Machine Learning using Genomics for Cancer Prognosis- SNP Detection in Liquid Biopsy

THESIS

Submitted in partial fulfillment of the requirements of BITS F421T, Thesis

by
Prahalad Atreyaa A
ID No- 2015B4A30823G

Under the supervision of
Dr. Ramesh Hariharan
Strand Life Sciences, CEO
Adjunct Professor, IISc
& Dr. Himadri Mukherjee,
Assistant Professor,
Department of Mathematics,
BITS-Pilani, K. K. Birla Goa Campus.

BITS, PILANI -K K BIRLA GOA CAMPUS

December 2019

ACKNOWLEDGEMENT

I would like to thank Dr. Ramesh Hariharan for supervising my project and giving this opportunity to work on an interdisciplinary topic. I would also like to thank Dr. Vamsi Veeramachaneni and Dr. Shanmukh Katragadda for their timely and valuable inputs. Also, Dr. Swaraj Basu for working closely with me and helping me clear any roadblocks that prop up by connecting me with the right people across different teams. Additionally, I would like to thank Dr. Himadri Mukherjee for his patient guidance as well as supervision.

CERTIFICATE

This is to certify that the thesis entitled, 'Machine Learning using Genomics for Cancer Prognosis- SNP Detection in Liquid Biopsy', is submitted by Prahalad Atreyaa A ID No. 2015B4A30823G in partial fulfilment of the requirements of BITS F421T Thesis embodies the work done by him under my supervision.

Signature of the supervisor

Name

Date_____ Designation

LIST OF SYMBOLS & ABBREVIATIONS USED

1. TNR-True Negative Rate
2. PCR - Polymerase Chain Reaction
3. CIGAR - Compact Idiosyncratic Gapped Alignment Report
4. M - Match
5. I - Insertion relative to reference
6. D - Deletion relative to reference
7. N - skipped region from the reference
8. S - Soft clip, not aligned but still in sam file
9. H - hard clip, not aligned and not in sam file
10. = - sequence match
11. X - sequence mismatch
12. hg19- Human reference genome
13. Germ cells - Cells that become sex cells, also known as gametes or, in humans, sperm and eggs.
14. Somatic cells - Non-reproductive cells.
15. SVR - Supporting Variant Read.
16. NSVR- Non-Supporting Variant Read.
17. SNP- Single Nucleotide Polymorphism
18. FPR-False Positive Rate
19. StrandNGS-Properiatery software used for alignment ans SNP calling.
20. AUC- Area Under Curve.

THESIS ABSTRACT

Thesis Title: Machine Learning using Genomics for Cancer Prognosis- SNP Detection in Liquid Biopsy
Supervisor: Dr. Ramesh Hariharan
Year: 2019
Semester: First
Name of the student: Prahalad Atreyaa A
ID No: 2015B4A30823G
Variation when compared to a reference genome (i.e. hg19) can happen at various different frequencies. For instance, in somatic cells, variations occur less frequently (that is, one in thousand reads supports the variation). The errors made by the Alignment process, PCR process and the Sequencing process could can introduce variations of the same frequency. These errors can be construed as noise. This noise could be mislabelled as cancer, and give rise to erroneous SNP calls. Hence, the prime objective is to develop a machine learning model to differentiate a variation in a read due to cancer from a variation due to noise (generated mainly due to errors in sequencing and/or alignment and/or PCR process) in a query. Cancer analysis using genomic sequence requires accurate identification of somatic and germline variants in sequencing data. Manual intervention to refine the variant calls is required for applying the SNP filters. However, manual variant SNP filter is tedious, expensive in terms of manual efforts, very unstructured, not easily replicable, and highly reliant on read features solely. Hence, we use our ML model (i.e. Random Forest) to automate the process, and consider base specific features to increase accuracy. The final model takes in reads at a pileup columns (i.e. given base position) and assigns a probability score for each base to filter out erroneous bases. And this probability score is used to more accurate SNP calls. Thus, this model labels test data set with high specificity and sensitivity. The model improves on the manual SNP calling and/or aids the same.

Contents

1. Introduction

Cancer development is driven by the culmination of a lot of variations affecting the structure and function of the genome. Genetic variations disrupts normal patterns of gene expression, sometimes leading to the expression of abnormal, constructively active protein. These changes are inheritable at the cellular level, and hence contribute to the cloning of cancer cells.

1.1. History

Genetics in the field of medicine has historically been a very skill intense specialized field, and studies have shown an important connections between SNPs and multitudes of ailment. In due passage of time, there has been an increase in the assessment of genetic risk factors that cause disease(s). Although we are inching closer to achieving personalized medical care (as predicted by the planners of the Human Genome Project), that dream will only be actualized through continued research into all of the many genetic factors that contribute to complex disease susceptibility.

1.2. Relevant Concepts

1.2.1. Read

A read is a sequence of A,G,C,T nucleotides called by the sequencer corresponding to the end of a fragment. The sequence is called out in the direction of growth of the DNA strand, i.e., from $5^{'}$ to $3^{'}$. If the template (reference, in our case, hg19) strand is the forward one, then the read created will align in the negative direction to the reference. Conversely, reverse template strands on sequencing give rise to positively aligning reads.

1.2.2. DNA Sequencing

DNA sequencing, which determines the order of nucleotides in a DNA strand, allows the end user to read the genetic code so they can compare the disease-causing versions of a gene from normal versions of the genes (Fig.1).

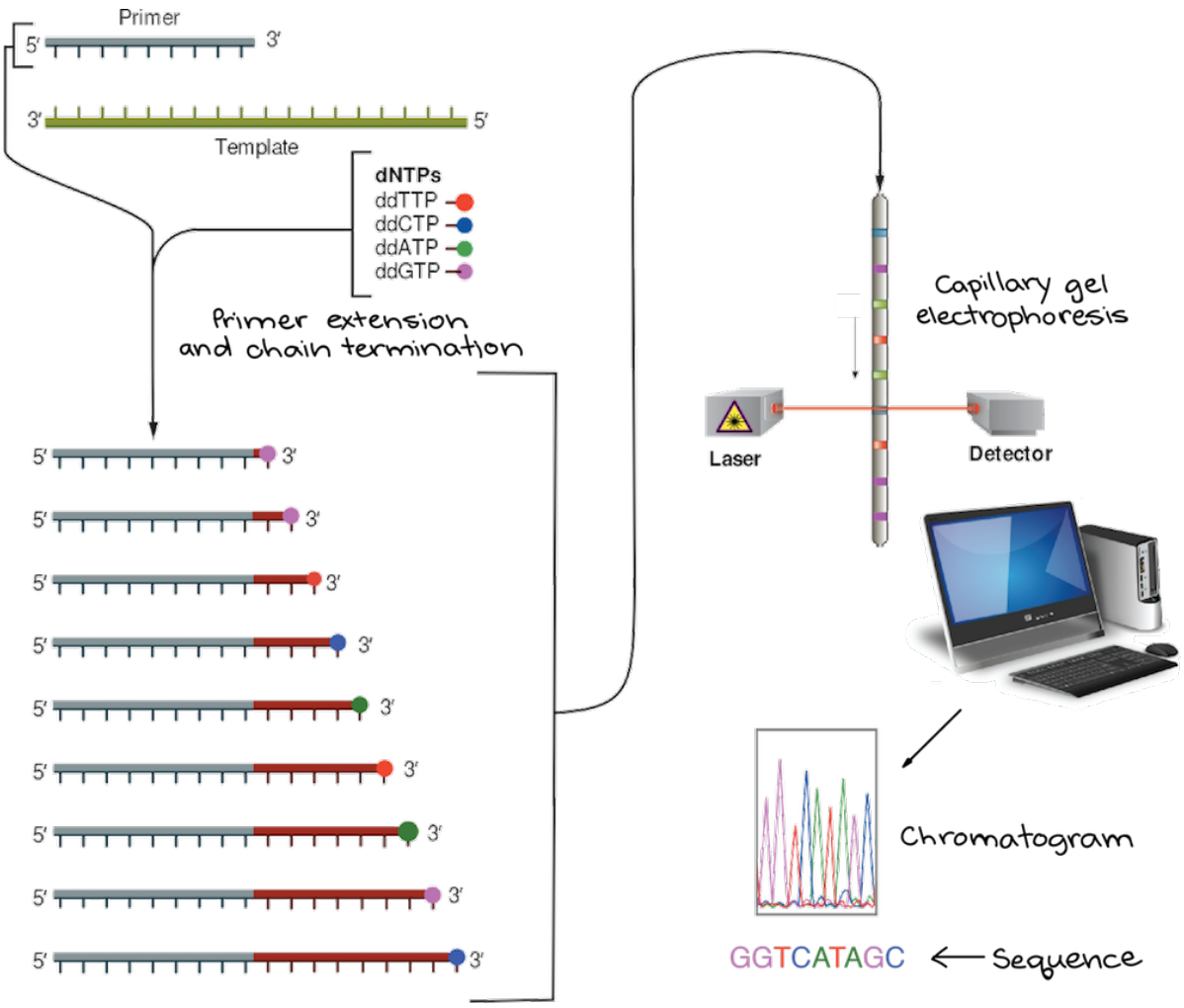

Figure 1: DNA Sequencing Process.

