Discovering Combinations of Genomic Aberrations Associated with Cancer

Michael A. Newton

Journal of the American Statistical Association, Vol. 97, No. 460. (Dec., 2002), pp. 931-942.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28200212%2997%3A460%3C931%3ADCOGAA%3E2.0.CO%3B2-S

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Wed Jan 16 15:05:57 2008

Discovering Combinations of Genomic Aberrations Associated With Cancer

Michael A. NEWTON

This article introduces a model-based statistical methodology for the analysis of copy-number variations in cancer genomes measured by comparative genomic hybridization. The methodology allows one to infer combinations of genomic aberrations associated with the cancer phenotype. The stochastic model conjoins two features of cancer biology to infuse some context into an otherwise unsupervised learning problem. It asserts random genomic instability in a potential progenitor cell, followed by selection into a tumor of the descending cell lineage if the lineage experiences certain ensembles of genomic aberration. Disease heterogeneity is reflected in the possibility of a network containing multiple ensembles. The network of ensembles is an identifiable parameter. By forming the sampling model conditionally on selection, statistical dependencies (both positive and negative) can be induced between aberrations, and the model entails heterogeneity in the marginal rate of occurrence of aberrations. A double-Pólya distribution is introduced as a prior over the network of ensembles, and Markov chain Monte Carlo is developed to enable posterior computation. As an example, the methodology is used to reanalyze genomic aberrations from 116 renal cell carcinomas. It produces posterior probabilities that any given aberration is relevant to oncogenesis, posterior probabilities that pairs of aberrations reside in a common ensemble, and a point estimate of the network of ensembles. The methodology provides a model-based clustering of all measured aberrations according to these estimated ensembles and a model-based clustering of tumors according to the probable ensembles of genomic aberration that they have experienced. Although it is formulated here to analyze aberrations in cancer genomes, the instability-selection-network model may provide an approach to modeling dependence in correlated binary data on various biological systems. Limitations and possible extensions of the methodology are discussed.

KEY WORDS: Comparative genomic hybridization; Correlated binary data; Ensemble of genomic aberration; Genetic instability; Markov chain Monte Carlo; Model-based clustering; Selection.

1. INTRODUCTION

The extensive body of research in cancer biology shows unequivocally that cancer tumors exhibit a wide variety of aberrations in the organization and content of their genomes as compared with the genomes of normal cells. Advances in measurement technology allow investigators to record these aberrations at ever-increasing levels of resolution, and efforts to catalog these aberrations have been critical to understanding the heterogeneity of cancer, guiding studies of tumor biology, and enhancing diagnosis and treatment (Knuutila et al. 1998, 1999). It is a statistical problem to identify patterns in these data that may have some biological significance.

Comparative genomic hybridization (CGH) is a technique used to measure changes in DNA copy number created by cancerous tumor growth (Kallioniemi et al. 1992; Gray and Collins 2000). Briefly, genomic DNA obtained from tumor cells is labeled with a fluorescent tag and combined with differently labeled genomic DNA from normal cells. The mixture is competitively hybridized to immobilized probe DNA that is formed in the original chromosome-based CGH from a set of metaphase chromosomes anchored to a glass slide. Lasers excite the complex and enable the measurement of relative abundance of the two source DNAs at each chromosomal locus. Deletion of a genomic region in the tumor cells is indicated if the tumor channel fluoresces at a relatively low level, and amplification is indicated otherwise. Signal-processing techniques reduce the fluorescence data from raw quantitative intensities to discrete estimates of DNA copy number along the genome for each tumor (e.g., Piper et al. 1995; Carothers 1997). Quite often, preliminary analyses further reduce the data to aberrations at the resolution of a chromosome arm. Figure 1 is a graphical representation of CGH profiles from 116 renal cell carcinomas (RCCs) collected by H. Moch and colleagues at the Institute of Pathology, University of Basel (Jiang et al. 2000). Following Jiang et al., here the data have been preprocessed to the point shown in the figure; that is, there are $n = 52$ distinct aberrations (amplifications and deletions) that either occur (dark shading) or do not occur in each of the 116 tumors. The marginal empirical frequency (EF) of each aberration is recorded in Table 2. For example, the aberration $-3p$ (i.e., deletion of some genomic DNA on the short arm of chromosome 3), the most frequent aberration, occurs in 72 of the 116 tumors. (Note that CGH data are different from microarray gene expression data, which have received so much attention in recent statistical literature. Both kinds of measurement involve hybridization, but CGH measures genomic DNA rather than the expression of messenger RNA.)

CGH profiling provides an approach to mapping cancer genes. Oncogenes are genes whose activation is associated with cancer development; such genes may be present at excess copy number in tumor DNA. Similarly, tumor-suppressor genes regulate or maintain normal cell function,

Michael A. Newton is Professor, Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison WI 53792 (E-mail: *newton@stat.wisc.edu*). The author thanks A. Schäffer for discussing oncogenic tree models, and for kindly providing the renal cell carcinoma data set originally obtained by F. Jiang and H. Moch, supported by Swiss National Science Foundation grant 97.3100-050752 and Krebsforschung Schweiz grant KFS 367-9-1996. W. F. Dove, C. Kendziorski, B. Larget, A. Moser, A. Schäffer, and anonymous referees and editors provided insightful comments leading to significant improvements in this study, which was supported by National Institutes of Health grant R01 CA64364. The author also thanks David Dahl and Hyuna Yang for research assistance, and R. Roylance for providing CGH profiles of breast tumors with which the methods could be further tested. All computations were carried out using the **R** system, version 1.3.0, *http://www.r-project.org/*. An **R** library of functions for performing ISN model calculations is freely available at *http://www.stat.wisc.edu/~newton/*.

© 2002 American Statistical Association
Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Applications and Case Studies
DOI 10.1198/016214502388618726

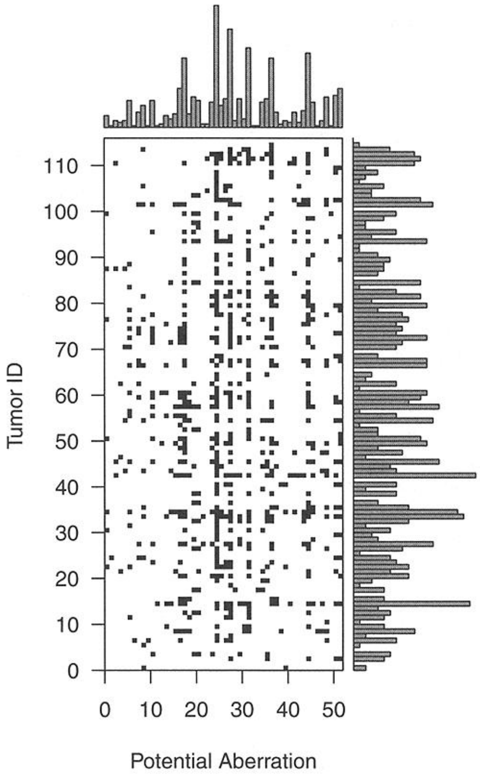

Figure 1. CGH Profiles: Moch's RCC Data.

and their inactivation (by, e.g., deletion) is associated with the cancer phenotype. By this reasoning, oncogenes and tumor-suppressor genes may reside in genomic regions of frequent amplification or deletion. Kainu et al. (2000) combined CGH data and more traditional linkage data to implicate a genomic region as harboring a putative breast cancer susceptibility gene. Earlier, Hemminki et al. (1997) used this strategy to study Peutz–Jeghers syndrome. Beyond gene mapping, CGH data may provide insights into the etiology of cancer, as the work of Roylance et al. (1999) has demonstrated. In all of these efforts, an ongoing challenge is to distinguish important genomic aberrations from noise accumulated in genetically unstable tumors (Gray and Collins 2000).

Elementary data analysis methods are often used to identify single genomic regions that are either amplified or deleted at an unusually high rate in the tumors under study. For instance, an investigator might graph a barplot of empirical aberration frequencies and then report a few of the most commonly aberrant regions (e.g., the horizontal barplot in Fig. 1). Some investigators have recognized the statistical nature of this type of inference and have adopted procedures that try to separate oncogenic signal from sporadic (neutral) changes caused by genetic instability (Brodeur, Tsiatis, Williams, Luthardt, and Green 1982; Newton, Wu, and Reznikoff 1994; Jarrard et al. 1999; Jiang et al. 2000).

As larger CGH datasets are obtained, it is becoming clear that dependencies exist among events in different genomic regions. Underlying this observation is the notion that certain combinations of genetic events have some biological significance in the context of cancer development. Several genomic aberrations could correspond to elements of a multistep pathway that brings a normal cell lineage to its observed tumorigenic state, and thus, because some tumors follow that pathway, positive dependence among these aberrations may be expected. Likewise, disease heterogeneity could be a source of negative dependence among aberrations. Tumors that are analyzed together because they are morphologically and histopathologically homogeneous may be molecularly distinct and may have arisen through different developmental pathways.

The pioneering work of R. Desper, A. Schäffer, C. Papadimitirou, and colleagues has addressed the problem of how to analyze dependencies among multiple genomic aberrations measured by CGH. This group proposed tree models for oncogenesis and computational approaches to analyze CGH profiles (Desper et al. 1999, 2000; Jiang et al. 2000). Essentially, these approaches rely on empirical pairwise and marginal event frequencies among a selected set of marginally most frequent aberrations. The pairwise event frequencies may be transformed into pairwise distances and then processed by a distance-based phylogenetic tree-building algorithm. In a second method, the data are used in a *maximum weight branching algorithm* to reconstruct a different kind of tree in which both internal nodes and leaf nodes correspond to aberrations. Both of these methods are readily applied and provide an informative view of positive associations present in the data. For example, tree-based calculations were central in the analysis of a breast cancer susceptibility gene by Kainu et al. (2000).

