our main point

1. section 1: introduction begins here
2. previous comparative gene set enrichment analysis does not take....
3. we propose a method that allows DE within the test set as well as the background gene set.

Comparative gene set enrichment analysis for correlated expression data

Abstract

To be filled

1 Introduction

Let's get started.

2 Methods

Overview of our method (denoted as OurMethod, will be easily replaced when we have a better new name)
Different from CAMERAWu and Smyth (2012) or GSEA (Subramanian et al., 2005)
Our method is based on case-control

2.1 The general assumptions for expression data

In a treatment-control gene expression experiment, we denote Y_{ijk} as a random variable for the expression level of gene i from observational unit j in treatment group k, with i taking the values $1, \ldots, m$ (the number of genes), j taking the values $1, \ldots, n_k$ (the total number of biological samples), and k being either 1 for control or 2 for treatment. Correspondingly, Y^*_{ijk} represents the standardized expression levels (described in REF???) for gene i of sample j, with $Y^*_{ijk} \sim N(0,1)$ (??? Normal assumption necessary here???) if sample j comes from the control group, and $Y^*_{ijk} \sim N(\Delta_i, 1)$ if it comes from the treatment group. Here, Δ_i is an *DE effect*: compared to the control group, gene i is not DE if $\Delta_i = 0$, up-regulated if $\Delta_i > 0$ and down-regulated if $\Delta_i < 0$. In a gene expression experiment, the DE effect Δ_i consists of two parts: 1) the treatment which determines whether a gene is DE or not; and 2) the strength when the gene is DE. For 1), we let $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ be a vector of DE indicators, where $Z_i = 1$ if gene i is DE and $Z_i = 0$ otherwise, and (DO WE NEED TO ASSUME Z_is TO BE INDEPENDENT OF EACH OTHER?

$$Z_i \sim \text{Binom}(1, p_i) \tag{1}$$

For 2), we denote δ_i as the *DE effect size* for gene i and δ_i follows some distribution f_δ with mean and variance

$$E(\delta_i) = \mu_\delta, \quad \text{Var}(\delta_i) = \sigma^2_\delta \tag{2}$$

We further assume that the DE indicator Z_i is independent of the DE effect size δ_i for gene $i = 1, \ldots, m$. Therefore, the DE effect can be expressed as

$$\Delta_i = Z_i \delta_i, \tag{3}$$

It can be shown that (details in Appendix 6),

$$E(\Delta_i) = p_i \mu_\delta, \quad \text{Var}(\Delta_i) = p_i \sigma^2_\delta + p_i(1-p_i)\mu^2_\delta, \quad i = 1, \ldots, m. \tag{4}$$

We assume that conditioning on the DE effects, expression levels for different samples are independent, but expression levels for different genes of the same sample may be correlated. Denote $C_{m\times m}$ as the gene correlation matrix, with entry ρ_{i_1,i_2} being the correlation between genes i_1 and i_2. Note that the between-gene correlation ρ_{i_1,i_2} is a constant, regardless of whether the sample is from the treatment or from the control group.

2.2 Gene set enrichment test

many method propose using a test statisitic as the measure of DE effec, and test the set against the backgroud genes.

Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ be a indicator vector of whether or not gene i belongs to the GO term being tested, and $I_g = \{i : x_i = 1\}$ be the set of GO term genes and $I_b = \{i : x_i = 0\}$ the set of *background genes*. We assume that the DE probability is p_g for the GO term genes and p_b for the background genes. For gene i, denote $U_i = \bar{Y}_{i.2} - \bar{Y}_{i.1}$ as the difference in mean expression levels between the treatment group and the control group, where $\bar{Y}_{i.k} = \sum_{j=1}^{n_k} Y_{ijk}/n_k$. It follows from equation 4 that $\boldsymbol{U} = (U_1, \ldots, U_m)$ has mean

$$E(U_i) = \begin{cases} p_g\mu_\delta, & \text{if } i \in I_g \\ p_b\mu_\delta, & \text{if } i \in I_b \end{cases} \tag{5}$$