1.2.3. Alignment Quality

The match of a read to a reference sequence can be represented in the form of two (maybe with gaps in between) strings placed one over the other. The following figure (Fig.2) shows the optimal alignment of a read, ATATTAGCC, with a portion of the reference with the sequence ATTAAGGC. The total length occupied by the alignment is 10, and it consists of:

-6 positions where the read and reference characters match exactly.

-a deletion of length one at position 2 of the alignment

-an insertion of length 2 at positions 5,6, and a mismatch at position 9.

The alignment score of a match is defined as the percentage of the alignment length which is occupied by exactly matching characters (60% in the

```
pos :  1234567890
read:  A-TATTAGCC
ref :  ATTA--AGGC
```

Figure 2: Aligned read with respect to a reference.

above example).

1.2.4. Base Quality

Each nucleotide output by the sequencer has a corresponding number accorded to it. This is the quality value and is assigned based on the confidence with which a particular base is called by the base-calling software. If the actual base in the sequence was denoted by B, the one as estimated by the base caller software is denoted by $B^{'}$, the base error is denoted by ϵ, the query length is denoted by n and the corresponding base quality is denoted by Q_B.

$$\epsilon = P[B \neq B^{'}] \tag{1}$$

$$Q_B = -10log_{10}\epsilon \tag{2}$$

And the average base quality of the read is computed as follows:

$$Q_{B_{avg}} = \frac{\sum_{i=1}^{i=n} Q_{B_i}}{n} \tag{3}$$

2. Approach

Pysam is used to extract the data for all the features from the bam file (which is exported from the StrandNGS software post alignment. Please refer to the flowchart given below to get an overview of the features that are included in the model (Fig.3). The cigar string and flag have been parsed to give additional features. The positive controls files contains both somatic and germline variants. These files give us the reference base and the expected variant allele at certain genome positions for select chromosomes, where a known variant has been introduced. The bed file contains all the aligned regions data. This is obtained by using StrandNGS (Version 3.4). The negative controls regions is defined as the regions left after the exclusion of all ignore regions and all those positive control regions that doesn't overlap with the ignore regions. The negative and psotove regions are used to create our training data set by encoding labels. M and SVR are encoded as positive label (class 1)and NSVR as negative label (class 0) for positive regions. And NSVR in negative regions are encoded with a negative label (class 0). Different ML models are tested and hypertuned to find the best parameters. The best model is chosen with optimized parameter set to ensure high specificity and sensitivity.

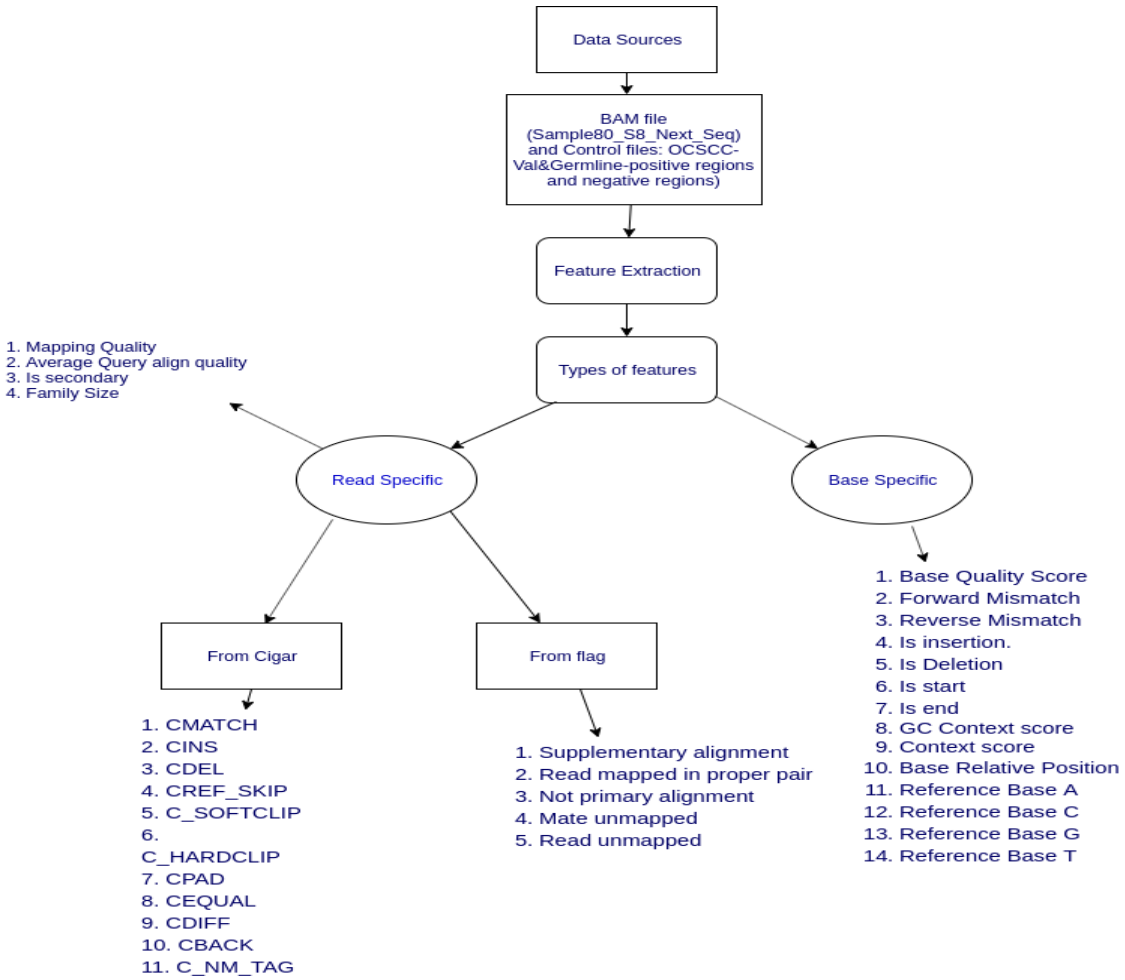

Figure 3: Flowchart of the approach followed to get the features data.

3. Model

The data set obtained is imbalanced. Thus, class weights will be used in the ML model to balance the effect of the positive and negative class data imbalance. The confusion matrix will be made for a default threshold of 0.5 using a customized scoring function. The customized scoring function consists of fpr, tpr, tnr, log loss, and auc score. The hyper-tuning of parameters is done using Grid Search CV to find the best parameters for the decision tree model, the logistic regression model, and the random forest model. Then, the ML model that gives high specificity and high sensitivity is selected with its best parameters.

3.1. Feature Engineering

The features are selected carefully to make the model more accurate and reliable. Features that had a high correlation were dropped so that over fitting doesn't occur. The below given subsections describes some of the important features that are a part of the final model, that is, after dropping few features either due to high correlation or no variation. Additionally, the feature importance chart is used to further eliminate features that are low in feature importance with respect to the model in question. Refer Fig.3 for a comprehensive list of all feature and refer Appendix A for the visualization of all the features. Please use the reference STrandNGS manual for further information on other features.

3.1.1. GC content

This is basically the percentage of G and C for the previous ten bases. This is a base specific feature.

3.1.2. Features extracted from flag

The below given Fig.4 contains all the possible properties that can be extracted from the flag. The below given code snippet is to extract the requisite flag properties using Pysam.

```
def extractFLAG(df_comb_features):
```

```
    def get_binary(x):
        return '{0:012b}'.format(x)

    df_comb_features.FLAG=df_comb_features.FLAG.apply(lambda x:get_binary(x))

    df_comb_features['F_SUPP_ALIGN']=df_comb_features.FLAG.apply(lambda x:int(x[0]))
    df_comb_features['F_PROPER_PAIR']=df_comb_features.FLAG.apply(lambda x:int(x[10]))
    df_comb_features['F_NOT_PRIMARY_ALIGN']=df_comb_features.FLAG.apply(lambda x:int(x
        [3]))
    df_comb_features['F_MATE_UNMAPPED']=df_comb_features.FLAG.apply(lambda x:int(x[8]))
    df_comb_features['F_READ_UNMAPPED']=df_comb_features.FLAG.apply(lambda x:int(x[9]))

    df_comb_features.drop(columns=['FLAG'],axis=1,inplace=True)

    return df_comb_features
```

#	Binary	Decimal	Hexadecimal	Description
1	1	1	0x1	Read paired
2	10	2	0x2	Read mapped in proper pair
3	100	4	0x4	Read unmapped
4	1000	8	0x8	Mate unmapped
5	10000	16	0x10	Read reverse strand
6	100000	32	0x20	Mate reverse strand
7	1000000	64	0x40	First in pair
8	10000000	128	0x80	Second in pair
9	100000000	256	0x100	Not primary alignment
10	1000000000	512	0x200	Read fails platform/vendor quality checks
11	10000000000	1024	0x400	Read is PCR or optical duplicate
12	100000000000	2048	0x800	Supplementary alignment
Sum	000000000000	0	0x0	

Figure 4: Flag properties.

