

Computational Analysis of Genome-Wide DNA Copy Number Changes

Lei Song

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State
University in partial fulfillment of the requirements for the degree of

Master of Science

In

Electrical Engineering

Yue Wang (Chairman)

Jianhua Xuan

Chang-Tien Lu

May 3rd, 2011

Arlington, VA

Keywords: DNA Copy Number Changes, Circular Binary Segmentation,
Haar Wavelet Transform, Chromosome Instability Index,
Georgetown Database of Cancer

Copyright © 2011, Lei Song

Computational Analysis of Genome-Wide DNA Copy Number Changes

Lei Song

Abstract

DNA copy number change is an important form of structural variation in human genome. Somatic copy number alterations (CNAs) can cause over expression of oncogenes and loss of tumor suppressor genes in tumorigenesis. Recent development of SNP array technology has facilitated studies on copy number changes at a genome-wide scale, with high resolution.

Quantitative analysis of somatic CNAs on genes has found broad applications in cancer research. Most tumors exhibit genomic instability at chromosome scale as a result of dynamically accumulated genomic mutations during the course of tumor progression. Such higher level cancer genomic characteristics cannot be effectively captured by the analysis of individual genes. We introduced two definitions of chromosome instability (CIN) index to mathematically and quantitatively characterize genome-wide genomic instability. The proposed CIN indices are derived from detected CNAs using circular binary segmentation and wavelet transform, which calculates a score based on both the amplitude and frequency of the copy number changes. We generated CIN indices on ovarian cancer subtypes' copy number data and used them as features to train a SVM classifier. The experimental results show promising and high classification accuracy estimated through cross-validations. Additional survival analysis is constructed on the extracted CIN scores from TCGA ovarian cancer dataset and showed considerable correlation between CIN scores and various events and severity in ovarian cancer development.

Currently our methods have been integrated into G-DOC. We expect these newly defined CINs to be predictors in tumors subtype diagnosis and to be a useful tool in cancer research.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Yue Wang for his constant support and patient guidance throughout my thesis project. The discussions, encouragement and critiques made by him throughout the course of my project, were of essence to the progress of this work and I am very grateful to him for letting me have this opportunity to work with him. I am blessed to have such a wonderful supervisor who provided me such challenging work that I never seem to get enough of it.

I would also like to express my gratitude towards my committee members Dr. Jianhua Xuan and Dr. Chang-Tien Lu for their guidance and encouragement. They discussed with me over emails and were critical towards helping me correct the mistakes in thesis. Dr. Jianhua Xuan had taken so much interest and enthusiasm in my project, now I can still remember how he talked with me about the TCGA dataset and how he gave me help patiently when I faced difficulties.

I wish to thank Dr. Subha Madhavan, Principal Investigator of Lombardi Comprehensive Cancer Center (LCCC) at Georgetown University, who always has a further view of the main trend in the whole bioinformatics field and gives our whole group promising ideas. It is my honor to work as an intern in her group for G-DOC. And Dr. Yuriy Gusev, my mentor when I worked as an intern in LCCC, who always appreciated my ideas with great encourage even sometimes there were some minor mistakes, we had lots of discussion to integrate the copy number work into G-DOC. His encourage helped me a lot and sometimes he is so nice that even seems like a grandfather.

A special thank to Dr. Yuanjian Feng, Guoqiang Yu, Bai Zhang, Chen Wang and Lu Jin, who are my best friends and a source of positive energy and support. I shall thank Yuanjian for his patient introduction of copy number background to me and also his help when I had difficulty in programming. The inspiring discussions with Guoqiang, Chen, Jin and Bai quickly clarified some confusing concepts I was having. I also wish like to thank my friends Li Chen, Ye Tian, Jinghua Gu and Lily for their good wishes and encouragement.

My parents, can never thank them enough for whatever they have done for me.

Table of Contents

ACKNOWLEDGEMENTS	III
TABLE OF CONTENTS	IV
LIST OF FIGURES.....	VI
LIST OF TABLES.....	VI
LIST OF ACRONYMS	VII
CHAPTER 1.....	1
INTRODUCTION	1
1.1 MOTIVATION	1
1.2 OVERVIEW ON COPY NUMBER CHANGES	2
1.2.1 <i>Biological Background</i>	2
1.2.2 <i>Microarray Technologies for Copy Number Measurement</i>	5
1.2.3 <i>Copy Number Estimation from Microarray Measurement</i>	7
1.3 CHALLENGES, OBJECTIVES AND CONTRIBUTIONS	8
1.4 ORGANIZATION OF THE THESIS.....	10
CHAPTER 2.....	12
PRE-PROCESSING SNP ARRAY DATA.....	12
2.1 SNP ARRAYS NORMALIZATION.....	12
2.2 MODEL-BASED EXPRESSION INDEX (MBEI) ESTIMATION IN SNP ARRAYS	15
2.2.1 <i>PM/MM Design in SNP Array</i>	15
2.2.2 <i>Statistical Model</i>	15
2.2.3 <i>Conditional Mean and Standard Error</i>	17
2.3 REAL COPY NUMBER ESTIMATION BASED ON MBEI	19
CHAPTER 3.....	21
COPY NUMBER NORMALIZATION USING STANDARD FINITE NORMAL MIXTURE MODEL	21
3.1 INTRODUCTION TO STANDARD FINITE NORMAL MIXTURE MODEL	21
3.1.1 <i>Mixture of Gaussians</i>	21
3.1.2 <i>Likelihood Function</i>	23
3.2 EXPECTATION MAXIMIZATION ALGORITHM.....	24
3.2.1 <i>General EM Algorithm</i>	24
3.2.2 <i>EM Solution for SFNM</i>	26
3.3 COPY NUMBER SIGNAL NORMALIZATION THROUGH SFNM.....	27
3.3.1 <i>Transform Function for Copy Number Signal Normalization</i>	27
3.3.2 <i>Fit Copy Number Signal through SFNM</i>	29
CHAPTER 4.....	32
CHROMOSOME INSTABILITY INDEX.....	32
4.1 INTRODUCTION	32
4.2 CIN DEFINITION BASED ON CBS COPY NUMBER DETECTION.....	33
4.2.1 <i>Circular Binary Segmentation</i>	33
4.2.2 <i>CIN Definition From Gain/Loss on CBS Segments</i>	35
4.3 CIN DEFINITION BASED ON HAAR WAVELET TRANSFORM	36
4.3.1 <i>Haar Wavelet Transform of Copy Number Profile</i>	37
4.3.2 <i>FDR Thresholding</i>	38

4.3.3 <i>L-Vector Haar-CIN</i>	39
CHAPTER 5	40
EXPERIMENTS AND DISCUSSION	40
5.1 OVARIAN CANCER SUBTYPES EXPERIMENT	40
5.1.1 <i>SNP Dataset of Ovarian Cancer Subtypes</i>	40
5.1.2 <i>CIN-SVM Classification on Ovarian Cancer Subtypes</i>	43
5.2 SURVIVAL ANALYSIS ON TCGA OVARIAN CANCER DATASET.....	44
5.2.1 <i>TCGA Ovarian Cancer Dataset</i>	44
5.2.2 <i>Pearson's Correlation test</i>	44
5.2.3 <i>Cox Proportional-Hazards Regression</i>	45
5.3 DISCUSSION	47
CHAPTER 6	48
CONCLUSIONS AND FUTURE WORK	48
6.1 CONCLUSIONS.....	48
6.2 FUTURE WORK	49
APPENDIX	50
INTEGRATED APPLICATIONS INTO G-DOC	50
A. INTRODUCTION TO G-DOC.....	50
B. COPY NUMBER VISUALIZATION	52
C. CHROMOSOME INSTABILITY INDEX HEATMAP.....	53
D. VIEW COPY NUMBER PROFILES THROUGH JWBROWSER	54
REFERENCES	55

List of Figures

Figure 1.1 Human genome double helix structure and SNP illustration	3
Figure 1.2 Variations in human genomes	4
Figure 1.3 Oligonucleotide based CGH for copy number change identification	6
Figure 1.4 The framework of computational analysis of copy number changes	10
Figure 2.1 Sample SNP array images	13
Figure 2.2 Normalization of probe intensities between SNP arrays.....	14
Figure 2.3 PM/MM illustration in SNP array	15
Figure 2.4 PM/MM data view.....	18
Figure 2.5 Sample copy number profile visualization	20
Figure 3.1 A sample normal copy number profile	28
Figure 3.2 The original signal profile (gray) and the normalized profile (red)	29
Figure 3.3 The fitted SFNM model and MDL model selection.....	31
Figure 4.1 Copy number detection result of CBS.....	35
Figure 5.1 The chromosome CIN heatmap (a) and genome-wide CIN distribution (b) based on CBS detection results. Each column of the heatmap in (a) corresponds to a chromosome and each row corresponds to a sample; each dot in (b) represents a sample and each row corresponds to a phenotype.	41
Figure 5.2 Haar-CIN of the normal and SBT, LG, and HG ovarian cancers samples.....	42
Figure Appendix.A G-DOC welcoming	51
Figure Appendix.B Sample copy number profile visualization.....	52
Figure Appendix.C Sample chromosome instability index heatmap.....	53
Figure Appendix.D View copy number profiles through Jbrowser.....	54

List of Tables

Table 5.1 p-values of genome-wide CBS-CIN between subtypes.....	42
Table 5.2 SVM classification performance based on CBS-CIN and Haar-CIN.....	43
Table 5.3 p-values of Pearson's Correlation test and Cox Proportional-Hazards Regression	46

List of Acronyms

CIN	Chromosome Instability
CNA	Copy Number Alternation
CNV	Copy Number Variation
EM	Expectation Maximization
FDR	False Discovery Rate
G-DOC	Georgetown Database of Cancer
HapMap	Haplotype Map Project
HMM	Hidden Markov Model
MBEI	Model-Based Expression Index
MDL	Minimum Description Length
NCI	National Cancer Institute
NHGRI	National Human Genome Research Institute
SD	Standard Deviation
SFNM	Standard Finite Normal Mixture
SNP	Single Nucleotide Polymorphism
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas

Chapter 1

Introduction

This chapter gives the basic introduction to DNA copy number and outlines our work in the thesis. In section 1.1, we introduce the motivation of our research work in the thesis. In section 1.2 we provide an overview of copy number research, including biological background of DNA copy number changes, microarray technologies for copy number measurement, and real copy number estimation from the biological measurement. In section 1.3, we talk briefly about the procedures and methods in our analysis framework and also summarize our major contributions. The organization of the thesis is outlined in the final section 1.4.

1.1 Motivation

Over the past few decades, major advances of biotechnologies have led to an explosive growth in biological information and have been changing the practice of biomedical research. The human genome project, which firstly showed people the entire map of human genome in 2000 [1], along with some other following projects such as HapMap (Haplotype Map Project) [2] and TCGA [3] et al. have produced abundant genomic data at gene or molecular level. This deluge of genomic information has, on one hand, provided the important base for biological and biomedical research, on the other hand, led to an absolute requirement for computational methods to analyze the huge springing-out data.

It is widely believed that common diseases and even cancers root wholly or partially from genetic factors, but the detail cause-and-effect is still a secret. Generally speaking, the goal of biomedical and bioinformatics research is to discover disease mechanisms, help medical

diagnosis, develop new cures to human diseases, and suggest prevention strategies to improve human health. Single nucleotide polymorphisms (SNPs) [5] and copy number change cover most important genome variations between different individuals [9] while somatic copy number alterations (CNAs) [7] can cause over expression of oncogenes and loss of tumor suppressor genes. In this thesis, we will focus on the analysis of copy number changes, mainly CNAs more accurately speaking, and propose a novel framework for effectively and practically process huge size copy number data. In the framework, we will try to extract the hidden information behind the variations of copy number, to characterize the chromosome instability indicated by these CNAs from the high throughput raw copy number data and explore their correlations with hallmarks in cancer development. This kind of information and correlation should be helpful for researchers to explore causality between diseases, cancers and genetic variations.

1.2 Overview on Copy Number Changes

1.2.1 Biological Background

DNA is a polymeric molecule carrying the genetic instructions for the development and maintaining of a living organism [4]. DNA has a double helix structure (See Figure 1.1) – it consists of two inter-twisted strands of nucleotide bases, including adenine (A), cytosine (C), guanine (G), and thymine (T). The two complement DNA strands are bound by hydrogen bonds in a base-by-base fashion and contain essentially the same genetic information: A/T in one strand is always paired with T/C in the other strand. In a single normal human cell, DNA molecules are packed into twenty-three pairs of chromosomes. Each pair of chromosomes contains genetic information inherited from the parents, one from mother and the other from father. For human

beings, the DNA sequences of the twenty-three pairs of chromosomes, which totally consist of about 3×10^9 nucleotide base pairs, uniquely determine an individual's genome.

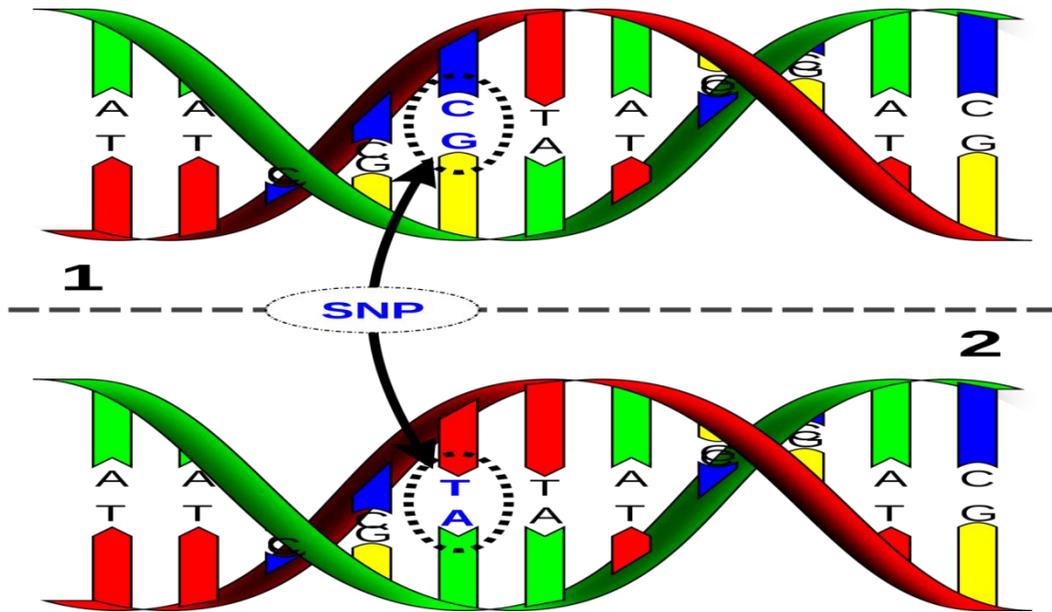


Figure 1.1 Human genome double helix structure and SNP illustration. (Public domain images via Wikipedia Commons GNU Free Documentation License.)

