S592 Literature Report 3

CAMERA: A competetive gene set test accounting for inter-gene correlation

Nick Sun

March 10, 2019

Background and Main Contributions

Gene set tests are statistical procedures which seek to determine if, given differential expression data, certain gene annotation categories are enriched. Gene set tests are particularly useful when we are interested not in a single gene, but a single functional group of annotate genes. These groups are gene pathways or represent key biological processes.

Gene set tests are divided into two subcategories: self-contained tests and competitive tests. Competitive tests are the far more popular type and are better suited to discovering the most important biological processes in pathways analysis. Other tests such as Gene Set Enrichement Analysis fall into this category, however, these existing tests do not incorporate correlation between genes in the test set of interest. Failing to incorporate this correlation leads to inflated false discovery rate. Additionally, these tests also find their p-values through permutation based methods which is computationally intensive and infeasible for data sets with low numbers of biological replicates.

The novel proposal in this paper is called CAMERA or Correlation Adjusted Mean Rank Gene Set test. The principle innovation is estimating variance inflation factors from the differential expression data and using this estimate to find an estimated correlation $\hat{\rho}$.

Methods

The novelty of CAMERA revolves around estimating ρ from the data. In order to do so, we first need the variance inflation factor (VIF) of the data which contributes to the overall variance of m genewise statistics.

The formula for this estimated VIF is given:

$$\hat{\text{VIF}} = \frac{m}{d}\sum_{k=1}^{d} u_k^2$$

where u are the column means from a matrix $\mathbf{U}$ which is found using QR decomposition on the design matrix $\mathbf{X}$. From this estimated VIF, we can then solve the following equation:

$$\hat{\text{VIF}} = 1 + (m-1)\hat{\rho}$$

to get our data estimate for ρ.

This ρ estimate is then incorporated into two modified tests which are based on the t-test and the Wilcoxon-Mann-Whitney test. The modifications on these tests are designed to incorporate this correlation factor. In practice, researchers can decide whether to implement the extended t-test or the extended Mann-Whitney test depending upon their data. The utilization of these tests is computationally beneficial since these test statistics all have an asymptotic distribution by which we can compute p-values. This eliminates the need for resource intensive permuation p-values.

Examples

The authors provide several examples of CAMERA in action in order to demonstrate the exactness and power of the test under different conditions. Using simulated data under a null hypothesis being true, the researchers were able to demonstrate that both the CAMERA modified t-test and Mann-Whitney test maintained exactness with correlate genes. Other gene set tests such as PAGE, geneSetTest, and sigPathway by contrast had very high type I error rates and a very skewed distribution of p-values clustering on the lower end of the unit interval.

The researchers were also able to demonstrate that CAMERA has good power in a variety of different circumstances while minimizing type I error. Again using simulated data, both the modified t-test and Mann-Whitney test had decent power for different values of intergene correlation, percentages of differentially expressed genes, and different values of log fold change. The modified t-test for almost all scenarios had higher power than the Mann-Whitney test, but it would have also been helpful to see the power performance of the other existing tests in the table presented in the paper.

Caveats and critical remarks

The authors did not provide as much power calculations and values as I would have liked to see. In particular, it was hard to judge whether CAMERA tests retain good power when the other older tests did not have their power calculations shown. It's a shame since this would have allowed readers to better judge the power trade-offs that might be associated with using a CAMERA test compared to some older methods and mars an otherwise really straightforward paper.

Additionally, a methodological limitation I can see is that there is only one ρ value being incorporated into these modified CAMERA tests. I am curious how well that captures genetic correlation structures in practice, since I imagine there might be gene that are upregulated and downregulated with each other even in the same test set! It would be interesting to see if there was a way to further extend this work so we could capture more of the correlation in these tests.