“Testing clonal relatedness of two tumors from the same patient based on their mutational profiles: update of the *Clonality* R Package”

Audrey Mauguen, Venkatraman E. Seshan, Colin B. Begg, Irina Ostrovnaya

Supplementary Materials

The main body of the article demonstrates how to execute the new software in our *Clonality* R Package. In the following we provide summaries of the rationale and technical detail of the two new methods that have been added to the package. Both methods concern the evaluation of clonal relatedness from anatomically distinct tumors in individual patients, where the tumors have been sequenced to identify somatic mutations. Methods are suitable for any amount of genotyping, e.g. from tumors in which only a single hotspot or gene is genotyped, to sequencing of a panel of selected genes, to exome or whole genome sequencing. The first method is a statistical significance test that is designed to be applied to an individual case, testing the null hypothesis that the tumors arose independently. Full details of this method can be found in Ostrovnaya et al (2015). The second method permits a combined analysis of a set of patients, each with two tumors that could be independent or clonal. This method uses empirical Bayes methods to share information and provides diagnostic probabilities of clonal relatedness for each case. Details of this method are available in Mauguen et al (2018).

The statistical test referenced in Section 2.3 of the main paper is constructed as follows. Each tumor will have a set of somatic mutations identified. These mutations can be classified as either shared (present in both tumors, and thus potentially “clonal” mutations that occurred in the originating clonal cell) or private (present in only one of the tumors). Although in this section we are considering only a test for a single case we nonetheless distinguish individual cases by the subscript j as we will be discussing global analyses of multiple cases in the next section. Let $\mathrm{A_j}$ denote the set of observed mutations in the j^th case that are shared and let $\mathrm{B_j}$ denote the set of private mutations. Let $\mathrm{G_j = A_j \cup B_j}$. Our data can thus be considered as $\mathrm{Y_j = (A_j, B_j)}$. Let $\mathrm{p_i}$ represent the marginal probability (assumed known) of a mutation at the i^th locus. Note that our strategy for estimating these probabilities is described in Section 2.2 of the main text. Our methods rely on a case-specific parameter, the clonality signal ξ_j. This represents, in the context of the evolution of the tumors, the relative duration of the period in which the original clonal cell accumulated mutations, prior to the period where the two tumors evolved separately and accrued additional independent mutations. Thus ξ_j essentially represents the probability that an observed mutation occurred during the clonal phase as opposed to the independent phase of tumor development. For independent tumors $\xi_j = 0$. It follows that for a case with a given clonality signal the probabilities of observing shared and private mutations at each locus are given by:

$$\begin{cases} P\left(i \in A_j \middle| \xi_j\right) = \xi_j p_i + \left(1 - \xi_j\right) p_i^2 \\ P\left(i \in B_j \middle| \xi_j\right) = 2\left(1 - \xi_j\right) p_i (1 - p_i) \\ P\left(i \in G_j \middle| \xi_j\right) = \xi_j p_i + \left(1 - \xi_j\right) p_i (2 - p_i) \end{cases}$$

The likelihood (conditioned on the observed mutations) for an individual case is thus given by:

$$\mathrm{P(Y_j \mid \xi_j) = \prod_{i \in A_j} \left\{ \frac{\xi_j + (1-\xi_j) p_i}{\xi_j + (1-\xi_j)(2-p_i)} \right\} \prod_{i \in B_j} \left\{ \frac{2(1-\xi_j)(1-p_i)}{\xi_j + (1-\xi_j)(2-p_i)} \right\}}. \qquad (1)$$

Consequently, a conditional likelihood ratio test statistic can be expressed as:

$$\mathrm{S = \sum_{i \in A_j} \log\left[\frac{\hat{\xi}_j}{1-\hat{\xi}_j} p_i^{-1} + 1 \right] - \sum_{i \in G_j} \log\left[\frac{\hat{\xi}_j}{1-\hat{\xi}_j} (2-p_i) + 1 \right]}.$$

To obtain a reference distribution we generate the distribution of S under the assumption that matches occur randomly via independent sampling and that at least one mutation is observed at each locus in the set $\mathrm{G_j}$.

Using simulations, we were able to show in Ostrovnaya et al (2015) that the test is valid and that its properties are very similar to the corresponding unconditional likelihood ratio test, demonstrating that the observed mutations contain essentially all of the available information relevant for testing the hypothesis, i.e. we can ignore loci at which no mutations occur on either tumor. We note that the use of the conditional test makes this a very flexible procedure, since the method allows a test of the hypothesis if only one locus is observed but can