3.1.3. Mapping Quality

It tells us to what extent a read is correctly matched with the genomic coordinates. Mathematically speaking, it is $-10log_{10}$(Probabilitymapping position is incorrect). For example, a mapping quality of 50 = 10 to the power of -5, which is 0.00001, which means there is a 0.001 percent chance that the read is aligned incorrectly. Also, a point to note would be that different mapping programs have subtle differences in the way they calculate the

mapping quality. Hence, the mapping quality values are comparable only if they are generated from the same program.

3.1.4. Context Score

The context score is computed based on the 3 character sequence that occurs before the given base. If there's a mismatch in any 3 character window with the character that's immediately next to the window of length 3, then the score is zero as it's highly unlikely that it is due to sequencing error. For instance, if all three are repeated bases, then it receives the highest score as sequencing error is directly proportional to the number of repeating bases.

3.1.5. Important features

The feature importance is calculated after finalizing the model. Random Forest with n_estimators = 100 and max_depth = 20 is the finalized model based on the comparison of ROC curves in (Fig.5). Random Forest model has the best AUC score, and therefore selected as the final ML model. Each feature is separated and passed to the finalized ML model to plot its ROC curve (Fig.7) to verify the findings in the feature importance chart (Fig.6). This is done to retain only the important features and improve the efficacy of our ML model.

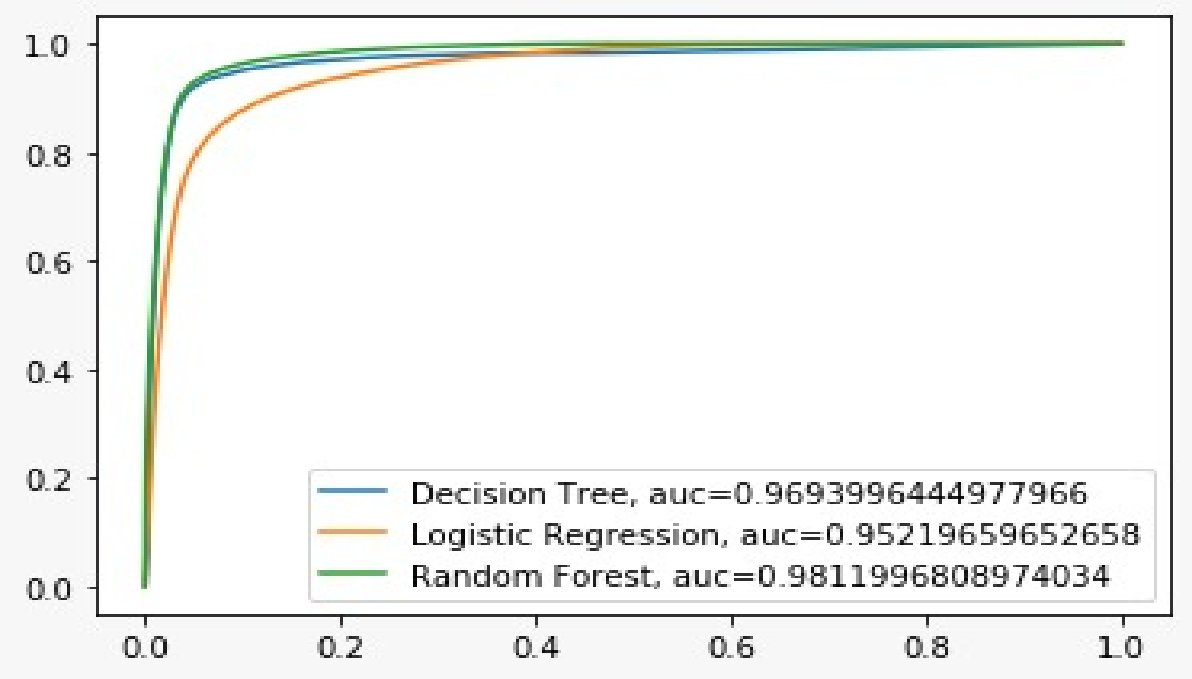

Figure 5: ML Models- ROC Curves

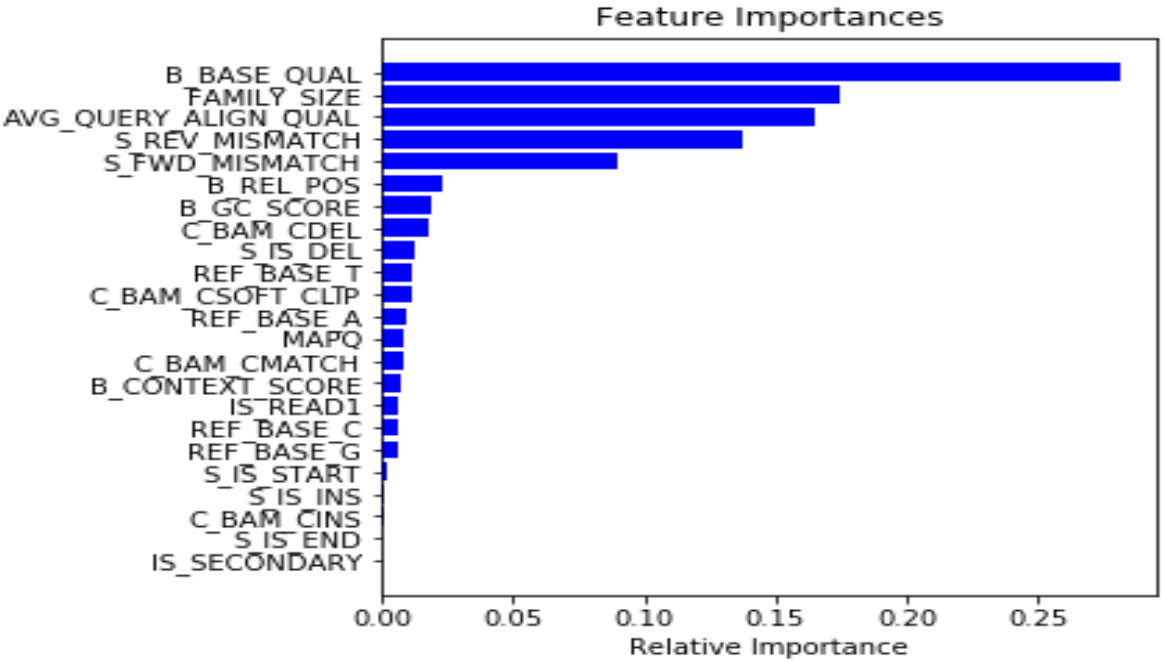

Figure 6: Feature Importance Chart

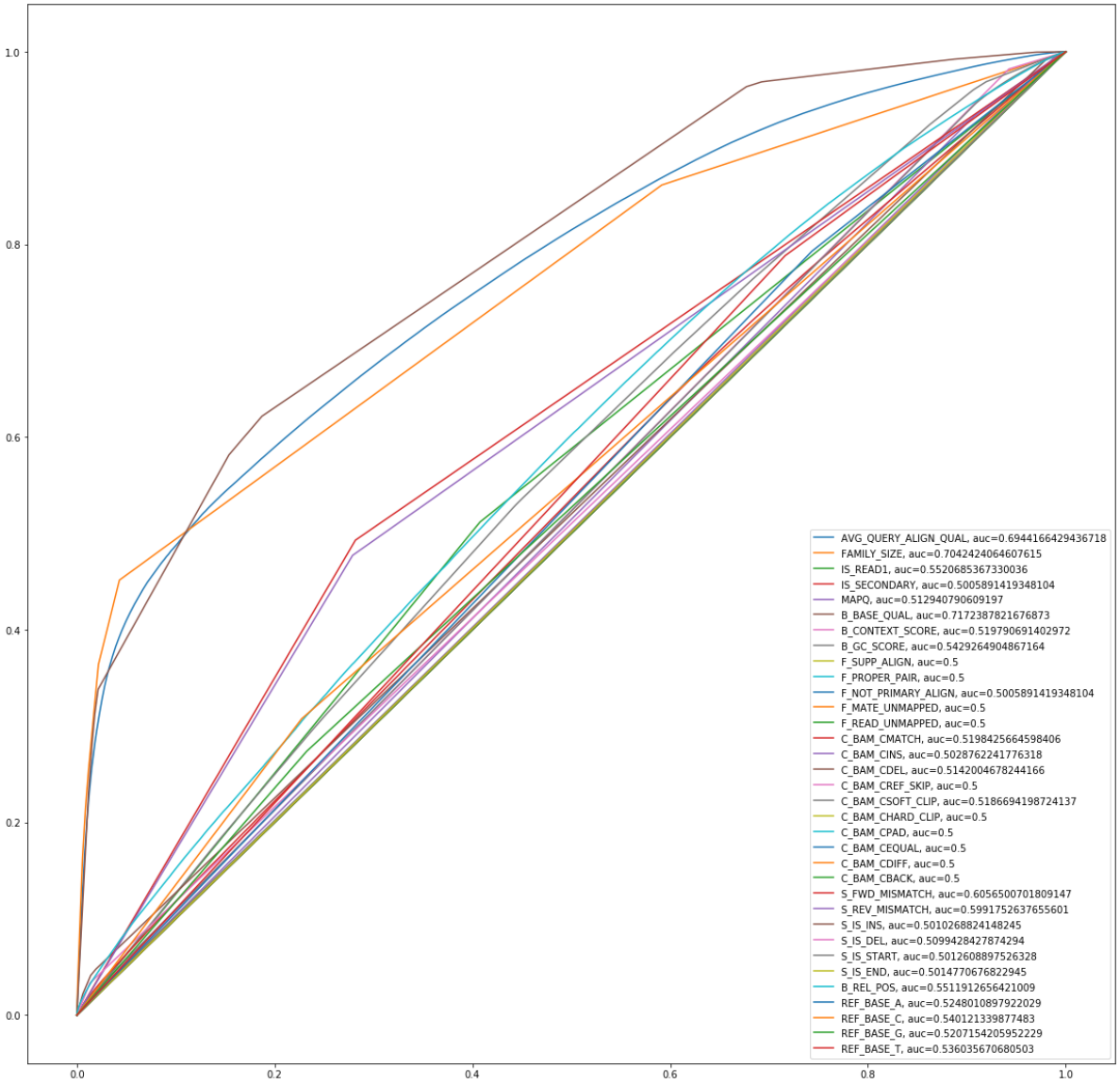

Figure 7: All features ROC.

