2

Uncover Cancer Genomics by Jointly Analysing Aneuploidy and Gene Expression

Lingling Zheng and Joseph Lucas
Duke University
USA

1. Introduction

Human cancers are heterogeneous due to combined effects of genetic instability and selection, where the accumulation of the most advantageous set of genetic aberrations results in the expansion of cancer cells (Pinkel & Albertson, 2005). There are many different types of instability that occurs during tumor development, such as point mutation, alteration of microsatellite sequences, chromosome rearrangements, DNA dosage aberrations and epigenetic changes such as methylation. These abnormalities acting alone or in combination alter the expression levels of mRNA molecules. However, the genetic history of tumor progression is difficult to decipher. Because it is only a sufficiently protumorigenic aberration or obligate products of a crucial alteration that results in tumor development (Pinkel & Albertson, 2005).

Genomic DNA copy number variations (CNVs), kilobase- or megabase-sized duplications and deletions, are frequent in solid tumors. It has been shown that CNVs are useful diagnosis markers for cancer prediction and prognosis (Kiechle et al., 2001; Lockwood et al., 2005). Therefore, studying the genomic causes and their association with phenotypic alterations is emergent in cancer biology. The underlying mechanism of CNV related genomic instability amongst tumors includes defects in maintenance/manipulation of genome stability, telomere erosion, chromosome breakage, cell cycle defects and failures in DNA repairs (Albertson, 2003). Consequential copy number aberrations of the above mentioned malfunctions will further change the dosage of key tumor-inducing and tumor-suppressing genes, which thereby affect DNA replication, DNA damage/repair, mitosis, centrosome, telomere, mRNA transcription and proliferation of neoplastic cells. In addition, microenvironmental stresses play a role in exerting strong selective pressure on cancer cells with amplification/deletion of particular regions of the chromosome (Lucas et al., 2010). Recently, high-throughput technologies have mapped genome-wide DNA copy number variations at high resolution, and discovered multiple new genes in cancer. However, there is enormous diversity in each individual's tumor, which harbors only a few driver mutations (copy number alterations playing a critical role in tumor development). In addition, CNV regions are particularly large containing many genes, most of which are indistinguishable from the passenger mutations (copy number segments affecting widespread chromosomal instability in many advanced human tumors) (Akavia et al., 2010). Thus analysis based on CNV data alone will leave the functional importance and physiological impact of genetic alteration ineluctable on the tumor. Gene expression has been readily available for profiling many tumors, therefore, how

to incorporate it with CNV data to identify key drivers becomes an important problem to uncover cancer mechanism.

This chapter is laid out as follows: Section 2 covers a variety of CNV data topics, starting with a range of different CNV measurement techniques, which includes a brief discussion of the data format. Practical examples are used to show collecting, generating and assessing data, plus several ways to manipulate data for normalization. In the end, different computational approaches are introduced for analyzing CNV data. Section 3 focuses on an algorithm for integrating CNV with mRNA expression data, which can be potentially extended to incorporate multiple genomic data. Basic concepts of Bayesian factor analysis are briefly mentioned. Case studies then provide detailed description for this particular approach. Section 4 provides a brief wrap-up of the main ideas in the chapter. It illustrates the advantage of our statistical models on studying cancer genomics, and discusses the significance of the approach for clinical application.

2. Copy number analysis

2.1 Copy number analyses techniques

Comparative genome hybridization (CGH) is a recently developed technology and profiles genome-wide DNA copy number variations at high resolution. It has been popular for molecular classification of different tumor types, diagnosis of tumor progression, and identification of potential therapeutic targets (Jonsson et al., 2010; McKay et al., 2011). The use of CGH array offers many advantages over traditional karyotype or FISH (fluorescence *in situ* hybridization). It can detect microduplications/deletions throughout genome in a single experiment.

BAC Array

The CGH array using BAC (bacterial artificial chromosome) clones has been widely used. The spotted genomic sequences are inserted BACs: two DNA samples from either subject tissue (target sample) or control tissue (reference sample) are labeled with different fluorescent dyes–for example, with the test labeled in green and reference in red. The mixture is hybridized to a CGH array slide containing hundreds or thousands of defined DNA probes. The probes targeting regions of the chromosome that are amplified turn predominantly green. Conversely, if a region is deleted in the test sample, the corresponding probes become red. However, given the resolution limitation on the order of 1Mb and array size of 2400 to $\sim$30000 unique elements, the BAC array data is relatively low density.

cDNA/oligonucleotide Array

cDNA and oligonucleotide arrays are designed to detect complementary DNA "targets" derived from experiments or clinics. It allows greater flexibility to produce customized arrays, and reduces the cost for each study. Since commercial arrays are often more expensive and contain a large number of genes that are not of interest to the researchers. The shorter probes spotted on these new arrays are less robust than large segmented BACs. But they provide higher resolution in the order of 50-100kb, where oligonucleotide array is a particular case.

Tiling Array

Tiling arrays are available now for finer resolution of specific CNV regions. These arrays are designed to cover the entire genome or contiguous regions within the genome. Number

of elements on the array ranges from 10000 to over 6000000. This relatively high resolution technique allows the detection of micro-amplifications and deletions.

SNP Array

SNP (single nucleotide polymorphism) arrays are a high-density oligonucleotide-based array that can be used to identify both loss of heterozygosity (LOH) and CNVs. LOH is the loss of one allele of a gene, which can lead to functional loss of normal tumor suppressor genes, particularly if the other copy of the gene is inactive. LOH is quite common in malignancies. Therefore, utilization of SNP arrays to detect LOH provides great potential for cancer diagnosis.

Array CGH

Array comparative genomic hybridization (array CGH, or aCGH) is a high-resolution technique for genome-wide DNA copy number variation profiling. This method allows identification of recurrent chromosome changes with microamplifications and deletions, and detects copy number variations on the order of 5-10kb DNA sequences. In the rest of this chapter, we will use the CNV data generated from the general Agilent Human Genome CGH microarray 244A.

2.2 Array CGH data

The CNV data is obtained from The Cancer Genome Atlas (TCGA) project. TCGA is a joint effort of the National Cancer Institute and the National Human Genome Research Institute (NIHGRI) to understand genomic alterations in human cancer. It aims to study the molecular mechanisms of cancer in order to improve diagnosis, treatment and prevention. The importance of DNA copy number variations has been demonstrated in many tumors. TCGA targets to perform high-resolution CNV profiling in a large-scale study, using diverse tumor tissues and across different institutes. In this section, we will show an example from TCGA project.

Sample collection

Biospecimens were collected from newly diagnosed patients with ovarian serous cystadenocarcinoma (histologically consistent with ovarian serous adenocarcinoma confirmed by pathologists), who had not received any prior treatment, including chemotherapy or radiotherapy. Technical details about sample collection and quality control are described in (*Integrated genomic analyses of ovarian carcinoma*, 2011). Raw copy number data was generated at two centers, Brigham and Women's Hospital of Harvard Medical School and Dana Farber Cancer Institute, using the Agilent Human Genome Comparative Genome Hybridization 244A platform.

