Multi-way Clustering of GTEx RNA-seq Data

Zhuofan Wang[a]

[a]Institute of Statistics and Big Data, Renmin University of China, Beijing, China

1 Introduction

In the past decades, there has been a rich literature showing a strong statistical association between genetic variation and human traits. But for the majority of these, the mechanisms underlying disease susceptibility remain unknown. It is crucial to further explore gene expression and patterns which can help to define mechanisms underlying disease susceptibility. The development of high-throughput sequencing technology enables measurements of RNA sequencing data across multiple individuals and tissues which can be seen as tensor data. Here we describe the Genotype-Tissue Expression (GTEx) project (Melé et al., 2015; Lonsdale et al., 2013), which will establish a resource database and associated tissue bank for the scientific community to study the relationship between genetic variation and gene expression in human tissues.

In analysis of RNA-seq data, clustering has played a major role in discovering co-regulated gene family and biologically-distinct subgroups of individuals. With the availability of GTEx data, multi-tissue RNA-seq clustering has the potential to improve the results of clustering by borrowing strength across tissues and to explore gene expression variations in different modules (e.g. different tissues sets and donor subgroups). Clustering, as a common task in statistical analysis, is widely

used in pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics, and machine learning. However, traditional clustering methods are based on similarities between vectors where every element are assigned equal weights. It fails when it comes to gene expression data since genes express distinctly across samples. Biclustering is then proposed to address these problems by considering the heterogeneity (Cheng and Church, 2000). Several attempts have been made to further develop the literature, especially in bioinformatics.

However, both traditional clustering aimed at order-1 case and biclustering aimed at order-2 case can not provide a satisfactory result of GTEx data since gene expression varies across multiple tissues and individuals. We aim to propose a Bayesian multi-way clustering approach to cluster simultaneously on each mode, which can help to define subgroups of donors and gene families, then further explains gene expression patterns.

2 GTEx RNA-seq Data

We can have a glance of the raw read count data of GTEx RNA-seq. This dataset is formed as matrix which has 56318 rows and 8555 columns with each row corresponds to an RNA and each column corresponds to a sample. Note that one sample comes from a slice of tissue in one donor.

Name <chr>	Description <chr>	GTEX.113IC.0226.SM.5HL5C <dbl>	GTEX.117YX.2226.SM.5EGJJ <dbl>	GTEX.11DXW.0326.SM.5H11W <dbl>
ENSG00000223972.4	DDX11L1	0	0	0
ENSG00000227232.4	WASH7P	391	823	552
ENSG00000243485.2	MIR1302-11	0	0	0
ENSG00000237613.2	FAM138A	0	0	0
ENSG00000268020.2	OR4G4P	0	0	0
ENSG00000238009.2	RP11-34P13.7	6	7	12

6 rows

Figure 1: First few rows and columns of GTEx RNA-seq data

2.1 Exploratory Data Analysis

2.2 Data Preprocessing

Before having a visualization of our data, we first need to preprocess the raw data.

Quality Control: We use the estimateSizeFactor function in R package DESeq2. This step takes into account sequencing depth and RNA composition and eliminates their impact on data.

Data selection: We laid our interest on the depression-related genes on brain subregion of GTEx-RNA data with curiosity about how these genes take effect. These genes are selected according to http://www.brainspan.org/ish. Then we got a $193 \times 13 \times 15$ tensor, that is 193 donors, 13 brain tissues, and 15 genes.

2.3 Data visualization

To see the data heterogeneity, we plot the heatmaps of six selected genes which shows that genes over/under express in different sets of tissues. For example, in figure 2, the first gene over-express in cerebellum and cerebellar hemisphere while under-expressing in three basal ganglia subtissues. But the second gene over-express in substantia while under-expressing in cerebellum and cerebellar hemisphere.

