HUI WANG
YOHAN LEE
STANLEY NELSON
CHIARA SABATTI

Inferring genomic loss and location of tumor suppressor genes from high density genotypes

INFERRING GENOMIC LOSS AND LOCATION OF TUMOR SUPPRESSOR GENES FROM HIGH DENSITY GENOTYPES

Hui WANG*, Yohan LEE**,

Stanley NELSON**, Chiara SABATTI * ** (1)

RÉSUMÉ

Des nouvelles technologies, comme les puces de génotypage d'Affymetrix (GeneChip Mapping 10K), permettent de mesurer des polymorphismes génétiques répartis sur le génome avec une grande densité. Cela permet une approche de questions classiques à une résolution auparavant inégalée. Simultanément, des nouveaux types d'analyses apparaissent pour tirer parti de l'augmentation des informations que contiennent ces données. Nous nous intéressons à la façon dont les génotypes des lignées cellulaires cancéreuses peuvent être utilisés pour reconstituer les événements de délétion génomique et cartographier des gènes suppresseurs de tumeur (TSG) potentiels. En utilisant le formalisme des modèles de Markov cachés, nous adaptons à cette nouvelle structure de données un modèle déjà décrit d'instabilité génomique au cours des cancers. Des simulations indiquent que notre procédure peut être puissante et précise. Son application à des données réelles montre des résultats préliminaires encourageants.

Mots clés : biopuces de génotypage, modèles de Markov cachés, perte de fragments chromosomiques, perte d'hétérozygotie.

ABSTRACT

Novel technologies, such as the 10k Affymetrix genotyping array, allow scoring of genetic polymorphisms at a very high density across the genome. This allows researchers to conduct traditional inquiries at an unprecedented resolution, while simultaneously motivates novel types of analysis, aimed at exploiting the increased information contained in these datasets. We consider how genotypes of cancer cell lines can be used to reconstruct genomic loss events and map putative tumor suppressor genes (TSG). Using a hidden Markov model framework, we adapt a previously described model for genomic instability in cancers to the current data structure. Simulations indicate that our procedure can be powerful and accurate and initial application to real data leads to encouraging results.

Keywords: Genotyping arrays, genomic loss, hidden Markov models, loss of heterozygosity.

* Department of Statistics, UCLA, Los Angeles, CA 90095-7088.
** Department of Human Genetics, UCLA, Los Angeles, CA 90095-7088.
(1) Corresponding author: csabatti@mednet.ucla.edu

1. Introduction

Large scale genomic variation is receiving increasing attention from the scientific community (Pollack *et al.*, 2002; Bignell *et al.*, 2004; Cox *et al.*, 2005; Iafrate *et al.*, 2004; Sebat *et al.*, 2004). Along the process that leads to tumorigenesis, genomic stability is impaired: cancer cells present a higher rate of genomic losses and duplications. Moreover, some of these variations in copy number may determine the cancerous status of the cells, for instance by inactivating a tumor suppressor gene. Observed cancer cells are often considered to be selected to favor genomic losses in particular regions harboring tumor suppressor genes (TSG). Indeed, genomic losses have often been studied in cancer cells with the precise intent of locating TSGs (Newton *et al.* (1998) provides a comprehensive review of the biological process). Generally, a link between copy number variation and altered behaviour of cancer cells has been established (Pollack *et al.* 2002, Cox *et al.*, 2005). Concurrently, it has recently become apparent that there are a number of genetic polymorphisms that consist of large scale genomic variations (e.g. considerably sized genomic deletes and duplicates) that arise or are transmitted through the germline (Iafrate *et al.*, 2004; Sebat *et al.*, 2004).

While these two biological processes are very different, their effects can be measured with the same technology. For a number of years, large scale studies of genomic instability were limited to the case of cancers, and variation in copy numbers was assessed by genotyping tumor tissue and a normal cell from the same individual. In these studies, the fact that a marker is heterozygous in the normal cell and a homozygous in the cancer cell, can be considered as evidence that one of the chromosomes of the individual under study experienced a loss of genetic material and, hence, only the allele residing on the other chromosome remains to be detected. (Symmetrically, one could explain this result in terms of increase in copy number of one of the chromosomes, such that the signal from the duplicated allele overwhelms the one from the other). Newton *et al.* (1998) (Newton and Lee, 2000; Newton, 2002, Newton *et al.*, 2003) propose a model for the analysis of this loss of heterozygosity (LOH) data and demonstrate how to apply it to locate putative tumor suppressor genes.

Recently, a number of other technologies have been developed to enable a more direct and high throughput assessment of copy number variations: these include comparative genomic hybridization (CGH), and array-based CGH (Pinkel *et al.*, 1998; Bignell *et al.* 2004). Due to these technological advancements, scientists are able to observe recently documented large scale genomic variation (Iafrate *et al.*, 2004; Sebat *et al.*, 2004). A number of statistical methods have been proposed to analyze the data from these experiments (Fridlyand *et al.*, 2004; Lai and Zhao, 2005; Wang *et al.*, 2005). Simultaneously, genotyping technology has evolved, and we are now able to measure polymorphic sites at a much higher density than just a few years ago. Thus LOH remains a powerful method of investigation. For example, recent studies have shown how genotyping arrays can be used to conduct effective copy number investigations (Lim *et al.*, 2004). One advantage of high density

genotyping is that one can potentially gather information on copy number without the concurrent need to type both a "normal" and "case" cell from the same individual. This is due to the aboundance of markers which provide a fine resolution snapshot of the genome, in which a "long" contiguous stretch of homozygous SNPs calls can be interpreted as the result of a genomic alteration (with appropriate caveats). When one considers this potential (already noted in Lim *et al.*, 2004), developing methods that enable the use of high density genotypes to study copy number variations becomes an important goal. While other techniques may lead to more direct and precise measurements, and should likely be preferred if the primary goal of the study is detection of copy number variations, genotyping is done routinely on large scale both for tumor and normal cells. Failing to identify detectable changes in copy numbers from these data may result in erroneous interpretation of the results of the study, in addition to a loss of useful information.

In this paper, we focus on the use of high density SNP genotyping for the study of genomic instability in cancer cells. In a companion study (Wang *et al.*, 2005) we analyze the case of large scale genomic variation. In the next section we present a model for genotypes in the presence of genomic instability and selection. We subsequently illustrate how to estimate the instability parameters of the model and how to reconstruct the most likely profiles of genomic alteration from an individual's genotype data. Section 4 contains the description of a likelihood ratio test to identify location of a tumor suppressor gene, and Section 5 illustrates our results with simulated and real data. We conclude with a discussion.

