QuantBio Express

ARTICLE

Identification of spatial discriminator genes by grade of membership model clustering on RNA-seq expression data

Received 20th March 2020,
Accepted 00th January 20xx

Lijia Wang, Thara Nallamothu, Dmitry Kondrashov

Understanding the key mechanism of morphogenesis during embryonic development is crucial to deciphering the guiding principles of body plan. Recently, many computational models have been created in order to identify spatial discriminator genes of embryos based on gene expression or prior spatial information. Two clustering models adopted for RNA-seq expression data analysis are the grade of membership (GoM) models and the self-organizing map (SoM) clustering. Our experiment compared the result identification from both models when applied to the same mouse embryo RNA-seq dataset. Results showed that the GoM model is able to identify some but not all of the 18 spatial discriminator genes identified by the SoM clustering, but was able to cluster genes by biological function and to identify spatial discriminator genes that plays an active role in morphogenesis. Our experiment suggests that GoM clusters might be able to provide an alternative to the Gene Ontology (GO) gene sets used as the SoM input and identify more influential and functionally relevant spatial discriminator genes.

Introduction

Ever since large-scale gene-expression analysis, such as RNA-seq and single-cell RNA-seq became available and widely applied, there has been a surge of collected gene-expression data as well as various methods of analysis, with clustering – of both genes and samples – being one of the popular methods. Clustering analysis on RNA-seq, such as the grade of membership (GoM) models, are capable of grouping genes by their distinct expression, and therefore creating functionally related gene clusters for further downstream analysis.[1]

One of the functions that has not been explored in the GoM model is its capability to group genes by their function on spatial organization of the cells, especially in embryo morphogenesis. Understanding the mechanism as well as the genes governing embryo morphogenesis is crucial as it informs the researchers of the guiding principles of body plan, which in turn assists research and development in biomedical fields. Clustering methods that can perform three-dimensional (3D) tissue reconstruction, such as the stochastic self-organizing map (stochastic-SoM) clustering, however, rely on either landmark genes or pre-grouped gene expression profiles to successfully reconstruct the correct spatial relationship between genes[2]. For instance, the later *ab initio* approach frequently uses principle component score calculated from gene expression profiles to assign each cell to its relative position in 3D place, providing a basis to facilitate the 3D reconstruction of SoM model. Without pre-grouping of genes by function, SoM cannot reconstruct the 3D structure with the gene expression data alone.

In our experiment, we aim to explore the possibility to identify the same spatial discriminator genes that were discovered with the SoM model using the GoM model instead. Since GoM model has the ability to accurately cluster genes by function, it could perform the same ability as isolating the gene clusters that are associated with distinct domains without relying on pre-grouped gene sets. Our experiment focuses solely on RNA-seq expression data analysis, but GoM also has the ability to analyse single-cell RNA-seq (scRNA-seq), so our findings may apply to scRNA-seq expression datasets as well.

Methods

Data Source Mouse embryo RNA-seq expression data is gathered from GSE65924, based on Peng 2016 paper[3]. The read counts of 23,361 genes by RNA-seq of 41 samples were normalized by FPKM (fragments per kilobase of transcript per million mapped reads).

Preliminary clustering of RNA-seq expression data using PCA RNA-seq expression data is collected and summarized into a data matrix $C_{N \times G}=(c_{ng})$ where N=41 is the number of samples and G=23,361 is the number of genes recorded in the dataset. Principle component analysis was performed on the matrix in order to visualize the relationship between the different domains (anterior A, posterior P, lateral left L, and lateral right R) as categorized in Peng 2016.

Clustering of RNA-seq expression data using GoM model
Topics are calculated with *maptpx* package. Each sample (n) has some proportion of its reads q_{nk} coming from cluster *k*, which is in tern characterized by a probability vector θ_k, whose *g*th element represents the relative expression of gene *g* in cluster *k*. The GoM model is thus:

$$(c_{n1}, c_{n2}, \dots, c_{nG}) \sim Multinomial(c_{n+}, p_{n1}, p_{n2}, \dots, p_{nG}),$$

where

$$p_{ng} \coloneqq \sum_{k=1}^{K} q_{nk}\theta_{kg}.$$

The number of clusters *K* is adjustable depending on dataset. Typically, larger *K*'s are preferable but results in longer calculation time. Since our sample number is small (N = 41), we chose a relatively large *K = 20* for our analysis.

