our main point

1. previous comparative gene set enrichment analysis does not take....
2. 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 *internal 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 $\Delta_i < 0$. By "internal" we mean that every gene has a tendency to be DE with a random DE size Δ, for example, for those stably-expressed genes $P(\Delta = 0) = 1$. We assume that the DE effects are mutually independent for all genes, and that whether gene i is DE or not is determined by a DE "trigger" — the treatment applied to gene i in the experiment. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ be a vector of DE indicators, where for gene i $Z_i = 1$ if there is DE and $Z_i = 0$ otherwise. Furthermore, we let $\delta_i \overset{i.i.d}{\sim} D(\delta)$ be the DE effect size and $E(\delta_i) = \mu_\delta$ and $\text{Var}(\delta_i) = \sigma^2_\delta$. Therefore, the a hierarchical model is imposed on the DE effect Δ_i

$$\Delta_i = Z_i \delta_i, \tag{1}$$

$$Z_i \sim \text{Binom}(1, p_i), \qquad \delta_i \sim D(\delta) \tag{2}$$

We also 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 the control group. In this paper, the between-gene correlations are estimated by the residual sample correlation after the treatment effects are nullified, and treated as known in the enrichment test procedure.

2.2 Hierarchical model for DE effect

3 Results

4 Conclusion

5 AcknowledgeMents

6 Appendix

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,

$$\begin{aligned}
\mathrm{Cov}(T_{i_1}, T_{i_2}) &= E\left[\mathrm{Cov}(T_{i_1}, T_{i_2}|\Delta_{i_1}, \Delta_{i_2})\right] + \mathrm{Cov}\left[E(T_{i_1}|\Delta_{i_1}), E(T_{i_2}|\Delta_{i_2})\right] \\
&= E\left(\frac{1}{n_1}\rho + \frac{1}{n_2}\rho\right) + \mathrm{Cov}(\Delta_{i_1}, \Delta_{i_2}) \\
&= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\rho_{i_1,i_2}
\end{aligned} \tag{3}$$

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

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

$$\begin{aligned}
\mathrm{Var}(\bar{Y}_{i,2}) &= \frac{1}{n_2^2}\left[\sum_{j=1}^{n_2}\mathrm{Var}(Y_{ij2}) + 2\sum_{1\le j_1<j_2\le n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2})\right] \\
&= \frac{1}{n_2}\mathrm{Var}(Y_{ij2}) + \frac{n_2-1}{n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}) \\
&= \frac{1}{n_2}\left[E\left(\mathrm{Var}(Y_{ij2}|\Delta_i)\right) + \mathrm{Var}\left(E(Y_{ij2}|\Delta_i)\right)\right] \\
&+ \frac{n_2-1}{n_2}\left[E\left(\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}|\Delta_i)\right) + \mathrm{Cov}\left(E(Y_{ij_1 2}|\Delta_i), E(Y_{ij_2 2}|\Delta_i)\right)\right] \\
&= \frac{1}{n_2} + \mathrm{Var}(\Delta_i)
\end{aligned} \tag{4}$$

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

$$Thisissomething \tag{5}$$

References

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.