Reviewer's report

Title: CDCOCA: A statistical method to define complexity dependence of co-occurring chromosomal aberrations

Version: 1 Date: 20 October 2010

Reviewer: Nancy Zhang

Reviewer's report:

- Major Compulsory Revisions

1. The main idea is that “sample heterogeneity” is introduced into the p-value computation. But you never compared to any existing method that does not consider this factor. The key question is: does introducing sample heterogeneity in the calculations improve things? There is no evidence in the paper to this effect. As a simple study, you can simply ignore the adjustment step in their algorithm, and show that the results are different and less desirable.

2. You evaluated the method on some real data sets, and as “proof that the method works”, you show that the regions found by the approach contain significant cancer genes. But this is an artifact of the data you started with! The original set of regions were already selected for those that appear in many samples, and this fact would make any region in that original set enriched for cancer genes. A simulation study, or some more carefully designed biological validation, comparing the new method to previous methods, would be more convincing.

3. How were the original regions selected? More details would be informative. How sensitive is the method to the initial segmentation?

4. How do you adjust for multiple testing in selecting the region-pairs? In your example you have 300+ regions, that would be 90000 tests! Furthermore the issue is complicated with the fact that the tests are highly correlated! How did you get the FDR on line -12, page 6?

- Minor Essential Revisions

1. P.3, line -3: “dimentions” should be “dimension”.

2. P.3, line -2: “D_{i,j} = 1” rather than “D=1”.

3. P.4, line 7: Maybe a term other than “overlay” is better, usually “overlap” between regions carries the connotation of physical overlap of these regions.

4. P.4, line 18-19: These two lines seem to be repeats of what appeared a few lines earlier.

5. P.4, line -1: “r=sample(N,sample=1,...)” This is R programming language. Please avoid using programming-language-specific notation in describing your algorithm.

6. Figures 2,3 are confusing: is every color for a *pair* of regions? Only one
region is listed in the legend.

**Quality of written English:** Not suitable for publication unless extensively edited

**Statistical review:** Yes, and I have assessed the statistics in my report.