Hidden Markov Model Inference Copy Number Change in Array-CGH Data

by

Yunyu Zhang

M.D.
Peking University Medical School, 1994

Submitted to Harvard-MIT Division of Health Science and Technology in Partial Fulfillment of the Requirements for the Degree Of

Master of Science in Biomedical Informatics
at the
Massachusetts Institute of Technology

MAY 2005 [June 2005]

Signature of Author.....
Harvard-MIT Division of Health Science and Technology
May 7th 2005

Certified by.....
Lynda Chin, M.D.
Associate Professor of Dermatology
Harvard Medical School
Thesis Supervisor

Certified by.....
Cheng Li, Ph.D.
Assistant Professor of Biostatistics and Computational Biology
Harvard School of Public Health
Thesis Co-supervisor

Certified by.....
Cameron W. Brennan, M.D.
Assistant Attending, Neurosurgery Service
Memorial Sloan-Kettering Cancer Center
Thesis Co-supervisor

Accepted by.....
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Science and Technology

Abstract

Cancer development and progression typically features genomic instability frequently resulting in genomic changes involving DNA copy number gains or losses. Identifying the genomic location of these regional alterations provides important opportunities for the discovery of potential novel oncogenes and tumor suppressors. Recently, array based competitive genomic hybridization (array-CGH) has become available as a powerful approach for genome-wide detection of DNA copy number changes. Array-CGH assesses DNA copy number in tumor samples through competitive hybridization on microarrays containing probes for thousands of genes. The datasets generated are complex and require statistical methods to accurately define discrete and uniform copy number from the data and to identify transitions between genomic regions with altered copy number. Several approaches based on different statistical frameworks have been developed. However, a fundamental informatic issue in array-CGH analysis remains unsolved by these methods. In particular, sample-specific data compression, a result of tumor cells being commonly admixed with normal cells in many tumor types, must be accounted for in each sample analyzed. Additionally, in order to accurately assess deviations from normal copy number, the copy number readout must be shifted to faithfully represent the baseline copy number in each tumor sample. Failure to appropriately address these issues reduces the accuracy of the data in hard-threshold based high-level analysis. By using the natural framework Hidden Markov Models (HMM) to model the distribution of array-CGH signals, a method infer the absolute copy number and identify change points has been developed to address the above problems. This method has been validated on independent dataset and its utility in inference on array-CGH data is demonstrated here.

Table of Contents

1. Background

Tumorigenesis progression in human require the accrual of genetic lesions that result in aberrantly functioning genes that control many aspects of cellular function including proliferation, apoptosis, genome integrity, angiogenesis, and invasion of metastasis [1]. The discovery and functional evaluation of these cancer-relevant genes is essential for understanding the biology of cancer and for clinical applications, including identification of therapeutic targets, early detection and improved prediction of cancer risk and disease course. Many factors can result in variant gene function including point mutations, epigenetic modifications, and changes in genome copy number and structure (chromosome aberrations).

1.1. Chromosome aberration, technology and functional significance

Chromosome aberrations are defined by a broad range of changes including alteration on ploidy, gain or loss of individual chromosomes or portions thereof and structural rearrangement (fig 1.1, modified from[2]). These structural changes may involve translocation of chromosome material from one chromosome to another. Equal exchanges of material between two chromosomal regions are referred to as balanced or reciprocal translocations. On the other hand, unequal exchanges may also occur and are termed unbalanced or non-reciprocal translocations. These unbalanced translocations and other forms of structural rearrangement may result in amplifications or deletions of chromosome material. Amplifications may present as small acentric fragments (double minute chromosomes) or may be incorporated into tumor chromosomes in nearly contiguous homogeneously staining regions (HSRs) or interspersed throughout the genome. Notably, individual HSRs or other sites of amplified DNA may include genomic DNA originating from multiple different regions.

An increasing number of genomic and molecular genetic technologies have been developed to detect chromosome aberrations. These include analysis of chromosome banding (Mitelman Database of Chromosome Aberration in Caner), high-throughput analysis of loss of heterozygosity (LOH; [3,4]), conventional and array-based comparative genomic hybridization (CGH[5-7]), fluorescence in situ hybridization (FISH; [8,9]), restriction

landmark genome scanning (RLGS; [10]) and representational differential analysis (RDA; [11]) . Some technologies, such as RLGS, analysis of LOH and RDA can also detect allelic imbalance that occurs by somatic recombination without net copy number change.

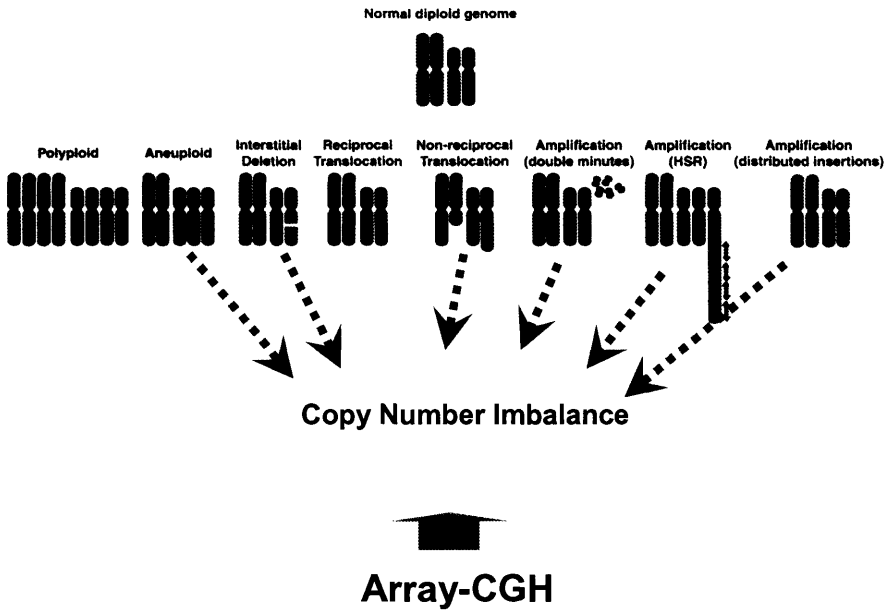

Figure 1.1 Schematic illustration of mechanisms of chromosomal aberrations and which will cause copy number change. Modified from [2].

It is widely believed that regions of recurrent genomic aberrations contain genes that are important for tumor initiation and development. In many cases, such aberrations contain known oncogenes or tumor suppressor genes whose expression levels are altered by the genomic change. Classic examples in solid tumors include amplifications of established oncogenes, such as EGFR [12], MYC[13], ERBB2[14], CCND1[15] and Ras family members [16]. Other aberrations involve loss of specific regions of the genome. Deletions involving specific loci are important in the inactivation of tumor suppressor genes, such as PTEN and CDKN2A. Elimination of the remaining normal alleles in cases of inherited mutation has been implicated in the inactivation of known tumor suppressor gene RB1, BRCA1, BRCA2, PTPRJ and TP53.

1.2. CGH and array-based CGH

Comparative genomic hybridization (CGH) was developed as a molecular cytogenetic technique that overcomes difficulties presented by conventional fluorescence in situ hybridization (FISH) analysis[5]. It allows the entire genome to be scanned, in a single step, for copy-number aberration in chromosomal material. In standard CGH procedures, genomic DNAs isolated from test and reference samples are labeled respectively with red and green fluorescent dyes. Each labeled DNA is subjected to competitive hybridization to normal metaphase chromosomes; hybridization of repetitive sequences is blocked by addition of Cot-1 DNA. The ratios of red and green fluorescent signals in paired samples, usually tumor-normal pairs, are measured along the longitudinal axis of each chromosome. Chromosomal regions involved in deletion or amplification in test DNA appear green or red respectively, but chromosomal regions that are equally represented in test and reference DNAs appear yellow.

CGH analyses of solid tumors have revealed a number of recurrent copy-number aberrations including amplifications that had not been detected previously by any other technique. In an early example, CGH revealed frequent tumor-specific amplifications at chromosomes 3q26-27 and 20q13 in various tumors where the oncogenic target genes were subsequently identified, *PIK3CA (*3q26) [17]in ovarian cancers and *ZNF217* (20q13) in breast cancers[18]. However, CGH to metaphase chromosomes can provide only limited resolution of 5–10 Mb for detection of copy-number losses and gains, and 2 Mb for amplifications. However, many of the most informative events are small to contain only involved a few genes and spans only a few hundreds of kilo-base pairs.

With the availability of human and mouse genome sequence and map, this limitation has been overcome by adapting the evolving microarray platform. Figure 1.2 [19] illustrates the procedure of how array based CGH is conducted with bacterial artificial chromosome (BAC). BAC-based arrays were the first to be proven highly effective in defining the location of regional copy number changes[6]. Current BAC arrays typically offer approximately 1 Mb of coverage (containing 3000 BACs), translating into a resolution limit of 2 Mb[20-22]. Using this platform, the additional delimitation of regional alterations is

made possible by custom microarrays containing BAC contigs that tile across the locus of interest in an iterative locus specific manner. Prior work has clearly documented the effectiveness of iterative BAC array-CGH profiles to identify candidate cancer genes residing in a focal amplicon. Several studies have documented the utility of cDNA-based microarrays for CGH profiling of human cancers. These studies have demonstrated that commercially available cDNA array-CGH platforms are sufficiently robust to detect regional single-copy changes [23], provided the high background probes are eliminated by empirical and bioinformatics means. Oligo-based CGH experiments were first introduced by using commercially available expression microarray. The median resolution of a 22K expression array is ~ 50kb for human and mouse. With elimination of ~5K probes not suitable for CGH hybridization, these oligo probes designed for expression do offer improved signals and noise when compared to cDNA platform. Genomic oligo microarrays (Agilent) recently developed specifically for CGH hybridization have been shown to perform with even higher signal to noise ratio while achieving a higher resolution of 30kb on 44kb probes.

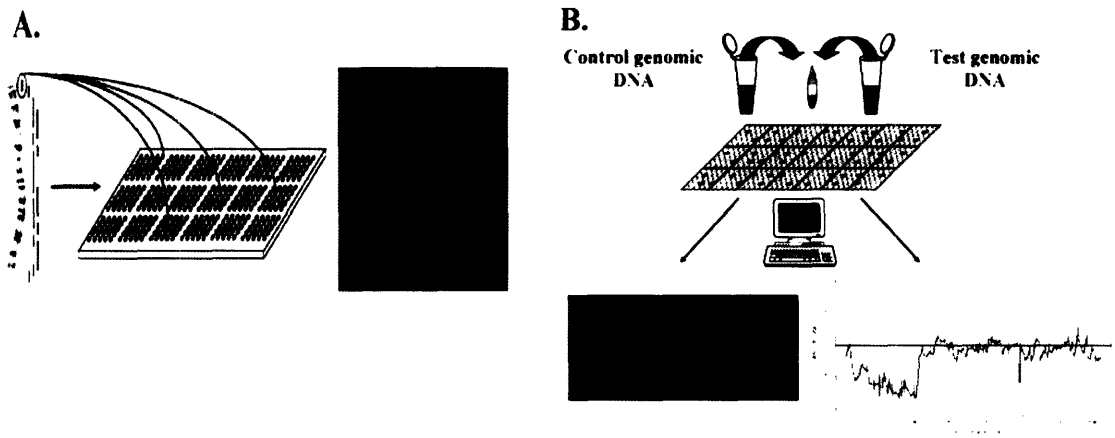

Figure 1.2. Procedures of array-CGH experiment [19].

(A) Large-insert clones, derived from a human chromosome, are printed onto a glass microscope slide (arrayed). The array can be stained to show the morphology and placement of each `spot' of cloned DNA (far right).

(B) Genomic DNA samples from a control (left) and test (right) are differentially labeled with two different fluorochromes. The labeled DNA is mixed and placed on the microarray. Computer imaging reveals a yellow hybridization color for all clones that are in equal proportion between the control DNA and test DNA (middle and lower left). Those clones deficient in the test DNA, will appear more green; those clones in excess in the test DNA, as compared to the control DNA, will appear, more red (middle). A plot of the ratio between control and test DNA for each clone (lower right) will reveal dosage differences, visualized as a deviation of the ratio from zero (horizontal red line).