4. Results

The class labels are imbalanced because they have the following distribution:
Class 1 (M+SVR): 17,41,375
Class 0 (NSVR): 69,99,019
This imbalance is accounted for in the class weights parameter of RF model. The features engineering is done to prevent over-fitting and under-fitting. The feature set is not normalized. Random forest with n_estimaors = 100 and max_depth = 20 is chosen as the ML model based on its performance (i.e., frp, tpr, and auc values). The Random forest model is used to make prediction on the all aligned bam file in the negative regions with a 20 % random sampling rate, and the whole of positive region. The results for the same are as follows for the training and testing:

Training:

FPR = 0.0422108221535341
Accuracy = 0.9475064025957958
AUC_ROC = 0.9911649537918791
TPR = 0.9471539009905384

Testing:

FPR = 0.0481185488570897
Accuracy =
AUC_ROC=0.9789841097021228
TPR = 0.9174014290411444

The below two equations will be used to compute the specificity and accuracy from the custom scorer made from the confusion matrix for the model. Although, under sampling and over sampling were explored. The model gave best results when the imbalance factor is addressed by class weights and not any sampling technique.

A customized scoring function is defined and passed to the Random Forest model for Grid search (hyper-tuning). It will give us different results for various Random forest parameters like max_features, max_depth, etc. From this we get the best parameters.

In our case, the best parameters are obtained at n_estimators = 100, and max_depth = 20.

Accuracy is defined as:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{4}$$

Specificity is defined as:

$$Specificity = TNR = \frac{tn}{tn + fp} \tag{5}$$

In our case, the specificity is around 0.96. The ROC curve (Fig.8) and the Precision-Recall curve is plotted (Fig.9) for the test data result to show the performance of the RF model.

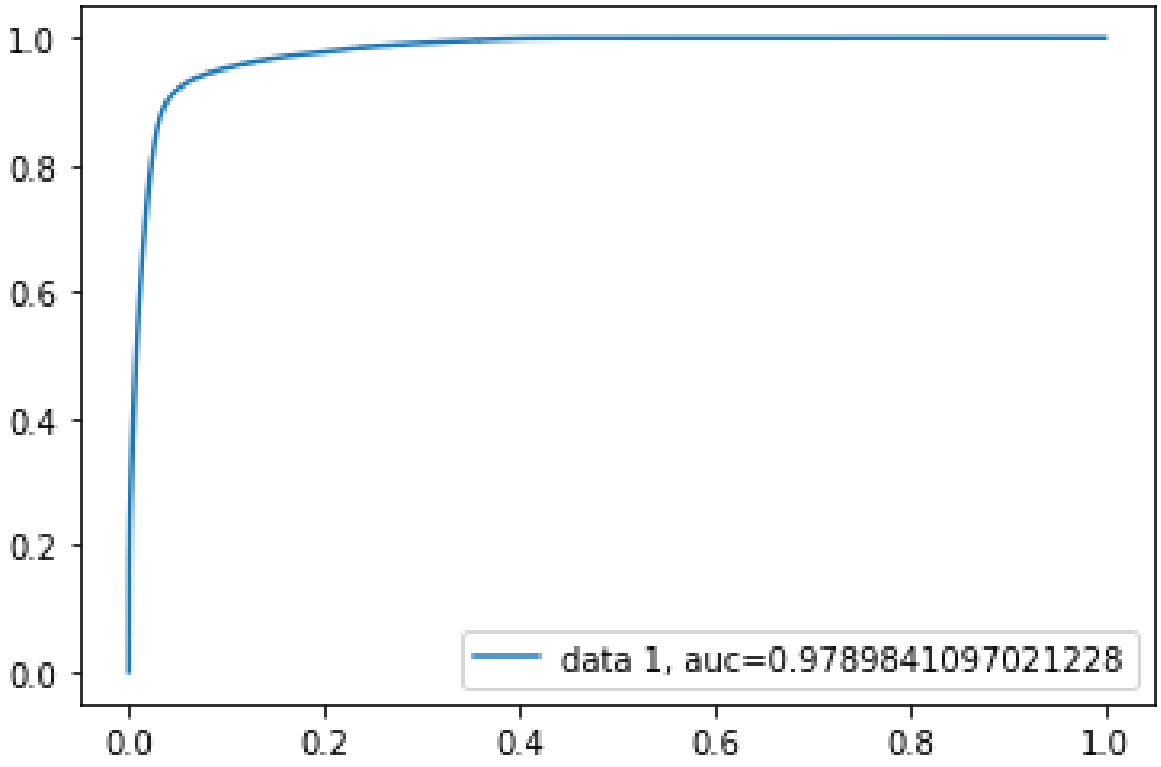

Figure 8: ROC Curve on Test data

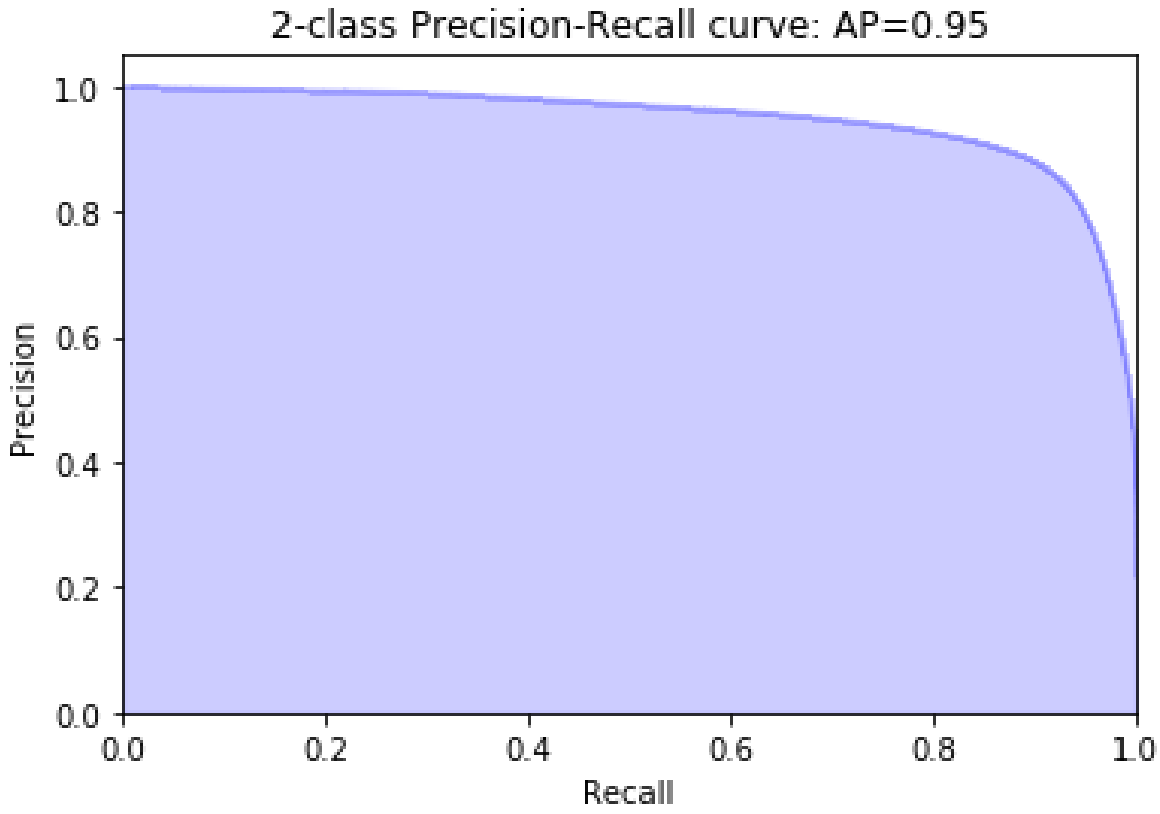

Figure 9: Precision Recall Curve

5. Conclusion

The results on the test data predicts that about 99.99% of the reads passed by StrandNGS will also be passed by the Random Forest model. And about 61.65% of all aligned reads that are passed, are passed only by the RF model and not by StrandNGS. This happens because StrandNGS operates on read level and filters the entire read based on certain specifications of the paramaters. Whereas, the RF model operates on the base level and hence, can retain part of the reads, even if the whole read isn't informative. As it assigns each base a predicted probability score which can be used to classify. RF model yields better results. Hence, RF model eliminates the need for manual intervention and improves accuracy, while meeting the objective of maintaining high specificity (i.e., 0.96) and high sensitivity (i.e., 0.92).