Data process

After the array CGH is constructed and tumor DNA samples hybridized to the platform, several steps need to be completed for detecting regions of copy number gains or losses: image scanning, image analysis (including gridding, spot recognition, segmentation and quantification, and low-intensified feature removal or mark), background noise subtraction, spot intensity ratio determination, log-transformation of ratios, signal normalization and quality control on the measured values. For Agilent 244K array, there are specific details on the data generation (*Comprehensive genomic characterization defines human glioblastoma genes*

and core pathways, 2008). First of all, the raw signal is obtained by scanning images using Agilent Feature Extraction Software (v9.5. 11), followed by image analysis steps mentioned above. *Background correction:* The background corrected intensity ratios for both channels are calculated by subtraction of median background signal values (median pixel intensities in the predefined background area surrounding the spot) of each channel from the median signal values (median pixel intensities computed over the spot area) of each probe in the corresponding channel. Since there are multiple copies of probes on an array, the final background corrected values are computed by taking the median across the duplicated probes. The $\log_2$ ratios of the above results are then estimated based on the background corrected values of sample channel over that of the reference channel. *Normalization of logarithmic ratio:* The normalization procedure involves the application of LOWESS (locally weighted regression and scatterplot smoothing) algorithm on $\log_2$ ratio data. This method assumes that the majority of probe $\log_2$ ratios do not change, and are independent of background corrected intensities of the probes. To develop the LOWESS model, a 21-probe window is applied for smoothing process after sorting the chromosome positions. It corrects the $\log_2$ ratio data so that the corresponding central tendency after normalization lies along zeros, assuming an equal number of up- and down- regulated features in any given intensity range. In addition, the artifact of the difference in the probe GC content on $\log_2$ ratios is considered for correction, in which case, the probe GC%, regional GC % (GC% of 20KB of genome sequence containing the probe sequence) and $\log_2$ ratio are used in the LOWESS model. *Quality control:* There are several criterions taken into account for quality assurance at various stages. 1) Probes that are flagged (marking spots of poor quality and low intensity) or saturated by the Agilent feature extraction software are eliminated; 2) Screening of the array image is conducted to exclude probes whose median signal values are lower than that of the background intensity; 3) Arrays with over 5% probes flagged out or being faint are considered as low quality; 4) The square root of the mean sum squares of variance in $\log_2$ ratio data between consecutive probes are calculated for quality assessment. Arrays with the value over 0.3 are considered as low quality.

The final result after these processes forms a data set containing 227614 probes with normalized $\log_2$ ratio values for every sample. The logarithmic ratios are computed as $\log_2(x) - \log_2(2)$, where x is the copy number inferred by the chip. Thus, ratios should be 0 for double loss, $\frac{1}{2}$ for a single loss, 1 for the normal situation, $\frac{3}{2}$ for a single gain, and $\frac{n}{2}$ for n copies. TCGA provides an Array Design Format file with annotation data, including information on chromosomal location and gene symbol for each probe.

Algorithms for CNVs detection

The main biomedical question for studying CNVs and downstream research is to accurately identify genomic/chromosomal regions that show significant amplification or deletion in DNA copy number. Satisfactorily solving this problem requires a method that reflects the underlying biology and key features of the technological platform. The array CGH data has particular characteristics: The status of DNA copy number remains stable in the contiguous loci, and the copy number of a probe is a good predictor for that of the neighboring ones, whereas for probes located far apart, it provides less information to predict the likely state of its neighboring probes (Rueda & Díaz-Uriarte, 2007). However, widely used array CGH platforms, such as cDNA/oligonucleotide arrays, do not have equally spaced probes, making it less informative based on consecutive probes. Furthermore, the identification of disease causal genes sometimes requires examining the amplitude of CNVs, especially when

high-resolution technologies are available, it can be valuable to distinguish between moderate copy number gains and large copy number amplification.

A number of well-known methods have been developed to carry out automatic identification of copy number gains/loss, and correlate that with diseases. These approaches are designed to estimate the significance level and location of CNVs. Models differ in distribution assumption and incorporation of penalty terms for parameter estimation. Subsequently, smoothing algorithms were derived for denoising and estimating the spatial dependence, such as wavelets (Hsu et al., 2005) and lowess methods (Beheshti et al., 2003; Cleveland, 1979). Later on, a binary segmentation approach, called circular binary segmentation (CBS) (Olshen et al., 2004), was proposed that allows segments in the aCGH data in each chromosome, and computes the within-segment means. CBS recursively estimates the maximum likelihood ratio statistics to detect the narrowed segment aberrations. A more complicated likelihood function was used with weights chosen in a completely data adaptive fashion (Adaptive weights smoothing procedure, AWS) (Hup et al., 2004). A different kind of modeling approach involves the hidden Markov model (HMM) (Fridlyand, 2004), which assigns hidden states with certain transition probabilities to underlying copy numbers. Thus, it adequately takes advantage of the physical dependence information of the nearby fragments. However, questions arise on how to appropriately select the number of hidden states. The sticky hidden Markov model with a Dirichlet distribution (sticky DD-HMM) (Du et al., 2010) was then developed to infer the number of states from data, while also imposing state persistence. Alternatively, the reversible jump aCGH (RJaCGH) (Rueda & Díaz-Uriarte, 2007) was introduced to fit the model with varying number of hidden states, and allow for transdimensional moves between these models. It also incorporates interprobe distance.

3. Joint analysis on copy number variation and gene expression

3.1 Overview

With the increasing availability of concurrently generating multiple different types of high throughput data on single samples, there is a lot of interest to jointly analyze this information and refine the generation of relevant biological hypotheses. This will lead to a greater, more integrated understanding of cellular mechanism, and will allow the identification of genomic regulators as well as suggest potentially synergistic drug targets for those regulators, which will lead to potential combination therapies for the treatment of human cancer. A number of approaches have demonstrated an ability to select specific genes from joint analysis and test specific hypotheses regarding the regulation of cellular responses, which is a tremendous advantage over the pathway analyses that can be obtained from gene expression or CNVs alone.

Recently, there are publications that highlight the impact of combining other types of DNA modification and gene expression. (Parsons et al., 2008) have identified a number of potential driver mutations in Glioblastoma through an analysis of mutation, copy number variation and gene expression. Their approach is designed around the use of currently available methods for the analysis of individual data types to create a compressed set of features which are then used independently in predictive models. They utilize tree models, however the compressed features are independent variables that can, in principle, be used in any type of predictive model. The approach does make use of correlation within each type of data, but not across different data types.

A similar approach to the integration of disparate types of data is outlined in (Lanckriet et al., 2004), but in this case features are compressed through the use of kernel functions. These must be predefined for each data type, but once that is done all of the different data types are mapped to the same vector space allowing joint analysis. The approach is particularly suited to the use of support vector machines, rather than tree models, for the generation of models from all of the different data types. The approach is remarkably general in that almost any type of data may be incorporated, and in the paper they include compelling examples of the integration of expression and protein sequence data. It, however, does suffer from the same flaws as (Parsons et al., 2008) in that there is no provision for dealing with correlation across data types.

Another approach to integrative analysis is through the use of data from different assays to filter lists of genes sequentially. (Garraway et al., 2005) describe such an approach, in the context of the identification of MITF as a genomic determinant in malignant melanoma. The algorithm first identifies genomic regions that show copy number variation in the condition of interest, and then searches for genes that are significantly over or under expressed in samples that have duplications or deletions in that region. This is a very powerful approach in cases where there are few genes that pass the filtering criteria and where the relationship between gene expression and CNV is direct. Through our own experimentation, we find that there are often many genes that pass both filtering criteria. Additionally, the approach is dependent on the order in which the data types are used to perform the filtering. This is because the filtering criterion on the second data set is determined by the behavior observed on the first.

