Integrative genomic and survival analysis of breast tumors

Szilárd Nemes

Department of Oncology
Institute of Clinical Sciences
Sahlgrenska Academy at University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2012

Integrative genomic and survival analysis of breast tumors

nemes.szilard@oc.gu.se

ISBN 978-91-628-8538-0

Printed in Gothenburg, Sweden 2012
Ale Tryckteam AB, Bohus

Live long and prosper!

Integrative genomic and survival analysis of breast tumors

Szilárd Nemes

Department of Oncology, Institute of Clinical Sciences
Sahlgrenska Academy at University of Gothenburg
Gothenburg, Sweden

ABSTRACT

With the continued accumulation of genomic data at ever increasing resolution, the challenge ahead lies in reading out meaningful clinical/biological information from the data that can contribute to a better understanding of the cancerous process. The need for novel approaches and new statistical methods is therefore strong.

The present thesis aims to contribute to the field with three problem-specific applications that hopefully will aid researchers in a better understanding of genomic data.

The first paper exemplifies the adaptation of a piecewise-linear regression framework for integrative analysis of DNA copy number aberrations and gene expression (mRNA) data. The method helps to identify the association between copy number and gene expression, but it takes a further step and allows detection of changing patterns and changepoints that could serve as a proxy for the degree of genomic instability that causes disruptions in feedback-mechanisms.

The second paper advocates the adaptation of a mediation analysis for a concomitant analysis of DNA copy number aberrations, mRNA and survival data. The paper offers ways of statistical inference by means of the Delta method applicable concomitantly on a large number of genes. If a mediation effect is observed for a specific gene, we hypothesize that the specific gene is a driver gene. If no mediation effect is observed, possible associations between DNA copy number aberrations and the outcome are likely to indicate passenger genes.

The third paper is a more applied/clinical work using applied statistics which identified a novel panel of 12-genes that can serve as a prognostic tool for breast cancer specific survival.

The thesis concludes with a methodological description in which we describe an easy permutation-based approach for testing the clonal origins of multiple tumors. The main assumption of the proposed method is that if two tumors that share a common origin, or if the alleged secondary tumor is clonally related to the primary tumor, they share a higher and tumor-specific amount of matching chromosomal aberrations (gains or deletions) than recurrent chromosomal aberrations can explain.

Keywords: DNA copy number aberrations, messenger-RNA, breast cancer, regression, survival analysis, mediation, permutations

ISBN: 978-91-628-8538-0

SAMMANFATTNING PÅ SVENSKA

Med den fortsatta ackumuleringen av genetiska data med allt högre upplösning ligger utmaningen framför allt i att extrahera meningsfull klinisk/biologisk information som kan bidra till en bättre förståelse av cancer. Behovet av nya tekniker och statistiska metoder är därför stort.

Denna avhandling syftar till att bidra till fältet med tre problemspecifika applikationer som förhoppningsvis kommer att hjälpa forskare till en bättre förståelse av genetiska data.

Den första artikeln ger exempel på användbarheten av en styckvis linjär regression för analys av avvikelser i antal DNA-kopior och genuttryck (messenger-RNA). Metoden hjälper till att identifiera sambandet mellan antalet DNA-kopior och genuttryck och tar ytterligare ett steg och tillåter detektion av förändrade genuttrycksmönster och ombytespunkter som kan fungera som en proxy för graden av genomisk instabilitet som orsakar störningar i feedback-mekanismer.

Den andra artikeln förespråkar en medieringsanalys för en samtidig analys av avvikelser i antal DNA-kopior, mRNA och överlevnadsdata. Detta delarbete presenterar en Delta metoden baserat statistiskt test för medieringseffekt som tillämpas parallellt på ett stort antal gener. Om en medieringseffekt observeras för en specifik gen, antar vi att den specifika genen är en driver-gen. Om ingen mediering observeras kommer det möjliga sambandet mellan antalet DNA-kopior och överlevnad sannolikt att indikera passagerargener.

Den tredje artikeln är ett mer tillämpat/kliniskt arbete som har identifierat en ny panel av 12-gener som kan tjäna som prognostiskt verktyg för bröstcancerspecifik överlevnad.

Avhandlingen avslutas med ett metodologiskt delarbete, där vi beskriver en enkel permutationstes för att undersöka det klonala ursprunget till multipla tumörer. Det huvudsakliga antagandet i den föreslagna metoden är att, om två tumörer som delar ett gemensamt ursprung eller om den påstådda sekundära tumören är klonalt relaterad till den primära tumören, de delar ett högre antal matchande kromosomavvikelser än vad återkommande kromosomavvikelser kan förklara.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Nemes Sz., Parris T, Danielsson A, Kannius-Janson M, Jonasson JM, Steineck G, Helou K. Segmented regression, a versatile tool to analyze mRNA levels in relation to DNA copy number aberrations. *Genes, Chromosomes and Cancer*. 2012, 51(1): 77-82.

II. Nemes Sz, Parris TP, Danielsson A, Einbeigi Z, Steineck G, Jonasson JM, and Helou K. Integrative genomics with mediation analysis in a survival context. (*Submitted*)

III. Nemes Sz, Parris TP, Danielsson A, Jonasson JM, Genell A, Karlsson P, Steineck G and Helou K. A novel 12-gene panel predicting clinical outcome of breast cancer. (*Submitted*)

IV. Nemes Sz, Danielsson A, Parris TP, Jonasson JM, Karlsson P, Steineck G and Helou K. Permutation test for the clonal origins of multiple tumors. (*Manuscript*)

CONTENTS

1 BACKGROUND

Chromosomal aberrations such as DNA losses (deletions), gains (duplications and amplifications), translocations, inversions or other forms of structural rearrangements have a major impact on tumor initiation and development. These types of genetic alterations often affect whole chromosomes, chromosome arms, and specific chromosome regions. The majority of these alterations may affect specific genes involved in key cellular pathways influencing patient clinical outcome and resistance to current treatment regimens. However, the biological mechanisms by which altered genes contribute to cancer pathophysiology and patient survival have not yet been fully elucidated.

Genomic screenings are an efficient way to portray the global state of tumors and provide a comprehensive overview of this complex heterogeneous and polygenous illness. Furthermore, this method can pinpoint specific chromosomal changes that characterize tumors. With the continuously accumulating published array-comparative genomic hybridization (array-CGH) and gene expression data, the challenge we face is to understand and find an efficient way to extract and summarize key biological information that can serve as novel prognostic markers and therapeutic targets. In this thesis I aim to contribute statistical tools applicable in integrative genomic settings. Specifically I look into how to integrate the two biological levels, DNA and RNA, and seek to provide insights into this complex relationship. Furthermore, I wish to ascertain the extent to which changes at the DNA and RNA levels manifest themselves in patients' survival status.

1.1 Cancer Etiology and Development

Cancer, an illness of modern times, is one of the most important causes of human death. It is often regarded as a single disease, while in reality it is a complex of diseases affecting different organs. Tumors are not simply an aggregation of clonal cancer cells but "abnormal organs" of multiple cell types and extracellular matrix [1]. Development and progression of tumors is a long step-wise process and may even take several years depending on the rate and type of specific mutations that accumulate in the cells [2]. Mutations may emerge as a result of external factors such as chemicals, radiations or viruses as well as internal factors such as hormones, immune system, and

inherited mutations. It is likely that the effects of external and internal factors are intertwined and act together to initiate and promote tumor initiation, progression, and development by inducing changes in the gene regulation process. Gene regulation includes the processes that cells use to regulate the way that the information encoded in DNA is turned into gene products. Although a functional gene product can be an RNA molecule, the majority of known biological mechanisms are regulated by protein coding genes. Any step of the gene's expression may be modulated, from DNA-RNA transcription to the post-translational modification of a protein with transcription rate being the prevalent regulatory point of gene expression [3]. It is especially important to recognize that transcription factors have biological functions related to the control of cell proliferation and differentiation. Two large classes of genes involved in carcinogenesis are (proto)oncogenes and tumor suppressors that often encode for transcription factors. The third class of genes with a prominent role in tumor initiation and progression is the class of caretaker or DNA repair genes involved in the detection of DNA-damages and activation of repair mechanisms and possibly inactivation of mutagenic molecules [4].

As Hanahan and Weinberg [5] noted in their seminal paper, normal cells evolve progressively towards a neoplastic state and acquire the characteristics that we call ‘hallmarks’ of cancer. The aforementioned paper and its 2011 reincarnation [6] describe six hallmarks that have both distinctive and complementary capabilities that enable tumor initiation, development, and metastatic dissemination. These hallmarks are sustained proliferate signaling, evasion of growth suppressors, replicative immortality, sustained angiogenesis, evasion of apoptosis and activation of invasion and metastasis. Of these six hallmarks, the first five are a common feature of both benign and malignant tumors, while the sixth solely characterizes malignant solid tumors [7]. Tumors are routinely classified as malignant or benign, and beyond that they are traditionally classified on the basis of the tissue of origin, e.g. epithelial carcinomas (breast, prostate, lung or colon cancer), mesenchymal sarcomas (fat, muscle and bone) or cancer types like leukemia, lymphomas, and hematopoietic cell cancers affecting the central nervous system.

Cancer may originate from one single somatic cell, but tumor progression results from the accumulation of genetic alterations within the original clone allowing a multistep clonal expansion of more aggressive cells. These cells ultimately acquire the capability of invasion and metastases and by means of the circulatory system spread inside the organ of origin, or to other organs. Metastasizing cells may form new tumors which sometimes can appear with a substantial time lag. Tumors may even redevelop from dormant cell clusters

left behind after an operation and subsequent therapy. Naturally, multiple tumors can develop independently from each other.

1.2 Data for Integrative Genomics

Efficient data management and data analysis constitute perhaps the greatest challenges in integrative genomic studies. The unprecedented amount of data alone confronts researchers with daunting tasks and the nature of the data adds an extra level of complexity.

Cells of normal tissues typically contain two copies of DNA material, that is two copies of each gene. Biologists generally consider four different categories: *i)* loss with less than two copies of DNA; *ii)* normal with exactly two copies of DNA; *iii)* gains with three or four copies of DNA and *iv)* amplifications with more than four copies of a DNA segment. These changes are routinely measured by genome wide screening methods such as array-comparative genomic hybridization (aCGH) [8]. Array-CGH measurements are continuous by nature and lack direct interpretation, and they represent the relative amount of genetic material of neoplastic cells compared to the normal genetic material extracted from x with a healthy tissue as reference. Similarly, gene expression is measured by expression microarrays that provide a continuous reading which is proportional to the true amount of messenger RNA (mRNA) present in tumor cells [9]. Patterns of gene expression are mostly described by two stages, down- or up-regulation. However, as DNA CNA and mRNA measurements are made on different platforms, matching the two is the most important preprocessing step of integrative genomic analysis. The most common difficulty researchers encounter is the differences in resolution between DNA and gene expression arrays. The Array-CGH platform uses artificial DNA constructs, Bacterial Artificial Chromosomes or BAC-clones, that characteristically cover several genes. Moreover, adjacent BAC-clones overlap; consequently the same chromosome fragment might be covered by two (or more) BAC clones, a BAC-clone can contain several genes and a gene can be covered by two or more BAC clones. Up to this point we lack well established and widely accepted procedures for matching the measurements from the two biological levels, though this represents an area of interest and systematic efforts have been undertaken to provide standardized procedures [10].

1.3 Statistics for Integrative Genomics

In this section we will cover the statistical aspects of the present thesis. As a rule, more attention will be paid to details about the methods at the core of papers I, II and IV, while methods for paper III, which represents a more applied/biological orientation, will be addressed only briefly and the reader will be referred to the relevant literature.

As the title suggests, our main goal is to integrate the two biological levels and to expose their effect on patient survival. To this end we commence with a brief review of the methodologies applied in integrative genomics. Thereafter, we describe a regression-based versatile approach for modeling the DNA copy number aberrations and mRNA relationship. Following this, we outline an approach for a mediation analysis in a survival analysis setting, where we assume that the effect of DNA copy number aberrations on survival is mediated by mRNA. We conclude the methodological description with a brief review of the methods used for assessing the similarities/differences between multiple tumors.

Assume that we have preprocessed and matched data. For each patient, or each tumor, we have a pair of measurements $\{x_i, y_i\}_{i=1}^{n}$ with x_i denoting the copy number measurement and y_i the mRNA gene expression measurement for the respective gene. Moreover, for each patient we know the follow-up time t_i and their survival status, δ_i.

A large variety of statistical methods have been employed in the integrative analysis of DNA copy number aberrations and mRNA levels with a preponderance for correlation [11-17] and regression analysis [18-21]. Other studies commenced with the determination of DNA copy number aberrations and changes in mRNA expression separately and then matched the located aberrations together to determine if aberrations in mRNA levels follow DNA copy number aberrations [22-25]. This two-step analysis occasionally is augmented with an assessment of relationship strength [21-24, 26]. Schäfer merged these two approaches and derived a modified correlation coefficient to measure equally directed derivations of CNA and mRNA from the median values in the reference samples [27].

Analyses employing correlation- or regression-based measures generally assume a simple linear relationship between DNA copy number aberrations and mRNA levels. However, this might not always be the case, and small scale changes in DNA CNA can result in unproportional changes in gene expression [28]. Moreover, genes displaying a linear DCN-mRNA

relationship in cancerous cells can be associated with substantially different biological processes from genes displaying a nonlinear relationship [29].

