protein interaction map inference using interacting domain profile pairs*. Bioinformatics, 2001. **17 Suppl 1**: p. S296-305.
325. Coleman, T.F. and Y.Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*. Siam Journal on Optimization, 1996. **6**(2): p. 418-445.
326. Cameron, A.C. and F.A.G. Windmeijer, *An R-squared measure of goodness of fit for some common nonlinear regression models*. Journal of Econometrics, 1997. **77**(2): p. 329-342.
327. Hamby, D.M., *A Review of Techniques for Parameter Sensitivity Analysis of Environmental-Models*. Environmental Monitoring and Assessment, 1994. **32**(2): p. 135-154.
328. Martins, J.R.R.A., P. Sturdza, and J.J. Alonso, *The complex-step derivative approximation*. Acm Transactions on Mathematical Software, 2003. **29**(3): p. 245-262.
329. Goeckeler, Z.M. and R.B. Wysolmerski, *Myosin light chain kinase-regulated endothelial cell contraction: the relationship between isometric tension, actin polymerization, and myosin phosphorylation*. J Cell Biol, 1995. **130**(3): p. 613-27.
330. Kolodney, M.S. and E.L. Elson, *Correlation of myosin light chain phosphorylation with isometric contraction of fibroblasts*. J Biol Chem, 1993. **268**(32): p. 23850-5.
331. Zhi, G., et al., *Myosin light chain kinase and myosin phosphorylation effect frequency-dependent potentiation of skeletal muscle contraction*. Proc Natl Acad Sci U S A, 2005. **102**(48): p. 17519-24.
332. Benardeau, A., et al., *Contribution of Na⁺/Ca²⁺ exchange to action potential of human atrial myocytes*. Am J Physiol, 1996. **271**(3 Pt 2): p. H1151-61.
333. Tran, Q.K. and H. Watanabe, *Calcium signalling in the endothelium*. Handb Exp Pharmacol, 2006(176 Pt 1): p. 145-87.
334. Jeng, J.H., et al., *Protease-activated receptor-1-induced calcium signaling in gingival fibroblasts is mediated by sarcoplasmic reticulum calcium release and extracellular calcium influx*. Cell Signal, 2004. **16**(6): p. 731-40.

335. Birukova, A.A., et al., *Role of Rho GTPases in thrombin-induced lung vascular endothelial cells barrier dysfunction*. *Microvasc Res*, 2004. **67**(1): p. 64-77.
336. Yazaki, A., et al., *Inhibition by Rho-kinase and protein kinase C of myosin phosphatase is involved in thrombin-induced shape change of megakaryocytic leukemia cell line UT-7/TPO*. *Cell Signal*, 2005. **17**(3): p. 321-30.
337. Kureishi, Y., et al., *Rho-associated kinase directly induces smooth muscle contraction through myosin light chain phosphorylation*. *J Biol Chem*, 1997. **272**(19): p. 12257-60.
338. Senger, D.R., et al., *Tumor cells secrete a vascular permeability factor that promotes accumulation of ascites fluid*. *Science*, 1983. **219**(4587): p. 983-5.
339. Nelken, N.A., et al., *Thrombin receptor expression in normal and atherosclerotic human arteries*. *J Clin Invest*, 1992. **90**(4): p. 1614-21.
340. Tellez, C. and M. Bar-Eli, *Role and regulation of the thrombin receptor (PAR-1) in human melanoma*. *Oncogene*, 2003. **22**(20): p. 3130-7.
341. Jin, H.G., et al., *Hypoxia-induced upregulation of endothelial small G protein RhoA and Rho-kinase/ROCK2 inhibits eNOS expression*. *Neurosci Lett*, 2006. **408**(1): p. 62-7.
342. Li, B., et al., *Involvement of Rho/ROCK signalling in small cell lung cancer migration through human brain microvascular endothelial cells*. *FEBS Lett*, 2006. **580**(17): p. 4252-60.
343. Price, J.T., M.T. Bonovich, and E.C. Kohn, *The biochemistry of cancer dissemination*. *Crit Rev Biochem Mol Biol*, 1997. **32**(3): p. 175-253.
344. Worthylake, R.A., et al., *RhoA is required for monocyte tail retraction during transendothelial migration*. *J Cell Biol*, 2001. **154**(1): p. 147-60.
345. Adamson, P., et al., *Lymphocyte migration through brain endothelial cell monolayers involves signaling through endothelial ICAM-1 via a rho-dependent pathway*. *J Immunol*, 1999. **162**(5): p. 2964-73.
346. Ferrara, N., H.P. Gerber, and J. LeCouter, *The biology of VEGF and its receptors*. *Nat Med*, 2003. **9**(6): p. 669-76.
347. Chua, C.C., R.C. Hamdy, and B.H. Chua, *Upregulation of vascular endothelial growth factor by H2O2 in rat heart endothelial cells*. *Free*

