Document S1: Appendix

Differential gene expression score

For each gene, the differential expression score can be computed by a variety of statistics. The five statistics we provide in the software are Signal2Noise, tTest, Ratio_of_Classes, Diff_of_Classes, and log2_Ratio_of_Classes.

Consider two phenotype classes, C_1 and C_2:

1, Signal2Noise is the difference of the class means scaled by the standard deviation

$$\frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$$

where (μ_1, μ_2) and (σ_1, σ_2) are the means and standard deviations of a gene's expression values in classes C_1 and C_2, respectively. The absolute magnitude of the statistic indicates the strength of the correlation between the gene expression profile and the phenotype, and the sign indicates the direction of this correlation.

2, tTest is the difference of the class means scaled by the standard deviation and number of samples

$$\frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where (μ_1, μ_2) and (σ_1, σ_2) are the means and standard deviations of a gene's expression values in classes C_1 and C_2, respectively. (n_1, n_2) are the number of samples in classes C_1 and C_2. The absolute magnitude of the statistic indicates the strength of the correlation between the gene expression profile and the phenotype, and the sign indicates the direction of this correlation.

3, Ratio_of_Classes is the ratio of the class means

$$\frac{\mu_1}{\mu_2}$$

where (μ_1, μ_2) are the means of a gene's expression values in classes C_1 and C_2, respectively. This statistic is used to calculate fold change for natural scale data. The absolute magnitude of the statistic indicates the strength of the correlation between the gene expression profile and the phenotype.

4, Diff_of_Classes is the difference of the class means

$$\mu_1 - \mu_2$$

where (μ_1, μ_2) are the means of a gene's expression values in classes C_1 and C_2, respectively. This statistic is used to calculate fold change for log scale data. The absolute magnitude of the statistic indicates the strength of the correlation between the gene expression profile and the phenotype, and the sign indicates the direction of this correlation.

5, log2_Ratio_of_Classes is the log2 ratio of the class means

$$\log_2\left(\frac{\mu_1}{\mu_2}\right)$$

where (μ_1, μ_2) are the means of a gene's expression values in classes C_1 and C_2, respectively. This statistic is used to calculate fold change for natural scale data. The absolute magnitude of the statistic indicates the strength of the correlation between the gene expression profile and the phenotype. In this case negative values are truncated to 0 since they are not biologically meaningful.

Single-SNP association score

Five different methods to calculate single-SNP association scores are provided in our software: a genotype-based chi-square statistic, an allele-based chi-square statistic, a statistic based on frequency differences in major/minor alleles in the two classes, and two statistics extended from genotype-based and allele-based chi-square statistics.

1, ChiSquare_Allele is an allele-based chi-square score

Suppose alleles for a bi-allelic SNP in the SNP dataset are encoded as A and B, or alternatively as groups G_1 and G_2. We first define two scaling factors, K_1 and K_2, to adjust for unequal sample sizes between class C_1 and C_2:

$$K_1 = \sqrt{\frac{\sum_{i=1}^{2} S_i}{\sum_{i=1}^{2} R_i}}, \qquad K_2 = 1 \Big/ K_1,$$

where R_i is the observed counts for group G_i in class C_1, and S_i is the observed counts for group G_i in class C_2.

The allele-based chi-square score is then computed as

$$\chi^2 = \sum_{i=1}^{2} \frac{(K_1 R_i - K_2 S_i)^2}{R_i + S_i}.$$

The test statistic value represents the degree to which the SNP is associated with phenotypic class distinction.

2, ChiSquare_Geno is a genotype-based chi-square score

Suppose genotypes for a bi-allelic SNP in the SNP dataset are encoded as AA, AB, and BB, or alternatively as groups G_1, G_2, and G_3. We first define two scaling factors, K_1 and K_2, to adjust for unequal sample sizes between class C_1 and C_2:

$$K_1 = \sqrt{\frac{\sum_{i=1}^{3} S_i}{\sum_{i=1}^{3} R_i}}, \qquad K_2 = 1 \big/ K_1,$$

where R_i is the observed counts for group G_i in class C_1, and S_i is the observed counts for group G_i in class C_2.

The genotype-based chi-square score is then computed as

$$\chi^2 = \sum_{i=1}^{3} \frac{(K_1 R_i - K_2 S_i)^2}{R_i + S_i}.$$

The test statistic value represents the degree to which the SNP is associated with phenotypic class distinction.

3, Diff_of_Alleles is the absolute value of frequency difference of the major/minor allele in two classes

$$| f_1 - f_2 |$$

where (f_1, f_2) is the frequency of the major/minor allele in classes C_1 and C_2, respectively. The test statistic value represents the degree to which the SNP is associated with phenotypic class distinction.

4, tTest_Allele is an allele-based score derived from **ChiSquare_Allele**

Suppose alleles for a bi-allelic SNP in the SNP dataset are encoded as A and B, or alternatively as groups G_1 and G_2. We first define two scaling factors, K_1 and K_2, to adjust for unequal sample sizes between class C_1 and C_2:

$$K_1 = \sqrt{\frac{\sum_{i=1}^{2} S_i}{\sum_{i=1}^{2} R_i}}, \qquad K_2 = 1 \big/ K_1,$$

where R_i is the observed counts for group G_i in class C_1, and S_i is the observed counts for group G_i in class C_2.

The **tTest_Allele** is then computed as

$$t = \sum_{i=1}^{2} \frac{| K_1 R_i - K_2 S_i |}{R_i + S_i}.$$

The test statistic value represents the degree to which the SNP is associated with phenotypic class distinction.

5, tTest_Geno is a genotype-based score derived from **ChiSquare_Geno**

Suppose genotypes for a bi-allelic SNP in the SNP dataset are encoded as AA, AB, and BB, or alternatively as groups G_1, G_2, and G_3. We first define two scaling factors, K_1 and K_2, to adjust for unequal sample sizes between class C_1 and C_2:

$$K_1 = \sqrt{\frac{\sum_{i=1}^{3} S_i}{\sum_{i=1}^{3} R_i}}, \qquad K_2 = 1 \big/ K_1,$$

where R_i is the observed counts for group G_i in class C_1, and S_i is the observed counts for group G_i in class C_2.

The **tTest_Geno** is then computed as

$$t = \sum_{i=1}^{3} \frac{| K_1 R_i - K_2 S_i |}{R_i + S_i}.$$

The test statistic value represents the degree to which the SNP is associated with phenotypic class distinction.