There are considerable differences between genomes of different individuals, and the diversity of individuals roots in these differences. Single nucleotide polymorphisms [6] and structural variations are two major types of genomic variations which are currently being intensively investigated. SNPs refer to single base differences among the genomes of population (Figure 1.1 above). For a particular SNP, there are usually a major allele with dominant frequency among population and a minor allele occurring less frequently in population. There are about 3.1 million SNPs among the 3 billion base pairs of the human genome. It was once believed that these 0.1% contents of the human genome account for the difference between individuals. Recent studies have revealed that the structural changes in the human genome occur

in a much larger scale compared with SNPs and may have more extensive impact on the diversity of population [7].

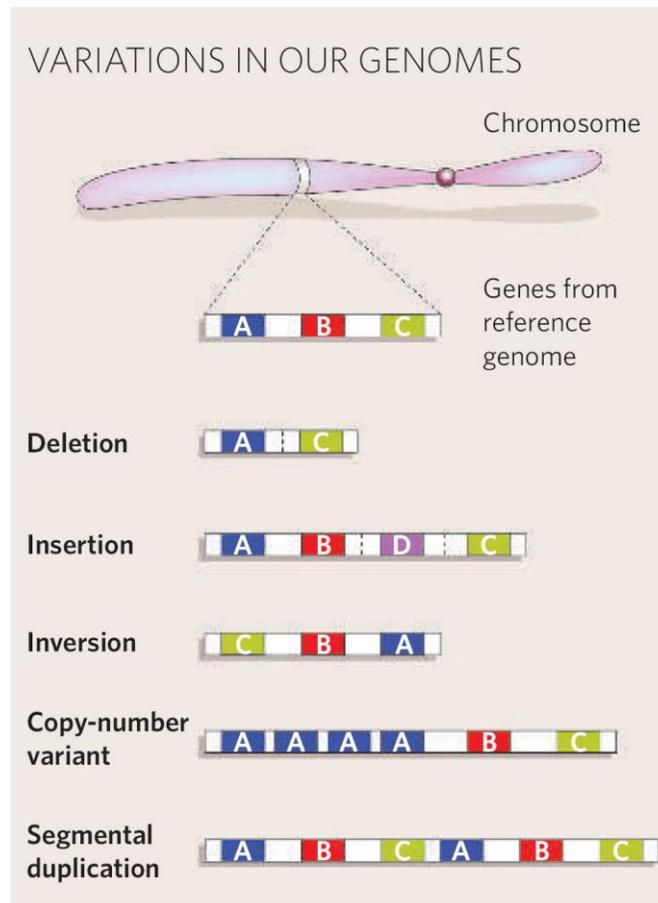


Figure 1.2 Variations in human genomes (Image from reference [8])

There are several forms of genomic structural variations, for instance, insertions, deletions, inversions, translocations of DNA segments and copy number changes (Figure 1.2). Among these structural variations, copy number changes have been mostly studied due to the recent advances in DNA microarray technologies. Copy number change refers to the phenomena that the number of copies of a particular DNA segment varies among individuals.

The scale of copy number changes spans from a few hundred nucleotide bases to millions of bases. It is estimated that more than 12% of the human genome are involved in copy number

changes [9]. Copy number changes are either acquired by heredity (germline copy number changes) or postnatal development (somatic copy number changes). Germline and somatic copy number changes are usually referred to as copy number variations (CNVs) and copy number alterations (CNAs), respectively. Inherited CNVs affect biological traits and susceptibilities to diseases. Similar to SNPs, the variants of a CNV (different copy numbers) can be cataloged in population and referred to as a copy number polymorphism. CNAs are usually found in cancers and other diseases. CNAs involving gains of cancer inducing genes and losses of tumor suppressor genes are considered as critical events in the development and progression of cancers.

1.2.2 Microarray Technologies for Copy Number Measurement

Technologies for measuring DNA copy number changes have been quickly evolving and the resolutions of measurements and coverage of the genome have been greatly improved. In the early days, fluorescence in situ hybridization (FISH) and comparative genomic hybridization (CGH) are two major technologies detecting variations in copy numbers that consist of more than several million bases [7]. In FISH, the copy number of a particular DNA segment is detected by hybridizing fluorescently labeled probes directly to the chromosome and observing the fluorescence intensity under microscope. CGH uses large-insert DNA clones as the probes, whose lengths are between 10^5 to 2×10^5 base pairs. A reference DNA sample and a target DNA sample are labeled with different fluorophores and competitively hybridized to a substrate spotted with probes. Measuring the fluorescence ratio along the length of each chromosome identified regions of relative loss and gain in the test sample (the reference DNA being assumed to be diploid in copy number). Although this method had a huge impact, a major drawback was the low resolution, typically 5–10 Mb which can be observed by microscopy [7], afforded by metaphase FISH. Such a scale is too coarse to reveal deeper levels of genetic variations.

Recent advances of microarray technologies enable researchers to study copy number changes in the sub-microscopic level, which covers several thousand bases up to three million bases. Representative high resolution methods include oligonucleotide based comparative genome hybridization and single nucleotide polymorphism arrays.

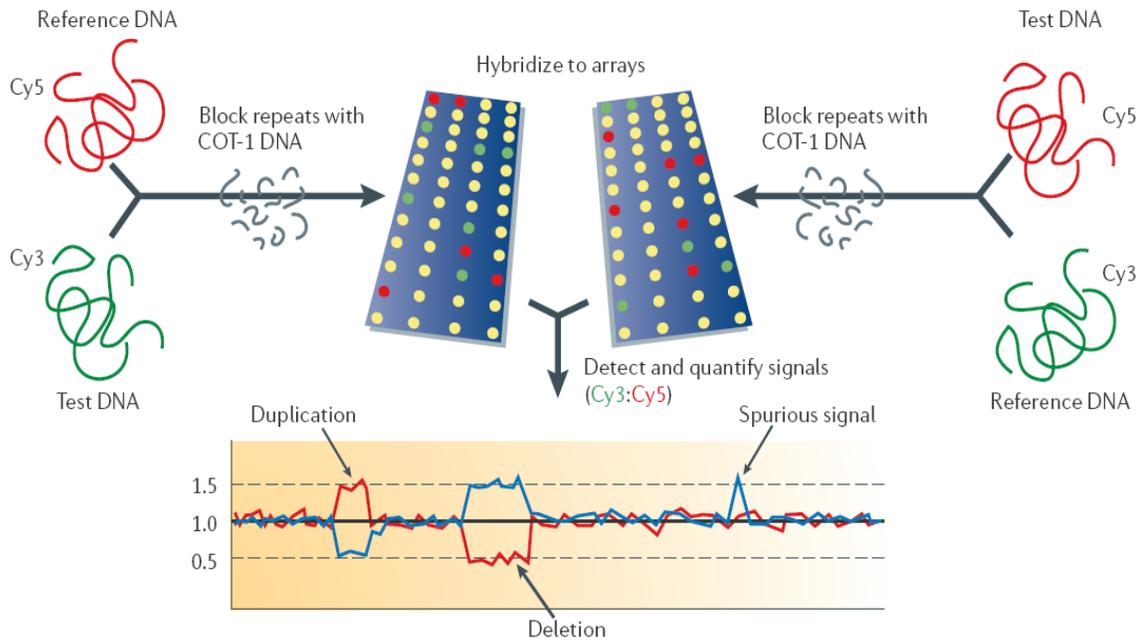


Figure 1.3 Oligonucleotide based CGH for copy number change identification (Image from reference [7])

Based on the same comparative hybridization mechanism, oligonucleotide based CGH replaces metaphase chromosomes for CGH with arrays of oligonucleotide probes accurately mapped onto the human genome and spotted robotically onto glass slides using split metal pins or glass capillaries. In oligonucleotide based CGH (see Figure 1.3), reference and test DNA samples are differentially labeled with fluorescent tags (Cy5 and Cy3, respectively), and are then hybridized to genomic arrays after repetitive-element binding is blocked using special COT-1 DNA. After hybridization, the fluorescence ratio (Cy3: Cy5) is determined, which reveals copy-number differences between the two DNA samples. Typically, it is carried out using a ‘dye-swap’ method, in which the initial labeling of the reference and test DNA samples is reversed for a

second hybridization (indicated by the left and right sides of the panel) for spurious signals in case the reciprocal ratio is not observed. For this method, germline copy number changes are suppressed in the measurements if the two samples are from the same subject.

SNP arrays are originally designed for genotyping genomes at pre-selected loci of single-base mutations [10]. Affymetrix SNP arrays use oligonucleotides as the probes as well, the probes are 25 base pairs long but each SNP is measured by a set of probes in order to increase the confidence of genotype calls. Usually a digested and labeled DNA sample is hybridized to the SNP array and the fluorescent intensities of the oligonucleotide probes are used as raw SNP signals. SNP arrays are always considered superior to other platforms in terms of resolution. For example, the latest Affymetrix SNP 6.0 array consists of about 1.8 million probe sets, and the mean and median distances between the target loci are only 500 bases and 2k bases respectively; additionally, a single experiment on a SNP array provides both copy numbers and genotype information, which can be used jointly to determine allele-specific deletions [11]. Furthermore, copy numbers can be extracted from existing SNP array data originally generated for genotyping. Due to these attractive properties, SNP arrays are becoming popular tools for copy number change analysis.

1.2.3 Copy Number Estimation from Microarray Measurement

From the measurements of DNA abundance in a sample, we can estimate the copy numbers of DNA segments. After hybridization, the fluorescence intensities of the probes on a microarray are captured as a digital image. For CGH arrays, the comparative hybridization signals effectively eliminate experiment-related biases across different arrays and can be analyzed directly to detect copy number changes. For SNP arrays, the copy number estimation

step is more complicated. Firstly there is only a single DNA sample hybridized to an array, the systematic difference between the intensity signal levels of different arrays has to be removed before analyzing the arrays as a batch. Secondly, as we talked in the previous section, to measure a particular genomic locus, a probe set consisting of multiple probes is designed to enhance the reliability, we need to estimate a single expression for the targeted SNP based on all the probes in the probe set. We focus on SNP array in the thesis and will provide a thorough discussion of pre-processing SNP array signal and further real copy number estimation in Chapter 2.

1.3 Challenges, Objectives and Contributions

Detecting copy number changes in microarray data is a challenging task due to the large numbers of observations and low SNRs in the signal profiles of high resolution SNP arrays. Detecting CNAs in tumor genomes further complicates the problem due to the complex signal patterns caused by normal tissue contamination and heterogeneous cell populations. We need effective computational methods for the high throughput copy number analysis, and our goal is to extract discriminative or mechanistic information conveyed by the CNAs, which should be helpful for cancers mechanism research.

In the thesis, we decompose the copy number analysis into following four tasks. Firstly, to remove systematic differences between different SNP arrays we employ dChip to pre-process the SNP array to get signal calls and estimate the real copy number signal. Secondly, copy number signal is normalized through SFNM to adjust the bias generated in pre-processing step. Thirdly, we will define the chromosome instability mathematically. We perform copy number variation/alternation detection and proposed two definitions of chromosomal instability (CIN) index based on Circular Binary Segmentation (CBS) and Haar wavelet transform respectively. CBS-CIN shows the instability trend of a whole chromosome based on the gain/loss calls, while

Haar-CIN describes more details of structural mutations on different scales through its sub band coefficients. Finally, in the experiment part, we tested our CIN indices on real ovarian cancer dataset and got promising results through SVM classification and survival analysis.

The major contributions of the thesis can be summarized as follows:

1. Introduced an entire practical and effective framework (Figure 1.4) for high throughput copy number analysis. starting from the raw SNPs arrays up to the extraction of discriminative information indicated by the copy number changes;
2. Pre-processed the raw SNP array and Normalized raw copy number profiles through Standard Finite Normal Mixture Models
3. Proposed two definitions of CIN to mathematically and quantitatively characterize chromosome instability, namely CBS-CIN and Haar-CIN;
4. Tested the proposed CIN indices on the real ovarian cancer dataset and obtained promising results;
5. Implemented the proposed framework in G-DOC and it can be used as a tool in cancer research.

The framework of computational analysis of copy number changes is illustrated in Figure 1.4. We will discuss each task in the following chapters.

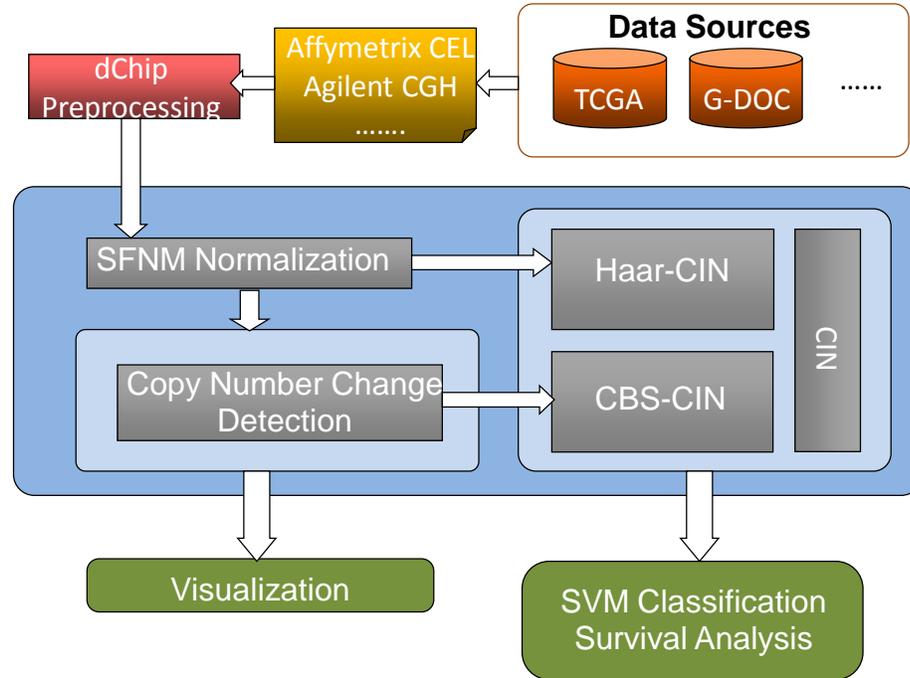


Figure 1.4 The framework of computational analysis of copy number changes

1.4 Organization of the Thesis

In Chapter 1, we have introduced the motivation, background knowledge and microarray technologies for copy number, and we finally showed our framework and contributions.

In Chapter 2, we will mainly discuss the dChip SNP signal pre-processing, including SNP array normalization, MBEI (Model-Based Expression Index) for probe set intensities estimation and finally the real copy number estimation.

In Chapter 3, we firstly introduce the general SFNM model, adjust it according to the characteristics of copy number signals and show its application in copy number profiles normalization.

In Chapter 4, we propose two definitions of chromosome instability indices to mathematically and quantitatively characterize chromosome instability: the CBS-CIN based on circular binary segmentation and Haar-CIN based on Haar wavelet transform.

In Chapter 5, we will test our analysis methods in the framework on real copy number datasets. We generated CIN indices on real copy number data of three ovarian cancer subtypes and used them as features to train a support vector machine (SVM) classifier. We also performed additional survival analysis based on the extracted CIN scores from TCGA ovarian cancer dataset and showed considerable correlation between these CIN scores and various events and severity in ovarian cancer development.

In Chapter 6, we will conclude our current analysis and discuss several possible future work.

The Appendix shows the integrated applications implemented into G-DOC.