2. A model for genotypes under genomic instability and selection

We assume that the results of high density SNP genotyping across the genome are available for T cancer cell lines. In particular, we consider the case where the 10k genotyping array from Affymetrix is used, leading to the scoring of 10913 markers across the genome, at approximately intervals of 0.3 megabases. The methodology we will describe can be used with allele calls produced with other platforms, but it is tailored for high density genotyping—necessary to inform the inference of genomic aberrations from the genotype of cancer cells only. We denote with $Y = \{y_i\}_{i=1}^{M}$ the sequence of genotypes at M markers for one cell line: each y_i can take on one of four possible values: $(AA, AB, BB, -)$, corresponding to the three genotypes and a "no call" value. We will assume independence between genotypes corresponding to markers on 22 different chromosomes. It is convenient to group markers according to the chromosome where they are located: $X_k = \{x_{ki}\}_{i=1}^{m_k}$ will indicate the genotypes of the m_k markers on chromosome k, so that $Y = \{X_k\}_{k=1}^{22}$. When unnecessary for clarity, we will avoid using the index k, so that $X = \{x_i\}_{i=1}^{m}$ the collection of genotypes for markers in an unspecified chromosome. We use the superscript t to identify cell lines, so that the entire collection of our data is $\{Y^t\}_{t=1}^{T}$.

Newton *et al.* (1998) and Newton and Lee (2000) describe the model for genomic instability and selection in cancer cells that we use for our analysis. We refer the interested reader to these original papers to fully appreciate the underlying biological hypotheses, and we limit ourselves to a brief exposition. The substantial difference between the work presented in Newton *et al.* (1998), Newton and Lee (2000) and our study resides in the fact that these previous authors asserted LOH calls, from the availability of genotypes from both a cancer and a normal cell from the same individual. Here we assume that only cancer cells are typed: the genomic loss process is unobserved, and we need to describe how it is reflected in the genotypes of cancer cells. The framework of hidden Markov models appears as a natural solution. We start considering the instability component of the model, and data from one cell line. Let $\Pi = \{\pi_i\}_{i=1}^m$ denote the genomic loss process at the positions corresponding to each of the m markers on a chromosome: $\pi_i = 0$ indicates no genomic alterations at the location of the i-th marker and $\pi_i = 1$ indicates an alteration. In the following we will refer to alterations as a loss, since the model by Newton *et al.* (1998) that we are adapting was conceived specifically for losses; however, it is quite possible to interpret the detected abnormalities as increases in genomic copy number. With d_i we indicate the distance in megabases (Mb) between marker i and $i + 1$. Thus set the notation, the transition probabilities of the hidden Markov process can be described in terms of two parameters δ and η:

$$\begin{pmatrix} t(\pi_{i+1} = 1 \mid \pi_i = 1) & t(\pi_{i+1} = 0 \mid \pi_i = 1) \\ t(\pi_{i+1} = 1 \mid \pi_i = 0) & t(\pi_{i+1} = 0 \mid \pi_i = 0) \end{pmatrix}$$

$$= \begin{pmatrix} 1 - (1 - \delta)(1 - e^{-\eta d_i}) & (1 - \delta)(1 - e^{-\eta d_i}) \\ \delta(1 - e^{-\eta d_i}) & 1 - \delta(1 - e^{-\eta d_i}) \end{pmatrix}.$$

The parameter δ represents the sporadic loss rate, which is the probability that any location in the genome is lost in a random individual. Hence we have $\delta = P(\pi_1 = 1)$, and to conform with the transition probabilites, we will denote it with $t(\pi_1 = 1)$. The parameter η is used to model the dependency of the Markov process and the length of the genomic losses. In this framework the distance between two change-points in the π process is modeled similarly to the distance between two recombination events (Lange, 2002). Note that $t(\pi_{i+1} = 1 \mid \pi_i = 1) \to 1$ and $t(\pi_{i+1} = 0 \mid \pi_i = 0) \to 1$ as $d_i \to 0$. The adequacy of this model to describe genomic instability in cancer cells has been discussed by Newton *et al.* (1998) and supported by its successful application in empirical studies (see, for example Miller *et al.*, 2003). In this model, δ and η are constant across a region spanned by linked markers. Depending on the nature of the data acquired, it may be sensible to assume that each chromosome is characterized by a specific value of δ and η; additionally, it may be appropriate to allow these parameters to be cell-line specific.

To link the unobserved loss process to the genotype data, we use the following emission probabilities:

$$\begin{pmatrix} e(x_i = AA \mid \pi_i = 1) & e(x_i = AB \mid \pi_i = 1) \\ e(x_i = AA \mid \pi_i = 0) & e(x_i = AB \mid \pi_i = 0) \end{pmatrix}$$

$$\begin{pmatrix} e(x_i = BB \mid \pi_i = 1) & e(x_i = - \mid \pi_i = 1) \\ e(x_i = BB \mid \pi_i = 0) & e(x_i = - \mid \pi_i = 0) \end{pmatrix} .$$

$$= \begin{pmatrix} P_{A,i}(1-\tau) & 0 & (1-P_{A,i})(1-\tau) & \tau \\ P_{A,i}^2(1-\kappa) & 2P_{A,i}(1-P_{A,i})(1-\kappa) & (1-P_{A,i})^2(1-\kappa) & \kappa \end{pmatrix}, \quad (1)$$

where P_{A_i} is the frequency of allele A for the ith marker, τ is the missing rate in loss regions, and κ is the missing rate in regions with no genomic aberrations. The difference in missing rate is due to the fact that the loss region may produce an "anomalous" intensity signal, filtered out by quality control mechanisms and resulting in "no call"s. In the presence of a genomic loss, only homozygous genotypes are observed, and the relative abundance of AA and BB depends on the allele frequencies $P_{A,i}$. While the emission probabilities (1) do not account for genotyping error, this effect can be easily incorporated. More substantial modification of the emission probabilities are needed, instead, to consider linkage disequilibrium across markers. This is important when the genotype density increases substantially, and there is clear association between alleles at nearby markers. We discuss such modifications in a companion paper (Wang et al., 2005) and in the discussion section. The likelihood of a genotype sequence X under the instability model can be evaluated with standard HMM recursion formulas. In particular, if we define $\alpha(\pi_i) = \Pr(x_1, \ldots, x_i, \pi_i)$, and $\beta(\pi_i) = \Pr(x_{i+1}, \ldots, x_m \mid \pi_i)$, we have:

$$\alpha(\pi_i) = \sum_{\pi_{i-1}=0,1} \alpha(\pi_{i-1}) t(\pi_i \mid \pi_{i-1}) e(x_i \mid \pi_i)$$

$$\beta(\pi_i) = \sum_{\pi_{i+1}=0,1} \beta(\pi_{i+1}) t(\pi_{i+1} \mid \pi_i) e(x_{i+1} \mid \pi_{i+1}).$$

Then, for example $\Pr(X) = \sum_{\pi_m=0,1} \alpha(\pi_m)$. To emphasize that this probability depends only on the instability component of the model, we will indicate it with $P_I(X)$. A version of the aforemented recursion formulas can also be used to evaluate conditional probabilities $P_I(X \mid \pi_s = k)$, which will be relevant in the following.