In the next step we integrate DNA copy number aberrations and mRNA levels with the survival status of the patients. Survival (overall, disease-specific or distant disease-free survival) is the natural endpoint for most cancer studies and has been used in countless studies. However, generally the effect of few chosen markers (DNA copy number aberrations, mRNA or protein levels) of survival is studied over time with the help of Cox-regression or survival plots. We are unaware of any efforts to model a DNA copy number aberrations - mRNA-survival pathway based on a biologically plausible model.

1.3.1 Regression analysis

Regression analysis in general assumes that the response Y (mRNA in our case) can be modeled as a function of the predictors X (DNA copy number aberrations here), with the general form for the model

$$Y = f(X) + \varepsilon$$

where f is an unknown function and ε is a mean zero error. Integrative genomics usually assumes a simple linear relationship, namely

$$Y = \alpha + \beta X + \varepsilon$$

where α represents the intercept and β the regression slope. Interpretation of α represents the intercept and β in classical analysis says that α is the value of the response when the predictor takes the value zero, while β represents the change in the response following a one-unit change in the predictor. Keeping in mind that both mRNA and DNA copy number aberrations measurements are $\log_2$ratios, it is easy to see that X will be zero only if the amount of genetic material in the cancerous cells equals the genetic material in the healthy cells used for normalization. Thus, α represents the amount of mRNA that cancerous cells would contain without chromosomal aberrations. Similarly, β denotes a unit change on the $\log_2$ratio scale. An increase from zero to one assumes double amounts of DNA in the cancerous cells compared to the healthy cells (approximately four copies), while an increase from one to two indicates a fourfold increase in the number of copies. The model parameters, α and β are estimated by minimizing the sum of squared errors

$$RSS = \sum (Y - \alpha - \beta X)^2$$

or by using the so called normal equations, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{X}$ is the predictor matrix and $\mathbf{y}$ is the response vector.

The aforementioned study by Solvang *et al* [29] introduced a second order term in the regression equation

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

which offers a more nuanced depiction of the relationship between DNA copy number aberrations and mRNA; however, the assumed structural form is somewhat restrictive. A positive second order term (β_2) assumes an initial decrease in mRNA expression with accumulation of gene copies followed by a rapid increase. A negative second order term assumes a rapid initial increase in mRNA expression with accumulation of gene copies followed by a rapid decrease in expression when a threshold given by $-\beta_1 / 2\beta_2$ is passed. While this model can be plausible for a number of genes, it is safe to assume that it does not apply to all possible non-linear DNA copy number aberrations-mRNA relationships. Naturally, one could consider adding even more higher-order terms to the equation

$$Y = \alpha + \sum_{k=1}^{m} \beta_k X^k + \varepsilon .$$

However, this can lead to a serious over-fitting, and a high number of possible models complicates model selection. Estimation of regression coefficients by penalized least-squares will shrink non-important terms towards zero, though the estimated coefficients are largely biased and lack direct interpretability.

So far we have only considered *f* to be a smooth and continuous function. In the following section we explore the idea of approximating a smooth and continuous function by a piecewise linear regression analysis.

1.3.2 Segmented regression

Segmented regression (also known as piecewise linear regression splines or two-phase regression) has a long history but with sparse application in genomics [18, 30]. Article **I** proposed the use of a framework based on segmented regression analysis to analyze DNA copy number aberrations-mRNA relationships. The proposed framework is aimed at describing patterns of the relationship between abnormally expressed genes due to aberrant DNA copy numbers, specifically to determine if the variation of gene expression pattern changes over the domain of DNA copy number aberrations. Statistically, this change in gene expression pattern is expressed as a change in regression slopes. The segmented regression framework

assumed the existence of one or more identifiable point(s) where the relationship between the number of DNA copy aberrations and mRNA levels (i.e., the slope of the regression line) changes.

Model formulation

Generally the literature recognizes two structural forms of segmented regression equations. The difference between the two is whether the regression equations are built with a continuity constraint or if they are disjoint. Here we only consider segmented regression with continuity constraint, which assumes that the model parameters are estimated under the restriction of $\alpha^{(k)} + \beta_1^{(k)}\tau = \alpha^{(k+1)} + \beta_1^{(k+1)}\tau$ for the change-points τ and segments $k = 1,...,K$.

For a given chromosome fragment we denote with $x_{i,j}$ the $\log_2$ratio normalized DNA copy number aberration measurement at probe j for individual i and we let $y_{i,j}$ denote the corresponding $\log_2$ratio normalized mRNA measurement. We assume that the pairs $\{x_i, y_i\}_{i=1}^n$ are ordered so that $x_{1,j} \le ... \le x_{n,j}$. For each probe j, we build a linear model for the relationship between DNA copy number aberration and relative mRNA levels

$$y_{i,j} = \alpha_j + \beta_j x_{i,j} + \varepsilon_{i,j}$$

where for any given j, $\varepsilon_{i,j}$ are independent and identically distributed normal errors with mean zero. We assume that for some probes the linear model is not adequate, and we approximate the unknown smooth and continuous non-linear function

$$Y_{i,j} = f(\beta_j X_{i,j}) + \varepsilon_{i,j}$$

by a sequence of joined linear sub-models

$$\begin{aligned} y_{i,j} &= \alpha_j + \beta_{j1} x_{i,j} + \delta_{j1}(x_{i,j} - \tau_{j1})^+ + ... + \delta_{jk}(x_{i,j} - \tau_{jk})^+ + \varepsilon_{i,j} \\ &= \mu_{i,j} + \varepsilon_{i,j} \end{aligned}$$

where $(x_{i,j} - \tau_{j.})^+ = (x_{i,j} - \tau_{j.})$ for $(x_{i,j} - \tau_{j.}) > 0$, τ_k's are unknown change-points and $\delta_{j,l} = \beta_{j,l} - \beta_{j,l-1}$.

Parameter estimation

The model coefficients, $\theta = (\alpha_j, \beta_{j.1,}, \delta_{j1}...\delta_{jk})$ are estimated by minimizing the residual Sum of Squares,

$$RSS = \sum_{i=1}^{n} \left(y_{i,j} - \mu_{i,j} \right)^2 .$$

Direct minimization of the RSS for segmented regression is not possible due to the existence of numerous local minima. The location of change-points, τ, can be set based on empirical knowledge and treated as known, in which case the estimation of model parameters is straightforward. If we cannot set a change-point without a reasonable doubt, we have to calculate it from the data. Estimating the change-point from the data is nontrivial and requires numerical optimization.

Commonly applied optimization routines cannot be used due to the numerous local minima that might occur [31]. Lerman's grid search is one of the first methods developed for parameter estimation [32] and is the method of choice of article **I**.

The grid search is a stochastic search belonging to the class of exploratory Monte Carlo optimization methods. The general solution would be to explore the entire space for $\tau(T)$ by simulating points over T according to an arbitrary distribution J, positive everywhere on T until a sufficient value of $RSS(\tau)$ is observed. In practice J is a uniform distribution over the domain T. Given a uniform distribution $u_1,...,u_m \sim U_T$ we use $RSS_m^* = \min(RSS(u_1),...,RSS(u_m))$ as an approximation to the solution of *RSS*. For the change-point of the segmented regression the domain T consists of the possible values that the $\log_2$ratio normalized CNA measurements can take. Instead of using a uniform distribution ranging between the minimum and maximum values of the $\log_2$ratio normalized CNA measurements we directly use $x_{i,j}$, thus $RSS(\tau) = \min(RSS(x_{1,j}),...,RSS(x_{n,j}))$. In summary, the change-point corresponds to the observed $\log_2$ratio normalized CNA measurement that gives the lowest possible RSS. An alternative to the Lerman's grid search is Hudson's continuous fitting algorithm [33]. Despite better asymptotic properties of the regression coefficients estimated by Hudson's continuous fitting algorithm [34] we chose to adopt Lerman's grid search at lower computational costs[35].

Without loss of generality we now refer to one change-point problem and illustrate the details of parameter estimation. The residual sum of squares (RSS) for the two-segment regression equation for probe j will be

$$RSS_j = \sum_{i=1}^{n} \left\{ (y_{i,j} - \alpha_{j,L} - \beta_{j,L} x_{i,j})^2 I(x_{i,j} \le \tau) + (y_{i,j} - \alpha_{j,R} - \beta_{j,R} x_{i,j})^2 I(x_{i,j} \ge \tau) \right\}$$

where $I(x_{i,j} \le \tau)$ and $I(x_{i,j} \ge \tau)$ are indicator functions that take the value 1 if the condition is met, otherwise 0.

We defined the following design matrix for the segmented regression with parameters $\boldsymbol{\beta} = (\alpha_{j,R}, \beta_{j,L}, \beta_{j,R})$ with the indices L and R denoting if the segment positions were on the left or right side of the estimated change-point

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,j} - \tau & \tau \\ 1 & x_{2,j} - \tau & \tau \\ . & . & . \\ 1 & x_{t,j} - \tau & \tau \\ 1 & 0 & x_{t+1,j} \\ . & . & . \\ 1 & 0 & x_{n-1,j} \\ 1 & 0 & x_{n,j} \end{pmatrix}$$

and $\mathbf{Y}_1 = \left(y_{1,j},...,y_{n,j}\right)^T$. The parameters are estimated as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ with $\alpha_{j,L} = \alpha_{j,R} + \tau(\beta_{j,R} - \beta_{j,L})$.

Feder [36] and Hinkley [37, 38] offer guidelines for statistical inference for the regression parameters of the segmented model. The approach advocated by Hinkley uses the standard errors computed without imposing the continuity constraint. Feder proposes deleting observations around the estimated change-point. For Lerman's grid search this corresponds to deleting the observations equal to the estimated changepoint(s). Statistical inference for the constrained regression parameters relies on the consistency of the estimates, which implies that the constrained regression parameters are distributed asymptotically like the unconstrained ones. The variance of the slopes of the constrained regression is

$$\sigma^2_{\beta_{j,L}} = \frac{\sigma^2}{C_{xx,\tau}} \text{ and } \sigma^2_{\beta_{j,R}} = \frac{\sigma^2}{C^*_{xx,\tau}}$$

where σ^2 is the variance of the unconstrained regression equation and $C_{xx,\tau}$ and $C^*_{xx,\tau}$ are the corrected sums of squares of the predictor in the segments

$$C_{xx,\tau} = \sum_i \left(x_{i,j} - \overline{x}_{.,j}\right)^2 I(x_{i,j} < \hat{\tau}) \text{ and } C^*_{xx,\tau} = \sum_i \left(x_{i,j} - \overline{x}_{.,j}\right)^2 I(x_{i,j} > \hat{\tau}) .$$

Bai [39], Bai and Perron [40], Liu et al [41] and Kim & Kim [42] have proved the consistency of the change-point estimator for both the unconstrained and constrained case. For the constrained regression the change-point has an asymptotic normal distribution, for the unconstrained regression the change-point involves a step function with unknown distribution form [42]. Confidence intervals for the constrained changepoint

can be built using the likelihood ratio statistic [37] based on the residual sum of squares and the F distribution as

$$\left\{\tau : RRS(\tau) \le RSS(\hat{\tau})\left(1+\frac{r-1}{n-p}F^{1-\alpha}_{r-1,n-p}\right)\right\}$$

with r denoting the number of segments, n the sample size and p the number of estimated parameters.

Theoretically, the grid search can be applied on models with multiple change-points, however the computational cost and the data requirements increase rapidly making the method unfeasible for genomic studies. Recently, an alternative approach emerged. Muggeo [43] proposed a re-parameterization of the changepoint that facilitates a straightforward iterative estimation. Moreover, simulation studies had shown that when the regression lines are continuous the algorithm proposed by Muggeo is superior to the alternatives [44].

1.3.3 Survival analysis

Researchers often augment their findings with the addition of a clinical endpoint, frequently survival status of the patients. This can result in study-to-study differences where survival status can refer to cancer-specific survival, overall survival or distant disease-free survival. Independently of the outcome, the most common tool of analysis is Kaplan-Meier curves. Kaplan-Meier curves offer a vivid descriptive depiction of the survivor status, assuming discretized aCGH or gene expression readings. Combining aCGH and gene expression readings in one prognostic factor is not straightforward, but it can be achieved with multivariate network analyses [45]. However Kaplan-Meier curves have considerable limitations that can be addressed with regression analysis.

The most common tool of choice for survival analysis is the Proportional Hazard Regression, or Cox-regression. Results of the Proportional Hazard Regression are summarized by calculating hazard ratios and associated confidence intervals. A hazard ratio greater than one indicates that a gene (or other marker) is positively associated with the event probability and is negatively associated with survival time.

