UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS

Technical Report #75
February 1993

ASSESSING THE SIGNIFICANCE OF CHROMOSOME-LOSS DATA:
WHERE ARE THE SUPPRESSOR GENES FOR BLADDER CANCER?

Michael A. Newton
Departments of Biostatistics and Statistics

Shi-Qu Wu
Department of Human Oncology

Catherine A. Reznikoff
Department of Human Oncology

MADISON, WISCONSIN

* Submitted to Statistics in Medicine, February 17, 1993
Assessing the significance of chromosome-loss data: Where are suppressor genes for bladder cancer?

Michael A. Newton \(^1\)
Department of Biostatistics
University of Wisconsin–Madison
6770 Medical Sciences Center
1300 University Avenue
Madison WI 53706-1532.
Phone: 608 263 0357
Fax: 608 263 0415
Internet: newton@stat.wisc.edu

Shi-Qi Wu
Department of Human Oncology
University of Wisconsin Comprehensive Cancer Center
600 Highland Avenue
Madison WI 53792
Phone: 608 263 4562
Fax: 608 263 8613

Catherine A. Reznikoff
Department of Human Oncology
University of Wisconsin Comprehensive Cancer Center
600 Highland Avenue
Madison WI 53792
Phone: 608 263 4526
Fax: 608 263 8613

\(^1\)To whom correspondence should be addressed.
Assessing the significance of chromosome-loss data:
Where are suppressor genes for bladder cancer?

Summary
Cytogenetic analysis reveals alterations of chromosome structure (losses, gains, and rearrangements of genetic material) in bladder cancer cells generated using an in vitro/in vivo transformation system. To predict possible locations of bladder cancer suppressor genes, a robust Bayesian analysis of the chromosome-loss data is performed. A simple stochastic model is postulated to describe chromosome loss during tumor progression. Posterior computations are enabled by a dynamic simulation algorithm. Ordered by decreasing posterior probability of putatively harboring a suppressor gene, significant losses are observed on chromosomes 3, 18, 13, 10, 11, and y.
1 Introduction

The loss of cytogenetically detectable genetic material from several chromosome arms is ubiquitous in clinical carcinoma cells [1], including bladder cancer [2]. A possible explanation for this phenomenon is that particular genes that regulate normal cell behavior (cancer suppressor genes) exist in these chromosomal regions and their loss during multistage carcinogenesis leads to tumor development. Although the initial events of tumorigenesis are unknown, any cell having lost such a suppressor gene might gain a selective advantage over other cells. The clonal selection of such a cell (with an initial growth advantage) could continue, leading to the development of cancers lacking specific gene functions. Cancers derived from the same cell type often show similar losses, but some level of background loss also exists; that is loss of chromosome material not carrying a suppressor gene. For example, colorectal carcinomas show losses from 20% of evaluable chromosome arms, but losses from particular regions are most significant [3, 4].

In a unique in vitro/in vivo transformation system [4], an immortalized human uroepithelial cell line has been observed as it progresses through various stages of bladder-cancer tumorigenesis. Cytogenetic analysis reveals which chromosomal regions are lost, and what regions remain intact in cells at each stage of this progression. The data record chromosomal aberrations at the level of chromosome arms (the short p arm and the long q arm) which together form single chromosomes. Data are not at the molecular level. This present paper describes a statistical analysis of the chromosome-loss data in this system, and, roughly speaking, separates the losses associated with putative suppressor-gene status from background losses. The result is a list of chromosomal regions most likely to harbor suppressor genes. In summary, significant arm losses are observed at: 3p, 18q, 13q, 10, 11q, and y. Losses at 22q, and 6q are somewhat less significant overall, but tend to be associated with progression to high grade tumors.

Data are categorical, have a number of intrinsic constraints and dependencies, and are further constrained by a stochastic model of the system. A robust Bayesian analysis is used to predict the probable locations and the number of putative suppressor genes based on the data and the model. Posterior computation is enabled by a Markov chain Monte Carlo (dynamic simulation) algorithm. Data are described in Section 2, and a stochastic model is proposed in Section 3. Bayesian inference for predicting the location of putative suppressor genes is done in Section 4 and a discussion follows. Details of the dynamic simulation algorithm are in an appendix.
2 Data and Experimental Methods

To establish the *in vitro/in vivo* transformation system used in this study, a nontumorigenic human uroepithelial cell line (SV-HUC) was produced after HUC infection *in vitro* with simian virus 40 (SV40) [5]. This initial cell line was tumorigenically transformed in a multistep fashion to 16 descendant lines after exposure of cells *in vitro* to different carcinogens: the polycyclic hydrocarbon, 3 methylcholanthrene (MC), or the human bladder carcinogen 4-aminobiphenyl (A), or its reactive N-hydroxy metabolites (HA) [6, 7]. Tumor cell lines were initiated from tumors produced in athymic, nude mice. Figure 1 shows the relationships between the 16 tumor cell lines. The 16 T-SV-HUC's used in the present study show different tumor growth kinetics (ranging from indolent to aggressive) and also different tumor histopathologies (including transitional, squamous, and undifferentiated carcinomas) that recapitulate the heterogeneous cancer phenotypes observed in clinical bladder cancers [6, 7]. Cytogenetic analysis indicates that each T-SV-HUC was clonal in origin and that all sustained nonrandom losses from several chromosome arms, including 3p, 18q, and 6q [8, 9, 10]. In the present report, we further assess the significance of chromosome loss data in the SV-IIUC transformation system using a robust Bayesian analysis and a simple stochastic model.