1.3. Data features of array-CGH log2 ratio

1.3.1 Type of changes defined in array-CGH log2 ratio

Array-CGH measures the relative copy number of the tumor (T) against reference(R) genomic DNA, which is reflected in its log2 T/R ratio. After the signal is normalized, they are aligned and plotted along the chromosome by their physical position (figure 1.3). Biologically, loss is defined as a relative decrease of 1 or 2 copies relative to a diploid reference (R), and gain is defined as a gain of 1 or more copies. However, since underlying ploidy of a tumor is not necessarily diploid, the T/R ratio in fact reflects merely the gain or loss relative to the underlying ploidy of a particular sample.

In our analysis, an array-CGH log2 ratio of +/-0.2 is often used as a threshold to make the call of changed clone, for most platform and samples with a profile standard deviation of 0.1~0.3. A threshold of +/- 1~1.5 is often used to define the high amplitude change like amplification and homozygous deletion. The low amplitude gain and loss tends to involve large region like entire chromosome or its arms, while amplification and homozygous deletion tends to be focal and only involves loci up to several mega bases and has more functional significance.

In a normal female - male hybridization profile (figure 1.4), autosomal chromosomes are present in 2 copies with aCGH probes showing log2 ratio around 0. One-copy or actually two-fold change gain on chromosome X of the female is represented by a plateau of elevated log2 ratio with mean around 0.5. Chromosome Y is completely deleted seen as scattered negative log2 ratios, with the wide scatter due to non-specific hybridization of non-Y labeled products in the female sample. Figure 1.4 is a typical tumor profile from a primary melanoma. All major types of change are presented in this profile: entire chromosome gain of 7, loss of 6q, amplification of 1q and deletion of 9p. The gain and loss change are classically wide giving a plateau shape and amplification and deletion events are much sharper.

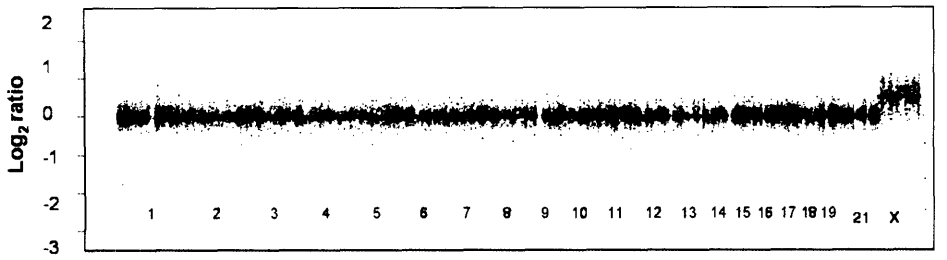

Figure 1.3. Genomic profiles of female versus male normal genomic DNAs. Array-CGH profiles of female DNA against male DNA as reference (both pooled samples) with X-axis coordinates representing oligo probes ordered by genomic map positions. Average log2 ratio of the probes on X-chromosome is around 0.5. Y chromosomes are deleted with log2 ratio scattered between 0~-3.

Figure 1.4. Genomic profiles of tumor (melanoma, versus male normal genomic DNAs. Array-CGH profiles of female DNA against male DNA as reference (both pooled samples) with X-axis coordinates representing oligo probes ordered by genomic map positions. Different types of change are present is this profile as (1) gain on 7, 11p. (2) loss of 6q and 8p. (3) amplification of 1q. (4) homozygous deletion of 9p (yellow arrow).

1.3.2 Spatial correlation of the array-CGH log2 ratio

In homogenous samples, array-CGH should report integer copy numbers change representative of the unit copy gains and losses seen, for instance, in SKY experiments. Biologically, there have to be limited numbers of genomic events for cancer cell to survive and proliferate, so large parts of cell chromosomal materials remains intact. This phenomenon translates visually into array-CGH profile as step-wise up and down like consecutive segments with some focal change as in figure 1.4. All the probes on a single segment are measuring a uniform copy number thus should share the same log2 ratio, aside from the effects of noise and artifact. This strong spatial correlation between the neighboring probes along the chromosome is a unique feature of array-CGH compared to other profiling, such as gene expression. The closer the two probes, the more likely they are detecting the same segments and reporting the same copy number. Also, a stepwise pattern in the profile suggests that a copy number change happens between 2 probes, and the change is discrete instead of continuous or gradient. Thus, the uniform log2 ratios carried by all probes are in a discrete distribution and can be translated into relative copy number.

1.3.3 Absolute quantification of array-CGH data

Since the normal male or female pooled genomic DNA is routinely used as common reference, the reference copy number is known to be 2 for all autosomes and 1 or 2 on the X chromosome. If relative gene copy number of each probe is definable, the absolute copy numbers are generally definable as well, except for cases of extreme aneuploidy. Unlike the expression data which is continuous and is comparable only gene wise with no inherent reference expression level, array-CGH has a "ground-truth" that permit comparison of the gene copy number comparable across samples. This justifies the strong need for development of appropriate analytical tools for CGH analysis which takes into account of these unique features and enable the definition of "ground-truth".

However, there're several hurdles to achieve such absolute quantitative analysis on array-CGH data. First, in the normal and tumor profiles shown previously, the observed signals are lower than theoretical, as phenomenon of so-called "data compression". For example, the log2 ratio signal mean of the probes on X-chromosome in normal

female/male hybridization is supposed to be 1, the actually observed is only around 0.5 (figure1.3). Mostly caused by cross-hybridization, this problem has been addressed in many previous studies. Pollack et al [24]hybridized samples with 1~5 copies of X-chromosome to normal female samples on cDNA platform. When mean fluorescence ratios of X-chromosomal cDNAs from each experiment were plotted against number of X chromosomes and fitted with a linear regression model, the slope of the model is only 0.7. This factor can be used to correct the data. Meanwhile, normal tissue contamination is another source commonly contributing to the reduction the signals. Hodgston et al showed the signal will decrease linearly with the increasing of the proportion of the contaminated normal tissue [25]. This reflects in real data as that the signals of primary tumors usually are slightly lower than the signals of cell lines. This is reflected in real data in that the signals of primary tumors, admixed with normal stroma and infiltrating leukocytes, usually are slightly lower than the signals of cell lines. Moreover, this factor depends on the degree of contamination and experiment variation: different samples show different levels of data compression. Figure 1.5 shows an example of BxPC from pancreatic cell lines. The log2 ratio has been median-filtered to reduce probe noise and make the change pattern easier to observe. The karyotype of the sample is 53<2n>, the signal mean of 2 copy resides at -0.17~0.18 and 0.14 for 3 copy and 0.3 for 4 copies. The log2 ratio 0 is in the middle of the signal mean of 2 copy and 3 copy. In other words, there is no direct translation between the uniform segment log2 ratio and absolute copy number.

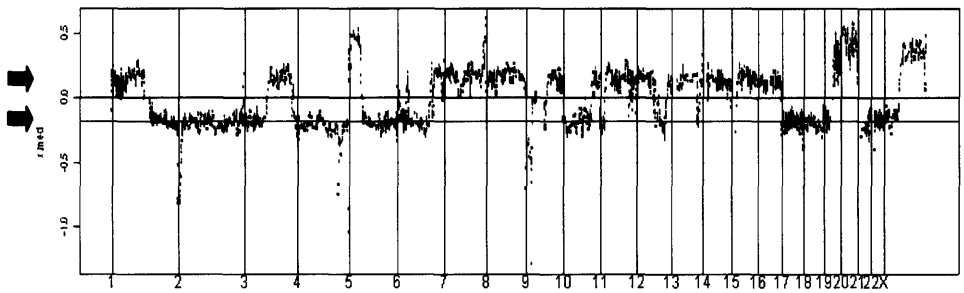

Figure 1.5. Genomic profiles of BxPC3, a pancreatic cell line versus male normal genomic DNAs. Array-CGH profiles of female DNA against male DNA as reference (both pooled samples) with X-axis coordinates representing oligo probes ordered by genomic map positions. Median-filtering has been applied to reduce the noise. DNA which 2 (red arrow) and 3 copy (green arrow) has an average log2 ratio of -0.18 and 0.14 respectively.

1.4. Current analysis methods and limitations

To date, several methods have been developed utilizing the special features of the array-CGH data. These methods report uniform relative copy number in terms of median log2 ratio of segments, but avoided defining the absolutely copy number of each profile.

1.4.1 Circular Binary Segmentation (CBS)

Oshen et al first developed a method called circular binary segmentation to define the change points on array-CGH data [26]. For indexed data set $x_1, x_2, ...x_i, x_j, x_n$ ($1 \le i \le j \le n$) and $S_{ij}=x_i+x_{i+1}+..+x_{j-1}+x_j$ be the partial sum. The circular binary segmentation is derived using the statistics Z_{ij} defined as.

$$Z_{ij} = \frac{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)}{\sqrt{1/(j-i)+1(n-j+1)}}$$

Under null hypothesis, Z_{ij} have zero mean. Basically, Z_{ij} is calculating the difference of the mean between the segments *i..j* and *j..n*, adjusted by the position of the *j*. A change is declared if $z=max(|z_{ij}|)$ exceeds an appropriate threshold level defined by normality of x_i's . As a modification of original binary segmentation method, they added a permutation approach to relax the normality assumption of the binary segmentation procedure.

1.4.2 Unsupervised HMM partitioning

Fridlyand et al proposed an unsupervised partitioning method to define the boundary of uniform copy number using Hidden Markov Model [27] (details about HMM will be given in the method section). The procedure starts with a predefined k-state HMM, each data points was first given a state they individually most close to in terms of the distance to the log2 ratio mean of those states. Then they use iterative Baum-Welch algorithm to re-estimate the parameters of the model. The procedures are repeated for all models with states from 1 to *k*-1 and the best model (*k*) was chosen based on the penalized maximum likelihood defined as:

$$\psi(K) = -\log(Lik(\lambda \mid O)) + q_K D(L)/L, \quad K = 1,..., K\max,$$

where q_K is the number of the parameters corresponding to the number of states, *K*; and *D(L)* is a function of the number of L clones (probes) on a chromosome. Note that *D(L) = 2* gives AIC or Akaike's information criterion [1] and *D(L) = log(L)* one

obtains the Schwartz BIC or Bayesian information criterion. They claimed usually a five-states HMMs are enough to describe almost all the changes even the complicate ones.

After the model fitting is done, two states with the closest median, estimated from the signals belong to them respectively, were merged if their difference is less than *d*, a pre-defined threshold. The procedure also goes iteratively until the difference exceeds *d*. When the state merging is done, the sample standard deviation $\boldsymbol{\sigma}$ computed as the median absolute deviation (MAD) of the clones in the states containing at least 20 clones located on the chromosomes partitioned in ≤ 3 states. A clone is identified as an outlier if its value differs from the median value of their state by $\geq 5\ \boldsymbol{\sigma}$. Focal changes was defined based on the outlier clones. In this method the states merging was done after the model fitting and subject to a predefined threshold *d*.

1.4.3 Other approaches

Jong et al. [28] applied a genetic local search algorithm to segment the clones into clusters. Autio et al. [29] and Picard et al. [30] used dynamic programming to define change points given known numbers of segments. Picard also implemented a penalized maximum-likelihood model, to automatically provide the global optimum of segments. There are also a few recent approaches; one is "Cluster Along Chromosomes" (CLAC) method. [31] It builds a hierarchical clustering-style trees bottom up along each chromosome arm (or chromosome), and then "interesting" clusters can be selected by controlling the FDR at a certain level.

1.4.4 Deficiency of the current methods and Motivation of this study

All the above methods are able to recognize the step-wise change pattern in the data, define change points and derive uniform underlying relative copy number. But the uniform copy number for a particular segment was obtained by estimating the mean or median of the log2 ratio on a particular segments; it does not translate to absolute copy numbers thus the results still carries data compression problem and does not define meaningful baseline of the reference. This could pose a problem in threshold cut off based high level analysis. Take BxPC as example, the gain of chromosome 1p, 3q will be missed if the usual cutoff of

+/- 0.2 or 0.25 is used for identify changes. For both single sample analysis and group comparisons especially with a smaller sample size, this problem will cause larger false positive/negative rate. Given the potential of absolute quantification based on the spatial correlation, discrete distribution and known reference copy number of array-CGH data, a method to infer biologically direct interpretable copy number will help improve the accuracy of the high level analysis.

2. Methods and Results

The goal of this thesis is to develop a method to derive absolute copy number from the array-CGH data. The approach is made by starting with a clean, well-annotated dataset to retrieve the signal distribution from the array-CGH log2 ratio and then train a HMM model on top. Testing on a variety of independent datasets with different noise levels and proportion of changes were conveyed to validate the method. This method was compared to other methods mainly circular binary segmentation on break points and focal change detection.

2.1. Datasets

Published and in-house datasets totaling 200 samples from 2 major platforms: BAC and oligonucleotide platforms are used in this study.