6. Refrences

[1] A journal paper:

Walid Korania, Josh P. Clevengera, Ye Chua and Peggy Ozias-Akins,'Machine Learning as an Effective Method for Identifying True Single Nucleotide Polymorphisms in Polyploid Plants', The Plant Genome, Mar.2019.

[2] A journal paper

M.A. DePristo, E. Banks, R.E. Poplin, and K.V. Garimella,'A framework for variation discovery and genotyping using next-generation DNA sequencing data', NCBI, Apr.2011.

[3] A journal paper

Li MM et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J. Mol. Diagn 19, 4–23 (2017).

[4] A Reference Manual

StrandNGS Reference Manual.

A. Appendix A

This appendix consists of the visualization of both numeric and binary features grouped by labels, and normalized for better segregation.

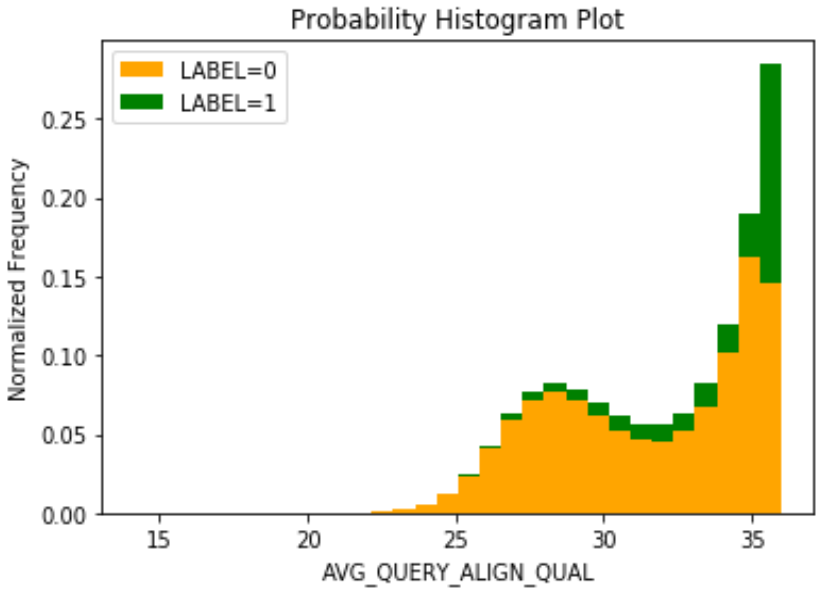

Figure 10: Average Query Alignment Quality Score.

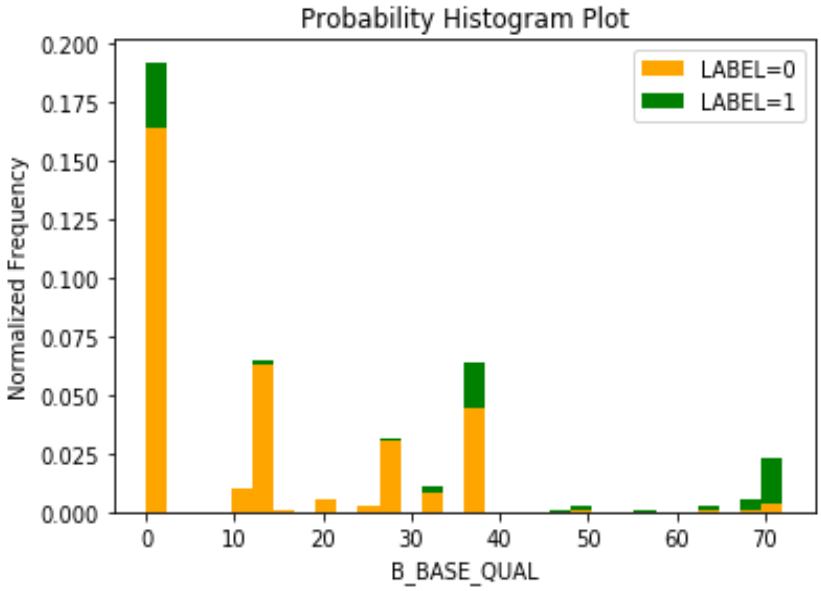

Figure 11: Base Quality Score.

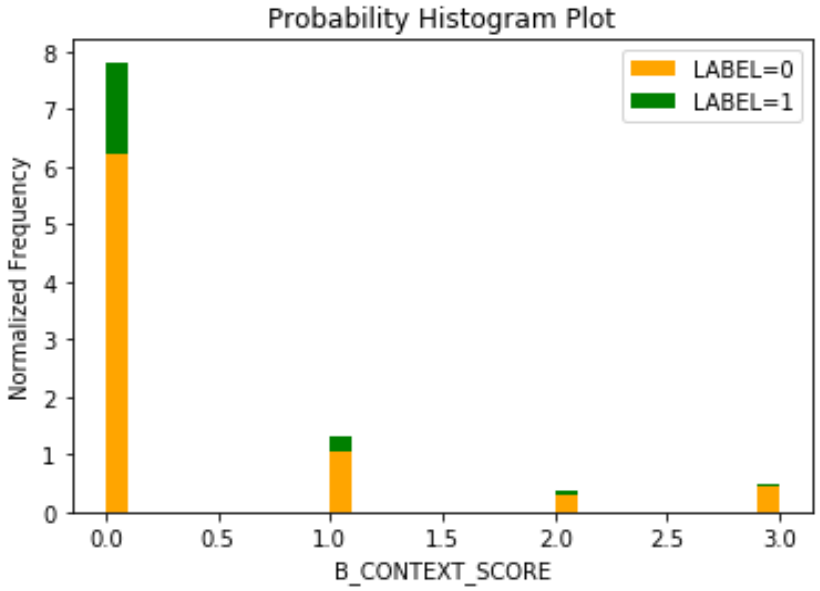

Figure 12: Base Context Score.

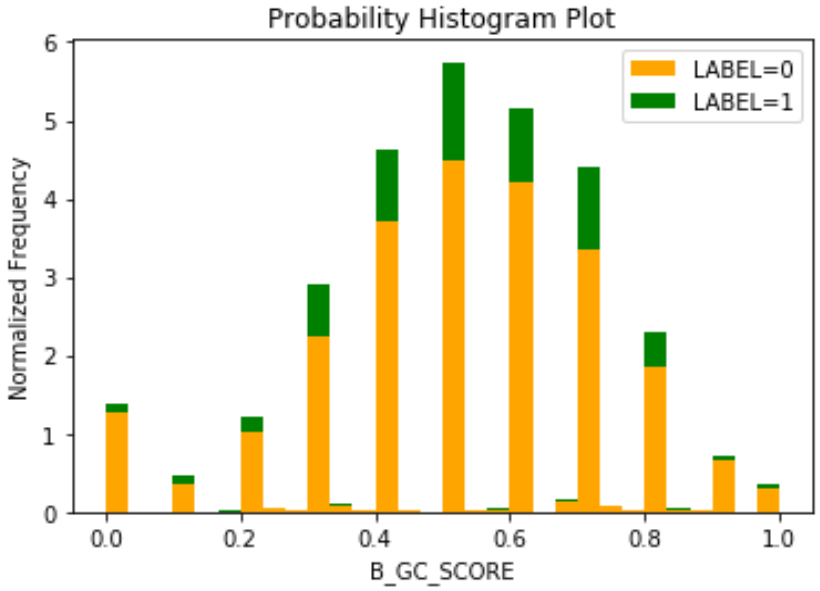

Figure 13: Base GC Content Score.

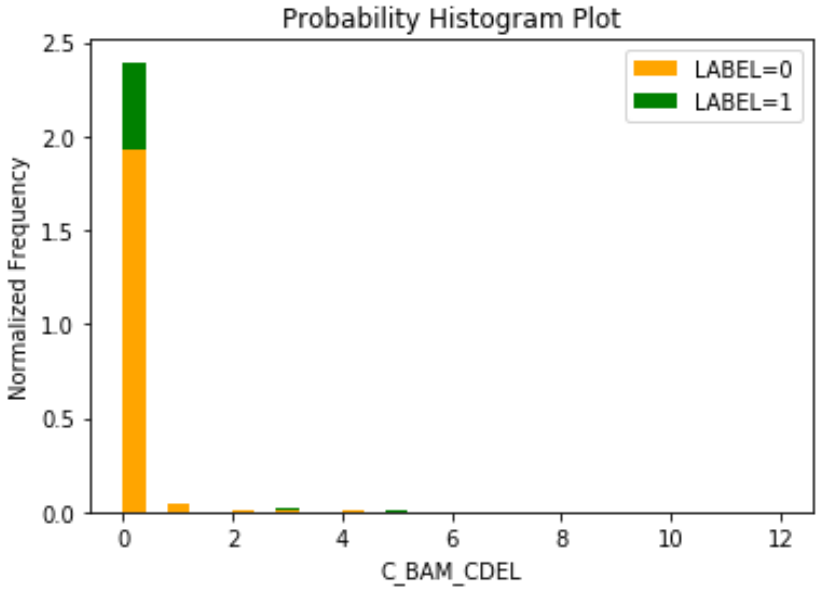

Figure 14: Deletions.

