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

Michael A. Newton*

April 1 2005

Abstract

Insight into the biology of early tumor formation is provided by studies which demonstrate through the use of cell-lineage markers that some tumors have a polyclonal composition. 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. On data from several published studies, our estimates of the polyclonal fraction are substantially smaller than Novelli's ratio.

Some key words: cancer biology; conditional probability; Novelli's ratio

1 Background

If all cells in a tumor are descended from a common ancestral cell which itself is aberrant relative to normal cells in the tissue of origin, then the tumor may be viewed as a monoclonal population of cells. The widely held view that tumors are monoclonal is challenged

*Department of Statistics; University of Wisconsin-Madison; 1300 University Avenue; Madison, WI 53706, USA; newton@stat.wisc.edu. This document is Statistics Department Technical Report #1099. The work was inspired by a project from W.F. Dove's lab and from meetings with L. Clipson, R. Halberg, R. Sullivan, S. Stanhope and A. Thliveris. The work was supported by grants from the National Cancer Institute: R01 CA63464 and R37 CA63677.

by studies which demonstrate through the use of cell-lineage markers that some fraction of tumors have a polyclonal composition. An important example comes from a study of familial colorectal cancer reported by 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 useful binary lineage marker; each cell inherited its type from its parent cell. 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 integrity of the marker labeling, none could have formed as cells descendant from a single initiated aberrant cell.

Polyclonality has been an intriguing possibility in cancer research. The conventional model is that a tumor develops from a monoclonal cell lineage within which genetic damage accumulates (e.g., Nowell 1976). However, if polyclonality is necessary for certain tumors to grow, then there would be intercellular interactions of importance to the initiation and maintenance of the tumor. There are also statistical questions raised by studies of polyclonality, though for some time these have not received considerable attention (Bühler 1967; Linder and Gartler 1967; Thliveris *et al.* 2005).

In the Novelli *et al.* study, it was possible that covertly polyclonal tumors were among the 250 homotypic tumors, but the actual number of such polyclonal tumors could not be measured because the binary lineage marker did not distinguish different clones within a tumor that happened to have the same value of the binary marker. All heterotypic tumors are polyclonal, but not all polyclonal tumors are heterotypic. Recognizing the inherent missing-data structure, Novelli *et al.* 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, and, strikingly, obtained the same value 76% when results from two mice were combined. These and related results are reviewed in Dove *et al.* (1998) and Garcia *et al.* (1999). A recent study by Thliveris *et al.* (2005) overcomes

biological limitations of earlier studies and provides another example of data relevant to the polyclonal fraction. Indeed, the question of polyclonal tumor origin is expected to recur as lineage-marker studies improve. A statistical analysis of methods for estimating the polyclonal fraction therefore seems to be warranted, especially considering a problem with the Novelli's ratio.

2 Novelli's ratio is not a valid bound

Lineage-marker studies of polyclonality offer two binary classifications of tumors in a tumor population: (1) clonality, i.e. whether the tumor is monoclonal or polyclonal, and (2) phenotype, i.e. whether the tumor is homotypic or heterotypic. 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. Logically no tumors can be both heterotypic and monoclonal (assuming integrity of the lineage marker). Note that the covertly polyclonal tumors are those satisfying $\mathrm{POL} \cap \mathrm{HOM}$ and the overtly polyclonal tumors satisfy $\mathrm{POL} \cap \mathrm{HET}$.

Table 1: Cross classification of a tumor population in terms of clonality (not directly observed) and phenotype (observable):

	clonality		
	monoclonal [MON]	polyclonal [POL]	
homotypic [HOM]	$P(\mathrm{HOM} \cap \mathrm{MON})$	$P(\mathrm{HOM} \cap \mathrm{POL})$	$P(\mathrm{HOM})$
heterotypic [HET]	0	$P(\mathrm{HET} \cap \mathrm{POL})$	$P(\mathrm{HET})$
	$P(\mathrm{MON})$	$\theta = P(\mathrm{POL})$	100%

The inference problem is to estimate or somehow bound $\theta = P(\mathrm{POL})$ using counts of homotypic and heterotypic tumors. The obvious bound $P(\mathrm{HET}) \leq \theta$ is valid, but it is not tight if a significant fraction of the homotypic tumors are polyclonal. For instance, $P(\mathrm{HET})$ was estimated to be 5% in the Novelli *et al.* study, but those authors estimated θ to exceed 76%.