and covariance (see Appendix 6 for detail)

$$\text{Var}(\boldsymbol{U}) = \boldsymbol{D} + \sigma_2^2\boldsymbol{C} \tag{6}$$

where $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_m)$ with $d_i = p_g\sigma_\delta^2 + p_g(1-p_g)\mu_\delta^2$ if $i \in I_g$ and $d_i = p_b\sigma_\delta^2 + p_b(1-p_b)\mu_\delta^2$ if $i \in I_b$, $\sigma_2^2 = \frac{1}{n_1} + \frac{1}{n_2}$ and $\boldsymbol{C}$ is the between-gene correlation matrix .

(**The test**) The GO term status affects both the mean vector in equation 5 and the covariance in equation 6. Under this framework, the GO term is not enriched only if the mean and variance of DE effects in the GO term genes are the same as those in the background genes. Therefore, the hypothesis for enrichment testing can be statistically formulated as

$$H_0\text{: } \mu_g = \mu_b \text{ and } \sigma_g^2 = \sigma_b^2 \overset{\text{def}}{=} \sigma_1^2 \text{ Versus } H_1\text{: at least one equation does not hold} \tag{7}$$

We can combine equation 5 and 6 into the following linear model

$$\boldsymbol{U} = \beta_0\mathbf{1}_m + \beta_1\boldsymbol{x} + \boldsymbol{\epsilon}, \quad \text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{D} + \sigma_2^2\boldsymbol{C} \tag{8}$$

with $\beta_0 = \mu_b$, $\beta_1 = \mu_g - \mu_b$ and $\mathbf{1}_m$ is a vector of ones. Under the null in 7, we have $E(\boldsymbol{U}) = \beta_0\mathbf{1}_m$ and $\text{Var}(\boldsymbol{U}) = \sigma_1^2\boldsymbol{I}_m + \sigma_2^2\boldsymbol{C}$ where $\boldsymbol{I}_m$ is an identity matrix.

(**Estimating the parameters**) In practice, we need to estimate β_0, σ_1^2 and $\boldsymbol{C}$ in 8 for enrichment test. Our strategy is to use *quasi-likelihood*, which requires only the mean and the variance of $\boldsymbol{U}$. The between-gene correlation matrix $\boldsymbol{C}$ is estimated by the residual sample correlations after the treatment differences have been nullified (the same as is done in Efron (2007) or Wu and Smyth (2012)), and is treated as known in estimating β_0 and σ_1^2. Denoting $\hat{\boldsymbol{C}}$ as the estimate of $\boldsymbol{C}$ and,

$$\boldsymbol{\Sigma} = \sigma_1^2\boldsymbol{I}_m + \sigma_2^2\hat{\boldsymbol{C}} \tag{9}$$

The score equations for β_0 and σ_1^2 are

$$\begin{aligned} (\boldsymbol{U} - \beta_0\mathbf{1}_m)^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_m &= 0 \\ (\boldsymbol{U} - \beta_0\mathbf{1}_m)^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{C}}(\boldsymbol{U} - \beta_0\mathbf{1}_m) &= \text{trace}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{C}}) \end{aligned} \tag{10}$$

.... something to catch up.....
The enrichment test statistic for the GO term is

$$T = \frac{\left[\boldsymbol{x}^T(\boldsymbol{U} - \hat{\beta}_0\mathbf{1}_m)\right]^2}{\left[\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{H})\right]\boldsymbol{\Sigma}\left[\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{H})\right]^T} \tag{11}$$

3 Results

4 Conclusion

5 AcknowledgeMents

6 Appendix

First $E(\Delta_i) = E(Z_i\delta_i) = E(Z_i)E(\delta_i) = p_i\mu_\delta$. Next

$$\mathrm{Var}(\Delta_i) = E[(Z_i\delta_i)^2] - [E(Z_i\delta_i)]^2 = \mathrm{Var}(Z_i)[E(\delta_i)]^2 + \left[(EZ_i)^2 + \mathrm{Var}(Z_i)\right]\mathrm{Var}(\delta_i) = p_i\sigma_\delta^2 + p_i(1-p_i)\mu_\delta^2$$