2.1.1 Long oligonucleotide-array datasets

Commercially available long oligo-array (50 ~70 mer) has been popular in array-CGH experiment recently. Originally designed for expression profiling, the noise level caused by cross-hybridization on this platform is slightly higher than BAC array. The average standard deviation is 0.2~0.3 of the log2 ratio in unchanged part. The following datasets have been generated in the lab with 60-mer expression microarray (Agilent Technologies).

2.1.1.1. Pancreatic cancer cell lines

The dataset contains 9 well-annotated profiles which have been used to demonstrate the feasibility of oligo platform in the earlier publication [7]. SKY has also been performed and indicates a high homogenous population for all samples. With SKY observable copy number ranges from 1 to 7 from 2 diploid, 6 triploid and 1 tetraploid sample with mediate level of genomic complexity, this is an ideal dataset to study the signal distribution of different copy numbers. Several interesting complicated focal loci, either high or low amplitude are present in the dataset. Real-time quantitative PCR has been carried out to give a more accurate measurement of the relative copy number to validate some of them.

2.1.1.2. Primary Multiple Myeloma

The 67 samples in this primary multiple myeloma dataset are typically diploid with less complicated changes in terms of number of events within a profile. This data set can be used as both for training set and testing set.

2.1.1.3. Primary Glioblastoma

Analyses of array CGH data have shown that the genomes of established tumors are remarkably stable, as evidenced by similarity of tumor recurrences to primary tumors. Of total 35 samples, there are 12 original and recurrent pairs and 2 duplicates in this dataset.

2.1.2 Public BAC-arrays dataset

BAC array datasets are selected. In both dataset, each array contained 2276 mapped partial BACs spotted in triplicates. Comparing to long oligo arrays, it has much sparse resolution (1/5) but lower noise with standard deviation around 0.1 for unchanged part. Since BAC dataset has been largely used to demonstrate other methods, the purpose to including them is to examine the method on a different platform but with sparser resolution.

2.1.2.1. Coriel cell lines

The data consists of single experiments on 15 fibroblast cell lines containing cytogenetically mapped partial or whole-chromosome aneuploidy (http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754 S1.html). There are only 1 or 2 characterized chromosome aberrations presented in each sample. It has been used to demonstrate both the CBS method and unsupervised HMM partition because of its simplicity.

2.1.2.2. MMR cell lines

The dataset includes 10 MMR deficient and 10 proficient cell lines. They are used to demonstrate the unsupervised HMM partition method by Fridlyand (Complete data set is available at http://cc.ucsf.edu/albertson/public.). Since many of the cell lines are from the NCI 60 panel, the Spectrum Karyotyping (SKY) of which are also available from NCBI's SKY/M-FISH/CGH database

(http://www.ncbi.nih.gov/sky/skyweb.cgi?form_type=submitters).

2.2. Data preprocessing

The public BAC dataset were downloaded from the web and was merged into a single table. Since the data was already normalized and all clones were annotated with the physical position, no further preprocessing was needed. For all oligo dataset, normalization procedures were performed to eliminate the common array experiment bias. The annotation of the probes in terms of their physical genome position and the gene they resides on are generated.

2.2.1 Normalization and filtering

Microarray data contain inherent systematic measurement errors arising from variations in labeling, hybridization, spotting or other non-biological sources. Normalization procedures, which adjust microarray data to remove such systematic variations, are therefore important for subsequent analysis. Within-slide normalization aims to correct dye incorporation differences which affects all the genes similarly, or genes with the same intensity similarly [32] One scatter-plot based normalization technique that is particularly suitable for balancing the intensities is called locally weighted scatter-plot smoothing (LOWESS) [33] and its original application was for smoothing scatter-plots in a weighted, least-squares fashion. Lowess normalization has been applied to all the oligo samples to remove 2 types of bias: intensity and GC content.

The intensity-dependent bias, caused by unbalanced dye efficiency, often appears as a curvature in MA-plot (figure 2.1 a). After lowess normalization, the log2 ratio become independent of the signal intensity. (figure 2.1b).

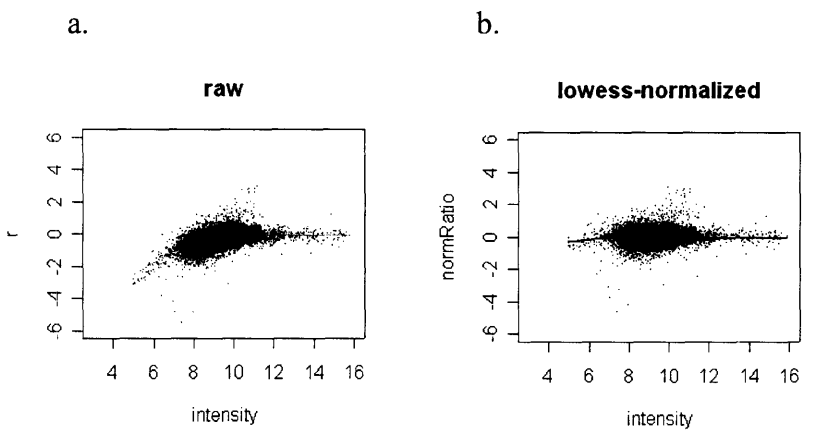

Figure 2.1. Lowess correction on intensities for array-CGH raw log2 ratio.

M-A plot of raw log2 ratio v.s. average log2 intensity from both channels

M-A plot for raw log2 ratio v.s. average log2 intensity from both channels after Lowess normalization.

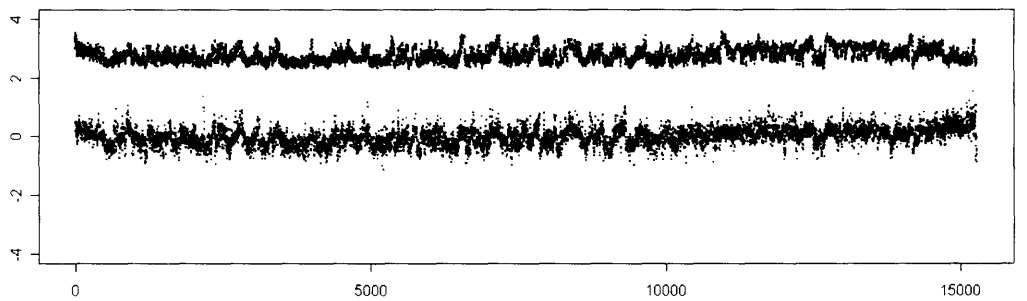

Figure 2.2. Correlations between genome GC content and array-CGH profiles.

The x-axis is the probes in the order of their genomic position. The y-axis is arbitrary units. The purple line is the profile of median smoothed (window size 5) array-CGH profile of a nevus sample. It's DNA was whole genome amplified for this experiment. no obvious genomic lesions. The blue line is the genome GC-content in a 70kb around of the probe, also median smoothed with window size 5. The correlation of two between median filter is 0.25 and thereafter is 0.61.

The probe/genome GC content related bias is unique in array-CGH experiment, especially on whole genome PCR amplified tumor DNA sample. Unlike a curvature with the intensity, usually the lower GC content, the lower the log2 ratio. The correlation between intensity bias corrected log2 ratio and genomic/probe GC content can be as high as 0.25 (figure 2.2). When those high GC correlated profile are visualized along the chromosome, they presents as a strong wavy local data trend, which might induce undesirable breakpoints when applying change point defining methods. Lowess correction on GC content effectively removed this bias.

Most of the experiment has a dye-swap hybridization replicate. If R_1, G_1 and R_2, G_2 are the intensity of the red (*Cy5*) and green (*Cy3*) channel on the same probe in both experiment, when measurement on that probe is 100% consistent, we should get $R_1/G_1=G_2/R_1$ and $log_2(R_1R_2/G_1G_2)=0$. For all the probes on the array, we calculate the statistics ($log_2(R_1R_2/G_1G_2)$), and obtain a distribution with mean and standard deviation (σ). For those probes with $log_2(R_1R_2/G_1G_2)$ exceeding 2 σ, were marked as outliers and discarded. It's important to use the dye-swap replicates to reduce random error introduced in the experiments.

2.2.2 Probe annotation

For oligo array, the 60 mer probe sequences were obtained from Agilent and mapped with BLAT (Blast-like Alignment Tools, Kent) method. Human genome (hg17) sequences were downloaded from UCSC website loaded into a local standalone BLAT server. The sequences were transformed into fasta format and aligned to genome with BLAT. The alignment results were filtered for each probe to specify its mapping position by the following criteria: 1) The perfect alignment is >55 mer. 2) The secondary hit should have an alignment less than 95% of the perfect alignment. For all 20K available sequences, 17Kb probes are mappable by the above standard. The genes the probes reside in were obtained based on the file seq_gene.md from NCBI.

2.3. Overview of Hidden Markov Model

Hidden Markov Model is a sophisticated but flexible probability model often used in time/space related pattern discovery [34]. A HMM can be visualized as a finite state machine, moving through a series of states and producing output either when the machine has reached a particular state or when it is moving from state to state. The HMM generates a state sequence by emitting certain observations as it progresses through a series of hidden states. HMM has been widely adapted to sequence analysis (gene identification, sequence alignment and protein structural prediction), [35,36] genetic linkage [37], LOH studies[38] and time-course microarray data [39]. For the above application, only time-course data analysis is based on continuous observations. For the rest, the observations are discrete. Since the array-CGH log2 ratio is continuous, here we only characterized a discrete time HMM on continuous observation by the following:

(1) **S**: the hidden states in the model. Typically, the states are interconnected in a way that any state can be reached from any other state. We denote the individual states as $S=S_1, .., S_K$ and the state at location t as s_t, $1 \le t \le T$, where T is the total length of the state sequence.

(2) The initial state distribution $\pi = \{\pi_k\}$, where $\pi_k = P\{s_1 = S_k\}$, $1 \le k \le K$,.

(3) The state transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = P\{ s_{t+1} = S_j \mid s_t = S_i \}, \quad 1 \le i, j \le K$$

(4) The emission distribution or probability density function $B = \{b_k(O)\}$ where $\{b_k(O)\} = G(O, \mu_k, U_k)$, $1 \le k \le K$. O is the vector being modeled. G is Gaussian density with mean vector μ_k and covariance matrix U_k: More generally, G is any log-concave or elliptically symmetric density and the probability density function $\{b_k(O)\}$ is a finite mixture.

Thus, an HMM with the fixed number of hidden states K can be characterized in terms of three parameters: (i) the initial state probabilities, π; (ii) the transition probability matrix, A; and (iii) the collection of Gaussian emission probability functions defined within each state, B: The parameters of the model may be represented in a compact way as $\lambda(A, B, \pi)$ and the sequence of values $O = (o_1, \ldots, o_T)$.

2.4. Defining HMM on array-CGH data

To infer distinctive copy numbers on the a chain of continuous log2 T/R ratios $O(o_1, \ldots, o_T)$ ordered by their chromosome position, the states and model parameter $\lambda(A, B, \pi)$ need to be clearly specified.

2.4.1 States

Since the goal of the method is to infer copy numbers, the state can be set to represent integers ranging from *0* to *N*. *N* is user defined maximum copy number as long as it clearly describes the change. The copy numbers do not have to be consecutive integers. As the copy number increases, the difference between log2 ratios of consecutive copy numbers will decrease. Ultimately the difference becomes so small that it becomes un-informative to distinguish the change in status or amplitude. For instance copy number 31 and 32. For a diploid sample, any copy number above *8* is clearly indicating an amplification, hence a practical set of integer states can be defined as *0*, .., *8,16,32*. In an ideal homogeneous population, this is enough to describe all the changes within a single profile. In a heterogeneous population, some changes are only carried by parts of the tumor cell population. The non-integer states like half and quarter states can also be added in to explain those changes. Though still discrete, the states are numerical yet not symbolic; this is markedly different from states in sequence or LOH study.

Within the integer copies, a ploidy copy is defined as the majority copy number carried by all probes within the profile. It usually agrees with the ploidy number defined by cytogenetic experiments. When the cytogenetically counted chromosomes are between 2 adjacent ploidy number such as 58<2n>, majority copy number might be 2 or 3 depending on where gains happened since the probes are not evenly distributed on each chromosome.

