Aiding low-frequency somatic variant calling by sequencing mapping features and spatial phylogenetic compatibility

Bioinformatics and Computational Biology

University of Minnesota

Fall 2020

Motivation

Understanding evolution in cancer plays a critical role in prevention and cure of various types of cancer, such as colorectal cancer. A population of organisms evolve through speciation, genetic drift and natural selection, similarly tumors evolve by accumulation of mutations and the nature of these mutations (positive, negative, neutral). There are several factors which affect the evolution of cancer such as carcinogens and aging and the evolutionary mechanism behind these cancers can help us to better understand the overall architecture of a tumor and ultimately its cure. Ongoing tumor evolution can cause intratumoral heterogeneity (ITH) but there are spaciotemporal constraints restricting the growth of a subclone to a specific region in the tumor. There is a need to understand how a subclone grows from a primordial tumor and what is the overall architecture of the final tumor. Somatic mutations accumulated in the tumor faithfully dictate the evolutionary history of a clinically detected tumor. Sequencing data of bulk samples derived through next-generation sequencing (NGS) technology can help us to explore and analyze the nature of variants present in the tumor. However, challenges remain to reliably detect these somatic variants, especially at the low frequency range. Whereas a "golden" set of true low frequency mutations from bulk tumor sampling is lacking, we reason that mutations showing phylogenetic compatibility in multi-sample sequencing experiments are more likely to be real, as opposed to the ones that are incompatible. By studying how sequencing mapping features relate to phylogenetic compatibility, here we explore the utility of these features for a better filtration of low frequency somatic variants calls from bulk tumor sequencing.

Aim and Significance

The aim is to develop a classification model to predict true variants in a tumor using features in a bulk sample.

Using the information generated through NGS technology for both bulk and single gland samples collected from colorectal tumor of two patients, a decision tree-based classifier is built to study how the features of these low-frequency variants are related and how they can be used to predict the "true variants" from the "technical artifacts". The variant labeling is derived from spatial phylogenetic compatibility between samples.

A decision-based classifier is trained on the features extracted from bulk and single gland samples, which is then validated and tested. The model is evaluated through various performance metrics such as accuracy, ROC, sensitivity, specificity and precision.

Accurately detecting somatic single nucleotide variants (SSNVs) through variant-calling tools and analyzing the mutations present in bulk and single gland samples can give us clues as to how a subclonal expansion happens in a tumor.

Preliminary Research

Somatic or acquired mutations occur due to harmful changes or alterations in the genes during a person's lifetime. These genetic alterations acquired by a cell can be passed to the progeny of the mutated cell in the course of cell division. Unlike germline mutations which occur in the gamete (sperms or eggs) and can be passed from the parent to the progeny during conception, somatic mutations are not usually passed on to the descendants and are not hereditary in nature, and while they are not passed down to an organism's offspring, they are present in all descendants of a cell within the same organism.

Somatic mutations can be caused by various environmental factors or mutagens (e.g., exposure to harmful chemicals or ultraviolet radiation) which accumulate in a person's genes overtime and modify the DNA sequence. The accumulation of these somatic mutations is associated with the development of cancer e.g., colorectal cancer.

Colorectal cancer affects the colon or the rectum, which starts as a growth on the inner lining of the colon or rectum known as *polyps*. Depending on the type of polyps (adenomas, hyperplastic, serrated polyps), they can turn cancerous over-time. These polyps grow into the wall of the colon or rectum over time starting from the innermost layer (the mucosa) and sometimes reaching the outermost layer.

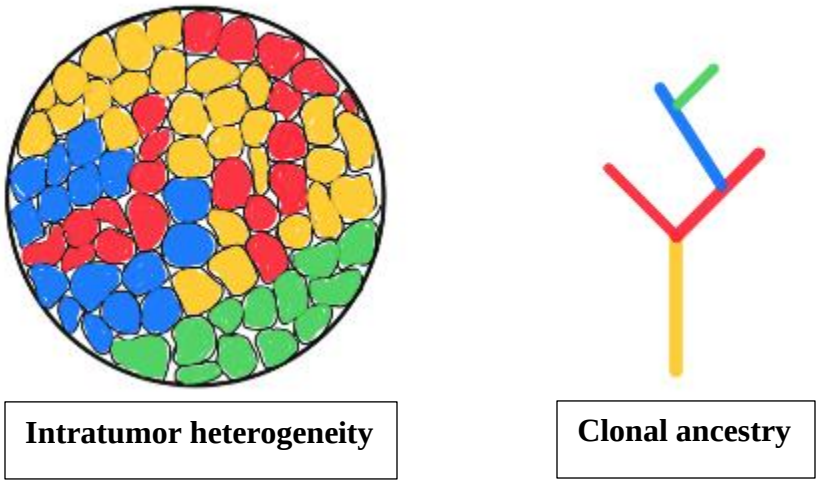

Figure 1: **Intratumoral heterogeneity refers to formation of subclones within a tumor population. It is caused due to several factors affecting cancer cells, including driver mutations. The colors in the figure represent various lineages of subclones formed within a tumor along with the phylogenetic tree of clonal ancestry.**