A karyotype is a systematized arrangement of the chromosomes. Karyotypes for 20 tumor cells from each cell line provide the raw data for statistical analysis. Each karyotype describes, among other things, which arms are missing and which are intact in each cell. Losses of intact chromosomes (p and q arms together), complete arms, and regions within arms are observed. A representative or modal karyotype is generated for each tumor cell line, by recording an arm as lost if any region of it is missing in 15 of the 20 cells examined. The modal karyotype, together with the modal karyotype of the parent line, indicates what new arm losses have occurred during progression of the tumor, and whether or not such losses are tied (i.e. loss of an intact chromosome or not). Arms lost in the SV-HUC parent, 15q, 5p, 6p, are never evaluated and so they do not enter further analysis. Satellites and short arms of acrocentric chromosomes, 13p, 14p, 15p, 21p, 22p, are also excluded.

Although human somatic cells are normally diploid (two copies of each chromosome), many neoplastic cells are tetraploid (4 copies). In fact, the modal karyotype of 7 out of the 16 T-SV-HUC cell lines studied is tetraploid. We avoid the diploid/tetraploid complication in this analysis by producing a diploid approximation to every tetraploid karyotype. This is arguably reasonable because even the loss of one arm out of four may lead to the loss of heterozygosity, and hence may be physiologically equivalent to one loss from two.

To be a viable cell, at least one copy of each chromosome arm must be present. Therefore, if only one copy of an arm exists in the parent line, then no new losses of this arm can occur in descendant lines. Thus the modal karyotypes are statistically dependent because of relationships between the different cell lines. Figure 2 is a *chromogram* summarizing the
marginal losses and potential losses from all 16 lines. The potential number of losses often is less than 16 for precisely the reason mentioned above: if an arm is lost in a parent line, then it is unavailable for loss in descendent lines. It is apparent from inspection of Figure 2 that some arms are lost more than other arms. Indeed if the theory outlined in the Introduction is true, then more losses would be expected in arms harboring suppressor genes. Deciding which arms in fact have significant losses, in some sense, is the purpose of this paper.

The cytogenetic analysis indicates whether or not simultaneous p and q losses are tied. Since tied losses are quite prevalent (35 tied losses were observed overall), it is important to account for them in the analysis. If each pair of tied losses is treated as two independent losses, then we may overstate the significance of the total losses of an arm. For example, if a suppressor gene exists on 18q, but not 18p, then we expect more losses of 18q than 18p. If, however, 18 is prone to ties, we may in fact observe many losses of 18p because it is getting dragged along with 18q. To handle this issue, we extend the notion of tied loss to a slightly more general notion of tied-ness. It is convenient to suppose that the p and q arms of the same single chromosome may or may not be tied, regardless of whether or not they are lost. Thus, tied arms will be lost together if they are lost at all, but a loss may not occur. This somewhat artificial construction is quite useful, as we shall see, because it facilitates the separation of two distinct types of arm loss. Tied losses are indicated in the chromogram (Figure 2) by the type of hashing. The data, after diploid approximation, are organized in Table 2.

3 A Stochastic Model

3.1 Assumptions

To determine which arms have an elevated loss probability (and thus putatively harbor suppressor genes), we construct a simple stochastic model of karyotype evolution. All probability computations are based on the following modeling assumptions:

1. Losses at a given chromosome depend on ancestral cell lines only through the parent karyotype.

2. Losses at different chromosomes are independent.

3. Two arms of a chromosome may or may not be tied. (A tie does not imply a loss.)

4. The tendency for ties is a property of each chromosome, and may vary (e.g. all losses of 10p are tied with losses of 10q, but none of the 3p losses are tied.)

5. If arms are not tied, then loss of the p arm is independent from loss of the q arm.
6. If arms are tied, then loss of the whole chromosome is treated like the loss of an untied arm.

7. Untied arms that are available to be lost (i.e. there are 2 copies in the parent karyotype) have two possible loss probabilities:

   (a) baseline (i.e. no suppressor gene on that arm)
   (b) elevated (i.e. a suppressor gene on that arm)

8. A suppressor gene may or may not exist on any arm.

The first assumption implies that the representative karyotype is a Markov process along any lineage of cell lines. The independence assumption is hard to justify on physical grounds, but it is important in the analysis because data from only 16 lines are available. If dependence is assumed, then its structure would have to be formulated for a high-dimensional space (nominally 46 dimensions). Indeed dependence may exist if there are interactions between the products of genes on different chromosomes, however no obvious patterns emerged from inspection of the complete loss data. On ties, a tied-loss is observable by the cytogeneticist, but we suppose that a tie can exist without any losses. Tied-ness thus becomes a partially observable quantity, because we cannot identify a tie between two arms that are not lost. To interpret this notion, suppose that tying events precede loss events. A chromosome that has been tied will necessarily lose p and q together if a loss occurs, but a loss may not occur. The last two assumptions are critical. They give us, in effect, a working definition of suppressor gene. By natural selection, an arm carrying a suppressor gene is one with an elevated loss probability.