2.4.2 Emission probability and initial probability

Each copy number states emits log2 ratio signal according to certain statistical distribution. As we mentioned above, the overall data compression and shift of the log2 ratio signals of ploidy copy form zero are two major sample specific fixed effects to transform the observation from its empirical log2 ratio. Thus, the emission probability of state S_i can be specified as:

$$G(b_i * \log2(C_i/C_p) + \mu_p,\ \sigma_i)$$

The C_i is copy number represented by S_i, b_i is a sample specific compression scaling factor for state S_i, C_p is the ploidy copy number, μ_p is the signal mean of the ploidy copy, σ_i is the sample specific standard deviation for state S_i.

Once the mean of the emission probability (called "signal mean" and so forth) are defined for all copy numbers states, all the data points can be partitioned with them into K sections by normal approximation, where K is the number of the copy number states in the model. So the mutually exclusive signal means of those states resides in one of the sections and has different numbers of observations around it. A background probability p_i, defined as the proportion of the observations in O (o_1, .., o_T) around the signal mean for states S_i.

$$p_i = \frac{\sum_{t=1}^{T} (o_t \in Si)}{T} \quad \text{where } o_t \in Si \text{ if } (\frac{\mu_{i-1} + \mu_i}{2} \le o_t \le \frac{\mu_i + \mu_{i+1}}{2})$$

The initial probability π_i defines the likelihood of being in state S_i at the initial of the sequence o_1. It's reasonable to setup π_i as its corresponding background.

2.4.3 Transition probability

The transition probability describes the correlation between different copy number states of adjacent probes. Apparently, the probability of sharing the same copy number decreases as distance between two neighboring probes increases. Haldane's map function $\theta=0.5*(1-e^{-2d})$ has been traditionally used in linkage analysis to convert the genetic distance d between two markers to the probability that the second maker will have a meiotic cross-over events. Recently it has been adapted to estimate the transition probability in inferring LOH states in SNP array. Lin et al demonstrated that the empirical transition probabilities estimated from observed LOH calls agreed well with the properly scaled Haldane's map function defined as $\theta=1-e^{-2(d/100)}$, where d is physical distance in megabase scale. The relationship can be illustrated in figure 2.3. Biologically, genomic events causing copy number changes have some similar mechanism to cross-over or LOH, the haplotype state in LOH is one kind of copy number change in CGH. The incentive to use it here is to give a guide line on how to define the probability to reflect the uneven spacing of

the probes on the array used in this study and many others. For some regions especially for two consecutive probes separated by a large region such as centromere, it's reasonable to loosen the restriction on free transition between different copy numbers. Meanwhile transition to a certain state should also be related to the background probability of the copy number of that state. The higher the background probability, the likelier it is that the state sequence will transfer to the next state. Thus, the transition probability from state i to j at position t can be formally defined as:

$$a_{ij} = p_j * \theta \qquad for \quad i \neq j$$
$$a_{ij} = p_j * \theta + (1 - \theta) \qquad for \quad i = j$$

p_j is the background probability at position $t+1$. a_{ij} only depends on whether the i, j are the same and the background probability of p_j, thus it's independent of the state at position t. For array with median resolution of 50kb and θ of 0.001, the relationship is naturally favoring the copy number sequence to continue in the same states for most of the probes at most of the positions.

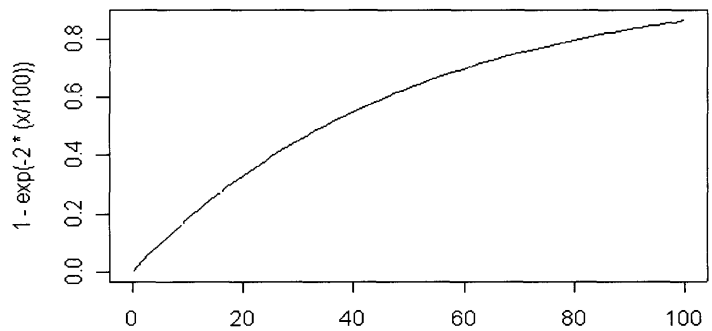

Figure 2.3. Scaled Haldane's map function: θ=1-$e^{-2(d/100)}$. X-axis is the distance in mega base pair unit. Y axis is θ.

2.4.4 Estimating the emission probability of the model

It has been shown that for HMM with continuous distribution observations, good estimation for emission probability is essential for making a correct inference, while a slight loosely defined initial probability and transition probability have no essential impact on inference results [34]. Given both the initial and transition probabilities have already been relatively well defined; the focus on inference is to fill the open slots in the emission probability definition. The best way to specify those parameters is to obtain them from a good training dataset, constructed by a set of clean and well-annotated samples in which the distribution of difference copy numbers can be studied after assigning copy number to clone/probes with a less labor intensive manner.

2.4.5 Constructing training datasets

To select samples well suited for observing the data distribution, 4 criterions were set for the task:

a) Cleanness: the derivative standard deviation <0.3.
b) High homogeneous: the agreement of the SKY and array-CGH profile in terms of relative change and copy number > 10 chromosome. SKY images are generated from a few cells and may represent subclonal population, while the genomic DNA is typically obtained from whole tumor, reflecting the average change across an entire population. But if the population is largely homogeneous, the SKY should have a good level of agreement with array-CGH relative copy number.
c) Range of observable copy number at SKY resolution level is greater than or equal to 4. Since sample specific distributions are expected, it's important to have more available copy numbers in a single profile to study the relationship among the signal distributions of wide range of copy numbers.
d) Medium complexity, defined as average copy number transitions, excluding focal changes within single chromosome that are less than 4. It's easier to assign copy numbers to an entire chromosome or arm under single level of changes than complex patterns.

Finally 17 samples were included in the training dataset: 5 pancreatic cell lines, 12 multiple

myeloma primary tumors. CBS were applied to all the samples to get a guide line of boundaries of the copy number changes. The plots of raw log2 ratio overlaid with uniform relative copy number (log2 ratio median of each segment) are generated for all 18 samples with single chromosomes. Based on SKY reports, only integer copy numbers were assigned to 16004 oligo probes on 22 autosomes with careful visual inspection. Probes on entire intact chromosomes and arms that matched with SKY were first identified to be assigned the copy number. The segments left were compared to segments assigned in terms of their median value and assigned a copy number with closest median. If the median is >0.02 from any of other assigned segments, the segment will be considered as carrying a non-integer copy number segment and will be excluded. Focal changes, usually with less than 50 probes, that were not observed in SKY are also excluded. Loci with complicated pattern (transitions >5) are also excluded from the training dataset.

2.4.6 Sample-wise signal distribution of all copy numbers in training dataset

For each profile, the count, mean, median, mode and standard deviation are calculated for all assigned integer copy numbers, referred as the "actual" or "nominal" mean and standard deviation.

2.4.6.1. Probes distributed on different copy numbers

First, the ploidy copy is the most dominating copy number for all probes across all samples except TU8902 (figure 2.4, upper right panel). As the only tetraploid sample with karyotype of 86(81-90)<4n> having gain/loss in many chromosomes, relative to its ploidy copy of 4, the count of the probes on ploidy copy was actually little lower than its one copy gain and loss. After the ploidy copy, one copy gain and one copy loss are next 2 major copy numbers within all gain/loss copy number across the genome(figure 2.4), suggesting that the most dominated genomic events are one copy loss or gains. This matches what has been observed with SKY.

2.4.6.2. Spread and density of the log2 ratio signals in different copy numbers

The log2 ratio signals for all copy number are symmetrically distributed in the density plot (figure 2.4, upper left panel), especially for ploidy copy number. As the nominal signal

mean deviates from 0, the spread of the signals become wider and tends to have a longer tail towards the log2 ratio 0. Overall, the spread suggests a t distribution should be appropriate to describe the data, and for copy numbers with signals off from zero should have lower degree of freedom as more outliers are present in those copy number.

2.4.6.3. .Mean of the log2 ratio signals in different copy numbers

5 samples have the mean log2 ratio of their plodiy copy within -0.05~0.05 (table 1), 3 samples are off from 0 with greater than 0.15, the rest 9 are between 0.05~0.15. In all the samples, the nominal log2 ratio mean of all copy numbers are in a linear relationship with their empirical log2 ratio mean (figure 2.4 lower left panel). When fitted with a linear regression model, slope of the regression line reflecting the data compression levels, are in a range of 0.4~0.7(table 1).

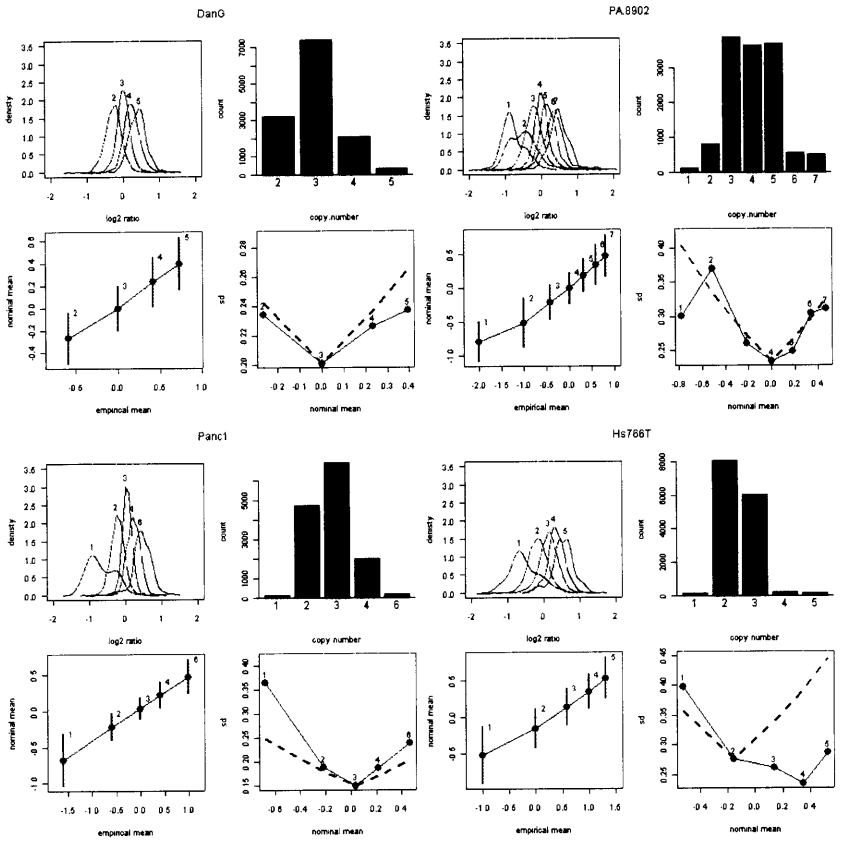

Figure 2.4. Examples of Log2 ratio signal distribution of different copy numbers across training samples. Each sample includes 4 graphs, the red line (or bar, points) marks the ploidy copy

Upper left: The density plot of the signals of all copy numbers which marked on top on their signal density line.

Upper right: Count of probes measuring the specific copy number.

Lower left: The log2 ratio signal mean and standard deviation of each copy number. The X-axis is the empirical log2 ratio of each copy number. The Y-axis is the nominal log2 ratio. The bar marks the one standard deviation.

