Biostatistics Advance Access published February 16, 2006

On estimating the polyclonal fraction in lineage-marker studies of tumor origin

Michael A. Newton[1]

Department of Statistics, University of Wisconsin–Madison,

1300 University Ave, Madison WI 53706, U.S.A., and

Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison,

600 Highland Ave., Madison WI 53792, U.S.A.

e-mail: newton@stat.wisc.edu

phone/fax: 608-263-0357/608-262-0032

Revision of MS B05-046: January 2006

[1]A draft of this work was presented in UW Statistics Department Technical Report # 1099. The work grew from a project in W.F. Dove's lab and from meetings with L. Clipson, R. Halberg, R. Sullivan, S. Stanhope and A. Thliveris, and was supported by grants from the National Cancer Institute: R01 CA63464 (PI MA Newton) and R37 CA63677 (PI WF Dove).

On estimating the polyclonal fraction in lineage-marker studies of tumor origin

Michael A. Newton

Summary

Insight into the biology of tumor formation is provided by studies which demonstrate through the use of cell-lineage markers that some tumors have a polyclonal origin. Novelli *et al.* 1996 proposed to use the proportion of heterotypic tumors among the tumors that are either heterotypic or pure and of the minority marker type as a lower bound on the marginal fraction of polyclonal tumors. Generally, Novelli's ratio does not provide a valid lower bound for the marginal polyclonal fraction, as we demonstrate by analyzing relevant conditional probabilities. Estimation of the polyclonal fraction requires modeling assumptions on the distribution of the number of involved clones. Using three elementary models, we develop maximum likelihood estimation of the polyclonal fraction. We establish robustness of our estimates to misspecification of the clone-marking process, though the estimates are sensitive to assumptions about polyclonal mechanisms. On data from several published studies, our estimates of the polyclonal fraction are substantially smaller than Novelli's ratio.

Key words: cancer biology; conditional probability; Novelli's ratio

1 Introduction

A tumor has a monoclonal origin if early in development its constituent cells descend from a single ancestral cell that is aberrant relative to normal tissue. Otherwise it has a polyclonal origin. Cellular events at the genesis of tumor growth are naturally difficult to measure; clonality studies have provided significant insights, but questions persist about the frequency and functional role of polyclonality. Many studies present evidence supporting the prevailing view, which is that polyclonality is the exception rather than the rule in tumor formation (e.g., Linder and Gartler 1967; Vogelstein *et al.* 1985; Fearon *et al.* 1987). Other studies

suggest that polyclonality may have an important role, especially for certain intestinal tumors (e.g., Buetler *et al.* 1967; Hsu *et al.* 1983; Novelli *et al.* 1996; Merritt *et al.* 1997; Thliveris *et al.* 2005).

The premiss of clonality studies is that cells presenting different states of a binary lineage marker belong to different clones. Thus, tumors presenting both states of the lineage marker presumably have polyclonal origin. Due to X-chromosome inactivation early in development, each tissue in a female who is heterozygous at a marker locus is a mosaic of cells presenting one or the other variant of the marker. X-chromosome-inactivation markers have been used in many clonality studies. A different marker was used in Novelli *et al.* (1996). Intestinal tumors (microadenomas) had been measured in an unusual patient who not only had inherited a defective tumor suppressor gene, making him susceptible to intestinal cancer, but whose tissues were mosaic with respect to the presence of the Y chromosome. The presence (XY) or absence (XO) of Y could be measured in cells, and this formed a binary lineage marker. Aggregation chimeras enable lineage marking in recent clonality studies using mouse models of intestinal cancer (Merritt *et al.* 1997; Thliveris *et al.* 2005). Briefly, two early mouse embyros (morulae) are fused and ultimately produce a single mouse in which the tissue is a mosaic of contributions from both embryos. One embryo is designed to carry a certain reporter gene in order to easily evaluate the embryonic origin of a cell of interest in the adult chimeric mouse.

In large part, evidence regarding the clonality of tumor origin has been inconclusive owing to limitations of lineage marking and possible measurement errors. Fearon *et al.* (1987) noted some problems in previously reported studies. For example, multiple clones would appear to exist in a monoclonal tumor if normal epithelial or stromal cells happened to contaminate the tumor sample; and in enzyme polymorphism studies the level of expression would not necessarily be uniform among different clones. The Fearon *et al.* study applied a DNA-based assay to 50 intestinal tumors. Every tumor presented a single state of the binary marker, in support of monoclonality. Subsequent calculations, however, showed that the Fearon *et al.* study had low power to detect polyclonality because a very small fraction of the tissue was near patch boundaries in the X-inactivation mosaic (Novelli *et al.* 2003).

Patch structure in the Novelli *et al.* (1996) (X0/XY) case was finer grained and thus provided a greater opportunity to detect polyclonality. However in that study it was possible that the Y chromosome could be lost sporadically; though the estimated rate was low, there was a small chance that tumors presenting both marker variants were actually monoclonal. Subsequent mouse chimera studies, on the other hand, did not suffer from problems with marker fidelity and clearly demonstrated polyclonality. The background rate of adenoma formation was relatively high in the Merritt *et al.* (1997) study. Thliveris *et al.* (2005) used similar methods but engineered the mice to have many fewer tumors overall. In spite of substantial challenges in measuring early events of tumor formation, there is now clear evidence supporting the polyclonal origin of a class of intestinal tumors.

The mechanisms responsible for polyclonality are not well understood. Some polyclonal tumors may emerge simply by the close proximity of distinct initiated clones without a requirement for clonal cooperation. Tumor multiplicity was quite low in the Thliveris *et al.* (2005) study of murine intestinal adenomas, and so this so-called random collision hypothesis was considered unlikely. If polyclonality is necessary for certain tumors to grow, then there are intercellular interactions of importance to the initiation and maintenance of the tumor. The existence of such interactions would strain the standard model which holds that a tumor develops according to a monoclonal cell lineage within which genetic damage accumulates (e.g., Nowell 1976). As improved methods are applied to study the earliest events of tumor growth, the precise role of polyclonality will be clarified. A pervasive statistical question in this effort is how to estimate the fraction of polyclonal tumors from data obtained in lineage-marker studies. This question is the focus of the present paper.