Newton et al. (1998) describe the selection effect that is the basis of the possible localization of tumor suppressor genes, and we adopt the same model. They consider the possibility of one tumor suppressor gene per chromosome. Two parameters are introduced: s represents the location of the tumor suppressor gene, and ω_s the probability that a cancer cell line has a loss

at position s, i.e. $\omega_s = P(\pi_s = 1)$. The likelihood of a genotype sequence, once the selection effect is introduced, can be written as:

$$\Pr(X) = \omega_s P_I(X|\pi_s = 1) + (1 - \omega_s)P_I(X|\pi_s = 0),$$

where $P_I(X|\pi_s)$ depends on the instability process. In order to map a tumor suppressor gene, one needs to acquire data on multiple cell lines, so that the complete data for analysis will be a collection of genotype sequences $X^1, X^2, \ldots, X^T$ on the same chromosome. The data likelihood is then:

$$\mathcal{L}(X^1, \ldots, X^T|\delta, \eta, \omega, \tau, s) = \prod_{t=1}^{T}(\omega P_I(X^t|\pi_s^t = 1) + (1 - \omega)P_I(X^t|\pi_s^t = 0)).$$

Given $\mathcal{L}(X^1, \ldots, X^T|\delta, \eta, \omega, \tau, s)$, we can estimate the model parameters, reconstruct the location of most likely losses, and attempt mapping the tumor suppressor gene. We will assume that allele frequencies $P_{A,i}$ are known: Affymetrix, for example, provides an estimate of allele frequencies for all the SNPs on its genotyping chip. We will also assume a known background "NoCall" rate: typically this can be estimated in each laboratory without much difficulty using a set of genotypes from normal individuals. The parameters we then need to estimate are η, δ, τ for the instability component of our model and ω_s and s for the selection effect.

3. The instability model: parameter estimation and reconstruction of genome losses

Our estimation strategy rests on the assumption that the large number of typed markers allows us to gather enough information on the sporadic loss process that its parameters η, δ, and τ can be estimated separately from ω and the TSG location s. This is certainly the case when one is typing SNPs at high density genomewide, η, δ, and τ are constants across chromosomes and individuals, and there is only one TSG per chromosome –which is the situation we consider in this paper. However, the same assumption may be appropriate when η, δ, and τ vary across chromosomes and individuals, and when there are multiple TSGs –depending on the marker density and the value of the parameters. In the cases in which this assumption appears inadequate, one would not resort to the two stage strategy for estimating $\eta, \delta, \tau, \omega$ and s that we describe below, but would need to simultaneously estimate all of these parameters. This is not difficult theoretically, and the likelihood derivatives given in the following can be used to describe such a maximization routine. We did not pursue this strategy because the nature of the data we were interested in made it unnecessary and computationally very intensive. Note that if a smaller number of markers is typed –so that the proposed assumption may be inadequate– the number of computations will also be significantly reduced, making the simultaneous estimation strategy more feasible (simultaneous estimation is carried out, for example, in Newton *et al.* 1998).

To estimate the parameters of the instability model, δ, η and τ, we use a maximum likelihood approach and a gradient algorithm. As anticipated, we consider a likelihood that is based solely on the instability component of our model:

$$\log \mathcal{L}_I(Y^1, \ldots, Y^T | \eta, \delta, \tau) = \sum_{t=1}^{T} \sum_{k=1}^{22} \log P_I(X_k^t).$$

To describe the form of the first derivatives of the loglikelihood with respect to the three parameters of interest, it is easier to focus initially on one term $P_I(X_k^t) = P_I(X)$. When the hidden state $\Pi = (\pi_i)$ is known, we obtain

$$P_I(X, \Pi) = P_I(x_1, \cdots, x_m, \pi_1, \cdots, \pi_m) = t(\pi_1) \prod_{i=1}^{m-1} t(\pi_{i+1} \mid \pi_i) \prod_{i=1}^{m} e(x_i \mid \pi_i).$$

Therefore,

$$\frac{\partial P_I(X, \Pi)}{\partial t(\pi_{i+1} \mid \pi_i)} = \frac{P_I(X, \Pi)}{t(\pi_{i+1} \mid \pi_i)} \quad \frac{\partial P_I(X, \Pi)}{\partial e(x_i \mid \pi_i)} = \frac{P_I(X, \Pi)}{e(x_i \mid \pi_i)}.$$

Now, recalling that the likelihood of the data is $P_I(X) = \sum_{\pi_1, \cdots, \pi_m} P_I(X, \Pi)$, and that we can carry out the summations with respect to π_i using the forward and backward recursions, we obtain the following expression for the partial derivative for η:

$$\frac{\partial P_I(X)}{\partial \eta} = \frac{\partial \sum\limits_{\pi_1, \cdots, \pi_m} P_I(X, \Pi)}{\partial \eta} = \frac{\sum\limits_{\pi_1, \cdots, \pi_m} \partial P_I(X, \Pi)}{\partial \eta}$$

$$= \sum_{\pi_1, \cdots, \pi_m} \sum_{i=1}^{m-1} \frac{P_I(X, \Pi)}{t(\pi_{i+1} \mid \pi_i)} \frac{\partial t(\pi_{i+1} \mid \pi_i)}{\partial \eta}$$

$$= \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \frac{1}{t(\pi_{i+1} \mid \pi_i)} \frac{\partial t(\pi_{i+1} \mid \pi_i)}{\partial \eta} \sum_{\substack{\pi_1, \cdots, \pi_{i-1} \\ \pi_{i+2}, \cdots, \pi_m}} P_I(X, \Pi)$$

$$= \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \frac{P_I(X, \pi_i, \pi_{i+1})}{t(\pi_{i+1} \mid \pi_i)} \frac{\partial t(\pi_{i+1} \mid \pi_i)}{\partial \eta}$$

$$= \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \alpha(\pi_i) e(x_i \mid \pi_{i+1}) \beta(\pi_{i+1}) \frac{\partial t(\pi_{i+1} \mid \pi_i)}{\partial \eta},$$

where $\alpha(\pi_i)$ and $\beta(\pi_i)$ are computed by the forward and backward algorithms. We initialize $\alpha(\pi_1)$ and $\beta(\pi_1)$ with $\alpha(\pi_1) = P(x_1, \pi_1) = e(x_1 \mid \pi_1) t(\pi_1)$ and

$\beta(\pi_1) = 1$, where $t(\pi_1 = 1) = \delta$ and $t(\pi_1 = 0) = 1 - \delta$. Similarly, we have

$$\frac{\partial P_I(X)}{\partial \delta} = \sum_{\pi_1} \alpha(\pi_1)\beta(\pi_1)\frac{\partial t(\pi_1)}{\partial \delta}$$

$$+ \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \alpha(\pi_i)e(x_i \mid \pi_{i+1})\beta(\pi_{i+1})\frac{\partial t(\pi_{i+1} \mid \pi_i)}{\partial \delta},$$