We also produce a correlation analysis and conduct a t-SNE analysis to see how the gene expression data distributes. T-SNE is a useful dimension reduction method that allows you to visualize data embedded in a lower number of dimensions, e.g. 2, in order to see patterns and trends in the data. Results show that the correlations differ across different tissue-sets and t-SNE analysis splits the samples also by tissue-sets. See in figure 3

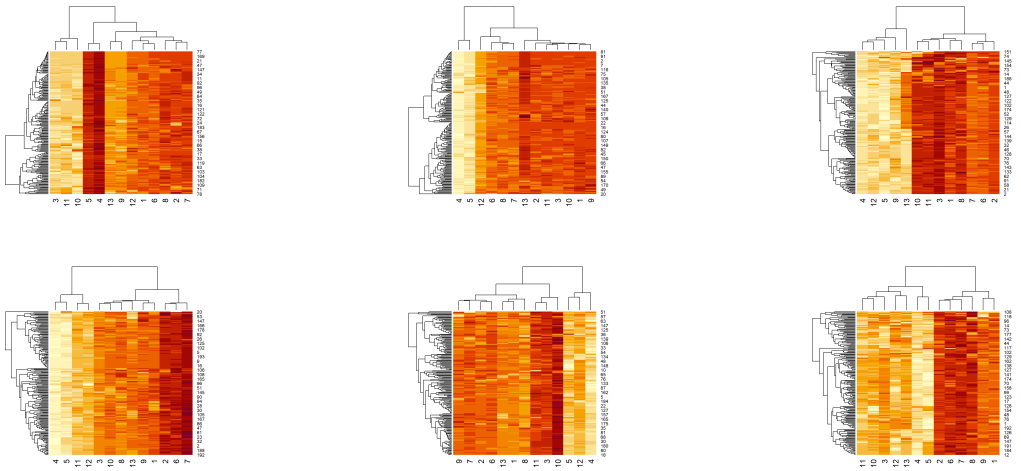

Figure 2: Heatmap of 6 selected genes

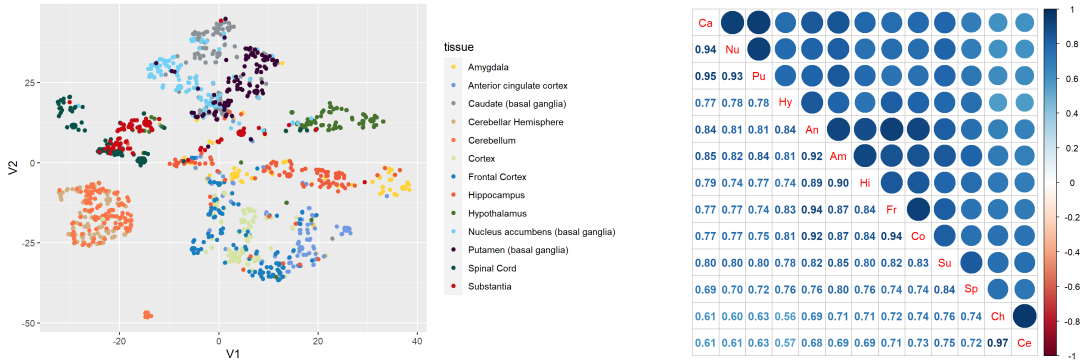

Figure 3: t-SNE Analysis and Correlation Plot

3 Multi-way clustering Model

Assume $\boldsymbol{Y}$ is a $D \times T \times G$ tensor, according to Parmigiani et al. (2002), we use

$$y_{dtg} \sim I(z_{dtg} = -1)\mathrm{U}(\mu_t+\mu_g - k_g^-, \mu_t+\mu_g)$$

$$+ I(z_{dtg} = 0)\mathrm{N}(\mu_t+\mu_g, \sigma_g^2)$$

$$+ I(z_{dtg} = 1)\mathrm{U}(\mu_t+\mu_g, \mu_t+\mu_g + k_g^+)$$

to model the gene expression data, where the latent indicator $z_{dtg} = -1$, 0, and 1 represent the case of under, normal, and over-expression of gene g respectively.

For the bayesian hierarchical model, we link z_{dtg} with θ_{dtg} by a latent multi-class logistic model:

$$z_{dtg} \sim \text{ Categorical } \left\{ M^{-1} \exp(\theta^-_{dtg}), M^{-1}, M^{-1} \exp(\theta^+_{dtg}) \right\},$$

where M is a normalizing constant.