In a system where each clone has one of two possible types, HOM is the union of tumors of the minority type HOM_1 and the majority type HOM_2. For instance XO was

the minority type in the Novelli *et al.* study. That paper proposed

$$\hat{B} = \frac{\#\{\text{HET}\}}{\#\{\text{HET}\} + \#\{\text{HOM}_1\}} \tag{1}$$

as a lower bound for θ. Here $\hat{B}$ estimates the population quantity

$$B = P(\text{HET})/P(\text{HET} \cup \text{HOM}_1). \tag{2}$$

A clear rationale for the claim that $B \leq \theta$ was not provided in Novelli *et al.*, but the idea may have been simply this: among the heterotypic and minority-homotypic tumors, the polyclonal fraction is

$$P(\text{POL} \,|\, \text{HET} \cup \text{HOM}_1) = \frac{P(\text{HET} \cap \text{POL}) + P(\text{HOM}_1 \cap \text{POL})}{P(\text{HET} \cup \text{HOM}_1)}. \tag{3}$$

According to Table 1, the first term in the numerator is $P(\text{HET})$. The second term, furthermore, 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 the second term in the numerator of (3), it is true that

$$B \leq P(\text{POL} \,|\, \text{HET} \cup \text{HOM}_1)\ . \tag{4}$$

Thus Novelli's ratio (2) does bound a certain polyclonal fraction, but it is not θ, the marginal polyclonal fraction of interest; rather B is a lower bound on the rate 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 the case. In fact, in the population of heterotypic and minority-homotypic tumors, polyclonality is more frequent than in the whole population of tumors (Theorem 1 below). There is a positive gap between the polyclonal fraction and the quantity being bounded below by Novelli's ratio B. This creates a problem, for if B lies in this gap, then it is not a lower bound for the marginal polyclonal fraction. Section 3 gives scenarios where the bound holds, and, more importantly, realistic scenarios where it does not. Section 4 describes an alternative statistical approach to estimate the polyclonal fraction.

To reiterate, the difficulty with Novelli's ratio B stems from a positive gap between θ and the larger quantity $P\,(\text{POL} \,|\, \text{HET} \cup \text{HOM}_1)$; if B lands in the gap, then it exceeds θ. Further, whether or not B lands in the gap depends on details of the tumor initiation process, and, therefore, B can not be a general purpose lower bound. The gap 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 the nature of polyclonality (e.g., the number of clones involved) nor the means by which clones are marked. However we do require a weak technical assumption about the stochastic process governing clonal marking which essentially ensures that heterotypic tumors are possible. Consider that in a population of tumors comprised of monoclonal tumors and, for various $k \geq 2$, tumors originating by the interaction of k clones, we have an overall proportion γ_t of clones that are of type t.

Definition: The clone marking process is *regular* if for each clonal type t, $0 < \gamma_t < 1$ and if for each $k \geq 2$,

$$P(\text{all } k \text{ clones have type } t \,|\text{POL}_k) < P(\text{single clone has type } t \,|\text{MON}) = \gamma_t.$$

POL_k is the event that the tumor is formed from exactly k clones, and MON is the event that the tumor is monoclonal.

The first condition implies, among other things, that there are at least two clonal types. A non-regular marking process would be one, for instance, in which the types of all clones in a polyclonal tumor are fully determined by the type of one of those clones. This might seem plausible for tumors forming in a large mono-typic region of the intestine, but the condition refers to the entire population of tumors, and so such determinism would eliminate the possibility of heterotypic tumors. The assumption of regular clonal marking includes a range of plausible stochastic processes, such as those in which the marking is neutral and thus independent, in a certain sense, from the process by which the tumors become monoclonal or polyclonal. We take up the point at the end of this section. First let us establish the positive gap.

