

Dear Editor,

Please consider our manuscript, "Clustering RNA-seq expression data using grade of membership models," for publication in PLoS Genetics.

Grade of membership models are a generalization of cluster models that have been widely used in population genetics to model admixture. For example, the widely-used software Structure (Pritchard et al, 2000) implements a grade of membership model for clustering samples on the basis of their genetic data. Here we demonstrate how a similar model can be used to cluster samples based on their gene expression (RNA-seq) data. Our example applications include both bulk RNA-seq data (the GTEx data) and single-cell RNA-seq data from mouse embryonic preimplantation development. In both cases the grade of membership model highlights interesting biological structure. Notably, in the single cell data, the cells do not cluster into entirely "distinct" groups, but rather show some continuous clines of variation, suggesting that classifying single cells into distinct cell "types", as traditional clustering methods would, is an oversimplification.

We believe that these methods have potential for widespread use in the analysis of RNA-seq data, particularly single cell data, and therefore that they will be of considerable interest to the readers of PLoS Genetics.

Our work has not been previously published and is not under consideration at any other journal or conference.

We thank you in advance for your consideration.

Sincerely,

Matthew Stephens, PhD
Professor of Statistics and Human Genetics
University of Chicago