Regarding statistical concerns, there is the problem that the fraction of polyclonal tumors may be different from the fraction of tumors that appear to have polyclonal origin according to lineage marker data. Take the data from Novelli *et al.* (1996) as illustrative. The patient presented with 263 microadenomas in his intestinal tract; 4 of these were pure (homotypic) and of the minority XO type, 246 of these were homotypic and of the majority XY type, and the remaining 13 were heterotypic in that they contained cells of both marker types. These 13 tumors were overtly polyclonal; assuming fidelity of the marker, none could have formed

as cells descendant from a single initiated aberrant cell. Quite possibly, covertly polyclonal tumors were among the 250 homotypic tumors, though the actual number of such polyclonal tumors cannot be assessed because a binary lineage marker does not have the resolution to distinguish different clones within a tumor that happen to have the same marker type. All heterotypic tumors are polyclonal, but not all polyclonal tumors are heterotypic.

Recognizing the inherent missing-data structure, Novelli *et al.* (1996) proposed, as a lower bound on the fraction of polyclonal tumors, the proportion of heterotypic tumors among those that are either heterotypic or homotypic of the minority marker type. That became $13/(13+4) = 76\%$ for these data. It is a rather impressive inference, since we know with confidence only that the polyclonal fraction exceeds the heterotypic fraction, estimated at $13/263 = 5\%$. Merritt *et al.* (1997) used the same ratio technique to bound the polyclonal fraction in tumor count data from mouse aggregation chimeras. The rates estimated by this Novelli ratio technique have been reported in various reviews (e.g., Playford 1998; Garcia *et al.* 1999). Through an analysis of conditional probabilities, we show that the Novelli ratio technique is flawed. In doing so we identify two key stochastic components of lineage marker data, and further show that model-free estimates of the polyclonal fraction cannot improve the proportion of heterotypic tumors as a lower bound on the polyclonal fraction. Model-based methods are developed, and we show that these are robust to certain forms of model violation, but not to others. These findings have guided some of the statistical calculations in Thliveris *et al.* (2005), and would seem to have relevance in future clonality studies.

2 The inference problem

Of interest are tumors that originate within intestinal epithelial tissue, though none of the statistical reasoning is restricted to this site. We suppose that cells in the tissue can be classified as either normal or abnormal; for our purposes the detailed distinctions among abnormal cells (e.g. adenomas/carcinomas) are not important, and we consider that abnormal cells populate tumors. Classification of cells may be based on histopathology

to detect abnormal cell morphology or immunohistochemistry to detect certain proteins produced in tumor cells (e.g. Merritt *et al.* 1997). To study tumor origin, one needs to consider initiation events, each of which irreversibly transforms a normal cell into an abnormal state. We equate a tumor clone with the full set of extant cells that descend from such an initiated cell via cell proliferation within the tumor. Cells comprising a tumor either form a single clone or partition the tumor mass into multiple clones (owing to multiple initiation events). Thus in the population of intestinal tumors under study, a fraction $f(c)$ of tumors are formed from exactly c clones, for $c = 1, 2, \ldots$. This forms the probability mass function of C, the number of clones in a randomly sampled tumor. Underlying f is a stochastic process governing how clones are bound together to form tumors. Three elementary, mechanistic models of this *clone-binding process* are presented in Section 4.

The sampled tumor is monoclonal if $C = 1$; otherwise it is polyclonal. The polyclonal fraction

$$\theta = P(C > 1) = \sum_{c=2}^{\infty} f(c) = 1 - f(1)$$

is the parameter of primary interest. Ideally we can consistently estimate θ from available data. Because any evidence that $\theta > 0$ is in conflict with the standard theory of monoclonal tumor origin, an informative lower bound is useful in conjunction with any point estimate of θ.

Lineage-marker studies provide partial information about the clonal structure of tumors and thus enable inference about the polyclonal fraction θ. Each cell in the tissue assumes one of a finite number of marker types which marks the cell and any descendant cells. All studies to date have used two types, say $\{1, 2\}$, though more are biologically possible and could be readily considered in our statistical analysis. Fidelity of the marker through cell proliferation is essential, otherwise we cannot, from measurements on extant cells, conclude much of anything about the type of ancestral cells that existed at the time of tumor initiation. Auxiliary data may support the marker-fidelity hypothesis, and we adopt this hypothesis in what follows.

For a tumor sampled from the population under study, let $N(t)$ denote the number of

clones of type t. Naturally $C = N(1) + N(2)$ in the case involving binary types. The tumor is homotypic of type t if $N(t) = C$. It is heterotypic if it is not homotypic for either type. We can observe which of the three mutually exclusive events has occurred:

$$\mathrm{HOM}_1 = [N(1) = C]\,, \quad \mathrm{HOM}_2 = [N(2) = C]\,, \quad \mathrm{HET} = [N(1) > 0] \cap [N(2) > 0]\,.$$

Current measurements do not allow us to know C or $N(t)$ for either t; they simply indicate the value of a trinomial random variable for each tumor. Some sort of *clone-marking process* characterizes the conditional distribution of $N(t)$ given $C = c$. Possible models for the clone-marking process are discussed in Section 5.

Lineage-marker studies of polyclonality offer two classifications of a tumor population: (1) clonality, i.e. whether $C = 1$ or $C > 1$, and (2) phenotype, i.e. whether $N(t) = C$ for some t, or not. Table 1 shows the cross classification of such a population in terms of these factors. Tumor count data provide direct information on the marginal row proportions, but complete data are not available on entries inside the table. Assuming marker fidelity, no tumors can be both heterotypic and monoclonal, and this forces a structural zero in the table.