The Proportional Hazards model assumes that the covariates of interest act multiplicatively on the baseline hazard as follows

$$\alpha(t \mid X) = \alpha_0(t)\exp\left\{\sum_{j=1}^{p}\beta_j X_j\right\}$$

where β is a $p \times 1$ vector of unknown parameters and $\alpha_0(t)$ is an arbitrary non-negative function, the baseline hazard. The difficulty in applying Proportional Hazards Regression to genomic data lies in the high dimensionality of the data. Classical model selection techniques such as stepwise selection or even Bayesian methods cannot cope with the setting when $p \gg n$. In order to reduce the set of predictors to manageable levels it is possible to fit a series of univariate models and retain the markers that show significance after adjustment for multiple testing and a multivariate model is fit on the preselected variables. Iterative Bayesian Model Averaging, a possible alternative, iterates through the predictor set in a fixed order. For each subset the Bayesian Model Averaging procedure retains variables with posterior probability greater than 0.5 [46]. The disregarded variables are replaced by new ones until the procedure iterated through the whole data set. Regularized/penalized regression methods have the advantage of being able to deal with high dimensionality. Panelized regression models shrink all regression coefficients towards zero and depending on the penalty exactly to zero, thus concomitantly performing estimation and variable selection. Regression parameters are estimated and defined in terms of penalized likelihood optimization $\hat{\beta} = \arg\max\{l(\beta) - P_\lambda(\beta)\}$ where $l(\beta)$ is the log-likelihood and the penalty term $P_\lambda(\beta) = \lambda|\beta|^m$ with $m \geq 1$ denoting the vector norm of the regression coefficients. Applicability of the penalized Proportional Hazards Regression with different penalty definitions has proved their feasibility but there is no general consensus concerning the optimal penalty term [47-49]. Moreover, optimization of penalized regression models can be done with respect to the global predictive power [50, 51]. Here we chose to apply the elastic-net with

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + (1-\alpha)|\beta|^2 \right)$$

and selected the optimal value for the penalty parameter with cross-validation [52, 53].

From a clinical-practical point of view penalized regression has the disadvantage of making largely biased estimates that make statistical inference meaningless and practical interpretation equivocal. Thus we used the elastic net as a diagnostic tool to detect estimation problems due to multicollinearity in the data.

The predictive power of the models can be assessed as time dependent Area Under the Receiver Operatic Characteristic Curves (AUC(t)) and summarized by the concordance index (C-index) [54]. Predictive power can be validated by ten-fold cross-validation and the 0.632 bootstrap [55].

Mediation and Aalen's Additive Model

Mediation analysis aims to identify and describe the structural form that underlines an observed relationship between an independent variable and the outcome by the inclusion of a third variable, the mediator. Mediation exists if the independent variable changes the mediator, and change in mediator is followed by change in the outcome when the independent variable is present [56]. Mediation explicitly assumes that the variables form a causal chain, and the mediator variable serves to clarify the nature of the relationship between the independent and outcome variables. Thus, the mediator accounts partially or totally for the relation between the independent variable and outcome, and the total effect of the independent variable on the outcome can be decomposed into effects due to mediated paths and effects due to non-mediated paths [57].

Decomposition of Cox-regression estimates into direct and mediated effects lack any straightforward analytical expression and there are no general measures for a mediated effect [58]. Lack of possibilities for decomposition implies that Proportional Hazard Regression can model the effect of DNA copy number aberrations and mRNA reading belonging to a gene as two independent predictors. If the effect of DNA copy number aberrations on survival is mediated by mRNA and there is no direct effect, a Cox-regression model will miss this effect and it will conclude that DNA copy number aberrations have no effect on survival status. In opposition to the Proportional Hazard Regression the model proposed by Aalen, the Additive Model [59], can be decomposed into direct and mediated effects. Aalen's Additive Model assumes that the covariates add additively on the hazard

$$\alpha(t \mid \mathbf{x}_i) = \beta_0(t) + \sum_j \beta_j(t) x_{i,j}(t)$$

Here, $\beta_0(t)$ is the baseline hazard and has similar interpretation to the intercept of any regression equation, namely the hazard rate of an individual when all covariates equal zero. The coefficients $\beta_0(t)$ represent the increase in the hazard at time *t* corresponding to a unit increase in the j^{th} covariate; x_{ij} denotes the value for the j^{th} covariate for the i^{th} patient. It might be hard to give intuitive biologically meaningful interpretations to the regression coefficients, but using the properties of hazards and survival functions allows for direct transformation to survival probabilities.

The survival function can be expressed in terms of the hazard as

$$S(t) = \exp\left\{-\int_0^t \alpha(u)du\right\}.$$

Substitution of the hazard with the Additive model formulation leads to

$$S(t) = \exp\left\{-\int_0^t \beta_0(u) + \sum_j \beta_j(u) x_{i,j}(u) du\right\}.$$

If we restrict our attention to one single covariate then we have

$$S(t) = \exp\left\{-\int_0^t \beta_0(u) + \beta_1(u) x(u) du\right\}.$$

Now β_1 equals zero, then the equation simplifies to

$$S(t) = \exp\left\{-\int_0^t \beta_0(u) du\right\} \text{ and } \exp\left\{-\int_0^t \beta(u) x(u) du\right\}$$

will be the excess probability of the event due to one unit increase in the covariate. Intuitively, $1{,}000 \times (S(t \mid \beta_i \neq 0) - S(t \mid \beta_i = 0))$ will be the expected number of patients in a group of 1,000 that experiences the event due to the studied covariate.

Returning to the problem at hand, the Additive Models coefficients can be interpreted as excess mortality due to one unit change on the DNA copy number aberrations or mRNA measurement scale.

Yet another attractive feature of the Additive Model is the possibilities of decomposition of the total effect into direct and mediated effect, while decomposition of Cox-regression estimates into direct and mediated effects lacks any straightforward analytical expression and there are no general measures for a mediated effect [58].

We assumed that mRNA levels are explained to a certain degree by DNA copy number aberrations and their relationship can be depicted as

$$mRNA = \alpha_0 + \alpha_m DCNA + \varepsilon$$

where ε is i.i.d. mean zero normally distributed noise with variance σ^2. This type of regression analysis has long served as an exploratory tool in integrative genomic analysis [21]. Furthermore we modeled the effect of DNA copy number aberrations and mRNA levels on the hazard as

$$\alpha(t \mid \mathbf{x}_i) = \beta_0 + \lambda_m mRNA + \lambda_c DCNA$$

where λ_m is the effect of the mediator on the hazard (in our case mRNA levels) while λ_c is the effect of the covariate on the hazard (in our case DNA copy number aberrations). Simplifying these two equations leads to

$$\begin{aligned}\alpha(t \mid \mathbf{x}_i) &= \beta_0 + \lambda_m mRNA + \lambda_c DCNA \\ &= \beta_0 + \lambda_m(\alpha_0 + \alpha_m DCNA) + \lambda_c DCNA \\ &= \beta_0 + \lambda_m \alpha_0 + DCNA(\alpha_m \lambda_m + \lambda_c)\end{aligned}$$

where $\alpha_m \lambda_m$ is the effect of DNA copy number aberrations on survival status mediated through mRNA, while $\alpha_m \lambda_m + \lambda_c$ is the total effect of DNA copy number aberrations on survival status. The residuals, ε, were omitted as they have an expected value of zero.

Consequently, the effect of DNA copy number aberrations on survival status is decomposed in a direct effect λ_c (natural direct effect or pure direct effect) and the effect mediated by mRNA $\alpha_m \lambda_m$ (natural indirect effect or pure indirect effect). Testing λ_c and λ_m is straightforward and it is based on the martingale central limit theorem. Inference for the indirect effect can be based on Normal Product Distribution [60] or on asymptotic results based on the multivariate Delta method [61]. Preacher and Hayes provide a comprehensive review of the subject [62]. The Delta method provides a framework for establishing the asymptotic distribution of a differentiable function, and we propose an asymptotic statistical inference based on it [63].

The Delta method

The Delta method is a method for deriving an approximate probability distribution for a function of an asymptotically normal statistical estimator from knowledge of the limiting variance of that estimator. If $\sqrt{n}(x - \mu_x) \xrightarrow{L} N(0, \sigma_x^2)$ then for a given function $f(x)$ with existing first-order derivative $\sqrt{n}[f(x) - f(\mu_x)] \xrightarrow{L} N(0, \sigma_x^2[f'(\mu_x)]^2)$, assuming that $f'(\mu_x)$ exists and it is non-zero [64]. The Delta method applies a Taylor expansion to linearize a non-linear relationship. If a function $f(x)$ has derivatives of order k, then for a constant a the Taylor series of order k about a is

$$T_n(x) = \sum_{j=0}^{k} \frac{f^{(j)}(a)}{j!}(x-a)^j \ .$$

Generally, the statistical literature and practical applications are interested mainly in the first order Taylor expansion and to a lesser extent in the second order expansion.

A second order expansion of $f(x)$ around μ_x gives

$$f(x) = f(\mu_x) + f'(x - \mu_x) + \frac{1}{2} f''(x - \mu_x)^2 + R_{j \geq 3}(x)$$

where the reminder $R_{j \geq 3}(x) = (x - \mu_x)^j f^{(j)}(\xi) / j!$ with $\xi \in (x, \mu_x)$ rapidly converges to zero. Following the notation of Preacher et al [65] we define the following parameters

1. $\hat{\boldsymbol{\theta}}$ a column vector of regression coefficients used in the estimation of the mediated effect
2. $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ the expected values of the regression coefficients, $\boldsymbol{\mu}_{\boldsymbol{\theta}} = E\left[\hat{\boldsymbol{\theta}}\right]$
3. $f\left(\hat{\boldsymbol{\theta}}\right)$ the effect of interest, the estimator for the mediation effect
4. $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$ the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$
5. $\mathbf{D} = \partial_{\theta} f\left(\hat{\boldsymbol{\theta}}\right)$ the first order derivatives of $f\left(\hat{\boldsymbol{\theta}}\right)$ evaluated at $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, the Jacobian matrix of $f\left(\hat{\boldsymbol{\theta}}\right)$
6. $\mathbf{H} = \partial^2_{\theta} f\left(\hat{\boldsymbol{\theta}}\right)$ the Hessian matrix of $f\left(\hat{\boldsymbol{\theta}}\right)$ evaluated at $\boldsymbol{\mu}_{\boldsymbol{\theta}}$

The Delta method based variance is defined as

$$Var[f(\hat{\boldsymbol{\theta}})] \approx E\left[f(\hat{\boldsymbol{\theta}})^2\right] - \left(E\left[f(\hat{\boldsymbol{\theta}})\right]\right)^2.$$

By the Taylor theorem we have

$$f\left(\hat{\boldsymbol{\theta}}\right) \approx f\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right) + \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\boldsymbol{\theta}}\right)\mathbf{D} + \frac{1}{2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\boldsymbol{\theta}}\right)^2 \mathbf{H}$$

Without explicitly going through the algebra we give

$$E\left[f(\hat{\boldsymbol{\theta}})^2\right] = f^2\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right) + \mathbf{D}^T\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{D} + \frac{1}{4}\left(tr(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}))\right)^2$$

$$+\frac{1}{2}\left(tr(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}))^2\right) + f\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)tr(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}))$$

and

$$E\left[f(\hat{\boldsymbol{\theta}})^2\right] = f^2\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right) + f^2\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)tr\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right) + \frac{1}{4}\left\{tr\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)\right\}^2,$$

consequently $Var\left(f(\hat{\boldsymbol{\theta}})\right) = \mathbf{D}^T\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{D} + \frac{1}{2}tr\left\{\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right\}$

Variance estimator and inference for the mediated effect

As noted above $\alpha_m\lambda_m$ is the effect of DNA copy number aberrations on survival status mediated through mRNA, while $\alpha_m\lambda_m + \lambda_c$ is the total effect of DNA copy number aberrations on survival status. Using the above outlined notation we have that $\hat{\boldsymbol{\theta}} = \left[\hat{\alpha}_m, \hat{\lambda}_m\right]^T$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \left[\alpha_m, \lambda_m\right]^T$ and $f\left(\hat{\boldsymbol{\theta}}\right) = \hat{\alpha}_m\hat{\lambda}_m$. The gradient matrix of $f\left(\hat{\boldsymbol{\theta}}\right)$ is $\mathbf{D} = \partial_{\theta} f\left(\hat{\boldsymbol{\theta}}\right)\Big|_{\mu}$ and the Hessian matrix equals to

$$\mathbf{H} = \begin{pmatrix} \partial^2_{\alpha_m\alpha_m} f\left(\hat{\boldsymbol{\theta}}\right) & \partial^2_{\alpha_m\lambda_m} f\left(\hat{\boldsymbol{\theta}}\right) \\ \partial^2_{\alpha_m\lambda_m} f\left(\hat{\boldsymbol{\theta}}\right) & \partial^2_{\lambda_m\lambda_m} f\left(\hat{\boldsymbol{\theta}}\right) \end{pmatrix}\Bigg|_{\mu}.$$

As α_m and λ_m are independent from each other the estimator for the covariance matrix is

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2_{\alpha_m} & 0 \\ 0 & \sigma^2_{\lambda_m} \end{pmatrix}$$ while the Hessian, $$\mathbf{H} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$.