also handle exome or even whole genome data, since the numbers of observed mutations in these settings will generally be relatively modest.
We acknowledge that there have been methods proposed by others for the purpose of clonality testing of mutational data. In very early work that is frequently cited, Teixera et al. (2004) proposed an index of clonal relatedness (ICR) to gauge the evidence for clonality. This index is defined as $\mathrm{ICR} = 1 - \prod_{i \in A_j} p_i$. Teixera et al. proposed using this as evidence of clonality if ICR>0.95. It can be shown that this is actually a likelihood ratio statistic (conditioned on A_j) of the hypothesis that the tumors are independent ($\xi_j = 0$) versus the restricted clonal alternative that the tumors are identical ($\xi_j = 1$). However, in our extensive experience analyzing data from breast and lung cancers this alternative is not credible generally since most clonal tumor pairs have numerous private mutations. In other words this test does not take account of the evidence for independence in any non-matches that are observed. The information in the non-matching mutations is clearly important for assigning diagnostic probabilities (see the following section). A more recent proposal is from Bao et al (2015) who advocate using a simple binomial distribution to represent the observed matches among the loci with an observed mutation in either tumor. That is, their test statistic is $X = \sum_{i \in G_j} I(i \in A_j)$, and this is assumed to be generated from a binomial distribution with $N = \sum_{i \in G_j} I(i \in G_j)$ trials with a pre-defined null value. This approach fails to account for the fact that the probability of a match at loci i depends on p_i although this kind of strategy has been used by others (Harms et al 2017). To clarify the loss of power from this assumption we have conducted simulations to examine the properties of this test in comparison to our proposed test using the framework of a study we published concerning the clonal relatedness of lobular carcinomas in situ (LCIS) with invasive breast cancers where the data were generated from exome sequencing (Begg et al. 2016). That is, we generated mutations based on marginal probabilities of the mutations in the breast study such that the average number of mutations per case was 34. The Bao et al. test critical value was calibrated to be similar to that of our proposed clonality test. The results are in the top panel of the Table. In the lower panel we present results based on the features of a study of bilateral breast cancer that employed panel sequencing of the major breast cancer genes, resulting in a mean number of mutations per case of 10 (Begg et al 2018). In both settings the Bao et al. (2015) test is shown to be relatively inefficient, especially when ξ is relatively small and there are few (typically one) matches. The likelihood ratio approach makes much more efficient use of the matches by taking account of the marginal probabilities, while the Bao et al. approach does not distinguish matches at common loci from those at rare loci.

Table: Comparison of Clonality Tests

Mean # Mutations	Test	Size	Power	
		$\xi = 0$	$\xi = 0.1$	$\xi = 0.25$
34	Ostrovnaya	4.0%	86%	99%
	Bao	4.1%	75%	98%
10	Ostrovnaya	3.4%	32%	66%
	Bao	3.3%	19%	53%

The great advantage of our likelihood ratio test is that it can be applied in clinical settings without the need of a reference dataset, i.e. all that is needed are the marginal mutation probabilities. However, it only provides a partial solution to the problem at hand, in that it provides evidence against the null hypothesis of independence but does not speak convincingly to the evidence in favor of independence. This is because the p-value is always 1 if there are no matches when in fact if relatively few genes are sequenced we may simply not be observing clonal matches that exist elsewhere in the genome. Intuition suggests that tumor pairs with many non-matching mutations are more likely to be independent than those with few non-matching mutations. This issue led us to pursue a random effects modeling approach that makes more efficient use of all of the available data and permits the estimation of diagnostic probabilities for each case (outlined in Section 2.4 of the main paper). We frame the problem in the context of the study of bilateral breast cancer cited above that we recently conducted in which we obtained tissue samples from both breasts of women who had experienced bilateral breast cancer and where both surgeries were performed at our center (Begg et al 2018). These tissues were subjected to sequencing using a special panel of 254 genes known to be important in breast cancer. A total of 49 cases (i.e. tumor pairs) made it through the sequencing pipeline and were analyzed. We are interested primarily in two issues: (1) What proportion of cases are clonal? (2) For each case, what are the diagnostic probabilities of clonal relatedness (i.e. one tumor is a metastasis of the other one) versus independence (both tumors are primary cancers that developed independently)? We define π to be the proportion of clonal cases in the population, i.e. the proportion of cases for which $\xi_j > 0$. We also denote by C_j the event {case j is clonal} and by $\bar{C}_j$ its complement. We consider the clonality signal ξ_j to be a random effect with probability density $g(\xi_j)$, where $\xi_j = 0$ with probability $1-\pi$ and, with probability π, $\phi_j = -\log(1-\xi_j)$ follows a lognormal distribution with parameters μ and σ^2. The corresponding density of the clonality signal is thus zero-inflated but has a flexible structure for modeling the positive random effects in the range $0 < \xi_j < 1$. Our goal is to maximize the likelihood:

$$L = \prod_{j=1}^{n} P\left(Y_j \mid \pi, \mu, \sigma\right), \tag{2}$$

where n is the number of cases. To perform this estimation we first recognize that