Chapter 2

Pre-processing SNP Array Data

In this chapter we will discuss the pre-processing of SNP Array through dChip¹ software. Once the tissue samples are measured by a SNP array experiment, the foremost task is to pre-process the observed SNP array to get the real copy number signals. The first step is to normalize the SNP array, whose purpose is to make the signals of different arrays comparable. Given the image of the SNP array, what we need to do is to normalize the fluorescence intensities of individual probes with respect to a reference array. The second step is to generate the probe set intensity or expression index from the adjusted intensities of its member probes, which usually have different contributions to the overall intensity. The last step is to infer the real DNA copy number based on these SNP probe set intensities. We will discuss these steps in the following sections.

2.1 SNP Arrays Normalization

Since SNP array images usually have different overall image brightness, especially when they are generated at different times and places (Figure 2.1 gives an example of two SNP array images, (a) is brighter than (b) generally), so proper normalization is required before comparing the probe intensities of SNPs between arrays.

¹ <http://biosun1.harvard.edu/complab/dchip/>

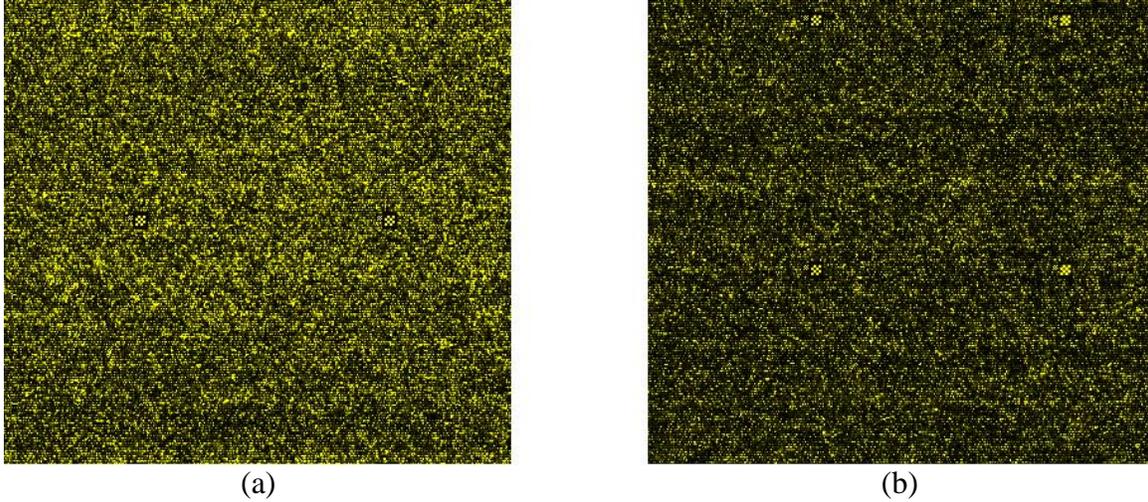
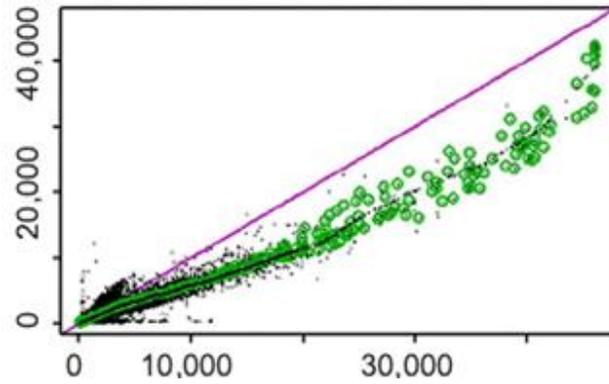


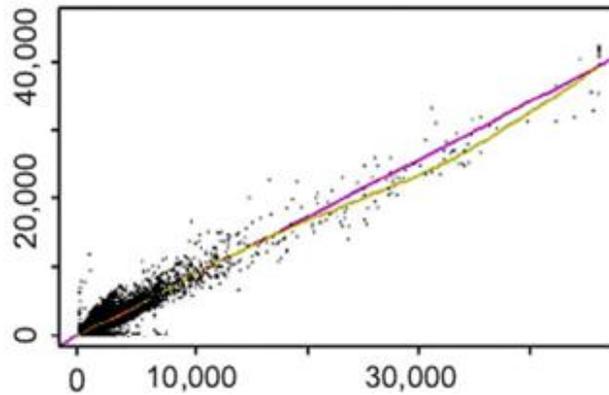
Figure 2.1 Sample SNP array images

Given a group of arrays, we use dChip to normalize all arrays to a common baseline array having the median overall brightness (as measured by the median probe intensity in an array). A normalization relation can be understood as a curve in the scatter plot of two arrays with the baseline array drawn on the y-axis and the array to be normalized on the x-axis (see Figure 2.2). We should base the normalization only on probe values that belong to “non-differentially” expressed SNPs, but generally we do not know which ones are “non-differentially” expressed (control or housekeeping ones may also be variable across arrays). Nevertheless, we expect that probes of a “non-differentially” expressed SNP in two arrays to have similar intensity ranks. dChip uses an iterative procedure to identify a set of probes (called an invariant set), which presumably consists of points from “non-differentially” expressed SNPs. Specifically, dChip starts with points of all Perfect Match (PM) probes. If a point's proportional rank difference [12] is small enough (smaller than a threshold), it is kept for the new set. These thresholds were chosen empirically to make the selected points in the invariant set thin enough to naturally determine a normalization relation. In this way, a new set of about 10K points will be obtained (For SNP 250K array), and the same procedure is applied to the new set iteratively until the

number of points in the new set does not decrease any longer. A piecewise-linear running median curve [13] is then calculated and used as the normalization curve. After normalization, the two arrays have similar overall brightness. We can see the whole process in the Figure 2.2.



(a) Not normalized



(b) Normalized

Figure 2.2 Normalization of probe intensities between SNP arrays

In Figure 2.2(a), the probe intensities of two arrays are plotted against each other. The baseline array shown on the y-axis is not as bright as array shown on the x-axis, indicating the need for normalization; and the probes from the invariant set are plotted as green circles. Based on these probes from the invariant set a piecewise linear normalization relationship is determined. Figure 2.2(b) showed the results after normalization, the scatter plot centers around the diagonal line and the array is adjusted to have the similar overall brightness as the baseline array.

2.2 Model-Based Expression Index (MBEI) Estimation in SNP Arrays

2.2.1 PM/MM Design in SNP Array

As we discussed in Chapter 1, each SNP is measured by a set of probes in order to increase the confidence of reliability. So to estimate the real copy number we need first estimate a single intensity for the target SNP (here we use expression index to distinguish from member probe intensity) based on all the probe intensities in the probe set. For SNP arrays, 25 probe pairs are used to locate the target SNP, each probe pair has a Perfect Match (PM) and Mismatch (MM) signal (see Figure 2.3 below), and PM-MM differences for all probe pairs in a probe set are used to calculate an expression index for the target SNP [11].

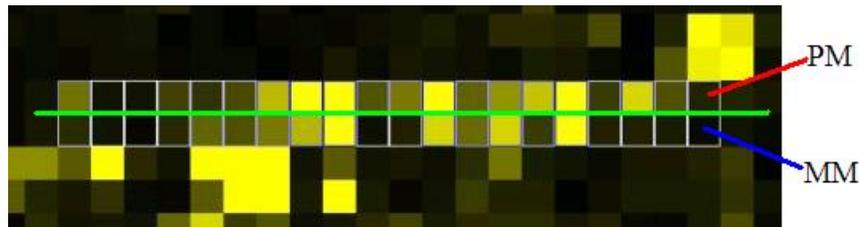


Figure 2.3 PM/MM illustration in SNP array

2.2.2 Statistical Model

Suppose we have N samples to profile in an experiment. Then, for any given SNP, our task is to estimate its expression index. The expression index estimates are constructed from the $2*N*25$ intensity values for the PM and MM probes corresponding to each SNP if we assume each probe set has 25 probe pairs. The estimation procedure is based on a model of how the probe intensity values respond to changes of the expression levels [14] of the whole probe set for the target SNP. We denote θ_i by an expression index for the SNP in the i_{th} sample. We assume

that the intensity value of a probe in its probe set will increase linearly as θ_i increases, but that the rate of increase will be different for different probes. It is also assumed that within the same probe pair, the PM intensity will increase at a higher rate than the MM intensity. We then have the following simple model:

$$\begin{aligned} MM_{ij} &= v_j + \theta_i \alpha_j + \varepsilon \\ PM_{ij} &= v_j + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon \end{aligned} \quad (2.1)$$

Here PM_{ij} and MM_{ij} denote the PM and MM intensity values for the i_{th} array and the j_{th} probe pair for this SNP, v_j is the baseline response of the j_{th} probe pair due to nonspecific hybridization, α_j is the rate of increase of the MM response of the j_{th} probe pair, ϕ_j is the additional rate of increase in the corresponding PM response, and ε is a generic symbol for a random error. The rates of increase are assumed to be nonnegative.

The model for individual probe responses implies an even simpler model for the PM–MM differences:

$$y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij} \quad (2.2)$$

The above (2.2) model is introduced since currently there is a computational advantage in reducing to differences, as the fitting of the full data is a more difficult numerical task. Thus, in the rest of this chapter our discussion will be focused mainly on the analysis of PM–MM differences directly.

The foregoing model for the differences is identifiable only if we constrain it in some way. Here we simply make the sum squares of ϕ s to be J :

$$\begin{aligned}
y_{ij} &= PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij} \\
\sum_j \phi_j^2 &= J, \quad \varepsilon_{ij} \sim N(0, \sigma^2)
\end{aligned} \tag{2.3}$$

Least square estimates for the parameters are carried out by iteratively fitting the set of θ_s and ϕ_s , regarding the other set as known.

2.2.3 Conditional Mean and Standard Error

Suppose for a particular SNP, the ϕ_s have been learned from a large number of arrays, we can then treat them as known constants and analyze the mean and variance of the expression index estimate. For a single array, the model becomes:

$$y_j = PM_j - MM_j = \theta \phi_j + \varepsilon_j \tag{2.4}$$

Given the ϕ_s , the linear least square estimate for θ is

$$\begin{aligned}
\tilde{\theta} &= \frac{\sum_j y_j \phi_j}{\sum_j \phi_j^2} = \frac{\sum_j y_j \phi_j}{J} \\
E(\tilde{\theta}) &= \theta, \quad \text{Var}(\tilde{\theta}) = \sigma^2 / J
\end{aligned} \tag{2.5}$$

Now, an approximate standard error for the least square estimate can be computed:

$$\begin{aligned}
\text{Std}(\tilde{\theta}) &= \sqrt{\hat{\sigma}^2 / J} \\
\hat{\sigma}^2 &= \left(\sum_j (\text{fit} - \text{obs})^2 \right) / (J - 1)
\end{aligned} \tag{2.6}$$

Further when we regard the estimated θ_s as fixed, we can calculate the standard errors of ϕ_s in the same way. These standard errors will play an important role in outlier detection and probe selection. For example in the standard analysis, the mean and standard deviation of the PM-MM differences in a probe set are computed after excluding the maximum and minimum values. If a

difference deviates by more than $3 \times \text{Std}$ from the mean, a probe pair is marked as an outlier in this array and discarded in future analysis.

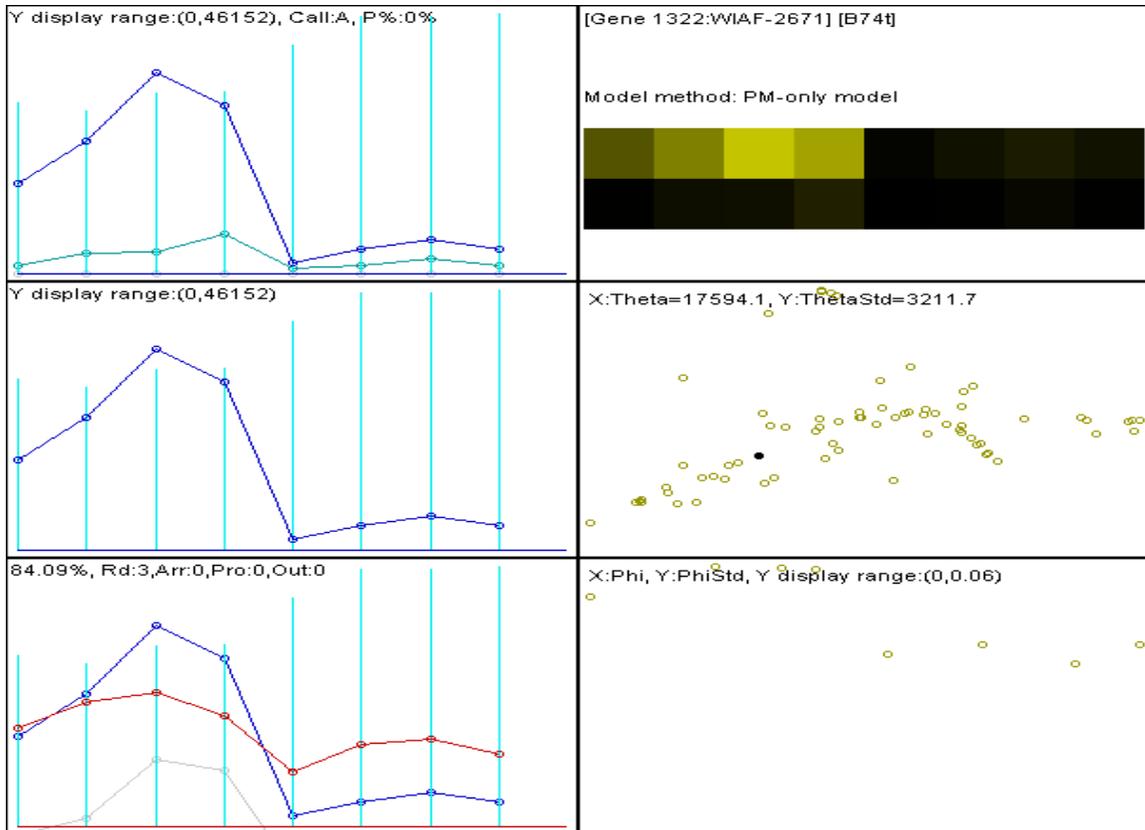


Figure 2.4 PM/MM data view

Figure 2.4 show the PM/MM data view in dChip. There are 6 grids in this "PM/MM Data View". Let us use (x, y) to denote different grids, with (1, 1) the upper-left grid and (2, 3) the lower-right grid. Grid (1, 1) displays the PM and MM data in blue and green curves, with the x-axis ordering probe sets from 1 to n, and y-axis for probe intensities with range (0, 46152). Grid (1, 2) is the PM/MM difference curve, the horizontal blue line is $y=0$. Grid (1, 3) overlays red fitted curve to the blue PM/MM difference curve, and also shows the residual curve in light gray. From this pane we can also read that the explained energy is 84.09% after fitting the model to the PM/MM difference data of this probe set, and it takes 3 rounds of iteration for model fitting and

outlier identification. Here we found 0 array outlier, 0 probe outlier and 0 single outliers. Grid (2, 1) is the intensity image of the current probe set in current array. Grid (2, 2) displays the scatter plot of MBEI θ , versus the standard error of θ . In Grid (2, 2), the black dot represents the current array.