Insofar as the tree-based methods are estimating underlying properties of a cancer, it would seem to be beneficial to have some measures of uncertainty or standard error associated with the tree estimates, but these are not yet available. Also, existing methods do not cope with potential negative associations among aberrations, or the possibility that the aberrations are not arranged in a tree structure. The methods reduce the CGH data to marginal event frequencies and pairwise distances, and thus lose potentially important higher-order information. Further, the methods require that the relevant aberrations to be placed on the tree be preselected on the basis of their marginal frequencies. If there is significant disease heterogeneity, then such preselection might omit from consideration aberrations that affect a subset of tumors. Nevertheless, the tree-based calculations represent a significant methodology for studying dependencies among genomic aberrations, and they go far beyond earlier efforts that considered aberrations to arise independently (e.g., Newton et al. 1994).

Presented here is a complementary and rather more statistical approach to the problem of analyzing multiple genomic aberrations. The approach involves a joint probability distribution for the measured CGH profiles—one that is derived from some elementary structural features of cancer biology. The joint distribution is parameterized by *ensembles* of genomic aberrations. Each ensemble is an unordered collection of aberrations whose co-occurrence in a progenitor cell lineage is somehow beneficial to the tumor, in a sense specified in Section 2. Elements of an ensemble may correspond to elements of a multistep pathway of tumor development, although the methodology presented here makes no attempt to infer the order in which aberrations occur in such a pathway. The possibility of disease heterogeneity corresponds to the possibility of multiple ensembles. Together, multiple ensembles of genomic aberrations make a network object that determines the distribution of the CGH data. Owing to the complexity of the parameter space of networks, a Bayesian inference strategy is adopted and posterior computations are implemented using Markov chain Monte Carlo (MCMC). This not only enables the calculation of point estimates, but also provides a full posterior distribution describing uncertainty in aspects of the network. The range of posterior inferences is quite extensive and includes model-based clustering of both aberrations and tumors. Further, the proposed method does not reduce data to pairwise summaries; rather, the full joint information is encoded in a likelihood function. Preselection of relevant abnormalities is not required either; rather, the relevant abnormalities are inferred simultaneously with the network structure. The method rests on stochastic elements of CGH data, allowing sporadic aberration and measurement error. For technical reasons, the calculations presented here are limited to nonoverlapping ensembles; Section 5 discusses this limitation and ways to overcome it.

The stochastic model central to my inference calculations is constructed in Section 2. Part of the rationale for the model is that it encodes sampling properties that are evident in CGH data. These properties are reviewed in Section 3. Model-fitting techniques are summarized in Section 4, and are applied to the RCC data in Section 5. The inferences obtained by instability-selection modeling are compared with inferences obtained in earlier analyses. A brief discussion follows.

2. INSTABILITY-SELECTION-NETWORK MODEL

To start, list the potential aberrations to be measured by CGH as $\{1, 2, \ldots, n\}$ and let $x = (x_1, x_2, \ldots, x_n)$ denote the CGH profile from one tumor, where x_i is a binary indicator for the ith aberration (i.e., x is one row in Fig. 1). Simply, $x_i = 1$ means that the aberration occurs and $x_i = 0$ means that it does not occur. In the RCC example, $n = 52$ (after some initial preprocessing); the data are recorded at the rather coarse resolution of the chromosome arm, and there are separate records for amplifications and deletions. The vector x describes for each arm whether or not a deletion was observed in the tumor cells and also whether or not an amplification was observed. Here x is viewed as the realization of a random vector $X = (X_1, X_2, \ldots, X_n)$ whose joint probability distribution represents all that can be known about the cancer under study from the CGH data.

In the instability-selection-network (ISN) model, the joint probability distribution $p(x)$ is parameterized by a network, $\mathcal{C} = (C_0; \{C_1, C_2, \ldots, C_K\})$, and two scaler parameters, α and β, both in $(0, 1)$. The network consists of K ensembles $C_1, \ldots, C_K$, each of which is a subset of $\{1, 2, \ldots, n\}$. An aberration i is said to be *relevant* if it is in some ensemble; otherwise, it is *neutral*. The special set C_0 in the definition of the network $\mathcal{C}$ is the collection of all neutral aberrations. By definition, C_0 is disjoint from every ensemble C_k, $k = 1, 2, \ldots, K$. It is meaningful to allow different ensembles to overlap, but at present calculations are feasible only in the special case of nonoverlapping ensembles. In what follows, therefore, $\{C_0, C_1, \ldots, C_K\}$ forms a set partition of $\{1, 2, \ldots, n\}$. (See Section 6 for more on this restriction.) Using the nonoverlapping ensembles assumption, the joint probability mass function for an aberration profile $X = (X_1, \ldots, X_n)$ becomes

$$p(x) = \alpha^{\sum_i x_i}(1-\alpha)^{\sum_i (1-x_i)} \left\{ \frac{1 - \prod_{k=1}^{K}(1-\beta^{t_k})}{1 - \prod_{k=1}^{K}(1-\theta^{m_k})} \right\}, \qquad (1)$$

where $t_k = \sum_{i \in C_k}(1 - x_i)$, m_k is the cardinality of ensemble C_k, and $\theta = 1 - (1-\alpha)(1-\beta)$.

The joint distribution (1) is derived by noting that for x to have been observed, the cell lineage in which x occurred must have survived in the tumor cell population to the time of observation. Some well-accepted cancer biology is encoded mathematically; genetic instability creates somatic genomic aberrations, and cell-level selection subsequently determines whether or not the affected lineage will pass descendants into an observable tumor. [The ideas of instability (see, e.g., Lengauer, Kinzler, and Vogelstein 1998) and selection (see, e.g., Tomlinson, Novelli, and Bodmer 1996) have deep roots in the cancer literature.] Thus $p(x)$ is really a conditional probability of a profile given that the progenitor cell lineage, having incurred damage x, is selected by oncogenesis to exist at the observation time. Keeping things simple, the model makes no claims about the size of the tumor cell population or anything about the cell division dynamics, which seems reasonable in light of the information available from the CGH profiles.

The exact form (1) is determined by a particular model for the instability and selection components. The instability component is a simple model for neutral, random genomic damage (Volpe 1990). I say that a potential aberration i can occur either overtly, in which case $X_i = 1$, or covertly, in which case $Y_i = 1$, or both. Here $\{X_i\}$ are taken to be independent and identically distributed (iid) Bernoulli trials with success probability α. Independently, $\{Y_i\}$ are iid Bernoulli(β) trials. The overt damage $X = (X_1, \ldots, X_n)$ is potentially observable, but the covert damage is completely unobservable; it represents forms of damage that may occur but that are not measurable by CGH, such as suppressor gene silencing, somatic recombination, or other factors. In summary, the instability component entails two forms of random damage that occur independently of each other and independently across the genome. Aberration i occurs somehow with probability $\theta = 1 - (1-\alpha)(1-\beta)$.

The random genetic instability is filtered by the process of cellular selection. Selection, denoted by SEL, occurs if for some ensemble C_k, genetic instability causes all aberrations $i \in C_k$ to occur. Otherwise, SEL does not occur, and no tumor becomes available to be observed (Fig. 2). To emphasize a point made earlier, the ensemble C_k is a collection of aberrations whose co-occurrence in a progenitor cell lineage is beneficial to the tumor. Also, the joint distribution $p(x)$ in (1) is seen to be a conditional distribution $p(x) = P(X = x|\text{SEL})$. It is interesting that Bayes rule is used here to derive the sampling model for data (see App. A).

3. PROPERTIES OF THE MODEL

The ISN model captures a range of statistical properties present in real data. By filtering through the event SEL, interesting *nonrandom* features emerge in $p(x)$ that are not

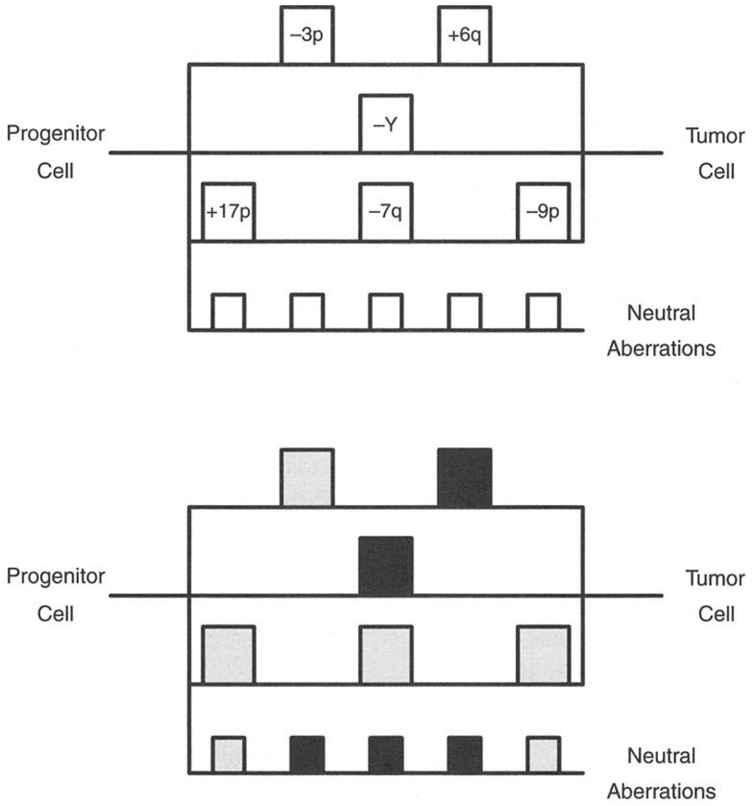

Figure 2. Schematic of the ISN Model. For this hypothetical cancer, disease heterogeneity is represented by three ensembles, containing a total of six relevant aberrations, indicated by boxes in the top three rows of each panel. Other aberrations are neutral. A potential progenitor cell (left) that incurs all of the aberrations in at least one ensemble will become the ancestor of cells in an observable tumor (right). That is, if all boxes in one row are gray (indicating that aberration occurs, either overtly or covertly), then the full complement of overt aberrations in the corresponding tumor can be measured. There is no tumor to be measured otherwise. Evidently, genetic instability in this example leads to a tumor by a pathway containing the $\{+17p, -7q, -9p\}$ *ensemble (lower panel).*

present in the instability component before SEL. For example, there is heterogeneity in the marginal aberration rate. For any relevant aberration i, $P(X_i = 1|\text{SEL}) > \alpha$, and this probability decreases as the size of the corresponding ensemble increases, making large ensembles difficult to detect. Naturally, $P(X_i = 1|\text{SEL}) = \alpha$ for neutral aberrations $i \in C_0$, and these aberration indicators are independent of all other measurements. Selection also induces dependencies between relevant aberrations. When there are multiple ensembles (i.e., $K > 1$), the following results hold:

1. $\text{cov}(X_i, X_j|\text{SEL}) > 0$ if i and j are in the same ensemble.
2. $\text{cov}(X_i, X_j|\text{SEL}) < 0$ if i and j are in different ensembles.