The version of integrative genomic analysis that is most similar to our own proposal is CONEXIC, detailed in (Akavia et al., 2010). CONEXIC is based on gene modules, which was initially developed for the analysis of gene expression data in isolation. Gene modules consist of groups of genes that are coexpressed, and these are embedded as leaves in a binary tree structure where the nodes are populated by putative gene expression regulators. In its original incarnation, the approach was intended to identify important regulators of groups of genes in the context of experimental interventions. As such, expression is assumed to be constant within any particular experimental group. Also, the original approach depends on a list of putative regulators, which can be tricky to generate. With CONEXIC, the identification of lists of potential regulators is generated from regions of the genome that demonstrate consistent copy number variation, and the gene module algorithm is largely retained. Fundamental to a binary tree model is the assumption that the expression pattern of a leaf, conditional on the expression pattern of its parent node, is independent of all other elements in the tree. This is a shortcoming of the CONEXIC approach. It is quite reasonable to expect that there are many ways that a cell has available to control the expression of a particular gene, including CNV, methylation, inactivation of promoters, and RNA interference, and multiple different regulators may combine to ultimately regulate gene expression. Because each node of the tree contains only one putative regulator, the model assumes that only one regulator is responsible for the observed expression pattern of a module.

3.2 Bayesian factor analysis

Bayesian factor analysis is a dimension reduction method to decompose variability among observations into a lower number of unobserved, uncorrelated factors. It has been widely applied in microarray analysis (Carvalho et al., 2008; Lucas et al., 2009), where the data usually comes with a much higher dimension than the number of observed samples. Therefore, it

is desirable to select important genes that should bear some biological meanings. Recent developments in Bayesian multivariate modeling has enabled the utility of sparsity induced structure in genomic studies (Lucas et al., 2006). Such a sparse factor model implies that only those genes with non-zero loadings on those factors are relevant, and higher values indicate more significant gene-factor relationship.

3.3 Sparse regression model of Bayesian factor analysis

Our statistical framework utilizes high-dimensional sparse factor model, and is extended to incorporate gene expression, CNVs and other high-throughput genomic data. The underlying hypothesis is that the gene signatures of expression variation can be represented by the estimated factors. Furthermore, given the potential contribution of chromosomal aneuploidy and CNVs to the altered mRNA expression of relevant genes during oncogenesis, we could use the factor model to test for the association between gene expression signatures and CNVs. The model assumes that the input data are from the same organism. Suppose the data structure is given as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ with dimension $n \times p_x$, where n denotes the sample size, p_x the number of genes, and $\mathbf{x}_i$ the fluorescence level from probes of gene expression measurements. The CNV data is represented by $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ with similar strucutre. Therefore, the linear regression model for sample i can be expressed as

$$\mathbf{x}_i = \mathbf{B}_h \mathbf{h}_i + \mathbf{B}\mathbf{F}_i + \boldsymbol{\epsilon}_i \tag{1}$$

$$\mathbf{y}_i = \mathbf{A}_h \mathbf{h}_i + \mathbf{A}\mathbf{G}_i + \boldsymbol{\zeta}_i \tag{2}$$

with the following components:

• $\mathbf{B}$ is the $p_x \times k$ factor loadings matrix for sample $\mathbf{x}_i$, with elements $\beta_{g,j}$ for $g = 1, \ldots, p_x$ and $j = 1, \ldots, k$.

• $\mathbf{F}_i = [\mathbf{f}_i^C; \mathbf{f}_i^{(r)}]^{\mathrm{T}}$. $\mathbf{f}_i$ is a k-dimension vector of factor scores, where $\mathbf{f}_i^{(r)}$, the r-th factor for sample i, are specific to data $\mathbf{x}_i$, and $\mathbf{f}_i^C$ consists of the factors ***common*** between both data.

• $\mathbf{B}_h$ is the $p_x \times r$ regression matrix for dataset $\mathbf{x}_i$, with elements $b_{g,j}$ for $g = 1, \ldots, p_x$ and $j = 1, \ldots, r$.

• $\mathbf{h}_i = \left[h_{1,i}, \ldots, h_{q,i}\right]^{\mathrm{T}}$ is the q design factors of sample i.

• $\boldsymbol{\epsilon}_i = \left[\epsilon_{1,i}, \ldots, \epsilon_{p_x,i}\right]^{\mathrm{T}}$ is the idiosyncratic noise vector with dimension p_x.

The priors for each parameters are defined as follows:

$$\beta_{g,j} \sim (1 - \rho_j)\delta_0(\beta_{g,j}) + \rho_j \mathcal{N}(\beta_{g,j}; 0, \tau_j) \tag{3}$$

$$\rho_j \sim Beta(\rho_j; s_0, l_0); \tau_j \sim Gamma({\tau_j}^{-1}; \frac{a_\tau}{2}, \frac{b_\tau}{2}) \tag{4}$$

$$b_{g,j} \sim (1 - \pi_j)\delta_0(b_{g,j}) + \pi_j \mathcal{N}(b_{g,j}; \mu_{0,j}, {\sigma_0}^2) \tag{5}$$

$$\pi_j \sim Beta(\pi_j; t_0, v_0) \tag{6}$$

$$\mathbf{f}_i^{(r)} \sim \mathcal{N}(\mathbf{f}_i^{(r)}; \mathbf{0}, \mathbf{I}) \qquad \mathbf{f}_i^C \sim \mathcal{N}(\mathbf{f}_i^C; \mathbf{g}_i^C, \Sigma) \tag{7}$$

$$\boldsymbol{\epsilon}_i^{(r)} \sim \mathcal{N}(\boldsymbol{\epsilon}_i^{(r)}; \mathbf{0}, \boldsymbol{\Phi}); \boldsymbol{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_{p_x}); \phi_g \sim Gamma(\phi_g; \frac{a_{\phi_x}}{2}, \frac{b_{\phi_x}}{2}) \tag{8}$$

The parameters and prior structures are similar for copy number data $\mathbf{y}_i$.

Prior Choices

• $\beta_{g,j}$: The regression coefficient. Here we consider the long-standing problem of variable selection in a multivariate linear regression model. That is, in gene expression analysis the number of gene features is huge (usually larger than 20,000) compared with the number of samples available. A direct way is to use regression model on the high-dimensional genomic data and impose sparseness on the coefficients. In this way, most of the coefficients will be shrunk towards zero. Bayesian spike and slab approaches (George & McCulloch, 1993; Ishwaran & Rao, 2005; Mitchell & Beauchamp, 1988) have been proposed to address the variable selection problem. As indicated in 3, it sets up a two-component mixture distribution with the spike part centered at zero and the slab part distributed diffusely without informed prior knowledge.