2.3 Real Copy Number Estimation Based on MBEI

Once we get the MBEI for each SNP, the next step is to infer the copy number [11] based on these MBEIs. There are two ways to do the job. One is quite simple and intuitive. In general, we assume a diploid genome for normal samples, then firstly for each SNP, the signal values of all of the normal sample arrays are averaged to obtain the mean signal of 2 copies, and the copy number is defined as (observed signal/mean signal of two copies)*2. The other way to infer the copy number is a little bit more complex but more accurate based on Hidden Markov Model (HMM) [15][16]. In our following discussion, we use HMM to infer the copy number signal in dChip in default. In Figure 2.5 below, we give a scatter plot of a sample copy number extracted through dChip. From the figure, we can get a global view of a sample copy number profile: the chromosome 23 (X) from an ovarian cancer sample, which is the shortest chromosome in human genome, even has 80K probes. The sample profile shows us the typical characteristics for copy number data, low signal to noise ration and huge size, which are very challenging. In following chapters, we will introduce our methods to effectively overcome these challenging characteristics and get promising results.

TCGA-24-1556 Chr 23

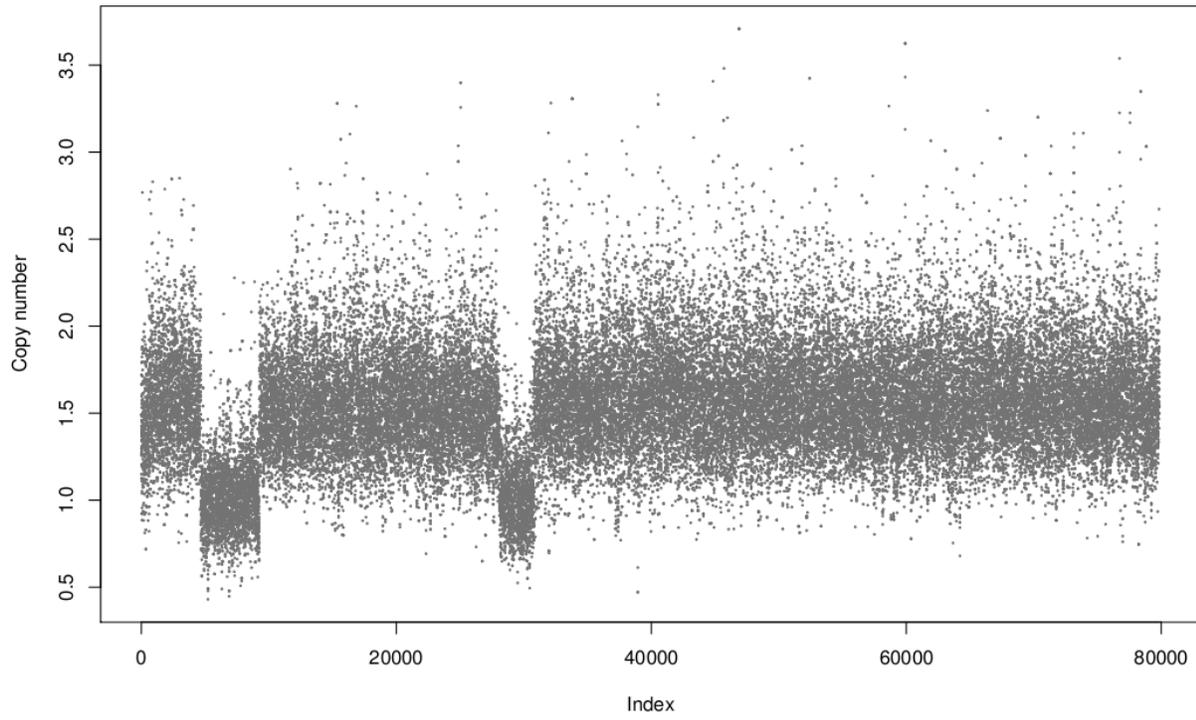


Figure 2.5 Sample copy number profile visualization

Chapter 3

Copy Number Normalization using Standard Finite Normal Mixture Model

As discussed in Chapter 2, in SNP array pre-processing step the signals of the target array are transformed to have the same median as the signals of the reference or baseline array, so the estimation of normal copy number profiles based on these transformed signals may have systematic bias and deviate from 2, which is the true value since the human genome is diploid. In this chapter we address this problem using standard finite normal mixture (SFNM) model [17]. We will first introduce the general SFNM model and the maximum likelihood solutions EM algorithm for the model, and then we adjust the model according to the characteristics of copy number signals, further we discuss how to normalize the copy number signals through the model.

3.1 Introduction to Standard Finite Normal Mixture Model

3.1.1 Mixture of Gaussians

In statistics, a mixture model is a probabilistic model for density estimation using a mixture distribution [18], and the observations in a mixture model are assumed to be distributed according to a mixture density.

Now let's discuss the general standard finite normal mixture model. Without loss of generality, here we discuss the model based on vector-valued random variables $\mathbf{x} \in R^D$ of D dimensions, while our copy number signal is the special case with $D = 1$, since copy number for each SNP loci is a scalar value.

Suppose in addition to the observed data set \mathbf{X} , corresponding to $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ consisting of N observations of $\mathbf{x}_n \in R^D$, we were also given the values of the corresponding latent label variables \mathbf{Z} as $\{z_1, \dots, z_N\}$. Further if we have K mixtures in the standard finite normal mixture model, then z_n is a K -dimensional binary random vector having a 1-of- K representation in which a particular element z_{n_k} is equal to 1 and all other elements are equal to 0. The values of z_{n_k} therefore satisfy $z_{n_k} \in \{0, 1\}$ and $\sum_k z_{n_k} = 1$, and we can see that there are K possible states for the vector z_n according to which element is nonzero. Now we can define the joint distribution $p(\mathbf{x}, \mathbf{z})$ in terms of a marginal distribution $p(\mathbf{z})$ and a conditional distribution $p(\mathbf{x} | \mathbf{z})$. The marginal distribution over \mathbf{z} is specified in terms of the mixing coefficients π_k , such that:

$$p(z_k = 1) = \pi_k \quad (3.1)$$

$$\text{where } 0 \leq \pi_k \leq 1 \text{ and } \sum_{k=1}^K \pi_k = 1 \quad (3.2)$$

The latent variable \mathbf{z} uses a 1-of- K representation, so we can also write this distribution in the form of $p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$ and the conditional distribution of \mathbf{x} given a particular value for \mathbf{z}

is a Gaussian:

$$p(\mathbf{x} | z_k = 1) = N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (3.3)$$

which can also be written in the form:

$$p(\mathbf{x} | \mathbf{z}) = \prod_{k=1}^K N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \quad (3.4)$$

The joint distribution is given by $p(\mathbf{x} | \mathbf{z})p(\mathbf{z})$,

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} | \mathbf{z})p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k} N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \quad (3.5)$$

and the marginal distribution of \mathbf{x} is then obtained by summing the joint distribution over all possible states of \mathbf{z} to give:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x} | \mathbf{z})p(\mathbf{z}) = \sum_{k=1}^K \pi_k N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (3.6)$$

where we have made use of (3.4) and (3.5).

3.1.2 Likelihood Function

For every observed data point \mathbf{x}_n there is a corresponding latent variable z_n , if we have several observations $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, then:

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \quad (3.7)$$

Taking the logarithm of (3.7), we get:

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \} \quad (3.8)$$

The expected value of the complete-data log likelihood function is therefore given by:

$$E[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K E[z_{nk}] \{ \ln \pi_k + \ln N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \} \quad (3.9)$$

While we can derive $E[z_{nk}]$ based on Bayes' theorem as:

$$E[z_{nk}] = \frac{\pi_k N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j N(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (3.10)$$

Here we define $\gamma(z_{n_k}) = \frac{\pi_k N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j N(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = E[z_{n_k}]$, then we get:

$$E[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{n_k}) \{ \ln \pi_k + \ln N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \} \quad (3.11)$$

We shall view π_k as the prior probability of $z_k = 1$, and the quantity $\gamma(z_{n_k})$ as the corresponding posterior probability once we have observed \mathbf{x}_n . $\gamma(z_{n_k})$ can also be viewed as the responsibility that component K takes for ‘explaining’ the observation \mathbf{x}_n .

Now our task is to maximize this likelihood, which refers to the next section, the EM (Expectation Maximization) algorithm.

3.2 Expectation Maximization Algorithm

The Expectation Maximization (EM) algorithm [19][20] is a parameter estimation method which falls into the general framework of maximum likelihood estimation and is applied in cases where part of the data can be considered to be incomplete or hidden. The goal of the EM algorithm is to find maximum likelihood solutions for models including latent variables [21][22]. It is essentially an iterative optimization algorithm which at least under certain conditions will converge to parameter values at a local maximum of the likelihood function.

3.2.1 General EM Algorithm

We denote the set of all observed data by \mathbf{X} , and similarly we denote the set of all latent variables by \mathbf{Z} . The set of all model parameters is denoted by $\boldsymbol{\theta}$. Now suppose that, for each observation in \mathbf{X} , we were told the corresponding value of the latent label variable \mathbf{Z} . We shall call $\{\mathbf{X}, \mathbf{Z}\}$ the complete data set and we shall refer to the actual observed data \mathbf{X} as incomplete.

The likelihood function for the complete data set simply takes the form $\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$. In practice, however, we are not given the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, but only the incomplete data \mathbf{X} . Our knowledge of the values of the latent variables in \mathbf{Z} is given only by the posterior distribution $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta})$. Because we cannot use the complete-data log likelihood [23], we consider instead its expected value (3.9) under the posterior distribution of the latent variables, which corresponds to the E-step of the EM algorithm. In the subsequent M-step, we maximize this expectation. If the current estimate for the parameters is denoted as $\boldsymbol{\theta}_{old}$, then a pair of successive E and M steps gives rise to a revised estimate $\boldsymbol{\theta}_{new}$.

The algorithm is initialized by choosing some starting value for the parameters $\boldsymbol{\theta}_0$. In the E step, we use the current parameter values $\boldsymbol{\theta}_{old}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}_{old})$. We then use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value $\boldsymbol{\theta}$. This expectation, denoted as $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$, is given by

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}_{old}) \ln p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}) \quad (3.12)$$

In the M step, we determine the revised parameter estimate $\boldsymbol{\theta}_{new}$ by maximizing this function: $\boldsymbol{\theta}_{new} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$.

The EM algorithm is summarized below:

Given a joint distribution $p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$ over observed variables \mathbf{X} and latent variables \mathbf{Z} , governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters θ_{old} .
2. E step evaluates $p(\mathbf{Z} | \mathbf{X}, \theta_{old})$.
3. M step evaluates θ_{new} given by $\theta_{new} = \arg \max_{\theta} Q(\theta, \theta_{old})$
4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let $\theta_{new} \leftarrow \theta_{old}$ and return to step 2.

3.2.2 EM Solution for SFNM

We can now precede the SFNM-EM solution as follows. First we choose some initial values for the parameters μ_{old} , Σ_{old} and π_{old} , and use these to evaluate $\gamma(z_{n_k})$ (the E step). We then keep $\gamma(z_{n_k})$ fixed and maximize $E[\ln p(\mathbf{X}, \mathbf{Z} | \mu, \Sigma, \pi)]$ with respect to μ_k , Σ_k and π_k (the M step):

1. Maximize $E[\ln p(\mathbf{X}, \mathbf{Z} | \mu, \Sigma, \pi)]$ with respect to μ_k .

Setting the derivatives of $E[\ln p(\mathbf{X}, \mathbf{Z} | \mu, \Sigma, \pi)]$ with respect to the means μ_k to zero, we obtain:

$$\begin{aligned}
& \frac{\partial E[\ln p(\mathbf{X}, \mathbf{Z} | \mu, \Sigma, \pi)]}{\partial \mu_k} \\
&= \sum_{n=1}^N \gamma(z_{n_k}) \left(\frac{\partial}{\partial \mu_k} \frac{1}{2} (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) \right) \\
&= \sum_{n=1}^N \gamma(z_{n_k}) \left(\frac{1}{2} (\Sigma_k^{-1} + \Sigma_k^{-T}) (\mathbf{x}_n - \mu_k) (-1) \right) \\
&= \sum_{n=1}^N \gamma(z_{n_k}) (\Sigma_k^{-1} (\mathbf{x}_n - \mu_k)) = 0 \\
&\Rightarrow \mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{n_k}) \mathbf{x}_n \quad \text{where} \quad N_k = \sum_{n=1}^N \gamma(z_{n_k}) \tag{3.13}
\end{aligned}$$

2. Maximize $E[\ln p(\mathbf{X}, \mathbf{Z} | \mu, \Sigma, \pi)]$ with respect to Σ_k

Setting the derivatives of $E[\ln p(\mathbf{X}, \mathbf{Z} | \mu, \Sigma, \pi)]$ with respect to the means Σ_k to zero, we obtain:

$$\begin{aligned}
& \frac{\partial E[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})]}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_{n=1}^N \gamma(\mathbf{z}_{n_k}) \left(-\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| - \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) \\
&= \sum_{n=1}^N \gamma(\mathbf{z}_{n_k}) \left(-\frac{1}{2} \boldsymbol{\Sigma}_k^{-T} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-T} (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-T} \right) = 0 \\
&\Rightarrow \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(\mathbf{z}_{n_k}) (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^T
\end{aligned} \tag{3.14}$$

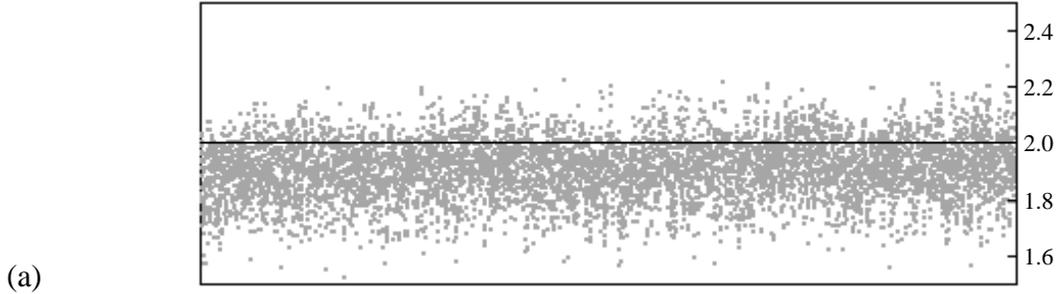
3. Maximize $E[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})]$ with respect to π_k . It can be done using Lagrange multiplier and maximizing the following quantity:

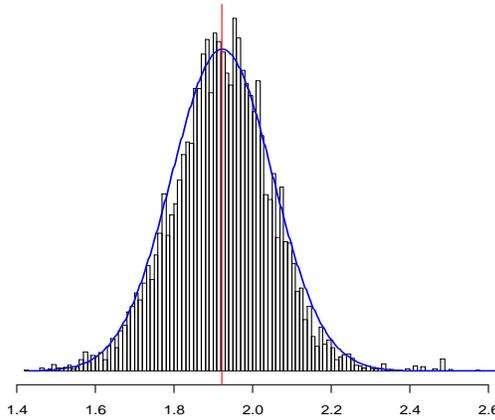
$$\begin{aligned}
& E[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) \\
& \frac{\partial E \left[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) \right]}{\partial \pi_k} = \sum_{n=1}^N \gamma(\mathbf{z}_{n_k}) \frac{1}{\pi_k} - \lambda = 0 \\
& \Rightarrow \pi_k = \frac{1}{\lambda} \sum_{n=1}^N \gamma(\mathbf{z}_{n_k}) \Rightarrow \lambda = N \Rightarrow \pi_k = \frac{1}{N} \sum_{n=1}^N \gamma(\mathbf{z}_{n_k})
\end{aligned} \tag{3.15}$$

3.3 Copy Number Signal Normalization through SFNM

3.3.1 Transform Function for Copy Number Signal Normalization

Since human genome is diploid, majority of the DNA segments should have two copies of themselves, each copy on one homologous chromosome. Therefore the observed copy numbers of a normal sample should distribute around 2.