3.2 The Sampling Model

Some notation is required. For i indexing the chromosome (i = 1, ..., 23) and j indexing the cell line (j = 1, ..., 16), the complete data are denoted

\[ X_{i,j} = 1 \text{ (p-arm loss in line } j, \text{ chromosome } i) \]
\[ m_{i,j} = 1 \text{ (2 p-arms in the parent karyotype of line } j, \text{ chromosome } i) \]
\[ Y_{i,j} = 1 \text{ (q-arm loss in line } j, \text{ chromosome } i) \]
\[ n_{i,j} = 1 \text{ (2 q-arms in the parent karyotype of line } j, \text{ chromosome } i) \]
\[ Z_{i,j} = 1 \text{ (tie in chromosome } i, \text{ line } j) \]

By the definition of ties, the \( Z_{i,j} \) are only partially observed. If \( m_{i,j} = n_{i,j} = 1 \) but there are no losses \( X_{i,j} = Y_{i,j} = 0 \), then \( Z_{i,j} \) is unobserved.
A number of parameters are implied by the eight assumptions of the last section. Key parameters of interest are suppressor gene indicators

\[ \theta_{p,i} = 1 \text{ (suppressor gene on p-arm of chromosome i)} \]
\[ \theta_{q,i} = 1 \text{ (suppressor gene on q-arm of chromosome i).} \]

The baseline and elevated loss probabilities are

\[ \alpha = P(\text{arm loss given no tie and no suppressor gene}) \]
\[ = P(X_{i,j} = 1|m_{i,j} = 1, Z_{i,j} = 0, \theta_{p,i} = 0) \]
\[ = P(Y_{i,j} = 1|n_{i,j} = 1, Z_{i,j} = 0, \theta_{q,i} = 0) \]

\[ \beta = P(\text{arm loss given no tie and a suppressor gene}) \]
\[ = P(X_{i,j} = 1|m_{i,j} = 1, Z_{i,j} = 0, \theta_{p,i} = 1) \]
\[ = P(Y_{i,j} = 1|n_{i,j} = 1, Z_{i,j} = 0, \theta_{q,i} = 1) \]

constrained so that \( \alpha < \beta \). The tendency for ties is summarized by tie-probabilities

\[ \tau_i = P(Z_{i,j} = 1|m_{i,j} = n_{i,j} = 1). \]

The tie probability is 0 if arms are not available to be tied (i.e. if \( m_{i,j} + n_{i,j} < 2 \)). Assumptions 5 and 6 imply that \( X_{i,j} \) and \( Y_{i,j} \) are conditionally independent given \( Z_{i,j} = 0 \), and equal if \( Z_{i,j} = 1 \). Also by assumption 6, tied arms are lost with probability \( \alpha \) if there are no suppressor genes on the entire chromosome, and with probability \( \beta \) otherwise. Assumption 2 implies that the triples \( (X_{i,j}, Y_{i,j}, Z_{i,j}) \) are independent across \( i \). These triples are dependent across cell lines \( j \) because some lines descend from others. However the Markov assumption (assumption 1) ensures that the likelihood function (based on the complete data) is readily computed:

\[
P(\text{data} | \text{parameters}) = \left( \prod_{i=1}^{23} \left[ \tau_i^{T_{1,i} + T_{2,i}}(1 - \tau_i)^{T_i} \right] \right) \alpha^a(1 - \alpha)^{\bar{a}} \beta^b(1 - \beta)^{\bar{b}} \quad (1)
\]

where

\[
a = \sum_{i=1}^{23} [P_i(1 - \theta_{p,i}) + Q_i(1 - \theta_{q,i}) + T_{1,i}1(\theta_{p,i} = \theta_{q,i} = 0)]
\]
\[
\bar{a} = \sum_{i=1}^{23} [\bar{P}_i(1 - \theta_{p,i}) + \bar{Q}_i(1 - \theta_{q,i}) + T_{2,i}1(\theta_{p,i} = \theta_{q,i} = 0)]
\]
\[
b = \sum_{i=1}^{23} [P_i\theta_{p,i} + Q_i\theta_{q,i} + T_{1,i}1(\theta_{p,i} + \theta_{q,i} > 0)]
\]
\[
\bar{b} = \sum_{i=1}^{23} [\bar{P}_i\theta_{p,i} + \bar{Q}_i\theta_{q,i} + T_{2,i}1(\theta_{p,i} + \theta_{q,i} > 0)]
\]
and

\[ P_i = \sum_{j=1}^{16} 1(m_{i,j} = 1, Z_{i,j} = 0, X_{i,j} = 1) \]

\[ \bar{P}_i = \sum_{j=1}^{16} 1(m_{i,j} = 1, Z_{i,j} = 0, X_{i,j} = 0) \]