Plugging in the estimators for the mediation effect into the algorithm of the Delta method leads to the following variance estimator for the mediation parameter

$$\begin{aligned} \sigma^2_{Med} &= \mathbf{D}^T \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{D} + \frac{1}{2} tr\left\{\left(\mathbf{H}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\right)^2\right\} \\ &= \underbrace{\alpha^2_m \sigma^2_{\lambda_m} + \lambda^2_m \sigma^2_{\alpha_m}}_{first-order} + \underbrace{\sigma^2_{\lambda_m}\sigma^2_{\alpha_m}}_{second\text{-}order} \end{aligned}.$$

The second order term is often omitted with the implicit assumption that it is small compared with the first order term [61]. The total effect is defined as $\alpha_m\lambda_m + \lambda_c$, a summation of the mediated and direct effect. Here, we can take advantage of the properties of variances, namely $\sigma^2_{Tot} = \sigma^2_{\lambda_c} + \sigma^2_{Med} + 2\sigma_{\lambda_c, Med}$, however the Delta method leads to the same variance estimator. Under mild regularity conditions, (λ_c, λ_m) are normally distributed and independent from α_m, thus $\sigma_{\lambda_c, Med} = \alpha_m \sigma_{\lambda_m \lambda_c}$ leading to a variance estimator for the total effect of $\sigma^2_{Tot} = \sigma^2_{\lambda_c} + \alpha^2_m \sigma^2_{\lambda_m} + \lambda^2_m \sigma^2_{\alpha_m} + 2\alpha_m \sigma_{\lambda_m \lambda_c}$.

Approximate confidence intervals for mediation effect are calculated as $(\alpha_m\lambda_m - Z_{\alpha/2}\sigma_{Med};\ \alpha_m\lambda_m + Z_{\alpha/2}\sigma_{Med})$. The procedure is similar for the total effect and ratio, just with the suitable changes. Confidence intervals for the direct effect can be obtained in a similar way based on the output of the Aalen's Additive model.

As it is not a straightforward matter to know how to adjust confidence intervals for multiple testing, we need to calculate P-values. We test the null hypothesis of no effect $H_0:\ \alpha_m\lambda_m = 0$ against the alternative $H_1:\ \alpha_m\lambda_m \neq 0$. Based on the approximately normal distribution of the estimates we can calculate a test statistics, Z-score as $\alpha_m\lambda_m / \sigma_{Med}$, with $Z \sim N(0,1)$. Inference for the total effect proceeds with the same steps, while the inference for the direct effect is provided by the Aalen's Additive model.

1.3.4 Statistics of clonal origins

The statistical challenge of clonal relatedness of two tumors is yet to be solved [66]. The complex nature of genomic data together with the dependency of two tumors belonging to the same patient poses considerable difficulties. Not only is it true that the data from two tumors belonging to the same patient are not independent, but the markers themselves tend to be correlated with each other. Independence of two events assumes that the probability of one is the same whether the other is given or not. While it is clear that two markers from the same chromosome cannot be considered

independent, the status of two markers from separate chromosomes is open to debate. Currently, studies of clonal origins assume that somatic changes occurring on different chromosomes are independent events. As we see it, this assumption is violated and one cannot ascertain without reasonable doubt the independence of markers. We argue that somatic changes of different chromosomes can be dependent, conditionally independent or random independent somatic changes. One cannot exclude the occurrence or random deletions or amplification, in which case markers affected by the observed somatic change will be independent from others. Additionally somatic changes of different chromosomes can be conditionally independent when their occurrence traces back to a common biological process and their initiation and development do not directly affect each other. Furthermore, markers of different chromosomes can be causally linked to each other when specific somatic changes trigger genomic events that cause further somatic changes on the same or different chromosomes propagating the genomic instabilities that characterize cancer cells. It is likely that when researchers consider markers from the whole genome spread on different chromosomes (whether one or a few markers per chromosome) they will have to deal with an array of complex relationships between markers ranging from independence to casual relationships. The effect of violations of the independence assumption on test of clonal origins is yet to be elucidated. To circumvent this problem researchers have refrained from using full genomic profiles and have restricted their attention to specific markers, ordinarily being the most characteristic aberrations per chromosome [67, 68] or single selected markers from each chromosome [69, 70]. This approach not only reduces the multidimensional data to a few values but it assumes that readings from different chromosomes are independent. This assumption might be violated in cancer cells. Gains, deletions, and rearrangements (translocations) of DNA segments from different chromosomes develop concomitantly in cancer cells, though these aberrations may not be causally related. Additionally, several recurrent aberrations are frequently co-identified in breast cancers such as those found on chromosome arms 1q/16p (gains and losses) and 8p/11q (losses and gains). The question is whether these coexisting aberrations should be classified as one event instead of two events. This makes the designation of a single characteristic aberration per chromosome arm difficult. Pre-selection of markers is rather subjective and will certainly influence the results.

AIMS

The main aim of this work was to describe and formalize three statistical approaches that were inspired by statistical analyses undertaken by the PhD candidate in previous efforts.

Specifically, the first goal was to describe a regression analysis-based approach for the integrative genomic analysis of DNA copy number aberrations and messenger RNA levels. The goal was not only to establish the association between the two biological levels but to describe the pattern of relationship between the two.

Having done this, the aim was to augment the analysis of the two biological levels with a clinically relevant endpoint, survival time. Herein we aimed to offer a statistical framework applicable in a genome-wide setting to assess the possible mediation when the effect of DNA copy number aberrations on survival is mediated by messenger-RNA. This depicts mathematically a biologically plausible model.

A third aim of the present work was to identify a novel panel of gene expression signatures predicting breast cancer-specific survival.

The last goal differed somewhat from the previous three. In this paper we considered only DNA and we aimed to present a framework that facilitates making inferences about the clonal origins of tumor pairs.

2 PATIENTS AND METHODS

2.1 Patients and genomic data

Primary invasive tumors (n = 141) from 141 breast cancer patients (Table 1) were selected from the fresh-frozen tissue tumor bank at the Sahlgrenska University Hospital Oncology Lab (Gothenburg, Sweden) [71-73]. All samples were assessed for DNA content at the time of diagnosis from 1991 to 1999 (data not shown) by flow cytometry at the Laboratory for Clinical Chemistry, Sahlgrenska University Hospital. The presence of malignant cells was assessed in all samples by evaluation of touch preparation imprints stained with May-Grünwald Giemsa (Chemicon). All procedures were done in accordance with the Declaration of Helsinki and approved by the Medical Faculty Research Ethics Committee (Gothenburg, Sweden).

Array-CGH and Gene expression analysis

Whole-genome tiling arrays with 38,043 reporters mapping to the UCSC May 2004 hg17: NCBI Build 35 were manufactured as previously described [74] at the SCIBLU Genomics DNA Microarray Resource Center (SCIBLU), Department of Oncology, Lund University. Images and raw signal intensities were acquired using an Agilent G2505B DNA microarray scanner (Agilent Technologies) and GenePix Pro 6.0.1.22 (Axon Instruments) image analysis software. Data preprocessing and normalization were done using the web-based BioArray Software Environment system (BASE) provided by SCIBLU (http://base2.thep.lu.se/onk/).

The RNA samples were processed at SCIBLU using Illumina HumanHT-12 Whole-Genome Expression BeadChips (Illumina), according to the manufacturer's instructions. The expression microarrays contained approximately 49,000 probes representing > 25,400 RefSeq (Build 36.2, Release 22) and Unigene (Build 199) annotated genes. Images and raw signal intensities were acquired using the Illumina BeadArray Reader scanner and BeadScan 3.5.31.17122 (Illumina) image analysis software, respectively.

Data preprocessing and quantile normalization were applied to the raw signal intensities using BASE. Further data processing was done in Nexus Expression 2.0 (BioDiscovery) using $\log_2$-transformed, normalized expression values and a variance filter. Normalized values from five normal breast samples profiled with Illumina HumanWG-6 Expression Beadchips (GEO, accession number GSE17072) were used as reference [75]. Further details the reader will find in Parris *et al* [71] and its supplementary material.

Table 1. Clinical and pathological characteristics of the 141 breast cancer patients studied

	Survivors	Deceased
Follow-up time	11 yrs	4 yrs
Mean Age at diagnosis	61.5	57.0
Tumor size	32.04 mm	32.38 mm
Tumor status		
T1	21	15
T2	29	32
T3	15	16
T4	1	4
Missing	6	2
Histology		
Ductal	53	45
Lobular	6	7
Other	7	13
Missing	6	4
ER status		
Positive	63	47
Negative	9	21
Missing	-	1
PR status		
Positive	52	40
Negative	20	28
Missing	-	1
HER2 status		
Positive	63	60
Negative	9	9
Molecular subtype		
Luminal A	1	0
Luminal B	64	45
HER2/ER-like	5	11
Normal	0	0
Basal-like	2	13
Surgery		
Lumpectomy	23	21
Mastectomy	36	36
Non	10	8
Missing	3	4
Hormonal treatment		
Yes	11	20
No	10	12
Missing	51	37
Radiotherapy		
Yes	36	36
No	20	20
Missing	16	13
Chemotherapy		
Yes	38	39
No	17	18
Missing	17	12

2.2 Simulation studies

In paper two we proposed an inference method for the mediation analysis in survival context. To assess the small sample behavior of the proposed method, we ran a Monte Carlo simulation. Through this Monte Carlo study we assessed the coverage probability of 95 % CIs at different sample sizes and compared the performance of the proposed approach to inferential procedures based on normal product distribution and non-parametric bootstrapping.

First, we designed a study to estimate the coverage probability of the proposed 95% confidence intervals [76]. To this end, we simulated data sets of sample sizes varying between 100 and 1000 with increment 50. For each sample size we simulated 1000 data sets. In a second simulation we generated 1000 samples of size 500. Based on these 1000 samples we constructed confidence intervals based on the Delta method, product normal distribution, and Monte Carlo confidence intervals, and for every sample we ran 1000 nonparametric bootstrap resampling and constructed confidence intervals based on normal approximation, the base bootstrap, the percentile method, and Bias Corrected and Accelerated (BCa) bootstrap. Confidence intervals were compared for coverage probability and width aiming to find the narrowest confidence interval with coverage closest to the nominal 95%.

Based on a previous analysis of 97 tumors [77] we estimated the mean $\log_2$ratio values for DNA copy number aberrations at $\mu = 0.248$ and $\sigma^2 = 0.047$. Thus, we simulated the DNA copy number aberrations as normally distributed at $\mu = 0.248$ and $\sigma^2 = 0.047$. Furthermore, based on the same data we generated the relative mRNA log2ratio values by $-0.578 + 0.775\text{DCNA} + \varepsilon$ with $\varepsilon \sim N(0, 0.158)$. Survival times were generated according to the additive hazard model with $\alpha(t \mid \mathbf{x}_i) = \beta_0 + \lambda_m mRNA + \lambda_c DCNA$, where $\lambda_m = 0.3$, $\lambda_c = 0.1$, leading to an indirect effect of 0.31 and a total effect of 0.41. The baseline β_0 was set to 1 and censoring was chosen to 0.9 to obtain a censoring around 60%, relevant for cancer studies. The coverage probability of the 95% CI was estimated as the proportion of confidence intervals covering the true values. Given that for every sample size we run 1000 simulations we would expect that the coverage lies between 0.936 – 0.9635. Lower or higher values indicate systematic under- or over-coverage.

3 RESULTS AND DISCUSSIONS

3.1 Segmented regression

The segmented regression approach proposed in paper **I** was intended to serve as an aid or enhancement. The versatility of the method arises from the possibility of a linear approximation to a large number of possible non-linear relationships. Naturally, it is desirable to find the true form of the structural relationship between DNA copy number aberrations and mRNA (or as close to the truth as possible). This is likely to be feasible if we restrict our attention to a single gene or marker. However, extended visual inspection or mathematical analysis of the DNA copy number aberrations-mRNA relationship for every gene is not practical or possible. A priori assumptions about the structural form of the relationship can lead to discovery of relevant biological/medical phenomena [29]; nevertheless this imposes a rigidity in assumptions that are likely to be violated. On the other hand, the increase in flexibility by adopting a segmented regression approach could easily lead to overfitting and could seriously impair any generalizations. Article **I** did not fully acknowledge these issues. The authors do test if a two-line segmented regression is better at describing the data than a simple one-line regression line; however the validity of the proposed approach is not fully explored. As noted by the authors, the testing procedure was a compromise due to limitation in computing power. Clearly significance testing based on permutations is the most appealing approach [78, 79]. However, adaptation of a permutation-based significance testing is hindered by the need for testing a large number of genes simultaneously. This latter aspect raises the question of multiple testing as well. Article **I** notes that based on biological reasoning we can expect two change-points, one on the limit between deletions and normal-like profiles and one between normal-like profiles and gains. So it is not only that we have to establish that mRNA-levels are related to DNA copy number aberrations but we have to decide if this relationship is best described by linear, one change-point or two change-point regression lines. This results in making it necessary for us to conduct six null-hypothesis tests for every gene, or four if we test sequentially in a predefined order from the simplest to the more complex relationships. This not only reduces statistical power, but adds an extra layer of complexity to an already complex problem.

3.1.1 Information Criteria for segmented regression

There is a vast body of literature on model selection procedures with Information Criterions (IC) occupying a prominent role. Information

Criterions trade-off fit and model complexity and the idea of parsimony suggest that one should prefer the least complex solution, and the IC offer mathematical justification for this major problem in the philosophy of science [80]. The two best known Information Criterions are the Aakike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The theoretical motivations behind the AIC and BIC are distinct [81]. AIC does not assume the existence of a true model. Models are by definition only approximations to the unknown and unidentifiable truth and AIC searches for the best approximation. In contrast with this approach, BIC compares the probability of each model assumed to be the true model that generated our observations.