Table 1: Cross classification of a tumor population in terms of clonality and phenotype.

		Clonality		
Phenotype		monoclonal $[C = 1]$	polyclonal $[C > 1]$	
homotypic	type 1 $[\mathrm{HOM}_1]$	$P(\mathrm{HOM}_1 \cap (C = 1))$	$P(\mathrm{HOM}_1 \cap (C > 1))$	$P(\mathrm{HOM}_1)$
	type 2 $[\mathrm{HOM}_2]$	$P(\mathrm{HOM}_2 \cap (C = 1))$	$P(\mathrm{HOM}_2 \cap (C > 1))$	$P(\mathrm{HOM}_2)$
heterotypic	$[\mathrm{HET}]$	0	$P(\mathrm{HET} \cap (C > 1))$	$P(\mathrm{HET})$
		$P(C = 1)$	$\theta = P(C > 1)$	100%

In summary, trinomial phenotype data are available from tumors sampled from a relevant population. Stochastic processes governing the biology of clone binding and clone marking affect the distribution of these data. There is substantial missing information, but also there are structural constraints which relate parameters and guide inference about the polyclonal

fraction θ.

3 No model-free lower bound improves $P(\mathrm{HET})$

Evidently $P(\mathrm{HET}) \leq \theta$ because all heterotypic tumors are polyclonal (Table 1). This assertion relies on the marker-fidelity assumption, but it requires no assumptions on either the process by which clones are bound into tumors or the process by which clones attain marks. In this sense it is model free. Though valid, the bound $P(\mathrm{HET}) \leq \theta$ is not tight when a substantial fraction of the homotypic tumors are also polyclonal.

First Novelli *et al.* (1996) and then Merritt *et al.* (1997) used a certain ratio aiming to produce a tighter lower bound for θ. From a sample of tumors, *Novelli's ratio* is the proportion of heterotypic tumors among those that are either heterotypic or homotypic and of the minority type. The empirical value is:

$$\hat{\beta} = \frac{\#\{\mathrm{HET}\}}{\#\{\mathrm{HET}\} + \#\{\mathrm{HOM}_1\}}, \tag{3.1}$$

where type $t = 1$ homotypic tumors are less frequent than type $t = 2$ homotypic tumors. This estimates the population quantity

$$\beta = P(\mathrm{HET})/P(\mathrm{HET} \cup \mathrm{HOM}_1). \tag{3.2}$$

A clear rationale for the claim that $\beta \leq \theta$ was not provided in Novelli *et al.*, but evidently there was no appeal to particular modeling assumptions. The idea may have been simply this: among the heterotypic and minority-homotypic tumors, the polyclonal fraction is

$$\begin{aligned} \theta^* &= P(C > 1 \,|\, \mathrm{HET} \cup \mathrm{HOM}_1) \\ &= \frac{P\{\mathrm{HET} \cap (C > 1)\} + P\{\mathrm{HOM}_1 \cap (C > 1)\}}{P(\mathrm{HET} \cup \mathrm{HOM}_1)} \\ &= \frac{P(\mathrm{HET}) + P\{\mathrm{HOM}_1 \cap (C > 1)\}}{P(\mathrm{HET} \cup \mathrm{HOM}_1)} \\ &= \beta + \epsilon. \end{aligned} \tag{3.3}$$

Here the development uses $\mathrm{HET} \subset (C > 1)$ as noted in Table 1. The term $\epsilon \geq 0$ is liable to be small if the minority cell type is a small proportion of the whole, since multiple clones from

that minority component have to somehow interact to form each tumor. Regardless of the magnitude of ϵ, we have a valid bound $\beta \leq \theta^*$. Thus Novelli's ratio β does bound a certain polyclonal fraction, but it is not θ, the marginal polyclonal fraction of interest; rather β is a lower bound on the rate θ^* of polyclonality among the heterotypic and minority-homotypic tumors. Were there some sort of conditional independence, it would follow that the bound also holds marginally. This is not so. In fact, in the population of heterotypic and minority-homotypic tumors, polyclonality is more frequent than in the whole population of tumors (see theorem below). There is a positive gap between θ and the larger θ^*, which creates a problem; for if β lies in this gap then it is not a lower bound for the marginal polyclonal fraction θ (see Figure 1). Further, whether or not β lies in the gap depends on details of the stochastic processes generating the data, and so β can not be a general purpose, model-free, lower bound.

The gap affecting Novelli's ratio is always non-negative. Some conditions are required to establish strict positivity. For one, we require $0 < \theta < 1$. But this is innocuous; if $\theta = 1$ then any quoted rate would provide a valid lower bound for θ; on the other hand if $\theta = 0$ then all tumors would be homotypic and the question of polyclonality would not have surfaced in the first place. We make no specific assumptions about clone binding or clone marking. However we do require a weak technical assumption about the latter. Consider that in a population of tumors comprised of monoclonal tumors and, for various $c \geq 2$, tumors originating by the interaction of c clones, we have an overall proportion γ_t of clones that are of type t. More formally,

$$\gamma_t = \frac{\sum_{c=1}^{\infty} f(c) \left[\sum_{n=1}^{c} n \, P\left\{N(t) = n \,|\, C = c\right\}\right]}{\sum_{c=1}^{\infty} c f(c)}, \tag{3.4}$$

which arises from consideration of size-biased sampling, as long as $E(C) < \infty$ (e.g., Patil and Rao 1978). Probabilities $P\left\{N(t) = n \,|\, C = c\right\}$ for various c and n reflect the possibly complex clone-marking process.

Definition: The clone-marking process is *regular* if for each clonal type t, $0 < \gamma_t < 1$, and also if for each $c \geq 2$ for which $f(c) > 0$, $P\left\{N(t) = c \,|\, C = c\right\} < P\left\{N(t) = 1 \,|\, C = 1\right\} = \gamma_t$.

