our main point

1. section 1: introduction begins here (suggestions: try writting your introductions last).
   (a) what is my topic (question)?
   (b) why is it important?
   (c) how will I plan to proceed with my
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 a *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 effect, and test the set against the backgroud genes.

We denote by I_t and I_b the gene set being tested and background set (i.e., the genes not in the test set). Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ be a indicator vector of whether or not gene i belongs to the test set and thus $I_t = \{i : x_i = 1\}$ and $I_b = \{i : x_i = 0\}$. We assume that the DE probability is p_t for genes in the test set and p_b for genes in the background set. For gene i, denote $U_i = \bar{Y}_{i.2} - \bar{Y}_{i.1}$ as the difference in mean expression levels between the treatment 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

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

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 $d_i = p_t\sigma_\delta^2 + p_t(1-p_t)\mu_\delta^2$ if $i \in I_t$ 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 DE probability affects both the mean vector in equation (5) and the covariance in equation (6). Under this framework, the test set is not enriched only if the probability of DE in the test set is the same as that in the background set. Therefore, the hypothesis for enrichment testing can be statistically formulated as

$$H_0\text{: } p_t = p_b \overset{\text{def}}{=} p_0 \text{ Versus } H_1\text{: } p_t \neq p_b \tag{7}$$

We can combine equations (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 $\beta_0 = p_b\mu_\delta$, $\beta_1 = (p_t - p_b)\mu_\delta$ and $\mathbf{1}_m$ being a vector of ones. Now the hypothesis testing problem in (7) becomes

$$H_0\text{: } \beta_1 = 0 \text{ Versus } H_1\text{: } \beta_1 \neq 0. \tag{9}$$

Under the null of (9), 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 and $\sigma_1^2 = p_0\sigma_\delta^2 + p_0(1-p_0)\mu_\delta^2$.

(**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,

$$\boldsymbol{\Sigma} = \sigma_1^2\boldsymbol{I}_m + \sigma_2^2\hat{\boldsymbol{C}} \tag{10}$$

The score equations for β_0 and σ_1^2 are

$$\begin{aligned} (\boldsymbol{U} - \beta_0\mathbf{1}_m)^T\boldsymbol{\Sigma}^{-1}\mathbf{1}_m &= 0 \\ (\boldsymbol{U} - \beta_0\mathbf{1}_m)^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{C}}(\boldsymbol{U} - \beta_0\mathbf{1}_m) &= \text{trace}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{C}}) \end{aligned} \tag{11}$$

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

$$T = \frac{\left[\boldsymbol{x}^T(\boldsymbol{U} - \hat{\beta}_0 \mathbf{1}_m)\right]^2}{\left[\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{H})\right] \boldsymbol{\Sigma} \left[\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{H})\right]^T} \tag{12}$$

Under the null, $T \sim \chi^2(1)$.

2.3 Other competitive gene set test

We will compare OurMethod to three existing gene set tests: GSEA (Subramanian et al., 2005) modified from the original R-GSEA script (`http://software.broadinstitute.org/gsea/index.jsp`) to allow single gene set test, two versions of the geneSetTest procedure and two versions of the CAMERA procedure (Wu and Smyth, 2012) in the limma package (Smyth, 2005). The two versions are, respectively, parametric and rank based, and will be denoted by geneSetTest-modt and geneSetTest-ranks for geneSetTest, and by CAMERA-modt and CAMERA-ranks for CAMERA. Because GSEA and OurMethod do not support linear models, the implementations are restricted to two-group comparisons.