Cancer growth and evolutionary process is being compared to Darwinian natural selection in which genetic and epigenetic changes occur in individual cancer cells over time. If these changes provide a particular selective advantage, they will allow individual clones of cancer cells to out-compete other clones and thus give rise to subclones. Natural selection works similarly for both cancer cells and other organisms – growth by competing for space and resources.

The development of these subclones depend on interaction of several factors like selectively advantageous “driver mutations”, selectively neutral “passenger mutation”, environmental factors and type of mutation seen (missense, nonsense, frameshift etc.)

Driver mutations are positively selected and increase prevalence of tumor in the body, drive malignant progression in a clone and is a major contributor of *intratumor heterogeneity* (ITH) (*fig. 1*), whereas passenger mutations are neutral mutations also known as “hitchhiking mutations” which increase in frequency due to hitchhiking alongside driver mutations. Due to evolution a cancer cell may develop into a subclone which has functional differences as compared to its neighboring cancer cells within a single patient. Genetic changes are not the only factor contributing to clonal evolution. Epigenetic changes which are inherited at cell division can also affect cell phenotypes. “selective sweeps” is a term used to define a series of clonal expansions which grows to dominate the neoplasm. These clones have to compete with each other for such an expansion as space and resources are limited in a tumor population. This phenomenon is known as “clonal interference”.

Andrea Sottoriva *et al.,* presented a “big-bang” model for human colorectal tumor growth, which validates that public(clonal), and private(subclonal) alterations arise early during tumor expansion and that “detectable” intratumoral heterogeneity (ITH) is the result of these earlier detected private alterations, not from later clonal growths.[1] The evolutionary process and ancestral history behind tumor expansion can provide us clues about genomic patterns, earlier events and growth of a tumor which in turn helps us to treat cancer effectively.

The initial growth of tumor is a result of accumulation of driver mutations, and the early private(subclonal) alterations, copy number aberrations (CNAs) and point mutations give rise to “pervasive” alteration in the neoplasm which are found throughout the final tumor but are not dominant. The paper suggests that it is possible to recover the genomic profile of primordial tumor because of these “pervasive” alterations.

Although a private alteration acquired in the tumor at later stages may have selectively advantageous “driver mutations” but that does not guarantee its pervasiveness in the final tumor, owing to space and resource constraints within the tumor. Only the earliest most pervasive alterations will be present throughout the tumor, and the later arising mutations will be localized to smaller tumor subpopulations. This means that the timing of a mutation will determine its pervasiveness in the final tumor rather than positively selected alterations in the colorectal cancer samples under study. This type of growth dynamic gives rise to a variegated tumor, in which earlier mutations become persistent and pervasive in the tumor in spite of being not dominant and the later arising mutations contained in small regions of the tumor. As soon as the tumor is initiated it starts acquiring private mutations which results in ITH forming within that newly formed small tumor. Using a statistical framework and spatial computational modelling, genomic profile of the primordial tumor can be detected from the final tumor.

Performing a whole-exome sequencing on bulk and single gland samples from two different sides of tumor (left and right) and analyzing the copy number data provides a spatial distribution of ITH which is classified as following (Table 1)

Public	All glands of the tumor
Private, side specific	All glands from one tumor side only
Private, side variegated	All glands from one tumor side and in some glands from the opposite side
Private, variegated	Subset of glands from both sides
Private, regional	More than one but not all glands from one tumor side only
Private, unique	Single gland

Table 1: Spatial distribution of intratumor heterogeneity as found in whole-exome sequencing (WES) data from bulk and single gland samples of two different sides of tumor (left and right).

Next-generation sequencing (NGS) of bulk and single gland samples provides information regarding the variant allele frequency (VAF) distribution or site frequency spectrum (SFS) which in turn along with spaciotemporal (space and time) patterns of clonal and subclonal alterations in tumor population can provide important clues about the underlying evolutionary process.