- Radic Biol Med, 1998. **25**(8): p. 891-7.
348. Fiallo, P., et al., *Overexpression of vascular endothelial growth factor and its endothelial cell receptor KDR in type 1 leprosy reaction*. Am J Trop Med Hyg, 2002. **66**(2): p. 180-5.
349. Caldwell, R.B., et al., *Vascular endothelial growth factor and diabetic retinopathy: role of oxidative stress*. Curr Drug Targets, 2005. **6**(4): p. 511-24.
350. Ferrara, N. and K. Alitalo, *Clinical applications of angiogenic growth factors and their inhibitors*. Nat Med, 1999. **5**(12): p. 1359-64.
351. Padro, T., et al., *Overexpression of vascular endothelial growth factor (VEGF) and its cellular receptor KDR (VEGFR-2) in the bone marrow of patients with acute myeloid leukemia*. Leukemia, 2002. **16**(7): p. 1302-10.
352. Beynon, H.L., et al., *Combinations of low concentrations of cytokines and acute agonists synergize in increasing the permeability of endothelial monolayers*. Clin Exp Immunol, 1993. **91**(2): p. 314-9.
353. Csermely, P., V. Agoston, and S. Pongor, *The efficiency of multi-target drugs: the network approach might help drug design*. Trends Pharmacol Sci, 2005. **26**(4): p. 178-82.
354. Maragoudakis, M.E., et al., *Effects of thrombin/thrombosis in angiogenesis and tumour progression*. Matrix Biol, 2000. **19**(4): p. 345-51.
355. Roselli, M., et al., *Vascular endothelial growth factor (VEGF-A) plasma levels in non-small cell lung cancer: relationship with coagulation and platelet activation markers*. Thromb Haemost, 2003. **89**(1): p. 177-84.
356. Gieseler, F., et al., *Activated coagulation factors in human malignant effusions and their contribution to cancer cell metastasis and therapy*. Thromb Haemost, 2007. **97**(6): p. 1023-30.
357. MacDonald, J.A., et al., *Dual Ser and Thr phosphorylation of CPI-17, an inhibitor of myosin phosphatase, by MYPT-associated kinase*. FEBS Lett, 2001. **493**(2-3): p. 91-4.
358. Ohama, T., et al., *Chronic treatment with interleukin-1beta attenuates contractions by decreasing the activities of CPI-17 and MYPT-1 in intestinal smooth muscle*. J Biol Chem, 2003. **278**(49): p. 48794-804.

359. Sakai, H., et al., *Possible involvement of CPI-17 in augmented bronchial smooth muscle contraction in antigen-induced airway hyper-responsive rats*. Mol Pharmacol, 2005. **68**(1): p. 145-51.
360. Dakshinamurti, S., L. Mellow, and N.L. Stephens, *Regulation of pulmonary arterial myosin phosphatase activity in neonatal circulatory transition and in hypoxic pulmonary hypertension: a role for CPI-17*. Pediatr Pulmonol, 2005. **40**(5): p. 398-407.
361. Chang, S., et al., *Increased basal phosphorylation of detrusor smooth muscle myosin in alloxan-induced diabetic rabbit is mediated by upregulation of Rho-kinase beta and CPI-17*. Am J Physiol Renal Physiol, 2006. **290**(3): p. F650-6.
362. Woodsome, T.P., et al., *Expression of CPI-17 and myosin phosphatase correlates with Ca(2+) sensitivity of protein kinase C-induced contraction in rabbit smooth muscle*. J Physiol, 2001. **535**(Pt 2): p. 553-64.
363. Aslam, M., et al., *cAMP/PKA antagonizes thrombin-induced inactivation of endothelial myosin light chain phosphatase: role of CPI-17*. Cardiovasc Res, 2010. **87**(2): p. 375-84.
364. Angus, D.C., et al., *Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care*. Crit Care Med, 2001. **29**(7): p. 1303-10.
365. Hotchkiss, R.S. and I.E. Karl, *The pathophysiology and treatment of sepsis*. N Engl J Med, 2003. **348**(2): p. 138-50.
366. Lever, A. and I. Mackenzie, *Sepsis: definition, epidemiology, and diagnosis*. BMJ, 2007. **335**(7625): p. 879-83.
367. Zambon, M., et al., *Implementation of the Surviving Sepsis Campaign guidelines for severe sepsis and septic shock: we could go faster*. J Crit Care, 2008. **23**(4): p. 455-60.
368. *Biomarkers and surrogate endpoints: preferred definitions and conceptual framework*. Clin Pharmacol Ther, 2001. **69**(3): p. 89-95.
369. Larsson, T.P., et al., *Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery*. FEBS Lett, 2005. **579**(3): p. 690-8.
370. Zhang, W., R. Rekaya, and K. Bertrand, *A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer*.