Let $T_i = \bar{Y}_{i,2} - \bar{Y}_{i,1}$ be the difference in mean expression levels between the treatment group and the control group. We have

$$E(T_i) = E(\bar{Y}_{i,2}) - E(\bar{Y}_{i,1}) = E(\Delta_i) = E(Z_i\delta_i) = p_i\mu_\delta$$

The covariance between two genes i_1 and i_2 is given by (I HAVE CONCERNS HERE, IS IT VALID TO ASSUME THAT DE EFFECTS ARE INDEPENDENT BETWEEN GENES? WE SEE CO-EXPRESSION!! OR WE'VE ALREADY TAKEN THAT INTO ACCOUNT BY "CORRELATION BETWEEN GENES"),

$$\begin{aligned}
\mathrm{Cov}(T_{i_1}, T_{i_2}) &= E\left[\mathrm{Cov}(T_{i_1}, T_{i_2}|\Delta_{i_1}, \Delta_{i_2})\right] + \mathrm{Cov}\left[E(T_{i_1}|\Delta_{i_1}), E(T_{i_2}|\Delta_{i_2})\right] \\
&= E\left(\frac{1}{n_1}\rho + \frac{1}{n_2}\rho\right) + \mathrm{Cov}(\Delta_{i_1}, \Delta_{i_2}) \\
&= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\rho_{i_1,i_2}
\end{aligned} \tag{12}$$

For gene i, the variance $\mathrm{Var}(T_i) = \mathrm{Var}(\bar{Y}_{i,1}) + \mathrm{Var}(\bar{Y}_{i,2})$, with

$$\mathrm{Var}(\bar{Y}_{i,1}) = \frac{1}{n_1}$$

$$\begin{aligned}
\mathrm{Var}(\bar{Y}_{i,2}) &= \frac{1}{n_2^2}\left[\sum_{j=1}^{n_2}\mathrm{Var}(Y_{ij2}) + 2\sum_{1\le j_1<j_2\le n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2})\right] \\
&= \frac{1}{n_2}\mathrm{Var}(Y_{ij2}) + \frac{n_2-1}{n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}) \\
&= \frac{1}{n_2}\left[E\left(\mathrm{Var}(Y_{ij2}|\Delta_i)\right) + \mathrm{Var}\left(E(Y_{ij2}|\Delta_i)\right)\right] \\
&+ \frac{n_2-1}{n_2}\left[E\left(\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}|\Delta_i)\right) + \mathrm{Cov}\left(E(Y_{ij_1 2}|\Delta_i), E(Y_{ij_2 2}|\Delta_i)\right)\right] \\
&= \frac{1}{n_2} + \mathrm{Var}(\Delta_i)
\end{aligned} \tag{13}$$

Therefore $\mathrm{Var}(T_i) = \frac{1}{n_1} + \frac{1}{n_2} + \mathrm{Var}(\Delta_i)$, and it follows

$$\mathrm{Cov}(\boldsymbol{T}) = \sigma_1^2\boldsymbol{D}_1 + \sigma_2^2\boldsymbol{D}_2 + \sigma_2^2\boldsymbol{C} \tag{14}$$

where $\sigma_1^2 = \mathrm{Var}(\Delta_i)$, $\sigma_2^2 = \mathrm{Var}(\Delta_j)$ and $\sigma_3^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

Under the assumption $Z_i \sim \text{Bernolli}(1, p_g), i \in \boldsymbol{I}_g$ and $Z_i \sim \text{Bernolli}(1, p_b), i \in \boldsymbol{I}_b$. It immediately follows that

$$\mathrm{Var}(\Delta_i) = \begin{cases} p_g\sigma_\delta^2 + p_g(1-p_g)\mu_\delta^2, & \text{if } i \in I_g \\ p_b\sigma_\delta^2 + p_b(1-p_b)\mu_\delta^2, & \text{if } i \in I_b \end{cases} \tag{15}$$

References

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477).

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133.