**Spatial reconstruction of single cell gene expression data**

i) Spatial patterning of gene expression in developing embryos (as, in this case, with zebrafish) is much easily modeled due to significant amount of prior research in landmark genes regulating differentiation of the blastula and other germline layers. It is also much easier to verify the findings of such a study in zebrafish embryos, on account of their near-transparent form during early development. However, when trying to determine the spatial origin of single cells derived from, for ex., a bulk tissue extract from a single organ in an adult human, the gene expression signals are going to be confounded by latent processes shared by all the cells. In such a scenario, it might also be difficult to determine a robust set of landmark genes delineating within-organ distinctions. Would it okay to only consider the most highly variable genes (which, in the current method, they do) to conduct a decomposition of the data? Or should the methodology be adapted to normalize for latent processes when considering biological specimen that are not actively dividing and/or cannot be distinctly profiled by known landmark genes?

ii) The authors make a bold assumption that the expression levels of the different landmark genes are independent of each other, when establishing the probability distribution of each of these genes in the spatial context. This over-simplification of the cell expression state seems to defy their own logic by which these landmark genes as a set are supposed to present different joint expression profiles based on the spatial origin of a cell. Is this because they use an L1-constrained gene expression model, which automatically introduces sparsity in the gene expression fit for each landmark gene across all the cells? If so, does that mean that we would always need to have a high number of single cells profiled before we can make the assumption of independence within the landmark genes’ set (and that too only after LASSO based shrinkage)?

**Model based clustering for RNA-seq data**

i) It is not clear to me how their poisson distribution model differs significantly from their negative binomial based model. The dispersion parameter in the NB is the only ‘known’ factor they model for each gene; and from the description (vague, at that!) provided, it doesn’t seem tobe any different from the αg, also modeled as a known, in the poisson distribution. All the unknowns in these two distributions are the same by their own description. What, if any, is the value addition in using one of these two distributions in analyzing RNA-seq data in this model, in that case?

ii) It is interesting that the authors do not compare the hierarchical clustering results of the MB approach against the result of unsupervised clustering, which will be formed as an output during the Self Organizing Map algorithm. In Figure 3, I am curious that since the SOM has lower NMI scores in general, is the reason primarily because of lack of fine-tuning (increasing/decreasing the number of nodes in the SOM)? A demonstration of the bias introduced by using a high cluster count, in an unsupervised learning algorithm, can be seen in the k-means and NB comparison (pg. 204). While the authors try to present it in a different light, you get gene sets with the same MapMan functional annotation, in multiple different clusters, when using either of the two approaches.