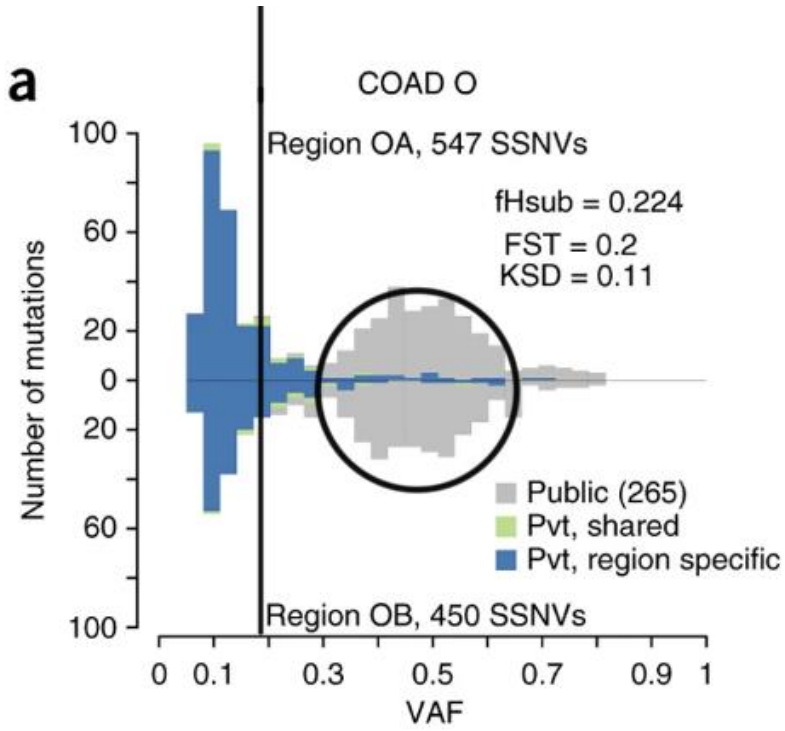

Figure 2: Pairwise histogram of SFS for colorectal patient O comparing two regions (OA VS OA). The histogram has a bimodal structure representing two mutational clusters (public and private). Majority of SSNVs with VAF < 0.2 are region specific.

The distribution of VAF depends on the mode of evolution for a tumor as shown by simulating virtual tumors under different modes of tumor evolution (neutral model, neutral CSC model, and various levels of positive selection).[2] The SFS histogram *(fig. 2)* for these evolutionary modes

have a bimodal structure which represent the two mutational clusters, there is a public cluster centered at VAF = 0.5 and a right-skewed private cluster for VAF < 0.25. The public cluster consists of mutations that are present in all tumor cells (fixed) and the private cluster consists of subclonal alterations which are generally region specific.

According to the SFS histograms *(fig. 2)* from the bulk sample sequencing data, the majority of subclonal single nucleotide variants (SSNVs) with VAF < 0.2 are region specific, which suggests that mutations falling under a VAF cluster does not belong to distinct clones and that VAF does not necessarily provide clues about the architecture of a subclone in a tumor.

When a WES is performed on multiple single gland samples and bulk samples (OA and OB) from left and right region of a colorectal tumor for two different patients (O and U), it is found that private SSNVs specific to bulk samples (OA or OB) are found only in glands within the same tumor region, while majority of later-arising SSNVs are region specific.[2] These finding further validate that later-arising subclones have certain spatial constraints during tumor expansion.

Another finding to be noted is that the SSNVs which are specific to bulk sample OA with VAF values < 0.2 are also found in various grouping of single glands with VAF values > 0.2.[2] This suggests that the subclones having unique lineages can have similar VAFs in bulk tumor. The idea for this project is to use the features from the bulk sample (VAF < 0.2) and train a decision-based classifier which can classify the 'true variants' from the 'technical artifacts'. The variant labeling is derived from spatial phylogenetic compatibility between bulk and single-gland samples.

Methods

Data collection

The data used for this project comes from the Whole-Exome-Sequencing of "bulk" and "single gland" samples of two colorectal cancer patients (O and U) on Illumina platform.

The metadata file contains the following information regarding samples:

- Bulk samples "OA" and "OB" for patient O's tumor, whereas CRCTumorOA1, CRCTumorOA2, CRCTumorOA3, CRCTumorOA4 and CRCTumorOB1 are single gland samples.
- Similarly, UA, UB are bulk sampling of patient (U)'s tumor, whereas CRCTumorUA_L and CRCTumorUB_L are single gland sampling.
- A and B refer to two different sides (e.g., Left and Right) of the tumor.

