# Phenotypic Prediction from Transcriptomic Features

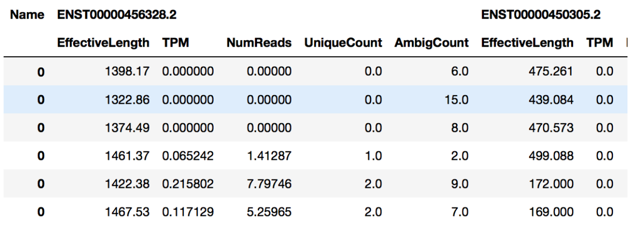
Arjun Mathew Dan, Ajin Oommen Kuriakose, Manu Mathew, Shilpa Mary George

***Abstract*—A central problem in machine learning is identifying a representative set of features from which to construct a classification model for a particular task. We are provided with output from Salmon, an RNA-seq mapping and quantification tool. The various samples come from different phenotypes; types of population in this case. Our goal is to perform feature selection/reduction and to come up with a model that takes the Salmon output and predicts the original label and to achieve the maximum accuracy in classifying the samples using multi-class classification models.**

***Keywords*—*Transcript Quantification, TPM, Prediction, Salmon***

## I. Data Format

We initially created the input data by selecting basic features from the Salmon output provided in the quant.sf. Here there are 369 samples each having around 200k features. Apart from quant.sf, we have considered the auxilary file ”ambig_info.tsv” as well. This file contains information about the number of uniquely-mapping reads and the total number of ambiguously-mapping reads for each transcript. Then we merged these two files and generated a new file ”quant1.sf” with the required features which is then used to train the model (using Merge.sh file).
Then we performed data formatting by flattening the features, such that each row corresponds to a sample having the entire feature list.



| Name | ENST00000456328.2 | | | | | ENST00000450305.2 | |
|---|---|---|---|---|---|---|---|
| | EffectiveLength | TPM | NumReads | UniqueCount | AmbigCount | EffectiveLength | TPM |
| 0 | 1398.17 | 0.000000 | 0.00000 | 0.0 | 6.0 | 475.261 | 0.0 |
| 0 | 1322.86 | 0.000000 | 0.00000 | 0.0 | 15.0 | 439.084 | 0.0 |
| 0 | 1374.49 | 0.000000 | 0.00000 | 0.0 | 8.0 | 470.573 | 0.0 |
| 0 | 1461.37 | 0.065242 | 1.41287 | 1.0 | 2.0 | 499.088 | 0.0 |
| 0 | 1422.38 | 0.215802 | 7.79746 | 2.0 | 9.0 | 172.000 | 0.0 |
| 0 | 1467.53 | 0.117129 | 5.25965 | 2.0 | 7.0 | 169.000 | 0.0 |

Flattened data from ”quant1.sf”

## II. feature selection

Upon analyzing the feature set, we found that there is a co-relation among the features provided ie, between length and effective length, so we omitted the length.
Also the feature set was enhanced using 2 additional properties from ”ambig” file namely ”Unique_count” and ”Ambig_count”.

### A. Dimensionality Reduction

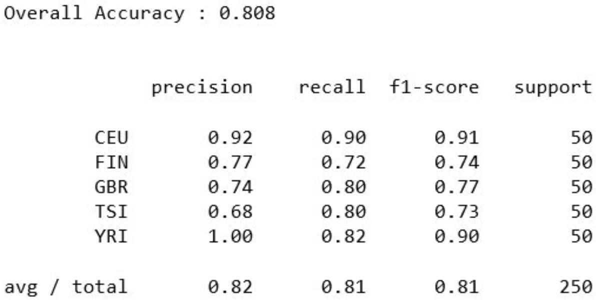
We used Principal Component Analysis(PCA) to change the feature domain and map the input feature list to an enhanced set of features. The data was transformed using the enhanced feature set from PCA.The model was trained using this transformed data set as training input.

## III. Machine Learning Model

From comparing the results obtained from various classification models such as Decision tree, Random Forest and Logistic Regression, we observed better results with Logistic Regression. Also 5-fold cross validation is used to train and test the model.

## IV. Results and Findings

- We observed that removing the length attribute from each of the feature didn’t affect the overall accuracy.
- Adding 2 additional attributes from Ambiginfo file resulted in an improved accuracy. For instance training the model using 300 samples without including attributes from ”Ambiginfo” gave an accuracy of 0.78 in comparison to 250 samples with ”Ambiginfo” attributes included giving an accuracy of **0.81**
- In our experiments Logistic regression gave better accuracy when compared to other classifiers that we tried.
- Equally sub-sampling from all the types gave a slightly better accuracy



```
Overall Accuracy : 0.808


             precision    recall  f1-score   support

        CEU       0.92      0.90      0.91        50
        FIN       0.77      0.72      0.74        50
        GBR       0.74      0.80      0.77        50
        TSI       0.68      0.80      0.73        50
        YRI       1.00      0.82      0.90        50

avg / total       0.82      0.81      0.81       250
```

Classification Report

## V. Other approaches tried

- We randomly took around 10000 features from the set of all features which gave around 0.75 accuracy.
- We tried other classifiers such as Decision Tree, SVM and Random Forest but all of those gave lower precision.

## VI. In pipeline

- Further reduce the feature size by comparing and contrasting the co-relation among the features by considering abundance measures such as TPM and NumReads.
- We also plan to use the information from the equivalence class files to come up with possibly better features.
- We also plan to try other classifiers including Neural networks on the data set.