See Appendix B for proofs.

Figure 3 shows sampling properties of profiles obtained on the hypothetical cancer in Figure 2. The values $\alpha = .10$ and $\beta = .05$ were used, and 10^4 aberration profiles were simulated according to (1). As predicted by theory, the simulation shows both heterogeneity of marginal rates and negative and positive covariance between aberrations.

From the perspective of model development, it is interesting to ask whether or not the network of ensembles is identifiable—that is, do two different networks necessarily correspond to different joint distributions (1). If so, then it is known, for example, that the sequence of posterior distributions over network space computed from an ever-growing sample of CGH profiles will concentrate on the true underlying network, in the context of the ISN model. Appendix B sketches the proof of identifiability in the case where α and β are known.

4. BAYESIAN ANALYSIS

4.1 Overview

The likelihood function from a set of CGH profiles $x^1, x^2, \ldots, x^N$ is obtained naturally as the product

$$L(\mathcal{C}, \alpha, \beta) = \prod_{s=1}^{N} p(x^s),$$

where $p(x)$ is as in (1), $\mathcal{C}$ denotes the unknown network of ensembles, (α, β) are rates of overt and covert damage, and N is the number of tumors ($N = 116$ in the RCC example.) An effective approach to extracting information from this likelihood is to form the posterior distribution over the parameter space,

$$\pi(\mathcal{C}, \alpha, \beta|x^1, \ldots, x^N) \propto L(\mathcal{C}, \alpha, \beta)\pi(\mathcal{C}, \alpha, \beta), \qquad (2)$$

where $\pi(\mathcal{C}, \alpha, \beta)$ is a prior distribution to be specified. All inferences arise from this posterior distribution. One may attempt to find the global maximum [the maximum a posteriori (MAP) estimate], although the extent of posterior uncertainty

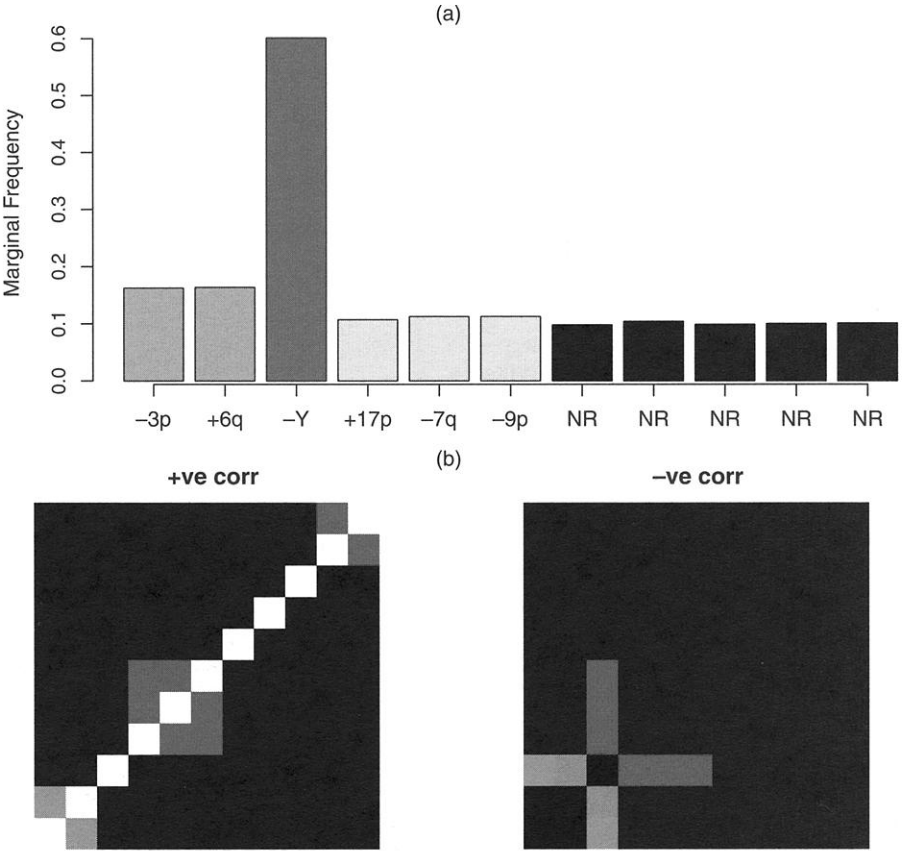

Figure 3. Sampling Properties of Data From the Hypothetical Network in Figure 2. (a) A barplot indicating the marginal aberration frequency for each potential aberration. Gray levels distinguish aberrations in different ensembles; black corresponds to neutral aberrations. (b) The correlation between pairs of aberrations. Black corresponds to 0, and lighter shades indicate increased correlation magnitude (positive correlation on the left; negative correlation on the right). Positive correlation within ensembles and negative correlation between ensembles is indicated. Plots are based on 10^4 simulated tumors.

may require that additional features be reported. Reporting several marginal posterior summaries is useful. For instance, the posterior probability that aberration i is relevant [i.e., $P(i \in C_k$, some $k \geq 1|\text{data})]$ can be approximated by the empirical frequency of this event in networks sampled from the posterior. Ensemble information is contained in an interesting summary matrix that has, for each pair (i, j), the posterior probability that both i and j are relevant and in the same ensemble. Typically, the posterior distribution is also computed for the number of ensembles, K, and the number of relevant aberrations, $m = \sum_{k=1}^{K} m_k$.

4.2 Prior

A prior distribution for the network $\pi(\mathcal{C})$ is specified separately from that of the rate parameters $\pi(\alpha, \beta)$, and the full joint prior is obtained by multiplication, thus encoding prior independence. A prior restriction that the covert aberration rate β be smaller than the overt aberration rate α is not necessary theoretically, but it is helpful, because the likelihood can become quite flat if β is too large. So this restriction is taken, but otherwise the prior for (α, β) is uniform. The prior $\pi(\mathcal{C})$ contains a single hyperparameter, $\tau > 0$, which affects the amount of clustering expected in the network $\mathcal{C}$. This is called a *double-Pólya* prior; one piece governs the set partition, and another piece governs whether or not potential aberrations are relevant. The functional form is

$$\pi(\mathcal{C}) = \frac{\tau^K \Gamma(\tau)\left[\prod_{k=1}^{K} \Gamma(m_k)\right]}{\Gamma(\tau + m)} \frac{\Gamma(m+1)\Gamma(n-m+1)}{\Gamma(n+2)}, \quad (3)$$

where K is the number of ensembles, m_k is the size of ensemble C_k, $m = \sum_{k=1}^{K} m_k$ is the total number of relevant aberrations, $n - m$ is the number of neutral aberrations, and $\Gamma()$ is the gamma function. Note two facts: (1) the induced prior distribution for the network size m is uniform between 0 and the total number of potential aberrations n, but still (2) the prior tends to penalize larger networks.

For computational reasons, a simple form of data augmentation is used to represent the network $\mathcal{C}$. $\mathcal{C}$ is represented using two vectors: a binary relevance vector $a = (a_1, \ldots, a_n)$ characterizing C_0 and a label vector $c = (c_1, \ldots, c_n)$ encoding ensemble structure. Thus $a_i = 0$ means $i \in C_0$, and $a_i = 1$ means that i is relevant. Elements c_i reside in some label space (here the unit interval). The particular numerical values have no meaning beyond serving as labels for ensembles. Two relevant aberrations, i and j, are in the same ensemble iff $c_i = c_j$, and so clearly one can obtain $\mathcal{C}$ from the pair (c, a). Use of the pair (c, a) constitutes a mild form of data augmentation, because there is superfluous information in the labels on neutral aberrations. The Markov chain calculations are easier to set up in this augmented space. Interestingly, the prior (3) on

$\mathcal{C}$ is induced by a simple prior on the pair (c, a) in which c and a are independent, c has a uniform–Pólya prior, and a has a Bernoulli–Pólya prior (see App. C).

4.3 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) may be the only practical approach to posterior analysis (e.g., Tierney 1994; Gilks, Richardson, and Spiegelhalter 1996), but several factors make it difficult to construct an effective MCMC sampling tool. For one thing, the network $\mathcal{C}$ resides in an enormous discrete space. The number of networks is greater even than the Bell number of set partitions (e.g., van Lint and Wilson 1992, p. 105), owing to the labeling of some aberrations as neutral. For example, with $n = 11$ potential aberrations as in Figure 2, there are 4,213,596 different networks. The likelihood function also creates certain irregularities that make obvious choices for move types quite ineffective. For instance, two networks that differ by one large ensemble may have much closer likelihoods than otherwise similar networks that differ by a small ensemble.