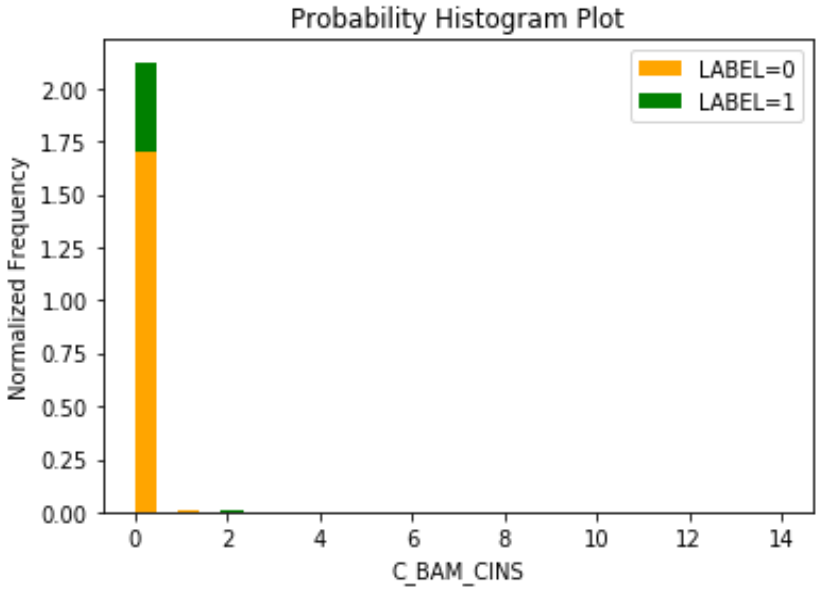

Figure 15: Insertions.

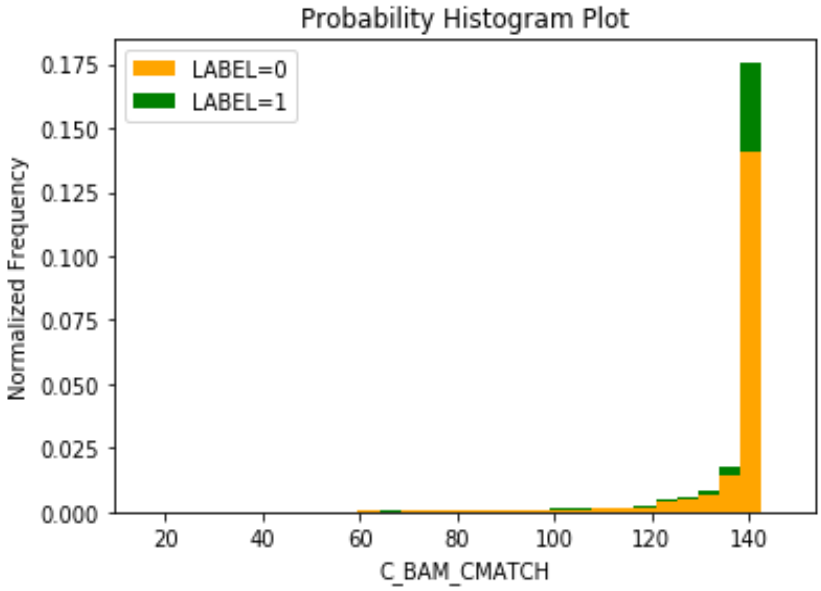

Figure 16: Matches.

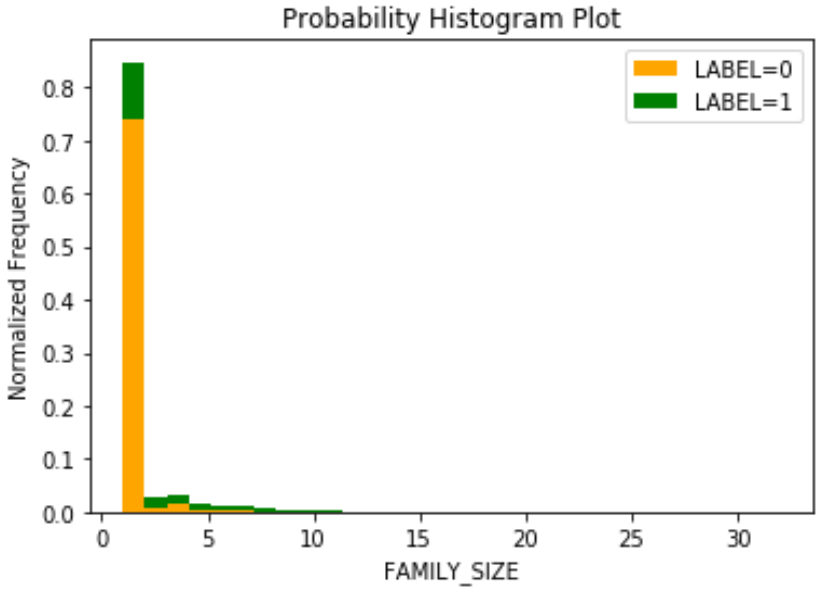

Figure 17: Family Size (Consensus).

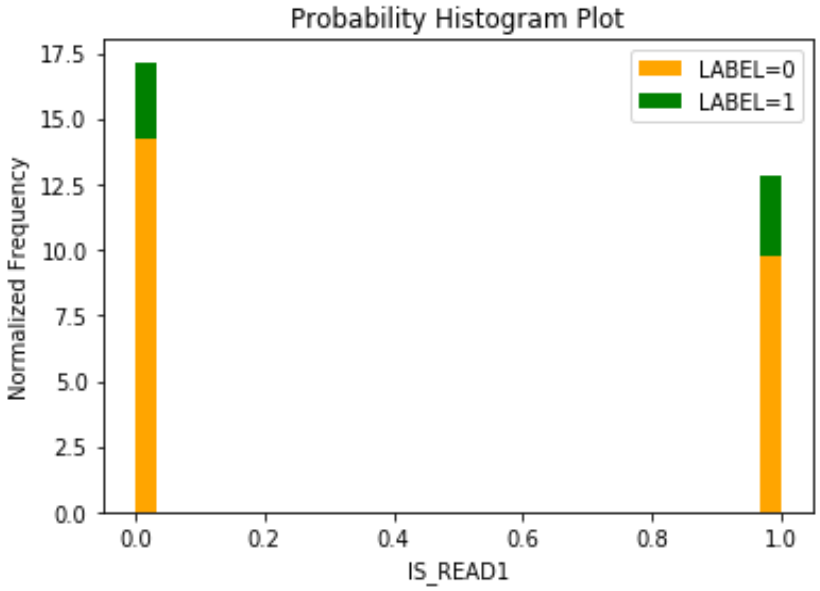

Figure 18: Read 1 (Boolean feature).

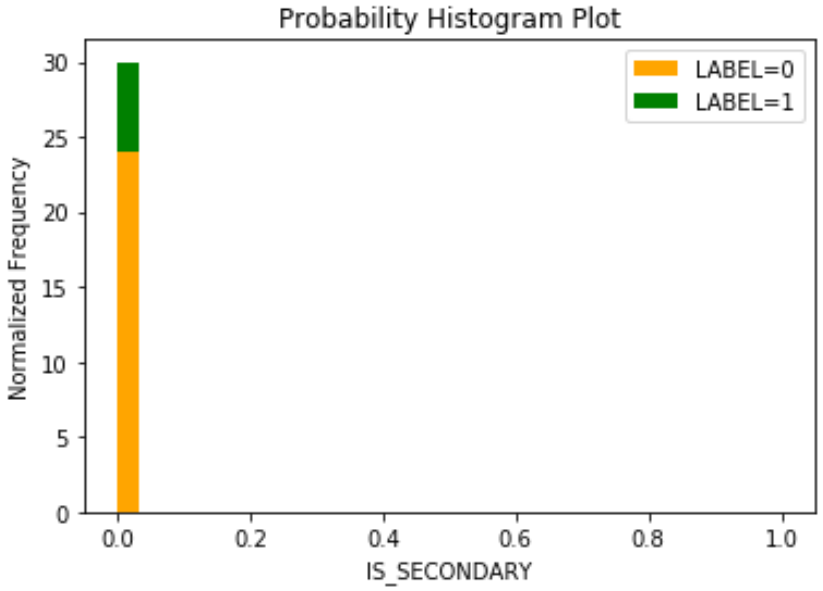

Figure 19: Secondary alignment (Boolean).