All of the three tests use local genewise statistics as observations to conduct global tests comparing genes in the test set to the background genes. However, they may differ either in terms of the local statistics used to compare factors of interest (e.g. treatment vs. control) at the gene level, or in terms of the global statistics used to summarize the significance of the test set compared to the background set. For GSEA, the local statistics are the rankings of genes according to a ranking metric (e.g. signal-to-noise ratio, t-statistic), then based on the rankings an enrichment score for the test set is calculated, and the significance of the enrichment score is determined by randomly permuting the sample labels. Both the parametric geneSetTest-modt and CAMERA-modt use local statistics (e.g., the moderated t-statistics (Smyth, 2004)), and determine whether the mean of the local statistics is significantly different for genes in the test set versus genes in the background set. The difference is how they evaluate the global statistics: geneSetTest-modt evaluates p-values by comparing the observed mean of the local statistics in the test set, to those (??? is it clear?) obtained by randomly permuting the gene labels; CAMERA-modt uses a t-statistic that allows the local statistics in the test set to be correlated by first estimating a variance inflation factor, and then incorporating it into the t-statistic to adjust for between-gene correlation. geneSetTest-rank and CAMERA-rank conduct a Wilcoxon-Mann-Whitney rank sum test, and they amount to, respectively, geneSetTest-modt and CAMERA-modt in that they compare the rankings instead of the local statistics themselves for genes in the test set to those for genes in the background set.

(other methods such as sigPathway or PAGE will be mentioned in the introduction part.)

3 Examples and Numerical Results

3.1 Simulations

In this section, we present results from type I error and power simulations under a range of between-gene correlation structures.

The simulation setup is as follows: first, we simulate an entire gene set containing $m = 500$ genes, from which we sample $m_1 = 100$ genes to represent those in the test set, and the remaining $m_2 = 400$ genes those in the background set; second, for gene $i = 1, \ldots, m$, we simulate the DE effect size δ_i from $N(2, 1)$ and DE indicators Z_i from Binom$(1, p_i)$, where $p_i = p_t$ if gene i belongs to the test set and $p_i = p_b$ otherwise; third, we set the "true" mean expression values $\mu_1 = \mathbf{0}_m$ and $\boldsymbol{\mu}_2 = \boldsymbol{\Delta}$, respectively, for the control and treatment groups; fourth, n_1 samples are simulated from MVN$(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for the control group and n_2 samples from MVN$(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ where the covariance $\boldsymbol{\Sigma}$ may represent one of the following cases:

(a0): the genes are independent of each other.

(a): only genes in the test set are correlated, with exchangeable correlation structure, that is, $\text{Cor}(Y_{i_1}, Y_{i_2}) = \rho$ for $\forall i_1, i_2 \in I_t$.

(c): all genes are correlated, with exchangeable correlation structure, that is, $\text{Cor}(Y_{i_1}, Y_{i_2}) = \rho$ for $\forall i_1, i_2 \in I$.

(e): genes are correlated within the test set and within the background set, but any two genes, one from each set, are independent. That is, the correlation structure is block diagonal, with $\text{Cor}(Y_{i_1}, Y_{i_2}) = \rho_1$ for $i_1, i_2 \in I_t$, $\text{Cor}(Y_{i_3}, Y_{i_4}) = \rho_2$ for $i_3, i_4 \in I_b$, and $\text{Cor}(Y_{i_5}, Y_{i_6}) = 0$ for $\forall\ i_5 \in I_t, \forall\ i_6 \in I_b$.

(f): all genes are correlated, but the correlation between two genes depend on whether they belong to the test set or not. Specifically, $\text{Cor}(Y_{i_1}, Y_{i_2}) = \rho_1$ for $i_1, i_2 \in I_t$, $\text{Cor}(Y_{i_3}, Y_{i_4}) = \rho_2$, for $i_3, i_4 \in I_b$, and $\text{Cor}(Y_{i_5}, Y_{i_6}) = \rho_3$ for $\forall\ i_5 \in I_t, \forall\ i_6 \in I_b$.

(g): genes are correlated in the same way as those from a real data.

3.1.1 Type I error simulations

In (the paper to be finished), we have shown that the sample correlation is not equivalent to test statistics correlation, unless the null is true (i.e., none of the genes is DE). blablabla...

For the above simulation setup, the test set is not enriched if DE probabilities are the same for the genes in the test set and for those in the background set (i.e., $p_t = p_b = p_0$). In this section, we evaluate the type I error rates by simulations on two scenarios: case 1) none of the genes is DE with $p_0 = 0$ and, case 2) the genes in the test set and in the background set are DE with $p_0 = 0.2$.