• ρ_j: This parameter controls the prior probability of a coefficient being non-zero. We assume coefficients that are promising have posterior latent variables $\hat{\rho}_j = 1$ (the slab). The opposite occurs when $\hat{\rho}_j = 0$ with a delta function $\delta_0(\cdot)$ indicating the point-mass at zero (the spike). Here we use beta priors, defining the probability ρ_j distributed on the interval (0,1). The hyperparameters s_0 and l_0 determine the domain of the beta distribution. Small values of ρ_j reflect high prior skepticism about the coefficients, while large ρ_j means the knowledge of more theoretical importance of the variables and more skeptical about the sampling of the data.

• τ_j: the variance for the slab part of the mixture prior for $\beta_{g,j}$. This gamma distribution is the conjugate prior for the precision of the normal distribution $\mathcal{N}(\beta_{g,j}; 0, \tau_j)$. In addition, it allows the Markov chain to identify and adjust the appropriate sample space for updating coefficients. Different combinations of ρ_j and τ_j prior choices are usually required to obtain desirable mixing and shrinkage in $\beta_{g,j}$.

• $\mathbf{f}_i$: Unknown latent factors for sample i. For factors unique for each data, we use a diffuse, conjugate prior distribution such that $f_{j,i} \sim \mathcal{N}(0,1)$, in order to alleviate issues with identifiability of $\mathbf{f}_i$ and $\boldsymbol{\beta}$ due to scaling. On the other hand, since high-throughput data can vary in size by orders of magnitude, e.g. CGH data is approximately ten times larger than gene expression. Thus one data set may dominate the factor model given a large size discrepancy. Therefore, rather than utilizing the uninformative prior, we link individual factors from each data using $\mathbf{f}_i^C \sim \mathcal{N}(\mathbf{g}_i^C, \Sigma)$ based on the hypothesis that gene expression is directly influenced by CNVs. This will prevent difference in data size from overwhelming the information available on associations between them. In addition, the systematic error between two data sets will be considered by estimation of the covariance matrix Σ.

Updated Distributions

• $p(\beta_{g,j}|-)$:

For factor j, let $x^*_{g,j} = x_{g,j} - \sum_{j=1}^{r} b_{g,j} h_{j,i} - \sum_{l \neq j}^{k} \beta_{g,l} f_{l,i}$, so that $x^*_{g,j} \sim \mathcal{N}(\beta_{g,j} f_{j,i}, \phi_g)$. In order to be mathematically identifiable for $\mathbf{B}$, we assume the regression coefficients a lower triangular matrix with positive diagonal elements (Carvalho & West, 2006). This gives the following posterior updates where $g \neq j$:

$$
\begin{aligned}
p(\beta_{g,j}|-) &\propto \prod_{i=1}^{n} p(x^*_{g,j}|\beta_{g,j}f_{j,i},\phi_g)p(\beta_{g,j}) \\
&= \prod_{i=1}^{n} \mathcal{N}(x^*_{g,j};\beta_{g,j}f_{j,i},\phi_g)((1-\rho_j)\delta_0(\beta_{g,j}) + \rho_j\mathcal{N}(\beta_{g,j};0,\tau_j)) \\
&= (1-\hat{\rho}_j)\delta_0(\beta_{g,j}) + \hat{\rho}_j\mathcal{N}(\beta_{g,j};\mu_{g,j},\Omega_{g,j})
\end{aligned}
$$

where $\Omega_{g,j} = (\tau_j^{-1} + \sum_{j=1}^{k} f_{j,i}^2/\phi_g)^{-1}, \mu_{g,j} = \Omega_{g,j}(\sum_{i=1}^{n} x^*_{g,i}f_{j,i})\phi_g^{-1}$ and $\beta_{g,j} \neq 0$ with probability

$$
\hat{\rho}_j = \frac{\rho_j}{\rho_j + (1-\rho_j)\frac{\mathcal{N}(0;0,\tau_j)}{\mathcal{N}(0;\mu_{g,j},\Omega_{g,j})}}
$$

For the constrained diagonal elements of $\mathbf{B}$, the posterior conditional distribution is given as

$$
p(\beta_{j,j}|-) \sim \mathcal{N}(\mu_{j,j},\Omega_{j,j})\mathbf{I}(\beta_{j,j} > 0)
$$

with similar forms of $\mu_{j,j}$ and $\Omega_{j,j}$.

• $p(\rho_j|-)$:

$$
\begin{aligned}
p(\rho_j|-) &\propto \prod_{j=1}^{k} p(\beta_{g,j}|\rho_j)p(\rho_j) = (1-\rho_j)^{p_x-j-S_j}\rho_j^{S_j}Beta(\rho_j;s_0,l_0) \\
&\sim Beta(s_0+S_j, l_0+p_x-j-S_j)
\end{aligned}
$$

with $S_j = \sum_{g=j}^{p_x}\mathbf{I}(\beta_{g,j} \neq 0)$.

$$
\begin{aligned}
p(\tau_j|-) &\propto \prod_{g=1}^{p_x} p(\beta_{g,j}|\rho_j,\tau_j)p(\tau_j) = \prod_{g=1}^{p_x}\mathcal{N}(\beta_{g,j};0,\tau_j)Ga({\tau_j}^{-1};\frac{a_\tau}{2},\frac{b_\tau}{2}) \\
&\sim InvGamma(\tau_j;\frac{a_\tau+\omega_j}{2},\frac{b_\tau+\sum_{g=1}^{p_x}\beta_{g,j}^2}{2})
\end{aligned}
$$

with $\omega_j = \sum_{g=j}^{p_x}\mathbf{I}(\beta_{g,j} \neq 0)$.

• $p(\mathbf{f}_i|-), p(\mathbf{g}_i|-)$:

Let $\mathbf{F} = [\mathbf{f}_1,\ldots,\mathbf{f}_n]$. The posterior distribution of $\mathbf{F}$ can be updated as:

$$
\begin{aligned}
p(\mathbf{F}|-) &\propto p(\mathbf{X}|\mathbf{F},\mathbf{B},\mathbf{\Phi})p(\mathbf{F}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{f}_i,\mathbf{B},\mathbf{\Phi})p(\mathbf{f}_i) \\
&= \prod_{i=1}^{n}\mathcal{N}(\mathbf{x}_i - \mathbf{B}_h\mathbf{H}_i;\mathbf{B}\mathbf{f}_i,\mathbf{\Phi})\mathcal{N}(\mathbf{f}_i;\mathbf{g}_i,\Sigma) \\
&\propto \prod_{i=1}^{n}\mathcal{N}(\mathbf{f}_i;\mathbf{E}1_i,\mathbf{V}1_i)
\end{aligned}
$$

where $\mathbf{V}1_i = (\Sigma^{-1} + \mathbf{B}'\mathbf{\Phi}^{-1}\mathbf{B})^{-1}$, $\mathbf{E}1_i = \mathbf{V}1_i(\mathbf{B}'\mathbf{\Phi}^{-1}(\mathbf{x}_i - \mathbf{B}_h\mathbf{H_i}) + \mathbf{G}_i\Sigma^{-1})$.

Similarly $p(\mathbf{g}_i|-)$ takes the form

$$p(\mathbf{G}|-) \propto \prod_{i=1}^{n} \mathcal{N}(\mathbf{g}_i; \mathbf{E}2_i, \mathbf{V}2_i)$$

where $\mathbf{V}2_i = (\mathbf{I} + \mathbf{A}'\mathbf{\Psi}^{-1}\mathbf{A})^{-1}$, $\mathbf{E}2_i = \mathbf{V}2_i(\mathbf{A}'\mathbf{\Psi}^{-1}(\mathbf{y}_i - \mathbf{A}_h\mathbf{H_i}))$. $\mathbf{\Psi}$ is the covariance matrix of $\mathbf{y}_i$.