Variant calling pipeline

MuTect is used as a tool to detect the somatic mutations in the above samples. MuTect applies a Bayesian classifier to detect mutations with very low frequency allele fractions.[3]

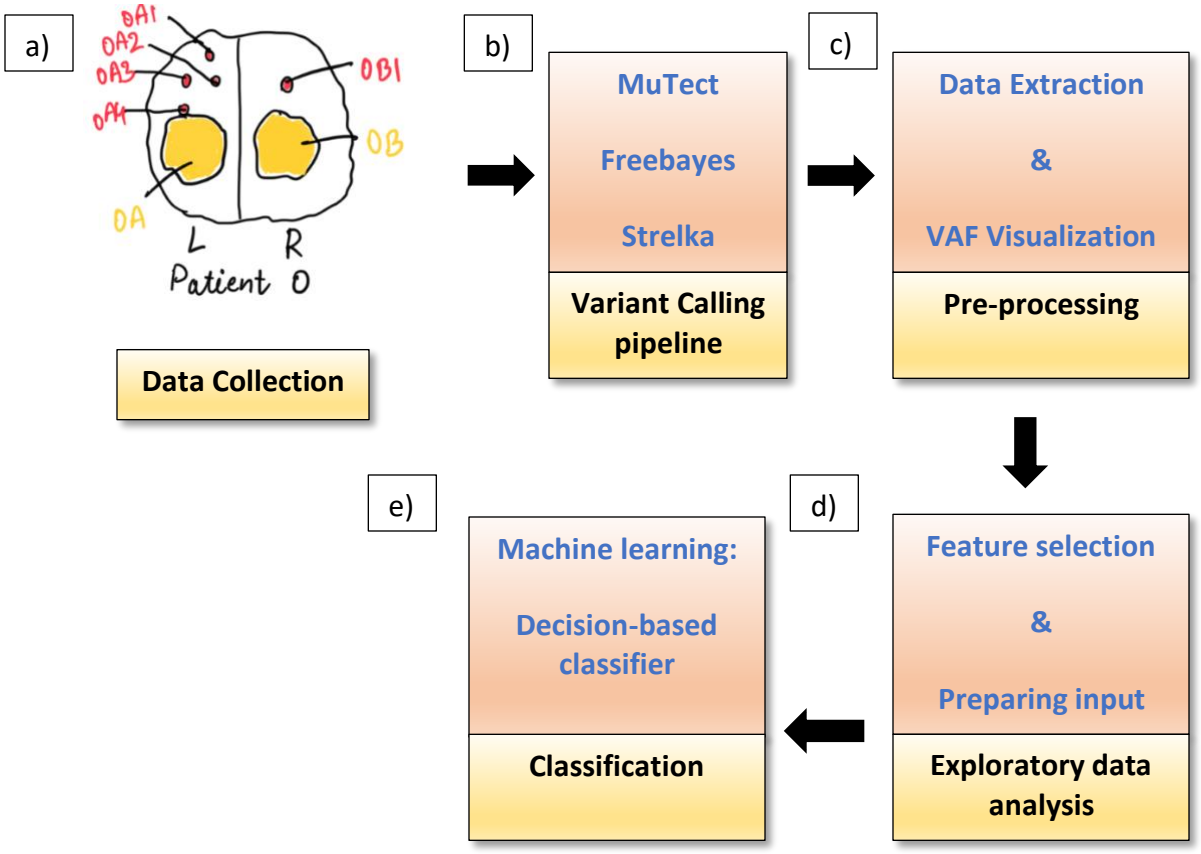

Figure 3: Workflow to build a classifier. a) The workflow starts with collection of samples from colorectal patients O and U (bulk and single gland samples), which are sequenced using WES. b) The sequenced data is then processed using tools such as MuTect etc. c) The output of the variant calling pipeline is used to extract data for features used in the model. d) the input data for classification is cleaned and prepared for modelling. e) Final model is trained, tested and validated.

The output file from MuTect processing generates a master table of mutation calls for these samples along with the features extracted. The master table contains the following information regarding the bulk and single gland samples.

- At a genomic position, for e.g. Two columns 'OAmaf' and 'OAd' dictates the pattern information and the depth at this genomic position in sample OA respectively.
- The information regarding features is provided in Table 2.

Variant allele frequency	variant read depth / total depth
endsratio	fraction of variant base located at the end of reads
cmean, cmedian	summary of the number of mismatches in the reads that carry the variant. First is the mean, the second is the median

strandRatio	fraction of variant carrying reads which are from positive strand
strandRatioRef	fraction of other reads (without variant) which are from positive strand
strandFisherP	p-value of Fisher exact test indicates if there is a difference in strand bias between variant-carrying reads and other reads
badQualFrac	fraction of reads with bad quality
log-likelihood	log-likelihood that it is somatic (specific to tumor sample, rather than the normal)
indels mean, indel median	number of indels in the reads carrying the variant
vrlen	median length of the reads carrying the variant

Table 2: Information regarding features extracted by MuTect.