Theorem 1. *Suppose that* $0 < \theta < 1$ *and further that the stochastic process which marks clones is regular. Then* $\theta < P\left(\text{POL} \,|\, \text{HET} \cup \text{HOM}_1\right)$.

Proof of Theorem 1: We prove something slightly more general than is stated. Let t denote any one of the clonal types, and consider the event

$$T = \{\text{tumor is homotypic of type } t\},$$

which has probability $0 < P(T) < 1$ by regularity. Observe that the polyclonal fraction θ is a weighted average

$$\theta = P(\text{POL}|T)\, P(T) + P(\text{POL}|T^c)\, P(T^c),$$

where T^c is the compliment of T. In a two-type system where t is the majority type, $T^c = \{\text{HET} \cup \text{HOM}_1\}$, for example. To establish the positive gap it is sufficient to prove $P(\text{POL}|T) < \theta$. By Bayes's rule, this is equivalent to $P(T) > P(T|\text{POL})$. Now the marginal $P(T)$ is decomposed into non-zero terms according to clonality:

$$\begin{aligned} P(T) &= P(T|\text{POL})\,\theta + P(T|\text{MON})\,(1-\theta) \\ &= a\theta + \gamma_t(1-\theta) \end{aligned}$$

where $\gamma_t \in (0,1)$ is the marginal rate of type t clones and $a = P(T|\text{POL})$ has to do with the stochastic process by which multiple clones are marked in a polyclonal tumor. Thus the difference

$$P(T) - P(T|\text{POL}) = (\gamma_t - a)(1-\theta).$$

We have assumed $\theta < 1$, and $\gamma_t \in (0,1)$ in the statement of the theorem, so the theorem is true if $a < \gamma_t$. Polyclonality itself may be expressed as a union over the number $k \geq 2$ of contributing clones. Given k, some joint distribution over clone marking governs whether or not all the clones are of type t. With suggestive notation,

$$\begin{aligned} a &= P(T|\text{POL}) \\ &= \sum_{k=2}^{\infty} P(T|\text{POL}_k)\, q_k/\theta \\ &= \sum_{k=2}^{\infty} P(\text{all } k \text{ clones have type } t\,|\text{POL}_k)\, q_k/\theta \\ &< \gamma_t \end{aligned}$$

where $q_k = P(\text{POL}_k)$ is the fraction of tumors comprised of k clones where $k \geq 2$. (Note $\theta = \sum_{k=2}^{\infty} q_k$.) The last inequality follows from the definition of a regular marking process. □

We end this section by returning to the assumption that the clonal marking process is regular. The value of lineage-marker studies derives in part from the possibility that the markers are neutral in the sense that the marking itself does not alter the polyclonal structure. To be specific, let $N_k(t)$ denote the number of clones of type t that are bound together in a tumor, which itself is considered to be sampled randomly from the population of tumors formed from $k \geq 1$ clones. These are count random variables, and $k = \sum_t N_k(t)$. The tumor is homotypic of type t if $N_k(t) = k$, for example. One definition of neutral marking is to insist that the expected proportion of type t clones in clonality-k tumors does not depend on k; i.e. $E\{N_k(t)/k\} = \gamma_t$. It follows by routine algebra that if the marking

process is neutral and if it allows heterotypic tumors (i.e., if $\sum_t P(N_k(t) = k) < 1$), then the marking process is regular.

An elementary 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 $P(N_k(t) = k)$ above the independence homotypic rate γ_t^k.

3 Random marking in monoclonal and biclonal tumors

Taking the concepts from the last section to a concrete example, suppose that tumors are monoclonal with probability $1-\theta$ or are formed from two clones, and thus are biclonal with probability θ. 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), the proportion of biclonal tumors among the heterotypic or pure type-1 tumors is

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

As promised by (4), Novelli's ratio B does provide a lower bound for a certain conditional polyclonal fraction. However there is a gap between that conditional fraction and the smaller marginal polyclonal fraction θ of interest, and so B can fail. Figure 1 charts the difference $\Delta = \theta - B$ 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.

