Reviewer's report

Title: Comprehensive profiling of DNA methylation in colorectal cancer reveals three subgroups with distinct clinicopathological and molecular features

Version: 1 Date: 31 December 2009

Reviewer: Karen Curtin

Reviewer's report:

General comments: The background section is concise, well-written (as is the manuscript as a whole), and comprehensively references the previous work of other research groups that have studied CIMP and methylation markers in CRC samples. The methods are generally appropriate and well-described, however more statistical detail and discussion is warranted given that in the data analysis, there were a large number of variables (1505 CpG sites) and only a small number of samples (91 CRCs total); see comments below. Given the small number of samples, the authors should either provide further support for their assertion of a CIMP-M phenotype (one could argue it is not clearly distinct from CIMP-L given Table 1 and limitations of the ad hoc clustering analyses), or acknowledge that there is a lack of compelling evidence that CIMP-M may be a distinct phenotype pending further study with larger sample sizes. In their discussion, the authors could acknowledge any work upon which they are currently building; for example, describe how they foresee the utility of the Illumina methylation array technology in determining CIMP; what are potential diagnostic and clinical implications, given its additional expense compared to conventional methylation panels? Overall, the manuscript adds to the literature in the field, and should be published pending revision.

Major Compulsory Revisions

1. In the methods (statistical analysis) or discussion, provide further justification for use of unsupervised hierarchical clustering for their analysis (references from other groups using the Illumina methylation platform and similar analysis methods; results of the authors’ assessment of the robustness of their clustering using non-parametric bootstrap resampling). Why were only the results for unsupervised hierarchical clustering analyses reported, given that supervised hierarchical clustering analyses were performed as mentioned by the authors?

2. The authors need to acknowledge in the discussion that the three “distinct clusters” they observed may have occurred by chance, given the large number of comparisons and their small CRC sample size. In the results or in additional file 2, they should provide nominal or their Bonferroni-corrected p-values (as the authors allude to in the methods) to support their assertion of three distinct clusters. To this reviewer, a visual inspection of additional file 2 indicates two clearly distinct clusters with a third rather “iffy.”

3. Given the high correspondence between the authors' CIMP-H and CIMP-high
In their samples and agreement of CIMP-high and clinicopathological features in other previous studies, the authors should address in their discussion: is the additional expense and complexity of using the Illumina methylation platform potentially beneficial in some regard, given their results appear in large measure to validate the results of smaller methylation panels?

Minor Essential Revisions

1. Methods (statistical analysis), first paragraph, first sentence: “allele” should be plural, as the beta-value is a ratio based on fluorescent signals from methylated and unmethylated alleles at each data point.

2. Discussion (p. 13, last paragraph, 3rd sentence should read “..70% of these just one CpG site per gene was evaluated.” rather than “sites.”

Discretionary Revisions

1. Title may be overly explicit; perhaps remove “three.”

2. In the Conclusion paragraph of the abstract, final sentence: revise to “This study provides further confirmation that both KRAS and BRAF mutations are involved with the CIMP-H pathway of CRC.” Previous research groups have suggested and confirmed this as referenced by the authors in the text; thus it is not a new finding, as the current sentence implies.

3. In the Conclusion paragraph of the abstract, the authors should consider stating that their research appears to validate the smaller sets of methylation markers that others (e.g. Weisenberger) have been running in CRC samples.

4. Methods, statistical analysis: Either provide more detail (formula) as to how the beta value was calculated, or a reference that provides the formula (e.g. Martin-Subero, et al., PLoS one, 2009). Was background intensity computed from negative controls subtracted from each analytical data point?

5. Regarding comment 1 under major compulsory revisions, the authors should acknowledge limitations of the unsupervised hierarchical clustering approach and perhaps reference the work of Houseman, et al. (BMC Bioinformatics, 2008) that concludes via simulation analysis of data from the Illumina methylation platform, their recursively-partitioned mixture model is an effective (with respect to classification error) and computationally efficient method for clustering DNA methylation data and is more reliable than other clustering approaches.

6. Provide more detail in the statistical analysis as to the determination of the binarized cut-off value of 0.297 for beta-values which was set based on a 5% FDR for the methylated control. Was the FDR based on a simulation analysis, and how many simulations?

7. Discussion: Given the relatively high proportion (65%) of CIMP-H tumors in their study especially compared to large population-based studies of CIMP in CRC, does the hospital in which the CRCs were sampled see a certain profile of patient (older, large proportion of smokers, etc.)?
8. Discussion (page 12, paragraph 3): the authors state that “the absence of MSI and KRAS and BRAF mutations in the 13 CIMP-M tumors suggests this subgroup may have a distinctive molecular and clinical phenotype.” The authors should also acknowledge that the rates of MSI, BRAF, and KRAS mutation in larger population-based studies of incident CRC (see Slattery, et al. Dis Colon Rectum 2009) would indicate that at most in a sample of 13 “CIMP-M” colon tumors (12 of which are distal), it would be expected that only 1 or 2 tumors would have any of these alterations; the absence of these alterations may have occurred by chance or because they are distal site, and not from an underlying distinct CIMP phenotype.

9. Table 1: Regarding distal tumor site, please footnote the numbers of samples that were rectal site (vs. distal colon).

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.