Let θ_{dtg} be the (dtg)-th element of $\mathbf{\Theta}$. We propose

$$\theta^-_{dtg} = \sum_{r=1}^{R} c^r_{1d} c^r_{2t} \omega^{r-}_{3g} I\left(c^r_{3g} = -1\right) + b^-, \tag{1}$$

$$\theta^+_{dtg} = \sum_{r=1}^{R} c^r_{1d} c^r_{2t} \omega^{r+}_{3g} I\left(c^r_{3g} = 1\right) + b^+ \tag{2}$$

We set $\tilde{\boldsymbol{c}_3}^+ = \omega_3^+ \circ I(\boldsymbol{c}_3 = 1^{G\times R})$,$\tilde{\boldsymbol{c}_3}^- = \omega_3^- \circ I(\boldsymbol{c}_3 = -1^{G\times R})$ the Hadamard product (element-wise) of ω_3 and $\boldsymbol{c}_3$. Then (1)(2) can be denoted as

$$\mathbf{\Theta}^- = \sum_{r=1}^{R} \boldsymbol{c}^r_1 \circ \boldsymbol{c}^r_2 \circ \tilde{\boldsymbol{c}_3}^{r-} + \boldsymbol{B}^- \tag{3}$$

$$\mathbf{\Theta}^+ = \sum_{r=1}^{R} \boldsymbol{c}^r_1 \circ \boldsymbol{c}^r_2 \circ \tilde{\boldsymbol{c}_3}^{r+} + \boldsymbol{B}^+ \tag{4}$$

The proposed model (3)(4) coincides with the CP (CANDECOMP/PARAFAC) decomposition (Kolda and Bader, 2009) for its simplicity of representation and meaningful interpretation of clustering.

Prior: We set prior with $\mu_g \sim N\left(m_\mu, \sigma_\mu^2\right)$,$k_g^-, k_g^+ \sim \text{Gamma}\left(\alpha_k, \beta_k\right)$, and

$$\sigma_g^2 \sim \text{IG}\left(\alpha_\sigma, \beta_\sigma\right) I\left(\sigma_g < \min\left(k_g^-, k_g^+\right)/k_0\right)$$

$\boldsymbol{C_1}$ follows an IBP process (Ghahramani and Griffiths, 2005) which is determined by parameter m, elements of $\boldsymbol{C_1} \sim \text{Bernoulli}(\rho)$, each element of $\boldsymbol{C_3}$ follows the categorical distribution $C_{3g}^r \sim \text{Categorical}\ (\boldsymbol{\gamma})$ with $\gamma = (\gamma_{-1}, \gamma_0, \gamma_1)$. We also assume $\omega_{3g}^{r+}, \omega_{3g}^{r-} \sim \text{Gamma}\left(a_w, b_w\right)$, $b^+, b^- \sim N\left(\mu_b, \sigma_b^2\right)$, and $m \sim \text{Gamma}\left(\alpha_m, \beta_m\right)$

4 Results

The implement details are stated as follows:

Initial start point: we set elements of C_1, C_2, C_3 generating from Bernoulli distribution with parameter 0.5 and they correspond to the donor, tissue, gene respectively. We set the prior of C_1 to be IBP, C_2 Bernoulli and C_3 categorical in $\{-1, 0, 1\}$. The weight matrix is generated from the Gamma distribution.

MCMC: We run the MCMC algorithm for 10,000 iterations with one random initial cluster. The first 5,000 iterations are discarded as burn-in and posterior samples are retained every 10th iteration after burn-in. And We summarize R, C_1, C_2, C_3 with same procedure as described in inference section. We derive the In-sample error 7.4% and In-sample correlation between fitted value and observation 96%. We derive 3 clusters overall and details in each component are shown below.