$$\begin{aligned} P\left(Y_j, \xi_j, C_j \mid \pi, \mu, \sigma\right) &= P\left(Y_j \mid \xi_j, C_j\right) \times g\left(\xi_j, C_j \mid \pi, \mu, \sigma\right) \\ &= g\left(\xi_j, C_j \mid Y_j, \pi, \mu, \sigma\right) \times P\left(Y_j \mid \pi, \mu, \sigma\right), \end{aligned} \tag{3}$$

where $g(\bullet)$ is a generic density function. Note that the likelihood contribution of case j to (2) is a component of the second term in (3). The EM algorithm permits us to maximize (iteratively) the expectation of the logarithm of the likelihood averaged over the latent variables conditioned on the data. That is, the expected likelihood is given by

$$E = \prod_{j=1}^{n} \int_0^1 \log\left\{P\left(Y_j, \xi_j, C_j \mid \pi, \mu, \sigma\right)\right\} g\left(\xi_j, C_j \mid Y_j, \tilde{\pi}, \tilde{\mu}, \tilde{\sigma}\right) d\xi_j dC_j,$$

where $\tilde{\pi}, \tilde{\mu}$ and $\tilde{\sigma}$ are the current estimates of the parameters. After choosing starting values for these parameters the expectation and maximization steps proceed iteratively until convergence. To calculate E we recognize that $P\left(Y_j, \xi_j, C_j \mid \tilde{\pi}, \tilde{\mu}, \tilde{\sigma}\right)$ is obtained easily from the defined terms on the right-hand side of (3), represented by (1) and the normal parametric model used for the distribution of ξ_j. Further, $g(\xi_j, C_j \mid Y_j, \tilde{\pi}, \tilde{\mu}, \tilde{\sigma})$ can be obtained from Bayes Theorem, i.e.

$$g\left(\xi_j, C_j \mid Y_j, \tilde{\pi}, \tilde{\mu}, \tilde{\sigma}\right) = \frac{g\left(\xi_j, C_j \mid \tilde{\pi}, \tilde{\mu}, \tilde{\sigma}\right) P\left(Y_j \mid \xi_j, C_j\right)}{\int_0^1 g\left(\xi_j, C_j \mid \tilde{\pi}, \tilde{\mu}, \tilde{\sigma}\right) P\left(Y_j \mid \xi_j, C_j\right) d\xi_j dC_j}.$$

Simulations to evaluate bias in the estimates of the model parameters based on this method demonstrate that to estimate the parameters with low bias the dataset needs to be sufficiently large and data rich, i.e. there needs to be sufficient numbers of both clonal and independent cases.

References

Bao L, Messer K, Schwab R, Harismendy O, Pu M, Crian B, Yost S, Frazer KA, Rana B, Hasteh F. Mutational profiling can establish clonal or independent origin in synchronous bilateral breast and other tumors. PLoS One 2015;10(11):e0142487.

Begg CB, Ostrovnaya I, Carniello JVS, Sakr RA, Giri D, Towers R, Schizas M, De Brot M, Andrade VP, Mauguen A, Seshan VE, King TA. Clonal relationships between lobular carcinoma in situ and other breast malignancies. Breast Cancer Research 2016; 18:66. doi: 10.1186/s13058-016-0727-z.

Begg CB, Ostrovnaya I, Geyer FC, Papanastasiou AD, Ng CKY, Sakr RA, Bernstein JL, Burke KA, King TA, Piscuoglio S, Mauguen A, Orlow I, Weigelt B, Seshan VE, Morrow M, Reis-Filho JS. Contralateral breast cancers: Independent cancers or metastases? Int J Cancer. 2018;142(2):347-356.

Girard N, Ostrovnaya I, Lau C, Park B, Ladanyi M, Finley D, Deshpande C, Rusch V, Orlow I, Travis WD, Pao W, Begg CB. Genomic and mutational profiling to assess clonal relationships between multiple non-small cell lung cancers. Clinical cancer research : an official journal of the American Association for Cancer Research 2009;15(16):5184-90.

Harms KL, de la Vega LL, Hovelson DH, Rahrig S, Cani AK, Liu C-J, et al. Molecular profiling of multiple primary merkel cell carcinoma to distinguish genetically distinct tumors from clonally related metastases. JAMA Dermatology 2017;153:505-512.

Mauguen A, Seshan VE, Ostrovnaya I, Begg CB. Estimating the probability of clonal relatedness of pairs of tumors in cancer patients. Biometrics. 2018;74:321-330.

Ostrovnaya I, Seshan VE, Begg CB. Using somatic mutation data to test tumors for clonal relatedness. Annals of Applied Statistics 2015;9:1533-48.

Teixeira MR, Ribeiro FR, Torres L, Pandis N, Andersen JA, Lothe RA, Heim S. Assessment of clonalrelationships in ipsilateral and bilateral multiple breast carcinomas by comparative genomic hybridisation and hierarchical clustering analysis. British Journal of Cancer 2004;91(4):775-82. PMCID: PMC2364777.