Jones [82] used a modified version of the AIC for selection of change points. Liu et al [41] proposed a framework based on a modified version of the Bayesian Information Criterion. The authors [41] note that there is no reasonable way of picking the best information criterion other than that the one they proposed may also give consistent parameter estimators. We chose here to discuss the Akaike's Information Criterion as a model selection tool. AIC can be estimated as $AIC = -2\log L(\hat{\theta}) + 2p$, where $L(\hat{\theta})$ is the likelihood and p the number of parameters in the model. AIC has been reported to find the "true" model more reliably than F-test [83, 84]. The model with lowest AIC value is considered to be closest to the unknown truth. If the sample size is small, one could consider further penalizing complex models to avoid overfitting [85]. For small sample sizes (e.g. $n / p <\sim 40$) AIC should be corrected as

$$AIC_C = AIC + \frac{2(p+1)(p+2)}{n-k-2}.$$

The correction term rapidly converges to zero with increasing sample size. Model selection proceeds with direct comparison of the estimated AIC values and the model with the lowest is preferred. The drawback of AIC and Information Criterions in general is the lack of a straightforward universal interpretation and a proper scale with easily interpretable values. The lack of scale makes it hard to get insight into just how much statistical importance we can attach to a difference in AIC between two models. Raw comparison of AIC values does not provide sufficient evidence in favor of the chosen model. If the models considered have almost equal AIC values, raw comparison becomes even more difficult. In this case it is attractive to calculate the Akaike weights that serve as estimates for the conditional probabilities for each model. First we estimate the differences in AIC between models

$$\Delta_i(AIC) = AIC_i - \min(AIC)$$

Based on the results of Bozdogan [86] we can estimate the relative likelihood of model i given the data as

$$L(M_i \mid \text{data}) \propto \exp\left\{-0.5\Delta_i(AIC)\right\}.$$

The normalized relative likelihoods function as Akaike weights (w_i)

$$w_i(AIC) = \frac{\exp\left\{-0.5\Delta_i(AIC)\right\}}{\sum_k \exp\left\{-0.5\Delta_k(AIC)\right\}}$$

so that $\sum_i w_i(AIC) = 1$. The interpretation is straightforward, the probability that the chosen model best describes the data given the candidate models considered. Dividing the Akaike weights of two competing models gives the strength of the evidence for choosing one model over the other [87, 88].

AIC will not give a degree of belief about the model's truthfulness; it merely gives us an objective tool we can use to compare the degree to which the data support the various models we wish to consider. This immediately highlights a practical issue that needs to be addressed before Information Criterions can be applied in integrative genomics. The classical null-hypothesis testing procedures have clear-cut widely accepted limits (significance levels). These significance levels have well known frequentist properties. While calculating Akaike weights facilitates direct comparison of competing models there are no guidelines for thresholds that would help us determine how much better a more complex model has to be in order to be preferred over the simple ones. Moreover, it is not a straightforward matter to determine how to deal with multiple testing. The initial version of the article **I** used AIC as model selection tool. However, this approach was dropped in later stages due to the two concerns mentioned above. Nevertheless, the applicability of ICs for segmented regression in integrative genomics settings will be revisited in the more or less distant future. A recent paper by Leday revisited the idea of segmented regression as an exploratory tool for mRNA DNA copy number aberrations relationship with Information Criterions based model selection, reinforcing the idea of better feasibility of Information Criterions as selection tools over null-hypothesis testing [89].

3.1.2 Biological meaning of the change-point(s)

In article **I** we argue that the identified change-points and the pattern of relationship between DNA copy number and gene expression can generate additional hypotheses. To be specific, investigating what causes swift

changes in mRNA production, what causes over-production of mRNA, or the leveling out of the mRNA levels in spite of accelerated DNA accumulation in tumors could generate relevant research questions. Tumors characteristically consist of a mixture of cells, cells with aberrations in particular chromosome fragments, cells with high degrees of genomic instability and normal cells without copy number aberrations. In some cases, the change-point could simply divide the sample in subsets of tumors with high levels of normal cells without copy number aberrations and tumors with substantial copy number aberrations. As Chen *et al.* [90] demonstrated, segmented regression can simultaneously classify observations and provide statistical inference. More importantly, change-points could represent the degree of genomic instability that causes disruptions in negative-feedback mechanisms. Defects in feedback mechanisms might enhance proliferate signaling, thus inducing an ever increasing gene expression [6].

3.1.3 Application to breast cancer

Multiple analysis

We applied a segmented regression analysis of 1161 chromosome fragments from 97 tumors from the study of Parris *et al* [71]. Of the 1161 chromosome fragments examined after multiple adjustments, 341 showed significant associations between DNA copy number aberrations and relative mRNA levels. For 269 of the 341 significant relationships (78%), addition of change-point and subsequent segmented regression provided no genuine improvement over linear regression, while for the remaining 72 chromosome segments the two-segment regression had a significantly better fit. For 59/72 chromosome fragments (82%), we observed an initial increase in mRNA levels due to changes in DNA copy number aberrations. After the change-point was passed, the mRNA levels reached a plateau and a further increase in DNA copy numbers did not induce further elevation in mRNA levels. For 12 chromosome fragments, the change-point marked the point where mRNA production accelerated and accumulation was faster than DNA levels suggested.

The case of gene *HDGF*

The *HDGF* gene encodes a member of the hepatoma-derived growth factor family. *HDGF* increases the tumorigenic, mitogenic and angiogenic activity of a variety of cancer cells [91] and participates in the pathogenesis of breast cancer by promoting cell growth [92]. Overexpression of *HDGF* mRNA levels has been observed in nasopharyngeal carcinomas [93]. The mean $\log_2$ratio of the 1161 chromosome fragments we analyzed was –0.082, suggesting that neoplastic cells had mRNA levels similar to normal cells.

Compared with that the $\log_2$ratio relative mRNA levels for the *HDGF* gene was –0.954, indicating a down-regulation. The $\log_2$ratio relative mRNA levels for the *HDGF* gene were partially explained by copy number aberrations. We fitted three working models, the null model with only a single-line linear regression and a piecewise-linear regression with two segments. We observed that both the single-line linear regression and a piecewise-linear regression with two segments significantly improved fit (F-test, p= 0.0000006 and p= 0.0000005). Moreover the two-segment regression line represented an improvement over the single-line regression (P-val=0.03). To gain a better insight into the appropriateness of the two-segment regression line over the competing models we calculated the Akaike weights. It turned out that with a probability of 0.789 the two segment-regression best describes the data. The single-line regression achieved a posterior probability of 0.21while the intercept model had a posterior probability close to zero.

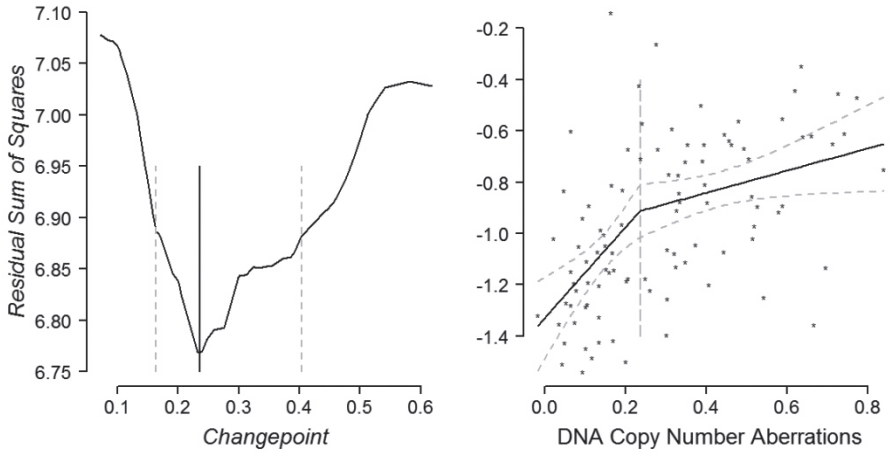

Figure 1. DNA copy number and messenger-RNA relationship for gene HDGF depicted by a two-segment regression equation. The image on the left side illustrates the Residual sum of squares over the domain of the change-point and the estimated change-point with the associated 95% Confidence Interval. The right side image depicts the changes in the regression line.

Using the algorithms outlined in a previous section we estimated the change-point and the regression parameters. We observed that the expression velocity changes over 0.232 on the normalized copy number scale (**Figure** 1). Below that limit, one unit increase on the normalized copy number scale resulted in

1.76 unit increase of mRNA levels (p= 0.0003). After passing the change-point the mRNA accumulation slows down; one unit increase on the normalized copy number scale resulted in 0.432 unit increase of mRNA levels (p= 0.04). True, DNA copy number aberrations explain only 29% of the variation in mRNA levels.

A non-parametric bootstrap analysis with 1000 resamplings with replacement revealed a small bias in the change-point estimation (0.004); however, the bias was 50-fold lower than the parameter estimate. The estimated standard error for the change-point was similar in magnitude, 0.006. Interestingly, confidence intervals based on bootstrapping were genuinely narrower than confidence intervals based on the *F*-distribution.

3.2 Mediation analysis

In this section we offer numerical results demonstrating the feasibility of the proposed inference algorithm. Additionally, we offer an extension for multiple mediator and multiple mediating pathways.

3.2.1 Properties of the proposed confidence interval

The coverage of the 95% confidence interval was close to the nominal value and was in between the acceptance limits at sample sizes as low as 100 (**Figure 2**). Coverage of 95% confidence interval based on the Delta method was similar to confidence intervals based on non-parametric bootstrapping, Monte Carlo simulation and on the normal product distribution (Table 2). Moreover, the confidence interval based on the Delta method was symmetrical; intervals that failed to cover the true population value fell roughly equally into the lower and upper tail of the distribution. However, the estimated statistical power was low and achieved the generally accepted level of 0.8 only at sample sizes around 300. True, the choices we made in this simulation study directly influence the estimated statistical power, and had we used other parameter values the power curve plotted in **Figure 3** might have had steeper (or perhaps lower) slopes. Nevertheless, low power generally characterizes confidence intervals based on the Delta method. Somewhat surprisingly, confidence interval width based on the Delta method was superior to confidence intervals based on the alternatives. Bootstrapping is thought to be superior to the Delta method [94], however this is not always the case [95] and bias-corrected and accelerated bootstrap are known to have elevated Type I errors with sample sizes under 500 [96].

Table 2. Numerical results for coverage probability and confidence interval width of different types of 95% confidence intervals at different sample size for the mediation effect.

	n=100		*n=200*		*n=300*		*n=400*		*n=500*	
	Cov	Width	Cov	Width	Cov	Width	Cov	Width	Cov	Width
Delta method	0.946	1.662	0.946	1.140	0.945	0.904	0.943	0.790	0.946	0.701
Normal Product	0.942	1.687	0.942	1.149	0.945	0.909	0.942	0.793	0.945	0.703
Monte Carlo	0.942	1.686	0.945	1.148	0.944	0.908	0.941	0.793	0.941	0.703
Bootstrap: Normal	0.969	1.812	0.948	1.178	0.950	0.927	0.950	0.804	0.950	0.710
Base	0.982	1.853	0.956	1.195	0.957	0.938	0.956	0.812	0.954	0.716
Percentile	0.934	1.853	0.943	1.195	0.940	0.938	0.943	0.812	0.939	0.716
BCa	0.945	1.843	0.943	1.193	0.939	0.938	0.948	0.811	0.941	0.715

Moreover, based on asymptotic expansions it is known that the Delta method and bootstrap estimator of variance coincide at least at first order terms [97].

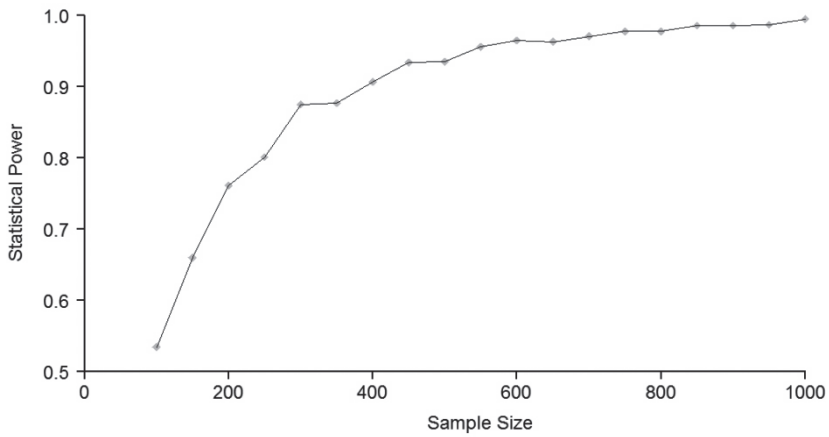

Figure 2. Evolution of the empirical statistical power as a function of the sample size.

We believe that the normality assumption for the $\log_2$ratio DNA copy number aberrations and mRNA data is plausible. However, deviation from this assumption is likely at least for a number of genes, given that we might concomitantly study up to around 20,000 genes. Consequently, we simulated DNA copy number aberrations and mRNA data using non-normal distribution (t-distribution, Weibull and log-normal). Apart from small variations that can be attributed to randomness, the behavior of the 95% confidence intervals was similar to the ones based on DNA copy number

aberrations and mRNA with normal distribution with slight under-coverage at sample sizes of 100 (data not shown). If the elements of the design matrix (in our case DNA copy number aberrations readings) are evenly spaced the overall error due to skewed distribution will be $O(n^{-1})$, thus as $n \to \infty$ normal approximation of the least squares regression slope estimates is appropriate [98]. Parameter estimation of Aalen's Additive Model also uses least squares, however generalization of the above described results is not straightforward. Nonetheless, the asymptotic efficiency of the derived estimators is expected.