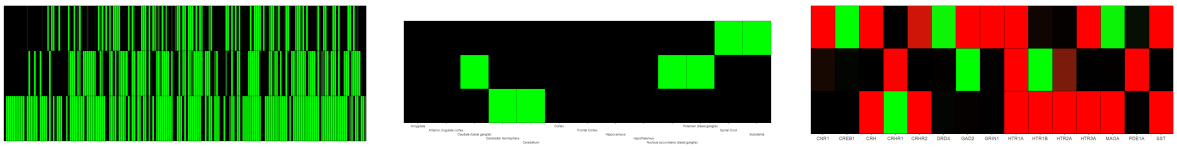

Figure 4: Experiment Results

4.1 Interpretation

The first cluster is a cerebellum related cluster. It only contain cerebellar hemisphere and cerebellum. Genes underexpressed in this cluster is highly correlated to regulation of serotonin secretion ($p = 3.67 \times 10^{-9}$ under Bonferroni correction).

Serotonin is an important neuromodulatory hormone that makes us feel happy and happy. Increasing serotonin levels can improve sleep, calm people, reduce impatience, bring pleasure and happiness, and bring more happiness. The natural way to boost serotonin is through diet, regular exercise, stress reduction, reminiscing about happiness, sunbathing, yoga, meditation, deep breathing exercises, and less well known: keeping a journal can also increase serotonin levels in the body.

The second cluster is a basal ganglia related cluster. It contain three basal ganglia subtissues: caudate, putamen and nucleus accumbens. Genes underexpressed in this cluster is highly correlated to regulation of amine transport ($p = 5.67 \times 10^{-4}$ under Bonferroni correction) while genes overexpressed in this cluster is highly related to neurotransmitter secretion ($p = 2.96 \times 10^{-2}$ under Bonferroni correction).

The third cluster is a spinal cord and substantia related cluster. Genes underexpressed in this cluster is highly correlated to anterograde trans-synaptic signaling ($p = 8.37 \times 10^{-7}$ under Bonferroni correction) while genes overexpressed in this cluster is highly related to chemotaxis to arachidonic acid ($p = 2.94 \times 10^{-2}$ under Bonferroni correction) and phenylethylamine metabolic

process ($p = 2.94 \times 10^{-2}$ under Bonferroni correction). Hypergeometric test shows females ($p = 0.017$) are enriched in this cluster , which may suggest an age and gender effect on this cluster.

5 Discussion

Motivated by GTEx RNA-seq data, we proposed a unified identifiable multi-way clustering approach which can cluster high-order tensor data from each mode simultaneously. Coincided with CP (CANDECOMP/PARAFAC) decomposition (Kolda and Bader, 2009), our model can fully explore the tensor structure and show interaction among multi-aspects. Our approach can also determine the cluster number from the posterior samples via a nonparametric Bayesian prior – IBP process (Ghahramani and Griffiths, 2005). Applying on GTEx RNA-seq data, we discover three gene expression modules within brain region. The three modules, cerebellum related cluster, asal ganglia realted cluster, the spinal cord and substantia related cluster, together with correlated gene families, shows underlying biological functions of depression, which may further assist in uncovering disease machanism.

References

Cheng, Y. and Church, G. M. "Biclustering of expression data." In *Ismb*, volume 8, 93–103 (2000).

Ghahramani, Z. and Griffiths, T. "Infinite latent feature models and the Indian buffet process." *Advances in neural information processing systems*, 18 (2005).

Kolda, T. G. and Bader, B. W. "Tensor decompositions and applications." *SIAM Review*, 51(3):455–500 (2009).

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N. J., Nicolae, D. L., Gamazon, E. R., Im, H. K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E. T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J. M., Wilder, E. L., Derr, L. K., Green, E. D., Struewing, J. P., Temple, G., Volpi, S., Boyer, J. T., Thomson, E. J., Guyer, M. S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T. R., Koester, S. E., Little, A. R., Bender, P. K., Lehner, T., Yao, Y., Compton, C. C., Vaught, J. B., Sawyer, S., Lockhart, N. C., Demchok, J., and Moore, H. F. "The Genotype-Tissue Expression (GTEx) project." *Nature Genetics*, 45(6):580–585 (2013).

URL http://www.nature.com/articles/ng.2653

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Consortium, T. G., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Der-

mitzakis, E. T., Ardlie, K. G., and Guigó, R. "The human transcriptome across tissues and individuals." *Science*, 348(6235):660–665 (2015).

URL https://www.science.org/doi/10.1126/science.aaa0355

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. "A statistical framework for expression-based molecular classification in cancer." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):717–736 (2002).