Roughly speaking, regularity means that homotypic type t tumors are more frequent among

monoclonal tumors than they are among polyclonal tumors. The assumption holds for a range of plausible stochastic processes, such as those in which the marking is neutral and thus independent, in a certain sense, from the clone-binding process. We take up the point shortly. First we state the main theoretical result which is key to the flaw in Novelli's ratio.

Gap theorem: *If* $0 < \theta < 1$ *and the clone-marking process is regular, then* $\theta < \theta^*$.

The value of lineage-marker studies derives in part from the possibility that the marking process itself does not alter the polyclonal structure. This concept of neutrality is stronger than the concept of regularity required in the gap theorem. To be specific, reconsider $N(t)$, the number of clones of type t that are bound together in a sampled tumor. One definition of neutral marking is to have that the expected proportion of type t clones in clonality-c tumors does not depend on c; i.e. $E\{N(t)/C \,|\, C = c\} = \gamma_t$. Owing to discreteness we cannot have that $N(t)/C$ is independent of C, but we can ask that on average the proportion of type t clones in a tumor matches the proportion of type t clones overall. If the marking process is neutral and allows heterotypic tumors, i.e. if $P(\mathrm{HET}|C = c) > 0$ for all $c > 1$ for which $f(c) > 0$, then from (3.4), it follows routinely that the marking process is also regular. Thus neutrality implies regularity.

As an example of a non-regular marking process, suppose that tumors can be either monoclonal (with probability $f(1) = 1/2$), or biclonal (with probability $f(2) = 1/2$). Suppose further that all monoclonal tumors are marked with type 1, and all biclonal tumors are marked with type 2. The phenotypes and the clonality are highly dependent in this case and seem far from a neutral marking process. Overall among clones, $\gamma_2 = 2/3$ are of type 2, yet $P\{N(2) = 2|C = 2\} = 1$, $P\{N(2) = 1|C = 1\} = 0$, which clearly violates the definition of a regular marking process.

An elementary, though useful, neutral marking process entails independent type assignments according to distribution $\{\gamma_t\}$ over types. Independence requires few parameters, but it conflicts with the spatial patterning evident in real tissue that is a mosaic of different types (Griffiths *et al.* 1989; Novelli *et al.* 2003; Thliveris *et al.* 2005). Neutral marking can respect this sort of patterning through positive association by boosting the

homotypic rate $P\{N(t) = c \,|\, C = c\}$ above the independence homotypic rate γ_t^c.

Taking these concepts to a concrete example, consider a simplified model in which tumors are monoclonal with probability $f(1) = 1 - \theta$ or are formed from two clones, and thus are biclonal with probability $f(2) = \theta$. Tumor-bound clones are marked independently by one of two types, with the minority type $t = 1$ having frequency $\gamma_1 < 1/2$. Evaluating (3.3), the proportion of biclonal tumors among the heterotypic or pure type-1 tumors is

$$\begin{aligned} P(C > 1 \,|\, \mathrm{HET} \cup \mathrm{HOM}_1) &= \frac{2\theta\gamma_1(1-\gamma_1) + \theta\gamma_1^2}{2\theta\gamma_1(1-\gamma_1) + (1-\theta)\gamma_1 + \theta\gamma_1^2} \\ &> \frac{2\theta\gamma_1(1-\gamma_1)}{2\theta\gamma_1(1-\gamma_1) + (1-\theta)\gamma_1 + \theta\gamma_1^2} \\ &= \frac{2\theta(1-\gamma_1)}{1 + \theta(1-\gamma_1)} \\ &= \beta \end{aligned}$$

As ensured by (3.3), Novelli's ratio β does provide a lower bound for a certain conditional polyclonal fraction. However there is a gap between that conditional fraction θ^* and the smaller marginal polyclonal fraction θ of interest, and so the bound $\beta \leq \theta$ can fail. Figure 2 charts the difference $\Delta = \theta - \beta$ for different polyclonal fractions and different minority type frequencies γ_1. When both θ and γ_1 are large, Novelli's ratio provides a legitimate bound because $\Delta > 0$. The bound fails when $\Delta < 0$. In terms of state-space area, the bound fails for most scenarios. The error is particularly extreme in the realistic situation where the minority fraction is small.

What statistical recourse is there for inference about θ? The weak lower bound $P(\mathrm{HET}) \leq \theta$ is the best one can do without adopting modeling assumptions on clone binding and marking. Mathematically, for example, it is possible that tumors are either monoclonal or polyclonal of some large degree c, and are marked by some simple marking scheme. If this were the case, virtually all the polyclonal tumors would be heterotypic, and so the simple lower bound $P(\mathrm{HET}) \leq \theta$ would be tight.

Curiously, there is a modification of the Novelli ratio which provides a valid lower bound for θ in the special monoclonal/biclonal model, though not generally. Peter Sasieni (personal communication) proposed to replace the denominator in Novelli's ratio (3.1) with the number

of tumors that are heterotypic plus twice the number of minority homotypic tumors.

4 Model-based inference is sensitive to clone-binding assumptions

Three elementary models of clone binding are:

1. **Monoclonal/Biclonal**: As in Section 3, polyclonality is equivalent to biclonality. This is the simplest form of polyclonality. One justification is parsimony; the model is a minimal representation of interacting clones.

2. **Conditional Poisson**: The number C of clones in a tumor has probability mass function

$$f(c) = \frac{\lambda^c \exp(-\lambda)}{c!} \frac{1}{1 - \exp(-\lambda)}$$

for $c = 1, 2, \ldots$ and $\lambda > 0$. This is a Poisson distribution conditioned on at least one clone, and could be justified under some model of random collision or random collision followed by selection if there is sufficient tumorigenic potential (Newton *et al.* 2006). Here, the polyclonal fraction is $\theta = 1 - \lambda / \{\exp(\lambda) - 1\}$.