Recall that an MCMC implementation involves realizing a Markov chain $s_1, s_2, \ldots, s_B$ in the space of the unknown (augmented) state $s = (c, a, \alpha, \beta)$. Each of the $B - 1$ scans is built from a series of proposal-test steps, and on such a step, a proposal state s^* is drawn from some distribution $q(s, s^*)$. The Metropolis–Hastings ratio is

$$r = \frac{\pi(s^*|\text{data})q(s^*, s)}{\pi(s|\text{data})q(s, s^*)},$$

where $\pi(s|\text{data})$ is the data-augmented posterior associated with (2). With probability $\min(r, 1)$, the chain moves to s^*; otherwise, it continues at s. Typically the chain is run for several million scans, subsampling to reduce the number of states used in posterior calculations.

The proposed implementation has two move types affecting the vector c, two move types affecting the relevance vector a, and a small-box uniform random-walk proposal affecting the rate parameters α and β. Appendix C gives details of the network move types.

The potential for mistakes in a complex MCMC implementation merits a series of basic tests. One test is to remove the likelihood component and use the algorithm to simulate the prior distribution. Another is to run the calculations on simulated data. Newton (2001) presented a range of such checks; one is reported here. The simulation involved a network like the one in Figure 2, $(K = 3,\ m = 6)$, but with many more $(n - m = 44)$ neutral aberrations, and thus a total of $n = 50$ potential aberrations. Four datasets each comprising 100 aberration profiles were generated according to (1) using $\alpha = .10$ and $\beta = .05$. Two independent MCMC runs of length 500,000 were applied to each of the four datasets and were subsampled every 500 scans, giving a sample of 1,000 networks for each run. This replication provides some information about Monte Carlo error in the MCMC in addition to sampling error from data. Summaries from each run included the MAP estimate of the network, the vector of marginal posterior probabilities that each aberration is relevant, and estimates of the overt/covert aberration rates.

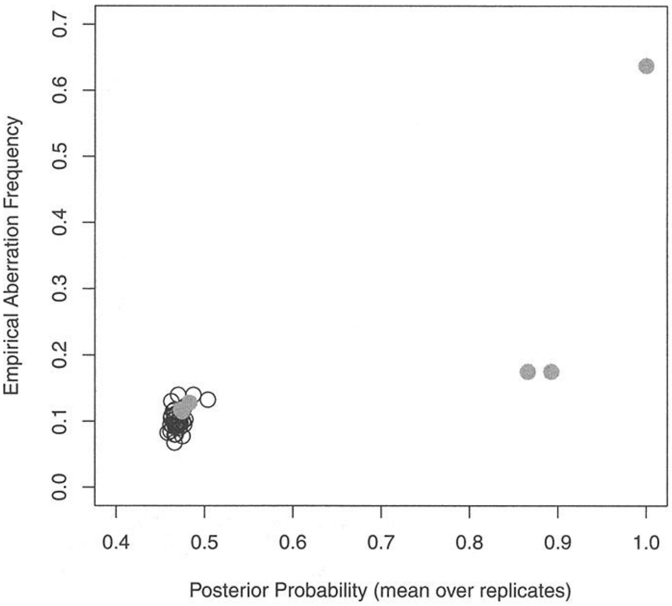

Figure 4. MCMC Results, Simulated Data. For each of $n = 50$ aberrations, the empirical aberration frequency versus the average marginal posterior probability of relevance across the 8 MCMC runs is plotted. Solid (gray) dots indicate the six relevant aberrations.

As an overall summary, Figure 4 compares the average (over 8 runs) marginal posterior probability of relevance to the empirical aberration frequency for all $n = 50$ potential aberrations. The six relevant aberrations are indicated by solid shaded circles. Indeed, the association between the posterior probability and the empirical frequency is quite strong, as is to be expected. The aberration $-Y$ in this hypothetical network is in an ensemble by itself and presents the highest marginal rate of occurrence. It also appears in every one of the 8,000 posterior-sampled networks. The aberrations $-3p$ and $+6q$ constitute another ensemble; their marginal rate of occurrence is high, and in datasets of 100 tumors they tend to show high posterior relevance probability. In contrast, the three aberrations on the largest ensemble have statistics close to the background and do not show high posterior relevance probability. Further summaries, not shown, indicate that the Monte Carlo error of the MCMC tool is fairly low.

5. DATA ANALYSIS

Using the ISN methodology, this study reanalyzed CGH profiles from 116 RCCs collected by H. Moch and colleagues at the Institute of Pathology, University of Basel (Jiang et al. 2000). Following Jiang et al., data on arms 1p, 16p, 19p, 19q, and 22q and Y were excluded from the analysis because of potential inaccuracies in the CGH measurements on these arms. This leaves 36 chromosome arms for which there is both amplification and deletion information. Thus there is a total of $n = 72$ potential aberrations. As a minor filter, a set of 20 potential aberrations that never occurred in the dataset was removed from further consideration, reducing the problem to $n = 52$ (Fig. 1). In the first calculations reported here, the data were not reduced further to a set of $n = 12$ *nonrandom* aberrations as was done by Jiang et al. (2000) using the method of Brodeur et al. (1982); instead, the analysis considered the full spectrum of observed abnormalities.

Table 1. Network Estimates, RCC Example

τ	Run	Ensemble	Relevant aberrations							Mean Pr(Relevant)
1	a	C_1	+17q	−3p	−4q	−6q	−9p	−13q		1.00
1	a	C_2	+Xp	+Xq	−8p	−8q	−Xp	−Xq	−3q	.81
1	b	C_1	+17q	−3p	−4q	−6q	−9p	−13q		1.00
1	b	C_2	+Xq	+Xp	−8p	−8q	−Xp	−Xq	−3q	.77
5	a	C_1	+17q	−3p	−4q	−6q	−9p	−13q		1.00
5	a	C_2	+Xq	+Xp	−8p	−8q	−Xp	−Xq	−3q	.72
5	b	C_1	+17q	−3p	−4q	−6q	−13q	−9p		1.00
5	b	C_2	+Xp	+Xq	−8p	−8q	−Xp	−Xq	−3q	.74
10	a	C_1	−3p	−4q	−9p	−13q	+17q	−6q		1.00
10	a	C_2	+Xq	+Xp	−8p	−8q	−Xp	−Xq	−3q	.49
10	b	C_1	+17q	−3p	−4q	−6q	−9p	−13q		1.00
10	b	C_2	+Xq	+Xp	−8p	−8q	−Xp	−Xq	−3q	.55

NOTE: τ is a hyperparameter, "Run" indicates the two independent MCMC realizations, and "Ensemble" labels the two different ensembles found in the MAP network in each case. Aberrations are sorted from left to right by decreasing marginal probability of relevance. The final column shows the average (across aberrations) posterior probability of relevance for those aberrations in the particular ensemble.

It is straightforward to demonstrate that patterns of rate heterogeneity and dependence exist in these data to a much greater degree than would be expected by chance alone. Some elementary permutation tests were performed to demonstrate this (data not shown). One test involved shuffling the columns in Figure 1 and repeatedly recomputing the sample covariance among potential aberrations. The magnitude of positive sample correlations was particularly significant. The use of permutation procedures provides an initial assessment of structure in data and naturally precedes more elaborate model-based calculations (e.g., Besag and Clifford 1989).

Next, the MCMC sampler was applied to fit the ISN model, using two replicate runs on each of three different hyperparameter values, $\tau = 1, 5$, and 10. These values capture a broad range of prior variation over network space. Large ensembles are expected for $\tau = 1$ (e.g., on average fewer than 5 ensembles are expected given 50 relevant aberrations). With $\tau = 10$, more smaller ensembles are expected (see fig. 4 of Newton 2001). Chains of length $B = 2 \times 10^6$ were initiated at a random network and were subsampled every 1,000 scans, yielding samples of 2,000 states for output analysis. There was quite good mixing of the chains. (See Newton 2001 for some output analysis.)

There is a striking perfect agreement across runs and priors in the MAP estimate of the network (Table 1). The estimate contains $\hat{m} = 13$ potential aberrations arranged in two ensembles. As expected, these relevant aberrations exhibit very high empirical frequency of occurrence. Table 2 compares the EF with the marginal posterior relevance probability for all 52 potential aberrations. Here the replicate runs are combined, but the results from different priors are shown separately. The numerical value of posterior relevance probability is somewhat sensitive to the prior, although the most relevant abnormalities are clear in each case.

Because the calculations attempt to use all joint information in the sample, the posterior relevance probability is not perfectly correlated with marginal EF. A useful posterior summary that provides some dependence information is the pairwise probability that a given pair of potential aberrations i and j are both relevant and in the same ensemble. Treating the resulting $n \times n$ matrix as a similarity matrix, distance-based hierarchical clustering was applied to obtain the trees shown in Figure 5. Such posterior probability trees may provide some useful inferences beyond the simple point estimates of Table 1. For one thing, they tend to be stable as the prior τ changes. They all strongly indicate a single ensemble, $C_1 = \{-13q, -9p, -6q, -4q, +17q, -3p\}$, but yet provide evidence for and structure of a second ensemble. Notably, the potential aberrations that are probably neutral appear bunched together and are well separated from those that are probably relevant. One can find a signal much more simply by clustering these posterior probabilities than by running, say, a hierarchical clustering on the raw CGH profiles.