3.2.2 Application to Breast Cancer

Regression analysis of breast cancer tumors showed that in 128 out of 8,349 chromosome fragments a significant DNA copy number aberrations -mRNA and subsequent mRNA-survival association exists. After adjusting for multiple testing, none of these 8,349 genes showed a significant mediation effect. If we had only tested the 128 genes with both DNA copy number aberrations -mRNA and mRNA-survival association, then all 128 genes would have shown significant mediation effects. Among these 128 genes, the mRNA levels for 124 genes mediated completely the effect of DNA copy number aberrations on survival, and no significant direct effect of DNA copy number aberrations on survival was recorded. For four fragments we observed that mRNA levels exhibited significant mediation effect but DNA copy number aberrations exerted an effect on survival that was not mediated by mRNA levels belonging to that specific fragment.

3.2.3 Extension to more than one mediator

In Article **II** we did not consider the case when more than one mediator or more than one mediation pathway is present. It is acknowledged that it is not mRNA but proteins that are the building blocks and functional elements of the human body. Thus, it would be natural to extend the DNA copy number aberrations-mRNA-survival pathway to DNA copy number aberrations-mRNA-protein-survival. Moreover, it is natural to expect that proteins (especially enzymes) might affect the expression of other genes that influence the functionality of other proteins.

The DNA copy number aberrations-mRNA-protein-survival pathway

The information stored in the DNA is transcribed to mRNA which in turn is translated to proteins. As a result it would be desirable to consider this full

pathway, but due to technical limitations protein data are rarely available for the researchers. Reliable genome-wide protein expression patterns are yet to be developed. The extension of the mediation analysis advocated in paper II for this pathway is straightforward. In agreement with the methodology previously described we assume that we modeled the effect of DNA copy number aberrations, mRNA and protein levels on the hazard as

$$\alpha(t \mid \mathbf{x}_i) = \beta_0 + \lambda_{m1} Prot + \lambda_{m1} mRNA + \lambda_c DCNA$$

where $\lambda_{m.}$ is the effect of the mediators on the hazard (in our case mRNA levels) while λ_c is the effect of the covariate on the hazard (in our case DNA copy number aberrations). Moreover we assume that levels are explained to a certain degree by DNA copy number aberrations and their relationship can be depicted as

$$mRNA = \alpha_0 + \alpha_m DCNA + \varepsilon_{mRNA}$$

and protein levels are explained to a certain degree by mRNA levels

$$Prot = \gamma_0 + \gamma_m mRNA + \varepsilon_{Prot}.$$

Simplifying these equations leads to

$$\alpha(t \mid \mathbf{x}_i) = \beta_0 + DCNA(\alpha_m(\lambda_{mRNA} + \lambda_{Prot}\gamma_m) + \lambda_{DCNA})$$

where $\alpha_m(\lambda_{mRNA} + \lambda_{Prot}\gamma_m)$ is the effect of DNA copy number aberrations on survival status mediated through mRNA and protein, while $\alpha_m(\lambda_{mRNA} + \lambda_{Prot}\gamma_m) + \lambda_{DCNA}$ is the total effect of DNA copy number aberrations on survival status. Applying the Delta method leads to a variance estimator for the mediated effect of

$$\sigma^2_{med} = \alpha_m^2 \sigma^2_{\lambda_{mRNA}} + \left(\lambda_{mRNA} + \lambda_{Prot}\gamma_m\right)^2 \sigma^2_{\alpha_m} + \left(\alpha_m \gamma_m\right)^2 \sigma^2_{Prot} + \left(\alpha_m \lambda_{Prot}\right)^2 \sigma^2_{\gamma_m}.$$

This assumed that mRNA might have an effect on survival that is not mediated through proteins. If we reject this assumption the estimator for the mediator effect simplifies to $\alpha_m \lambda_{Prot} \gamma_m$ and its variance will be

$$\left(\lambda_{Prot}\gamma_m\right)^2 \sigma^2_{\alpha_m} + \left(\alpha_m \gamma_m\right)^2 \sigma^2_{Prot} + \left(\alpha_m \lambda_{Prot}\right)^2 \sigma^2_{\gamma_m}.$$

Multiple mediating pathways

Following Fosen *et al* [99] we formalize the following notation. Let ψ_{hj} be the regression coefficient when X_j is regressed X_h and λ_j the Aalen's Additive regression coefficient when the survival status ($\mathrm{d}Y$) is regressed

onto X_j. The direct effect of X_j on the survival is represented by λ_j while the mediated effect is a path of length longer than one with one or more mediators . We assume that there are r indirect paths for X_h to dY , and denote them as P_i with $i = 1,...,r.$ Then the indirect path P_i takes the following form

$$P_i = \left\{ \left(X_{i_1}, X_{i_2} \right), \left(X_{i_2}, X_{i_3} \right), ..., \left(X_{i_{w_i}}, \mathrm{d}Y \right) \right\}$$

where w_i is the length of the path and the mediated effect will be

$$\sum_{i=1}^{r} \left\{ \sum_{j}^{w_i - 1} \left(\lambda_j \left(\prod_{l=1}^{w_i - 1} \psi_{i_l, i_{l+1}} \right) \right) \lambda_i \right\}.$$

Again, if we assume that all mediators but the last on the path have no effect on the survival the formula simplifies to

$$\sum_{i=1}^{r} \left(\prod_{l=1}^{w_i - 1} \psi_{i_l, i_{l+1}} \right) \lambda_i .$$

It is easy to see that this formulation is identical with the Dynamic Path models formulation without time depending effects. The variance estimator for the mediated effect will be

$$\sum_{i=1}^{r} \left\{ \left(\prod_{l=1}^{w_i - 1} \psi_{i_l, i_{l+1}} \right)^2 \sigma_{\lambda_i}^2 + \left(\lambda_i \prod_{l=1, l \neq k}^{w_i - 1} \psi_{i_l, i_{l+1}} \right)^2 \sigma_{\beta_k}^2 \right\}.$$

While we did not attempt to assess the plausibility of normal approximation for the multiple pathways or multiple mediators setting, we assume that the estimator would work, but we recognize that the simple size requirement will be substantial.

3.2.4 Mediation for ill-conditioned regression equations

Epidemiological literature routinely warns against adjusting regression equations for variables that are chains in the causal link, e.g. the mediator itself. From a mathematical point of view, the reasoning behind this recommendation is straightforward; if the independent variable causes the mediator then a regression equation with the independent and mediator as predictors for the outcome will be more or less ill-conditioned.

For the simple linear regression equation as we noted above the regression coefficients can be estimated by minimizing the residual sum of squares

$$\hat{\beta} = \arg\max \left\{ \sum_{i} \left(y_i - \sum_{j=1}^{p-1} x_{ji} \beta_j \right)^2 \right\}$$

or by solving the normal equations, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The least-squares estimates are not only the maximum-likelihood values for the parameter vector $\boldsymbol{\beta}$ but also the best linear unbiased estimators (BLUE). However, if the regression equation contains an independent variable and a mediator that are highly correlated to each other then $\mathbf{X}^T\mathbf{X}$ will have a determinant close to zero, making inversion imprecise and inflates the variance of the parameter estimates. If the mediator is a deterministic linear combination of the independent variable then $\mathbf{X}^T\mathbf{X}$ will be singular and there are no unique least-squares estimates of $\boldsymbol{\beta}$. If $\mathbf{X}^T\mathbf{X}$ is close to singular the problem is ill-conditioned and we have to deal with collinearity and variance inflation. A possible solution would be to relax the desire for an unbiased parameter estimate and use an estimator which is biased but has considerably smaller variance. Here we seek to minimize the mean squared error of the model. The mean squared error assesses an estimator in terms of its variation and unbiasedness, $MSE(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + E(\hat{\theta} - \theta)^2$ and for the ill-conditioned problems with inflated variance we seek a solution with lower MSE than the unbiased model, a bias-variance trade-off. We assume that inducing bias in parameter estimates yields a decrease in MSE, thus less variability in parameter estimates. Addition of a penalty term to the parameter estimates will inevitably shrink them towards zero and if the penalty term is properly used will lower their variability considerably. For the linear-regression equation the estimate will be

$$\hat{\beta} = \arg\max \left\{ \sum_i \left(y_i - \sum_{j=1}^{p-1} x_{ji}\beta_j \right)^2 \right\} + P_\lambda(\beta) .$$

One of the oldest and most frequently applied penalties is the ridge penalty,

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p-1} \beta_j^2 .$$

Addition of the penalty term will make possible the inversion of the $\mathbf{X}^T\mathbf{X}$ and $(\mathbf{X}^T\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ will have unique solutions.

So far we have focused on the linear regression problem; however in our case a survival regression equation will contain a pair of correlated predictors. Penalized estimation is applicable to the additive hazards model as well [100, 101]. While in certain high dimensionality problems penalized regression is a must in the frame of mediation analysis, its feasibility has not yet been fully explored. The main drawback of penalized regression methods lies in the biased estimates themselves. Bayesian formulation can produce standard errors [102] and as always computer intensive methods can provide frequentist standard errors as well. However, as the parameter estimates are

inevitably biased, any effort to get inference about the value of the unknown true parameter is futile.

Working with large scale genome-wide data will inevitably lead to situations when the independent variable (DNA copy number aberrations) and mediator (mRNA) will be highly correlated. In that case most likely the methodology outlined by article **II** will fail. The estimates of Aalen's additive model will have vastly inflated variances and this inevitably inflates the variance of the estimated mediated effect as well. Considering a penalized regression framework might be a path worth exploring in a mediation context.

3.2.5 Concluding remarks

The methodology presented in article **II** is largely based, and expands on, the current advances in mediation analysis in a survival context [99, 103-105] and adheres to the effort to infer a causal association between genes and disease [106-110]. The core biological assumption is that significant mediation effect ensures implication of DNA copy number aberrations in the progression to the outcome. The absence of mediation effects coupled with a significant mRNA-outcome relationship could indicate the presence of genes with dosage-independent effects and might guide researchers to other relevant biological processes. Similarly, the absence of a mediation effect coupled with the presence of a significant direct DNA copy number aberrations-outcome relationship indicates possible passenger genes whose effect depends upon the physical proximity of genuine driver genes. The aforementioned article does not discuss the issue of multiple testing, but simply mentions that the false discovery rate was controlled with Benjamini & Hochberg adjustment. This issue deserves more attention. The low empirical power in the simulation study and the fact that after multiple adjustments none of the 8,349 genes showed a significant mediation effect clearly indicates problems with statistical power. Decreased statistical power due to multiple test adjustment is a known phenomenon in genomic studies [111]. Our simulation results showed that the proposed inference based on the Delta method is equivalent to inferential procedures based on normal product distribution or re-sampling, the staple methods of inference in mediation studies [112, 113]. Thus, it seems likely that independently of the chosen testing method we would end up with a large number of false negative results.

3.3 A novel 12-gene predictive panel for breast cancer specific survival

3.3.1 Patients and data

Based on sequence homology, differential expression of 12,252 transcripts was matched to corresponding altered DNA segments generated from corresponding tumors for the 141 tumors. Our primary concern was gene expression and its relation to cancer specific survival. Copy number aberration data were only used for depicting the complex relationships among genes.

3.3.2 Results and interpretations

Univariate Cox-regression analysis identified 54 genes that can stratify patients into groups with a favorable or an unfavorable disease course. These 54 genes retained their significance and effect after adjustment for a set of clinicopathological variables (age at diagnosis, estrogen and progesterone receptor status, HER2 status, molecular subtypes, endocrine treatment, radiotherapy, and chemotherapy) suggesting an effect independent of prognostic factors and therapies. Due to the sample-size, the predictor-number multivariate model based on these 54 genes did not produce interpretable effect sizes. By means of Bayesian Model averaging, we reduced the 54 genes to a panel of 12 for the sake of efficiency and interpretability. Naturally this reduction resulted in loss of predictive power which, however proved to be marginal ($Z=0.75$, $P= 0.588$) and the two models had similar C-indexes (0.88 for the model with 54 genes and 0.83 for the model with 12 genes). The predictive power of the multivariate model based on the 12-gene signature was high, especially in the first five years following initial diagnosis, and remained at levels previously indicated by gene expression profiling [114, 115]. Internal validation showed that the predictive power of the model was stable.

3.3.3 Concluding remarks

Gene expression profiling is an efficient way to portray the global state of tumors and provides a comprehensive overview of this complex heterogeneous and polygenous nature of cancer and has led to important, but so far incremental and somewhat controversial clinical advancements [116, 117]. The results outlined above are not free from controversies or contradictions. We observed that univariate and multivariate predictive models indicated opposing effects for specific genes. Moreover, findings about *LETMD1* gene clearly contradict current knowledge. These issues are

not specific to the findings presented in paper **III** but they are more or less a general aspect of genomic studies. The 12-gene panel proposed in this paper will have to withstand external validation and will have to prove its worth against well-established gene signatures such as MammaPrint and Oncotype Dx. From the dawn of high-throughput data, hundreds if not thousands of gene signature panels have been proposed to stratify patients to predict clinical outcome. Many of these studies failed to have been validated on an external data set. Internal validation by the means of resampling or cross-validation is feasible as a means of assessing the accuracy gain compared with known clinical predictors; however, it is in any case desirable to validate the model on an independent data set [118]. This is more easily said than done, partly because of the unavailability of clinical data, non-overlapping gene sets and the instability of model building in this high-dimensional setting [119]. As a consequence, translating the proposed gene panel (or other published gene panels) to clinical praxis is not straightforward task.