Results

PCA analysis of counts matrix PCA analysis was carried out for the C_{NxG} counts matrix.

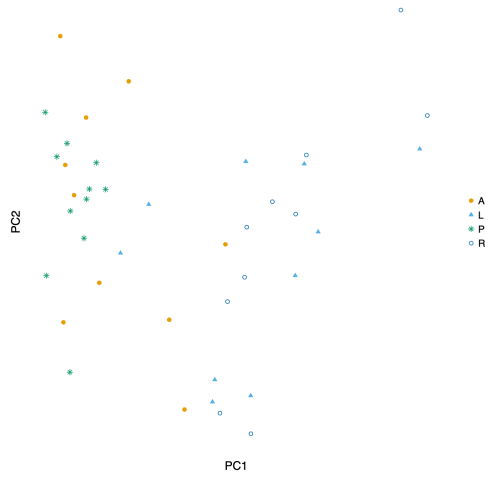

Figure 1 | PCA plot for RNA-seq expression counts matrix, components 1 and 2

From figure 1, we could observe that there is significant clustering between anterior and posterior samples' gene expression as well as between lateral left and right. PC's 3-10 does not show significant information distinguishing the sample features in the 4 domains (figure 2). From the above observation, we know that the 4 domains have significant differences that we should be able to recover with the GoM model; we should also expect to see more similarity between the cluster composition of A, P domains and L,R domains as presented in figure 1.

Figure 2 | PCA plot for RNA-seq expression counts matrix, components 3 through 10

Clustering using the GoM model

Clustering with GoM model is performed using K = 20 and plotted in figure 3. There are unique clusters that are identified in each domain: cluster 3 (red) is mostly observed in posterior domain; cluster (light green) 10 is observed in lateral left, cluster 4 (salmon) in anterior, and cluster 8 and 6 (orange and light blue, respectively) in lateral right domains. We could also see a great portion of cluster 2 in posterior and anterior domains, which is potentially the reason we observed clustering between these two domain groups in figure 1.

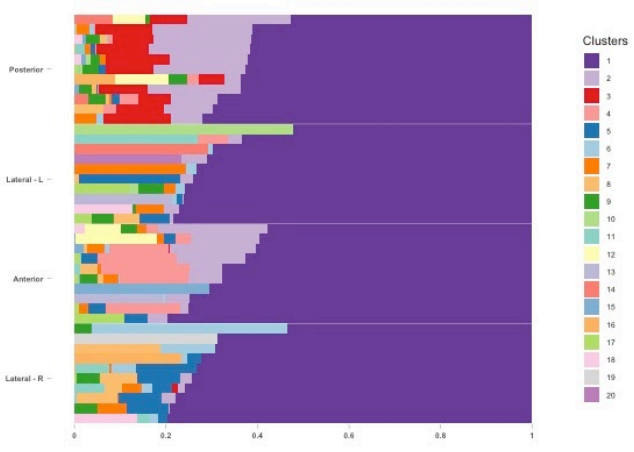

Figure 3 | Cluster composition of each domain according to GoM model clustering

Top gene identification and recovery

Mori et. al. identified 18 top genes that are essential to structural reconstruction through SoM clustering and regarded those 18 genes as the essential spatial discriminator genes. Among them *Id2* is regarded as the most influential contributing to reconstruction.

When top 100 driving genes (highest θ) from each cluster are extracted from the GoM analysis, we discovered that only *Id2* (cluster 15) and *Coro1b* (cluster 9) from the list of 18 genes were recovered. Cluster 15 is predominantly located in the anterior domain, and according to iTranscriptome there is a gene expression gradient of *Id2* from the anterior proximal side to posterior distal side of the mid-gastrula embryo[3], we can regard this identification as accurately reflecting the function of *Id2* in guiding morphogenesis in anterior domain.

Since we only recovered 2 out of 18 genes when top 100 θ's is extracted, we decided to expand the range of acceptance. Among the top 200 genes identified in each cluster, we recovered 6 out of 18 genes in total, including *Fgfrl1* (cluster 11), *Igfbp4* (cluster 5), *Nrp2* (cluster 15), and *Pdgfra* (cluster 3). *Nrp2* was identified to have a similar expression gradient pattern as Id2 and shows high expression in anterior proximal region, therefore its categorization in cluster 15 shows that genes with similar expression and function are indeed grouped together through GoM clustering.