\[ Q_i = \sum_{j=1}^{16} 1(n_{i,j} = 1, Z_{i,j} = 0, Y_{i,j} = 1) \]

\[ \bar{Q}_i = \sum_{j=1}^{16} 1(n_{i,j} = 1, Z_{i,j} = 0, Y_{i,j} = 0) \]

\[ T_{1,i} = \sum_{j=1}^{16} 1(m_{i,j} = n_{i,j} = Z_{i,j} = X_{i,j} = Y_{i,j} = 1) \]

\[ T_{2,i} = \sum_{j=1}^{16} 1(m_{i,j} = n_{i,j} = Z_{i,j} = 1, X_{i,j} = Y_{i,j} = 0) \]

\[ \bar{T}_i = \sum_{j=1}^{16} 1(m_{i,j} = n_{i,j} = 1, Z_{i,j} = 0). \]

In words, \( P_i \) is the number of non-tied \( p \) arm losses at chromosome \( i \), and \( \bar{P}_i \) is the number of \( p \)-arms not tied and not lost. A similar interpretation holds for \( Q_i \) and \( \bar{Q}_i \). The number of tied losses at \( i \) is \( T_{1,i} \), and the number of ties where no losses occur is \( T_{2,i} \). The number of times a tie is possible \( (m_{i,j} = n_{i,j} = 1) \), but no tie occurs is \( \bar{T}_i \). The total number of lines in which both \( p \) and \( q \) arms are available to be lost is \( T_{1,i} + T_{2,i} + \bar{T}_i \). These sufficient statistics can be read directly from the appropriate contingency table. For example, \( P_i \) is the sum of elements 11111 and 11100 and 10100 from Table 1.

Note how the tie-tendency parameters \( \tau_i \) factor nicely out of the likelihood, but the suppressor-gene indicators are wrapped up in the exponents of \( \alpha \) and \( \beta \).

## 4 Inference

### 4.1 The Posterior

The primary goal of inference in this application is to identify those chromosome arms which harbor suppressor genes. That is, which \( \theta_{p,i} \) or \( \theta_{q,i} \) equal 1. A secondary but related goal is to estimate how many suppressor genes there may be altogether. The approach taken here is Bayesian, essentially because it provides a rather direct way to estimate the suppressor-gene indicators. The Bayesian analysis is robust in the sense that posteriors are studied under a range of prior assumptions. Manipulation of the likelihood itself is quite difficult, making inference based on maximum likelihood estimates impractical.
The model specifies the distribution of the triples \((X_{i,j}, Y_{i,j}, Z_{i,j})\) given the parameters. A convenient prior in the Bayesian analysis is one having three independent sets of parameters:

1. loss probabilities \((\alpha, \beta)\)
2. suppressor gene indicators \(\{\theta_{p,i}, \theta_{q,i}\}\)
3. tie-probabilities \(\{\tau_i\}\).

Further, a uniform prior is placed on \(\{(\alpha, \beta) : \alpha < \beta\}\). Independent Beta\((\phi_1, \phi_2)\) priors are placed on each \(\tau_i\). The critical prior specification is on the suppressor-gene indicators. We assume that these parameters are iid, with

\[
\eta = P(\theta_{p,i} = 1) = P(\theta_{q,i} = 1).
\]

As we shall see, the magnitude of posterior probabilities is quite sensitive to \(\eta\) but much less so to \(\phi_1\) and \(\phi_2\). Certain summaries of the full posterior are not sensitive to the prior.

Posterior computations are enabled by a dynamic simulation algorithm described in detail in the appendix. This algorithm integrates over the high-dimensional parameter space and over the space of missing data. For each prior specification, one long run of a Markov chain is produced and subsampled to approximate the various marginal posterior quantities of interest. All computed probabilities are Monte Carlo estimates of actual posterior probabilities, and hence are subject to a small error.

4.2 All 16 cell lines

4.2.1 Suppressor-gene indicators

Figure 3 shows a simple posterior summary for two prior specifications. Plotted for each of the 23 chromosome pairs is the posterior probability that a suppressor gene exists on either the \(p\) arm or the \(q\) arm of that chromosome. The posteriors differ because the prior probability of a suppressor gene, \(\eta\), takes two different values. A uniform prior is placed on the \(\tau_i\). We see immediately that certain arms, 3, 18, 10, show very high posterior probabilities of at least one suppressor gene. Perhaps as expected, the absolute value of these probabilities is quite sensitive to the prior expectation. Figure 4 shows that the posterior ranking of the chromosomes is quite insensitive to the prior. Plotted for each chromosome pair, and for values of \(\eta\) from .01 to .75, are Bayes factors (on a log scale):

\[
\text{Bayes factor} = \frac{\text{Posterior odds of a suppressor gene on } p \text{ or } q}{\text{Prior odds of a suppressor gene on } p \text{ or } q} = \frac{P(\text{data|gene on } p \text{ or } q)}{P(\text{data|no gene on } p \text{ or } q)}
\]

The Bayes factor [11, for example] is the amount by which the data change ones prior opinion on the existence of an elevated loss probability. Contrary to being subjective, these Bayes
factors are a very informative summary of the data. We see from Figure 4 that, in order of decreasing posterior probability, chromosomes 18, 3, 10, 13, y, 11, 22, and 6 are always most likely to harbor suppressor genes, regardless of one’s prior opinion. Furthermore, except for chromosomes 22 and 6, the Bayes factors of these chromosomes always exceed 10 for all priors \( \eta \geq .01 \).