Clinical-epidemiological thinking approaches validity from another angle but nevertheless its core requirements applies high-dimensional genomic settings. The hierarchical-step model by Steineck [120] identifies four key steps that influence the validity of a result. Confounding, the first of the four steps, is a likely candidate for problems in high-dimensional genomic studies. Both measured and unmeasured confounders are expected and controlling for them is nearly impossible. Misrepresentation is another source of error as patients are included more opportunistically than based on a rigorous selection procedure. Erroneous measurements, misclassified patients and misspecified prediction models further erode the validity.

3.4 Testing clonal origin

In the manuscript submitted as paper **IV** we present a combination of two well-proven and known strategies for testing the clonal origin of tumor pairs. The main assumption is that if two tumors share a common origin then they will both exhibit chromosomal aberrations at distinct locations. These are somatic changes such as deletions or gains of DNA material ranging from one single nucleotide mutation up to whole chromosomes [6]. This can be sporadic aberrations occurring in single patients' tumors or recurrent chromosomal aberrations that are nonrandom changes preferentially involving particular chromosomes [121]. Inevitably, due to these recurrent aberrations two tumors developing independently of each other will share common chromosomal aberrations. Metastases or clonal secondary tumors tend to develop from dominant cell populations of a primary tumor [122],

thereby inheriting not only recurrent chromosomal aberrations characteristic for specific types of solid tumors but also patient-specific chromosomal aberrations that develop during the evolution of neoplastic cell populations. Indeed, we cannot rule out that secondary tumors originate from one or more cells that deviate in genomic profile from the primary tumor. The heterogeneous nature of cancer explains why not all cells within a tumor mass harbor the same chromosomal aberrations [123].

With this in mind we heuristically derived a simple index that offers insights into the similarities of two tumors and developed a permutation based significance testing. The testing procedure assesses if two tumors exhibit a higher percentage of common aberration that might be explained by recurrent chromosomal aberrations or randomness.

4 SUMMARY AND CONCLUSIONS

Statistics, more specifically mathematical statistics and statistical inference, stand as a research field on its own and in addition serve as a backbone to many empirical studies. With the emergence of large-scale, genome-wide studies, researchers faced amounts of data and research settings that cannot be handled with classical statistics. Cancer genetics and genomics is a relatively young branch of biology which blurs the distinction between theoretical statistical research and applied statistics and acts as a driver for the refinement of old and development of new techniques. This thesis aims to contribute to the literature with three novel tools that might find usefulness and applicability in cancergenetic and genomic studies.

First, we proposed segmented regression analysis as a way to depict the possible complex DNA copy number aberration and messenger RNA relationships in an efficient way concomitantly for a large number of genes. This methodology enables researchers not only to model effectively possible non-linear DNA copy number aberration and messenger RNA relationships, but the detected changepoints can generate further hypotheses. Specifically, I can offer insights into the degree of genomic instability that induces altered gene expression patterns.

Second, we extended the integrative analysis of DNA copy number aberrations and messenger RNA with the inclusion of an endpoint with direct clinical interest. This paper mostly focuses on statistical/methodological aspects. Namely, it describes a computationally efficient null-hypothesis testing procedure based on the Delta method for the mediation effect. The method offers the possibility for researchers to weed out passenger genes whose copy number aberrations depend on the proximity of driver genes. Driver gene copy number aberrations ought to manifest in altered gene expression patterns and ultimately affect patient survival.

Third, we build a permutation-based procedure that might aid researchers who seek to elucidate the clonal origins of multiple tumors. This method is a computationally efficient straightforward way of testing tumor clonality that assumes that two clonal tumors have more

shared chromosomal aberrations than recurrent aberrations could explain.

The thesis concludes with an applied biological study and provides yet another gene signature for survival prediction. In this study we identified 12 genes that together can predict with a good predictive power breast-cancer specific survival up to 1ten years after diagnosis. This gene panel has the potential to stratify patients to favorable and non-favorable prognosis groups. However, as with other similar studies its applicability needs to be validated on an independent data set.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Khalil Helou for his guidance and encouragement during my years as a PhD student. I would like to thank my co-supervisor Gunnar Steineck for making possible my work and for the encouragements to explore ideas and thoughts outside my comfort zone. I am grateful for the friendship, support and weekly talks that we had with my co-supervisor Junmei Miao Jonasson.

This thesis could not be conceived without the help and support of all co-authors and I would like to thank you for sharing your expertise and thoughts with me.

I would like to thank all my co-workers and friends at the Division of Clinical Cancer Epidemiology, Regional Cancer Center West, Sahlgrenska Cancer Center and Oncolab at the Department of Oncology for your help and support, being that scientific or everyday routines.

I am thankful for all Romanian and Swedish tax payers for contributing to my education!

Last but not least I would like to my family, my mother Erzsébet and my sister Ágnes for their interest and support. My wife Lina showed great patience, support and understanding, I thank you for that!

REFERENCES

1. Egeblad M, Nakasone ES, Werb Z: **Tumors as Organs: Complex Tissues that Interface with the Entire Organism.** *Dev Cell* 2010, **18:**884-901.
2. Vogelstein B, Kinzler KW: **The multistep nature of cancer.** *Trends Genet* 1993, **9:**138-141.
3. Dolores Delgado M, León J: **Gene expression regulation and cancer.** *Clinical and Translational Oncology* 2006, **8:**780-787.
4. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nature Medicine* 2004, **10:**789-799.
5. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100:**57-70.
6. Hanahan D, Weinberg Robert A: **Hallmarks of Cancer: The Next Generation.** *Cell* 2011, **144:**646-674.
7. Lazebnik Y: **What are the hallmarks of cancer?** *Nat Rev Cancer* 2010, **10:**232-233.
8. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer.** *Nature Genet* 2005, **37:**S11-S17.
9. Quackenbush J: **Microarray data normalization and transformation.** *Nature Genet* 2002, **32:**496-501.
10. van Wieringen W, Unger K, Leday G, Krijgsman O, de Menezes R, Ylstra B, van de Wiel M: **Matching of array CGH and gene expression microarray features for the purpose of integrative genomic analyses.** *BMC Bioinformatics* 2012, **13:**80.
11. Green MR, Aya-Bonilla C, Gandhi MK, Lea RA, Wellwood J, Wood P, Marlton P, Griffiths LR: **Integrative genomic profiling reveals conserved genetic mechanisms for tumorigenesis in common entities of non-Hodgkin's lymphoma.** *Genes, Chromosomes and Cancer* 2011, **50:**313-326.
12. Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, Gwadry F, Jain A, Kouros-Mehr H, Fridlyand J, et al: **Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel.** *Mol Cancer Ther* 2006, **5:**853-867.
13. Gu W, Choi H, Ghosh D: **Global Associations between Copy Number and Transcript mRNA Microarray Data: An Empirical Study.** *Cancer Informatics* 2008, **2008:**17-23.
14. Horlings HM, Lai C, Nuyten DSA, Halfwerk H, Kristel P, van Beers E, Joosse SA, Klijn C, Nederlof PM, Reinders MJT, et al: **Integration of DNA Copy Number Alterations and Prognostic Gene Expression Signatures in Breast Cancer Patients.** *Clin Cancer Res* 2010, **16:**651-663.

15. Lee H, Kong SW, Park PJ: **Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes.** *Bioinformatics* 2008, **24:**889-896.
16. Salari K, Tibshirani R, Pollack JR: **DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data.** *Bioinformatics* 2010, **26:**414-416.
17. Tayrac Md, Etcheverry A, Aubry M, Saïkali S, Hamlat A, Quillien V, Treut AL, Galibert M-D, Mosser J: **Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression.** *Genes, Chromosomes and Cancer* 2009, **48:**55-68.
18. Menezes RX, Boetzer M, Sieswerda M, van Ommen GJB, Boer JM: **Integrated analysis of DNA copy number and gene expression microarray data using gene sets.** *BMC Bioinformatics* 2009, **10:**203.
19. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315:**848-853.
20. Peng J, Zhu J, Bergamaschi A, Han W, Noh DY, Pollack JR, Wang P: **Regularized multivariate regression for identifying master predictors with application to integrative genomic study of breast cancer.** *Ann Appl Stat* 2010, **4:**53-77.
21. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proc Natl Acad Sci U S A* 2002, **99:**12963-12968.
22. Bicciato S, Spinelli R, Zampieri M, Mangano E, Ferrari F, Beltrame L, Cifola I, Peano C, Solari A, Battaglia C: **A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets.** *Nucleic Acids Res* 2009, **37:**5057-5070.
23. Heidenblad M, Lindgren D, Veltman JA, Jonson T, Mahlamaki EH, Gorunova L, van Kessel AG, Schoenmakers E, Hoglund M: **Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic implications.** *Oncogene* 2005, **24:**1794-1801.
24. van Wieringen WN, Belien JAM, Vosse SJ, Achame EM, Ylstra B: **ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data.** *Bioinformatics* 2006, **22:**1919-1920.
25. Oudejans JJ, van Wieringen WN, Smeets SJ, Tijssen M, Vosse SJ, Meijer CJLM, Meijer GA, van de Wiel MA, Ylstra B: **Identification of genes putatively involved in the pathogenesis of diffuse large**

B-cell lymphomas by integrative genomics. *Genes, Chromosomes and Cancer* 2009, **48:**250-260.
26. Tsukamoto Y, Uchida T, Kaman S, Noguchi T, Nguyen LT, Tanigawa M, Takeuchi I, Matsuura K, Hijiya N, Nakada C, et al: **Genome-wide analysis of DNA copy number alterations and gene expression in gastric cancer.** *J Pathol* 2008, **216:**471-482.
27. Schafer M, Schwender H, Merk S, Haferlach C, Ickstadt K, Dugas M: **Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities.** *Bioinformatics* 2009, **25:**3228-3235.
28. Mileyko Y, Joh RI, Weitz JS: **Small-scale copy number variation and large-scale changes in gene expression.** *Proc Natl Acad Sci U S A* 2008, **105:**16659-16664.
29. Solvang H, Lingjaerde O, Frigessi A, Borresen-Dale A-L, Kristensen V: **Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer.** *BMC Bioinformatics* 2011, **12:**197.
30. Garcia-Alegria E, Ibanez B, Minguez M, Poch M, Valiente A, Sanz-Parra A, Martinez-Bouzas C, Beristain E, Tejada MI: **Analysis of FMR1 gene expression in female premutation carriers using robust segmented linear regression models.** *RNA-Publ RNA Soc* 2007, **13:**756-762.
31. Julious SA: **Inference and estimation in a changepoint regression problem.** *J R Stat Soc Ser D-Stat* 2001, **50:**51-61.
32. Lerman PM: **Fitting Segmented Regression Models by Grid Search.** *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1980, **29:**77-84.
33. Hudson DJ: **Fitting segmented curves whose join points have to be estimated.** *J Am Stat Assoc* 1966, **61:**1097-&.
34. Kim HJ, Yu B, Feuer EJ: **Inference in segmented line regression: a simulation study.** *J Stat Comput Simul* 2008, **78:**1087-1103.
35. Yu B, Barrett MJ, Kim H-J, Feuer EJ: **Estimating joinpoints in continuous time scale for multiple change-point models.** *Computational Statistics & Data Analysis* 2007, **51:**2420-2427.
36. Feder PI: **Asymptotic distribution theory in segmented regression problems- Identified case.** *Ann Stat* 1975, **3:**49-83.
37. Hinkley DV: **Inference in 2-phase regression.** *J Am Stat Assoc* 1971, **66:**736-743. .
38. Hinkley DV: **Inference about intersection in 2-phase regression.** *Biometrika* 1969, **56:**495-504.
39. Bai J: **Estimation of a change point in multiple regression models.** *Rev Econ Stat* 1997, **79:**551-563.
40. Bai J, Perron P: **Computation and analysis of multiple structural change models.** *J Appl Econom* 2003, **18:**1-22.

