STAT 530 Bioinformatics: Homework 6

Due Apr 24, 2017

For problems using R, turn in your answers in the form of a compiled R notebook PDF.

Problem 1 (5 points)

Recall the honeybee RNA-seq data you downloaded in homework 5. You can use STAR to align these reads. The resulting log file shows that only 72.79% of the reads were aligned to the honeybee genome. Come up with a plausible biological reason for why this is happening.

Problem 2

We will analyze the *Drosophila* data from Dr. Saul's lecture. We will use the count files generated by Dr. Saul, which are available on the course website as `dm_counts.zip`. Use edgeR to perform the following analyses. Consider only genes that are expressed above 1 cpm in at least 3 samples. Recompute the total library size after removing genes. Use TMM normalization.

Part (a) (5 points)

Let y_{gi} be the read count from the gth gene from the ith subject. Fit the saturated two-way ANOVA model

$$\log E(y_{gi} \mid sex_i, mating_i) = \beta_0 + \beta_1 I(sex_i = male) + \beta_2 I(mating_i = polygamous) + \beta_3 I(sex_i = male, mating_i = polygamous) \quad (1)$$

using tagwise dispersion. Report the estimated coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ for genes FBgn0000018, FBgn0000042, and FBgn0025712.

Part (b) (10 points)

Use the results from part (a) to identify:

- Male-biased genes: genes that are more highly expressed in males than in females among polygamous flies.
- Female-biased genes: genes that are more highly expressed in females than in males among polygamous flies.
- Unbiased genes: genes that are not differentially expressed between males and females among polygamous flies.

Use an FDR of 0.1 to distinguish significant from insignificant differential expression. How many genes are in each category? Report all of the code you used to answer this question.

Part (c) (5 points)

Does the behavior of the male- and female-biased genes in male flies support the conclusions of this paper? How do you know? Provide all of your code.

Problem 3 (5 points)

Consider the normal means problem $X_i \sim N(\mu_i, 1), i = 1, \ldots, n$. How does the James-Stein estimator for the μ_i compare to the MLE? Answer this question using simulation.

Let $n = 1000$ and use R to fill in the following table.

Table 1: Average error $n^{-1} \sum_i (\mu_i - \hat{\mu}_i)^2$ over 200 simulations

	MLE	James-Stein
`mu=rep(0,length=n)`		
`mu=seq(0,5,length=n)`		

Which method is more accurate?