Reviewer’s report

Title: Low-penetrance alleles predisposing to sporadic colorectal cancers: a French case-controlled study

Version: 1 Date: 16 May 2008

Reviewer: Victor Moreno

Reviewer’s report:

This study tries to replicate in a relatively large population some associations on genetic polymorphisms previously reported. This is a review of the statistical aspects.

1) Sample size. It is large, but not enough to reach conclusions that could be considered definitive. This study just adds data that will be compiled in future meta-analysis to solve inconsistencies among studies.

2) Populations. It is not clear if the analysis done in subjects from Vendée are a different sample or are included in the 2144.

3) The analysis of main effects. Since the design of the study is an individually matched case-control study, the authors should comment if a conditional logistic regression analysis restricted to matched pairs modify their results.

Some correction for multiple testing should be done at this stage. A Bonferroni approach would mean multiplying all p-values by 50 (or compare with p<0.001), with no significant results due to lack of power for multiple testing. I would recommend an estimation of false discovery rate and report the proportion of the 5 significant results that are probably false positives.

4) Analysis of haplotypes. Since the polymorphisms selected are in different genes, this analysis is useless. Haplotypes make sense when there is LD among markers, then haplotypes might capture better where is the risk. If there is no LD, then the algorithm will result in a haplotype for each combination of SNPs, with frequency the product of allele frequencies of each marker. I would delete this analysis.

5. Analysis of combined SNPs. The results obtained are anticipated by the analysis strategy since there is an exhaustive search of all possible combinations and the same population used to define the search set is used. This is the results of over-fitting.

The authors correct p-values dividing by N, the number of markers used, but they should divide by the number of independent tests performed. This number is very difficult to know because the same markers go into different models in the search, but is clearly greater than 5. A very conservative approach would be to divide by the number of tests performed, a figure that is not easy to know from
the description of the methods but is larger than N. After proper correction, probably the p-value will be non-significant. It is surprising that no other combinations were significant. I would expect a few more just by chance and finding only one that combines all significant results in main effects denotes the weakness of this profile.

Significance, however, is not as important as validity. Validity of the results could be assessed testing this profile in an independent population. If this is not possible, the analysis of internal consistency gives some clues. The authors should make clearer that this analysis, shown in supplementary table 2 points that the consistency of the risk profile is very weak, with many subgroups were significance is lost. I would downplay the significance of this profile.

6) Population stratification. Although the population is theoretically homogeneous Caucasian, the structure software identifies 3 populations, possibly more. These should be used as confounder factors in the analysis. An alternative is an adjustment by genomic control or by the principal components that identity clear groups in a PCA analysis. Some of the associations could be due to this problem, possibly the odd result for GSTM1 disappears if structure is accounted for.

7) Conclusion. What means "in a sub-group of the French population"?

Level of interest: An article of importance in its field

Quality of written English: Acceptable

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

Declaration of competing interests:

I declare that I have no competing interests