**UNCERTAINTY QUANTIFICATION**

**Midway Result Report**

**Aman Agarwal (111471742), Aditya Srivastava (111424337), Mahima Parashar (111463071)**

**Separating Transcripts:**

For given samples, poly\_ro and poly\_mo, we performed following method to separate transcripts:

We had ‘quant\_bootstrap.csv’, which is the output file of the bootstrap process. Each column of the file represents a transcript and each subsequent row represents the result of a bootstrap run. So, we calculated mean and standard deviation of each transcript from this file. Next we compared this standard deviation with its count value from the values present in ‘poly\_truth.csv’. This gave us the deviation of the count from the mean value. Finally we segregated the transcript into two class- ‘good’ which had |deviation| <= 2 and ‘faulty’ which had |deviation| > 2.

**Analysis on Properties:**

1. Our first step in this direction is to perform feature reduction. We calculated correlation among the four properties which are Length, Effective Length, NumReads and TPM and found that NumReads and TPM are highly correlated and so are Effective Length and Length. Thus, now the 4 features reduces to 2 features because of the high correlation.
2. Then regression model is used to predict which property highly contributes to a transcript being good or faulty. The features fed in this system are effective length and tpm.
3. The results obtained from the regression model were that TPM contributes more to a transcript being good or faulty.
4. To compare good and faulty transcripts, we plotted scatter graph between their tpm and deviation. Next, we tried to look for any anomalous pattern among faulty transcripts.
5. Similarly, we repeated the step 4 for effective length.

**RESULTS**

Upon analysing, we came to the following conclusions:

**Poly\_mo:**

**TPM:**

1. We saw 13084 out of 14176 (~92.3%) good transcripts had a TPM **below** **10**.
2. And 6421 out of 12713 (~50%) faulty transcripts had a TPM **above 10.**
3. Thus we can conclude that for **TPM > 10**, **85.46%** transcripts are **faulty.**

**Effective Length:**

1. We saw 13121 out of 14176 (~92.55%) good transcripts had EffectiveLength **less than 5000.**
2. And, 11307 out of 12713 (~88.9%) faulty transcripts had EffectiveLength **less than 5000.**
3. Hence, there was no clear range where either of them had a majority.

**Poly\_ro:**

**TPM:**

1. We saw 7917 out of 9097 (~87%) good transcripts had a TPM **below 10.**
2. And, 11140 out of 17792 (~62.61%) faulty transcripts had a TPM **below 10.**
3. Thus we can conclude that for **TPM > 10**, **85.01%** transcripts are **faulty**.

**Effective Length:**

1. We saw 8540 out of 9097 (~93.8%) good transcripts had EffectiveLength **below 5000.**
2. And, 15986 out of 17792 (~89%) faulty transcripts had EffectiveLength **less than 5000.**
3. Hence, we can conclude that for **EffectiveLength > 5000**, **76.4%** are **faulty.**