These samples involve both bulk and single-gland sampling of the same patient tumor O and U, the input for labeling is based on the presence and absence of the variants in these different samples.

Pre-processing

To explore the variant allele frequency distribution in the bulk and single gland samples, an input file is prepared which contains the VAF for each genomic position along with the read depth.

A heatmap and scatterplot is visualized for every bulk sample and single gland sample pair. For e.g., variant allele frequency distribution for bulk sample "OAmaf" VS "CRCTumorOA2" is plotted *(fig. 4)* and the results are following:

- The SSNVs for single gland sample CRCTumorOA2 have a high frequency (VAF > 0.15).
- The SSNVs for bulk sample OA have a low frequency (0.01 < VAF < 0.2).

For single gland sample CRCTumorOA2 (VAF > 0.15) and for bulk sample OAmaf (VAF < 0.2) is a good indication of presence of variants in these samples.

To prepare the input data file for classification, data is extracted from the MuTect output file. The input data file contains the VAFs and "features" at each genomic point for each sample.

The data input file is then cleaned and prepared for labelling the target classes. For the purpose of classification, two target classes are used: "true variant" label is used for the positive class and "technical artifact" label is used for labelling negative class.

Starting with the bulk sample for e.g., OAmaf, data is filtered for loglikelihood > 4 and for 0.01 < VAF < 0.2. The data is then labelled according to VAF in the single gland sample.

CRCTumorOA2. The reads were labelled as “technical artifact” for VAF < 0.15 and labelled “true variant” for VAF > 0.15. This method is applied to every bulk sample (OBmaf, UAmaf, UBmaf) and single gland sample (CRCTumorOB1, CRCTumorUA_L, CRCTumorUB_L) pair. The aggregate of these pairs creates the dataset for the classification model.

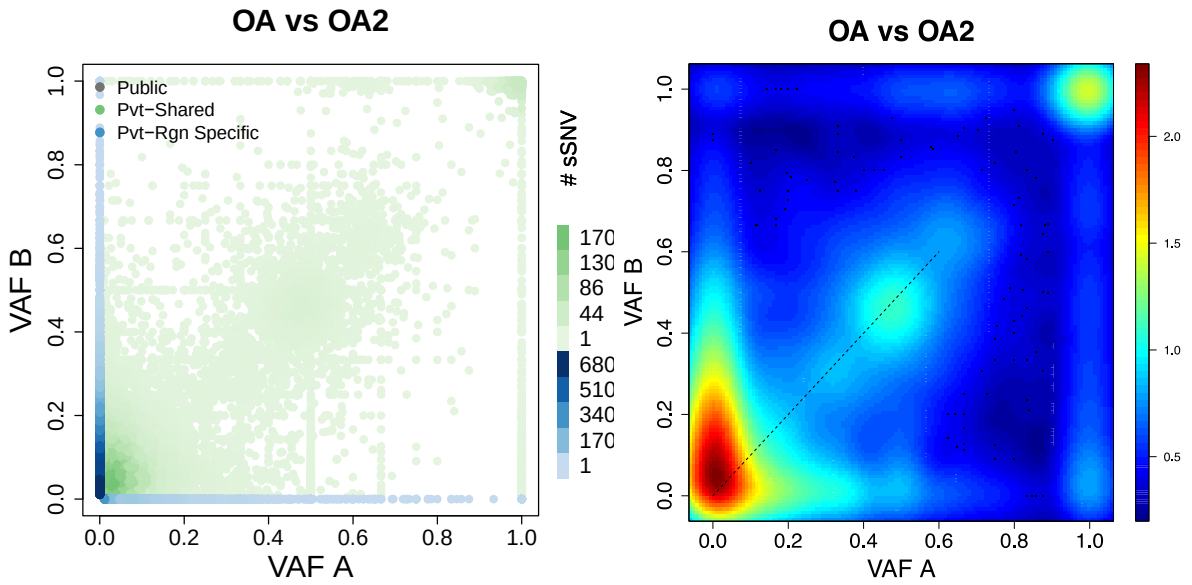

Fig 4: Scatterplot and heatmap of OAmaf (bulk) VS CRCTumorOA2(single gland). For bulk sample VAF < 0.2 and for single gland VAF > 0.15 is considered good and variants with these VAFs are labelled as ‘true variant’.

