Santhanakrishnan Ramani

SID: 105720585

**Bioinformatics & Genomics**

**Assignment – 4**

1. a) The lpt25 gene must be near the 5’ end, while using the polyT primer it will start the reverse transcription of

mRNA to cDNA from the polyA tail and the probability that the reverse transcriptase enzyme not reaching the

5’ end is more, so the process could have stopped after 250 reads in the lpt25 gene. Whereas, the random

priming has higher probability of converting the 5’ end of mRNA to cDNA, and would have got 45000 reads

mapping to lpt25.

b) There shouldn’t be any difference while using the polyT primer, because it is going to start from the same place, whereas there could be effect on the fold change of other genes while using random primer.

1. a) Sensitivity = True Positive / (True Positive + False Negative) = 125 / (125 + 9475) = 0.0130

b) Specificity = True Negative / (True Negative + False Positive) = 375 / (375 + 25) = 0.9375

c) Positive predictive value = True Positive / (True Positive + False Positive) = 125 / (125 + 25) = 0.8333

3. a) Gene A which changes 10-fold between the two conditions will have more variability than expected and a higher p-value, so statistically it won’t be of any use. Whereas Gene B which changes 1.2-fold is very close to the expected value and have a lower p-value making it statistically significant. The significance can be inferred from the volcano plots which plots the negative log of the p-value (significance) versus log of fold- change on the y- and x-axes, respectively results in data points with low p-values (highly significant) appearing toward the top of the plot.

b) Window of length 10 kb:

Pros:

Since we are sliding the window by 10 kb, it takes into consideration all the factors affecting the differential expression of a gene, and will help us identify the regions that are correlated.

Cons:

It is time consuming, as we need to look at nearly 3 x 106 windows and regions that will be of no interest. Since the window length is 10kb, the max base pairs looked will be 10000 which might give false predictions as it might be able to consider the whole gene at a time.

20,000 annotated protein coding genes:

Pros:

Since the protein coding genes are annotated it will easier to find the statistically significant differential expression regions.

Cons:

Since we looking only at the annotated protein coding regions, the effect of other non-coding regions nearby on differential expression is neglected, which might affect the p-value by going above or below the statistical cutoff and make the significant regions insignificant or vice versa.

4. a)



b)



c)



d) According to the Scoring above scheme, the sequence ACTGAG seems to be a better fit to this motif model.



5.

6. a) Fragment Assignment & Density Deconvolution.

b) The estimate will have very high variance.

c) 98

d) By “No sample is an island”, Pachter means every experiment performed is just a replicate of another, just with parameters changed. So in the computational stand point, if you have done one experiment, the other experiments will mostly be an extension of the previous work, or run with a different parameters.

e) Impute means assign (a value) to something by inference from the value of the products or processes to which it contributes. He says “it will make you queasy” because the statistics won’t look the way it needs to be.

f) RPKM/FPKM are more of a unit than a metric, and they are proportional to the relative abundance (ρ) which is experiment specific. TPM is a better metric.

g) They don’t look or give feedback on the supplement section of the paper, which may have errors.