4 Model-based estimation of the polyclonal fraction

What statistical recourse is there for inference about θ? It seems that the weak lower bound $P(\text{HET}) \leq \theta$ is the best one can do without adopting modeling assumptions on the nature of polyclonality. Mathematically, for example, it is possible that tumors are either monoclonal or polyclonal of some large degree k, and are marked by some simple marking scheme. If this were the case, virtually all the polyclonal tumors would be heterotypic, and so the the simple lower bound $P(\text{HET}) \leq \theta$ would be tight. Biological intuition suggests that we reject this extreme model of polyclonality for something more sensible. Three elementary models of clonality are:

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

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

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

   for $k = 1, 2, \ldots$ 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. Here, the polyclonal fraction is $\theta = 1 - \lambda / \{\exp(\lambda) - 1\}$.

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

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

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

Likelihood-based inference for θ is possible if we also invoke a model for clone marking. 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. It could be computationally challenging to incorporate detailed information about positive association.

Maximum likelihood estimation has some validity even in the absence of random marking. We argue at the end of this section that the maximum likelihood estimate obtained under the random 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.

To compute the marginal probability of a homotypic tumor of type t, we must sum over the unknown clonality. For the three clonality models presented above, and with random marking, these sums can be solved explicitly.

1. **Monoclonal/Biclonal**: $P(\text{homotypic tumor of type } t) = (1-\theta)\gamma_t + \theta\gamma_t^2$
2. **Conditional Poisson**: $P(\text{homotypic tumor of type } t) = \frac{\exp(\lambda\gamma_t)-1}{\exp(\lambda)-1}$
3. **Geometric**: $P(\text{homotypic tumor of type } t) = \frac{\gamma_t\psi}{1-\gamma_t(1-\psi)}$

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 (standard errors estimated from observed information were relatively small and are not shown; model-specific bias may be larger).

Table 2: Estimation of polyclonal fraction.

	tumor counts			ML estimate of θ			Novelli's ratio
	pure $t=1$	pure $t=2$	mixed	biclonal	poisson	geometric	$\hat{B}$
Novelli data	4	246	13	0.64	0.55	0.53	0.76
Merritt.112	5	93	7	0.46	0.40	0.37	0.58
Merritt.113	1	139	15	0.94	0.83	0.77	0.94

Maximum likelihood estimates obtained by independent marking as described above 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, this bias is expected to be conservative, (i.e., the estimates ought to be low) since the 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 random marking, a tumor will be homotypic type t with probability $\alpha_t(\theta) = (1-\theta)\gamma_t + \sum_{k\geq 2} q_k \gamma_t^k$. Positive association of clonal marking amounts to an increased homotypic rate $\beta_t(\theta) \geq \alpha_t(\theta)$. The rate of heterotypic tumors under random 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 random-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 random-marking is incorrect, the random-marking estimate of ψ will be consistent for this population heterotypic frequency; but in fitting closely to the data, an incorrect value $\theta^* = a^{-1}(\psi) \neq \theta$ 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 θ^* to which the random-marking estimator converges must be no greater than the true polyclonal fraction θ.

References

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.

Dove, W. F., Cormier, R. T., Gould, K. A., Halberg, R. B., Merritt, A. J., Newton, M. A., and Shoemaker, A. R. (1998). The intestinal epithelium and its neoplasms: genetic, cellular, and tissue interactions. *Phil. Trans. R. Soc. Lond. B*, **353**, 915–923.

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.

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.

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.

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**, In press.