And finally, for the parameter τ:

$$\frac{\partial P_I(X)}{\partial \tau} = \sum_{i=1}^{m} \frac{\alpha(\pi_i = 1)\beta(\pi_i = 1)}{e(x_i \mid \pi_i = 1)} \frac{\partial e(x_i \mid \pi_i = 1)}{\partial \tau}.$$

The derivatives of the log-likelihood can be easily computed from the expressions given above and used to set up a gradient algorithm:

$$\delta^{(t+1)} = \delta^{(t)} + \lambda_\delta \frac{\partial \log P_I(X)}{\partial \delta}\bigg|_{\delta = \delta^{(t)}}$$

$$\eta^{(t+1)} = \eta^{(t)} + \lambda_\eta \frac{\partial \log P_I(X)}{\partial \eta}\bigg|_{\eta = \eta^{(t)}}$$

$$\tau^{(t+1)} = \tau^{(t)} + \lambda_\tau \frac{\partial \log P_I(X)}{\partial \tau}\bigg|_{\tau = \tau^{(t)}},$$

with λ indicating the step size. It is also quite clear how different assumptions on variability of the parameters η, δ, τ across chromosomes and individuals will result in derivatives of the log-likelihood obtained using summations across different index sets.

Once an estimate of the instability parameters is obtained, this can be used to reconstruct the most likely genomic aberration profile Π_k^t for each of the individuals and chromosomes in the sample, using a standard Viterbi algorithm (see, Durbin et al., 1998). This represents by itself an interesting output of our procedure, as it allows scientists to gather information on location and size of genomic losses from data that consist only in genotypes of cancer cells. Indeed, some researchers may consider this as the only output of interest, and not subscribe to the selection model for identification of TSGs that we will describe in the following section. For this reason, we preferred to opt for the 2-stage procedure.

4. The selection effect: likelihood ratio test to identify tumor suppressor gene locations

For the purpose of identifying the location s of a tumor suppressor gene and the increased probability of genomic loss ω_s associated with this location, we conduct a series of likelihood ratio tests, where s is allowed to vary position across the entire chromosome, and the hypothesis $H_0^s : \delta = \omega_s$ vs $H_1^s : \delta < \omega_s$ is tested at each of the examined locations s. The genomic region for which the hypothesis H_0^s is rejected will be considered as likely regions to harbor tumor suppressor genes, and the corresponding ω_s will describe the selection effects. At this stage, the parameters δ, η, τ are considered known, so that for each explored location s we have to maximize the likelihood only with respect to the parameter ω_s. Recall that for a given location s, the likelihood of the chromosome data is:

$$\mathcal{L}(X^1, \ldots, X^T | \delta, \eta, \omega_s, \tau, s) = \prod_{t=1}^{T} (\omega_s P_I(X^t | \pi_s^t = 1) + (1 - \omega_s) P_I(X^t | \pi_s^t = 0)).$$

Notice that $\mathcal{L}(X^1, \ldots, X^T | \delta, \eta, \omega_s = \delta, \tau, s) = \prod_{t=1}^{T} P_I(X^t \mid \eta, \delta, \tau)$. Furthermore, taking the logarithm of $\mathcal{L}$ we obtain a concave function of ω_s; indeed $\log \mathcal{L}(\omega_s) = \sum_{t=1}^{T} \log(\omega_s(P_I(X^t | \pi_s^t = 1) - P_I(X^t | \pi_s^t = 0)) + P_I(X^t | \pi_s^t = 0))$ is the sum of logarithms of affine functions, hence the sum of concave functions is thus concave. This allows us to conclude that there is only one maximum for $\log \mathcal{L}(\omega_s)$ and for $\mathcal{L}(\omega_s)$. In carrying out this series of likelihood ratio tests we follow the convention used in linkage genome scans of recording the logarithm base 10 of the inverse of the likelihood ratio, called LOD score; precisely, at each examined location s our test statistics L_s will be

$L_s =$

$$\begin{cases} \log_{10} \dfrac{\mathcal{L}(X^1, \ldots, X^T | \delta, \eta, \omega_s^*, \tau, s)}{\displaystyle\prod_{t=1}^{T} P_I(X^t \mid \eta, \delta, \tau)} & \omega_s^* = \mathrm{argmax} \mathcal{L}(X^1, \ldots, X^T | \delta, \eta, \omega_s, \tau, s) > \delta \\ \\ \qquad\qquad 0 & \omega_s^* < \delta \end{cases}$$

In terms perhaps more familiar to statisticians, that L_s can be interpreted as a profile log-likelihood ratio. The s^* location that maximizes L_s represents a candidate tumor suppressor gene location. To decide if the evidence in favor of $\omega_s^* > \delta$ is sufficient, one needs to further examine the value of L_{s^*} and determine if such a difference in likelihood values is attributable to random chance or not. We will return to this point in the following.

To maximize $\mathcal{L}(X^1, \ldots, X^T | \delta, \eta, \omega_s, \tau, s)$ as a function of ω_s it is convenient to use an EM algorithm. If we consider $\pi_s^1, \ldots \pi_s^T$, loss status at location s for the T cell-lines in the sample as missing data, we obtain the following complete data likelihood:

$$P(X^1, \ldots, X^T, \pi_s^1, \ldots, \pi_s^T | \omega_s, s)$$

$$= \prod_{t=1}^{T} (\omega_s P_I(X^t \mid \pi_s^t = 1))^{\pi_s^t} ((1 - \omega_s) P_I(X^t \mid \pi_s^t = 0))^{(1 - \pi_s^t)},$$

leading to the complete data log-likelihood below:

$$\log P(X^1, \ldots, X^T, \pi_s^1, \ldots, \pi_s^T | \omega_s, s)$$

$$= \sum_{t=1}^{T} \pi_s^t \log(\omega_s) + (T - \sum_{t=1}^{T} \pi_s^t) \log(1 - \omega_s) + \sum_{t=1}^{T} \log P_I(X^t \mid \pi_s^t).$$

The last term does not depend on ω_s, so we can omit it from consideration in the following. The EM iterations will be based on the following expected complete data log-likelihood:

$$Q(\omega|\omega^{(l)}) = \sum_{t=1}^{T} E(\pi_s^t | X^t, \omega_s^{(l)}) \log(\omega_s) + (T - \sum_{t=1}^{T} E(\pi_s^t | X^t, \omega_s^{(l)})) \log(1 - \omega_s).$$

For the expectation step, we will have:

$$E(\pi_s^t | X^t, \omega_s^{(l)}) = \Pr(\pi_s^t = 1 | X^t, \omega_s^{(l)})$$

$$= \frac{P_I(X^t | \pi_s^t = 1)\omega_s^{(l)}}{P_I(X^t | \pi_s^t = 1)\omega_s^{(l)} + P_I(X^t | \pi_s^t = 0)(1 - \omega_s^{(l)})}.$$

Maximizing $Q(\omega_s | \omega_s^{(l)})$ leads to $\omega_s^{(l+1)} = \sum_{t=1}^{T} E(\pi_s^t | X^t, \omega_s^{(l)})/T$.