- Bioinformatics, 2006. **22**(3): p. 317-25.
371. <http://rana.lbl.gov/EisenSoftware.htm>. Available from: <http://rana.lbl.gov/EisenSoftware.htm>.
372. Struyf, S., et al., *PARC/CCL18 is a plasma CC chemokine with increased levels in childhood acute lymphoblastic leukemia*. Am J Pathol, 2003. **163**(5): p. 2065-75.
373. Cao, G., et al., *Involvement of human PECAM-1 in angiogenesis and in vitro endothelial cell migration*. Am J Physiol Cell Physiol, 2002. **282**(5): p. C1181-90.
374. Casanova, M.L., et al., *Epidermal abnormalities and increased malignancy of skin tumors in human epidermal keratin 8-expressing transgenic mice*. Faseb J, 2004. **18**(13): p. 1556-8.
375. Song, S., et al., *Galectin-3 modulates MUC2 mucin expression in human colon cancer cells at the level of transcription via AP-1 activation*. Gastroenterology, 2005. **129**(5): p. 1581-91.
376. Mizoshita, T., et al., *Loss of MUC2 expression correlates with progression along the adenoma-carcinoma sequence pathway as well as de novo carcinogenesis in the colon*. Histol Histopathol, 2007. **22**(3): p. 251-60.
377. Liloglou, T., et al., *A T2517C polymorphism in the GSTM4 gene is associated with risk of developing lung cancer*. Lung Cancer, 2002. **37**(2): p. 143-6.
378. DeRubertis, F.R., R. Chayoth, and J.B. Field, *The content and metabolism of cyclic adenosine 3', 5'-monophosphate and cyclic guanosine 3', 5'-monophosphate in adenocarcinoma of the human colon*. J Clin Invest, 1976. **57**(3): p. 641-9.
379. Ushigome, M., et al., *Up-regulation of hnRNP A1 gene in sporadic human colorectal cancers*. Int J Oncol, 2005. **26**(3): p. 635-40.
380. Shang, L. and T.B. Tomasi, *The heat shock protein 90-CDC37 chaperone complex is required for signaling by types I and II interferons*. J Biol Chem, 2006. **281**(4): p. 1876-84.
381. Futreal, P.A., et al., *A census of human cancer genes*. Nat Rev Cancer, 2004. **4**(3): p. 177-83.
382. Meltzer, P.S., *Spotting the target: microarrays for disease gene discovery*. Curr Opin Genet Dev, 2001. **11**(3): p. 258-63.