(b)

Figure 3.1 A sample normal copy number profile

In Figure 3.1(a) we plot a signal profile of a normal sample processed by the dChip software. We can see that the profile systematically deviates from the normal copy number 2. Figure 3.1(b) shows that the histogram of the signals can be nicely fitted by a normal distribution $N(\hat{\mu}, \hat{\sigma}^2)$, where $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and sample variance of the observations. The blue curve is the probability density function of the normal distribution. Based on the two observations, we have a simple transformation:

$$y'_i = \frac{y_i - \hat{\mu}}{\hat{\sigma}} \times \sigma + 2 \quad i = 1, 2, \dots, N \quad (3.16)$$

to adjust the copy number signals. Here $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and sample standard deviation of the observations $\{y_i\}_{i=1}^N$; the transformed signals $\{y'_i\}_{i=1}^N$ has mean 2 and standard deviation σ , which can be set to the standard deviation of the signals of normal reference arrays.

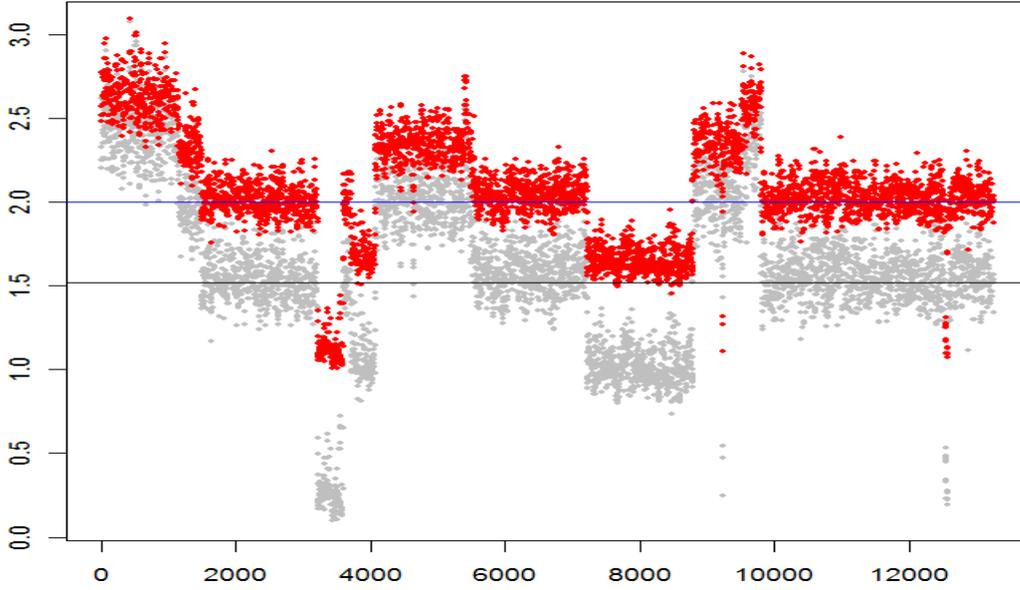


Figure 3.2 The original signal profile (gray) and the normalized profile (red)

For samples with copy number changes, we can apply the same transform as long as the cluster of signals of normal copy number can be identified. In Figure 3.2, we plot the copy number profile of the first chromosome of a high-grade ovarian cancer sample. We can see none of the major clusters of the signals has normal copy number 2. Based on the assumption that majority of the DNA segments should have two copies, we fit the signals by SFNM and assign the component with the largest mixing proportion as the normal cluster.

3.3.2 Fit Copy Number Signal through SFNM

We assume there is finite number (K) of copy number states, and then the signal can be fit by SFNMs as:

$$p(y_i | \theta) = \sum_{k=1}^K \pi_k N(y_i | \mu_k, \sigma_k^2), \quad i = 1, 2, \dots, N \quad (3.17)$$

where π_k , μ_k and σ_k^2 are the mixing proportion, mean and variance of the k_{th} component in the mixture, $\theta = \{\pi_k, \mu_k, \sigma_k^2\}_{k=1}^K$ are the parameters of all components. According to EM algorithm, the mixture parameters in SFNMs are estimated so as to maximize the log-likelihood:

$$Q(\theta', \theta) = \sum_{i=1}^N \sum_{k=1}^K \tau_k(y_i, \theta) \ln(\pi'_k N(y_i | \mu'_k, \sigma_k'^2)) \quad (3.18)$$

where $\tau_k(y_i, \theta)$ is the membership value of y_i to the k_{th} component, $\theta' = \{\pi'_k, \mu'_k, \sigma_k'^2\}_{k=1}^K$ are the parameters to be estimated in the upcoming M step. The E-step and M-step are described as follows:

E-step:

$$\tau_k(y_i, \theta) = \pi_k N(y_i | \mu_k, \sigma_k^2) / \sum_{k=1}^K \pi_k N(y_i | \mu_k, \sigma_k^2) \quad (3.19)$$

M-step:

$$\begin{aligned} \mu'_k &= \sum_{i=1}^N \tau_k(y_i, \theta) y_i / \sum_{i=1}^N \tau_k(y_i, \theta) \\ \sigma_k'^2 &= \sum_{i=1}^N \tau_k(y_i, \theta) (y_i - \mu'_k)^2 / \sum_{i=1}^N \tau_k(y_i, \theta) \\ \pi'_k &= \sum_{i=1}^N \tau_k(y_i, \theta) / N \end{aligned} \quad (3.20)$$

The EM algorithm terminates when the absolute difference between the conditional expectations of the log-likelihood of two consecutive steps is smaller than a pre-defined threshold. And we determine the number of the mixed component K based on the Minimum Description Length (MDL) principle [24][25]:

$$MDL(K) = -Q + \frac{3K-1}{2} \ln N \quad (3.21)$$

where Q is the value of (3.18) in the final step of the EM algorithm.

Denote the optimal number of components by K^* and the corresponding estimates of parameters by $\theta^* = \{\pi_k^*, \mu_k^*, \sigma_k^{*2}\}_{k=1}^K$, we can set the mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ in (3.16) as $\hat{\mu} = \mu_k^*$ and $\hat{\sigma} = \sigma_k^*$ with $k = \arg \max_{i=1, \dots, K} \pi_k^*$.

The density function of the fitted SFNMs is plotted in Figure 3.3, based on MDL we choose $K = 8$. Using the sample mean $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$ of the maximum cluster, we can normalize the entire profile by (3.16). The normalized profile is plotted in Figure 3.2 in red. We can clearly see that after normalization the majority of the profile is stretched to 2.

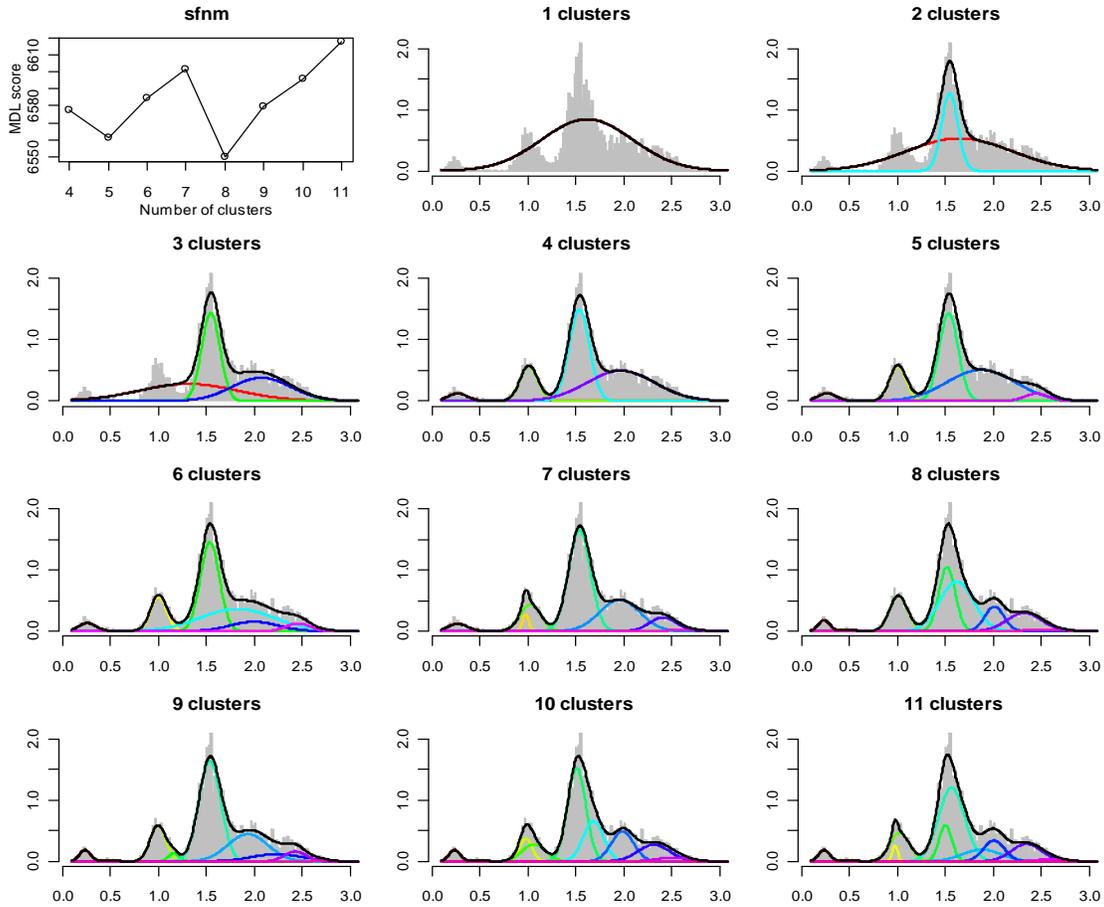


Figure 3.3 The fitted SFNM model and MDL model selection.

Chapter 4

Chromosome Instability Index

4.1 Introduction

The real copy number profile is generated through pre-processing in Chapter 2 and normalized in Chapter 3. In this Chapter we will discuss the analysis of estimated copy number profiles and extraction of informative features. As we discussed in the introduction, genomic instability is known to be a fundamental trait in the development of tumor; and most human tumors exhibit this instability in structural and numerical alterations: deletions, amplifications, inversions or even losses and gains of whole chromosomes, all of which will result in copy number changes. It is well recognized that the deletions of cancer suppressor genes and amplifications of oncogenes are considered as hallmarks of the formation and progression of cancers. Specifically, the chromosome instability indicated by these copy number alternations is associated with various events in the development or the severity of cancers. We can extract some quantitative measure such as the scales, frequencies, and amplitudes of copy number changes across a chromosome or the whole genome, namely the chromosome instability (CIN) index, and we can assess the impacts of copy number changes on various biological events by studying the association of CIN indices with those events.

In this chapter, we will introduce two chromosome instability index definitions. The instability across the chromosomes exhibits variations in the forms of deletions and amplifications, and to mathematically and quantitatively describe these variations we should first locate their positions and measure their ranges through copy number detection methods, here we use Circular Binary Segmentation (CBS) [28]. Then an intuitive yet effective CIN index, CBS-CIN, could be defined by summarizing the frequency and amplitudes of copy number variations. In signal processing's perspective, copy number signals are piecewise constant with heavy noises, which can be handled well by Haar wavelet transform [38] due to its natural characteristics. We decomposed the copy number signals based on Haar wavelet transform and defined Haar-CIN on sub band wavelet

coefficients. Essentially speaking, both the CINs we proposed can be viewed as the high level features extracted from the huge size copy number profiles, and our finally goal is to explore the relationship between these high level features and the biological and clinical characters.

4.2 CIN Definition Based on CBS Copy Number Detection

An intuitive CIN can be defined based on the amplitudes including gains and loss (amplifications and deletions as compared to the normal copy numbers). However, given the normalized signals and their corresponding locations on a chromosome, we need to first determine which parts of the genome have abnormal copy numbers, which seems easy but difficult since there is high noise in the copy number profile (we can see the Figure 2.5 and 3.2). Various methods have been proposed for detecting copy number variation/alternation. For CNV/CNA detection in a single profile, representative methods include Gain Loss Analysis of DNA (GLAD) [26], Circular Binary Segmentation [27], and Hidden Markov Models [29]. Lai [30] conducted a comparison of methods for analyzing Copy Number data that included CBS and 10 other approaches. They concluded that CBS has the best operational characteristics among the ten peer methods in terms of its sensitivity and appear to perform consistently well. Here we employed CBS to perform the detection of copy number variation/alternation.

4.2.1 Circular Binary Segmentation

The data from SNP biological experiments are the reference and testing sample intensities for each marker, which come from the normal and tumor tissues respectively of the sample patient. In Chapters 2 and 3, we pre-processed the SNP array and normalized the copy number signal of each marker around 2. Since we assume that the reference sample does not have any copy number aberrations, markers with normalized testing sample intensities significantly greater than the reference intensities are the indicators of copy number gains in the testing sample at those positions. Similarly, significantly lower intensities in the testing sample are the signs of copy number losses. The statistical methods for analyzing copy number data are thus aimed at identifying locations of gains or losses of copy number profile.

In CBS CNV/CNA detection process, given the normalized signals of N SNPs and their corresponding locations on a chromosome, our objective is to determine which parts of the genome have abnormal copy numbers. Firstly, the CBS assumes that gains or losses of copy number are discrete. These aberrations occur in contiguous regions of the chromosome that often cover multiple markers up to whole chromosome arms or chromosomes. In addition, the copy number data can be noisy, so that some markers will not reflect the true copy number in the testing sample. CBS tries to split the chromosomes into pieces of common regions, in each of these regions, the noises are neutralized and all probes in this region will have equal copy number value. Formulating the analysis of copy number variations as a problem of detecting change-points [32][33], and recursively detecting changes from coarse to fine resolutions, CBS provides a natural way to segment a chromosome into contiguous regions and bypasses parametric modeling of the data by using a permutation reference distribution [28].