• $p(\phi_g|-)$:

$$\begin{aligned} p(\phi_g|-) &\propto \prod_{i=1}^{n} p(x_{g,i}|\beta_g f_i, \phi_g) p(\phi_g) \\ &= \prod_{i=1}^{n} \mathcal{N}(x_{g,i} - \sum_{j=1}^{r} b_{g,j} h_{j,i}; \beta_g f_i, \phi_g) Ga({\phi_g}^{-1}; \frac{a_{\phi_x}}{2}, \frac{b_{\phi_x}}{2}) \\ &\sim InvGamma(\phi_g; \frac{a_{\phi_x} + n}{2}, \frac{b_{\phi_x} + \sum_{i=1}^{n}(x_{g,i} - b_g h_i - \beta_g f_i)^2}{2}) \end{aligned}$$

3.4 Example: joint analysis of ovarian cancer gene expression and CNVs

We applied our joint factor model on ovarian cancer gene expression and CNV data from TCGA project. This study is aimed to detect correlations between them, which will lead to the identification of pivotal genomic determinants of cancer phenotypes. We adopted 74 ovarian cancer individuals and 1 disease-free patient's data. In order to capture genes with differential expression patterns and their association with the CNVs in the narrowed chromosomal regions, we established a filtering criteria: 1) select Affymetrix HT_HG-U133A probes with sample mean above 8, and standard deviation above 0.6; take out probes without matched gene symbols. It results in a gene expression data set downsized from 22277 to 921 probes; 2) apply the basic Bayesian factor model 1, i.e., the one that only analyzes one data set, and generate signature expression factors; 3) remove CNV segments (Agilent Human Genome CGH 244A probes) not showing significant correlation (p-value < 0.01 after Bonferroni correction) with the gene expression factors. It reduced the CNV data dimension from 227613 to 7278. Therefore, we fitted our joint factor model 1 and 2 to the shrunk data.

We obtained 11 factors in the two data sets, i.e., $\mathbf{F}_{11*75}$ and $\mathbf{G}_{11*75}$, and selected the most strongly associated pair using Pearson correlation. It turns out that the largest factor loadings in the corresponding CNV factor come mostly from the long arm of chromosome 8 (figure 1A), that the factor correlates well with the paired gene expression factor (figure 1B), and that the gene expression factor correlates with individual SNP observations in the long arm of chromosome 8 (figure 1C). Based on these results, we further examined the genes loaded on this correlated CGH factor and gene expression factor. By ranking the squared factor loadings, we selected the top 16 Affymetrix probe sets (Table 1) and 178 CGH probe sets, because the variance in these probes are best explained by the corresponding factors compared with all other data. Pearson correlation between the values of mRNA expression levels and copy number variations were calculated on these heavily loaded genes. We noted that the copy number gains of EBAG9 (CGH probe position: 8q23.2, size 60 bp; mean copy number 2.63 (1-6)) and MTDH (CGH probe position: 8q22.1, size 60 bp; mean copy number 2.38

(1-6)) significantly accompanies their overexpression of mRNAs in the corresponding regions, where correlation coefficients indicate a good linearity between CNVs and gene expression with r = 0.758 for EBAG9 and r = 0.806 for MTDH. Interestingly, in the same factor, 3 CGH loci with duplicated DNAs show significant correlation with MTDH overexpression (r>0.8, p-val<0.01) and are located 0.2M upstream, 5M and 12M downstream of MTDH CGH locus, respectively; and 11 CGH loci are identified with copy number gain and 3Mb upstream of EBAG9 CGH clone (r>0.75, p-val<0.01). These findings may provide evidence for distant regulatory of transcription elements or interactions within a potential gene network.

Gene symbol	Gene
MTDH	LYRIC/3D3 (UID: 92140)
EBAG9	estrogen receptor binding site associated, antigen, 9 (UID:9166)
YWHAZ	tyrosine 3-monooxygenase (UID:7534)
LAPTM4B	lysosomal protein transmembrane 4 beta (UID:55353)
ESRP1	epithelial splicing regulatory protein 1 (UID:54845)
NBN	nibrin (UID:9048)
RAD21	RAD21 homolog (S. pombe) (UID:5885)
RNF139	ring finger protein 139 (UID:11236)
ZNF706	HSPC038 protein (UID:51123)
AZIN1	antizyme inhibitor 1 (UID:51582)
DERL1	Der1-like domain family, member 1 (UID:79139)
ENY2	enhancer of yellow 2 homolog (Drosophila) (UID:56943)
EXT1	exostoses (multiple) 1 (UID:2131)
CTSB	cathepsin B (UID:1508)
DECR1	2,4-dienoyl CoA reductase 1, mitochondrial (UID:1666)
PTDSS1	phosphatidylserine synthase 1 (UID:9791)

Table 1. Genes on chromosome 8 showing significantly differential expression in ovarian cancer. The list is ranked by the squared factor loadings.

The product of EBAG9 has been identified as an estrogen receptor binding site associated antigen 9 identical to RCAS1 (Nakashima et al., 1999). Overexpression of EBAG9/RCAS1 inhibits growth of tumor-stimulated host immune cells and induces their apoptosis (Nakashima et al., 1999). Furthermore, it has been reported that RCAS1 is expressed with high frequency in ovarian and lung cancers (Akahira et al., 2004; Iwasaki et al., 2000), and the copy numbers of the region increase in breast cancer (Rennstam et al., 2003). These lines of evidence, together with the results obtained above, imply that overexpression of EBAG9 in ovarian serous cystadenocarcinoma may be triggered by increased gene copy number, which is likely to play an important role in the immune escape of tumor cells and causing cancer progression.

In addition, MTDH, also known as AEG1, is an oncogene cooperating with Ha-ras as well as functioning as a downstream target gene of Ha-ras and may perform a central role in Ha-ras-mediated carcinogenesis (Lee et al., 2007). Overexpression of this gene has been reported in various cancers including breast, brain, prostate, melanoma and glioblastoma multiforme (Emdad et al., 2007; Kikuno et al., 2007). In particular, it has been revealed that MTDH overexpression is associated with 8q22 genomic gain in breast cancer, and has been considered as an important therapeutic target for enhancing chemotherapy efficacy and reducing metastasis risk (Hu et al., 2009). Therefore, we believe that, our results along with the

