Multi-way Clustering on GTEx RNA-seq data via Bayesian approach

Zhuofan Wang

Institute of Statistics and Big Data

April 26. 2022

中國人民大學
RENMIN UNIVERSITY OF CHINA

1. Data

2. Exploratory Data Analysis

3. Model

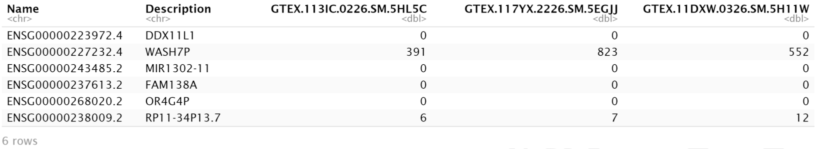
4. Results

1. Data
2. Exploratory Data Analysis
3. Model
4. Results

The Genotype-Tissue Expression (GTEx) project

- Genome-wide association studies have identified thousands of loci for common diseases, but, for the majority of these, the mechanisms underlying disease susceptibility remain unknown.
- Most associated variants are not correlated with protein-coding changes, suggesting that polymorphisms in regulatory regions probably contribute to many disease phenotypes.
- Here we describe the Genotype-Tissue Expression (GTEx) project, 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.

GTEx RNA-seq Data Revisit



Name	Description	GTEX.113IC.0226.SM.5HL5C	GTEX.117YX.2226.SM.5EGJJ	GTEX.11DXW.0326.SM.5H11W
<chr>	<chr>	<dbl>	<dbl>	<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

1. Data
2. Exploratory Data Analysis
   - Data Preprocessing
   - Data visualization
3. Model
4. Results

1. Data
2. Exploratory Data Analysis
   - Data Preprocessing
   - Data visualization
3. Model
4. Results

Data Preprocessing

- **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 × 13 × 15 tensor, that is 193 donors, 13 brain tissues, and 15 genes.

1 Data

2 Exploratory Data Analysis

- Data Preprocessing
- Data visualization

3 Model

4 Results

Heatmap

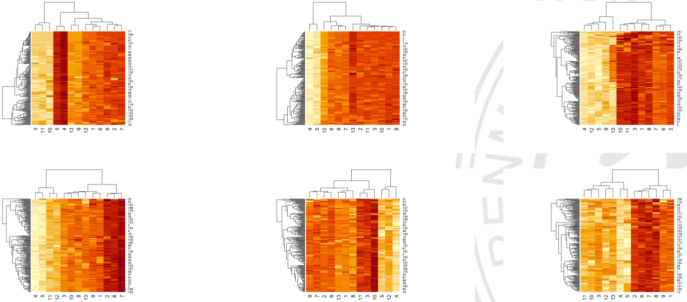


Figure 2: Heatmap of 6 selected genes

t-SNE Analysis and Correlation Plot

We also conduct a t-SNE analysis to see how the gene expression data distributes. T-SNE is a useful dimensionality reduction method that allows you to visualise data embedded in a lower number of dimensions, e.g. 2, in order to see patterns and trends in the data.

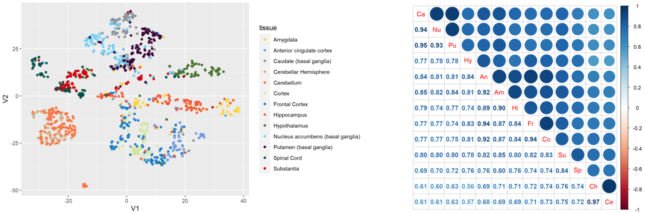


Figure 3: t-SNE Analysis and Correlation Plot

1. Data
2. Exploratory Data Analysis
3. Model
4. Results

Bayesian Hierachical Model

Assume $\boldsymbol{Y}$ is a $D \times T \times G$ tensor, according to [2], we use

$$
\begin{aligned}
y_{dtg} \sim I(z_{dtg} = -1)\mathrm{U}(\mu_t + \mu_g - k_g^-, \mu_t + \mu_g) \\
+ I(z_{dtg} = 0)\mathrm{N}(\mu_t + \mu_g, \sigma_g^2) \\
+ I(z_{dtg} = 1)\mathrm{U}(\mu_t + \mu_g, \mu_t + \mu_g + k_g^+)
\end{aligned}
$$

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.

Logistic Link

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 $\boldsymbol{\Theta}$. We propose

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

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

CP (CANDECOMP/PARAFAC) decomposition

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

$$\boldsymbol{\Theta}^{-} = \sum_{r=1}^{R} \boldsymbol{c}_1^r \circ \boldsymbol{c}_2^r \circ \tilde{\boldsymbol{c}}_3{}^{r-} + \boldsymbol{B}^{-} \tag{3}$$

$$\boldsymbol{\Theta}^{+} = \sum_{r=1}^{R} \boldsymbol{c}_1^r \circ \boldsymbol{c}_2^r \circ \tilde{\boldsymbol{c}}_3{}^{r+} + \boldsymbol{B}^{+} \tag{4}$$

The proposed model (3)(4) coincides with the CP (CANDECOMP/PARAFAC) decomposition [1] for its simplicity of representation and meaningful interpretation of clustering.

Prior Setting

- 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 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 $C_{3g}^r \sim \text{Categorical}\ (\gamma)$ 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 $m \sim \text{Gamma}\left(\alpha_m, \beta_m\right)$

1. Data
2. Exploratory Data Analysis
3. Model
4. Results

Implement Details

- 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.
- 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%.

Experiment Results

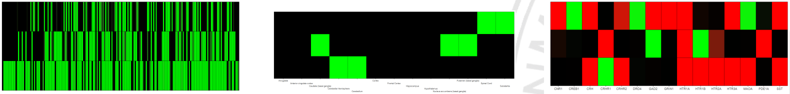


Figure 4: Experiment Results

Interpretation

We derive 3 clusters overall and details in each component are shown below.

- 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.

Interpretation

We derive 3 clusters overall and details in each component are shown below.

- 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).

Interpretation

We derive 3 clusters overall and details in each component are shown below.

- 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.

[1] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[2] Giovanni Parmigiani, Elizabeth S Garrett, Ramaswamy Anbazhagan, and Edward Gabrielson. 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.

Thanks!