Reviewer's report

**Title:** Exploring multilocus associations of inflammation genes and colorectal cancer risk using hapConstructor

**Version:** 1  **Date:** 8 June 2010

**Reviewer:** Bob O'Hara

**Reviewer's report:**

I have been asked to comment on the statistics, so I will concentrate on that (and make a couple of other comments at the end).

The manuscript uses hapConstructor to select SNPs that are correlated with colon cancer. It then uses these SNPs in a logistic regression, and compares the results. This is little different to what the original hapConstructor paper did, albeit with a much smaller data set.

One problem is that it is not clear what hapConstructor is actually doing: this should be made clearer, as should the difference between the hapConstructor and logistic regression analysis. As far as I can see, the models are similar, the differences are that hapConstructor imputes missing data, and no model selection is done with the logistic regression (this used to be done through PROC STEPWISE, but that might have changed now).

The model selection is curious - the authors change the inclusion criteria depending on the model complexity. However, the only reason given is "due to multiple comparisons". I have never seen this done in model selection, so I'm extremely dubious about this. The purpose of model selection is to find an optimal model, according to some criterion. The reason for using a stepwise selection procedure is that exhaustively searching the model space would often take too long (in this case, SAS should be able to do it in a reasonable amount of time: there are only 16 loci, so 2^16 possible models). Changing the criterion as the model selection proceeds is bizarre. If the authors are to do this, they must provide a stronger motivation for this: they have to show mathematically what they are doing and why, and show why the change in entry criterion is justified, and what the value should be (why use 0.05, 0.001, and 0.0001? these values look as if they have been pulled from thin air).

The authors also need to provide more references, e.g. in the last paragraph of p9, they mention several procedures and statistics, but one has to guess at what exactly was done.

The logistic regression includes several other covariates, all of which apparently have an effect (I can't find their effects reported, though). The model is thus different to that tested by hapConstructor. This limits any comparison: one might expect differences simply due to this. All that could be said, then, might be that hapConstructor is unable to fit the model one wants. It is difficult to see what
hapConstructor can do that can't be done more efficiently in a statistics package (even SAS, although I suspect R would be able to do much more. But I am an R junkie).

The manuscript claims to evaluate hapConstructor, but it is not clear what this really means. The authors use it, and show that it gives some numbers, and these are similar to those obtained by logistic regression. But this tells us little about how hapConstructor actually performs. The timings aren't impressive - SAS should be able to search through all possible models with all 3 genes in much less than a day (there are $2^{16}=65536$ models, which at 1s per model would take just over 18 hours).

The writing needs improvement. Whilst it is grammatically correct, it is difficult to read: some sentences are unnecessarily convoluted and the meaning slips away from them. The impression is that nobody has read the paper for meaning, e.g. the first sentence of the third paragraph of the Statistical Methods section (p9) reads (simplifying the nounal phases) "Using data, statistics were performed". I would suggest the authors try to write in a more direct style, and simplify their sentences.

**Level of interest:** An article of limited interest

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

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

**Declaration of competing interests:**

I declare that I have no competing interests