Exploratory data analysis

The dataset has 1260 observations and 7 columns, one of which is our response/target variable.

The target variable “label” has two levels: true variant(n=552) and technical artifact(n=708).

The proportion of true variant is 0.44 and technical artifact is 0.56. Even though technical artifact has a higher proportion, there doesn’t seem to be a huge class imbalance.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
endsratio	0	0.307	0.50	0.489	0.68	1
cmedian	1	1.000	1.00	1.520	2.00	5
strandFisherP	0	0.008	0.41	0.435	0.81	1
badQualFrac	0	0.000	0.12	0.181	0.29	1
log-likelihood	4	7.759	16.81	30.411	38.87	212
Indels mean	0	0.000	0.00	0.038	0.00	1

Table 3: Summary Statistics for the features used in the final model

Some of the features like cmean, strandRatio, strandRatioRef are not used in the final model due to high collinearity and features such as indel median, vrlen were also dropped for being redundant. Summary statistics and visual representation for the features used in the final model are provided in *(table 3) & (fig. 5).*

Checking for outliers and other inconsistent data points using cook's distance *(fig. 6)* and further exploration shows that there are 34 data points considered outliers according to cook's distance test. Since this is a decision-tree based model, they are not removed from the dataset.

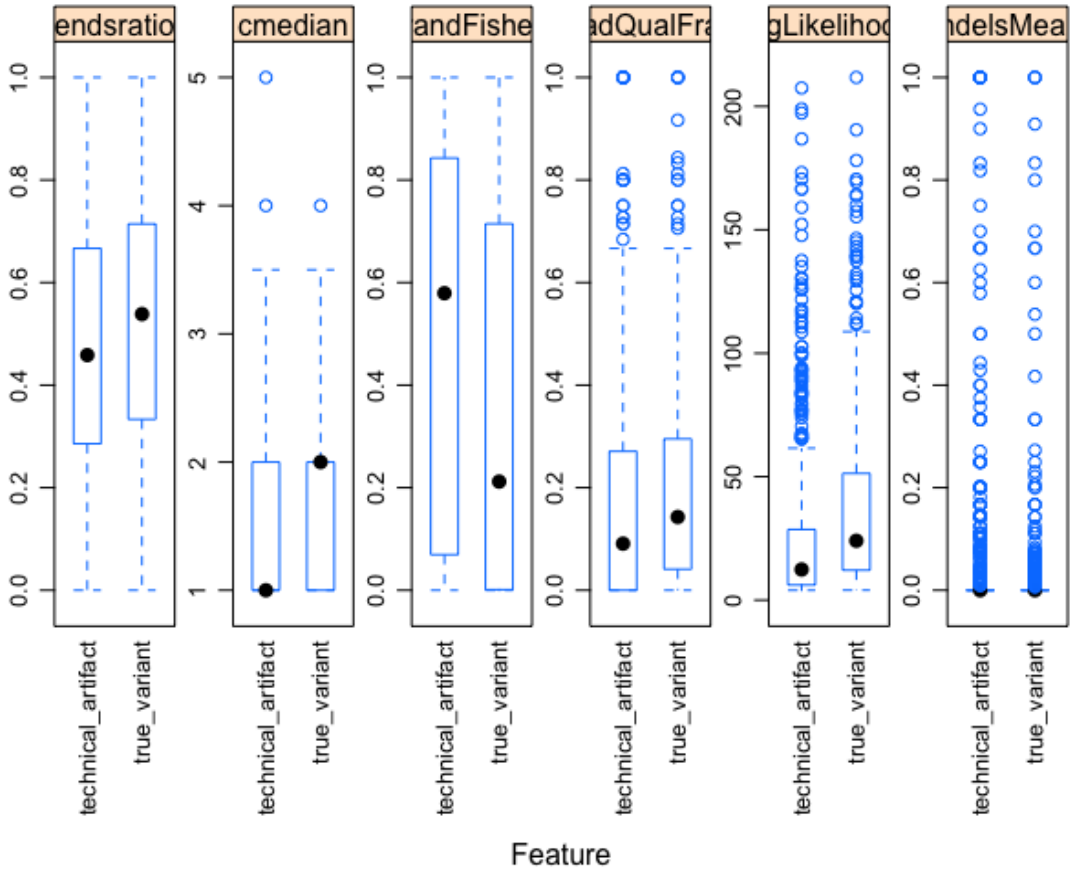

Figure 5: Boxplot representation of the features used in the final model.

The goal is to classify the "true variants" from the "technical artifacts" using the features from the bulk samples. Since it is a supervised learning and the target variable is categorical, a decision tree-based model is selected as the classification algorithm.