Furthermore, functions of top driving genes within each cluster are inspected. It is observed that genes within the same cluster tend to have very similar functions. For example: *Rps29, Rpl36, and Rpl41* in cluster 6 all code for ribosomal proteins; *Ndufa4 and Cox6a1* in cluster 10 code for cytochrome c oxidase. One of the key discoveries is that the top driving genes in cluster 3: *Mixl1, Mesp1*, and *Tdgf1* are all genes governing the epithelialization or general development of the mesoderm in the posterior. GoM not only is able to place them in the right domain, but also highlighted cluster 3 genes' active role in morphogenesis. These genes, however, were not identified in the Mori et. al. paper.

Discussion

Clustering is one of the most popular methods for RNA-seq data analysis. Different methods have different emphasis on the information it provides: GoM tends to group genes by function while SoM can provide more spatial information. Our goal in our experiment is to see if GoM is able to isolate the same genes that SoM identified to be the top spatial discriminator genes. This will allow us to compare the efficiency and applicability between GoM and SoM. Basically, it will tell us which one is better at its job as a clustering method.

Our results showed that GoM able to recover some but not all of the top spatial discriminator genes identified in SoM. However, comparing the top 100 and top 200 driving genes in each cluster, we can see that these genes are still among the top driving genes identified for the clusters, but the GoM model regarded them as less important functionality wise. This shows that GoM is not capable of completely prioritizing spatial discriminating function like SoM is capable of.

One alternative interpretation of this observation is that, as we discussed above, SoM requires the genes in the RNA-seq dataset to be pre-grouped by functionality. In the Mori et. al. experiment, the authors were not able to perform 3D reconstruction on the 23,361 gene set, therefore they used the GO:0060412 (ventricular septum morphogenesis) gene set. This could potentially cause its failure in isolating the top driving genes *Mixl1, Mesp1,* and *Tdgf1* that are clearly playing an active role in morphogenesis. In fact, I personally think that the genes isolated in Mori et. al. might only be the genes that were proven to be necessary for spatial reconstruction in the specific GO used, which might have resulted in them not giving equal weight and consideration to other genes in the dataset. Furthermore, few of the genes demonstrated a functional correlation to spatial organization or direct regulation on morphogenesis, so their claim on these identified genes being "top spatial discriminator genes" seem to be a bit of a stretch to me.

One option worth exploring in the future is using GoM and SoM in parallel. GoM has proven to be able to accurately provide functional clusters in each domain; if we replace the GO sets used in the SoM model with these gene clusters, we might be able to explore the exact genes that are directly involved in morphogenesis in each spatial domain.

On an unrelated note: there is a function in GoM code to make a really beautiful annotation table of the function of each gene identified in the cluster. However, I was not able to get it to work, so I had to search through UniProt one gene at a time. It made me just a little sad.

Code Availability

The experimental method including processing RNA-seq dataset, principle component analysis, and grade of membership model clustering are implemented in R and are available at *https://github.com/lwa19/GOM2SOM*

Conclusions

In conclusion, while our experiment demonstrated the superiority of GoM in clustering genes by function and SoM's superiority in isolating genes that is crucial for only spatial reconstruction, we found GoM to be an overall more useful method. GoM does not rely on pre-grouping, can be performed directly on RNA-seq datasets, and is very robust when it comes to matching the top-driving clusters to each cluster on top of accurate aggregation of genes by biological functionality. It will be an overall more fitting method to use in identifying spatial discriminator genes that are actively governing morphogenesis in a given RNA-seq dataset.

Acknowledgements

We thank Peter Carbonetto from the Matthew Stephens lab for providing guidance and updated code for the GoM model, and Kushal K. Dey and Tomoya Mori (authors of the GoM and SoM papers, respectively) for providing code and explanation of the two models. We would also like to show our gratitude to Professor Dmitry Kondrashov for guidance on this project and encouragement along the way.

References

1 Dey, Kushal K., Chiaowen Joyce Hsiao, and Matthew Stephens. "Visualizing the structure of RNA-seq expression data using grade of membership models." *PLoS genetics* 13.3 (2017).

2 Mori, Tomoya, et al. "Novel computational model of gastrula morphogenesis to identify spatial discriminator genes by self-organizing map (SOM) clustering." *Scientific reports* 9.1 (2019): 1-10.

3 Peng, Guangdun, et al. "Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo." *Developmental cell* 36.6 (2016): 681-697.