Figure 1 shows the

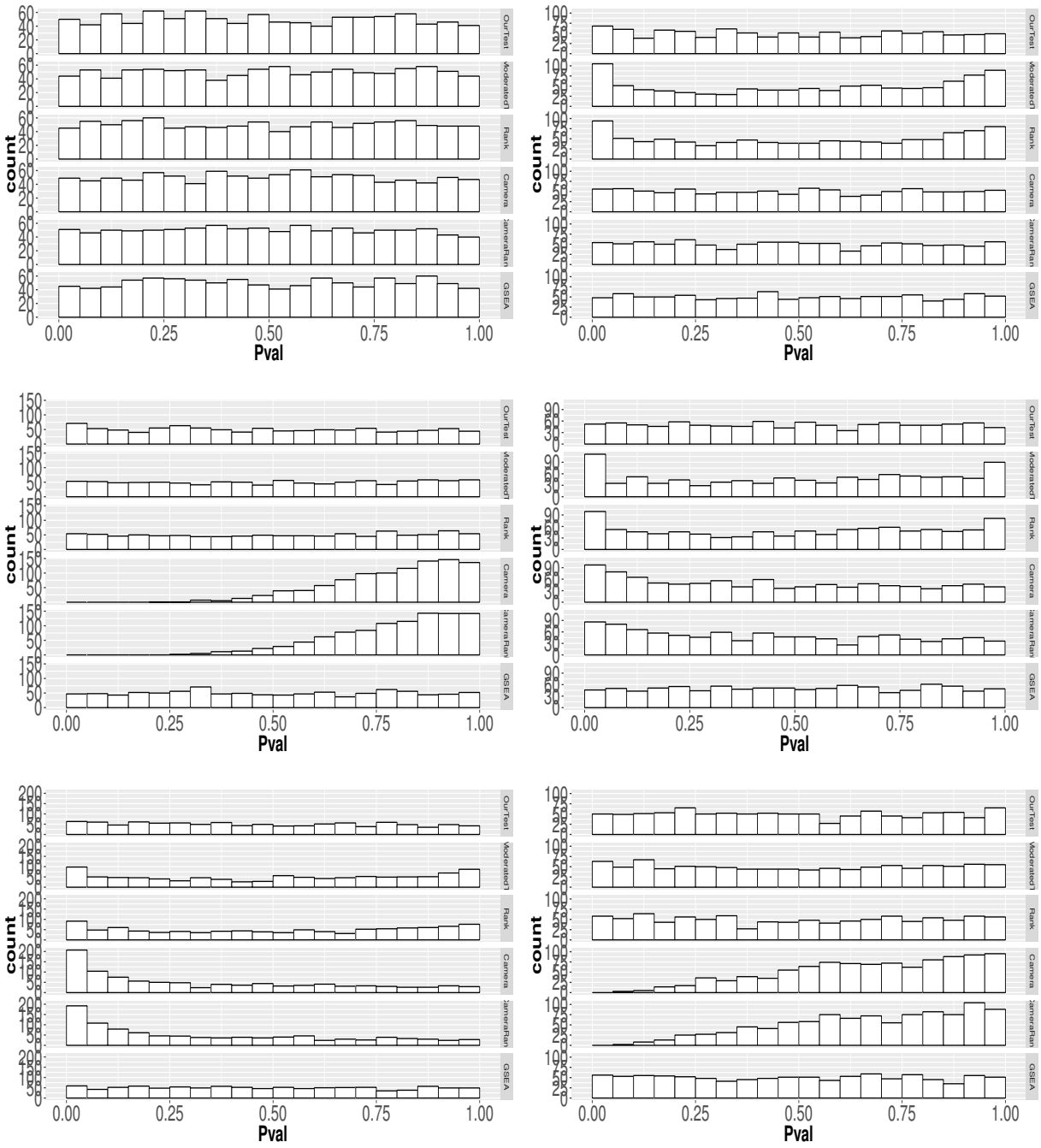
Table 1: Type I error rate of gene set tests for correlated expression values