-
383. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes*. Nature, 2008. **453**(7191): p. 56-64.
384. Redon, R., et al., *Global variation in copy number in the human genome*. Nature, 2006. **444**(7118): p. 444-54.
385. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome*. Nat Rev Genet, 2006. **7**(2): p. 85-97.
386. Dumas, L., et al., *Gene copy number variation spanning 60 million years of human and primate evolution*. Genome Res, 2007. **17**(9): p. 1266-77.
387. Nahon, J.L., *Birth of 'human-specific' genes during primate evolution*. Genetica, 2003. **118**(2-3): p. 193-208.
388. Bailey, J.A. and E.E. Eichler, *Primate segmental duplications: crucibles of evolution, diversity and disease*. Nat Rev Genet, 2006. **7**(7): p. 552-64.
389. Lupski, J.R., *Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits*. Trends Genet, 1998. **14**(10): p. 417-22.
390. Yang, L., et al., *Statistics on cancer in China: cancer registration in 2002*. Eur J Cancer Prev, 2005. **14**(4): p. 329-35.
391. Tchatchou, S. and B. Burwinkel, *Chromosome copy number variation and breast cancer risk*. Cytogenet Genome Res, 2008. **123**(1-4): p. 183-7.
392. Li, J., et al., *DNA copy number aberrations in breast cancer by array comparative genomic hybridization*. Genomics Proteomics Bioinformatics, 2009. **7**(1-2): p. 13-24.
393. Bush, N.J., *Advances in hormonal therapy for breast cancer*. Semin Oncol Nurs, 2007. **23**(1): p. 46-54.
394. Wang, H., et al., *Chemical data mining of the NCI human tumor cell line database*. J Chem Inf Model, 2007. **47**(6): p. 2063-76.
395. Workman, P., *Genomics and the second golden era of cancer drug development*. Mol Biosyst, 2005. **1**(1): p. 17-26.
396. Collins, I. and P. Workman, *New approaches to molecular cancer therapeutics*. Nat Chem Biol, 2006. **2**(12): p. 689-700.

397. Vogelstein, B. and K.W. Kinzler, *Cancer genes and the pathways they control*. Nat Med, 2004. **10**(8): p. 789-99.
398. de Castro Junior, G., et al., *Angiogenesis and cancer: A cross-talk between basic science and clinical trials (the "do ut des" paradigm)*. Crit Rev Oncol Hematol, 2006. **59**(1): p. 40-50.
399. Mancuso, A. and C.N. Sternberg, *Colorectal cancer and antiangiogenic therapy: what can be expected in clinical practice?* Crit Rev Oncol Hematol, 2005. **55**(1): p. 67-81.
400. Irish, J.M., N. Kotecha, and G.P. Nolan, *Mapping normal and cancer cell signalling networks: towards single-cell proteomics*. Nat Rev Cancer, 2006. **6**(2): p. 146-55.
401. Muller, A.J. and P.A. Scherle, *Targeting the mechanisms of tumoral immune tolerance with small-molecule inhibitors*. Nat Rev Cancer, 2006. **6**(8): p. 613-25.
402. Braun, P., et al., *An experimentally derived confidence score for binary protein-protein interactions*. Nat Methods, 2009. **6**(1): p. 91-7.

LIST OF PUBLICATION

1. Wei XN, Han BC, Zhang JX, Liu XH, Tan CY, Jiang YY, Low BC, Tidor B, Chen YZ*. An integrated mathematical model of thrombin-, histamine- and VEGF-mediated signalling in endothelial permeability. BMC Syst Biol. 2011 Jul 15; 5:112.
2. Wei XN. Mechanism of EGER-related cancer drug resistance. Anticancer Drugs. 2011 Nov; 22(10):963-70
3. Wei XN, Chen YZ*. Computational model of VEGF, thrombin and histamine signalling network. IEEE International Conference on Bioinformatics & Biomedicine (IEEE BIBM 2010), Hong Kong, IEEE Press.
4. Zhang JX, Han BC, Wei XN, C.Y. Tan, Y.Y. Jiang, Chen YZ. A two-step Target Binding and Selectivity Support Vector Machines Approach for Virtual Screening of Dopamine Receptor Subtype-selective Ligands. PLoS ONE 7(6): e39076. doi:10.1371/journal.pone.0039076 (2012).
5. X.H. Liu, H.Y. Song, J.X. Zhang, B.C. Han, X.N. Wei, X.H. Ma, W.K. Chui, Y.Z. Chen. Identifying Novel Type ZBGs and Non-hydroxamate HDAC Inhibitors Through a SVM Based Virtual Screening Approach. Mol Inf. 29(5): 407-20(2010).
6. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C, Chen Y*. Update of TTD: Therapeutic Target Database. Nucleic Acids Res. 2010 Jan; 38(Database issue):D787-91. Epub 2009 Nov 20
7. Zhang JX, J Jia, Ma XH, Han BC, Wei XN, C.Y. Tan, Y.Y. Jiang, Chen YZ. Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors. Mol. BioSyst., Advance Article, DOI: 10.1039/C2MB25165E. (2012)