The model formulation provides a range of data analysis possibilities. Figure 6 shows one example. The MAP network from Table 1 is taken as fixed, and for each tumor

Table 2. Marginal Posterior Relevance Probabilities, RCC Data

i	EF	1	5	10	i	EF	1	5	10
+1q	7	.40	.22	.04	−4p	17	.40	.22	.04
+2p	1	.41	.22	.04	−4q	58	1.00	1.00	1.00
+2q	4	.42	.22	.04	−5p	4	.48	.32	.12
+3q	3	.49	.33	.12	−5q	17	.60	.48	.26
+5p	4	.42	.23	.04	−6p	5	.40	.22	.02
+5q	16	.44	.24	.07	−6q	47	1.00	1.00	1.00
+6p	1	.40	.21	.03	−7p	1	.42	.20	.03
+7p	9	.44	.24	.06	−7q	1	.48	.32	.12
+7q	13	.44	.28	.10	−8p	15	.95	.96	.73
+8p	1	.40	.22	.03	−8q	17	.84	.80	.60
+9q	16	.50	.34	.14	−9p	41	1.00	1.00	1.00
+10q	1	.40	.22	.03	−9q	9	.48	.32	.12
+11p	2	.42	.20	.03	−10p	2	.40	.22	.04
+11q	7	.40	.20	.03	−10q	4	.42	.24	.04
+12q	5	.41	.22	.03	−11p	3	.39	.22	.03
+16q	8	.46	.29	.10	−11q	9	.42	.22	.04
+17p	23	.43	.24	.04	−12p	3	.40	.22	.04
+17q	41	1.00	1.00	1.00	−12q	10	.41	.20	.03
+20q	8	.41	.22	.03	−13q	44	1.00	1.00	1.00
+Xp	18	.96	.96	.74	−14q	15	.43	.26	.06
+Xq	16	.96	.96	.74	−15q	1	.40	.22	.03
−1q	2	.40	.21	.03	−18p	4	.42	.22	.03
−2p	2	.39	.22	.04	−18q	18	.42	.22	.04
−2q	15	.53	.39	.18	−20q	1	.41	.22	.03
−3p	72	1.00	1.00	1.00	−Xp	19	.68	.58	.40
−3q	13	.52	.38	.18	−Xq	23	.60	.45	.29

NOTE: i indicates the aberration out of $n = 52$, EF stands for empirical frequency of occurrence of the aberration in 116 profiles, and the columns 1, 5, and 10 indicate the hyperparameter value τ.

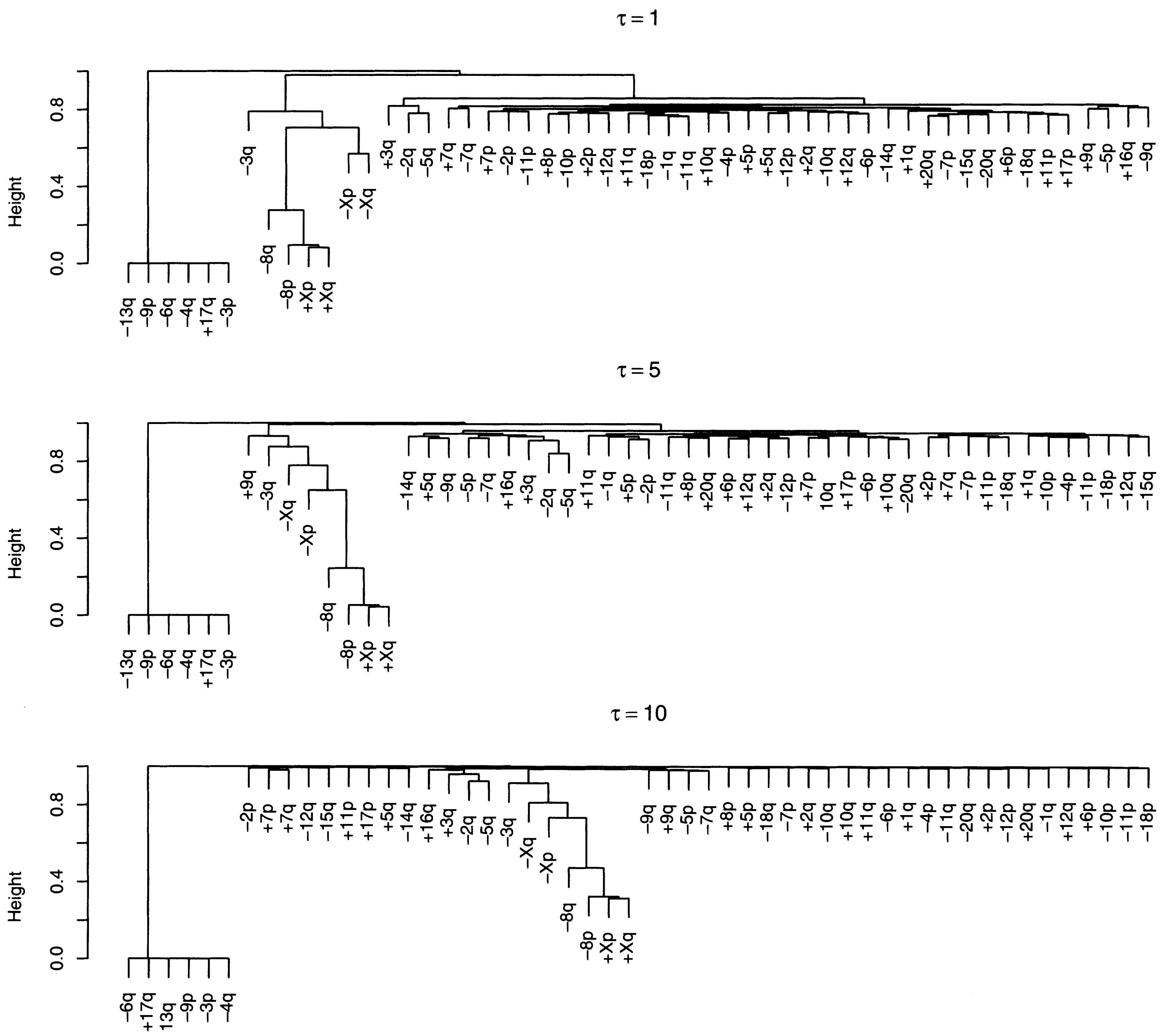

*Figure 5. Tree Summary of Pairwise Posterior Probabilities. Leaves correspond to potential aberrations. Aberrations plotted at a height close to 0 are probably in an ensemble with neighboring aberrations. Computations used default settings in the hclust function in **R**.*

the posterior probability that it experienced all aberrations in each of inferred ensembles, given the observed CGH profile, is determined. (This is simply the covert rate β raised to the number t_k of requisite covert aberrations for that ensemble.) The figure shows potential aberrations rearranged according to the two ensembles C_1 and C_2 and the neutral aberrations C_0. Then the tumors are reorganized according to the probability that they experienced all of the aberrations for that ensemble. It can be inferred that 102 of the 116 tumors probably experienced C_1, 8 probably experienced C_2, and C_1 and C_2 were tied for another 6 tumors. Effectively, the computations provide a model-based clustering of both the aberrations and the tumors. In cases where the ensemble predictions are more balanced, one might use this information when attempting to correlate clinical outcomes with the genomic profiles.

In their analysis, Jiang et al. (2000) used the subset of $n = 12$ aberrations $\{-3p, -4p, -4q, -6q, -8p, -9p, -13q, -18q, -Xp, +17q, +Xp\}$ that were deemed significant by a preselection procedure. For the most part, these aberrations also have a high posterior probability of being relevant. One exception is $-18q$, which, although it occurs in 18 patients (16%), is probably neutral. As Jiang et al. had concluded, it can also be inferred that $-8p$ is probably relevant and is in a separate ensemble from some of the other important aberrations. On the other hand, the present analysis does not support the conclusion of Jiang et al. that there may be two groups of RCCs, one group characterized by $\{-6q, +17q, +17p\}$ and the other by $\{-9p, -13q, -18q\}$. This may be due to the fact that the proposed model does not allow overlapping ensembles (e.g., a network with these two ensembles each augmented by $-4q$ or something else might fit well), but it may also reflect

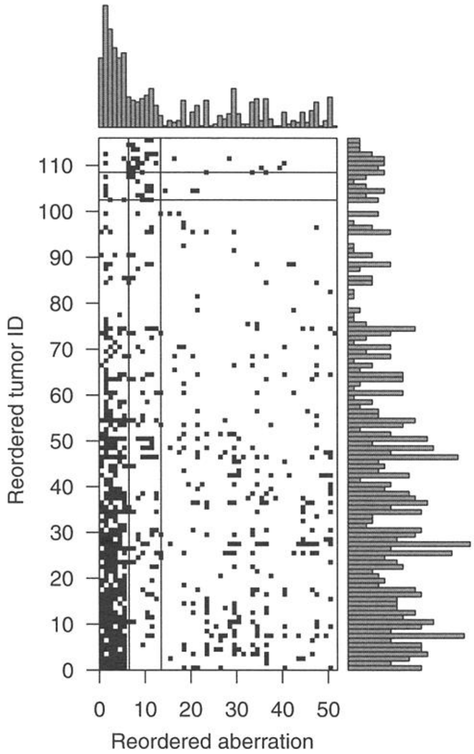

Figure 6. RCC Data Reorganized by Model Fit. Same raw data are used as in Figure 1, but the columns (aberrations) or organized from left to right according to the estimated ensembles, and the rows (tumors) are reorganized according to posterior probability of experiencing that ensemble. That is, in the first 102 tumors, ensemble C_1 (the left block of columns) has a higher probability of having been experienced than ensemble C_2 (the second block of columns). In the next six rows, probability is a tossup, and in the top eight rows, ensemble C_2 is more probable. Most aberrations (the right block of columns, C_0) are considered to be neutral.

the limited signal available in these data given the level of neutral aberration.