The dataset is split into 70% training and 30% testing with a Ten-fold cross validation. The method used for this decision tree model is "random forest" and the metric used to evaluate the model performance is ROC.

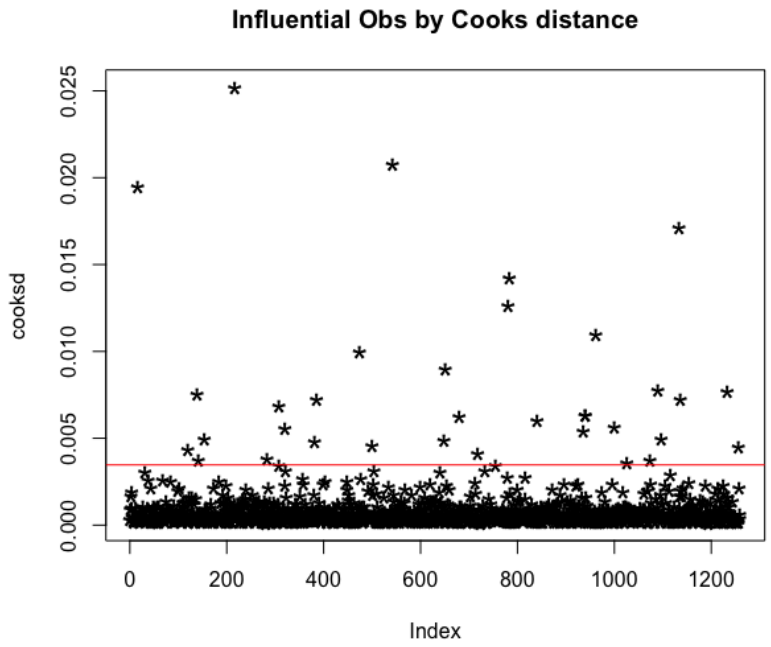

Figure 6: 34 data points detected by cook's distance as outliers.

Results

The model produces the following confusion matrix:

Prediction	Reference		
		technical_artifact	true_variant
	technical_artifact	161	71
	true_variant	51	94

Figure 7: Confusion matrix showing True Positives (TP), True Negatives (TN), False Positives (FP), False Positives (FP) for "true variant" as a positive class and "technical artifact" as negative class.

- **True Positives (TP):** *correctly* predicted 94 true variants.
- **True Negatives (TN):** *correctly* predicted 161 technical artifacts.
- **False Positives (FP):** *incorrectly* predicted 71 technical artifacts.
- **False Negatives (FN):** *incorrectly* predicted 51 true variants.

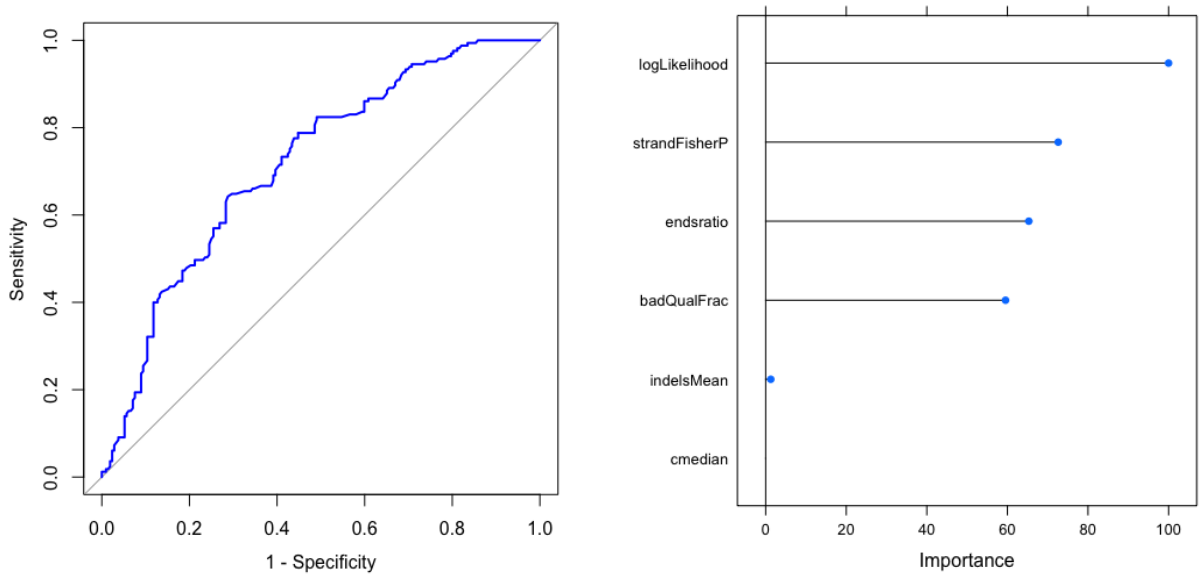

Figure 8: ROC plot and importance of features with 100% being the highest importance percentage. The area under curve is .72.