3. **Geometric**: The number C of clones in a tumor has probability mass function

$$f(c) = \psi(1 - \psi)^{c-1} \qquad \text{for } c = 1, 2, \ldots$$

and $\psi \in (0, 1)$. This model might be justified if aberrant clones engage in some sort of recruitment and conversion of additional clones (Shih *et al.* 2001). Here the polyclonal fraction is $\theta = (1 - \psi)$.

Likelihood-based inference for θ is possible if we invoke a clone-marking model on top of the clone-binding model. The simplest one is to mark the clones that are bound in a tumor independently and according to a common distribution over types $\{\gamma_t\}$. A better model would entail some positive association among bound clones since they are constrained

spatially and there may be a semi-regular patchwork pattern of lineage markers within the tissue. However, it could be computationally challenging to incorporate detailed information about positive association. Maximum likelihood estimation has some validity even in the absence of independent marking. We argue in Section 5 that the maximum likelihood estimate obtained under the independent-marking assumption is conservatively biased, in the sense of converging to a lower bound on θ, regardless of the positive association among clones bound in a polyclonal tumor.

Likelihood-based inference requires the marginal probability of a homotypic tumor of type t, which is obtained by summing over the unknown clonality C. For the three binding models presented above, and with independent marking, these sums can be solved explicitly.

1. **Monoclonal/Biclonal**: $P(\mathrm{HOM}_t) = (1-\theta)\gamma_t + \theta\gamma_t^2$
2. **Conditional Poisson**: $P(\mathrm{HOM}_t) = \frac{\exp(\lambda\gamma_t)-1}{\exp(\lambda)-1}$
3. **Geometric**: $P(\mathrm{HOM}_t) = \frac{\gamma_t\psi}{1-\gamma_t(1-\psi)}$

The tumor sample is viewed as a multinomial draw according to these type probabilities, allowing for the heterotypic class to have probability equal to the complement of the sum of these homotypic class probabilities. We have not found a closed form expression for the maximum likelihood estimates, but they may be obtained routinely by numerical methods. One may either use external estimates of the clonal marker frequencies $\{\gamma_t\}$, or these may be also estimated from the count data.

Table 2 shows the maximum likelihood estimates of θ for data from Novelli *et al.* (1996) and for data from Merritt *et al.* (1997). The estimates are rather different from Novelli's ratio in these examples. We obtained approximate 95% confidence intervals by first computing a profile likelihood function in each case (optimizing numerically in the rate parameter γ_1) and then normalizing the profile likelihood to be an approximate marginal posterior distribution for θ. Confidence intervals mark the central 95% of these distributions. Results from one data set are amplified in Figure 3, which reveals the lack of robustness of estimates for θ to changes in the clone-binding model.

Table 2: Estimation of polyclonal fraction θ: For three data sets shown on left, reported are MLEs and approximate 95% confidence intervals (CI) using three different models for how clones are bound into tumors: monoclonal/biclonal (MB), conditional Poisson (CP), and Geometric (Geo). Also shown are values of the Novelli ratio $\hat{\beta}$ and the naive lower bound LB which is simply the observed proportion of heterotypic tumors. All proportions are shown as percentages.

	Tumor Counts					MLE $\hat{\theta}$ (95% CI)		
Data set	HOM_1	HOM_2	HET	LB	$\hat{\beta}$	MB	CP	Geo
Novelli	4	246	13	5	76	64 (40,97)	55 (35,78)	53 (34,72)
Merritt 112	5	93	7	7	58	46 (23,76)	40 (21,66)	37 (20,60)
Merritt 113	1	139	15	10	94	94 (62,99)	83 (55,96)	77 (52,92)

5 Model-based inference is robust to clone-marking assumptions

Maximum likelihood estimates obtained under the independent-marking model will be biased if there is positive association amongst the types of the bound clones. Such positive association is expected owing to the typical patchy structure of mosaic tissue. However, we show that this bias is expected to be conservative, (i.e., the estimates ought to be low) since independent marking puts more probability mass on heterotypic tumors than would a more realistic positive-association marking. To establish the conservative bias, suppose that clone-type frequencies $\{\gamma_t\}$ are known or can be consistently estimated. Under independent marking, a tumor will be homotypic type t with probability $\alpha_t(\theta) = (1-\theta)\gamma_t + \sum_{c\geq 2} f(c)\gamma_t^c$. Positive association of clonal marking amounts to an increased homotypic rate $\beta_t^*(\theta) \geq \alpha_t(\theta)$. The rate of heterotypic tumors under independent marking is $a(\theta) = 1 - \sum_t \alpha_t(\theta)$ and under positive association is $b(\theta) = 1 - \sum_t \beta_t^*(\theta)$, both positive by regular marking, and satisfying $a(\theta) \geq b(\theta)$. Both functions are in 1-1 correspondence with the polyclonal fraction θ, and so either could be used to parameterize a likelihood computation for the independent-marking model. Suppose that the maximum likelihood estimate for $\psi = a(\theta)$ is derived from a

binomial model on the heterotypic frequency. Even though the independent-marking model is incorrect, the independent-marking estimate of ψ will be consistent for this population heterotypic frequency; but in fitting closely to the data, an incorrect value $\bar{\theta} = a^{-1}(\psi) \neq \theta$ will be converged upon. The correct polyclonal fraction is what we would have converged to using the positive-association model, namely $\theta = b^{-1}(\psi)$. Since $a(\theta) \geq b(\theta)$, the value $\bar{\theta}$ to which the independent-marking estimator converges must be no greater than the true polyclonal fraction θ.