To extend the present analysis, posterior sampling was repeated with all three priors for the dataset formed by restricting to the specially selected $n = 12$ potential aberrations. Again, a striking consistency was seen across priors and MCMC runs in the MAP network estimate. This is a one-ensemble network comprising exactly those aberrations deemed to make up ensemble C_1 from the first analysis (see Table 1). So the methodology applied to the full set of aberrations picked up two potential ensembles, one of which was found again in a specially selected subset. Interestingly, the minor ensemble C_2 was not picked up in the small dataset; but such detection could very well have been hampered by the preselection, which is based only on empirical frequency. This preselection inflates the estimate of the neutral rate α, and thus makes it more difficult to identify relevant aberrations. From the large dataset, the estimate is $\hat{\alpha} = .07$, but for the reduced set, it is about $\hat{\alpha} = .20$. The estimate of α is more sensitive to the prior τ in the latter case.

One way to account somewhat for the nonoverlapping ensembles limitation of the model is to apply the methodology to certain data subsets. Of the 116 tumors, 72 exhibited the $-3p$ aberration. The foregoing MCMC calculations were applied to these 72 profiles in an attempt to infer the conditional distribution of CGH profiles given that $-3p$ has occurred. The MAP network has just a single ensemble, $C_1 = \{+17q, -4q, -6q, -9p, -13q\}$, for all runs and priors. Indeed, this is the very same dominant ensemble from the first analysis, with $-3p$ withdrawn. A more refined network structure could not be inferred, perhaps owing to the reduced sample size. Interestingly, the probable relevance of $-8p$ is low in this analysis; but of the eight tumors that probably experienced C_2, only two exhibited $-3p$. In other words, the probable relevance of $-8p$ is linked to tumors that had been removed in this conditional test.

6. CLOSING REMARKS

Advances in molecular technology are providing oncologists with an unprecedented view of the genomic abnormalities presented by cancerous tumors. The goal of this study was to develop model-based statistical tools that may help oncologists identify and characterize significant combinations of these aberrations. The model assumptions react to basic elements of cancer biology, and the resulting calculations provide a range of inferential summaries that may complement existing tree-based analysis methods. They infuse some biological context into an otherwise unsupervised learning task.

As with any such effort, there is a tension between complexity and validity of the stochastic model. This study has attempted to identify central biological elements affecting variation in CGH profiles so that the model formulation remains quite simple and yet still captures the key statistical features of CGH data. In striking this balance, certain decisions have been made that seem appropriate but that could be revisited in further work. For instance, the instability component of the model might be extended to allow different prior aberration rates between amplifications and deletions, or even among chromosome arms. [The method of Brodeur et al. (1982) makes a certain length-based adjustment, but this does not seem to be appropriate for the CGH profiles analyzed.] One could also allow a second form of measurement error in which measured aberrations might correspond to false positives. The network and selection structure could be enhanced. For example, one could allow a *bypass* path on which no measurements are taken but that corresponds to another way for a progenitor cell to become a tumor. The consistency of the conclusions obtained across different priors and using different subsets of the data provide some assurance that the present method gives reasonable results.

Statistical dependencies evident in CGH profiles have at least two sources. The type of dependence modeled in the present study is dependence induced by selection; aberrations $-3p$ and $+17q$, for example, are deemed to be correlated in RCC, because their joint occurrence in a potential progenitor cell enhances the probability that descendants of this cell will populate a tumor. The instability component of the model considers that these aberrations occur independently in the progenitor cell; the association that occurs is caused by selection. One thing that has not been accounted for is any dependence inherent in the instability component, which is carried through into the tumor. The physical process leading to chromosomal deletion, for example, may make the event $-8p$ more likely if

$-8q$ is known to have occurred, quite aside from any advantages provided in oncogenesis. This issue has been addressed in simpler versions of the instability-selection model (Newton Gould, Reznikoff, and Haag 1998; Newton and Lee 2000), but not yet with the network version. It will need to be addressed if this methodology is to be applied to high-resolution, array-based CGH.

A limitation of the present calculations is that the ensembles of genomic aberrations may not overlap. The problem can be partially alleviated by doing the kind of conditional analysis done herein, but quickly the sample gets subdivided too much to be informative. In principle, the proposed methodology will carry over to more general networks; one key problem is the problem of computing the likelihood function for a general network. Dependence between ensembles invalidates the sampling formula (1). An interesting fact may point to a solution. Consider the graph with m nodes equal to the set of relevant aberrations and with an edge between i and j if i and j are in a common ensemble. Evidently, the ensembles are cliques of this graph. Instability amounts to realizing random variables on the nodes of this graph, and selection amounts to checking each of the cliques to see whether the cell experiences all of the corresponding aberrations. Exploring these calculations is a topic of ongoing research.

APPENDIX A: DERIVATION OF SAMPLING PROBABILITIES

By Bayes's rule,

$$P(x_1,\ldots,x_n|\mathrm{SEL}) = P(x_1,\ldots,x_n)P(\mathrm{SEL}|x_1,\ldots,x_n)/P(\mathrm{SEL}),$$

where

$$\begin{aligned}
P(\mathrm{SEL}) &= P(\text{at least one ensemble is fully aberrant})\\
&= 1 - P(\text{no ensembles are fully aberrant})\\
&= 1 - \prod_{k=1}^{K} P(\text{ensemble } C_k \text{ is not fully aberrant})\\
&= 1 - \prod_{k=1}^{K} [1 - P(\text{ensemble } C_k \text{ is fully aberrant})]\\
&= 1 - \prod_{k=1}^{K} \left[1 - \prod_{i\in C_k} P(\text{aberration } i \text{ occurs})\right]\\
&= 1 - \prod_{k=1}^{K} \left[1 - \prod_{i\in C_k} \{1 - P(\text{aberration } i \text{ does not occur})\}\right]\\
&= 1 - \prod_{k=1}^{K} \left[1 - \prod_{i\in C_k} \{1 - (1-\alpha)(1-\beta)\}\right]\\
&= 1 - \prod_{k=1}^{K} [1 - \theta^{m_k}],
\end{aligned}$$

and where $\theta = 1 - (1-\alpha)(1-\beta)$ is the marginal probability of either covert or overt aberration. By a similar argument, with $x = (x_1,\ldots,x_n)$,

$$P(\mathrm{SEL}|x) = 1 - \prod_{k=1}^{K} (1 - \beta^{t_k}),$$

where $t_k = \sum_{i\in C_k} (1 - x_i)$.

APPENDIX B: SAMPLING PROPERTIES

Marginal Rates

Consider one ensemble, say C_1, containing m_1 aberrations, and let A be the event that every aberration in C_1 occurs somehow (i.e., either overtly or covertly,) and so $A \subset \mathrm{SEL}$. By the instability assumptions, $P(A) = \theta^{m_1}$, where $\theta = 1 - (1-\alpha)(1-\beta)$ is the probability that a given aberration occurs somehow. Take a particular aberration $i \in C_1$. It is straightforward to show that the overt damage X_i is conditionally independent of SEL given A, and also given A^c. (Ensemble-level information reveals everything one needs to know about selection.) Thus, with suggestive notation,

$$\begin{aligned}
p(x_i|\mathrm{SEL}) &= p(x_i|A)p(A|\mathrm{SEL}) + p(x_i|A^c)p(A^c|\mathrm{SEL})\\
&= \frac{p(A|x_i)p(x_i)}{p(A)} \frac{p(\mathrm{SEL}|A)p(A)}{p(\mathrm{SEL})}\\
&\quad + \frac{p(A^c|x_i)p(x_i)}{p(A^c)} \frac{p(\mathrm{SEL}|A^c)p(A^c)}{p(\mathrm{SEL})}\\
&= \frac{p(x_i)}{p(\mathrm{SEL})}\{p(A|x_i) + p(A^c|x_i)p(\mathrm{SEL}|A^c)\}.
\end{aligned}$$

Evaluating this at $x_i = 1$ gives

$$p(1|\mathrm{SEL}) = \frac{\alpha}{p(\mathrm{SEL})}\{\theta^{m_1-1} + (1-\theta^{m_1-1})p(\mathrm{SEL}|A^c)\}, \qquad \text{(B.1)}$$

where $p(1|\mathrm{SEL})$ is shorthand for $P(X_i = 1|\mathrm{SEL})$. This formula uses the fact that $P(A|X_i = 1)$ is the probability that the remaining $m_1 - 1$ aberrations also occur.

Now, taking a second aberration j in another ensemble, C_2, say, of size $m_2 > m_1$, and with B the event that every aberration in C_2 occurs, the difference of marginal aberration rates can be considered,

$$\begin{aligned}
d &= P(X_i = 1|\mathrm{SEL}) - P(X_j = 1|\mathrm{SEL})\\
&= \frac{\alpha}{p(\mathrm{SEL})}\{\theta^{m_1-1} + (1-\theta^{m_1-1})p(\mathrm{SEL}|A^c)\\
&\qquad - \theta^{m_2-1} - (1-\theta^{m_2-1})p(\mathrm{SEL}|B^c)\}.
\end{aligned}$$

The aim is to show that $d > 0$, and clearly it suffices to show that the quantity between the braces is positive. To do so, it is convenient to introduce the probability

$$\begin{aligned}
\psi &= P(\text{no other ensemble is fully aberrant})\\
&= \prod_{k\in\text{other}} (1 - \theta^{m_k}).
\end{aligned}$$

The product here is over any other ensembles different from C_1 and C_2, and m_k is the size of the kth such ensemble. If there are no other ensembles, then $\psi = 1$. Having ψ allows a connection between $p(\mathrm{SEL}|A^c)$ and $p(\mathrm{SEL}|B^c)$ to be made. Specifically,

$$p(\mathrm{SEL}|A^c) = 1 - (1 - \theta^{m_2})\psi$$

and

$$p(\mathrm{SEL}|B^c) = 1 - (1 - \theta^{m_1})\psi.$$

Thus to show $d > 0$, it suffices to show that

$$\theta^{a-1} + (1-\theta^{a-1})[(1-(1-\theta^b)\psi] > \theta^{b-1} + (1-\theta^{b-1})[(1-(1-\theta^a)\psi]$$

where $a = m_1$ and $b = m_2$ are introduced for notation. Expanding both sides and cancelling terms, this is equivalent to

$$(1-\theta^b)(1-\theta^{a-1}) < (1-\theta^a)(1-\theta^{b-1}).$$

This clearly holds for $b > a$ and $\theta \in (0, 1)$, and so $d > 0$. That is, in a given network, an aberration in a larger ensemble has a lower rate of occurrence than an aberration in a smaller ensemble. By a similar argument, it may be concluded that this marginal probability must exceed the instability rate α.