We now return to the problem of determining which locations s, corresponding to high values of L_s, should be considered serious candidates for a TSG. First, note that L_s has a reasonable interpretation in terms of how much more likely the data is under H_1^s then H_0^s, and that the researcher may want to select a threshold value that better represents his/her interpretation of the study results. On purely statistical grounds, the determination of an appropriate cut off depends on the distribution of L_s under the null H_0^s and on the necessity of taking into account that multiple tests are being performed. Furthermore, notice that the tests L_{s_1} and L_{s_2} corresponding to two locations on the same chromosome are not independent. To determine a significance cut-off one ideally would like to know the distribution of the entire process $\{L_s\}_s$ under the complete null hypothesis. Unfortunately, this is unknown at this stage. The marginal distribution of L_s, as $T \to \infty$, can be roughly approximated using the known results for likelihood ratio tests: under H_0^s, $2\ln L_s$ is asymptotically distributed as a 50:50 mixture of a mass at zero and $\chi_{(1)}^2$ (the mass at zero derives from the fact that we place a constraint on the values of $\omega_s > \delta$, and the 0.5 mixing coefficient can be derived from the consistency and gaussianity of the MLE of ω_s). While this approximation of the distribution of L_s is rather crude, it provides us a guideline of what a reasonable significance cut-off may be. The appropriate cut-off for L_s depends on the distribution of L_s and, roughly speaking, on the number of "effectively independent" tests, which is determined by the length of the segment of the genome studied and the value of the η parameter. We suggest that once the instability parameters are estimated, a small scale simulation study be conducted where genotype data

with the same structure as the real one is generated from the instability model, with no selection effect, and a cut-off for L_s that controls the desired measure of error rate to be determined. It may be of use to refer once again to the analogy with linkage mapping which carries through in terms of distribution for L_s: in these genetic mapping studies, a value of L_s greater than 3, or 3.5 is typically considered strong evidence in favor of H_1^s (Lander and Kruglyak, 1995).

5. Simulations and data example

In this section we illustrate the proposed methodology with a small scale simulation study and by applying it to a real dataset. When we simulate data, we consider the same structure of the Affymetrix 10k mapping arrays: 10913 SNPs covering 22 autosomal chromosomes with an average distance of 230 kb. To generate genotypes we use the population allele frequencies provided by Affymetrix. Our simulations are small scale, in that we did not consider a series of possible values for the model parameters, but we simply chose one and evaluated the performance of our model in one case. Our intention is not to investigate the performance of our model overall, but simply to illustrate its potential. We chose instability parameter values that are realistic for at least some biological settings: $\delta = 0.1, \eta = 0.2, \kappa = 0.08, \tau = 0.13$. As far as the selection component of our model, we postulated one tumor suppressor gene at location 49.10 Mb on chromosome 1.

The first goal of our simulation was to evaluate the effectiveness of our procedure for the estimation of the instability parameters. We attempted estimation of these parameters using data from one individual only. To ensure that the presence of a TSG would not introduce distortions, we constrained the location 49.10 Mb on chromosome 1 to be lost and generated the rest of the data from the instability model. We repeated this 100 times and estimated in each case the three parameters δ, η, and τ. Results are presented in Figure 1: mean and median of the estimators are both concordant with the true parameter values, and the spread is reasonably small.
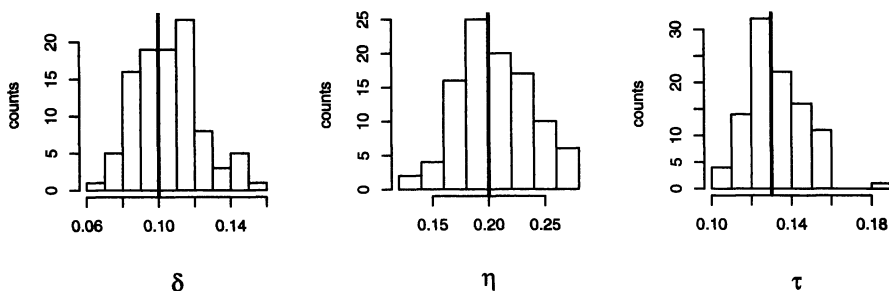

FIG 1. — **Instability Parameters Reconstruction.** Histograms of the estimated values for each of the instability parameters using genotypes from one cell line in 100 simulations. The vertical bold lines indicate the true value of the parameters

Using the estimates of the instability parameters and the Viterbi algorithm, we subsequently reconstructed the loss status for each of the simulated cancer cell lines. We evaluated this reconstruction in terms of sensitivity (fraction of SNPs in regions of genomic loss that are correctly described) and specificity (fraction of SNPs in normal genomic regions that are correctly described). Figure 2 presents the histograms of sensitivity and specificity across the 100 cases: the performance is quite satisfactory, with average sensitivity 0.86 and specificity 0.99.
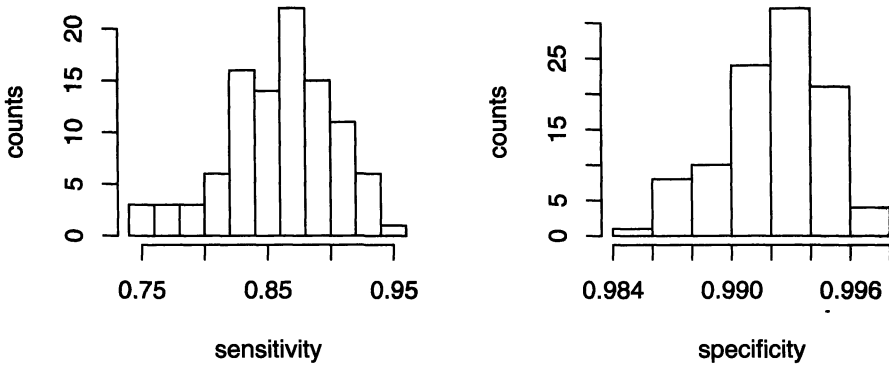

FIG 2. — **Reconstruction of Loss Status.** Histograms of sensitivity and specificity of the loss process reconstruction using instability model in 100 simulations