A more detailed analysis of the full posterior shows which particular arms are likely to have an elevated loss probability. Figure 5 is similar to Figure 3, but it shows the marginal posterior probabilities of a suppressor gene on each of the 37 evaluated arms. Especially for the more conservative prior (\( \eta = .05 \)), the graph of marginal posterior probabilities provides an effective filtering of noise from the original chromogram (Figure 2) by downweighting the probable background losses. Again the raw probabilities are sensitive to the prior specification, as a comparison of the two panels in Figure 5 shows. Ranking arms by Bayes factors shows the arms most likely to have an elevated loss probability, regardless of prior opinion (Figure 6). For small values of \( \eta \), the posterior ranking has significant losses at 3p, 18q, 13q, 11q, 10p, 10q, y, with somewhat less significant losses at 22q and 6q. This ranking is not completely stable as \( \eta \) increases because of uncertainty in the location of a suppressor gene created by ties in the losses. For example, all 8 of the losses of 10p are tied with losses of 10q. The tie probability \( \tau_{10} \) is quite high, and so the losses of 10p may very well be explained by the existence of a suppressor gene on 10q only, and vica versa. There are seven losses of 18p, but this arm is unlikely, a posteriori, to harbor a suppressor gene because all of those losses are tied to 18q which exhibits two untied losses. So the losses at 18 are best explained by a suppressor gene on 18q and a high tie probability.

4.2.2 Loss probabilities

The probable values of baseline and elevated loss probability are quite insensitive to the prior assumptions about suppressor gene status, as indicated by Figure 7. Nontied arms carrying a suppressor gene are lost about 50% of the time, and those without such a gene are lost about 18% of the time. The uncertainty in \( \beta \), the elevated loss probability, is quite high when \( \eta \) is small. This is because small \( \eta \) indicates a relatively small number of arms which have suppressor genes. Since there are fewer arms with the such genes, there is less data on the loss probability associated with those arms.

4.2.3 The total number of suppressor genes

Another by-product of the model and subsequent posterior computations is an estimate of the total number of arms (out of 37) which harbor suppressor genes. Figure 8 is a plot of the posterior expected number of such arms versus the prior expected number. If we assume, a priori, that there are less than about 14 such arms, then the data tell us to increase our
estimate. Similarly if we assume, a priori, that there are more than about 14, the data tell us to lower our estimate. This suggests 14 as an estimate of the total number of arms which harbor suppressor genes. It is not clear how to assess uncertainty in this empirical Bayes estimate without a rather sophisticated simulation study. Work on colorectal cancer [12] suggests multiple changes in tumorigenesis. Our data are consistent with this, but indicate the possibility that a greater number of genes are involved.

4.2.4 Tie probabilities

The marginal posterior distribution of each tie parameter \( \tau_i \) is quite insensitive to the prior \( \eta \). In fact, if there were no missing data, then these parameters would be independent of the other parameters, a posteriori. Figure 9 shows the posterior means of the \( \tau_i \)'s when the prior is iid uniform \((0,1), \text{ and } \eta = .5 \). Arms like 10 and 18, that have a lot of tied losses, understandably have rather high posterior mean tie probabilities. These posterior mean tie probabilities are high for other chromosomes that do not exhibit many losses. For example, even chromosome 12 has a rather high posterior mean tie probability, but no losses were observed at 12. This behavior is a consequence of the model which allows ties not followed by losses. When no losses are observed, the arms are acting the same, and so they may very well be tied. For example, if no suppressor genes are on either arm, then the chance of a tie given no loss exceeds the chance of a non-tie given no loss when the prior tie probability is larger than .44 and \( \alpha = .2 \).

Figure 10 shows the marginal posterior probabilities of a suppressor gene at each arm for two different priors on the \( \tau_i \), and with \( \eta \) fixed at .05. These priors are quite extreme, with means .05 and .95, respectively. The scale of the marginal probabilities is somewhat sensitive to \( \eta \), especially if we expect a large number of ties, but again the ranking of arms is not affected.

4.3 High-grade tumor lines

Tumors tend to progress spontaneously in grade over time, or more rapidly if exposed to carcinogens. Eleven of the 16 cell lines represent progression to high-grade tumor status. Since losses which lead to high-grade status may differ from other losses, it is meaningful to analyze the data from these 11 lines separately. The methodology is identical to that of the previous analysis, and so details are omitted. What we learn is that significant losses occur at the same arms as in the overall data set, but that arms 22q and 6q also stand out as likely spots for suppressor genes. Further, there are relatively more losses in these 11 lines as compared to all 16 lines. One might expect, therefore, that the baseline and elevated loss-probabilities would be higher. In fact this does not happen. Instead, the baseline and elevated loss probabilities are about the same as for the 16 lines, but more arms are suspected
of harboring suppressor genes.