Within-Ensemble Correlation

Consider an ensemble C_1 containing $m_1 > 1$ aberrations. As before, let A be the event that every aberration in C_1 occurs somehow, and now let i and j denote distinct aberrations within C_1. Because X_i and X_j are identically distributed Bernoulli trials, their covariance is

$$d = P(X_i = X_j = 1|\mathrm{SEL}) - [P(X_i = 1|\mathrm{SEL})]^2,$$

and so the aim is to show $d > 0$. The conditional independence of aberration level measures and SEL given A or given A^c is used to show that

$$p(1, 1|\mathrm{SEL}) = \frac{p(1, 1)}{p(\mathrm{SEL})}\{p(A|1, 1) + p(A^c|1, 1)p(\mathrm{SEL}|A^c)\}.$$

Combining this with (B.1), $d > 0$ is equivalent to

$$p(\mathrm{SEL})[p(A|1, 1) + p(A^c|1, 1)p(\mathrm{SEL}|A^c)] \\ > [p(A|1) + p(A^c|1)p(\mathrm{SEL}|A^c)]^2.$$

Evaluating these probabilities further, note that

$$p(\mathrm{SEL}) = 1 - \psi(1 - \theta^{m_1}),$$

where ψ is the probability that no other ensemble is fully aberrant and θ is the probability that an aberration occurs. (One would take $\psi = 1$ if there is just one ensemble.) Further,

$$p(A^c|1, 1) = 1 - \theta^{m_1 - 2}$$

and

$$p(A^c|1) = 1 - \theta^{m_1 - 1},$$

because for the ensemble under consideration to be fully aberrant, all other aberrations in that ensemble must have occurred, besides the ones on which information is available. Taken together, $d > 0$ is thus equivalent to

$$[1 - \psi(1 - \theta^a)][\theta^{a-2} + (1 - \theta^{a-2})(1 - \psi)] > [\theta^{a-1} + (1 - \theta^{a-1})(1 - \psi)]^2$$

using the notation $a = m_1$. This may be verified routinely because $a \geq 2$ and $\theta \in (0, 1)$. Note that $\psi < 1$ only if there are multiple ensembles, and so only in that case will the covariance between aberrations on the same ensemble be positive.

Negative covariance between X_i and X_j in different ensembles is calculated similarly. It was verified by Newton (2001) under the condition that both ensembles are of the same size.

Identifiability

To prove this in the standard way, let $p_1(x)$ and $p_2(x)$ denote the joint probability mass functions for two different networks, $\mathcal{C}_1$ and $\mathcal{C}_2$, as defined in (1). The aim is to show that p_1 and p_2 are different; so the opposite, $p_1(x) = p_2(x)$, is assumed for all length-n binary sequences x, and the attempt is made to derive a contradiction. Taking the particular realization $x = (1, 1, \ldots, 1)$, observe that the numerator in both p_1 and p_2 reduces to α^n, because the event SEL is implied regardless of the network. The assumed equality between p_1 and p_2 thus forces equality in the denominators, and this in turn forces equality of the numerators for any x. In other words, for any realization x,

$$\prod_{k=1}^{K^1}[1 - \beta^{t_k^1}] = \prod_{k=1}^{K^2}[1 - \beta^{t_k^2}], \tag{B.2}$$

where superscripts distinguish features of the two networks. Recall that t_k is the sum of $(1 - x_i)$ for aberrations i on ensemble k.

By careful choice of realizations x, the equality (B.2) reveals some useful information. For example, suppose that the two networks $\mathcal{C}_1$ and $\mathcal{C}_2$ have no relevant aberrations in common. This is one way in which they can differ, of course. Select a realization x that has $x_i = 1$ for relevant aberrations i in $\mathcal{C}_1$ but $x_i = 0$ for relevant aberrations i in $\mathcal{C}_2$. This choice forces the left side of (B.2) to equal 0 while the right side is a polynomial in β that is strictly positive. Thus a contradiction occurs. Through further careful choices of x, a complete proof is obtained.

APPENDIX C: PRIOR AND MARKOV CHAIN MONTE CARLO DETAILS

Prior

A Bernoulli–Pólya prior for the relevance vector $a = (a_1, \ldots, a_n)$ is

$$\pi(a) = \frac{\Gamma(m+1)\Gamma(n-m+1)}{\Gamma(n+2)},$$

where $m = \sum_{i=1}^n a_i$ is the total number of relevant aberrations, $n - m$ is the number of neutral aberrations, and $\Gamma()$ is the gamma function. One way to obtain this is by assuming that the a_i are iid Bernoulli(κ) given some value κ that itself has a uniform distribution. Alternatively, one may realize a Pólya–urn scheme in which the initial urn contains two tickets, one ticket labeled "0" and one labeled "1" (e.g., Hartigan 1983). The uniform–Pólya prior for the label vector $c = (c_1, \ldots, c_n)$ is obtained similarly. Let $c_1 \sim \mathrm{Uniform}(0, 1)$. For $i > 1$, let c_i be an independent uniform(0, 1) draw with probability $\tau/(\tau + i - 1)$. Otherwise make c_i identical to one of the previous values c_j, sampled uniformly from $j = 1, 2, \ldots, i - 1$. This creates the well-known cluster structure in Dirichlet process mixture calculations (e.g., Lo 1984). It may also be viewed as a product partition model (Hartigan 1990).

Markov Chain Monte Carlo

There are four network move types:

1. `MOVE`. Sample `nmove` aberrations at random (without replacement) from the full set $\{1, 2, \ldots, n\}$. Call this index set I. The proposed vector c^* is identical to c except possibly at indices $i \in I$. Draw the subvector labels $c_I^* = \{c_i : i \in I\}$ from their prior predictive distribution given the remaining subvector $\{c_i : i \in I^c\}$; that is, sample them as the last `nmove` steps in a uniform–Pólya sequence. This move type attempts to change the ensemble structure by wholesale movement of a subset of aberrations. The proposal mechanism is not symmetric, but, because it corresponds to prior-type draw, the Metropolis–Hastings ratio reduces to a ratio of likelihoods.
2. `SHUFFLE`. Sample `nperm` aberrations at random (without replacement) from the full set $\{1, 2, \ldots, n\}$. Call this index set I. As before, the proposed vector c^* is identical to c except possibly at aberrations $i \in I$. Draw the subvector labels c_I^* by randomly permuting the existing labels c_I. This amounts to permuting aberrations among ensembles. The proposal mechanism is symmetric; the Metropolis–Hastings ratio becomes a ratio of posterior masses of (c^*, a) to (c, a).
3. `AD-RANDOM`. Realize n iid Bernoulli trials with success probability `padI`. Call I the random set of aberrations corresponding

to successes. Propose a new state (c, a^*) by changing the relevance status of each $i \in I$; that is, if $a_i = T$, then make $a_i^* = F$, and if $a_i = F$, then propose $a_i^* = T$. This activate–deactivate proposal mechanism is symmetric, hence the Metropolis–Hastings ratio is a ratio of posterior masses of (c, a^*) to (c, a).

4. `AD-PATH`. From the list of unique labels in c, sample one at random. Let I denote the ensemble sharing this common label, and say that M is the cardinality of this set. Attempt to modify the state only if $M \geq 2$. If so, then take one of three possible actions. With probability `padII`, activate the entire ensemble; that is set $a_i^* = T$ for all $i \in I$. Alternatively, with probability `padII` again, deactivate the entire ensemble; set $a_i^* = F$ for all $i \in I$. Otherwise, (i.e., with probability $1 - 2$`padII`), replace the relevance subvector a_I with one of the $2^M - 2$ possible mixed arrangements, sampled at random. Do not modify the vector c. This move type is not symmetric. On moving to one of the purified ensembles (i.e., all neutral or all relevant), the probability is proportional to `padII`. On moving to a mixed ensemble, the probability is proportional to $1 - 2$`padII` divided by $2^M - 2$. Ratios of these chances enter the Metropolis–Hastings ratio.

The calculations presented here used `nmove` $= 4$, `nperm` $= 5$, `padI` $= (1/30)$, and `padII` $= (1/4)$. The rate-parameter update was attempted on a fraction `pup` $= (1/10)$ of the scans, and the box had side length .016. Making less dramatic moves adversely affects mixing, whereas more dramatic moves diminish the acceptance rate excessively. (Because it samples from the conditional prior of subsets, `MOVE` by itself enables movement through the entire space of c vectors, and, similarly, `AD-RANDOM` enables movement through all possible a vectors, so the resulting chain is irreducible.)

In the first efforts, the `MOVE` update affected only a single aberration, as did the predecessor of `AD-RANDOM`. Having the ability to rearrange multiple aberrations makes the sampler quite flexible. It is especially helpful to modify the relevance status of entire ensemble in `AD-PATH`. Adding new very small ensembles can dramatically reduce the likelihood, whereas the same aberrations placed on a larger ensemble fit well into the network. (Theory supports this, because in large ensembles, the marginal aberration rate is near the background level.) Note that `AD-PATH` allows mixed ensembles, so that reverse steps are possible from any current state.