Classification Accuracy	Sensitivity (TPR)	Specificity (TNR)	Precision
0.68	**0.57**	**0.76**	**0.65**
False Positive Rate (FPR)	**False Negative Rate (FNR)**	**Area Under Curve**	**F1 Score**
0.24	**0.43**	**0.72**	**0.61**

Table 4: Performance metrics for classification model

log-likelihood	**100.00**
strandFisherP	**71.55**
endsratio	**66.24**
badQualFrac	**61.19**
Indels mean	**0.52**
cmedian	**0.00**

Table 5: Importance of features (%)

Discussion

Overall, the classifier has an accuracy 68% and a misclassification rate of 32%. The sensitivity of the model which is also known as true positive rate or recall is 57%, meaning when the target "true variant" is positive, the prediction is correct 57% of times. Specificity is 77%, which means when the target "technical artifact" is negative, the prediction is correct 77% of times. The model is able to predict "technical artifact" slightly better than it predicts "true variants".

Clearly, the classification model does not have great accuracy and precision and the overall performance is not the best. There are few possible reasons for that, one being unavailability of better data points. The data used in this study is not specifically collected for classification purposes. There doesn't seem to be a concrete relation between the features and classes that are being predicted, which might be the reason behind the low accuracy of the model. The prediction accuracy depends a lot on the similarity between training and target data sets and since the data set used for modelling is prepared by aggregating the data points of all the single gland and bulk samples, it creates inconsistencies. The size of the dataset also contributes to the accuracy of model, and due to high number of dirty calls diluting the signal, the dataset had to be filtered which resulted in very few data points for every sample. Presence of more reliable data results in better and accurate models instead of working with assumptions and weak correlations. There is also the presence of some outliers in the model which require further exploration and proper handling. In addition, as we applied fixed VAF thresholds for labeling the variants, the copy number alterations in the tumor samples might affect the VAF distribution, henceforth lead to inaccurate labeling.

To improve the accuracy of the model, better data points and features are required. The number of features used in the final mode were reduced from 11 to 6, which did help to improve the model in some way, but not by a lot. Looking at the feature importance *(fig. 8 & table 5)*, some features are weighted more than other, such as Log-likelihood and strandFisherP which have a high importance percentage (100% and 71%) while features such as Indels mean and cmedian have very low importance percentage (0.52 and 0.00%). This imbalance in importance of features has also contributed to low accuracy in the trained model. Perhaps application of ensemble methods such as bagging or boosting can help to improve the model performance.

Other than decision-tree, some different classifiers were also used to predict the classes such as logistic regression, which did not yield any improved results.

For future work, to improve the accuracy of the model, any further data collection should be conducted with the results of this study in mind.

References

1. Sottoriva, A., Kang, H., Ma, Z. *et al.* A Big Bang model of human colorectal tumor growth. *Nat Genet* **47,** 209–216 (2015). https://doi.org/10.1038/ng.3214
2. Sun, R., Hu, Z., Sottoriva, A. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet* **49,** 1015–1024 (2017). https://doi.org/10.1038/ng.3891
3. Cibulskis, K., Lawrence, M., Carter, S. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31,** 213–219 (2013). https://doi.org/10.1038/nbt.2514
4. Fang, L.T., Afshar, P.T., Chhibber, A. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol* **16,** 197 (2015). https://doi.org/10.1186/s13059-015-0758-2
5. Greaves, M., Maley, C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012). https://doi.org/10.1038/nature10762
6. Casás-Selves M, Degregori J. How cancer shapes evolution, and how evolution shapes cancer. *Evolution (N Y)*. 2011;4(4):624-634. doi:10.1007/s12052-011-0373-y
7. Chowell D, Napier J, Gupta R, Anderson KS, Maley CC, Sayres MAW. Modeling the Subclonal Evolution of Cancer Cell Populations. *Cancer Res*. 2018;78(3):830-839. doi: 10.1158/0008-5472.CAN-17-1229
8. Unexpectedly High Subclonal Mutational Diversity in Human Colorectal Cancer and Its Significance Lawrence A. Loeb et al, bioRxiv 672493; doi: https://doi.org/10.1101/672493