Let $\{y_1, y_2, \dots, y_N\}$ be the normalized copy number signals, which are indexed by the locations of the N markers and let $Y_i = y_1 + y_2 + \dots + y_i, 1 \leq i \leq N$ be the partial sums, then the statistic for testing the null hypothesis that there is no change against the alternative that there is exactly one change at an unknown location i is the maximal t-statistic given by $Z = \text{MAX}_{1 \leq i < N} |Z_i|$, where Z_i is the two-sample t-statistic to compare the mean of the observations with index from 1 to i , to the mean of the rest of the observations. That is:

$$Z_i = \frac{Y_i/i - (Y_N - Y_i)/(N - i)}{s_z \sqrt{1/i + 1/(N - i)}} \quad (4.1)$$

$$s_z^2 = \frac{(i - 1)s_i^2 + (N - i - 1)s_{N-i}^2}{N - 2} \quad (4.2)$$

According to the above definition, binary segmentation declares a change to be statistically significant if the p-value of Z is smaller than a threshold level α and estimates the locations of the change-points as the i . The test is repeatedly applied on each of the parts and their sub-parts until no additional aberrant regions can be found.

There is a problem with the binary segmentation in that it looks for only one change point at a time. If we consider the segment to be spliced at the two ends to form a circle, the test statistic for testing the hypothesis that the arc from $i + 1$ to j and its complement have different means is given by:

$$Z_{ij} = \frac{(Y_{i+1} + \dots + Y_j)/(j-i) - (Y_1 + \dots + Y_i + Y_{j+1} + \dots + Y_N)/(N-j+i)}{s_z \sqrt{1/(j-i) + 1/(N-j+i)}} \quad (4.3)$$

$$s_z = \sqrt{\frac{(j-i-1)S_{ij}^2 + (N-j+i-1)S_{ij}^2}{(N-2)}} \quad (4.4)$$

This modification of the binary segmentation procedure is called circular binary segmentation and the test statistic is changed to $Z = \text{MAX}_{1 \leq i < j \leq N} |Z_{ij}|$.

Figure 4.1 below is the CBS detection result of a high grade ovarian cancer sample on chromosome 8. The red green line is the normal copy number 2, the blue line indicating losses and red line indicating gains.

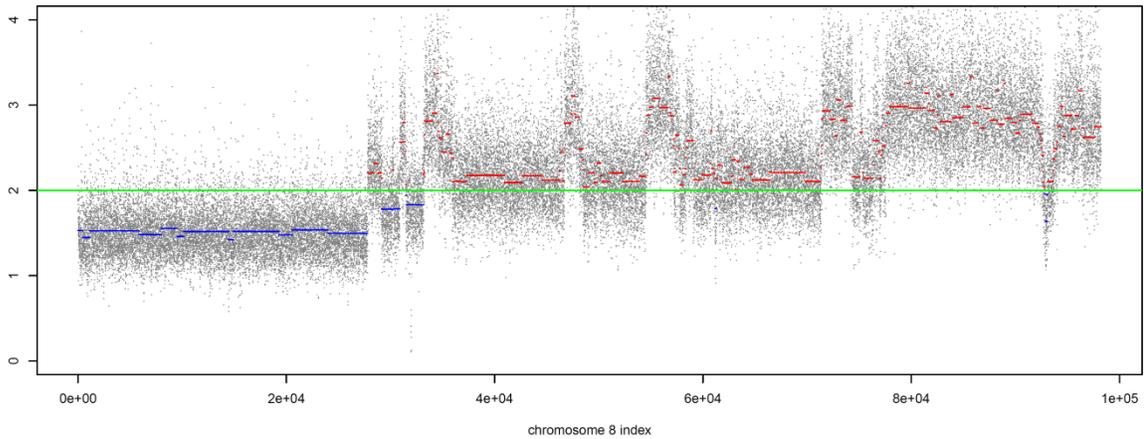


Figure 4.1 Copy number detection result of CBS

4.2.2 CIN Definition From Gain/Loss on CBS Segments

With the segmentation results from CBS, we propose an intuitive yet quantitative definition of chromosome instability CIN index based on the amplitudes of copy number alternations and explore the

feasibility of using such quantitative instability measures for disease association study and cancer diagnosis (see chapter 5 for more focused discussions).

The chromosome instability index CIN is calculated by the following steps. Given the signal profile of chromosome i and the segments generated by a CBS detection method:

1. Make gain/loss calls on the segments. A segment with mean signal intensity greater than t_{gain} is a gain while smaller than t_{loss} is a loss, the biologically experiential value of t_{gain} and t_{loss} are 2.5 and 1.5 respectively;
2. For each gain segment, its amplitude is the mean signal intensity;
3. Get the maximum gain amplitude A across all samples;
4. For each loss segment, convert its amplitude a to the new value a' based on the relationship given by $(t_{loss} - a)/a = (a' - t_{gain})/(A - t_{gain})$; the copy number of a loss segment is stretched from the range $t_{loss} \sim 0$ to the range $t_{gain} \sim A$.
5. Compute the chromosome-specific instability index $CIN_i = (\sum_k a_k + \sum_j a'_j)/N$ for each sample, where N is the number of probes (SNPs) on the chromosome.

The genome-wide CIN for each sample is defined as $\sum_{i=1}^{23} CIN_i$. In section 5.1 we show the chromosome CIN and genome-wide CIN in Figure 5.1, which will be discussed in details in next chapter.

4.3 CIN Definition Based on Haar Wavelet Transform

If the copy number profiles are viewed as time series along the chromosome coordinates, the goal of chromosome instability analysis can be interpreted as extracting distinctive information through the curve, for example, the sharp peaks and drops (correspond to CANs or CNVs) of the signal in the high noisy background. The Fourier transform [34] is a useful and important tool in signal processing domain, it can translate a time-domain signal into a frequency domain signal; however it will lose the information regarding position of signal

changes, further it cannot accurately model the sharp peaks and drops which is quite common in copy number profiles. In contrast, wavelet transform [35] can represent the signal simultaneously in both the frequency and time (position) domain and be well suited for approximating these piecewise constant signals with sharp discontinuities. Here we propose to use Haar wavelet transform as a tool to characterize the chromosome instability.

Inspired by HaarSeg [37], we use Haar coefficients to compose CIN measures. We first decompose a signal profile into a family of multi-resolution sub-bands using Haar wavelets. For each sub band, we assign p-values to the Haar coefficients based on a null-distribution estimated from normal reference samples. We select significant coefficients by controlling the False Discovery Rate (FDR) [36] and use the sum of their absolute values as the CIN signature of the sub-band. Finally, for each chromosome, we have an L -vector of CIN scores, each corresponding to a single sub band.

4.3.1 Haar Wavelet Transform of Copy Number Profile

The discrete wavelet transform (DWT) [35] can decompose a given signal into an approximation sub band and a set of detail sub bands at different resolution scales. The approximation sub band is a coarse or smooth version of the original signal, containing the scale coefficients. The detail sub bands describe the higher frequencies of the signal, and are composed of the wavelet coefficients. Due to all these properties the DWT is well suited for the task of our copy number signal with sharp discontinuities and high background noise.

The Haar wavelet is a natural choice [38] for analyze these piecewise constant copy number signals, its mother wavelet function $\psi(x)$ and scaling function $\phi(x)$ can be described as (4.5) below:

$$\psi(x) = \begin{cases} 1 & 0 \leq x < 1/2 \\ -1 & 1/2 \leq x < 1 \\ 0 & \textit{otherwise} \end{cases} \quad \phi(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & \textit{otherwise} \end{cases} \quad (4.5)$$

The complete set of Haar basis functions [39] $\psi_i^j(x) = \psi(2^j x - i)$ consists of all possible copies of the mother wavelet function, shifted and expanded by powers of 2 to cover the domain of interest (the whole

chromosome). The region where the Haar wavelet is nonzero defines both a characteristic position and length scale. The discrete Haar wavelet transform decomposes the input signal into L levels, $0 \leq L \leq \log_2 N$, N is the number of observations. At each level, the signal is decomposed into two sub-signals of half its length, namely approximation coefficients from scaling function and detail coefficients from wavelet function. According to (4.5), we can easily get Haar wavelet transform's discrete scaling and wavelet function as $\frac{1}{\sqrt{2}}(1,1)$ and $\frac{1}{\sqrt{2}}(1,-1)$ respectively, and the signal is recursively averaged and differenced on neighborhood approximation coefficients. For a copy number profile $\{y_0, y_1, \dots, y_{N-1}\}$, the detail coefficients of sub band L can be derived as:

$$\omega_L^n = \frac{1}{\sqrt{2^{L+1}}} \left(\sum_{i=n}^{n+(2^L-1)} y_i - \sum_{i=n+2^L}^{n+2^{L+1}-1} y_i \right) \quad (4.6)$$

$$i = 0, 1, \dots, N-1; \quad 0 \leq L \leq \log_2 N;$$

In the above formula (4.5), ω_L^n is the detail coefficient of sub band L at loci n , y_i is the copy number signal at loci i on the chromosome. The wavelet coefficients ω_L^n in (4.5) can be viewed as the difference between two averages. In places where no change point occurred in the signal, we expect ω_L^n to be zero, as it is the difference between two identical averages. When zero mean additive noise is present it will typically average out for large enough L , so that ω_L^n will still be close to 0. In places where a change point occurred, we expect a high absolute value of ω_L^n , as the two averages are different.

4.3.2 FDR Thresholding

Given a list of coefficients ω_L from a specific sub band L , we wish to keep just the larger ones, which in our case correspond to significant change points in the data. To this end we consider the False Discovery Rate (FDR) thresholding procedure [40], where FDR is defined as the proportion of false-positives out of all positives. FDR thresholding is a data-adaptive procedure, which controls the false discovery rate. Specifically,

we perform multiple hypotheses testing, where we get the null distribution of coefficients from the normal reference samples. We select the maximum number of coefficients such that the estimated FDR is kept under a predefined level q_{FDR} , where $0 < q_{FDR} < 0.5$. To apply FDR thresholding we first sort all $|\omega_L^i|$ in descending order, such that:

$$|\omega_L^{(1)}| \geq |\omega_L^{(2)}| \geq \dots \geq |\omega_L^{(i)}| \geq \dots \geq |\omega_L^{(m)}| \quad (4.7)$$

For each measurement $|\omega_L^{(i)}|$ we calculate the two-sided p-value: $p^{(i)}$. Starting from $i=1$, we then find the largest index i for which

$$p^{(i)} \leq (i/m)q_{FDR} \quad (4.8)$$

In the end we will keep the i largest coefficients, $|\omega_L^{(1)}|, \dots, |\omega_L^{(i)}|$.

4.3.3 L -Vector Haar-CIN

Since these i largest coefficients $|\omega_L^{(1)}|, \dots, |\omega_L^{(i)}|$ represent the i significant changes in sub band L resolution under the FDR level q_{FDR} , we can define CIN_L , the CIN in sub band L as the sum of these coefficients:

$$CIN_L = \sum_{k=1}^i |\omega_L^{(k)}| \quad (4.9)$$

Finally the CIN for each chromosome can be defined as a L -vector $\{CIN_{L_{\min}}, CIN_{L_{\min}+1}, \dots, CIN_{L_{\max}}\}$, including all the sub band CINs. In section 5.1 we show the Haar-CIN in Figure 5.2, we will discuss it in details in next chapter.

Chapter 5

Experiments and Discussion

In this chapter we will introduce our experiments and data sets, generate the proposed chromosome instability (CIN) indices, and discuss their applications on real copy number datasets. In section 5.1 we will perform ovarian cancer subtype classification through SVM classifier and in section 5.2 we will show the CIN application in survival analysis. In the end of the chapter we will have the conclusion and discussion on our current work.

5.1 Ovarian Cancer Subtypes Experiment

5.1.1 SNP Dataset of Ovarian Cancer Subtypes

We run our whole analysis on a SNP array copy number dataset of ovarian cancers provided by John Hopkins Medical Institutions [41]. The dataset has 54 samples, including 39 ovarian cancer samples (12 SBT (Serous Borderline Tumor), 12 LG (Low-Grade) and 15 HG (High-Grade)) as well 15 Normal samples.

To infer the copy number signals, all the tissue samples are first genotyped using Affymetrix 250K SNP arrays, then these SNP array data were processed using the dChip software: SNP data was first normalized to a baseline array with median signal intensity at the probe intensity level using the invariant set normalization method, then a model-based expression index method was employed to obtain the signal expression index for each SNP probe set in all arrays, finally the expression index for each SNP probe set was adjusted according to 15 normal samples arrays such that their median value equal to that of the reference arrays. The real copy number profiles generated by the Hidden Markov Model using dChip based on the

assumption of diploid for normal samples are later normalized by SFNM model. Due to the different lengths of the 23 chromosomes, there is different number of probes assigned on each chromosome, from 3189 on chromosome 21 to 20347 on Chromosome 1; totally we have about 250,000 probes for each sample on all the 23 chromosomes. For example, the copy number signal profile on chromosome X of a tumor sample processed by the dChip software is plotted in Figure 2.5.

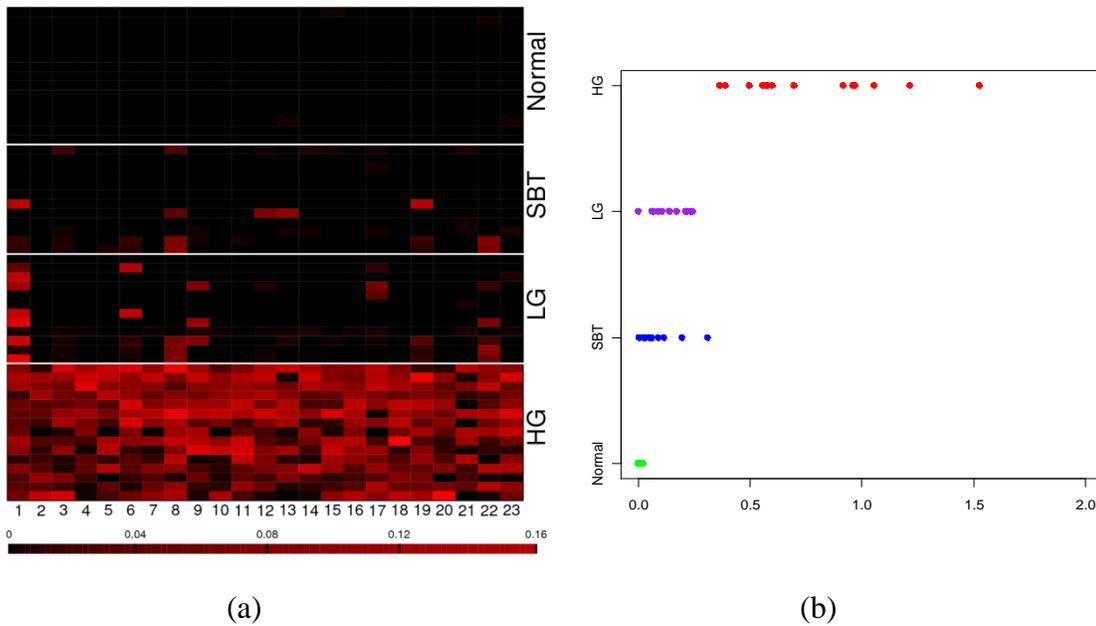


Figure 5.1 The chromosome CIN heatmap (a) and genome-wide CIN distribution (b) based on CBS detection results. Each column of the heatmap in (a) corresponds to a chromosome and each row corresponds to a sample; each dot in (b) represents a sample and each row corresponds to a phenotype.