We then turn to investigate the effectiveness of our algorithm for the localization of the tumor suppressor gene. For this problem we considered the same parameter settings described above, ω_s=0.3, and genotypes from 50 cancer cell lines. We generated 100 datasets, and applied to each of them the two-step procedure consisting of estimating the instability parameters first and then attempting localization of the TSG. Our estimates of the instability parameters are quite good (unsurprisingly better than the one described above, due to the larger abundance of data). We evaluated the LOD curve at roughly 10,000 locations in the genome corresponding to the mid-point between each of the markers available. To determine the appropriate significance threshold for the LOD score curve, we conducted a simulation study using the true values of the instability parameters. Given the quality of our estimates, this does not represent a significant short-cut: the value 3.5 resulted in control of the family wise error rate at the 0.05 level. Figures 3 and 4 illustrate the results in one of our simulations. The LOD score curve across the entire genome is presented in Figure 3: only the area on chromosome one corresponding to the location of the TSG leads to values higher than 3.5. Figure 4 provides a more detailed illustration of the data on chromosome one and our analysis for this simulation: comparing panels (a) and (b) one can appreciate the de-noising achieved with the reconstruction of the most likely loss profile. Simultaneously, the ω_s estimates and the LOD curve (based on the simultaneous analysis of all cell lined) detect signals that are not enriched sufficiently at the level of individual tumors to be interpreted as losses.
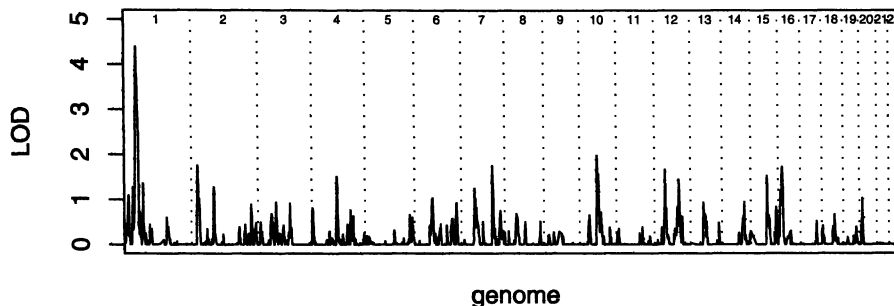
FIG 3. — **LOD score for TSG**. Profile log-likelihood ratio for the location s of the tumor suppressor gene across the genome for one case of the simulation

When we analyze the results of the 100 simulations, we obtain a power to detect the TSG on chromosome 1 equal to 0.5. The actual FWER is 0.05 and the FDR 0.037 (to estimate these, we considered as a false discovery any location leading to a LOD > 3.5 and more than $1/\eta$ away from the true TSG). The relatively low power can be explained considering that ω and the average size of losses (as determined by η) are both relatively small. Furthermore, it is noteworthy that we generated our dataset setting $\omega=0.3$, but we did not fix the actual fraction of cancer cells in the sample to be 0.3; this reduces the overall power. Figure 5 illustrates how the LOD score at the TSG location varies as a function of the actual proportion of losses in the sample. From Figure 5 is also evident that larger ω values will increase power; a larger sample size will also obviously do so. The effect of η and δ on power cannot be deduced from our simulations. Higher values of δ would make sporadic losses more common, and higher values of η would make sporadic losses shorter (and hence less likely to overlapp): changes in these parameters will affect the significance cut-off level and likely the power.

While the results of this small scale simulation study appear fairly positive, one must recall that we generated the data according to the model we used to analyze them. Our simulation allows us to verify the accuracy of our algorithm and evaluate the power of our procedure in an ideal situation, but it does not reflect the difficulties that we may encounter in the analysis of real data. To partially address this concern, we now turn to the analysis of a small dataset, collected by our colleagues at UCLA that motivated this methodology development. In a pilot study, Affymetrix 10k arrays were used to genotype 11 samples of primary brain tumors. In the initial analysis, whose results we report here, we assumed that the instability model parameters δ, η, and τ were constant across chromosomes and cell lines. Their estimated values were $\delta = 0.1638$, $\eta = 0.6777$, $\tau = 0.1398$. A simulation study suggested the value of 3.8 as a cut-off for L_s in order to control FWER at 0.05. Only one region in the genome reached that level, on the short arm of chromosome 9: data and LOD score are reported in Figure 6.
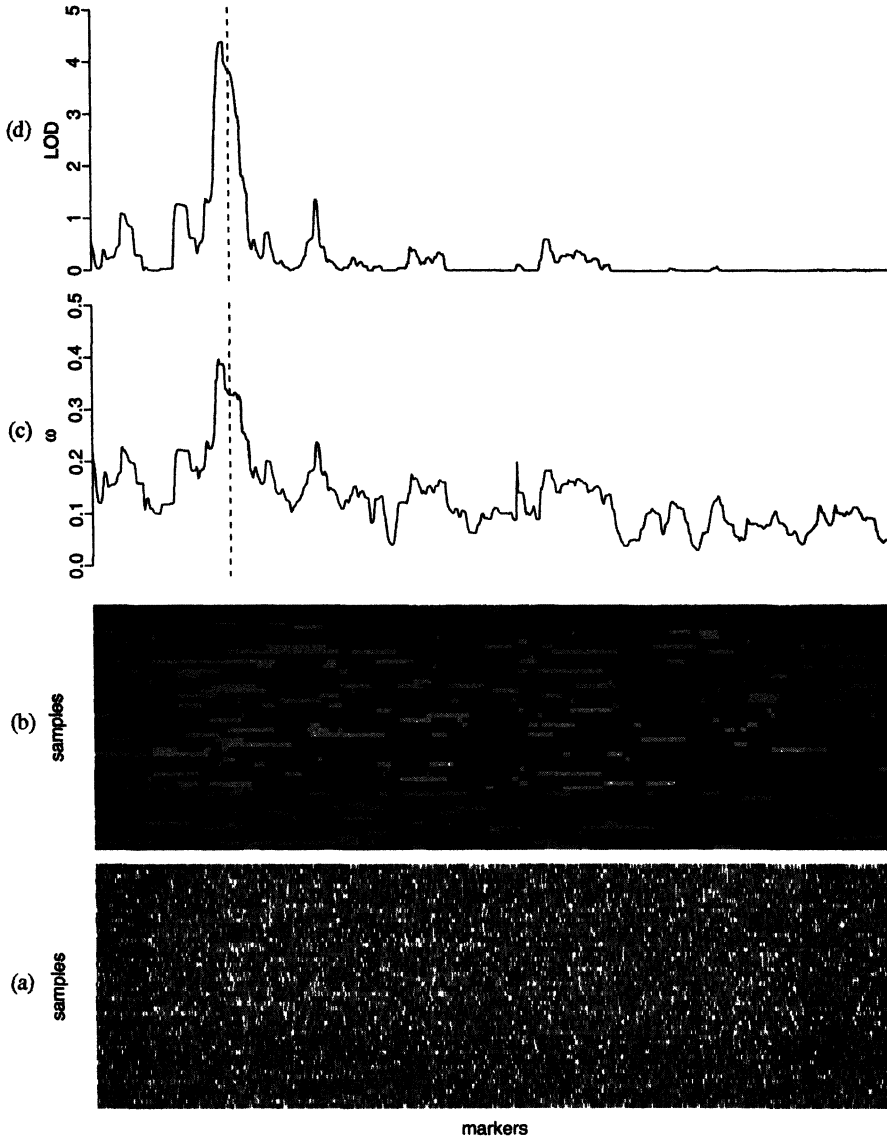
165