5 Discussion

A simpler approach to analyze these data is to first compare the total number of loss events to the total number of possible events. Then look at each chromosome arm in turn, regarding the number of events at that arm as a binomial random variable with some fixed (arm-independent) probability of a loss event, computed from the proportion above under a random loss assumption. A p-value could be computed for each arm by comparing the observed number of losses to that binomial distribution. Arms that get lost a lot will correspond to small p-values, completing the analysis. There are several drawbacks to this method.

First, it is difficult to deal with the ties in a way which discounts the effect of excessive tied losses. So, for example, 18p might be given a small p-value because it has many losses, but using this method one cannot account for the fact that these losses are all tied with 18q.

Second, the chromosome, rather than the arm, becomes the focus of the analysis by any reasonable definition of loss event. However, the arm is arguably the more relevant experimental unit.

Third, while a set of p-values gives a ranking of important arms related to the number of losses, this ranking is rather soft. One observes a gradual increase in p-value as the arms get ranked. The ranking produced by the methods of this paper, on the other hand, clearly delineates arms that show excessive losses from arms that do not. This removes some of the subjectivity in selecting a cutoff for sufficiently small p-values.

Fourth, the problem of multiple comparison plagues the simpler analysis. There are 37 evaluated arms, corresponding to many possible hypotheses one could test. Exactly how to proceed is unclear, especially if you look at the data before you set your hypotheses. It is tempting to test for excess arm loss at arms which show many losses! The Bayesian analysis presented here takes the bull by the horns, so to speak, and allows a hyperparameter describing the prior belief of the investigator. Calculations show just what the data say, no matter what is assumed prior to having the data.

Finally, the strength of the data is measured by the insensitivity of the posterior to the prior. With much more data, not only the ranking, but also the posterior probabilities of gene-status, would be very insensitive to the prior. The extent to which they are not insensitive indicates, at least heuristically, the amount of information in the data.

Various extensions of our stochastic model can be considered. More than two levels of loss-probability would allow differing intensities of suppressor genes. If such genes are uniformly scattered throughout the genome, you would expect more on longer arms than shorter arms. Arm length has been ignored here. Dependence between different chromosomes is a natural
extension, but more data are required to realistically estimate gene interactions. The ploidy of cells could also be explicitly modeled, so that a diploid approximation is unnecessary.

Acknowledgements

The authors gratefully acknowledge Barry Storer whose earlier work laid the foundation for this statistical analysis. Computations were done using the S statistical language, with ralfor subroutines attached.

A The Markov chain Monte Carlo algorithm

The likelihood function based on the complete data (both observed and missing) is given in equation 1. A consideration of this likelihood drives the posterior simulation. The algorithm is a dynamic simulation [13, 14, 15, 16, 17], combining the Gibbs sampler [18, 19] with Metropolis updates [20, 21] of the suppressor-gene indicator parameters.

A.1 Parameter updates

A number of complete conditional distributions used by the Gibbs sampler are immediate by inspection of equation 1. Firstly, the tie-tendency parameters $\tau_i$ have complete conditional density

$$p(\tau_i|\text{the rest}) \propto \tau_i^{T_1,i+T_2,i}(1-\tau_i)^{\bar{T}_i} \times \text{prior}(\tau_i|\text{the rest})$$

which is Beta if we start with a conjugate prior. These conditionals are readily simulated by drawing independent Gamma's and forming the appropriate ratios. The baseline and elevated loss-probability parameters have truncated-Beta complete conditionals:

$$p(\alpha|\text{the rest}) \propto \alpha^a(1-\alpha)^b \times \text{prior}(\alpha|\beta)$$

$$p(\beta|\text{the rest}) \propto \beta^b(1-\beta)^b \times \text{prior}(\beta|\alpha)$$

Again simulation is straightforward. We use simple rejection sampling from the non-truncated Beta (rejecting if the constraint $\alpha < \beta$ is not met).

The suppressor-gene indicator parameters are updated one at a time using a Metropolis algorithm conditional on the complete data and the other parameters. Given this information, the indicator parameters on different chromosomes are independent. Because of ties, $\theta_{p,i}$ and $\theta_{q,i}$ are not necessarily conditionally independent. The driving distribution of the Metropolis chain is quite simple. Suppose we are updating $\theta_{p,i}$, which of course can take only two values. The candidate state is chosen by keeping everything else fixed, but
switching the value of $\theta_{p,i}$. Then $r$, the posterior odds of this candidate state to the current state is computed. With probability $\min(r,1)$, the Metropolis chain moves to the candidate state, otherwise it stays at the current state. The chain then proceeds to update $\theta_{q,i}$ in an analogous way, and then the indicators on other chromosomes until all the indicator parameters have been updated. The value of $r$, the posterior odds for each update, depends on the complete data, the loss probabilities, and the gene indicator parameters on the two arms of the chromosome under consideration. For example, if we are updating $\theta_{p,i}$ which is currently equal to 0, and if $\theta_{q,i}$ happens to equal 1, then $r$ is (from 1)

$$r = \frac{\text{likelihood of}(\theta_{p,i} = 1, \theta_{q,i} = 1)}{\text{likelihood of}(\theta_{p,i} = 0, \theta_{q,i} = 1)} \times \text{prior odds}$$

$$= \left(\frac{\beta}{\alpha}\right)^{P_i} \left(\frac{1 - \beta}{1 - \alpha}\right)^{P_i} \left(\frac{\eta}{1 - \eta}\right).$$

Table 3 gives a complete list of the odds ratios that define the Metropolis update.