On implementation, acceptance rates of all the move types are recorded. Chains are started at a state c drawn from the Pólya prior and a random a biased slightly toward neutral aberrations. Typically, two independent runs are made for each value of the hyperparameter τ, and trace plots and summary statistics are monitored for output analysis.

[Received October 2001. Revised September 2002.]

REFERENCES

Besag, J., and Clifford, P. (1989), "Generalized Monte Carlo Significance Tests," *Biometrika*, 76, 633–642.

Brodeur, G. M., Tsiatis, A. A., Williams, D. L., Luthardt, F. W., and Green, A. A. (1982), "Statistical Analysis of Cytogenetic Abnormalities in Human Cancer Cells," *Cancer Genetics and Cytogenetics*, 7, 137–152.

Carothers, A. D. (1997), "A Likelihood-Based Approach to the Estimation of Relative DNA Copy Number by Comparative Genomic Hybridization," *Biometrics*, 53, 848–856.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H., and Schäffer, A. A. (1999), "Inferring Tree Models for Oncogenesis From Comparative Genome Hybridization Data," *Journal of Computational Biology*, 6, 37–51.

——— (2000), "Distance-Based Reconstruction of Tree Models for Oncogenesis," *J. Comp. Bio.*, 7, 789–803.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, New York: Chapman and Hall.

Gray, J. W., and Collins, C. (2000), "Genome Changes and Gene Expression in Human Solid Tumors," *Carcinogenesis*, 21, 443–452.

Hartigan, J. A. (1983), *Bayes Theory*, New York: Springer-Verlag.

——— (1990), " Partition Models," *Communications in Statistics, Part A—Theory and Methods*, 19, 2745–2756.

Hemminki, A., Tomlinson, I., Markie, D., Jarvinen, H., Sistonen, P., Bjorkqvist, A.-M., Knuutila, S., Salovaara, R., Bodmer, W., Shibata, D., de la Chapelle, A., and Aaltonen, L .A. (1997), "Localization of a Susceptibility Locus for Peutz–Jeghers Syndrome to 19p Using Comparative Genomic Hybridization and Targeted Linkage Analysis," *Nature Genetics*, 15, 87–90.

Jarrard, D. F., Sarkar, S., Shi, Y., Yeager, T. R., Magrane, G., Kinoshita, H., Nassif, N., Meisner, L., Newton, M. A., Waldman, F. M., and Reznikoff C. A. (1999), "p16/pRb Pathway Alterations Are Required for Bypassing Senescence in Human Prostate Epithelial Cells," *Cancer Research*, 59, 2957–2964.

Jiang, F., Desper, R., Papadimitriou, C. H., Schäffer, A. A., Kallioniemi, O.-P., Richter, J., Schraml, P., Sauter, G., Mihatsch, M. J., and Moch, H. (2000), "Construction of Evolutionary Tree Models for Renal Cell Carcinoma From Comparative Genomic Hybridization Data," *Cancer Research*, 60, 6503–6509.

Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D. Gray, J., Waldman, F., and Pinkel, D. (1992), "Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors," *Science*, 258, 818–821.

Kainu, T., Juo, S.-H. H., Desper, R., Schäffer, A. A., Gillanders, E., Rozenblum, E., Freas-Lutz, D., Weaver, D., Stephan, D., Bailey-Wilson, J., Kallioniemi, O.-P., Tirkkonen, M., Syrjäkoski, K., Kuukasjärvi, T., Koivisto, P., Karhu, R., Holli, K., Arason, A., Johannesdottir, G., Bergthorsson, J. T., Johannsdottir, H., Egilsson, V., Barkardottir, B. R., Johannsson, O., Haraldsson, K., Sandberg, T., Holmberg, E., Grönberg, H., Olsson, H., Borg, A., Vehmanen, P., Eerola, H., Heikkila, P., Pyrhönen, S., and Nevanlinna, H. (2000), "Somatic Deletions in Hereditary Breast Cancers Implicate 13q21 as a Putative Novel Breast Cancer Susceptibility Locus," *Proceedings of the National Academy of Sciences*, 97, 9603–9608.

Knuutila, S., Aalto, Y., Autio, K., Björkqvist, A.-M., El-Rifai, W., Hemmer, S., Huhta, T., Kettunen, E., Kiuru-Kuhlefelt, S., Larramendy, M. L., Lushnikova, T., Monni, O., Pere, H., Tapper, J., Tarkkanen, M., Varis, A., Wasenius, V.-M., Wolf, M., and Zhu, Y. (1999), "DNA Copy Number Losses in Human Neoplasms. Review," *American Journal of Pathology Online* 155, 683–694.

Knuutila, S., Björkqvist, A.-M., Autio, K., Tarkkanen, M., Wolf, M., Monni, O., Szymanska, J., Larramendy, M. L., Tapper, J., Pere, H., El-Rifai, W., Hemmer, S., Wasenius, V.-M., Vidgren, V., and Zhu, Y. (1998), "DNA Copy Number Amplifications in Human Neoplasms, Review of Comparative Genomic Hybridization Studies," *American Journal of Pathology*, 152, 1107–1123.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998), "Genetic Instabilities in Human Cancers," *Nature*, 396, 643–649.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: (I) Density Estimates," *The Annals of Statistics*, 12, 351–357.

Newton, M. A. (2001), "A Statistical Method to Discover Significant Combinations of Genetic Aberrations Associated With Cancer Using Comparative Genomic Hybridization," Technical Report 148, University of Wisconsin-Madison, Dept. of Biostatistics and Medical Informatics.

Newton, M. A., Gould, M. N., Reznikoff, C. A., and Haag, J. D. (1998), "On the Statistical Analysis of Allelic-Loss Data," *Statistics in Medicine*, 17, 1425–45.

Newton, M. A., and Lee, Y. J. (2000), "Inferring the Location and Effect of Tumor Suppressor Genes by Instability-Selection Modeling of Allelic-Loss Data," *Biometrics*, 56, 1088–1097.

Newton, M. A., Wu, S.-Q., and Reznikoff, C. A. (1994), "Assessing the Significance of Chromosome-Loss Data: Where are the Suppressor Genes for Bladder Cancer?," *Statistics in Medicine*, 13, 839–858.

Piper, J., Rutovitz, D., Sudar, D., Kallioniemi, A., Kallioniemi, O.-P., Waldman, F. M., Gray, J. W., and Pinkel, D. (1995), "Computer Image Analysis of Comparative Genomic Hybridization," *Cytometry*, 19, 10–26.

Roylance, R., Gorman, P., Harris, W., Liebmann, R., Barnes, D., Hanby, A., and Sheer, D. (1999), "Comparative Genomic Hybridization of Breast Tumors Stratified by Histological Grade Reveals New Insights Into the Biological Progression of Breast Cancer," *Cancer Research*, 59, 1433–1436.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1728.

Tomlinson, I. P. M., Novelli, M. R., and Bodmer, W. F. (1996), "The Mutation Rate and Cancer," *Proceedings of the National Academy of Science U S A*, 93, 14800–14803.

Van Lint, J. H., and Wilson, R. M. (1992), *A Course in Combinatorics*, New York: Cambridge University Press.

Volpe, J. (1990), "Genetic Stability and Instability in Tumours," in *Molecular Biology of Cancer Genes*, ed. M. Sluyser, Chichester: Ellis Harwood, pp. 10–23.

http://www.jstor.org

LINKED CITATIONS

- Page 1 of 2 -

You have printed the following article:

Discovering Combinations of Genomic Aberrations Associated with Cancer
Michael A. Newton
Journal of the American Statistical Association, Vol. 97, No. 460. (Dec., 2002), pp. 931-942.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28200212%2997%3A460%3C931%3ADCOGAA%3E2.0.CO%3B2-S

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

References

Generalized Monte Carlo Significance Tests
Julian Besag; Peter Clifford
Biometrika, Vol. 76, No. 4. (Dec., 1989), pp. 633-642.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198912%2976%3A4%3C633%3AGMCST%3E2.0.CO%3B2-K

A Likelihood-Based Approach to the Estimation of Relative DNA Copy Number by Comparative Genomic Hybridization
Andrew D. Carothers
Biometrics, Vol. 53, No. 3. (Sep., 1997), pp. 848-856.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199709%2953%3A3%3C848%3AALATTE%3E2.0.CO%3B2-V

Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors
Anne Kallioniemi; Olli-P. Kallioniemi; Damir Sudar; Denis Rutovitz; Joe W. Gray; Fred Waldman; Dan Pinkel
Science, New Series, Vol. 258, No. 5083. (Oct. 30, 1992), pp. 818-821.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819921030%293%3A258%3A5083%3C818%3ACGHFMC%3E2.0.CO%3B2-W

http://www.jstor.org

LINKED CITATIONS

- Page 2 of 2 -

On a Class of Bayesian Nonparametric Estimates: I. Density Estimates

Albert Y. Lo

The Annals of Statistics, Vol. 12, No. 1. (Mar., 1984), pp. 351-357.

Stable URL:

http://links.jstor.org/sici?sici=0090-5364%28198403%2912%3A1%3C351%3AOACOBN%3E2.0.CO%3B2-D

Inferring the Location and Effect of Tumor Suppressor Genes by Instability-Selection Modeling of Allelic-Loss Data

Michael A. Newton; Yoonjung Lee

Biometrics, Vol. 56, No. 4. (Dec., 2000), pp. 1088-1097.

Stable URL:

http://links.jstor.org/sici?sici=0006-341X%28200012%2956%3A4%3C1088%3AITLAEO%3E2.0.CO%3B2-R

Markov Chains for Exploring Posterior Distributions

Luke Tierney

The Annals of Statistics, Vol. 22, No. 4. (Dec., 1994), pp. 1701-1728.

Stable URL:

http://links.jstor.org/sici?sici=0090-5364%28199412%2922%3A4%3C1701%3AMCFEPD%3E2.0.CO%3B2-6