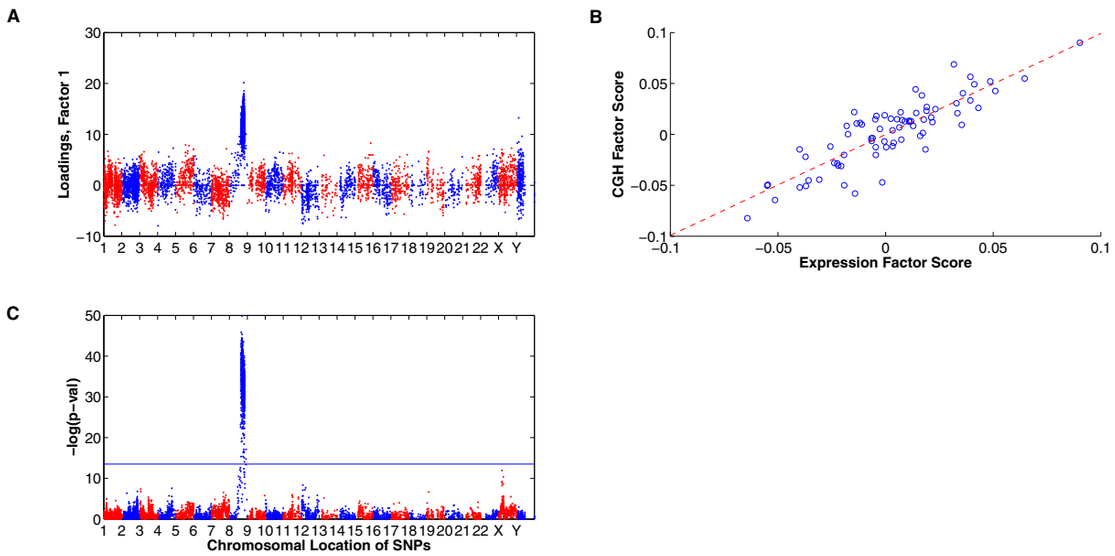


Fig. 1. Factor analytic relationship between CNV and gene expression. Panel A shows the factor loadings from the first factor of the joint factor model fit to CNV data. Panel B shows a scatterplot of significant correlation between gene expression factor and the CNV factor, of which it is linked to. Panel C shows the significance of correlation between the expression factor and each individual SNP from the high-density CGH array. The y-axis shows the -log(p-value) of the Pearson correlation between CNVs and gene expression factor. The horizontal line shows the threshold of p-value less than 0.01 after Bonferroni correction for multiple testing.

above findings suggest the copy number gain activated MTDH overexpression is a potential indicator in epithelial ovarian cancer.

Validations on the above hypotheses regarding critical genes in cancer progression and their regulation mechanisms can be carried out in several directions. A number of databases can be used to validate these hypotheses. For instance, GATHER and GOrilla are good resources to annotate gene functions; Tumorscape helps interpret copy number variations; DAVID Bioinformatics provides pathway analysis for genes identified by the model. In addition, experimental validation can be performed to quantitatively justify that the activation/inactivation of identified genes are caused by copy number variations. Moreover, we could identify drug susceptibilities of these candidates by searching against reference information from DrugBank (http://www.drugbank.ca), then using these results for experimental validation. The general approach is to grow cell lines in the presence of a particular treatment, whose genomic drivers are disrupted by the introduction of RNA inference and transfection with viral plasmids. Similar strategy can also been applied to predict potential therapies by the identification of new drug targets. Therefore, these will lead to a greater understanding of cancer progression, and allow the identification of combined therapies for individual tumors.

Tumor segmental aneuploidy association with gene expression factors has been demonstrated in a previous study (Lucas et al., 2010) that it makes significant contributions to variation in gene signature of breast cancer under the stress of lactic acidosis or hypoxia. We are

interested to test if this is consistent in other tumor tissues, which will provide potential treatment choices for different cancers. We used a similar approach (Lucas et al., 2010) by projecting the breast expression factors into TCGA ovarian and glioblastoma gene expression data and identified correlated CNVs under the same interventions of lactic acidosis/hypoxia. The ability of projecting the factor model into other data sets allows the possibility of comparing new experimental data to different genomic information, such as CNVs from aCGH. The underlying assumption is that genes showing shared expression patterns in tumors of different origins can be represented by the same loadings matrix. Therefore, in order to estimate the factor scores for the new data, this translates into a well known problem of inverse regression $\mathbf{F}_y = (\mathbf{I}_k + \mathbf{B}'\mathbf{\Phi}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{\Phi}^{-1}\mathbf{Y}$, where $\mathbf{B}$ is the loadings matrix and $\mathbf{\Phi}$ the diagonal matrix containing the gene by gene variance estimators in the original data, $\mathbf{Y}$ the new set of expression data and $\mathbf{F}_y$ the factor scores on the new data set. With this approach, we estimated factor scores for the TCGA data and calculated their correlations with CNVs. In our analysis, about half of the breast expression factors are also associated with copy number variations in ovarian cancer and that about a quarter are associated with CNVs in glioblastoma. For example, the CNV activated expression pattern in breast cancer (not shown) is also discovered in both ovarian cancer and glioblastoma within the same region

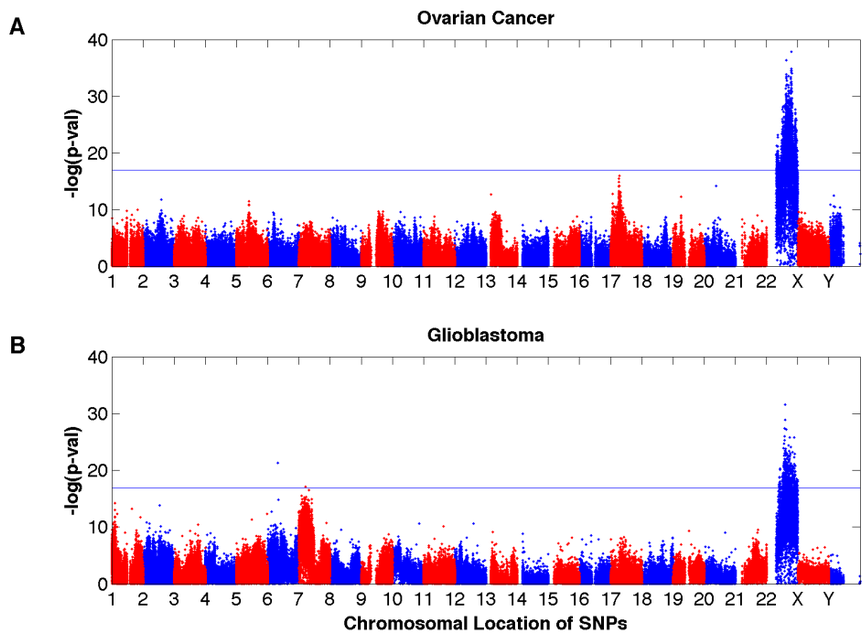


Fig. 2. Panel A and B show the the association between gene expression factor and CNVs across tumors of different origins. Each scatter plot indicates the evidence of association between the same factor that was learned on breast cancer data and copy number changes of different tumor tissues. Plot A shows correlation between the factor, projected onto ovarian cancer expression data, and ovarian CGH data. Plot B shows the same for Glioblastoma. Each point corresponds to one of the SNPs measured in the high-dimensional CGH array. The y-axis shows the -log(p-value) of the Pearson correlation between CNVs and gene expression factor. The horizontal line shows the threshold of p-value less than 0.01 after Bonferroni correction for multiple testing.

(figure 2A and 2B). Therefore, it is likely that similar CNVs might be selected under the same pressure of hypoxia/ lactic acidosis in difference cancers.

4. Conclusion

This chapter has built upon a basic understanding of a layout on the correlation between copy number variations and gene expression to deepen knowledge of key concepts and methods. By introducing and comparing a diverse range of techniques for measuring CNVs, we provide the scope of localizing cancer related genes using different platforms. By describing an appreciation of the use of several statistical methods to assist the positioning of CNV regions, we are aimed to better identify cancer driven mutations within the copy number gain/loss regions. Moreover, we have also included examples from TCGA project to show the unique features of CNV data.