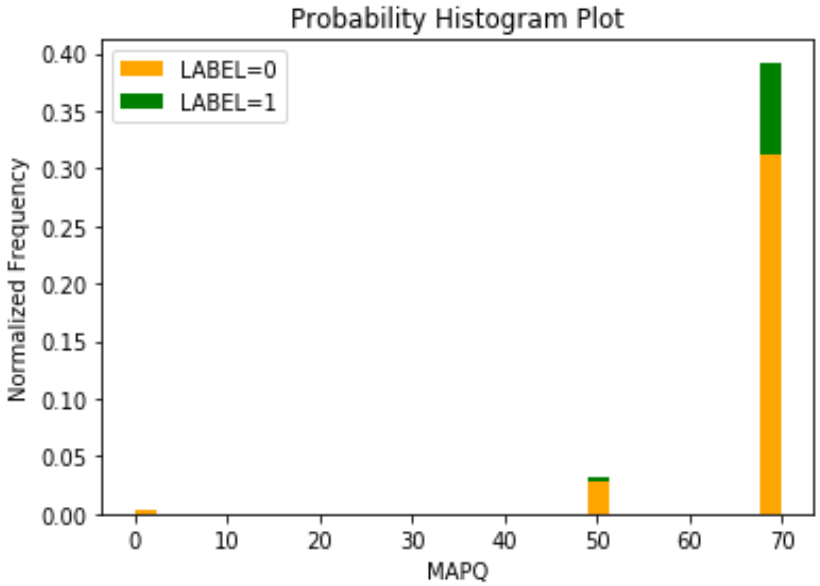

Figure 20: Mapping Quality.

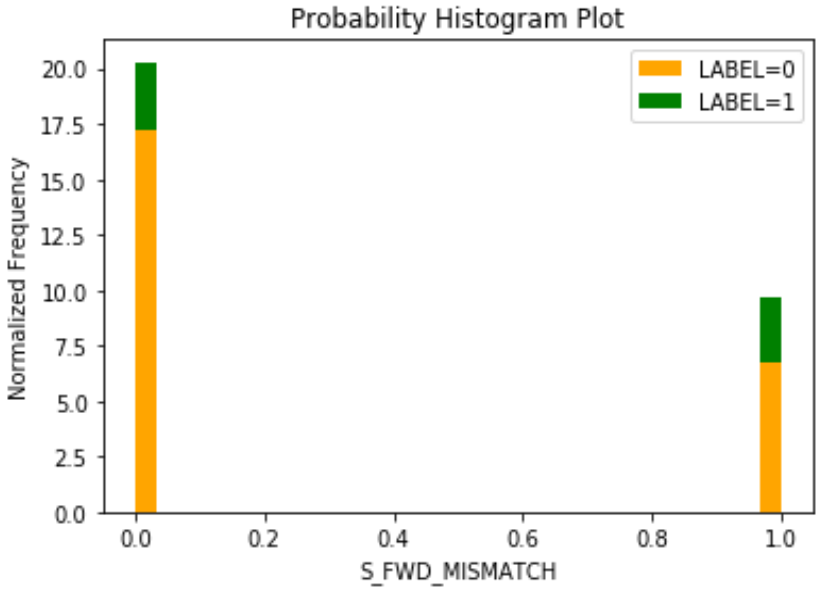

Figure 21: Forward Mismatch (Boolean).

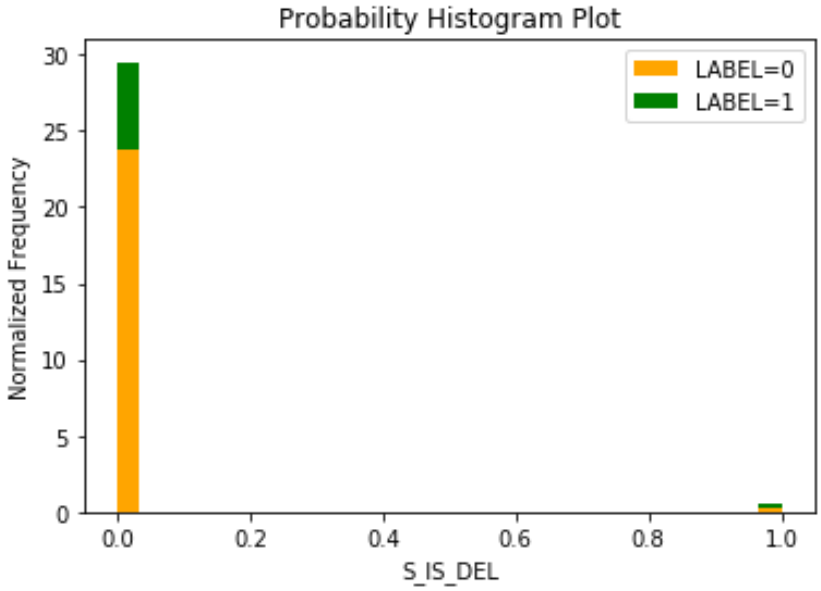

Figure 22: Is Deletion (Boolean).

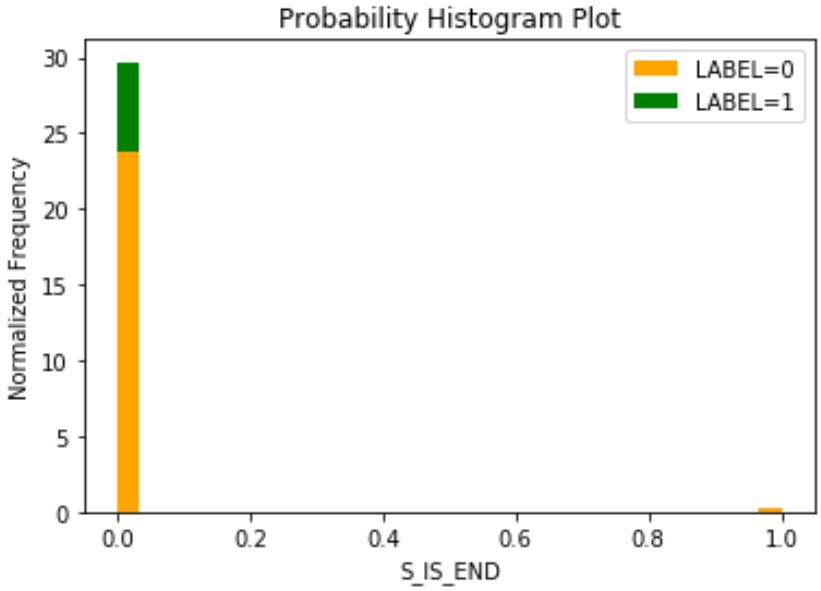

Figure 23: Is End (Boolean).

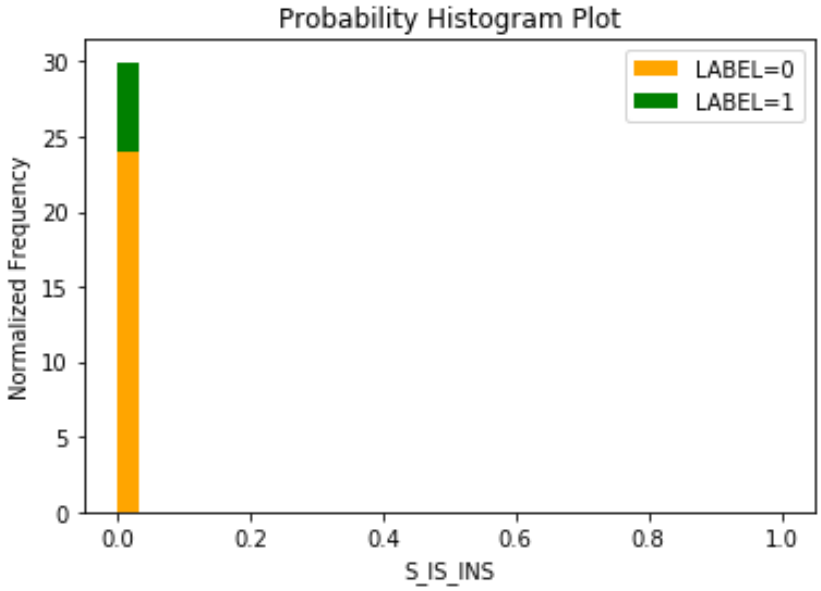

Figure 24: Is Insertion (Boolean).

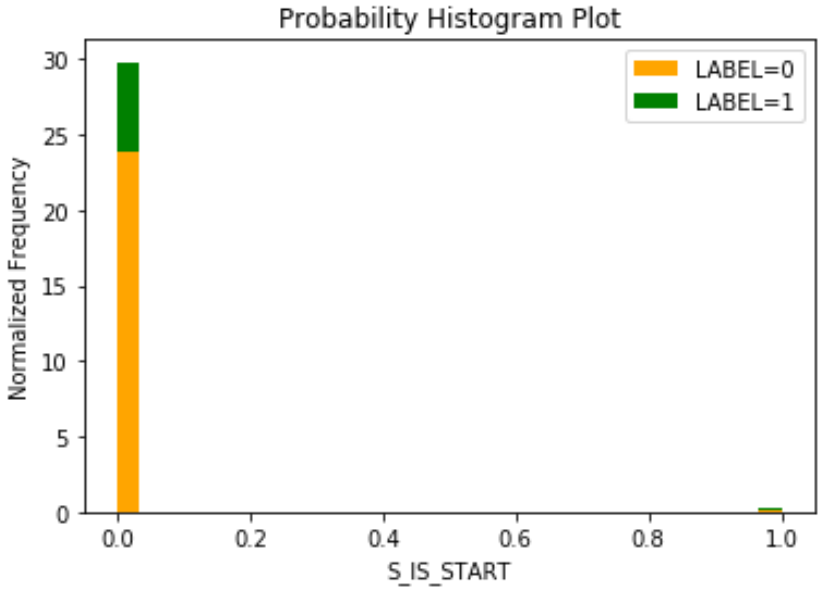

Figure 25: Is Start (Boolean).