Lower right: The variance of the signal of all copy numbers. The X-axis is the nominal mean of the signal, Y-axis is the nominal standard deviation of the signal. The black dash line is the curve fit by sj = 2(abs(mj)-abs(mp))*sp, where mj and sj is the mean and standard deviation of the signals of the copy number j, mp and sp is the mean and standard deviation of the signals of the ploidy copy p

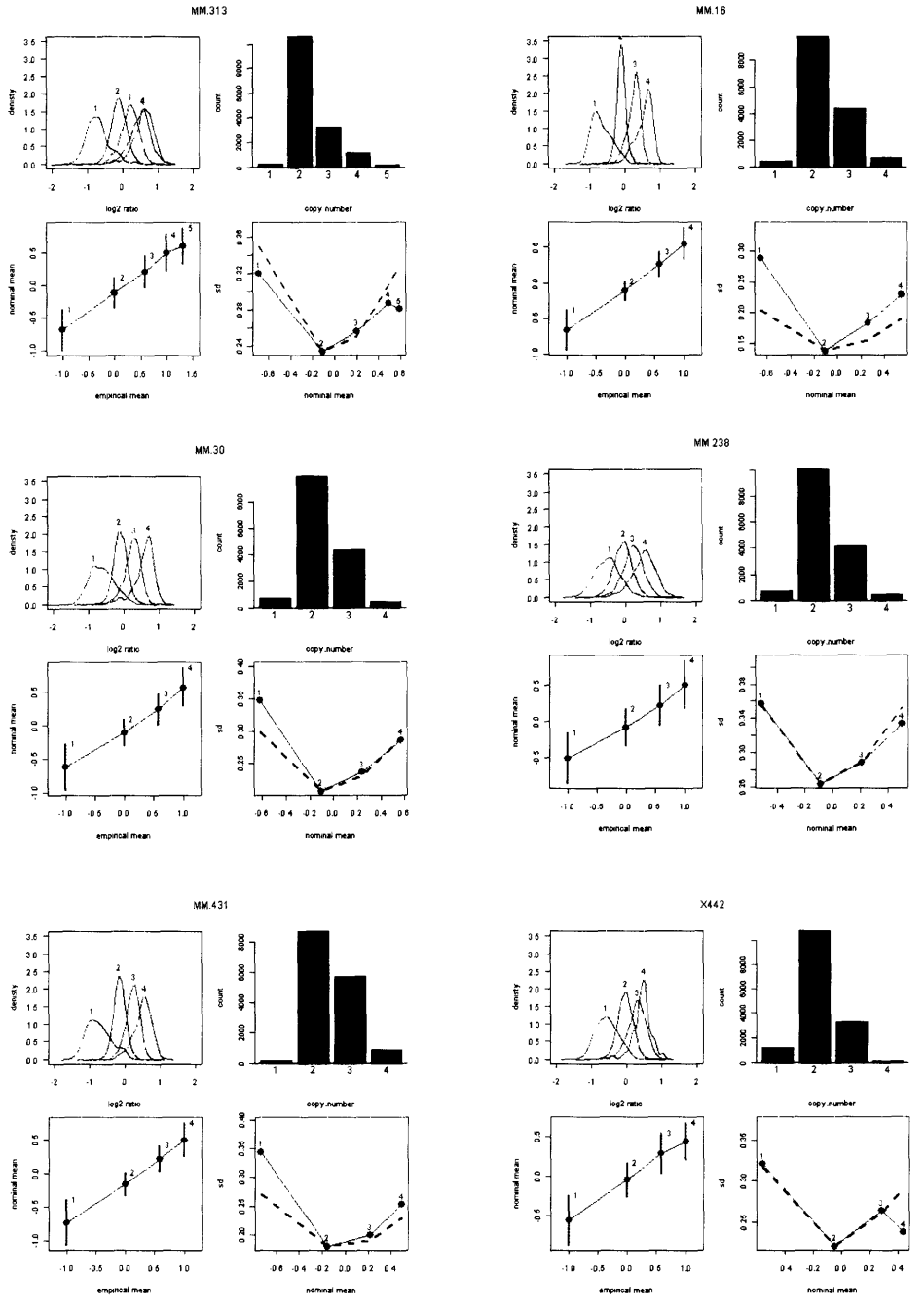

Figure 2.4 Examples of Log2 ratio signal distribution of different copy numbers across training samples.(Continued)

Table 1. The signal mean and compression scaling parameters of all training samples

Sample Name	signal mean of ploidy copy	Slope	P value of regression model
X1002	-0.11212	0.517348	0.001744
X442	-0.0487	0.507664	0.001504
X443	-0.10582	0.45616	0.002201
X454	-0.08789	0.591258	0.00091
MM.13	-0.10855	0.501518	0.004439
MM.16	-0.10406	0.59788	0.001319
MM.185	-0.17916	0.485883	0.005676
MM.238	-0.0862	0.498962	0.005601
MM.30	-0.10119	0.584003	0.003994
MM.313	-0.11167	0.565215	7.56E-05
MM.35	-0.07945	0.403268	0.002985
MM.389	-0.15678	0.495348	0.000308
DanG	0.001457	0.504379	0.000949
Hs766T	-0.1551	0.454197	0.000316
HUP.T4	0.199933	0.626521	0.000481
PA.8902	-0.00111	0.456991	2.75E-05
Panc1	0.038737	0.442532	1.69E-06
SU86.86	-0.03189	0.527175	6.08E-06

The signal mean of ploidy copy of all training samples and the compression scaling parameter obtained by fitting a regression model of the nominal signal mean to empirical signal mean. The third column is the p value of the fitting results.

2.4.6.4. Variance of the signals in different copy numbers

The standard deviation of the log2 ratio signals for most copy numbers increases with their absolute nominal signal mean but was harder to capture in a uniform linear formula (figure 2.4, lower right). One possible relationship could be used to explain this is

$$s_j = 2^{(abs(mj)-abs(mp))*sp},$$

where m_j and s_j are the mean and standard deviation of the signals of the copy number j, m_p and s_p is the mean and standard deviation of the signals of the ploidy copy p (black line in figure). It fits better in samples with lower noise in terms of the standard deviation of ploidy copy, but the standard deviation of one copy loss changes might be underestimated. Most of the copy number *sd* fits reasonably well; 2 samples appear to deviate from this relationship: Hs766T and HUPT-4 in pancreatic set.

Given the observed log2 ratio distribution described above and the fact that one copy gain/loss are the most frequently occurring changes across the genome, a sample specific loss/gain compression scaling parameter $a_{gain/loss}$ can be obtained from the nominal distance between the signal mean of one copy gain/loss with the ploidy copy number. In other words, the estimation of emission probability can actually be reduced down to the estimation of signal mean and standard deviation of ploidy copy and mean only for one copy gain and loss.

2.4.7 Applying HMM on training dataset

The compression scaling parameter *a* was calculated for all the training samples and HMM with integer copy number states 0..8, 16, 32 was setup based upon it. The assumed signal distribution of copy number 0 is calculated as ¼ copy. The standard deviation setup was based on the formula described above as $s_j = 2^{(abs(mj)-abs(mp))} * s_p$. For each sample, the model was fitted per chromosome each time, the posterior distribution of log2 ratio signals for each inferred copy numbers from all 22 autosomes were calculated after processing. The percentage of agreement of inferred copy number with assigned copy number was used as a measurement of the inference accuracy. Almost as expected, the average agreement is 99.8% for all 19 samples for those data points with assigned copy number. Meanwhile, the model is given a series of compression scaling parameter from 0.3~1.2 combined with offset range of -0.2 ~0.2 from signal mean of ploidy copy to see how the output results would deviate from the assigned copy numbers when mean of the emission probability specification deviates from the "real mean". Comparing accuracy for all inference results across all the combination of the 2 parameters, the closer the parameter to the original scale, the more accurate the result is. Figure 2.5 shows the results of inference accuracy on Panc1 for such procedure. The actual compression scaling parameter of Panc1 is 0.42, the figure clearly shows the best accuracy was with scaling parameter set as 0.4~0.5 and shifts from the true signal mean within 0.05. The results of the other samples showed similar results.

This observation indicates that the posterior mean of inferred ploidy copy, one copy gain and loss estimated from entire genome are uniformly close to their nominal distribution (figure2.6), given the initial scaling parameters, offsets from the ploidy copy signal mean, and ploidy copy signal standard deviation within a reasonable range (0.5~0.7 for scale and -0.05~0.05 offset). Since the emission probability specification agrees with the real data is essential in achieving the best inference, it suggests that we can use this re-estimated emission probability to arrive at a final model and re-fit this improved final model to make an accurate inference. In the case where there is no large segment of one copy gain or loss in the profile, this can be used as the reference. A default scaling parameter or most common ones can be use as a substitute. By this way, the parameter to specify the copy number mean in initial model has been largely simplified and reduced.

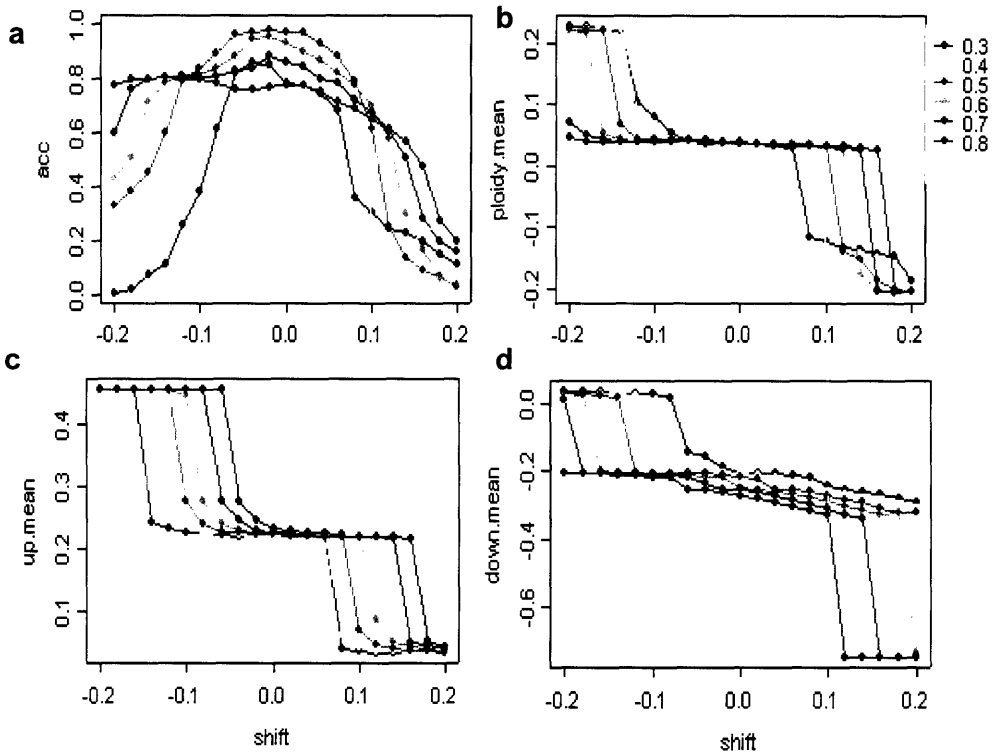

Figure 2.5 Accuracy of the inference results with posterior mean of ploidy copy, one copy gain/loss under different initial shift and compression scaling parameters.

X-axis of all panels represents the shift from estimated signal mean. Different color of lines in each figure represents different scaling parameters as in legend. The Y-axes are:

a) The accuracy of the inferred copy number.

b) The posterior signal mean of probes inferred as the ploidy copy.

c) The posterior signal mean of probes inferred as one copy up regarding the ploidy copy.

d) The posterior signal mean of probes inferred as one copy down regarding the ploidy copy.

2.4.8 Determine the log2 ratio signal distribution of ploidy copy number

The ploidy copy is the majority of the copy number in the genome, however, its log 2 ratio signal mean can be off from 0 as we showed above. Since array-CGH log2 ratio is spatially correlated, median filtering was applied to the data with a medium size window (k=21) to reduce the noise. Comparing mode of median filtered data with window size 21 on all samples, they matched quite well (figure 2.6a) with the signal mean of their ploidy copy with correlation 0.99 and slope of 1. This suggests that the mode of the median filtered data could be a reliable way to estimate the mean of ploidy copy number. But for the standard deviation of the log2 ratio signal of the ploidy copy, the profile noise and the signals from the real changes are confounding. We tried to estimate it from the part of the data which covers the 50% quartile around the log2 ratio mean of the ploidy copy. Figure 2.6b plotted the ploidy copy standard deviation against the standard deviation obtained from middle 50% quartile of the raw data. The linear relationship was fitted with a regression mode with R-square of 0.94, indicating a reasonable estimation for initial model setup. And this could be further improved by using the posterior standard deviation based on the signals inferred as the ploidy copy. Figure 2.6c is the actual standard deviation of the ploidy copy number signals against the standard deviation of the signals from inferred ploidy copy generated from the model with a default scaling parameter 0.7 and initial ploidy copy number mean and standard deviation estimated as above. The R-square is almost 1.