FIG 4. — **Localization of the TSG on Chromosome One.** Panel (a) represents the data generated by our simulation and used in the reconstruction. Each row represents an individual and each column a SNP. Polymorphisms are ordered according to their genomic position. Heterozygous genotypes are colored in black, homozygous genotypes in grey, and missing calls are left blank. For ease of display, inter-marker distances are depicted as constant. Panel (b) displays the corresponding loss profiles reconstructed with the instability model. Panel (c) reports the estimated value for ω_s for each of the genomic positions. Finally, in panel (d), we present the LOD score curve for s. The true location of the TSG is indicated with a dashed line
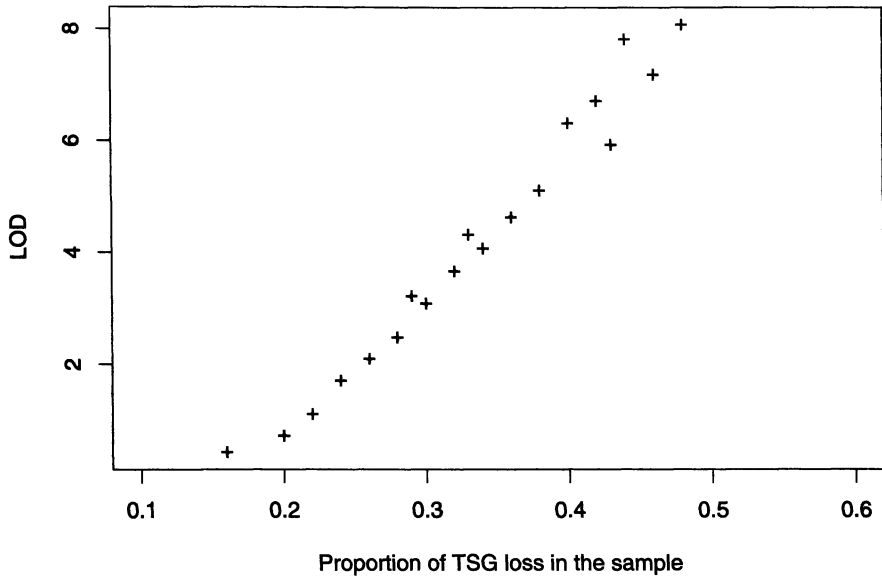
FIG 5. — **LOD value at the TSG.** Simulated datasets were grouped according to the actual proportion of samples with a loss in the TSG locations. The average LOD score at the TSG was computed within these groups and is plotted against the corresponding proportion of losses
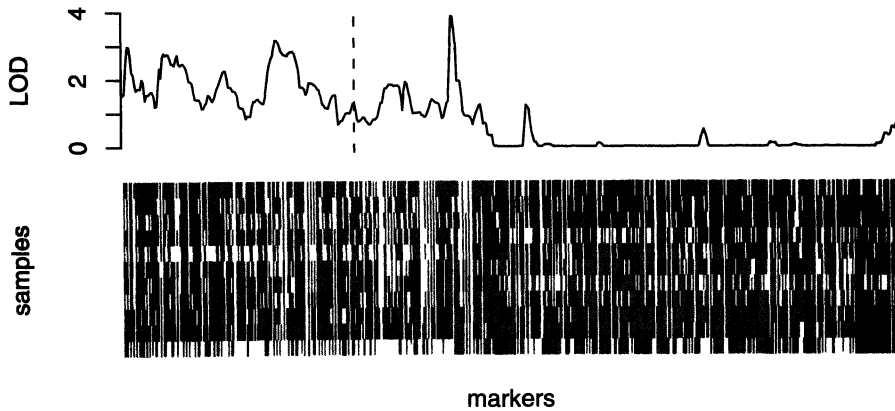

FIG 6. — **Data Analysis.** We present the data (coded in the same format as in Figure 4) and the LOD curve for the localization of a tumor suppressor gene relative to chromosome 9. The location of the known TSG p16 is indicated with a dashed line

Interestingly, this region is known to harbor a tumor suppressor gene, referred to as p16 (Lucas *et al.*, 1995). Considering the data in Figure 6 leads to a few remarks. There are four samples that are almost entirely homozygous on the

short arm of chromosome 9, and two with high loss rates: this is reflected in an overall higher level of the LOD curve. The maximum LOD score, however, is not reached in correspondence of the known TSG. The length of these losses is quite higher than what is predicted by the value of η: while this makes the signal quite strong in this region, it can also be interpreted as a model failure; we are currently exploring more careful analysis of the data, allowing instability parameters to vary across cell lines and chromosomes.

6. Discussion

We describe a model to reconstruct genomic losses in cancer cell lines from genotype data and to identify the likely location of tumor suppressor genes on the basis of genotype data from a collection of cancer cell lines. Our model for the process of genomic loss is taken from Newton *et al.* (1998), and we coupled it with a hidden Markov model to specify how the process of genomic loss is translated in observations on genotypes of the cancer cell line only. The study design we consider and the choices at the foundation of our computational strategy are based on the use of high density SNP genotyping, as obtained with Affymetrix 10k arrays, which allow scoring of 10,000 polymorphisms across the human genome.

When only genotype data from cancer cell lines are available, information on genomic losses is contained in stretches of homozygous markers: the longer the stretch of homozygosity, the less likely that such multilocus genotypes may be observed in absence of genomic aberrations. While in our model genomic loss is the only cause for increased multilocus homozygosity, there are other mechanisms that can potentially generate genotypes with similar characteristics: in particular, inbreeding (Leutenegger *et al.*, 2003) and linkage disequilibrium (Sabatti and Risch, 2002; Rosenberg and Calabrese, 2004) are important alternative explanations for increased homozygosity. We now briefly discuss the implications of these observations.

Linkage disequilibrium is the terminology used in genetics to indicate association between alleles at nearby markers. The model for genotypes given the genomic loss status Π that we proposed assumes linkage equilibrium, that is absence of association. As the distance between neighboring markers decreases, this assumption becomes inadequate and leads to identification of more losses than are really present in the data set. For data obtained with the Affymetrix genotyping array and described in this paper, this does not present a very serious problem: linkage disequilibrium, when present, typically extends at most across a window of 3-4 markers –which is generally too small to lead to the suspicion of a genomic loss. However, when the density of the scored polymorphisms increases, it becomes necessary to incorporate the effect of linkage disequilibrium in our model. Indeed, it is possible to modify the emission probabilities of the HMM in order to account for the association between alleles at neighboring markers. This leads to a likelihood that cannot be described in the typical framework of HMM, but can still be evaluated

with recursive formulas. We describe this model in a companion paper (Wang H. *et al.*, 2005).