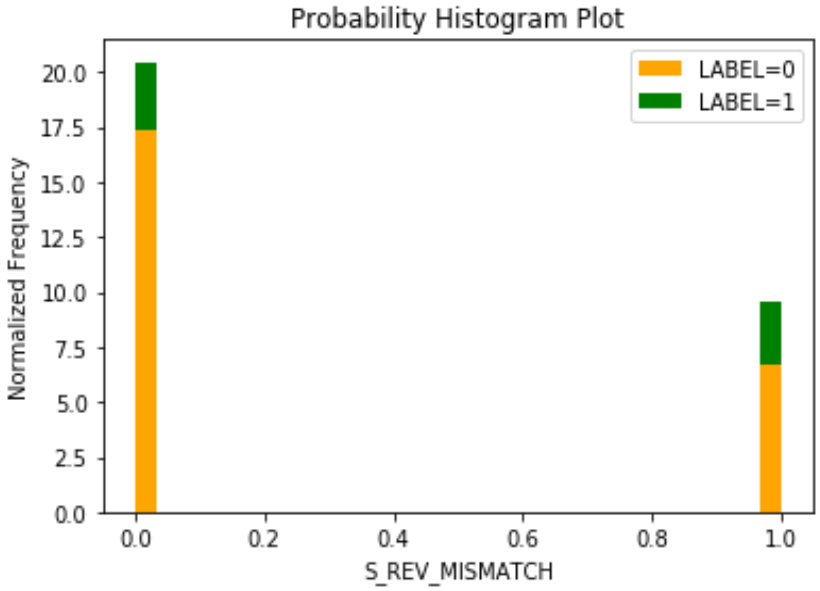

Figure 26: Is Reverse Mismatch (Boolean).

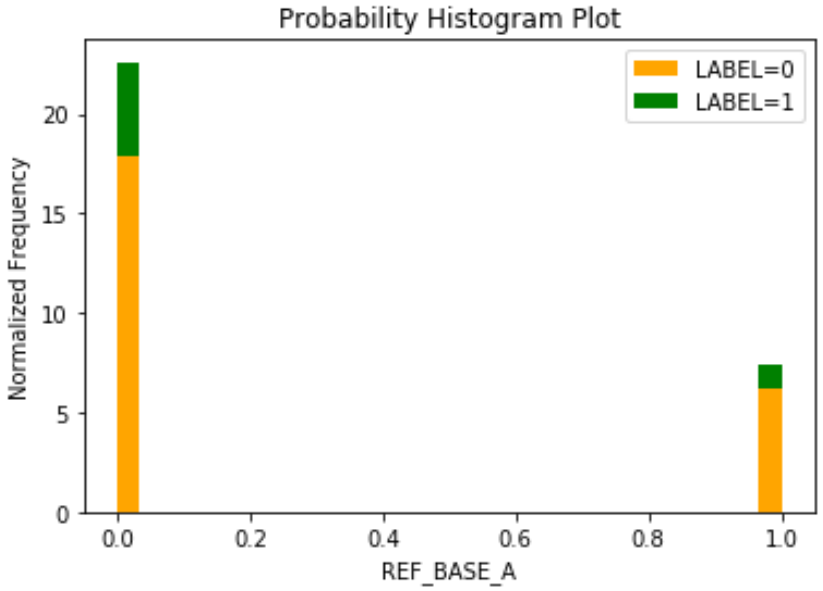

Figure 27: Reference Base A.

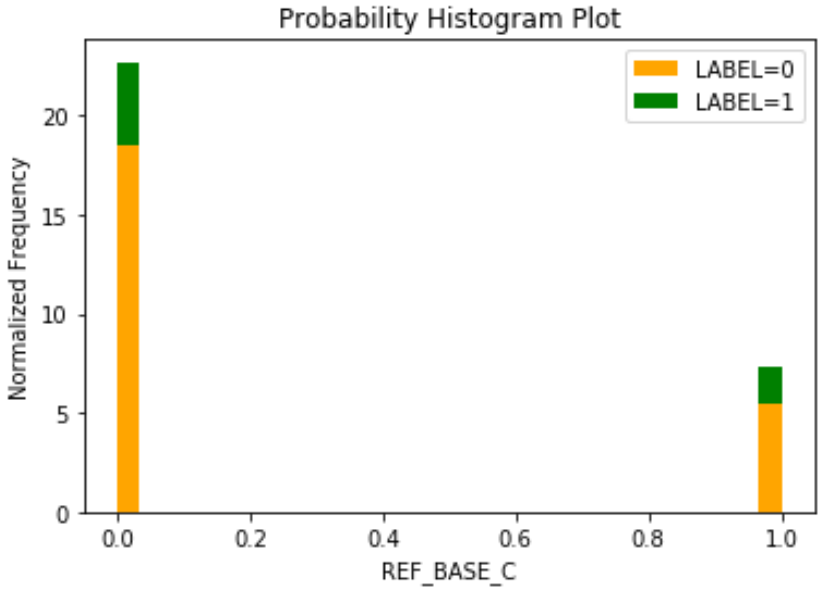

Figure 28: Reference Base C.

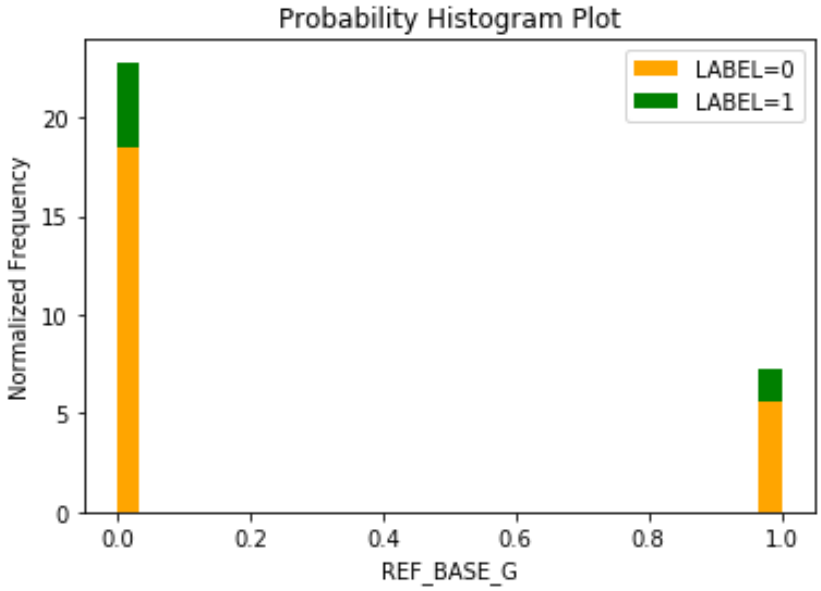

Figure 29: Reference Base G.

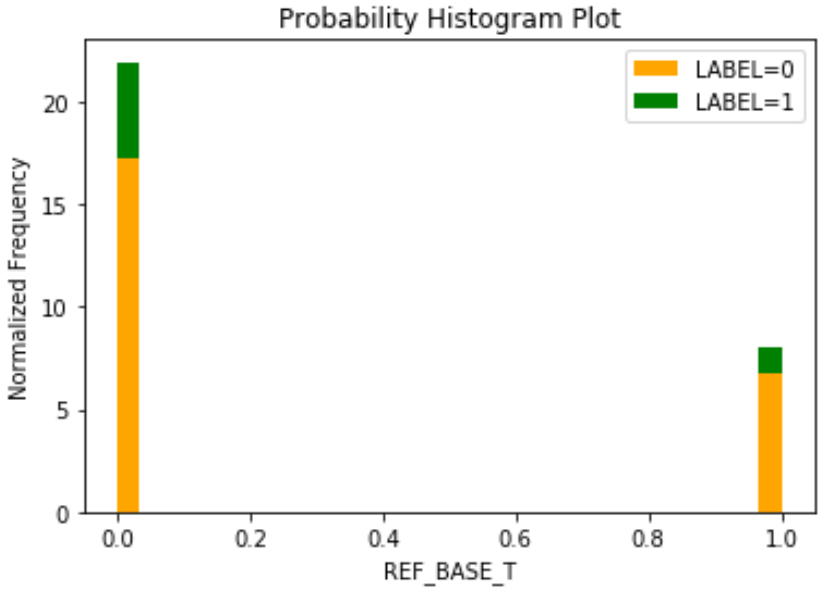

Figure 30: Reference Base T.

List of Figures