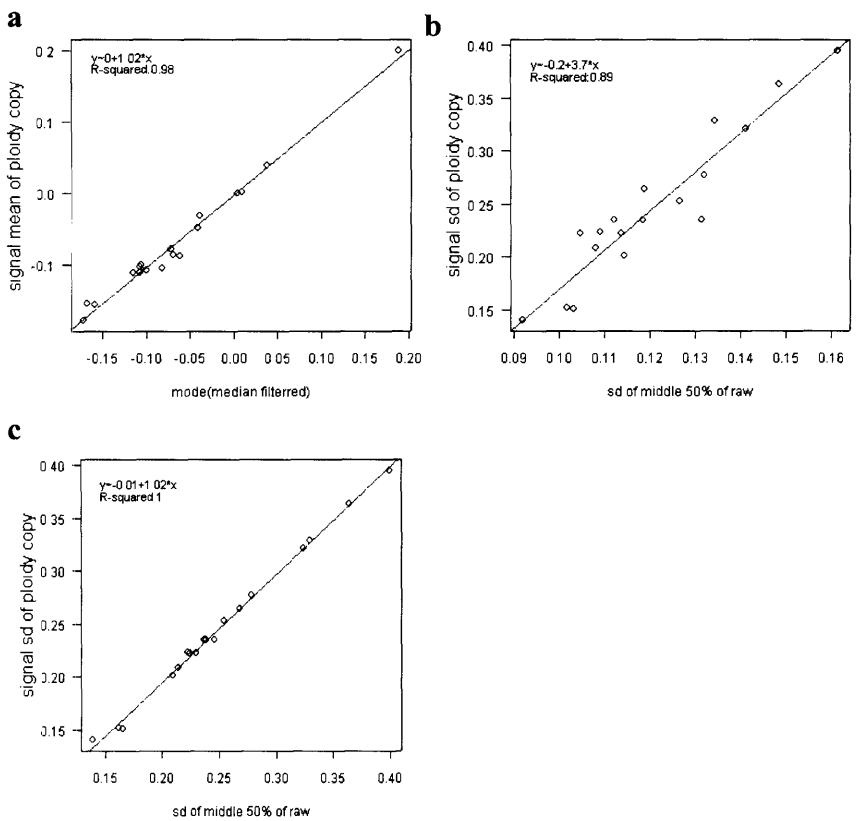

Figure 2.6. Estimation of the mean and standard deviation of the ploidy copy.

a) The estimate of the mean of plodiy copy. The X-axis is the mode of the median filtered data (window size 21).Y-axis is the actual mean of ploidy copy.

b) The estimation of the mean of the ploidy copy from the posterior mean

c) Estimate the standard deviation of the ploidy copy.

2.4.9 Adding Fractional Copy Number Status and its Model Selection

As array-CGH is detecting changes in entire population, there's always a possibility of heterogeneity in many changes resulting in non-integer copy numbers. Figure 2.7a shows a genome profile of a diploid sample on CL7. Most of the chromosome are intact 2 copies with mean of log2 ratio 0. Chromosome 7q, 9q, 15q has one copy gain with log2 ratio of 0.44~0.45. Chromosome 19q has 1 copy loss with log2 ratio -0.67. This gives a data compression of around 0.79 for gain and 0.67 for loss. The loss of 17q and 18p carry a log2 ratio of around -0.35. With only integer states, the changes on both the regions were missed.

With ½ copy added, 18q and 19p will both inferred as 1.5 copies. The sum of squares, measured as the square of distance between the raw log2 ratios to the median of inferred states, of both chromosomes reduced 70~80%. This clearly indicates a much better fit while the inference of rest chromosome remains about the same value. If we compare the joint maximum likelihood, both chromosomes gained 30% in log10 scale. Table 2 gives the statistics for all the chromosomes in this sample.

Table 2. Fractional copy improves the explanation of the intermediate states in BAC sample CL7.

chr	p.star	p.star (1/2)	Δp.star	Δp.star /p.star	SS	SS(1/2)	ΔSS	ΔSS/SS
1	75.170	75.286	-0.115	-0.002	0.812	0.812	0.000	0.000
2	111.941	112.050	-0.109	-0.001	1.485	1.485	0.000	0.000
3	54.849	54.995	-0.146	-0.003	0.611	0.611	0.000	0.000
4	111.171	111.269	-0.098	-0.001	1.505	1.505	0.000	0.000
5	73.362	73.456	-0.094	-0.001	1.263	1.263	0.000	0.000
6	50.948	51.093	-0.144	-0.003	0.708	0.708	0.000	0.000
7	91.540	91.588	-0.049	-0.001	0.871	0.871	0.000	0.000
8	89.654	89.713	-0.058	-0.001	1.050	1.050	0.000	0.000
9	82.296	82.355	-0.059	-0.001	1.035	1.035	0.000	0.000
10	76.902	77.001	-0.100	-0.001	0.995	0.995	0.000	0.000
11	112.492	112.572	-0.080	-0.001	1.522	1.522	0.000	0.000
12	56.579	56.720	-0.141	-0.002	0.765	0.765	0.000	0.000
13	31.329	31.446	-0.117	-0.004	0.422	0.422	0.000	0.000
14	72.776	69.879	2.897	0.040	1.429	1.243	0.186	0.130
15	42.872	42.912	-0.039	-0.001	0.737	0.737	0.000	0.000
16	46.573	46.638	-0.064	-0.001	0.808	0.808	0.000	0.000
17	45.876	45.952	-0.076	-0.002	0.566	0.566	0.000	0.000
18	52.276	36.338	15.938	0.305	1.638	0.485	1.153	0.704
19	41.223	27.162	14.061	0.341	1.896	0.334	1.562	0.824
20	59.520	59.569	-0.049	-0.001	0.856	0.856	0.000	0.000
21	32.431	32.519	-0.088	-0.003	0.904	0.904	0.000	0.000
22	21.862	22.019	-0.156	-0.007	0.816	0.816	0.000	0.000
X	64.445	64.688	-0.243	-0.004	1.374	1.374	0.000	0.000
Y	12.694	12.916	-0.222	-0.017	0.300	0.300	0.000	0.000

The improvement in p.star (joint likelihood of emission and transition probability) and sum of squares between the raw value and the emission mean of state inferred. Chromosme 17q and 18p, which by visual inspection needs a intermediate state between 1 and 2, inferred as

1.5 copy, their p.star improvement was more than 30% and 70% of sum of squares, while except for chromosome 14, others all remain about the same

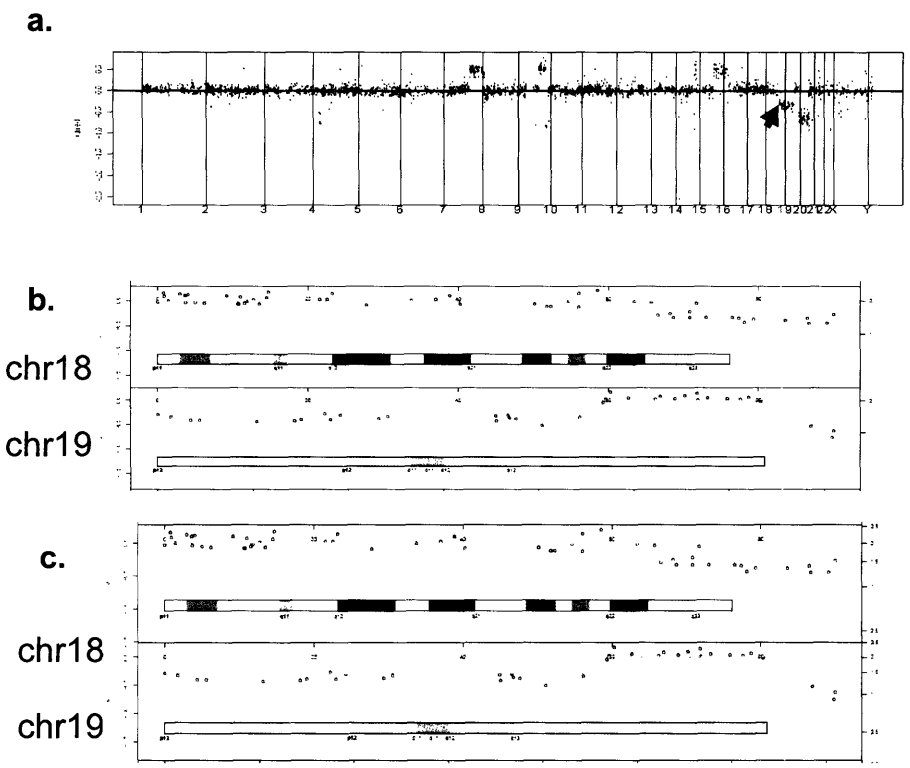

Figure 2.7. The fractional states setup improves fitting statistics.

a) The genome wide view of CL7 in Albertson dataset. The probes are plotted along the chromosome order with their physical position. Chromosome 20 shows one copy loss. 7q, part of 9 and 14, 15 are 1 copy gain. End of 18q and start of 19p has log2 ratio between one copy loss and no change.

b) HMM inference with integer states only. End of18q is inferred as partial no change and 1 copy loss. 19p was inferred as 1 copy loss.

c) HMM inference with half states. Both end 18q and start of 19q are inferred as 1.5 copies.

One concern over adding fractional copy numbers is that it might trigger spurious transitions between close neighboring copy number states caused by local data trend. Reducing the transition probabilities down by scaling on Haldane's map function might help, but it also restricts the proper transition to certain focal change, which is often important biologically. So it's a necessity to make a model selection between integer copy models and fractional copy number models. The fitting statistics for selection can be either sum of squares or p.star, the joint maximum likelihood of emission and transition probabilities calculated from Viterbi algorithm or simply, the number of transitions. p.star automatically punishes the transition between the states while measuring the good of fitness. To check how this scheme works on other samples, we picked 12 chromosomes with non-integer copy numbers from 9 multiple myeloma samples, whose copy number inference is alternating between 2 integer copies or missed (table 3). After fitting with a model with fractional copy status, all 15 chromosomes picked up a better fit with p.star improvement over 3%. While for the rest of 193 chromosomes their p.star improvements usually were less than 1% or negative. This suggests a 3% improvement on p.star could be a reasonable cutoff for model selection, but it's better for it to be leaved as an option for adjustment in real practice.

Table 3. Improvement of fitting statistics on selected chromosomes

sample	chr	Δp.star/p.star	ΔSS/SS	# probes changed in inference results	p.star change per count	Sum of squares on per probes
X449	2	0.099	0.319	1040	0.065	0.048
X449	5	0.080	0.299	574	0.062	0.048
X449	17	0.069	0.275	930	0.044	0.038
MM.10	9	0.090	0.317	460	0.077	0.030
MM.10	22	0.085	0.156	272	0.078	0.015
MM.341	6	0.115	0.429	363	0.150	0.043
MM.341	10	0.070	0.374	604	0.040	0.015
MM.341	2	0.080	0.093	577	0.082	0.006
MM.341	15	0.064	0.183	471	0.033	0.005
MM.342	4	0.059	0.240	241	0.084	0.027
MM.35	22	0.036	0.391	361	0.021	0.077
MM.386	19	0.041	0.223	1005	0.027	0.019
MM.389	8	0.047	0.275	213	0.075	0.056
MM.393	8	0.076	0.184	498	0.050	0.013

The improvement of fitting statistics: p.star (the joint likelihood of emission and transition) and sum of squares on selected chromosomes after fractional copy states has been added.

2.4.10 Assumption and Algorithm

The algorithm will take normalized array-CGH log2 ratio:

1. Estimate mean and standard deviation of the ploidy copy number:
   Median filter data with window size k (usually 1/40 of the total probe size).
   Calculate the mode as the estimate for mean of the pliody copy.
2. Setup a HMM with copy number states $0...K$, with a default scaling parameter (usually 0.7) to fit the model on a single chromosome each time.
3. Use the 50% quartile of the data around estimated signal mean of the ploidy copy number and pre-trained regression estimation parameter to estimate the standard deviation of the ploidy copy.
4. Summarize signal distribution of ploidy copy number, and one copy up and down. Only data from segments with greater than 50 probes are used.
5. Update the mean and standard deviation of the ploidy copy number .Calculate the sample and gain/loss specific compression scale $a_{gain/loss}$. If one of such change (gain or loss) is not available, use the estimation from the other change. If both type of change not present, use default scaling parameter.
6. Redefine the emission probability based on the new estimation and re-fit the model.

The assumption for samples to infer absolute copy numbers are:

1. Known ploidy and it is homogenous at ploidy level,
2. Heterogeneous changes are only small portions in the entire genome.
3. One copy gain and loss constitutes majority of the change in all gain/loss events.

For samples without known ploidy copy number, usually diploid is assumed. In this case, only relative nominal log2 ratio is recommended to use for results.

2.5. Validation on Independent Samples

Independent samples were tested on the method including 50 primary multiple myeloma and 31 recurrent giloma samples as along with the remaining 4 pancreatic cell lines. The states were setup as in the training set but fractional copy numbers were added between lower copy number states as: 1/8, ¼, ½, 1, 1½, 2, 2½, 3, .., 8, 16, 32. The initial scaling parameter was set to 0.7 and the regression formula used to estimate the initial ploidy variance was setup as $-0.2+3.7*s_p$ as obtained from the training dataset.

Table 4. HMM inferred copy numbers compare to the real copy number observed in SKY.