The key challenge of finding candidate drivers is to distinguish it from passenger genes, which are physically located close to the driver mutations and whose variations are not causal to convey growth advantage on cancer cells. In our analysis, we focus on genes with cis-regulated CNVs, and postulate that cancer driven mutation is associated with the expression of a group of genes, and it is likely to localize in DNA amplified or deleted regions in tumors. Since DNA dosage variations may result in functional changes of affected genes and cause expression change of downstream genes. We have proposed a generic framework to jointly analyze disparate data sets, which is extendable to incorporate diverse information such as proteomics data. This will allow for more robust analysis of the relationship between mRNA expression and protein abundance. Our results not only identify candidate genes whose mRNA expression is statistically significantly correlated with their CNVs, but also successfully recover the region where similar gene expression pattern is triggered by the same genomic program across tumors of different organ systems. This approach is able to estimate the probability of each gene regulated by genomic sources and the relative importance of each source. Additionally, two genes, EBAG9 and MTDH, suggest that abnormal abundance in their DNA copy numbers may contribute to proliferation in ovarian serous cystadenocarcinoma. For these two predicted drivers, we also find many CNVs in the same region but poorly correlated with their gene expression, thus consider them no apparent effect in cancer. Copy number variation is only one of many ways that gene expression can be altered. We believe that a number of complementary approaches are needed to validate possibly driving alterations, as illustrated in the previous section. Therefore, We envision that our model is used as screening guidance to assist the identification of potential cancer drivers with possibly therapeutic importance.

Our work presents a framework toward a broad understanding of the genomic determinants of cancer. With this approach, we anticipate to generate testable biological hypothesis regarding the regulation of cellular responses, which is a tremendous advantage over any single data analyses that can be obtained from gene expression or CNVs alone. This will lead to a greater, more integrated understanding of cellular mechanism, and will allow the identification of genomic regulators as well as enhancement of anticancer drug specificity targeting those regulators. This is key to the discovery of potential combination therapies for the treatment of human cancer. Moreover, genomic patterns related to therapeutic response and clinical outcomes can be identified as biomarkers, which will improve early cancer detection, prognosis and outcome prediction as well as treatment selection. All in all, this

will create a comprehensive picture of heterogeneity in tumor genomes, and offer a valuable starting point for new therapeutic approaches.

5. References

Akahira, J.-i., Aoki, M., Suzuki, T., Moriya, T., Niikura, H., Ito, K., Inoue, S., Okamura, K., Sasano, H. & Yaegashi, N. (2004). Expression of ebag9/rcas1 is associated with advanced disease in human epithelial ovarian cancer, *Br J Cancer* 90(11): 2197–2202.
URL: *http://dx.doi.org/10.1038/sj.bjc.6601832*

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A. & Pe'er, D. (2010). An integrated approach to uncover drivers of cancer, *Cell* 143(6): 1005 – 1017.
URL: *http://www.sciencedirect.com/science/article/pii/S0092867410012936*

Albertson, D. G. (2003). Profiling breast cancer by array cgh, *Breast Cancer Research and Treatment* 78(3): 289–298.
URL: *http://dx.doi.org/10.1023/A:1023025506386*

Beheshti, B., Braude, I., Marrano, P., Thorner, P., Zielenska, M. & Squire, J. (2003). Chromosomal localization of dna amplifications in neuroblastoma tumors using cdna microarray comparative genomic hybridization., *Neoplasia (New York, N.Y.)* 5(1).
URL: *http://ukpmc.ac.uk/abstract/MED/12659670*

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. & West, M. (2008). High-Dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics, *Journal of the American Statistical Association* 103(484): 1438–1456.
URL: *http://pubs.amstat.org/doi/pdf/10.1198/01621450800000086*

Carvalho, C. M. & West, M. (2006). Structure and sparsity in high-dimensional multivariate analysis, *ProQuest Dissertations and Theses* .
URL: *http://search.proquest.com/docview/305329670?accountid=10598*

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* 74(368): 829–836.
URL: *http://www.jstor.org/stable/2286407*

Comprehensive genomic characterization defines human glioblastoma genes and core pathways (2008). *Nature* 455(7216): 1061–1068.
URL: *http://dx.doi.org/10.1038/nature07385*

Du, L., Chen, M., Lucas, J. & Carin, L. (2010). Sticky hidden markov modeling of comparative genomic hybridization, *Signal Processing, IEEE Transactions on* 58(10): 5353 –5368.

Emdad, L., Sarkar, D., Su, Z.-Z., Lee, S.-G., Kang, D.-C., Bruce, J. N., Volsky, D. J. & Fisher, P. B. (2007). Astrocyte elevated gene-1: Recent insights into a novel gene involved in tumor progression, metastasis and neurodegeneration, *Pharmacology and Therapeutics* 114(2): 155 – 170.
URL: *http://www.sciencedirect.com/science/article/pii/S0163725807000332*

Fridlyand, J. (2004). Hidden Markov models approach to the analysis of array CGH data, *Journal of Multivariate Analysis* 90(1): 132–153.
URL: *http://dx.doi.org/10.1016/j.jmva.2004.02.008*

Garraway, L. A., Widlund, H. R., Rubin, M. A., Getz, G., Berger, A. J., Ramaswamy, S., Beroukhim, R., Milner, D. A., Granter, S. R., Du, J., Lee, C., Wagner, S. N., Li, C., Golub, T. R., Rimm, D. L., Meyerson, M. L., Fisher, D. E. & Sellers, W. R. (2005). Integrative genomic analyses identify mitf as a lineage survival oncogene amplified

in malignant melanoma, *Nature* 436(7047): 117–122.
URL: *http://dx.doi.org/10.1038/nature03664*

George, E. I. & McCulloch, R. E. (1993). Variable selection via gibbs sampling, *Journal of the American Statistical Association* 88(423): 881–889.
URL: *http://www.jstor.org/stable/2290777*

Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L. & Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets, *Biostatistics* 6(2): 211–226.
URL: *http://biostatistics.oxfordjournals.org/content/6/2/211.abstract*

Hu, G., Chong, R. A., Yang, Q., Wei, Y., Blanco, M. A., Li, F., Reiss, M., Au, J. L.-S., Haffty, B. G. & Kang, Y. (2009). Mtdh activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer, *Cancer Cell* 15(1): 9 – 20.
URL: *http://www.sciencedirect.com/science/article/pii/S1535610808003796*

Hup, P., Stransky, N., Thiery, J.-P., Radvanyi, F. & Barillot, E. (2004). Analysis of array cgh data: from signal ratio to gain and loss of dna regions, *Bioinformatics* 20(18): 3413–3422.
URL: *http://bioinformatics.oxfordjournals.org/content/20/18/3413.abstract*

Integrated genomic analyses of ovarian carcinoma (2011). *Nature* 474(7353): 609–615.
URL: *http://dx.doi.org/10.1038/nature10166*

Ishwaran & Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies.