6 Conclusions

The cellular and molecular events that characterize the earliest stages of intestinal tumor development are not fully understood. In particular, the question of tumor clonality – does a tumor derive from more than one initiated cell? – has remained somewhat elusive. Lineage-marker studies provide the approach to address clonality, but many factors affect the information which can be usefully extracted from lineage-marker data: (1) the marker must have fidelity otherwise it is not transmitted faithfully through cell division. Ideally the marker is not affected in any way by tumor growth, and simply records lineages (e.g., this fails if marker variants are created in subclones inside a developing tumor); (2) the marker's mosaic pattern in tissue must be fine-grained so that truly polyclonal tumors have sufficient opportunity to be heterotypic, otherwise there is insufficient power; (3) the measurements must be taken early in tumor development else a dominant clone may grow out and mask earlier polyclonal structure (e.g. Bühler 1967); and (4) measurements must be taken with great care to ensure that normal clones do not contaminate the tumor and lead to a false heterotypic determination. Not all clonality studies have satisfied these requirements, but existing data do indicate the polyclonal origin of a class of intestinal tumors.

Even the ideal lineage-marker study entails a statistical inference problem. We have discussed aspects of the problem to estimate the polyclonal fraction and have shown through an analysis of conditional probabilities that Novelli's ratio does not provide a valid lower bound for this fraction. Without assumptions on the process by which clones are bound into

a tumor, the heterotypic fraction is the best lower bound. Maximum likelihood estimates may be derived using simplified model assumptions, and under certain conditions these simplified estimates are robust. Though precise estimation of the polyclonal fraction is difficult, other parameters describing tumor initiation can be inferred when tumor-count data are combined with spatial information about the mosaic patch structure in lineage-marker studies. For example the extent of spatial interaction among clones was estimated in Thliveris *et al.* (2005).

References

Beutler, E., Collins, Z., and Iriwin, L. (1967). Value of genetic variants of glucose-6-phosphate dehydrogenase in tracing the origin of malignant tumors. *N Engl J Med*, **276**, 389–391.

Bühler, W. J. (1967). Single cell against multicell hypotheses of tumor formation. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume IV, Berkeley and Los Angeles, pp. 635–637. University of California Press.

Fearon, E. R., Hamilton, S. R., and Vogelstein, B. (1987). Clonal analysis of human colorectal tumors. *Science*, **238**, 193– 197.

Garcia, S. B., Park, H. S., Novelli, M., and Wright, N. A. (1999). Field cancerization, clonality, and epithelial stem cells: the spread of mutated clones in epithelial sheets. *Journal of Pathology*, **187**, 61–81.

Griffiths, D., Sacco, D., Williams, G. T., and Williams, E. D. (1989). The clonal origin of experimental large bowel tumors. *British Journal of Cancer*, **59**, 385–387.

Hsu, S. H., Luk, G. D., Krush, A. J., Hamilton, S. R., and Hoover, H. H. (1983). Multiclonal origin of polyps in gardner's syndrome. *Science*, **221**, 951–953.

Linder, D. and Gartler, S. M. (1967). Problem of single cell versus multicell origin of a tumor. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume IV, Berkeley and Los Angeles, pp. 625–633. University of California Press.

Merritt, A. J., Gould, K. A., and Dove, W. F. (1997). Polyclonal structure of intestinal adenomas in $Apc^{Min/+}$ mice with concomitant loss of Apc^{+} from all tumor lineages. *Proc. Natl. Acad. Sci. USA*, **94**, 13927–13931.

Newton, M. A., Clipson, L. C., Thliveris, A. T., and Halberg, R. B. (2006). A statistical test of the hypothesis that polyclonal intestinal tumors arise by random collision of initiated clones. *Biometrics*, **62**, in press.

Novelli, M., Williamson, J. A., Tomlinson, I. P. M., Elia, G., Hodgson, S. V., Talbot, I. C., Bodmer, W. F., and Wright, N. A. (1996). Polyclonal origin of colonic adenomas in an XO/XY patient with FAP. *Science*, **272**, 1187–1190.

Novelli, M. R., Cossu, A., Oukrif, D., Quaglia, A., Lakhani, S., Poulsom, R., Sasieni, P., Carta, P., Contini, M., Pasca, A., Palmieri, G., Bodmer, W., Tanda, F., and Wright, N. (2003). X-inactivation patch size in human female tissue confounds the assessment of tumor clonality. *Proceedings of the National Academy of Science*, **100**, 3311–3314.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, **194**, 23–28.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and sized biased sampling with application to wildlife populations and human families. *Biometrics*, **34**, 179–189.

Playford, R. J. (1998). Tales from the crypt – intestinal stem cell repertoire and the origins of human cancer. *Journal of Pathology*, **185**, 119–122.

Shih, I., Wang, T. L., Traverso, G., Romans, K., Hamilton, S. R., Ben-Sasson, S., Kinzler, K. W., and Vogelstein, B. (2001). Top-down morphogenesis of colorectal tumors. *Proceedings of the National Academy of Science*, **98**, 2640–2645.

Thliveris, A. T., Halberg, R. B., Clipson, L. C., Dove, W. F., Sullivan, R., Washington, M. K., Stanhope, S., and Newton, M. A. (2005). Polyclonality of familial murine adenoma: Analyses of chimeras at low tumor multiplicity reveal short-range interactions. *Proceedings of the National Academy of Science*, **102**, 6960–6965.

Vogelstein, B., Fearon, E. R., Hamilton, S. R., and Feinberg, A. P. (1985). Use of restriction fragment length polymorphisms to determine the clonal origin of human

tumors. *Science*, **227**, 642–645.

Appendix: Proof of gap theorem

We prove something slightly more general than is stated. Let t denote any one of the clonal types, and reconsider the event $\mathrm{HOM}_t = \{N(t) = C\}$, which has probability $0 < P(\mathrm{HOM}_t) < 1$ by regularity. Observe that the polyclonal fraction θ is a weighted average

$$\theta = P(C > 1|\mathrm{HOM}_t)\, P(\mathrm{HOM}_t) + P(C > 1|\mathrm{HOM}_t^c)\, P(\mathrm{HOM}_t^c),$$

where HOM_t^c is the complement of HOM_t. In a two-type system where t is the majority type, $\mathrm{HOM}_t^c = \{\mathrm{HET} \cup \mathrm{HOM}_1\}$, for example. Via convexity, it is sufficient to prove $\mathrm{P}(C > 1|\mathrm{HOM}_t) < \theta$. By Bayes's rule, this is equivalent to $P(\mathrm{HOM}_t) > P(\mathrm{HOM}_t|C > 1)$. Now the marginal $P(\mathrm{HOM}_t)$ is decomposed into non-zero terms according to clonality:

$$\begin{aligned} P(\mathrm{HOM}_t) &= P(\mathrm{HOM}_t|C > 1)\, \theta + P(\mathrm{HOM}_t|C = 1)\, (1 - \theta) \\ &= a\theta + \gamma_t(1 - \theta) \end{aligned}$$

where $\gamma_t \in (0, 1)$ is the marginal rate of type t clones and $a = P(\mathrm{HOM}_t|C > 1)$ has to do with the clone-marking process. Thus the difference

$$P(\mathrm{HOM}_t) - P(\mathrm{HOM}_t|C > 1) = (\gamma_t - a)(1 - \theta).$$

We have assumed $\theta < 1$, and $\gamma_t \in (0, 1)$ in the statement of the theorem, so the theorem is true if $a < \gamma_t$. Considering the possible levels of polyclonality C when $C > 1$,

$$\begin{aligned} a &= P(\mathrm{HOM}_t|C > 1) \\ &= \sum_{c=2}^{\infty} P(\mathrm{HOM}_t|C = c)\, f(c)/\theta \\ &= \sum_{c=2}^{\infty} P(N(t) = c\,|C = c)\, f(c)/\theta \\ &< \left[\sum_{c=2}^{\infty} \gamma_t\, f(c)/\theta\right] = \gamma_t \end{aligned}$$

where $f(c) = P(C = c)$ is the fraction of tumors comprised of c clones. The last inequality follows from the definition of a regular marking process. □

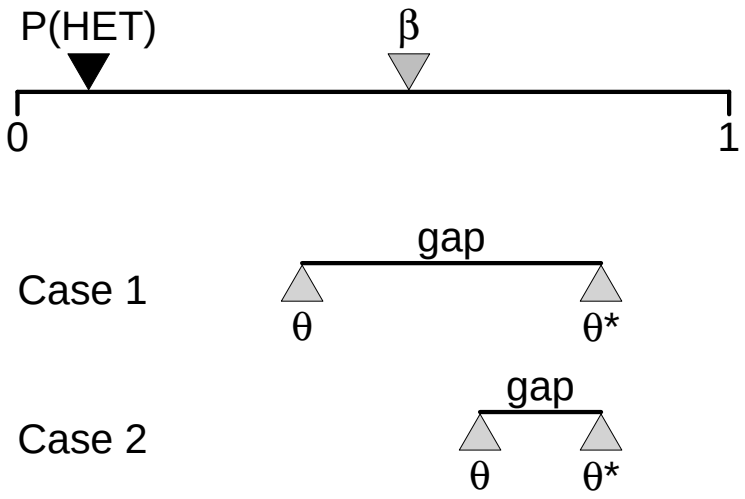

Figure 1: On the gap and Novelli's ratio β: β would be a valid bound if $\theta = \theta^*$, or at least if the gap between θ and θ^* is small (case 2), because $\beta \leq \theta^*$. The gap is of an unknown size, and may be large (case 1), in which case β is not smaller than θ.

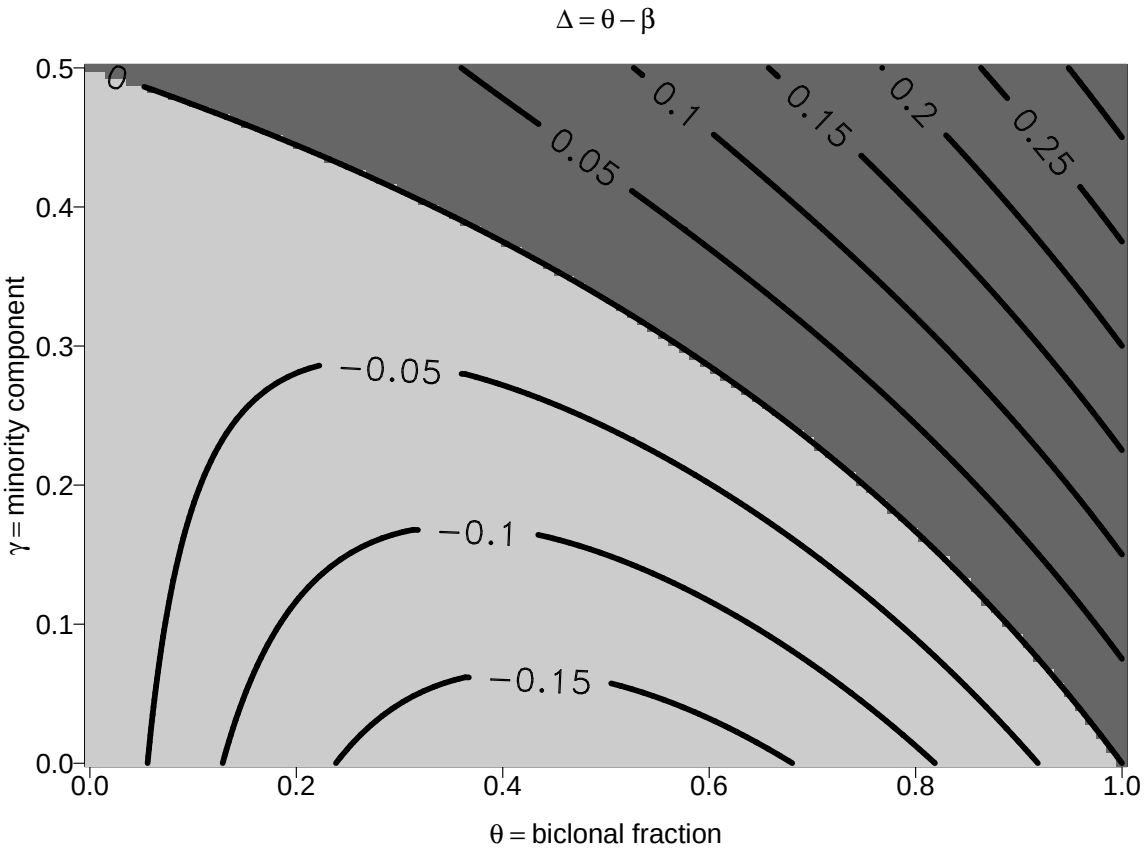

Figure 2: Discrepancy between the polyclonal fraction θ and Novelli's ratio β in the monoclonal/biclonal, independent-marking model as function of the minority fraction γ_1 and the biclonal fraction θ. In the lower left of the plot Novelli's ratio fails to bound the polyclonal fraction. The light gray shaded region corresponds to case 1 (Figure 1), and the dark to case 2.

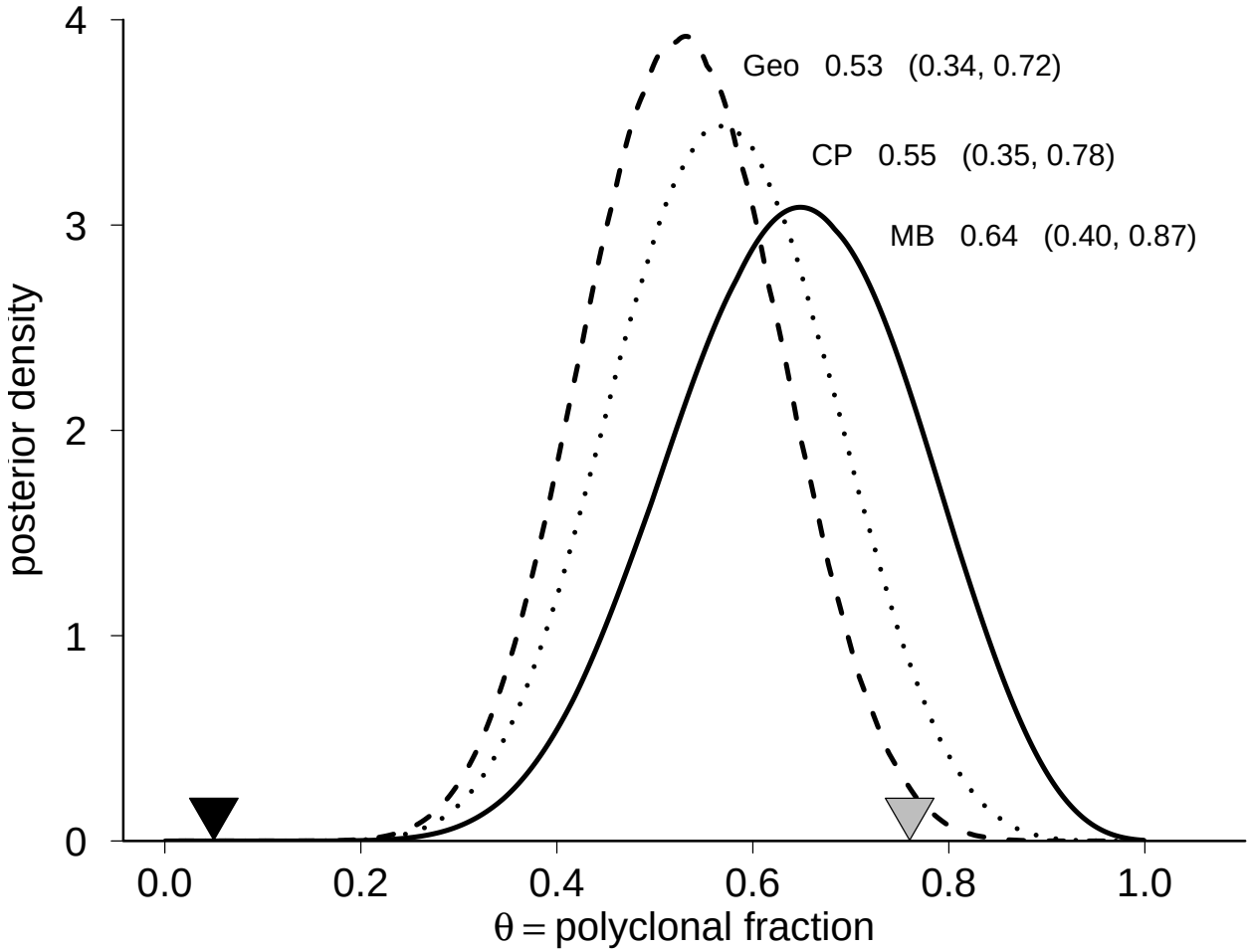

Figure 3: A comparison of three model-based estimates of θ using the clonality count data from Novelli *et al.* (1996): Plotted are profile likelihood functions that are normalized to integrate to one, and thus serve as approximate posterior distributions. The nuisance parameter γ_1, the proportion of blue clones, was removed by maximization in each case. MLEs are noted for each polyclonality model: MB (monoclonal/biclonal); CP (conditional Poisson); Geo (geometric). Approximate confidence intervals were computed as equi-tail 95% posterior intervals. The black triangle indicates the naive lower bound for θ which is the observed proportion of heterotypic tumors (5%). The grey triangle indicates the Novelli ratio $\hat{\beta} = 76\%$. Model-based inference about θ is highly sensitive to assumptions about process by which clones are bound into tumors. In the models considered, the probability is high that θ is less than the supposed lower bound $\hat{\beta}$.