We calculate and visualize Haar-CIN and CBS-CIN using the pre-processed ovarian datasets. From Figure 5.1 (a), HG tumors have high instabilities across all chromosomes while LG tumors are unstable only in some chromosomes, such as chromosome 1, 9, and 22. SBT tumors have lower instabilities compared with LG and HG tumors. The stabilities of the normal samples are reflected by the lowest intensities in the CIN heatmap. The transitions of stabilities from SBT to LG and then to HG are consistent with existing knowledge of ovarian cancers [41].

SBT and LG are usually considered indolent tumors and it is now believed that LG are developed from SBT. Compared with SBT and LG, HG patients are more aggressive and develop fast. Figure 5.1 (b) shows the global trend of genome-wide CIN indices. We also did t-test for genome-wide CIN between all subtypes and the p-values are listed in the Table 5.1. Based the p-values in the table, we can see that the genome-wide CINs for different subtypes are significantly different from each other in their distributions.

	Normal	SBT	LG	HG
Normal		0.0047	4.32E-07	5.76E-10
SBT			9.02E-02	1.43E-07
LG				5.63E-07
HG				

Table 5.1 p-values of genome-wide CBS-CIN between subtypes

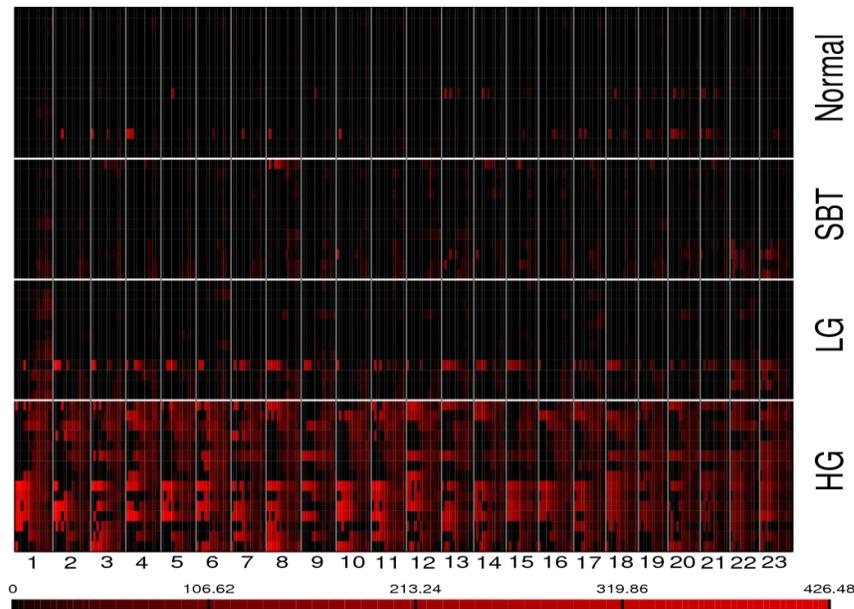


Figure 5.2 Haar-CIN of the normal and SBT, LG, and HG ovarian cancers samples

We also show the Haar-CINs for the cancer subtypes in Figure 5.2. The Haar-CIN also provides certain contrast between 4 subtypes, consistent with CBS-CIN, where HG again is

significantly different from the others qualitatively, in the next section 5.1.2 we will do subtypes classification using SVM to give quantitative results.

5.1.2 CIN-SVM Classification on Ovarian Cancer Subtypes

We apply SVM to perform classification based on our proposed predictive CBS-CINs and Haar-CINs to further test their association with the cancer subtypes. The SVM classification is done through two loops, the inner-loop is to perform SVM parameter selection, and the outer-loop is to assess the classification performance, both loops are executed with leave-one-out cross-validation. The data on the SVM classification performance are listed in Table 5.2. From the table, we can see the HG samples are easily separated from other subtypes using any of the two CIN definitions, but it is a little bit challenging to classify between Normal, SBT and LG samples, especially between SBT and LG although it is expected due to their close relationship. We obtained overall classification accuracies on the 4 subtypes, they are 61.11%, 72.22%, for CBS-CIN and Haar-CIN, respectively. Although it is imperfect, the initial result is promising in that our proposed CINs provide useful discriminatory power for subtype diagnosis.

	CBS-CIN	Haar-CIN
Normal VS. SBT	74.61%	73.37%
Normal VS. LG	81.48%	82.38%
SBT VS. LG	54.17%	70.83%
Normal VS. HG	100.00%	100.00%
SBT VS. HG	100.00%	100.00%
LG VS. HG	100.00%	100.00%
All 4 Subtypes	61.11%	72.22%

Table 5.2 SVM classification performance based on CBS-CIN and Haar-CIN

5.2 Survival Analysis on TCGA Ovarian Cancer Dataset

The Cancer Genome Atlas (TCGA), an ongoing project of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), is a comprehensive collection of genomic data for studying human cancers such as glioblastoma multiform, lung, and ovarian cancer. The project is unique in terms of the size of the patient cohort interrogated, scheduled are 500 patient samples, far more than most genomics studies, which is appropriate for our survival analysis.

5.2.1 TCGA Ovarian Cancer Dataset

We will again focus on the ovarian cancer data in TCGA, there are totally $596=298*2$ paired samples (all of which are genotyped using Affymetrix SNP 6.0 arrays), in which there are 157 'DECEASED' pairs samples with all the information available (such as when the patients took surgery, when they were followed up and when they died.). It is not quite clear to us whether 'DECEASED' and 'LIVING' samples are comparable or not, maybe it will lead to some computation bias if we combined those two types of samples. Therefore, we only performed CIN analysis based on those 157 'DECEASED' paired samples since their survival times are already determined.

As we did in section 5.1, we calculate the CBS-CINs following the processes discussed in the previous chapters, then we do survival analysis using Pearson's Correlation test [42] and Cox Proportional-Hazards Regression [43], the results are listed in the Table 5.3.

5.2.2 Pearson's Correlation test

We did correlation test for each chromosome between the CIN and the survival time to test the hypothesis of no correlation against the alternative that there is a non-zero correlation. The

population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (5.1)$$

In our case, X and Y are the CIN and the survival time respectively, the positive sign of correlation coefficient suggests that larger CIN value leads to longer survival time, while negative sign suggests that larger CIN leads to shorter survival time.

5.2.3 Cox Proportional-Hazards Regression

Cox Regression survival analysis typically examines the relationship of the survival distribution to covariates:

$$h_i(t) = h_0(t) \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}) \quad (5.2)$$

where $h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t}$ is the hazard function, which assesses the instantaneous risk of time t , conditional on survival to that time. Here X is the CIN, $h_0(t)$ is baseline hazard function and $h_i(t)$ is the hazard function for the i_{th} sample, positive sign of Cox-regression coefficient suggests that larger CIN value leads to higher risk (shorter survival time), while negative sign of Cox-regression coefficient suggests that larger CIN value leads to lower risk (longer survival time) .

Chr	correlation test		COX survival analysis	
	p-value	correlation	p-value	coefficients
1	0.7047	0.0305	0.793	-0.3215
2	0.6048	0.0416	0.749	-0.8056
3	0.0646	-0.1479	0.0907	2.201
4	0.0067	0.2155	0.0447	-3.6858
5	0.8108	0.0193	0.809	-0.3896
6	0.0212	0.1838	0.0396	-3.1598
7	0.8248	-0.0178	0.903	0.2202
8	0.9539	0.0047	0.861	-0.1524
9	0.3616	-0.0733	0.284	2.249
10	0.3111	-0.0814	0.304	1.921
11	0.7584	0.0247	0.71	-0.6395
12	0.5676	0.046	0.69	-0.6732
13	0.0321	0.1712	0.0445	-4.9922
14	0.6445	0.0371	0.85	-0.3983
15	0.0845	0.1381	0.261	-1.8796
16	0.6053	-0.0416	0.401	1.585
17	0.3068	0.0821	0.448	-0.9477
18	0.9023	-0.0099	0.869	0.2264
19	0.5262	-0.051	0.487	0.6321
20	0.0064	0.2168	0.7489	-4.1087
21	0.795	-0.0209	0.834	0.5405
22	0.027	0.1765	0.0395	-6.1563
X	0.8211	-0.0182	0.985	0.0379

Table 5.3 p-values of Pearson's Correlation test and Cox Proportional-Hazards Regression

The experimental results of the above two methods are listed in Table 5.3. In the table, we highlight the significant chromosomes (p-value<0.05) as bright yellow, and we found chromosome 4, 6, 13 and 22 are significant. For these 4 significant chromosomes, correlation coefficients are all positive and the Cox regression coefficients are all negative, so the results from both tests are consistent: larger CIN value leads to lower risk (longer survival time). The

results seem a little weird, since in our common sense bigger CIN means larger instabilities on the chromosomes, which are always the indicator of serious tumors. However, our collaborators at JHU made an alternative explanation that could serve as new hypothesis for further studies. The samples we analyzed are 'DECEASED' samples and all of them had accepted medical treatment in hospitals, these treatments are more effective for patients who have higher CINs because they are easily targeted by the medicines; while effects on patients, whose CINs are lower (many of them of relapse cancers), are not so clear since they are not as easily targeted by medicines. We need more clinical evidence for this, but at least the experiments show there are some correlations between our CINs with events in cancer development and response to therapy.

5.3 Discussion

We applied our analysis methods and tested our proposed CINs on two real ovarian cancer datasets. In the first experiment, the stability differences between cancer subtypes are clearly reflected by both the Harr-CIN and CBS-CIN heatmap, there is a trend that CIN gradually increases from normal, SBT, LG to HG, which is consistent with our existing knowledge. In later CIN-SVM classification, the CIN can be viewed as informative features since the HG subtypes can be perfectly distinguished from other subtypes, although the accuracies for other subtypes' classification are little lower. In the second experiment, the survival analysis results from correlation test and Cox regression are consistent; they both picked up chromosome 4, 6, 13 and 22 as significant chromosomes based on the CIN value and showed some correlations with the events in cancer development. Both experiments on real dataset show that our proposed CBS-CIN and Haar-CIN indeed have caught some underline information indicated by the copy number alternations and could be used as a promising analysis tool for future cancer research.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

As we discussed in the introduction character, DNA copy number analysis in SNP array data is a challenging task due to its unique characteristics. The most important one is the large numbers of observations, for example, the SNP 6.0 array currently can genotype 1.8 million probes. The uncertainty in the oligonucleotide competitive hybridization and the following imaging procedure will definitely introduce more noise, thus significantly low SNRs is another important characteristic. Detecting CNAs in tumor genomes further complicates the problem due to the complex signal patterns caused by normal tissue contamination and heterogeneous cell populations.

In this thesis, we proposed a simple but effective framework to perform copy number analysis. From the raw data process such as pre-processing of the SNP array and normalization of copy number signal, to high level process such as CNV/CNA detection, wavelet transform, CIN calculation, and the final CIN-SVM based classification in real tumor data set. The SNP array pre-processing and SFNM copy number normalization reduced part of system bias and improved the accuracy of follow-up analyses. Our proposed analysis methods effectively extracted informative features CBS-CIN and Haar-CIN using the huge size copy number profiles, and showed the promising capability of CINs in distinguishing cancer subtypes, as well as promising experimental results in survival analysis. Currently our methods have been integrated

in to G-DOC. We expect these newly defined CIN features to be useful predictors in tumors subtype diagnosis and hope it can be a useful tool in cancer research.

6.2 Future Work

There is one challenging task that we will face while our current analysis framework is not able to address. That is the heterogeneity of tissue samples. Usually a tumor tissue sample is a mix of normal cells (normal tissue contamination) and tumor cells at different stages of cancer development. The heterogeneity and normal tissue contamination highly influence the copy number analysis since it in fact introduces noise, and what we get is just a mixture signal. So what we need to do is to define a computational model to decompose the mixed copy number measurements on a tumor sample to different copy number signals of multiple cell populations to remove the heterogeneity.

Furthermore, people conduct a lot of gene expression based prognosis or association studies, so we can use copy number data for the same purpose. Although copy number data have much higher dimension than gene expression data, those signals are spatially correlated across genome, so the down-sampling or other dimension reduction methods can be considered. One limitation of our CIN is its resolution, currently it is at chromosome and genome level, but researcher are always more interested in some specific locations and genes. So our next possible research work is to define some local measurement of chromosome instability, or we can explore using gene expression and copy number data jointly for diagnosis/prognosis.

Appendix

Integrated applications into G-DOC

Major part of the work described in the thesis has already been integrated into Georgetown Database of Cancer (G-DOC), now it is public and free, users can login to perform their own biological data analysis. Here we just introduce our work which has been integrated in G-DOC and show how it looks like.

A. Introduction to G-DOC

The G-DOC is designed to serve as a cutting-edge data integration platform and integrative knowledge discovery system for the oncology and translational research communities. By aggregating public and proprietary data from across the Georgetown University Medical Center, G-DOC is expected to help bring about significant advances in personalized medicine for patients and to promote identification of new drug targets and therapeutic modalities. Using unified data portals, G-DOC allows researchers, to access and analyze clinical and research data across multiple trials and studies. The framework can be used to import data from multiple studies, to access biomedical research data, to perform analysis, and generate ad hoc queries and customized reports.

The web-based platform also contains a genome viewer that let users visualize multiple data types, including gene expression, copy number variation, and clinical outcome data. The viewer also supports flexible clinical criteria browsing to enable specific cohort selection and generate detailed reports. Users can also browse drugs of interest using chemical structure and

molecular property search functions, and study the molecular interactions of cancer drugs in a three-dimensional viewer.