The effect of inbreeding and genomic losses are more difficult to distinguish. Indeed, Leutenegger *et al.* (2003) have a model to estimate inbreeding coefficients that is practically identical to the HMM we use to describe genome instability. In this regard, the genomic loss profile that we reconstruct for each cell line under the instability model is valid only under the assumption of no inbreeding (which, incidentally, is likely to hold in the type of population samples collected for LOH studies). The instability component of our model, instead, captures an effect that is clearly distinct from inbreeding: the propensity of multiple cell lines to share losses in the same region. The inference on TSG location, hence, should be robust to the presence of inbreeding when sufficient numbers of independent cancers are analyzed.

Finally, we note that there are some extensions to our model that can be easily carried through and that we did not implement in the present paper partly because of time constraints and partly because they did not appear necessary to perform the analysis of the data we set out to study, namely genotypes obtained with the Affymetrix 10k array. One such extension involves the introduction of genotyping error parameter in the emission probabilities. This parameter can either be assumed known or estimated from the data (however, estimating it from another dataset is likely to produce better results). Similarly, the κ parameter can be varied across markers or cell lines. Another extension consists in the simultaneous estimation of all the parameters $\eta, \delta, \tau, \omega_s$, which, as we have discussed in the previous section, may be important for datasets that do not provide genome-wide genotyping. Furthermore, while we described our model assuming the instability parameters held constant across cell-lines and chromosomes, one can easily relax this assumption.

References

BIGNELL G., HUANG J., GRESHOCK J., WATT S., BUTLER A., WEST S., GRIGOROVA M., JONES KW., WEI W., STRATTON M., FUTREAL P., WEBER B., SHAPERO M., WOOSTER R. (2004). "High-resolution analysis of DNA copy number using oligonucleotide microarrays," *Genome Res.* 14: 287–95.

COX C., BIGNELL G., GREENMAN C., STABENAU A., WARREN W., STEPHENS P., DAVIES H., WATT S., TEAGUE J., EDKINS S., BIRNEY E., EASTON D., WOOSTER R., FUTREAL P., STRATTON M. (2005). "A survey of homozygous deletions in human cancer genomes," *PNAS* 102: 4542–4547.

DURBIN E., EDDY S., KROGH A., MITCHINSON G. (1999). *Biological sequence analysis*, Cambridge University Press.

FRIDLYAND J., SNIJDERS A., PINKEL D., ALBERTSON D., JAIN A. (2004). "Hidden Markov models approach to the analysis of array CGH data," *Journal of Multivariate Analysis* 90: 132–153.

HUANG J., WEI W., ZHANG J., LIU G., BIGNELL G., STRATTON M., FUTREAL P., WOOSTER R., JONES K., SHAPERO M. (2004). "Whole genome DNA copy number changes identified by high density oligonucleotide arrays," *Hum. Genomics* 1: 287–99.

IAFRATE A., FEUK L., RIVERA M., LISTEWNIK M., DONAHOE P., QI Y., SCHERER SW, LEE C. (2004). "Detection of large-scale variation in the human genome," *Nature Genetics* 36: 949–51.

LAI Y., ZHAO H. (2005). "A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data," *Comput. Biol. Chem.* 29: 47–54.

LANDER E., KRUGLYAK L. (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results," *Nature Genetics* 11: 241–247.

LEUTENEGGER A., PRUM B., GENIN E., VERNY C., LEMAINQUE A., CLERGET-DARPOUX, F., THOMPSON E. (2003). "Estimation of the inbreeding coefficient through use of genomic data," *Am. J. Hum. Genet.* 73:516–23.

LIN M., WEI L., SELLERS W., LIEBERFARB M., WONG W., LI C.(2004). "dChip-SNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data," *Bioinformatics* 20: 1233–40.

LUKAS J., PARRY D., AAGAARD L., MANN D. J., BARTKOVA J., STRAUSS M., PETERS G., BARTEK J. (1995). "Retinoblastoma-protein-dependent cell-cycle inhibition by the tumour suppressor p16," *Nature* 375: 503–506.

MILLER B., WANG D., KRAHE R., WRIGHT F. (2003). "Pooled analysis of loss of heterozygosity in breast cancer: a genome scan provides comparative evidence for multiple tumor suppressors and identifies novel candidate regions," *Am. J. Hum. Genet.* 73: 748–67.

NEWTON M., LEE Y. (2000). "Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data," *Biometrics* 56: 1088–97.

NEWTON M., GOULD M., REZNIKOFF C., HAAG J. (1998). "On the statistical analysis of allelic-loss data," *Stat. Med.* 17: 1425–45.

NEWTON M.A. (2002). "Discovering combinations of genomic alterations associated with cancer," *Journal of the American Statististical Association* 97: 931–942.

OLSHEN A., VENKATRAMAN E., LUCITO R., WIGLER M. (2004). "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics* 5: 557–72.

PINKEL D., SEGRAVES R., SUDAR D., CLARK S., POOLE I., KOWBEL D., COLLINS C., KUO W., CHEN C., ZHAI Y., DAIRKEE S., LJUNG B., GRAY J., ALBERTSON D. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nature Genetics* 20: 207–11.

POLLACK J., SORLIE T., PEROU C., REES C., JEFFREY S., LONNING P., TIBSHIRANI R., BOTSTEIN D., BORRESEN-DALE A., BROWN P. (2002). "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences* 99: 12963–12968.

ROSENBERG N., CALABRESE P. (2004). "Polyploid and multilocus extensions of the Wahlund inequality," *Theoretical Population Biology* 66: 381–391.

SABATTI C., RISCH N. (2002). "Homozygosity and linkage disequilibrium," *Genetics* 160: 1707–1719.

SEBAT J., LAKSHMI B., TROGE J., ALEXANDER J., YOUNG, J., LUNDIN P., MANER S., MASSA H., WALKER M., CHI M., NAVIN N., LUCITO R., HEALY J., HICKS J., YE K., REINER A., GILLIAM T., TRASK B., PATTERSON N., ZETTERBERG A., WIGLER M. (2004). "Large-scale copy number polymorphism in the human genome," *Science* 305: 525–8.

WANG H., SERVICE S., FREIMER N., SABATTI C. (2005). "Detecting large scale genomic variation through high density SNP genotyping," *manuscript in preparation*.

WANG P., KIM Y., POLLACK J., NARASIMHAN B., TIBSHIRANI R. (2005). "A method for calling gains and losses in array CGH data," *Biostatistics* 6: 45–58.