41. Liu J, Wu SY, Zidek JV: **On segmented multivariate regression.** *Stat Sin* 1997, **7:**497-525.
42. Kim J, Kim HJ: **Asymptotic results in segmented multiple regression.** *J Multivar Anal* 2008, **99:**2016-2038.
43. Muggeo VMR: **Estimating regression models with unknown break-points.** *Stat Med* 2003, **22:**3055-3071.
44. Chen CWS, Chan JSK, Gerlach R, Hsieh WYL: **A comparison of estimators for regression models with change points.** *Stat Comput* 2011, **21:**395-414.
45. Jornsten R, Abenius T, Kling T, Schmidt L, Johansson E, Nordling TEM, Nordlander B, Sander C, Gennemark P, Funa K, et al: **Network modeling of the transcriptional effects of copy number aberrations in glioblastoma.** *Mol Syst Biol* 2011, **7**.
46. Annest A, Bumgarner R, Raftery A, Yeung KY: **Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data.** *BMC Bioinformatics* 2009, **10:**72.
47. Waldron L, Pintilie M, Tsao M-S, Shepherd FA, Huttenhower C, Jurisica I: **Optimized application of penalized regression methods to diverse genomic data.** *Bioinformatics* 2011, **27:**3399-3406.
48. Gui J, Li H: **Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.** *Bioinformatics*, **21:**3001-3008.
49. van Wieringen WN, Kun D, Hampel R, Boulesteix A-L: **Survival prediction using gene expression data: A review and comparison.** *Computational Statistics & Data Analysis* 2009, **53:**1590-1603.
50. Liu Z, Magder L, Hyslop T, Mao L: **Survival associated pathway identification with group Lp penalized global AUC maximization.** *Algorithms for Molecular Biology* 2010, **5:**30.
51. Liu ZQ, Gartenhaus RB, Chen XW, Howell CD, Tan M: **Survival Prediction and Gene Identification with Penalized Global AUC Maximization.** *J Comput Biol* 2009, **16:**1661-1670.
52. Goeman JJ: **L1 Penalized Estimation in the Cox Proportional Hazards Model.** *Biometrical Journal* 2010, **52:**70-84.
53. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Stat Soc Ser B-Stat Methodol* 2005, **67:**301-320.
54. Heagerty PJ, Zheng Y: **Survival Model Predictive Accuracy and ROC Curves.** *Biometrics* 2005, **61:**92-105.
55. Efron B: **Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation.** *Journal of the American Statistical Association* 1983, **78:**316-331.
56. VanderWeele TJ: **Subtleties of explanatory language: what is meant by "mediation"?** *Eur J Epidemiol* 2011, **26:**343-346.

57. Suzuki E, Yamamoto E, Tsuda T: **Identification of operating mediation and mechanism in the sufficient-component cause framework.** *Eur J Epidemiol* 2011, **26:**347-357.
58. VanderWeele TJ: **Causal Mediation Analysis With Survival Data.** *Epidemiology* 2011, **22:**582-585.
59. Aalen OO: **Further results on the nonparametric Linear-regression model in survival analysis.** *Stat Med* 1993, **12:**1569-1588.
60. Aroian LA: **The probability function of the product of 2 normally distributed variables.** *Annals of Mathematical Statistics* 1947, **18:**265-271.
61. Sobel ME: **Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models.** *Sociological Methodology* 1982, **13:**290-312.
62. Preacher KJ, Hayes AF: **Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models.** *Behav Res Methods* 2008, **40:**879-891.
63. Oehlert GW: **A Note on the Delta Method.** *The American Statistician* 1992, **46:**27-29.
64. Casella G, Berger RL: **Statistical Inference, 2nd ed.** 2002.
65. Preacher KJ, Rucker DD, Hayes AF: **Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions.** *Multivariate Behavioral Research* 2007, **42:**185-227.
66. Ostrovnaya I, Begg CB: **Testing Clonal Relatedness of Tumors Using Array Comparative Genomic Hybridization: A Statistical Challenge.** *Clin Cancer Res* 2010, **16:**1358-1367.
67. Ostrovnaya I: **Testing clonality of three and more tumors using their loss of heterozygosity profiles.** *Stat Appl Genet Mol Biol* 2012, **11**.
68. Ostrovnaya I, Olshen AB, Seshan VE, Orlow I, Albertson DG, Begg CB: **A metastasis or a second independent cancer? Evaluating the clonal origin of tumors using array copy number data.** *Stat Med* 2010, **29:**1608-1621.
69. Begg CB, Eng KH, Hummer AJ: **Statistical tests for clonality.** *Biometrics* 2007, **63:**522-530.
70. Ostrovnaya I, Seshan VE, Begg CB: **Comparison of Properties of Tests for Assessing Tumor Clonality.** *Biometrics* 2008, **64:**1018-1022.
71. Parris TZ, Danielsson A, Nemes S, Kovács A, Delle U, Fallenius G, Möllerström E, Karlsson P, Helou K: **Clinical Implications of Gene Dosage and Gene Expression Patterns in Diploid Breast Carcinoma.** *Clin Cancer Res* 2010, **16:**3860-3874.
72. Möllerström E, Delle U, Danielsson A, Parris T, Olsson B, Karlsson P, Helou K: **High-resolution genomic profiling to predict 10-year**

overall survival in node-negative breast cancer. *Cancer genetics and cytogenetics* 2010, **198:**79-89.

73. Karlsson E, Delle U, Danielsson A, Olsson B, Abel F, Karlsson P, Helou K: **Gene expression variation to predict 10-year survival in lymph-node-negative breast cancer.** *BMC Cancer* 2008, **8:**254.
74. Jönsson G, Staaf J, Olsson E, Heidenblad M, Vallon-Christersson J, Osoegawa K, de Jong P, Oredsson S, Ringnér M, Höglund M, Borg Å: **High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization.** *Genes, Chromosomes and Cancer* 2007, **46:**543-558.
75. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat M-L, Gyorki DE, Ward T, Partanen A, et al: **Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers.** *Nat Med* 2009, **15:**907-913.
76. Jennings DE: **How Do We Judge Confidence-Interval Adequacy?** *The American Statistician* 1987, **41:**335-337.
77. Nemes S, Parris TZ, Danielsson A, Kannius-Janson M, Jonasson JM, Steineck G, Helou K: **Segmented Regression, a Versatile Tool to Analyze mRNA Levels in Relation to DNA Copy Number Aberrations.** *Gene Chromosomes Cancer* 2012, **51:**77-82.
78. Kim HJ, Fay MP, Feuer EJ, Midthune DN: **Permutation tests for joinpoint regression with applications to cancer rates.** *Stat Med* 2000, **19:**335-351.
79. Kim HJ, Fay MP, Yu BB, Barrett MJ, Feuer EJ: **Comparability of segmented line regression models.** *Biometrics* 2004, **60:**1005-1014.
80. Kieseppä IA: **Akaike Information Criterion, Curve-fitting, and the Philosophical Problem of Simplicity.** *The British Journal for the Philosophy of Science* 1997, **48:**21-48.
81. Kuha J: **AIC and BIC: Comparisons of Assumptions and Performance** *Sociological Methods & Research* 2004, **33:**188-229.
82. Jones RH, Dey I: **Determing one or more change-points.** *Chem Phys Lipids* 1995, **76:**1-6.
83. Glatting G, Kletting P, Reske SN, Hohl K, Ring C: **Choosing the optimal fit function: Comparison of the Akaike information criterion and the F-test.** *Med Phys* 2007, **34:**4285-4292.
84. Kletting P, Glatting G: **Model selection for time-activity curves: The corrected Akaike information criterion and the F-test.** *Z Med Phys* 2009, **19:**200-206.
85. Hurvich CM, Tsai C-L: **Regression and time series model selection in small samples.** *Biometrika* 1989, **76:**297-307.
86. Bozdogan H: **Model selection and Akaike Information Criterion (AIC) – The general theory and its analytical extensions** *Psychometrika* 1987, **52:**345-370.

87. Wagenmakers EJ, Farrell S: **AIC model selection using Akaike weights.** *Psychon Bull Rev* 2004, **11:**192-196.
88. Burnham KP, Anderson DR: **Multimodel inference - understanding AIC and BIC in model selection.** *Sociological Methods & Research* 2004, **33:**261-304.
89. Leday GGR, van der Vaart AW, van Wieringen WN, van de Wie MA: **Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines.** *Ann Appl Stat* 2012, **In Press**.
90. Chen CWS, Chan JSK, So MKP, Lee KKM: **Classification in segmented regression problems.** *Computational Statistics & Data Analysis* 2011, **55:**2276-2287.
91. Zhao J, Yu HX, Lin L, Tu J, Cai LL, Chen YM, Zhong F, Lin CZ, He FC, Yang PY: **Interactome study suggests multiple cellular functions of hepatoma-derived growth factor (HDGF).** *J Proteomics* 2011, **75:**588-602.
92. Guo ZY, He YZ, Wang SS, Zhang AX, Zhao P, Gao CF, Cao B: **Various effects of hepatoma-derived growth factor on cell growth, migration and invasion of breast cancer and prostate cancer cells.** *Oncol Rep* 2011, **26:**511-517.
93. Wang SA, Fang WY: **Increased expression of hepatoma-derived growth factor correlates with poor prognosis in human nasopharyngeal carcinoma.** *Histopathology* 2011, **58:**217-224.
94. Efron B: **Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods.** *Biometrika* 1981, **68:**589-599.
95. Hole AR: **A comparison of approaches to estimating confidence intervals for willingness to pay measures.** *Health Economics* 2007, **16:**827-840.
96. Fritz MS, Taylor AB, MacKinnon DP: **Explanation of Two Anomalous Results in Statistical Mediation Analysis.** *Multivariate Behavioral Research* 2012, **47:**61-87.
97. Parr WC: **A Note on the Jackknife, the Bootstrap and the Delta Method Estimators of Bias and Variance.** *Biometrika* 1983, **70:**719-722.
98. Hall P: **The Bootstrap and Edgeworth Expansion.** *Springer Series in Statistics* 1992.
99. Fosen J, Ferkingstad E, Borgan O, Aalen OO: **Dynamic path analysis - a new approach to analyzing time-dependent covariates.** *Lifetime Data Anal* 2006, **12:**143-167.
100. Ma SG, Huang J: **Methodology article - Additive risk survival model with microarray data.** *BMC Bioinformatics* 2007, **8**.
101. Gorst-Rasmussen A, Scheike TH: **Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model.** *Journal of Statistical Software* 2012, **47:**1-17.

102. Kyung M, Gill J, Ghosh M, Casella G: **Penalized Regression, Standard Errors, and Bayesian Lassos.** *Bayesian Anal* 2010, **5:**369-411.
103. Lange T, Hansen JV: **Direct and Indirect Effects in a Survival Context.** *Epidemiology* 2011, **22:**575-581.
104. Martinussen T: **Dynamic path analysis for event time data: large sample properties and inference.** *Lifetime Data Anal* 2010, **16:**85-101.
105. Martinussen T, Vansteelandt S, Gerster M, Hjelmborg JV: **Estimation of direct effects for survival data by using the Aalen additive hazards model.** *J R Stat Soc Ser B-Stat Methodol* 2011, **73:**773-788.
106. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang CS, et al: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nature Genet* 2005, **37:**710-717.
107. Li Y, Tesson BM, Churchill GA, Jansen RC: **Critical reasoning on causal inference in genome-wide linkage and association studies.** *Trends Genet* 2010, **26:**493-498.
108. Lee E, Cho S, Kim K, Park T: **An integrated approach to infer causal associations among gene expression, genotype variation, and disease.** *Genomics* 2009, **94:**269-277.
109. Lozano AC, Abe N, Liu Y, Rosset S: **Grouped graphical Granger modeling for gene expression regulatory networks discovery.** *Bioinformatics* 2009, **25:**I110-I118.
110. Chindelevitch L, Loh P-R, Enayetallah A, Berger B, Ziemek D: **Assessing statistical significance in causal graphs.** *BMC Bioinformatics* 2012, **13:**35.
111. Manly KF, Nettleton D, Hwang JTG: **Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses.** *Genome Research* 2004, **14:**997-1001.
112. MacKinnon DP, Lockwood CM, Williams J: **Confidence limits for the indirect effect: Distribution of the product and resampling methods.** *Multivariate Behav Res* 2004, **39:**99-128.
113. Mackinnon DP, Warsi G, Dwyer JH: **A simulation study of the mediated effect measures.** *Multivariate Behav Res* 1995, **30:**41-62.
114. Yu J, Sieuwerts A, Zhang Y, Martens J, Smid M, Klijn J, Wang Y, Foekens J: **Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer.** *BMC Cancer* 2007, **7:**182.
115. Fan C, Prat A, Parker J, Liu Y, Carey L, Troester M, Perou C: **Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures.** *BMC Medical Genomics* 2011, **4:**3.

116. Weigelt B, Pusztai L, Ashworth A, Reis-Filho JS: **Challenges translating breast cancer gene signatures into the clinic.** *Nat Rev Clin Oncol* 2012, **9:**58-64.
117. Colombo P-E, Milanezi F, Weigelt B, Reis-Filho J: **Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction.** *Breast Cancer Research* 2011, **13:**212.
118. Boulesteix AL, Sauerbrei W: **Added predictive value of high-throughput molecular data to clinical data and its validation.** *Brief Bioinform* 2011, **12:**215-229.
119. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21:**171-178.
120. Steineck G, Hunt H, Adolfsson J: **A hierarchical step-model for causation of bias-evaluating cancer treatment with epidemiological methods.** *Acta Oncol* 2006, **45:**421-429.
121. Mitelman F: **Recurrent chromosome aberrations in cancer.** *Mutation Research/Reviews in Mutation Research* 2000, **462:**247-253.
122. Patmore HS, James NEA, Cawkwell L, MacDonald A, Stafford ND, Greenman J: **Can a genetic signature for metastatic head and neck squamous cell carcinoma be characterised by comparative genomic hybridisation?** *Br J Cancer* 2004, **90:**1976-1982.
123. Heng HHQ, Liu G, Bremer S, Ye KJ, Stevens J, Ye CJ: **Clonal and non-clonal chromosome aberrations and genome variation and aberration.** *Genome* 2006, **49:**195-204.