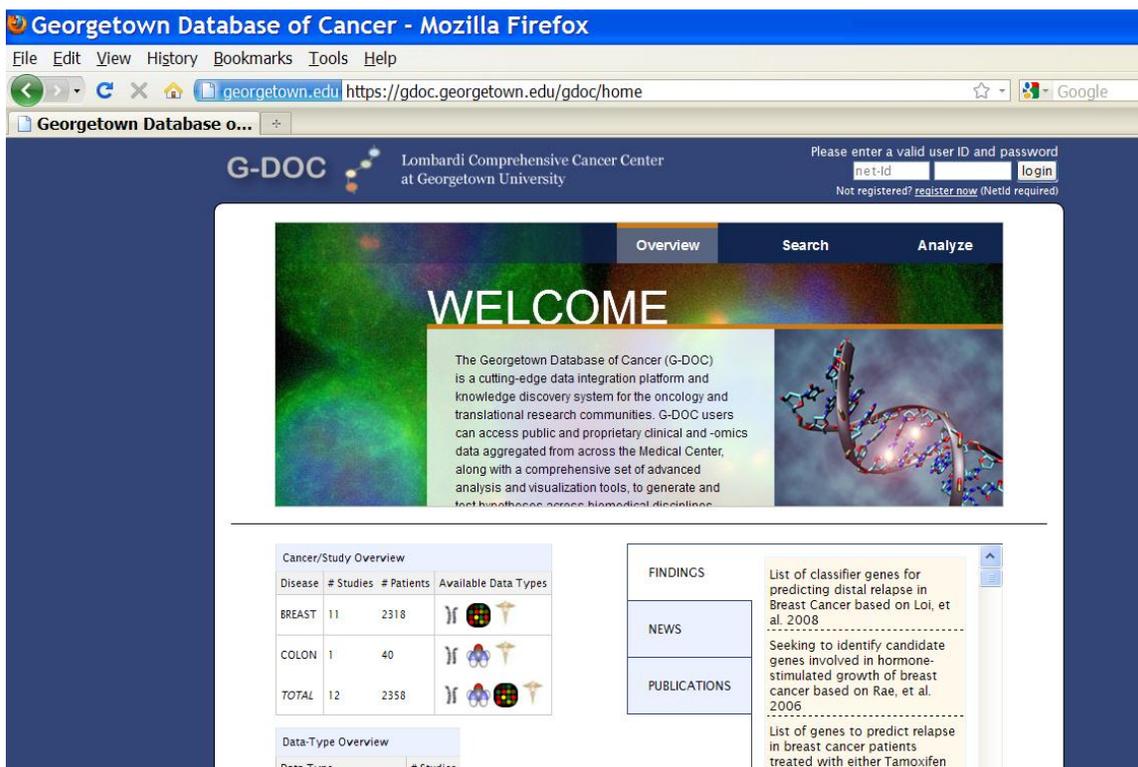


Figure Appendix.A G-DOC welcoming (Image form G-DOC² website)

The methods discussion in the thesis have already been integrated into G-DOC, such as copy number visualization according to the cytobands, chromosome instability index analysis and the copy number detection results display in Jbrowser.

² <https://gdoc.georgetown.edu/gdoc/>

B. Copy Number Visualization

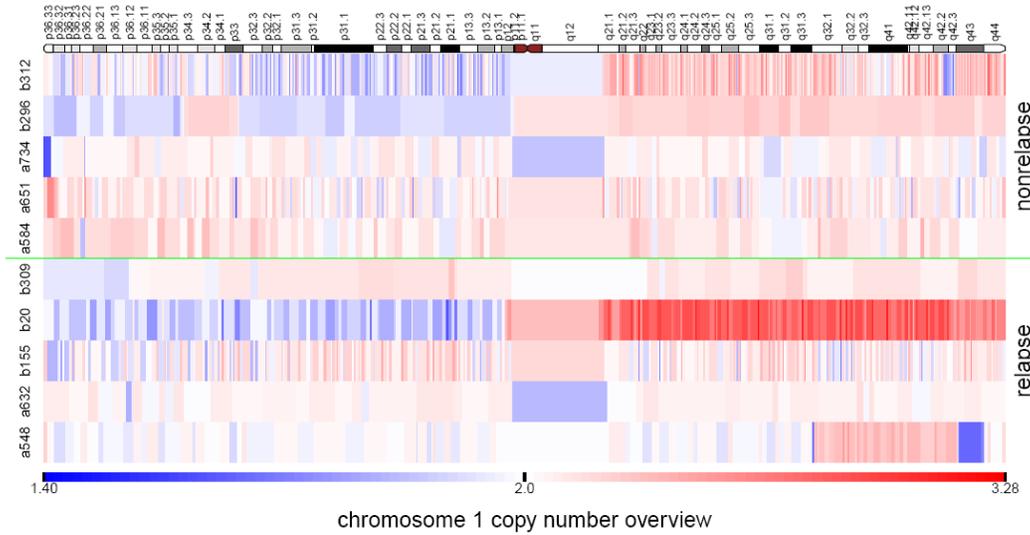


Figure Appendix.B Sample copy number profile visualization (Image form G-DOC website)

The copy number profiles can be visualized in G-DOC for group comparison. Like many commercial software, people can see the cytobands bar at the top.

C. Chromosome Instability Index Heatmap

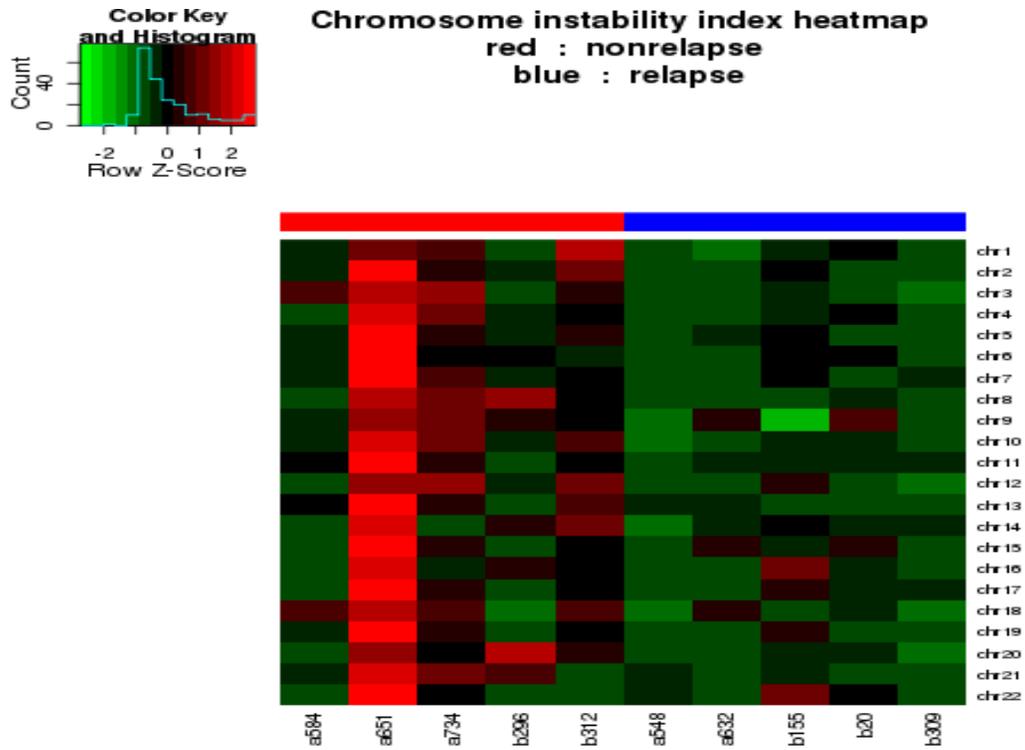


Figure Appendix.C Sample chromosome instability index heatmap (Image form G-DOC website)

D. View Copy Number Profiles through JBrowser

The copy number profiles can also be visualized in G-DOC through JBrowser, which will enable you to zoom in, zoom out and search specific position on the chromosome.

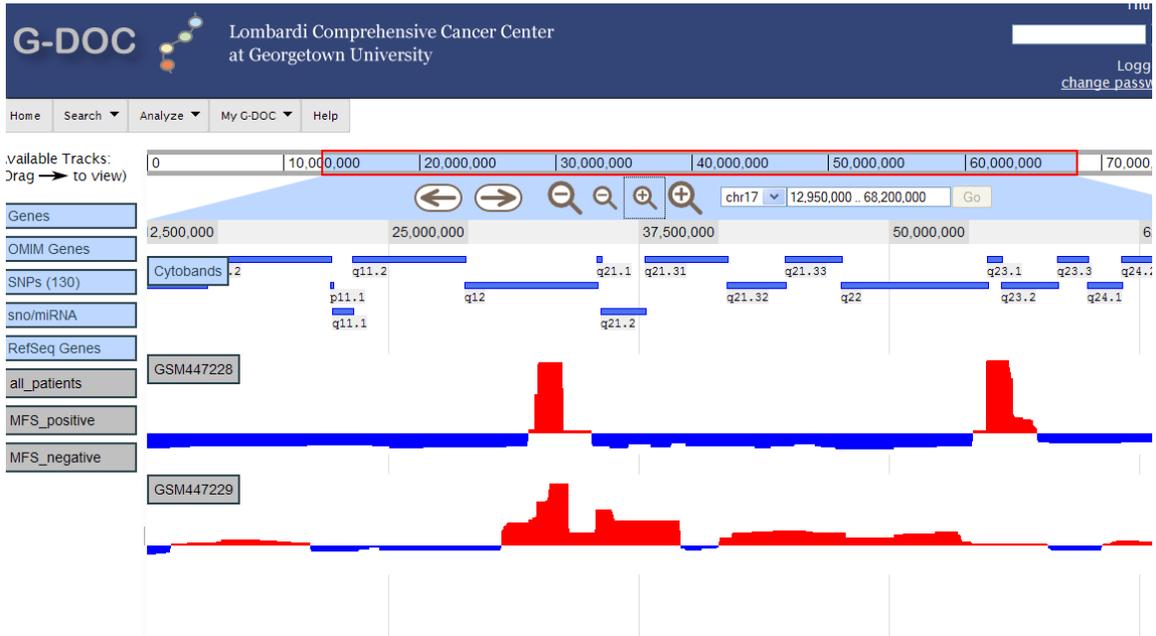


Figure Appendix.D View copy number profiles through Jbrowser (Image form G-DOC website)

References

- [1]. Lander, E.S., et al., Initial sequencing and analysis of the human genome. *Nature*, 2001. 409(6822): p. 860-921.
- [2]. Gabriel, S.B., et al., The structure of haplotype blocks in the human genome. *Science*, 2002. 296(5576): p. 2225-9.
- [3]. TCGA -(<http://wiki.nci.nih.gov/display/TCGA/TCGA+Home>).
- [4]. T. Strachan, and A. P. Read, *Human Molecular Genetics*, 2 ed., New York: John Wiley & Sons, 1999.
- [5]. SNP fact Sheet -(http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml).
- [6]. Collins, F.S., M.S. Guyer, and A. Charkravarti, Variations on a theme: cataloging human DNA sequence variation. *Science*, 1997. 278(5343): p. 1580-1.
- [7]. L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85-97, Feb, 2006.
- [8]. Check E. Human genome: patchwork people. *Nature*. 2005 Oct 20; 437(7062):1084-6.
- [9]. N. P. Carter, "Methods and strategies for analyzing copy number variation using DNA microarrays," *Nat Genet*, vol. 39, no. 7 Suppl, pp. S16-21, Jul, 2007.
- [10]. S. Kim and A. Misra (2007) SNP genotyping: technologies and biomedical applications, *Annu Rev Biomed Eng*, 9(289-320).
- [11]. X. Zhao, C. Li, J. G. Paez et al., "An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays," *Cancer Res*, vol. 64, no. 9, pp. 3060-71, May 1, 2004.
- [12]. Li C and Wong WH (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology* 2(8): research0032.1-0032.11
- [13]. Schadt EE, Li C, Su C, Wong WH: Analyzing high-density oligonucleotide gene expression array data. *J Cell Biochem* 2001, 80:192-202.
- [14]. Li C and Wong WH (2001a) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proc. Natl. Acad. Sci.* Vol. 98, 31-36.
- [15]. Dugad R, Desai UB. A tutorial on Hidden Markov Models. Technical report, No.SPANN-96.1, 1996.
- [16]. J. Fridlyand, A. M. Snijders, D. Pinkel et al., "Hidden Markov models approach to the analysis of array CGH data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 132-153, 2004, 2004.
- [17]. G. McLachlan, and D. Peel, *Finite Mixture Models*: Wiley-Interscience, 2000.
- [18]. Abbi R, El-Darzi E, Vasilakis C, and Millard P.H, A Gaussian mixture model approach to grouping patients according to their hospital length of stay, 21st IEEE International Symposium on Computer-based Medical Systems Finland, 2008.

- [19]. Pernkopf, F.; Bouchaffra, D., “Genetic-based EM algorithm for learning Gaussian mixture models”, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, volume 27, Issue 8, Aug. 2005 Page(s):1344 - 1348.
- [20]. Roberts, S.J.; Husmeier, D.; Rezek, I.; Penny, W. “Bayesian approaches to Gaussian mixture modeling”, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Volume 20, Issue 11, Nov. 1998 Page(s):1133 – 1142
- [21]. B. L. Pellom and J. H. L. Hansen, “An efficient scoring algorithm for gaussian mixture model based speaker identification,” *IEEE Signal Process. Lett.*, vol. 5, no. 11, pp. 281–284, Nov. 1998.
- [22]. C. Constantinopoulos, M. K. Titsias, and A. Likas, “Bayesian feature and model selection for Gaussian mixture models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1013–1018, Jun. 2006.
- [23]. Arthur Dempster, Nan Laird, and Donald Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977
- [24]. Tenmoto, H., Kudo, M., and Shimbo, M. 1998. “MDL-Based Selection of the Number of Components in Mixture Models for Pattern Classification”. *Lecture Notes In Computer Science*, vol. 1451. Springer-Verlag, London, 831-836.
- [25]. P. Gr ünwald. A tutorial introduction to the minimum description length principle. In P. Gr ünwald, I. J. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 3–81. MIT Press, 2005.
- [26]. Hup é P., et al. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions, *Bioinformatics*, 20, 3413-3422.
- [27]. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557–572 (2004)
- [28]. Venkatraman, E.S. and Olshen, A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data, *Bioinformatics*, 23, 657-663.
- [29]. Fridlyand, J., et al. (2004) Hidden Markov models approach to the analysis of array CGH data, *Journal of Multivariate Analysis*, 90, 132-153.
- [30]. Willenbrock, H. and Fridlyand, J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21, 4084–4091.
- [31]. Lai, W.R. et al. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21, 3763–3770
- [32]. SHABAN, S. A. (1980). Change-point problem and two phase regression: an annotated bibliography. *International Statistical Review* 48, 83–93.
- [33]. BASSEVILLE, M. (1988). Detecting changes in signals and systems—a survey. *Automatica* 24, 309–326.
- [34]. Stuart, R. D., *An Introduction to Fourier Analysis*, Science Paperback, Cambridge, 1966
- [35]. Mallat, S. “A wavelet tour of signal processing.” Academic Press, 1998, London.

- [36]. Benjamini, Y. and Hochberg, Y. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *J. Roy. Statist. Soc.*, 1995, Ser. B 57 289-300.
- [37]. Ben-Yaacov, E. and Eldar, Y.C. (2008) A fast and flexible method for the segmentation of aCGH data, *Bioinformatics*, 24, i139-145.
- [38]. Shore, J., "On the Application of Haar Functions, Communications," *IEEE Transactions on*, Volume 21, Issue 3, Mar 1973 Page(s):209- 216
- [39]. S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: A Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 11, No. 7, July 1989.
- [40]. Benjamini, Yoav; Yekutieli, Daniel (2001). "The control of the false discovery rate in multiple testing under dependency". *Annals of Statistics* 29 (4): 1165–1188.
- [41]. K.T. Kuo, B. Guan, Y. Feng, T.L. Mao, X. Chen, N. Jinawath, Y. Wang, R.J. Kurman, M. Shih Ie and T.L. Wang (2009) Analysis of DNA copy number alterations in ovarian serous tumors identifies new molecular genetic changes in low-grade and high-grade carcinomas, *Cancer Res*, 69(9):4036-4042.
- [42]. Mario F. Triola, *Essentials of Statistics* (3rd Edition), Addison Wesley, 2006
- [43]. Fox, John 2002. *Cox Proportional-Hazards Regression for Survival Data*. Appendix to *An R and S-PLUS Companion to Applied Regression*.