### A.2 Missing data updates

To complete the hybrid Metropolis-Gibbs sampler, we must specify how to update the missing data given everything else. Each chromosome with two evaluated arms has some missing data, according to the model. To see its structure, it is helpful to consider Table 1 which gives one way to organize the data at one chromosome. Although we observe the total number of lines in which both $p$ and $q$ arms are available but neither is lost, we do not observe the subset of those which are tied. In terms of Table 1, we observe $11001 + 11000$, but neither $11001$ or $11000$ separately. Let $S_i$ equal the total number of lines in which both the $p$-arm and the $q$-arm of chromosome $i$ exist in the parent line (and so are available to be lost). This total $S_i$ is partitioned multinomially into 6 cells of Table 1. Define

$$s_1 = \sum_{j=1}^{S_i} 1[Z_{i,j} = 0, X_{i,j} = 0, Y_{i,j} = 0]$$

$$s_2 = \sum_{j=1}^{S_i} 1[Z_{i,j} = 0, X_{i,j} = 0, Y_{i,j} = 1]$$

$$s_3 = \sum_{j=1}^{S_i} 1[Z_{i,j} = 0, X_{i,j} = 1, Y_{i,j} = 0]$$

$$s_4 = \sum_{j=1}^{S_i} 1[Z_{i,j} = 0, X_{i,j} = 0, Y_{i,j} = 1]$$

$$s_5 = \sum_{j=1}^{S_i} 1[Z_{i,j} = 1, X_{i,j} = 0, Y_{i,j} = 0]$$

14
\[ s_6 = \sum_{j=1}^{s_i} 1[Z_{i,j} = 1, X_{i,j} = 1, Y_{i,j} = 1]. \]

It is readily shown from the model assumptions that

\[ (s_1, \ldots, s_6) \sim \text{Multinomial}_6 (S_i, (p_1, \ldots, p_6)), \]

where

\[
\begin{align*}
p_1 &= (1 - \tau_i)(1 - u_i)(1 - v_i) \\
p_2 &= (1 - \tau_i)(1 - u_i)v_i \\
p_3 &= (1 - \tau_i)u_i(1 - v_i) \\
p_4 &= (1 - \tau_i)u_iv_i \\
p_5 &= \tau_i(1 - w_i) \\
p_6 &= \tau_iw_i
\end{align*}
\]

and

\[
\begin{align*}
u_i &= P(\text{p-arm loss | no tie}) = \alpha 1[\theta_{p,i} = 0] + \beta 1[\theta_{p,i} = 1] \\
v_i &= P(\text{q-arm loss | no tie}) = \alpha 1[\theta_{q,i} = 0] + \beta 1[\theta_{q,i} = 1] \\
w_i &= P(\text{tied loss | a tie}) = \alpha 1[\theta_{p,i} = \theta_{q,i} = 0] + \beta 1[\theta_{p,i} + \theta_{q,i} > 0].
\end{align*}
\]

Therefore, conditionally on the observed data and the other parameters, the missing counts \((s_1, s_5)\) also have a multinomial distribution:

\[ (s_1, s_5) | \text{the rest} \sim \text{Multinomial}_2 \left( S_i - s_2 - s_3 - s_4 - s_6, \left[ \frac{p_1}{p_1 + p_5}, \frac{p_5}{p_1 + p_5} \right] \right). \]

Thus \(s_1\), for example has a binomial distribution, conditionally on the observed data and parameters, and this is readily simulated by adding Bernoulli's. Note that upon updating the \(s_1\) and \(s_5\) we must update the statistics \(\bar{P}_i, \bar{Q}_i, T_i\) and \(T_{2,i}\).

### A.3 Running the algorithm

For each prior specification, a single "long run" realization of the Metropolis-Gibbs chain is used for posterior summaries. Each step of this chain is a scan over all the parameters and the missing data. The component updates of the scan are those described in the last section. Very little dependence was observed within realizations of the Markov chain, and independent runs were made for every prior specification. Each chain was relatively short, at 800 scans. Rather than invoke sophisticated stopping or subsampling rules, a simple approach was
taken. Marginal posterior probabilities were computed by averaging conditional probabilities at 200 states; every fourth step of the chain. Consistency of the results from similar priors showed that this procedure produced sufficiently accurate estimates in this setting. Several longer runs under particular priors confirmed this assessment.