True copy number	No. of chromosomes arms	Inferred copy number							
		1	2	3	4	5	6	7	other
1	23	23							
2	12		12						
3	45			45					
4	22				20	2			
5	3					3			
6	4						2	2	
total	109								

109 chromosomes are randomly selected to validate the result of HMM inference. Among all chromosomes only 4 was inferred as one copy number more than its actual SKY observation. But this does not change the result of gain/loss call.

2.5.1 Copy number inference on genome scale at low change level

Partial or entire chromosomal gain and loss are important genomic structure aberrations and may lead to pattern discovery if it is recurrent and associated with certain clinical phenotype. The copy number inference results on these large scale changes are compared with SKY data. 109 representative entire/partial chromosomes or arm length of copy number ranging from 1 to 7 were picked from 50 samples (4 pancreatic cell lines, 15 primary multiple myelomas) where array-CGH and SKY agrees closely. The method

predicts the copy number with 98% accuracy (table 2). For the 3 over estimated copy number, the aberration status does not change. Figure 2.8 shows the detailed comparison of psudo-karyotype ideogram (2.8b) generated with inferred copy number of AsPC-1 comparing with its SKY image (2.8c). The karyotype of AsPC-1 is 53<2n>, the relative change in array-CGH log2 ratio (2.8a) match with all the 24 chromosomes (2.8c). In this example, the inferred copy number shows exact match in entire/partial chromosome copy number change with SKY image side by side.

2.5.2 Focal change detection

Other than the large, high-confidence regions, the method has detected many focal changes in both training/testing dataset. Real-time quantitative PCR (qPCR) validated loci in the pancreatic cell line data set and has been collected and compared to the inference results (table5). The size of the loci in terms of number of probes ranges from >20 probes down to single probe. The method has detected all the changes presented from the raw data and inferred relative copy numbers were closer to qPCR results compared to raw log2 ratio. Figure 2.9 shows a few such examples in better detail: a single probe deletion on CDKN2A on chromosome 9p (Panc1), 2 probe EGFR amplicon on chromosome 7 (3R in RG), two homozygous deletion in BxPC.

Table 5. Focal changes by HMM in pancreatic dataset.

			Relative Gene Copy			
CytoBand	**Gene**	**Sample**	**qPCR**	**raw**	**HMM**	**Change**
19q13.2	PD2	Panc1	15	3.06165	16	amplification
7q22.1	TRAAP	ASPC1	14.23	9.902337	16	amplification
19q13.2	CLC	Panc1	13	9.747699	16	amplification
12p12.3	CPAZA3	DanG	12	8.58825	16	amplification
12p12.3	CPAZA3	TU8902	6.59	3.226567	8	amplification
12p12.3	CPAZA3	TU8902	4.36	3.646595	4	gain
12p11.21	MGC24039	TU8902	3.2	2.709247	4	gain
9p21.1	CDKN2A	ASPC1	2.23	0.586794	1.5	gain
7q21.12	CROT	ASPC1	2.145	0.938917	2	gain
9p21.3	CDKN2A	HPAC	0.56	0.763244	0.67	loss
9p21.3	CDKN2A	ASPC1	0.315	0.586794	0.5	loss
8p21.3	LZTS1	DanG	0.26	0.601449	0.33	loss
8p21.3	LZTS1	Panc1	0.258	0.644477	0.67	loss
9p21.3	CDKN2A	Panc1	0	0.362128	0	deletion

Comparison of relative gene copy obtained with real-time qPCR with the raw log2 ratio versus HMM inferred copy in pancreatic cancer cell lines.

2.5.3 Copy number inference on samples with unknown ploidy copy number

Of most samples used in this study, SKY results are available to provide the ploidy information. In RG dataset, the ploidy copies are unknown since no chromosome banding experiment information is available. Since diploid is often dominating in most of the primary tumors, we assumed all the samples are diploid and ran the HMM model with copy number states as we setup in the training dataset. Half copy states between copy numbers 0 and 3 (½ ,1½, 2½) are added to explain the possible heterogeneity in those samples. Instead of directly using inferred copy number, the median of the log2 ratio estimated from each segments are used as results to provide a relative quantification for the change. The change points in the data are identified at the position where transitions between different copy number states occurred. We compared all the position of change points with the CBS output, found the called change points are very similar for entire dataset. The major difference between the two is that CBS does not pick single probe change. When single probe segments are merged to the 2 neighboring segments if they agree, the two methods have agreement on most of the transition points. The disagreement often happens when distance related transition is involved like 2 probes across a region greater than 10Mb such as centromere region. This method will infer the transition at centromere and CBS often breaks the data one or two probe off it (figure 2.10a). Other type of agreement is often caused by local data trend and outliers (figure 2.10a and 2.10b). In focal change detection, our method shows better sensitivity compared to CBS, especially on loci with larger variance (figure 2.10b). Of the entire 35 profiles including 2 duplicates, EGFR amplicon is present in17 profiles, including 5 single probe amplifications. Along with the single probe amplicons, CBS also failed to identify 3 out of 5 two probe EGFR amplicons. Though the amplitude of the amplicons are very high (average log2 ratio >3), the variance in the 3 missed amplicons are roughly twice than those 2 that were not missed (0.7 vs. 0.4). Overall, under the circumstance where no ploidy information is available, the method still detects meaningful copy number transitions as well as focal change.

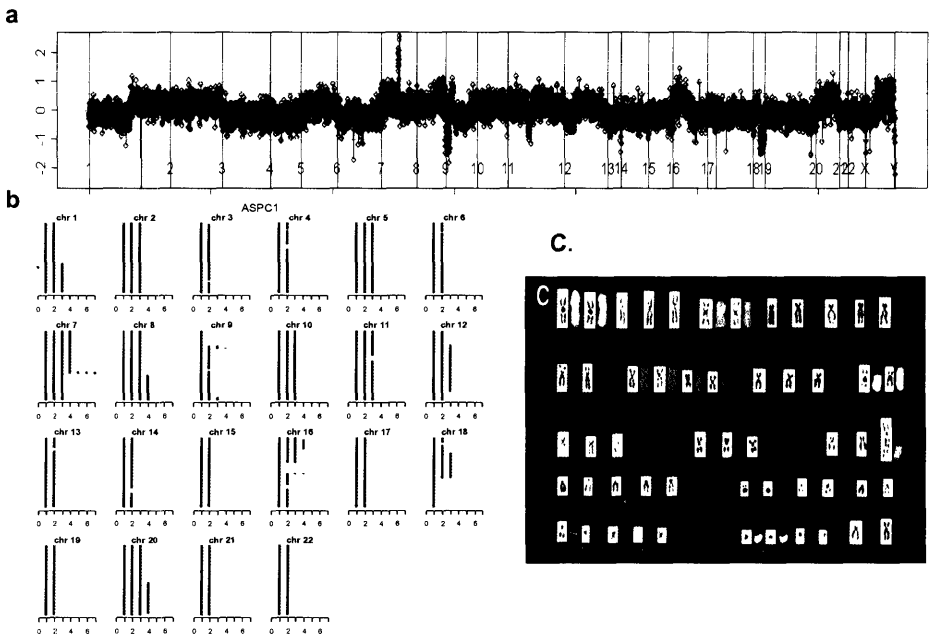

Figure 2.8. Example of HMM inference results of the pancreatic cell lines of AsPC-1

a) Genome wide view of the profile AsPC-1

b) The inference results for sample AsPC-1 in psudo-karyotype form. The x-axis is number of chromosomes

c) Published the SKY image for Aspc-1[40]

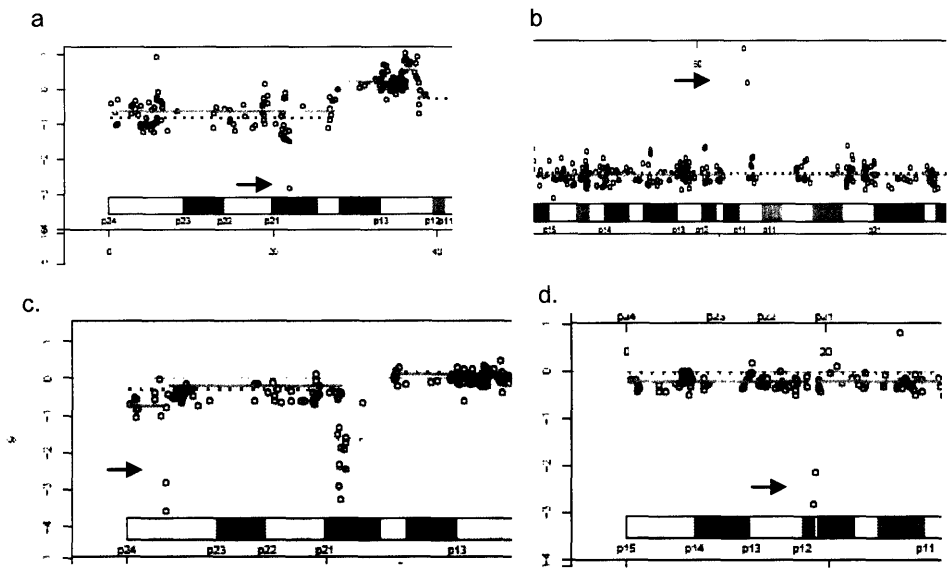

Figure 2.9. Examples of real-time PCR validated focal change identified by HMM. The navy dots are raw log2 ratio. The orange segments are inferred copy number draw at corresponding copy number states. X-axis is base pair position along a single chromosome.

a. Single probe showing homozygous deletion of CDKN2A on chromosome 9p in Panc1. Inferred copy number is 0.
b. 2 probe EGFR amplicon in 3R (RG dataset), inferred copy number is 32.
c. 2 probe homozygous deletions. Inferred copy number is 0.

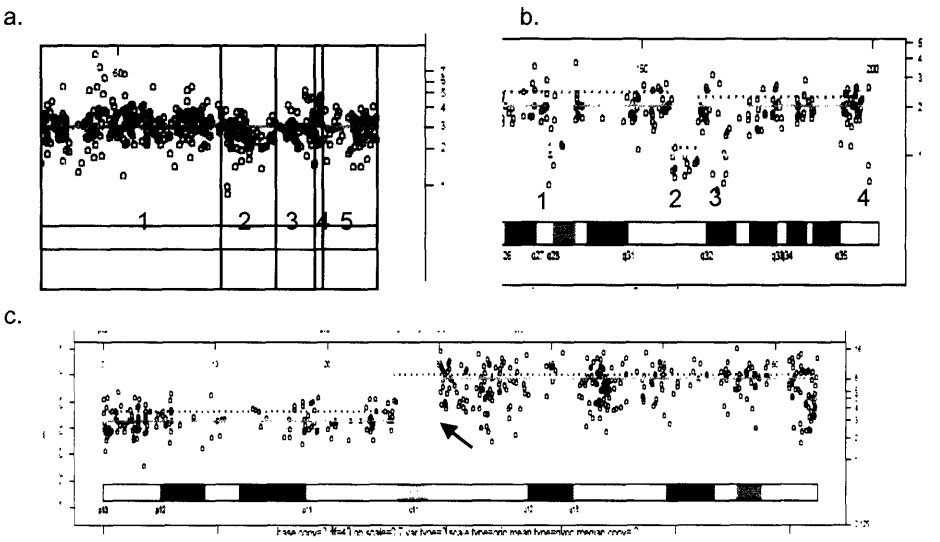

Figure 2.10. Comparison of CBS and HMM in RG sets.

a. Plot of chromosome 20 of 3F. The raw data is plotted along the physical position as navy circles. The HMM inferred copy number segments are in orange line overlay on top of the raw data. The CBS output is in red dotted line. The 20q has low-level amplification of entire arm. HMM identify the transition at centramere. CBS identify the change including centramere (red arrow 1). HMM has the change points at 20q end (red arrow2, probably due to the data trend).

b. Plot of chromosome in of... 5 segments are identified by CBS because of local data trend.

c. The EGFR amplicon in RG dataset. Of 33 profiles, 17 has EGFR amplicon including 5 only single probe amplification (black arrow). 3 of the two probe amplicon (blue arrow) are missed by CBS.

3. Discussion and perspective

This thesis has developed a novel method for analysis of array-CGH data based on Hidden Markov Modeling that enables automatic assessment of complex datasets for detection of structural aberrations in tumor genomes. By transforming the standard array-CGH output of log2 fluorescence ratios into absolute copy number values, this approach makes possible the meaningful and straightforward biological interpretation of array-CGH data. In contrast to other established methods of array-CGH analysis (e.g. segmentation, CLAC, etc), HMM overcomes two fundamental problems of copy number analysis: 1) Establishment of a baseline copy number that faithfully reflects the ploidy of each independent sample analyzed and 2) Elimination of the data compression seen on a sample-specific basis. When this HMM-based method of analysis was applied to multiple independent datasets/samples, it effectively predicted copy number values that showed strong concordance with results confirmed by other techniques such as SKY and qPCR.

