Dear Editors,

We are pleased to submit our manuscript entitled **"Consensus clustering for Bayesian mixture models"** for publication as a research article in *BMC Bioinformatics*.

Cluster analysis is a key part of precision medicine and systems biology. However, performing cluster analysis on high-dimensional 'omics data can be challenging. Prominent issues include choosing the number of clusters, and slow and limited exploration of different possible clustering solutions. An ensemble approach, such as consensus clustering, frequently explores the clustering space more thoroughly than individual models. Another tool for cluster analysis, Bayesian mixture models, has alternative advantages, including the ability to infer the number of clusters present, which avoids a user-chosen value. However, the most popular means to perform inference in Bayesian clustering, Markov chain Monte Carlo (MCMC) methods, are susceptible to becoming trapped in local optima when clustering high-dimensional data. This means that, in practice, Bayesian clustering models can produce inconsistent results across different choices of random initialisation, despite theoretical guarantees of ergodicity. We propose a novel application of consensus clustering to Bayesian mixture models to combine the strengths of both techniques. To our knowledge, consensus clustering has never previously been applied to Bayesian mixture models.

In a comprehensive simulation study, we compare consensus clustering of Bayesian mixture models to *Mclust*, a maximum likelihood implementation of Gaussian mixture models, as well as to traditional Bayesian inference. We show that the consensus clustering approach uncovers the generating structure and explores more modes in the likelihood surface than the other methods. Furthermore, consensus clustering yields significant reductions in runtime in a parallel environment compared to Bayesian inference.

We then perform an integrative analysis of three 'omics datasets relating to the cell cycle of budding yeast by applying consensus clustering to Multiple Dataset Integration, a Bayesian clustering method. We find sets of co-expressed genes that share regulatory proteins and validate these sets using external knowledge. We attempt a similar analysis using standard Bayesian inference but this does not produce consistent results in a runtime of 36 hours, with each chain getting trapped in different local optima. This shows the success of consensus clustering for the practical applicability of large Bayesian clustering methods.

Consensus clustering can be used as a wrapper to any existing MCMC-based clustering method, a process requiring no rewriting of the implementation, in contrast to changing to newer MCMC methods or variational inference. We believe that our method will be attractive to researchers attempting cluster analysis of biomedical data, as it reduces runtime and improves exploration, making for quicker, more consistent analyses. It is thus both easily applied and overcomes many of the difficulties associated with Bayesian clustering methods that might otherwise dissuade researchers from their use.

This research was funded in whole, or in part, by the Wellcome Trust [WT107881].

Yours sincerely,

Stephen Coleman, Paul DW Kirk and Chris Wallace