To ensure accurate estimates of marginal posterior probabilities, we average conditional probabilities rather than averaging indicators of a given state. For example, to estimate $P(\theta_{p,i} = 0|\text{the data})$, we average probabilities:

$$P(\theta_{p,i} = 0|\text{everything else}) = \begin{cases} 
1 + \left(\frac{\theta}{\alpha}\right)^{P_i + T_{1,i}} \left(\frac{1-\theta}{1-\alpha}\right)^{P_i + T_{2,i}} \left(\frac{\eta}{1-\eta}\right)^{-1} & \text{if } \theta_{q,i} = 0 \\
1 + \left(\frac{\theta}{\alpha}\right)^{P_i} \left(\frac{1-\theta}{1-\alpha}\right)^{P_i} \left(\frac{\eta}{1-\eta}\right)^{-1} & \text{if } \theta_{q,i} = 1
\end{cases}$$

Analogous expressions exist for the probability of a suppressor gene on the $q$ arm of chromosome $i$, and for the probability that a suppressor gene is on either the $p$ or the $q$ arm.
Table 1: Data from all lines can be arranged, for each chromosome, in a contingency table like this one. The five-digit binary code identifies a cell of the table by the rule $(m_{i,j}, n_{i,j}, X_{i,j}, Y_{i,j}, Z_{i,j})$. Cells which are empty by constraint are not shown, but include for example, 00111 indicating a tied loss where no arms are available to be lost. Although 11001 and 11000 are not observed, there sum 1100 (the number of non-losses when both $p$ and $q$ may be lost) is.

<table>
<thead>
<tr>
<th>p-arm available</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>lost</td>
<td>no</td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>tied</td>
<td>no</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>11111</td>
<td>11110</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>11010</td>
<td>01010</td>
</tr>
<tr>
<td>no</td>
<td>11100</td>
<td>01000</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>tied</td>
<td>no</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>11001</td>
<td>11000</td>
<td></td>
</tr>
<tr>
<td>10100</td>
<td>10000</td>
<td></td>
</tr>
</tbody>
</table>

17
<table>
<thead>
<tr>
<th>chromosome</th>
<th>11111</th>
<th>11110</th>
<th>11010</th>
<th>11100</th>
<th>1100</th>
<th>01010</th>
<th>01000</th>
<th>10100</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3: Odds ratios for the Metropolis update of the suppressor gene indicator parameters: Here, $u = \beta/\alpha$, $v = (1 - \beta)/(1 - \alpha)$, and $w = \eta/(1 - \eta)$. Notice that $u > 1$ and $v < 1$.

<table>
<thead>
<tr>
<th>current state</th>
<th>candidate state</th>
<th>odds ratio $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{p,i}$</td>
<td>$\theta_{q,i}$</td>
<td>$\theta_{p,i}$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
References


Figure 1: Schema of multistep strategy used for transformation of normal human uroepithelial cells (as reviewed in [4]). An immortal, non-tumorigenic cell line (SV-HUC) was established following infection with simian virus 40. Subsequent single or multiple stepwise tumorigenic transformation, as assessed by tumor formation in athymic nude mice, was established following infection with simian virus 40. Subsequent single or multiple stepwise tumorigenic transformation, as assessed by tumor formation in athymic nude mice. Tumors were classified as high or low grade after examination by two urologic pathologists [7]. Cell lines were established from tumors and used for cytogenetic and statistical analysis.
Figure 2: Chromogram: losses and potential losses from all 16 lines. Top half records information on $p$ arms, and the bottom half $q$ arms. Chromosomes are arranged horizontally. The total number of arm-losses at each chromosome is the number of shaded boxes. The outer box records the potential number of losses. Horizontal hashing indicates tied loss.
Figure 3: The two panels show posterior probabilities of a suppressor gene at either the $p$ arm or the $q$ arm for two values of the prior probability.
Figure 4: Lines show the log Bayes factor of at least one suppressor gene as a function of the prior probability $\eta$. The Bayes factor, equal to the ratio of posterior odds to prior odds of at least one gene, is the probability of the data given at least one gene divided by the probability of the data given no suppressor genes on either arm. Large values of the log Bayes factor indicate high posterior probability that the loss probability on one arm of that chromosome is elevated.
Figure 5: The two panels show posterior probabilities of a suppressor gene at each arm for two values of the prior probability. Dashed lines indicate the prior.
Figure 6: As in Figure 4, this plot shows the log Bayes factor for different priors. These Bayes factors relate to the probability of a suppressor gene on each individual arm, rather than on the whole chromosome.
Figure 7: This plot shows that the posterior distributions of both the baseline and elevated loss probabilities are quite insensitive to the prior $\eta$. Solid lines indicate the posterior means, and dashed lines indicate error bars at 2 posterior standard deviations from the mean.
Figure 8: This plot considers the total number of suppressor genes in the 37 evaluated arms. The prior distribution of this number is Binomial with mean $37\eta$. The posterior can be approximated by the empirical distribution of this number in the dynamic simulation algorithm. The dashed line is the identity.
Figure 9: In the first analysis, we have assumed that the $\tau_i$’s are iid uniform on $(0,1)$, a priori. This plot shows the estimated posterior means of the $\tau_i$ given the data.
Figure 10: Marginal posterior probabilities of suppressor-genes for two extreme priors on the tie-tendency parameters $\tau_i$: Compare this to the top panel of Figure 5 showing the same posteriors under a uniform prior for $\tau_i$. 