The goal of aCGH analysis is to define deviations from baseline genomic copy number (denoted as "0" on an aCGH profile) within a sample. This baseline copy number should reflect the ploidy of the sample, which is assumed to be diploid. However, for a large proportion of samples, the baseline ploidy is not diploid; thus, a shift of the baseline to account for such deviation is essential for identifying changes in those samples under a uniform hard threshold. To establish the baseline copy number for each individual sample, the distribution of log2 fluorescence values for all probes is examined and the mode value is determined, regardless of whether the actual ploidy copy number is known. This mode value is assigned as the baseline copy number. However, when supporting data (banding experiment such as SKY) regarding the sample ploidy is available, this baseline copy number can now be given its true copy number value.

Failure to reset the proper aCGH baseline can result in misinterpretation of gain and loss within a sample. In the example of MM.11 (figure 5c), as a diploid sample, the signal mean of ploidy copy (i.e. 2 copies) is -0.11 and 3 copies (one copy gain) is around 0.14. The gains on chromosomes 3, 9, 11, 15 and so on could be miscalled as insignificant changes if we simply adhere to the widely accepted threshold of +/-0.2 to mark gain and

loss within an aCGH sample. However, resetting the ploidy copy number as 0 results in one copy gain being designated as 0.24 and thus correctly falls beyond the threshold for classifying it as a gain of copy number. There are many similar examples within in the entire myeloma dataset and proper resetting of the baseline copy number in these samples is critical for recurrence and pattern analysis.

By inferring absolute copy number change, HMM eliminates the data compression issue present on a sample specific basis. Identical copy number levels are transformed to an identical y-axis value and thus are better representative of their ground truth. The uniform hard threshold can be applied to categorize changes without affecting those highly compressed samples (such as primary tumors). For categorical data analysis, this will improve the accuracy at the probe level, and reduce the false positive and negative rate especially for multiple testing and sample/group comparison with smaller sample size.

To infer absolute copy numbers for each sample, HMM analysis ideally requires two basic parameters: 1) that each sample consists of a homogeneous composition of tumor cells and with heterogeneous changes comprising only small proportions of the entire genome and 2) a knowledge of the ploidy level for each individual sample. Since ploidy information is not always available, this is certainly one significant limitation to usage of a HMM-based approach. Furthermore, HMM may not be appropriate for extremely heterogeneous samples when the one copy gain and loss might not be the most dominating changes. That said, for samples with unknown ploidy but relatively homogeneous genomic content, one may make the assumption of diploidy. Under this assumption, HMM identifies the change points which agree well with those identified by other methods of analysis such as CBS. In such cases of unknown ploidy, it is not possible for this HMM-based approach to report true copy number values for each probe. Instead, we utilize HMM as a method of identifying change points and report the baseline-adjusted log2 ratio of T/N fluorescence values for each probe on the array. While HMM cannot determine the data compression level in the absence of knowledge of the ploidy copy number, assessment of one copy deviation from the mode copy number within each individual sample enables effective determination of copy number change on a sample specific basis.

To describe the minor heterogeneity in the sample, fractional copy number states were added to bridge the gap between some neighboring states with a large difference in their means. However, adding too many states with close means can easily invoke spurious transitions between neighboring states that are caused by local data trends and may create artificial break points. This could make the results of some high level analysis misleading. For example, in subsequent automatic locus identification and prioritization algorithms, short segments are given larger weight than long ones because they are believed to be more informative on a biologic level. Thus, spurious short segments can significantly skew the results of these analyses. So far, using a p.star based model selection has successfully helped to decide whether partial copy states are needed. Additionally, one could choose whether to incorporate additional partial states based on the noise level of the profile -- fewer additional states would be appropriate for profiles with higher noise levels.

The outputs of this HMM method are copy number values which can be translated into log2 ratios that are equally representative for both gains and losses within a sample. Such equal representation of gain and loss is important for accurate descriptions of the biology of tumor specimens and is crucial for subsequent high level analyses at the informatic level. Compared with the relative log2 ratio outputs from other methods, the HMM log2 ratio data is more discrete, with copy number described in terms of a finite set of discrete log2 transformed values. Given that most current genomic data analysis methods are designed for continuous data, further new statistical methodology must be developed to better utilize this kind of discrete data in both probe level and segment level analyses. Such improved methodology will facilitate utilization of this data for addressing interesting biological problems such as pattern discovery and integrative analysis with data from other dimensions such as microarray gene expression profiles.

4. Software – R package HmmCGH

To fully facilitate the computation task in this study, the algorithm was implemented into a specific full function software system R package named HmmCGH.

4.1. Main Data structure

The main data structure includes 3 R class/object:

(1) acghSet:

This object is used to store the raw log2 ratio of the entire dataset and sample information. The median filtered data is store in this structure mainly used for visual inspection and estimation of mean for the ploidy copy.

(2) hmmParam

A set of parameters used to make copy number inference for a single profile: including initial states, the ploidy copy, initial compression scaling parameter, variance estimation model options and model selection on partial states. Wrapping all essential parameters in one single object makes the parameter easy to access in function calls.

(3) hmmInferRes

The inference results for a single profile. This object encapsulates a few data frames: the inferred copy number with initial and final states setup, the signal distribution of different copy numbers, the uniform copy number segments summary and the segmented data with its original log2 ratio scale. The structure provides a convenient and ready access for visualization and data summarization.

4.2. Functionality and work flow

The major functionality integrated with its work flow can be viewed as:

- IO: read in raw data, write inference result in table format
- HMM inference: copy number inference on array-CGH data
- Model selection on different parameters
- Visualization results: graphical display of the inferred results on genome-wise, chromosome and focal level.

4.3. Computational performance

With the fast increasing resolution of the microarray platform, efficient computation time is required to make durable analysis on a large dataset. The Viterbi algorithm used in this method requires O(3) ($l*k*k*N$, where l is the number of probes in entire profile, k is number of states and N is a number of fixed steps in each loop). Known to be inefficient on loop based procedures, 17K data points takes 2~3 minute to finish for one single sample when the inference procedure is first implemented in R. To reduce the computation time, the procedure was re-implemented in C and connected with R. The final running time is around 3 seconds for such sample. With the final visualization and summarization factored in, the entire procedure only takes around our hour for a dataset of 100 samples which is fairly efficient.

5. Acknowledgement

First of all I am very grateful to my research supervisor Dr. Lynda Chin for fully supporting me on this project. Dr. Cheng Li provided invaluable guide and help to this work as the technical mentor. Dr. Cameron Brennan brought me into this field and given me many insightful suggestions during the course of this project.

Also I would like to thank the following individuals for their support and help me on many aspects such providing datasets, draft editing and helpful comments: Drs. Ruben Carrasco, Giovanni Tonon, Elizabeth Maher, Andrew Aguirre, Alexei Protopopov, Bin Feng and Raktim Sinha as well as all other members in the bioinformatics group.

6. Reference

1. Albertson, D.G., Collins, C., McCormick, F. & Gray, J.W. Chromosome aberrations in solid tumors. *Nat Genet* **34**, 369-76 (2003).
2. Albertson, D.G. & Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* **12 Spec No 2**, R145-52 (2003).
3. Hampton, G.M. et al. Simultaneous assessment of loss of heterozygosity at multiple microsatellite loci using semi-automated fluorescence-based detection: subregional mapping of chromosome 4 in cervical carcinoma. *Proc Natl Acad Sci U S A* **93**, 6704-9 (1996).
4. Medintz, I.L. et al. Loss of heterozygosity assay for molecular detection of cancer using energy-transfer primers and capillary array electrophoresis. *Genome Res* **10**, 1211-8 (2000).
5. Kallioniemi, A. et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-21 (1992).
6. Pinkel, D. et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**, 207-11 (1998).
7. Brennan, C. et al. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res* **64**, 4744-8 (2004).
8. Fauth, C. & Speicher, M.R. Classifying by colors: FISH-based genome analysis. *Cytogenet Cell Genet* **93**, 1-10 (2001).
9. Schrock, E. et al. Multicolor spectral karyotyping of human chromosomes. *Science* **273**, 494-7 (1996).
10. Imoto, H. et al. Direct determination of NotI cleavage sites in the genomic DNA of adult mouse kidney and human trophoblast using whole-range restriction landmark genomic scanning. *DNA Res* **1**, 239-43 (1994).
11. Lisitsyn, N. & Wigler, M. Cloning the differences between two complex genomes. *Science* **259**, 946-51 (1993).
12. Merlino, G.T. et al. Amplification and enhanced expression of the epidermal growth factor receptor gene in A431 human carcinoma cells. *Science* **224**, 417-9 (1984).
13. Alitalo, K., Schwab, M., Lin, C.C., Varmus, H.E. & Bishop, J.M. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. *Proc Natl Acad Sci U S A* **80**, 1707-11 (1983).
14. Slamon, D.J. et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* **244**, 707-12 (1989).
15. Hinds, P.W., Dowdy, S.F., Eaton, E.N., Arnold, A. & Weinberg, R.A. Function of a human cyclin gene as an oncogene. *Proc Natl Acad Sci U S A* **91**, 709-13 (1994).
16. Sakaguchi, A.Y. et al. Human c-Ki-ras2 proto-oncogene on chromosome 12. *Science* **219**, 1081-3 (1983).
17. Shayesteh, L. et al. PIK3CA is implicated as an oncogene in ovarian cancer. *Nat Genet* **21**, 99-102 (1999).
18. Collins, C. et al. Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma. *Proc Natl Acad Sci U S A* **95**, 8703-8 (1998).
19. Shaffer, L.G. & Bejjani, B.A. A cytogeneticist's perspective on genomic microarrays. *Hum Reprod Update* **10**, 221-6 (2004).
20. Snijders, A.M. et al. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* **29**, 263-4 (2001).
21. Fiegler, H. et al. DNA microarrays for comparative genomic hybridization based on

DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* **36**, 361-74 (2003).
22. Chung, Y.J. et al. A whole-genome mouse BAC microarray with 1-Mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization. *Genome Res* **14**, 188-96 (2004).
23. Aguirre, A.J. et al. High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A* **101**, 9067-72 (2004).
24. Pollack, J.R. et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* **23**, 41-6 (1999).
25. Hodgson, G. et al. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat Genet* **29**, 459-64 (2001).
26. Olshen, A.B. & Venkatraman, E.S. Change-point analysis of array-based comparative genomic hybridization data. *Proceedings of the Joint Statistical Meetings*, 2530-2535 (2002).
27. Fridlyand, J., A.M., S., D., P., D.G., A. & A.N., J. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90**, 132–153 (2004).
28. Jong, K. et al. Chromosomal breakpoint detection in array comparative genomic hybridization data. in *In Applications of Evolutionary Computing: Evolutionary Computation and Boinformatics*, Vol. 2611 54–65 (Springer, Berlin, 2003).
29. Autio, R., S. , Hautaniemi, P., Kauraniemi, O., Yli-Harja, J. & Astola, M.W. CGH-plotter: MATLAB toolbox for cgh-data analysis. *Bioinformatics* **13**, 1714-1715 (2003).
30. Picard, F. et al. A statistical approach for CGH microarray data analysis. *INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE* **Mar**, 5139 (2003).
31. Wang, Y. & Guo, S.W. Statistical methods for detecting genomic alterations through array-based comparative genomic hybridization (CGH). *Front Biosci* **9**, 540-9 (2004).
32. Dobbin, K., Shih, J.H. & Simon, R. Statistical design of reverse dye microarrays. *Bioinformatics* **19**, 803-10 (2003).
33. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829-836 (1979).
34. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(1989).
35. Churchill, G.A. *Bulletin of Mathematical Biology*, 79-94 (1987).
36. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).
37. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* **84**, 2363-7 (1987).
38. Lin, M. et al. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20**, 1233-40 (2004).
39. Schliep, A., Schonhuth, A. & Steinhoff, C. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* **19 Suppl 1**, i255-63 (2003).
40. Ghadimi, B.M. et al. Specific chromosomal aberrations and amplification of the AIB1 nuclear receptor coactivator gene in pancreatic carcinomas. *Am J Pathol* **154**, 525-36 (1999).