Method	Norminal p-values				Norminal p-values			
	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
	(a0)				(a)			
OurTest	0.012	0.066	0.124	0.241	0.011	0.049	0.097	0.198
geneSetTest-modt	0.012	0.058	0.109	0.215	0.006	0.052	0.099	0.219
geneSetTest-rank	0.012	0.055	0.110	0.214	0.018	0.060	0.116	0.202
CAMERA	0.006	0.059	0.135	0.217	0.000	0.000	0.000	0.006
CAMERA-Rank	0.003	0.054	0.112	0.229	0.001	0.009	0.023	0.079
GSEA	0.989	0.995	0.997	0.997	0.228	0.608	0.794	0.927
	(c)				(e)			
OurTest	0.007	0.052	0.103	0.202	0.008	0.056	0.093	0.207
geneSetTest-modt	0.007	0.051	0.098	0.187	0.015	0.058	0.106	0.217
geneSetTest-rank	0.006	0.050	0.106	0.190	0.024	0.082	0.138	0.225
CAMERA	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.009
CAMERA-Rank	0.000	0.000	0.000	0.000	0.000	0.020	0.054	0.127
GSEA	0.942	0.984	0.992	0.995	0.108	0.469	0.731	0.899
	(f)				(g)			
OurTest	0.012	0.050	0.098	0.226	0.007	0.059	0.116	0.211
geneSetTest-modt	0.010	0.071	0.115	0.213	0.010	0.061	0.104	0.218
geneSetTest-rank	0.007	0.061	0.112	0.221	0.014	0.072	0.142	0.247
CAMERA	0.000	0.000	0.004	0.019	0.000	0.000	0.001	0.006
CAMERA-Rank	0.006	0.041	0.096	0.210	0.000	0.000	0.002	0.009
GSEA	0.015	0.188	0.434	0.770	0.943	0.975	0.983	0.992

3.1.2 Power simulation

this is two

Figure 1: Type I error rates for gene set tests, p-value distribution for case (a0) - (g) from left to right, from top to bottom, NO gene is DE



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 note that

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

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"),

$$\begin{aligned}
\mathrm{Cov}(T_{i_1}, T_{i_2}) &= E\left[\mathrm{Cov}(T_{i_1}, T_{i_2}|\Delta_{i_1}, \Delta_{i_2})\right] + \mathrm{Cov}\left[E(T_{i_1}|\Delta_{i_1}), E(T_{i_2}|\Delta_{i_2})\right] \\
&= E\left(\frac{1}{n_1}\rho_{i_1,i_2} + \frac{1}{n_2}\rho_{i_1,i_2}\right) + \mathrm{Cov}(\Delta_{i_1}, \Delta_{i_2}) \\
&= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\rho_{i_1,i_2}
\end{aligned} \tag{13}$$

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

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

$$\begin{aligned}
\mathrm{Var}(\bar{Y}_{i,2}) &= \frac{1}{n_2^2}\left[\sum_{j=1}^{n_2}\mathrm{Var}(Y_{ij2}) + 2\sum_{1\le j_1<j_2\le n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2})\right] \\
&= \frac{1}{n_2}\mathrm{Var}(Y_{ij2}) + \frac{n_2-1}{n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}) \\
&= \frac{1}{n_2}\left[E\left(\mathrm{Var}(Y_{ij2}|\Delta_i)\right) + \mathrm{Var}\left(E(Y_{ij2}|\Delta_i)\right)\right] \\
&+ \frac{n_2-1}{n_2}\left[E\left(\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}|\Delta_i)\right) + \mathrm{Cov}\left(E(Y_{ij_1 2}|\Delta_i), E(Y_{ij_2 2}|\Delta_i)\right)\right] \\
&= \frac{1}{n_2} + \mathrm{Var}(\Delta_i)
\end{aligned} \tag{14}$$

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

$$\mathrm{Cov}(\boldsymbol{T}) = \boldsymbol{D} + \sigma_2^2\boldsymbol{C} \tag{15}$$

where $\boldsymbol{D}$ is a diagonal matrix with $\mathrm{Var}(\Delta_i) = p_i\sigma_\delta^2 + p_i(1-p_i)\mu_\delta^2$ as its ith diagonal element, and $\sigma_3^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

References

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

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3, Article3.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

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.