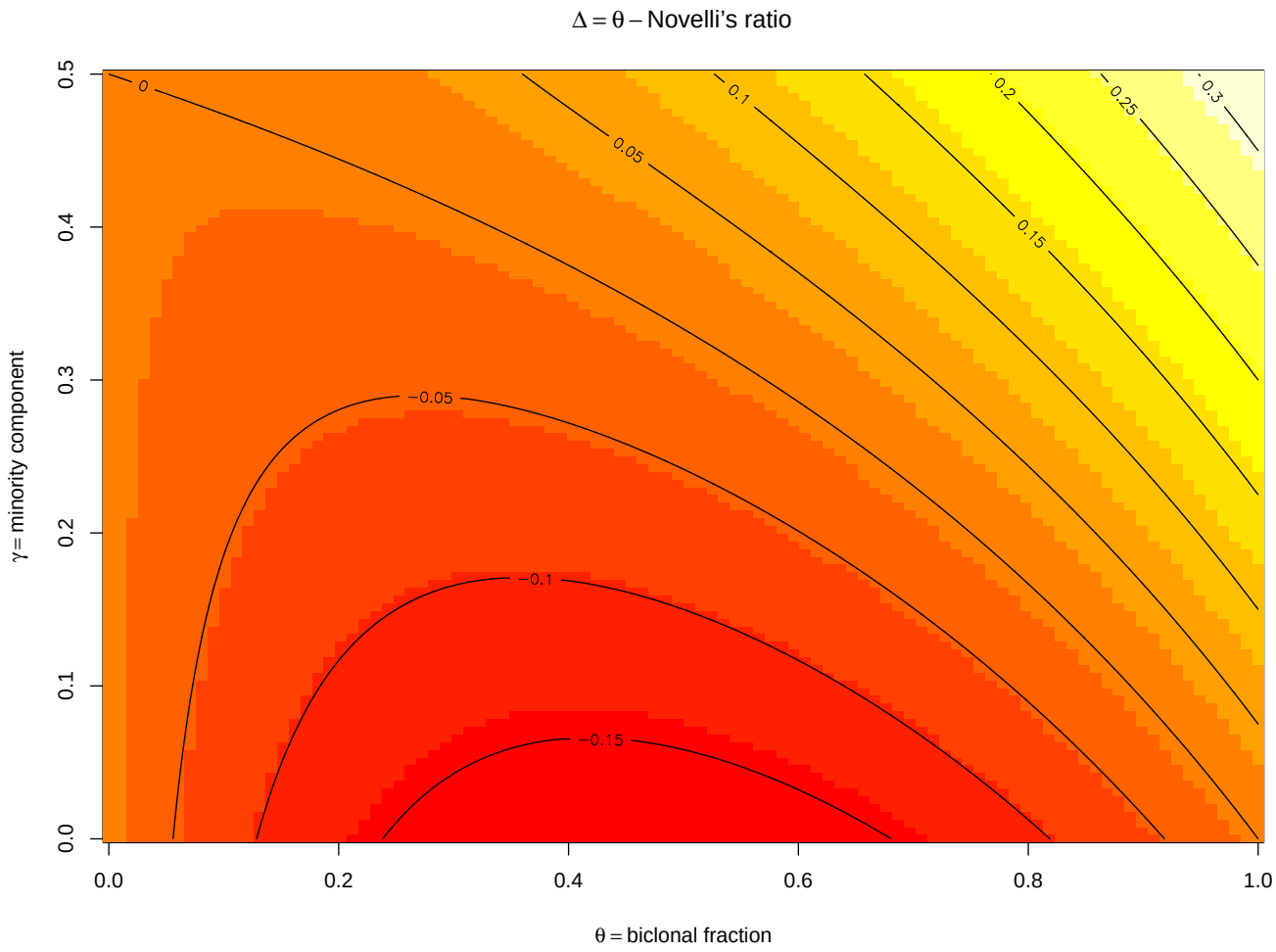

Figure 1: Discrepancy between polyclonal fraction and Novelli's ratio in the monoclonal/biclonal, random-marking model as function of the minority fraction γ_1 and the biclonal fraction θ. In the lower left of the plot Novelli's ratio fails to bound the polyclonal fraction.