Iwasaki, T., Nakashima, M., Watanabe, T., Yamamoto, S., Inoue, Y., Yamanaka, H., Matsumura, A., Iuchi, K., Mori, T. & Okada, M. (2000). Expression and prognostic significance in lung cancer of human tumor-associated antigen rcas1, *International Journal of Cancer* 89(6): 488–493.
URL: *http://dx.doi.org/10.1002/1097-0215(20001120)89:6<488::AID-IJC4>3.0.CO;2-D*

Jonsson, G., Staaf, J., Vallon-Christersson, J., Ringner, M., Holm, K., Hegardt, C., Gunnarsson, H., Fagerholm, R., Strand, C., Agnarsson, B., Kilpivaara, O., Luts, L., Heikkila, P., Aittomaki, K., Blomqvist, C., Loman, N., Malmstrom, P., Olsson, H., Th Johannsson, O., Arason, A., Nevanlinna, H., Barkardottir, R. & Borg, A. (2010). Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics, *Breast Cancer Research* 12(3): R42.
URL: *http://breast-cancer-research.com/content/12/3/R42*

Kiechle, M., Jacobsen, A., Schwarz-Boeger, U., Hedderich, J., Pfisterer, J. & Arnold, N. (2001). Comparative genomic hybridization detects genetic imbalances in primary ovarian carcinomas as correlated with grade of differentiation, *Cancer* 91(3): 534–540.
URL: *http://dx.doi.org/10.1002/1097-0142(20010201)91:3<534::AID-CNCR1031>3.0.CO;2-T*

Kikuno, N., Shiina, H., Urakami, S., Kawamoto, K., Hirata, H., Tanaka, Y., Place, R. F., Pookot, D., Majid, S., Igawa, M. & Dahiya, R. (2007). Knockdown of astrocyte-elevated gene-1 inhibits prostate cancer progression through upregulation of foxo3a activity, *Oncogene* 26(55): 7647–7655.
URL: *http://dx.doi.org/10.1038/sj.onc.1210572*

Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. (2004). A statistical framework for genomic data fusion, *Bioinformatics* 20(16): 2626–2635.
URL: *http://bioinformatics.oxfordjournals.org/content/20/16/2626.abstract*

Lee, S.-G., Su, Z.-Z., Emdad, L., Sarkar, D., Franke, T. F. & Fisher, P. B. (2007). Astrocyte elevated gene-1 activates cell survival pathways through pi3k-akt signaling,

Oncogene 27(8): 1114–1121.
URL: *http://dx.doi.org/10.1038/sj.onc.1210713*

Lockwood, W. W., Chari, R., Chi, B. & Lam, W. L. (2005). Recent advances in array comparative genomic hybridization technologies and their applications in human genetics, *Eur J Hum Genet* 14(2): 139–148.
URL: *http://dx.doi.org/10.1038/sj.ejhg.5201531*

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. & West, M. (2006). *Sparse Statistical Modelling in Gene Expression Genomics*, pp. 155–176.

Lucas, J., Carvalho, C. & West, M. (2009). A bayesian analysis strategy for cross-study translation of gene expression biomarkers., *Statistical applications in genetics and molecular biology* 8(1).
URL: *http://dx.doi.org/10.2202/1544-6115.1436*

Lucas, J. E., Kung, H.-N. & Chi, J.-T. A. (2010). Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers, *PLoS Comput Biol* 6(9): e1000920.
URL: *http://dx.doi.org/10.1371/journal.pcbi.1000920*

McKay, S. C., Unger, K., Pericleous, S., Stamp, G., Thomas, G., Hutchins, R. R. & Spalding, D. R. C. (2011). Array comparative genomic hybridization identifies novel potential therapeutic targets in cholangiocarcinoma, *HPB* 13(5): 309–319.
URL: *http://dx.doi.org/10.1111/j.1477-2574.2010.00286.x*

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression, *Journal of the American Statistical Association* 83(404): 1023–1032.
URL: *http://www.jstor.org/stable/2290129*

Nakashima, M., Sonoda, K. & Watanabe, T. (1999). Inhibition of cell growth and induction of apoptotic cell death by the human tumor-associated antigen rcas1, *Nat Med* 5(8): 938–942.
URL: *http://dx.doi.org/10.1038/11383*

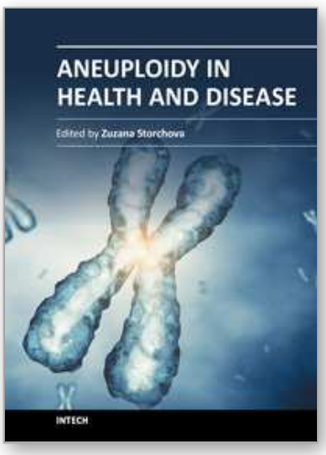
Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. (2004). Circular binary segmentation for the analysis of arrayâĂŘbased dna copy number data, *Biostatistics* 5(4): 557–572.
URL: *http://biostatistics.oxfordjournals.org/content/5/4/557.abstract*

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., Olivi, A., McLendon, R., Rasheed, B. A., Keir, S., Nikolskaya, T., Nikolsky, Y., Busam, D. A., Tekleab, H., Diaz, L. A., Hartigan, J., Smith, D. R., Strausberg, R. L., Marie, S. K. N., Shinjo, S. M. O., Yan, H., Riggins, G. J., Bigner, D. D., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E. & Kinzler, K. W. (2008). An integrated genomic analysis of human glioblastoma multiforme, *Science* 321(5897): 1807–1812.
URL: *http://www.sciencemag.org/content/321/5897/1807.abstract*

Pinkel, D. & Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer, *Nat Genet* .

Rennstam, K., Ahlstedt-Soini, M., Baldetorp, B., Bendahl, P.-O., Borg, Å., Karhu, R., Tanner, M., Tirkkonen, M. & Isola, J. (2003). Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. a study of 305 tumors by comparative genomic hybridization, *Cancer Research* 63(24): 8861–8868.
URL: *http://cancerres.aacrjournals.org/content/63/24/8861.abstract*

Rueda, O. M. & Díaz-Uriarte, R. (2007). Flexible and accurate detection of genomic copy-number changes from acgh, *PLoS Comput Biol* 3(6): e122.
URL: *http://dx.plos.org/10.1371%2Fjournal.pcbi.0030122*



Aneuploidy in Health and Disease

Edited by Dr Zuzana Storchova

ISBN 978-953-51-0608-1
Hard cover, 244 pages
Publisher InTech
Published online 16, May, 2012
Published in print edition May, 2012

Aneuploidy means any karyotype that is not euploid, anything that stands outside the norm. Two particular characteristics make the research of aneuploidy challenging. First, it is often hard to distinguish what is a cause and what is a consequence. Secondly, aneuploidy is often associated with a persistent defect in maintenance of genome stability. Thus, working with aneuploid, unstable cells means analyzing an ever changing creature and capturing the features that persist. In the book Aneuploidy in Health and Disease we summarize the recent advances in understanding the causes and consequences of aneuploidy and its link to human pathologies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Lingling Zheng and Joseph Lucas (2012). Uncover Cancer Genomics by Jointly Analysing Aneuploidy and Gene Expression, Aneuploidy in Health and Disease, Dr Zuzana Storchova (Ed.), ISBN: 978-953-51-0608-1, InTech, Available from: http://www.intechopen.com/books/aneuploidy-in-health-and-disease/joint-analysis-of-aneuploidy-and-gene-expression

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen