# CTAT Mutations: A Machine Learning Based RNA-Seq Variant Calling Pipeline Incorporating Variant Annotation, Prioritization, and Visualization

# CTAT Mutations: A Machine Learning Based RNA-seq Variant Calling Pipeline Incorporating Variant Annotation, Prioritization, and Visualization

Vrushali Dipak Fangal

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

November 2020

## Abstract

Cancer is a complex multi-factorial disease attributed to accumulation of diverse genetic variations that disrupt the genomic integrity. With the advent of genetic diagnostics in personalized medicine, gene panels have dramatically catapulted the diagnostic yield in cancer. While RNA-seq provides a cost-effective way of producing high-throughput data, the clinical application of single nucleotide polymorphism (SNP) arrays is limited by the high false-positive load concomitant with the variant detection pipelines. Here, we describe a robust end-to-end GATK-based Trinity Cancer Transcriptome Analysis Toolkit (CTAT) Mutations Pipeline that leverages a rich set of variant feature annotations with a collection of modern machine learning models to predict genetic variants from RNA-seq and reduce the burden of false positives. We demonstrate improved accuracy of our RNA-seq based variant prediction pipeline using the Genome in a Bottle (GIAB) reference data and RNA-seq and matched whole exome sequencing data from tumor cell lines. Cancer-relevant candidate somatic mutations are further selected based on feature annotations and reported in an interactive web application. As RNA-seq becomes more widespread in use for clinical diagnostics, we expect our CTAT variant detection pipeline to facilitate use of tumor RNA-seq in precision medicine.

# Dedication

I dedicate this achievement to my loving family without whom this would not have been possible. I would like to thank my parents Dipak Dhanji Fangal and Sugandha Dipak Fangal for always believing in me and supporting me in every situation.

Acknowledgments

The completion of my thesis would not have been possible without the participation and assistance of so many people who encouraged me throughout this journey. I am thankful to Dr. Aviv Regev, my thesis director, for her kindness, support and invaluable guidance with this thesis. Then, I would like to thank Brian Haas, my supervisor for his unwavering patience throughout my thesis project. You showed faith in me during the course of my entire Master's degree and showed generosity during hard times. Both of you have given me invaluable inputs and have greatly enhanced the quality of my education at Harvard Extension School, for that I am extremely grateful. Additionally, the Broad Institute of MIT and Harvard has been extremely encouraging in supporting me as their employee in pursuing my education.

Last but not least, I would like to thank Dr. James Morris, my thesis advisor at Harvard University, who has been extremely helpful in planning and judiciously managing the timeline of my thesis.

Table of Contents

List of Tables

List of Figures

# Chapter I

# Introduction:

## Importance of variants in cancer

Genomic instability resulting from genetic variants is responsible for increased susceptibility to tumorigenesis and is innate to cancer cells (Cibulskis et al. 2012). Single nucleotide variants (SNV), and Insertions-deletion (INDEL) mutations are the two most common forms of genetic variation in various driver genes in cancer (Lin M. et al. 2017). While SNPs are point mutations that change a single base, INDELs add or delete bases resulting in frameshifts in the genome that can either be inherited as germline mutations or accumulate somatically (Ramroop et al. 2019). Depending on their location in the genome such as exon, intron, 3'UTR, 5'UTR, CpG islands, or promoters, they can affect the alternative splicing, DNA methylation patterns, histone modification, and transcription binding sites, ultimately affecting the protein structure and cellular function (Deng N. et al. 2017). In complex genetic diseases like cancer, mutations in exonic regions garner more interest as they directly affect cell cycle regulation, metabolism, hormone regulation, DNA mismatch repair, and contribute to increased risk of the disease(Park et al. 2009) (Cunningham et al. 2009). More importantly, the variability in SNPs and INDELs contributes to intra-tumor heterogeneity, which drives drug resistance in cancer. Clinical interpretation of Thus, it is imperative to study SNPs and INDELs as foundational components of cancer progression and crucial to the development of reliable diagnostic tools (Zhao et al. 2019).

## RNA-seq for detecting SNVs

RNA-seq aptly captures many facets of transcribed regions such as gene expression, alternative splicing patterns, allele-specific expression, fusion transcripts, and RNA editing sites (Bosio M. et al. 2019, Adetunji M. O. et al. 2019), all of which become relevant to variant calling based on RNA-seq. As opposed to whole-exome sequencing (WES), RNA-seq captures the effect of variants on gene expression. For example, genes enriched by variations have higher read counts in RNA-seq as opposed to uniform read counts in WES (Chen J. et al. 2019, Rusch et al. 2018). Importantly, RNA-seq can capture the expression of various isoforms resulting from cancer heterogeneity (Piskol et al. 2013, Sheng Q. et al. 2016, Zhao Y. et al. 2019). Consequently, RNA-seq can capture multiple facets of cancer biology as compared to matched WES samples (Zhao Y. et al., 2019).

## Issues with predicting SNVs in RNA-seq data

Although RNA-seq has underlying benefits for cancer research, there are several challenges in leveraging RNA-seq for variant calling that differ significantly from WES samples (Sheng Q. et al. 2016, Piskol et al. 2013). False positive genetic variants can accumulate from biological sources or technical errors that occur during RNA sample preparation or sequencing, including replication errors during reverse transcription, variable read depth and skewed allele frequencies obtained from sequencing, or read alignment errors near splice junctions (Treangen et al. 2012, Gonorazky H. D., 2019). Similarly, innate properties of RNA such as RNA editing sites

that result from deamination of adenosines (A) to inosines (I) are often incorrectly detected as genetic variants (Picardi E. et al. 2017). Moreover, errors can arise from reads mapping at multiple sites in repeated regions of the genome (Sheng Q. et al. 2016, Piskol et al. 2013, Radenbaugh et al. 2014). Overall, the real challenge lies in distinguishing a true genetic variant from a biological or technical artifact given the diverse sources of false positives, thus presenting a challenging problem in bioinformatics (Xu C. et al. 2018).

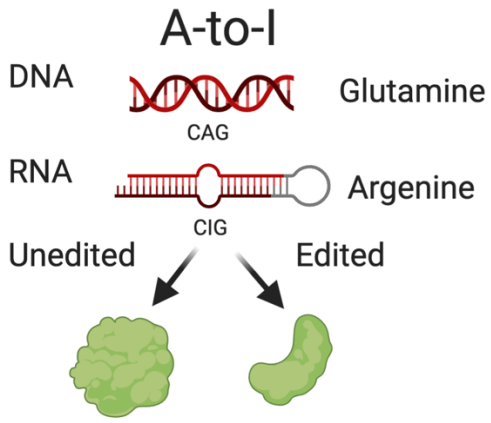



Figure 1: RNA editing. During transcription change of base A to I can result in an altered protein with function different than the original protein. Created with Biorender.com.

## Approaches for SNV detection

Recent advancement in high-throughput next-generation sequencing have provided large amounts of high-quality RNA-seq data (Alioto T. S. *et al.* 2015, Boudellioua T. S. *et al.* 2019, Xu C. *et al.* 2018, Zhao *et al.* 2019). Traditionally, rule-based methods as implemented in SNPiR, RADIA, and eSNV-Detect have been used to reduce the burden of false positives in SNV prediction (Bosio M. et al. 2019, Piskol et al. 2013, Radenbaugh A. J. et al. 2014). However, these heuristic algorithms are based on stringent filters which make them sensitive to the alignment

method, quality control measures, and variant calling algorithms. Similarly, ensemble approaches that integrate output from a collection of tools like VaDiR (Neums et al. 2018) and appreci8 (Sandmann et al. 2018) primarily combine outputs from diverse variant callers, and therefore, compromise on the candidate site counts to achieve a low false-positive rate. Both hard-filtering and ensemble approaches require features that are directly correlated with variant identity and do not capture the non-linear relationship between variant features and likelihood of being a true variant (Coudray A. et al. 2018). To address these issues, tools like RVBoost have framed SNP filtering as a ranking problem by using regression models based only on true SNPs and thus, loses the discriminative power that can be gained from the non-SNP samples. Similarly, supervised algorithm SmartRNASeqCaller (Bosio et al. 2019) uses a pre-trained Random Forest algorithm, which can suffer from class bias and sample bias (Ainscough et.al. 2018, Luo et al. 2019). Notably, a major aspect ignored in these automated algorithms is hyperparameter optimization to achieve the best performance for their chosen model architectures. The real challenge lies in framing the biology of variants as a befitting machine learning problem such that generalized models can be produced that adhere to biological appropriateness under diverse conditions (Wu C. et al. 2019).

Here, we present our Trinity Cancer Transcriptome Analysis Toolkit (CTAT) Mutations Pipeline, a comprehensive, robust, and automated pipeline that includes the GATK best practices variant calling pipeline coupled with machine learning-based variant refinement to maintain high sensitivity for variant prediction while producing a low false-positive rate (Romagnoni et al. 2019). We framed the variant refinement problem as a class-imbalanced classification problem to provide better accuracy than hard-filtering, regression, and variant-ranking methods. To achieve generalized models, we emphasized hyperparameter optimization, which cannot be inherently

trained by machine learning models. Broadly, high dimensional data with small sample size tend to overfit training data and generate biased models. Additionally, the low variance of small training data can lead to less generalizable models which under fit test data. To overcome this small training data problem, we used the publicly available dbSNP database to label variants instead of using the known variants as labels without requiring the construction of any specialized training dataset. This also makes our models potent to changes in cancer tissue types.

We compared several machine learning models - four ensemble tree-based models (Gradient Boosting, Stochastic Gradient Boosting, Adaboost, and Random Forest) and three linear models (SVM with linear kernel, SVM with Radial Basis Function kernel and Logistic Regression), we found that Gradient Boosting performed best compared to other models. We used Genome in the Bottle (GIAB) gold standard dataset for training and the reference high confidence variants for independent validation of our model. We also showed that despite being trained on high confidence SNPs, our model achieved the best accuracy even when benchmarked with tumor-matched exome samples and five types of cancer tissues. Interestingly, our model works equally well for INDELs, suggesting that rather than learning properties of SNPs, the Bayesian hyperparameter optimization on 10-fold cross-validation has learned to discriminate a variant from a non-variant. The accuracy in performance suggests that our models are trained and optimized as a generalized model that has appropriately captured the complex interactions between our chosen features that are sample and tissue independent. Additionally, we use Open-CRAVAT's CHASM, VEST, and ClinVar annotations to identify cancer-relevant variants. Finally, we provide an interactive stand-alone web-based Integrative Genome Viewer (IGV) based visualization to facilitate further evaluation of cancer-relevant variants in the context of the aligned RNA-seq read

alignment evidence. We believe that the prediction of reliable variants in our CTAT Mutations Pipeline can contribute to enhancing prognostics and early diagnosis in precision cancer treatment.

# Chapter II

# Research Methods

## Overview of CTAT Mutations pipeline

Starting with RNA-seq FASTQ files, the CTAT Mutations pipeline provides a complete analysis of genetic variants including alignment, quality control, variant calling, filtering, and cancer variant detection and visualization (Figure 1). We used the splice-aware STAR 2-pass sequence alignment procedure. Following the alignment, CTAT Mutations performs Genome Analysis Toolkit (GATK) specified best practices to extract maximum variants (Poplin R. et al., (2017), Piskol et al., (2013)). In CTAT Mutations, we leverage GATK's HaplotypeCaller algorithm (Poplin R. et al., (2017)) for initial variant prediction. In addition to GATK based variant attributes, we incorporated RNA-seq associated attributes to account for the distance from splice junction, variants in duplicate reads, variants in homopolymers, variants in repeats as given in UCSC RepeatMasker database, multi-mapped read fraction, variant allele fraction and entropy of a sequence within a window of 7-10 base pairs centered at variant position (Adetunji M. O. et al., 2019, M. Bosio et al., 2019, Jiarui Ding et al. 2011, Radenbaugh et al., 2014, Chen Wang et al., 2014) (Table 1). A major type of RNA modification happens through RNA-editing which can change the resulting protein sequence, splicing patterns, target miRNA sites, and RNA stability by facilitating A-to-I or C-to-U substitutions which reflects in the RNA-seq as apparent A-to-G transition or the reverse complement T-to-C variants. These substitutions can be misannotated as genetic variants and contribute to the high FP rate in RNA-seq (Hwang S. 2015, Kung et al. 2018). To avoid this, we used the REDIportal database, the largest and most comprehensive database for

RNA-editing sites (Ernesto Picardi et al., 2016), to annotate RNA-Editing events in the GATK variant calls. Additionally, we used the dbSNP database to annotate common variants. To further annotate the likely biological impact and relevance in cancer, we used CRAVAT (Masica, D. L. et al., (2017)), VEST (Carter et al. 2013), CHASM (Carter et al. 2009), and SNPeff (Cingolani, P. et al., (2012)) annotation tools along with COSMIC (Tate, J. G. et al., 2018) database. Finally, to facilitate the interpretability of variants in relation to the reads, we generated visualization reports of detected variants through Integrative Genomics Viewer (IGV) reports tool (Robinson et al., (2017)).

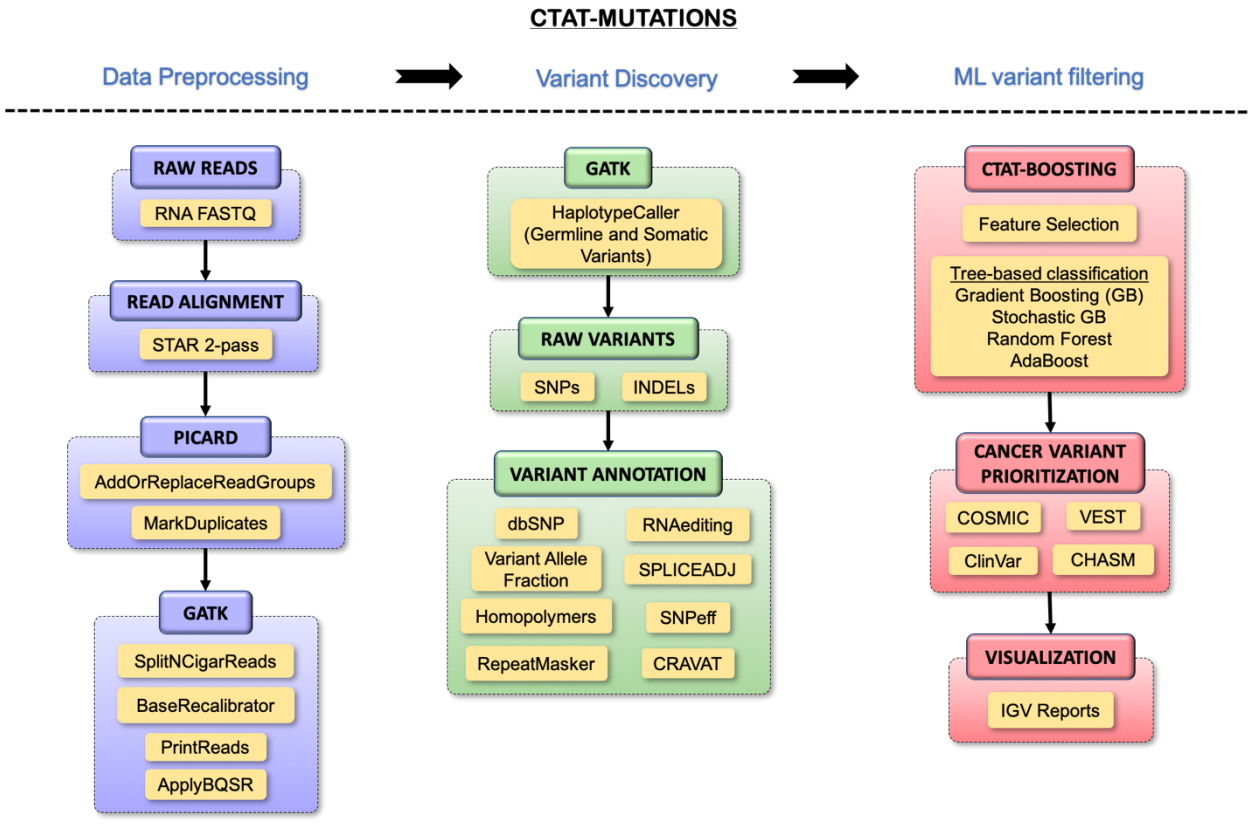



Figure 1: A schematic overview of the CTAT Mutations pipeline. The pipeline consists of three parts – Data preprocessing based on GATK Best Practices pipeline for variant calling (Blue),

GATK HaplotypeCaller for variant calling and variant annotations (Green) and Machine Learning-based variant classification and visualization (Red).

Following variant annotation, CTAT Mutations performs variant refinement through machine learning classification algorithms provided as CTAT Boosting Suite. After comparing seven algorithms - Gradient Boosting, Stochastic Gradient Boosting, Adaboost, Random Forest, Support Vector Machines (Linear and RBF kernel), and logistic regression, gradient boosting is used as the default variant refinement algorithm based on the performance measure. All algorithms have been optimized using Bayesian hyperparameter optimization to achieve the optimal F1 performance which helps reduce false positives in variant prediction. The resulting Variant Call Format (VCF) file is benchmarked by the CTAT Benchmarking suite which can compare the RNA-seq derived variants with either high confidence reference or tumor-matched exome sequencing data. To reduce the noise emanating from reads with low read depths, we used Samtools to remove variants with RNA-seq read depth and the corresponding exome read depth less than 10. The benchmarking performance is measured through Precision-Recall (PR) and Receiver Operating Curve (ROC) plots and under 5 accuracy statistics - Sensitivity, PPV, Accuracy, MCC, and F1 measure.

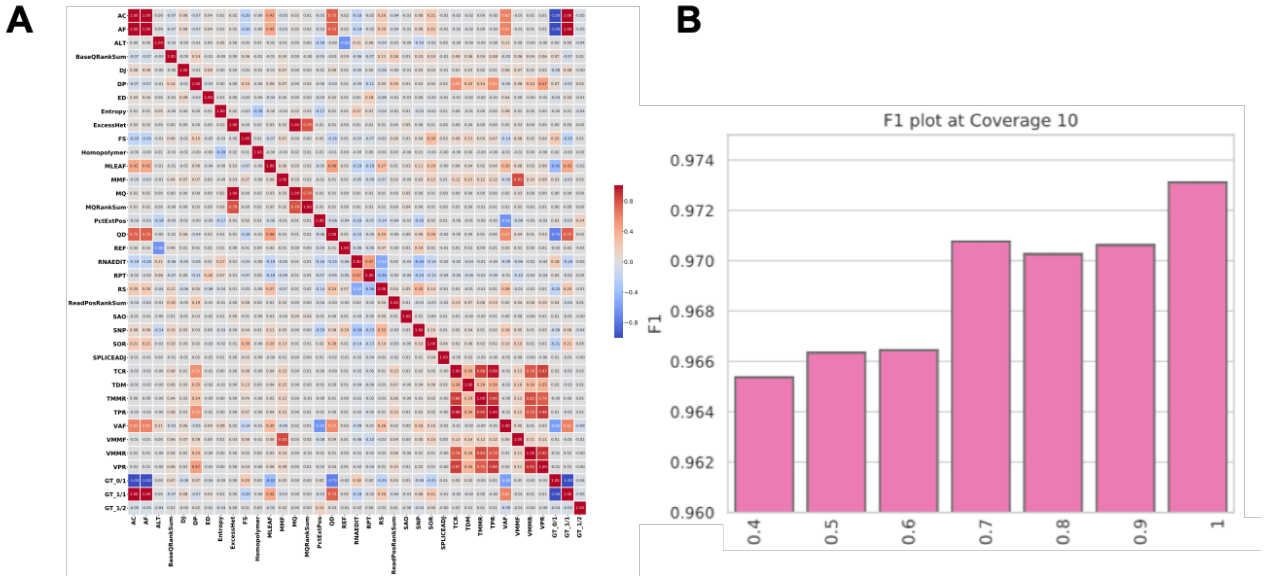
## Framing Variant Filtration as an ML problem:

The application of machine learning in biology is challenged by the difficulty in fitting the biology in an ML framework that is interpretable and can provide novel biological insights. Compared to a standard classification problem, variant detection lacks a clean and large dataset with ample labels of true and false variant sites for training unbiased ML models. Additionally, the lack of knowledge of true negative samples can skew training by lowering the discriminative power of classifiers. In biology, we largely deal with "True" vs "Unknown" class as opposed to a "True" vs "False" class. It is difficult to know if a non-variant label is a benign modification or a technical artifact that might come specifically from sequencing techniques, alignment method, or other technical aspects of data generation. If negative samples are available, they are usually available in low numbers, thus, creating a class imbalance during classification and consequently, increasing the risk of misclassification of the minority class. Even if both the issues are satisfied, due to inherent heterogeneity in cancer biology, the models are customarily not generalizable across samples raising a pressing need for appropriate problem framing. For example, training on somatic variants may not work well on germline variants, training on SNPs may not work well on INDELs, and training on a particular sequencing technology may not transfer to another. Unfortunately, it is a complicated procedure to develop generalized models with parameters that provide optimal performance universally. In such cases, the accuracy metrics are equally important in the optimization process because tools can provide high specificity at the cost of sensitivity. Considering these issues, we framed the SNP filtering problem as a class imbalance classification problem that is optimized for the F1 score so that the sensitivity is not compromised in an effort to achieve a high true positive rate.

To examine the performance of CTAT Mutations, we used GIAB gold standard dataset (NA12878) which is accompanied by corresponding known high confidence genetic variants (Zook et al. 2014, Zook et al. 2019). To address the training data size use, we used the publicly available common variant annotations from dbSNP as labels for training our models. This labeling ensures that the dataset is large enough to avoid overfitting and eliminates the need to construct any specialized dataset by combining multiple samples for model training. Note that the labels used for training on GIAB data are not GIAB specific but rather are the subset of common variants as defined in dbSNP. For training, we used the 2,000,789 variants predicted by the GATK variant calling step that included both SNPs and INDELs. Out of those, 108,572 variants present in dbSNP were labeled as true variants whereas the remaining 51,691 were labeled as non-variants. The number of positive samples are approximately twice the number of negative samples which creates a class imbalance, in which case, ML classifiers can have a higher probability of misclassification of minority class compared to majority class as it is more accurate to predict the majority class. Class imbalance is particularly important in variant refinement because predicting non-variant as a variant can generate higher false positives. To address this issue, we first stratify the training and test data to have equal distribution of both classes. Additionally, we penalize the model with the ratio of size of variant class to size of non-variant class every time the non-variant class is misclassified. For prediction on different datasets, we use only 10% of the training data and predict on the entire dataset. We performed independent validation by benchmarking the predictions against unseen high confidence variants provided by the GIAB Consortium (n=599,232). To adjust for lack of knowledge of true variants introduced by dbSNP common variant labels, we used a new reference set (GIAB high confidence variants) instead of using a hold-out set for validation.

## Feature Selection

We selected an initial set of 37 features that include a combination of GATK specific and RNA-seq specific features. In addition to GATK annotations, we added 19 RNA-specific features - Homopolymer, Entropy, PctExtPos, ED, DJ, RNAEDIT, RPT, RS, SAO, SOR, SPLICEADJ, TMMR, VPR, TPR, TCR, TDM, VAF, MMF, VMMF, VMMR, VAF features to appropriately frame the context of RNA transcripts (Table 1).



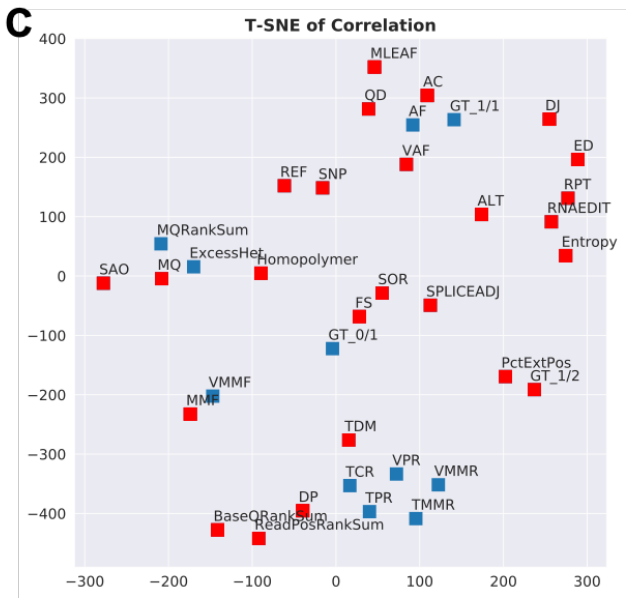



Figure 2: A) Correlation matrix of 37 features. A high positive correlation is shown in red, high negative correlation is shown in blue and weak correlation is shown in grey. B) Performance of gradient boosting at different correlation coefficients. C) t-SNE plot of feature based on correlation distance metric. The selected features are shown in red whereas rejected features are shown in blue. GIAB reference high confidence variants are used for benchmarking.

To capture non-redundant information that defines genetic variants in RNA-seq, we first performed correlation-based feature selection across 37 candidate variant features (Figure 2A). Highly correlated features not only provide duplicate information about variants but may also skew a classifier in favor of repeated features. Based on the correlation coefficient ranging from 0.4 to 1 at an interval of 0.1 we selected 17 features. For a chosen correlation coefficient, we removed one feature from a feature pair that had a higher correlation coefficient. We plotted the maximum F1 score for unoptimized gradient boosting classifier and found that a maximum allowed correlation coefficient 0.7 among selected features allowed for retaining high prediction accuracy with minimal correlation among features, yielding a set of 17 features (Figure 2B). On plotting the

correlation-based t-SNE plot of the complete feature set, we noticed that the rejected features lie within the clusters of selected features (Figure 3). These features serve as an input for the seven downstream classification models.

## Hyperparameter Optimization

After selecting the features, we performed hyperparameter optimization using Gaussian Processes on each of the seven classification algorithms: four tree-based ensemble models - Gradient Boosting (Boosting), Stochastic Gradient Boosting (SGB - variant of gradient boosting), Adaboost (Boosting), Random forest (RF - Bagging) and three linear classification models - Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel) and logistic regression. While logistic regression, SVM (Linear and RBF kernel) and bagging models (Random Forest) have fewer hyperparameters that can be optimized by exhaustively searching the parameter space, tree-based ensemble boosting (Gradient Boosting, Stochastic Gradient Boosting and Adaboost) models have a large number of hyperparameters which demand specialized Bayesian hyperparameter optimization to provide robust performance. Since boosting models depend on the sequential performance of trees, they may lead to overfitting if the parameters are not tuned properly. To account for overfitting, we split 33% of the dataset as test data and trained on the remaining samples with 10-fold cross-validation with F1 measure as the optimization criterion such that the sensitivity of prediction is not compromised. For implementation, we used Python library scikit-optimize (Skopt) with an acquisition function expected improvement (EI) for 100 rounds of optimization. The convergence plot of Gaussian processes optimizing

hyperparameters showed a consistent decrease in the negative F1 score (Figure 3A). Despite training on dbSNP common variant labels, the log loss function on training and test data showed that while training loss continued to decrease, the test loss was invariant at 20 epochs, suggesting that there is no overfitting (Figure 3B).

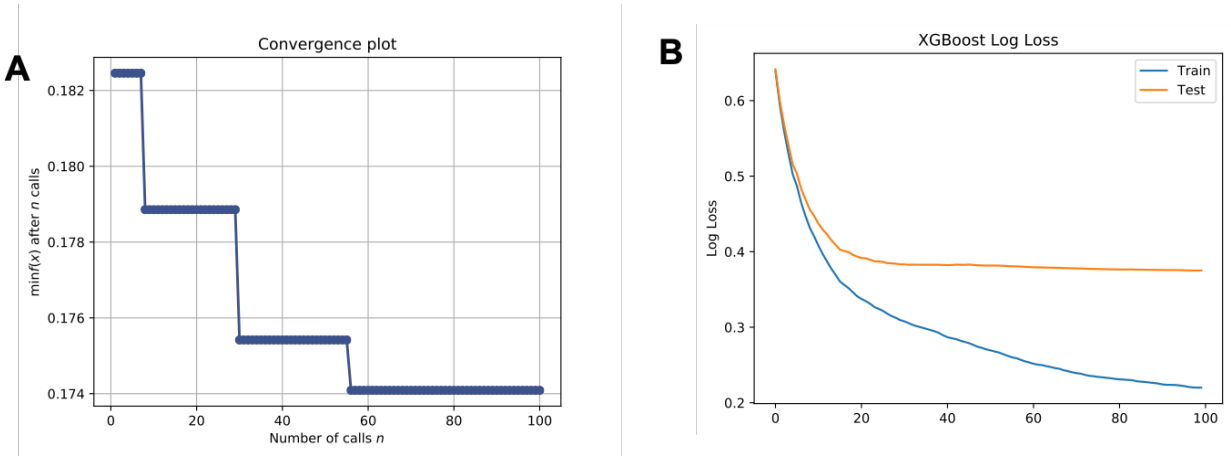



Figure 3: A) Convergence plot for gradient boosting optimization function. B) Log loss error on training and test data

## Benchmarking variant prediction accuracy across all methods

After inferring hyperparameters, we performed independent validation by benchmarking the model predictions on GIAB and five different cancer tissues. For GIAB data, we performed two-way benchmarking against reference high confidence regions provided by the GIAB consortium and matched exome sequencing data. Since reference high confidence data is generally not available for cancer samples, we followed benchmarking against WES samples.

For extracting variants in the matched-exome, we used BWA-MEM alignment tool to align WES samples to GRCh37 reference genome as the exome data exhibits low divergence. Following the alignment step, we used GATK best practices of PICARD tools for preprocessing read quality and annotation, and HaplotypeCaller for variant calling. Due to lack of knowledge of true negative variants, we used samtools to reduce noise in the data by removing variants with RNA-seq and corresponding WES read-depth less than 10. To support the issue of missing true negatives, we used five different accuracy metrics - Sensitivity, Positive Predictive Value, Accuracy, F1, and Mathew's Correlation coefficient, which are commonly used ML metrics for assessing the predictive power of a classifier. We generated a comprehensive report of model performance through Precision-Recall (PR) Curves and Receiver-Operating Curves (ROC).

The performance metric for model evaluation are defined as follows:

TP: Predicted variants also found in reference set and has the same genotype

FP: Predicted variants not found in the reference set or has a different genotype

FN: variants in the reference but not called by variant caller

$$SN \Rightarrow \frac{TP}{(TP + FN)}$$

$$PPV \Rightarrow \frac{TP}{(TP + FP)}$$

$$Accuracy \Rightarrow \frac{TP}{(TP + FP + FN)}$$

$$F1 \Rightarrow \frac{2 * SN * PPV}{(SN + PPV)}$$

$$MCC \Rightarrow \frac{(TP * TP) - (FP * FN)}{(TP + FP) * (TP + FN)}$$

Finally, we used IGV-reports visualization for easy navigation of the predicted variants in the context of RNA transcripts on an HTML browser.

## Model Selection

For our analysis, we chose GATK HaplotypeCaller used for variant calling as it showed significantly high performance compared to Mutect2 variant caller which is more specifically targeted to somatic variant calling as opposed to calling all likely variant sites (Supplementary Figure 1), previously also shown in (Liu et al, 2019). When benchmarked with both GIAB high confidence variants and matched-exome variants, we see similar trends in performance with Gradient Boosting (XGBoost) exhibiting the best F1 score, followed by Random Forest, SGBoost and Adaboost with similar F1 score, and linear models exhibiting the lowest F1 score (Figure 4A, 4B). Compared to gradient boosting, we noticed that SGB and RF models provided slightly higher PPV but at the cost of reduced sensitivity, which led to an overall decrease in their F1 scores (Figure 4C, 4D). Of note, even after training SNPs and INDELs together, all four tree models perform better than the linear models suggesting that the tree models are able to capture the non-linear relationships between the features that define a true variant.

Apart from the performance, the ensemble models handled various data types including numerical, non-numerical, mixed, categorical as opposed logistic regression and SVM which required data normalization. Tree-models can also handle non-normal distributions, so we did not

have to scale the data before running the models. Out of 17 selected features, 7 were categorical or nominal variables which couldn't be utilized by SVM even with a non-linear kernel. Thus, we retained the tree-based models as primary models in our CTAT Boosting suite which performs variant refinement in CTAT Mutations pipeline. We used Python XGBoost library to implement Gradient Boosting and Python Sklearn library to implement the rest of the six models.

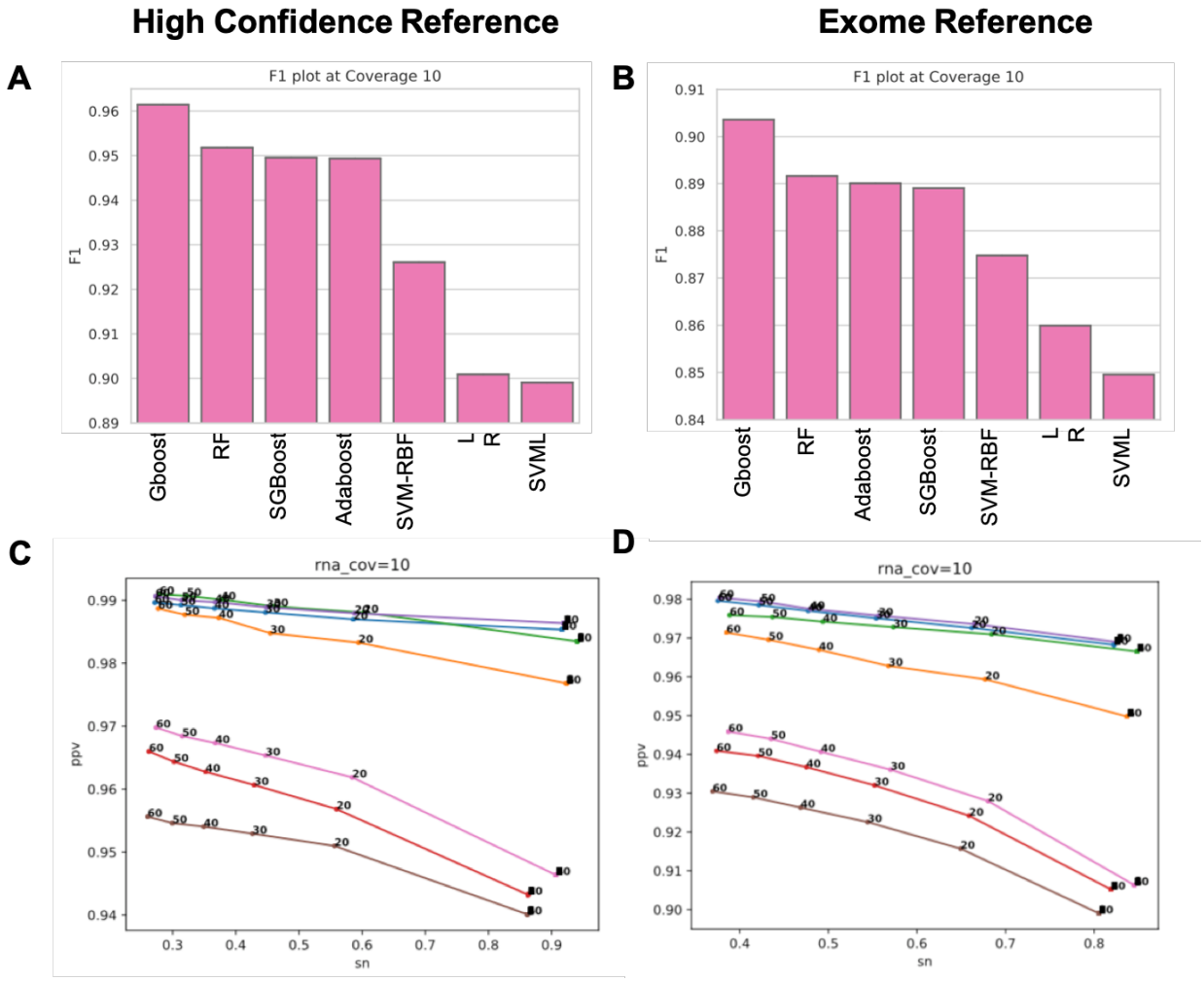



Figure 4. Model Selection: Comparison of 7 machine learning classifiers for variant refinement A) F1 score of GIAB benchmarked with high confidence reference variants B) F1 score of GIAB benchmarked with matched WES samples C) Precision-Recall curve of GIAB benchmarked with

high confidence reference variants D) Precision-Recall curve of GIAB benchmarked with matched WES samples. At a minimum read coverage of 10, the curves are plotted up to minimum coverage of 60 with an interval of 10.

Apart from the performance, the ensemble models handled various data types including numerical, non-numerical, mixed, categorical as opposed logistic regression and SVM which required data normalization. Tree-models can also handle non-normal distributions, so we did not have to scale the data before running the models. Out of 17 selected features, 7 were categorical or nominal variables which couldn't be utilized by SVM even with a non-linear kernel. Thus, we retained the tree-based models as primary models in our CTAT Boosting suite which performs variant refinement in CTAT Mutations pipeline. We used Python XGBoost library to implement Gradient Boosting and Python Sklearn library to implement the rest of the six models.

# Chapter III

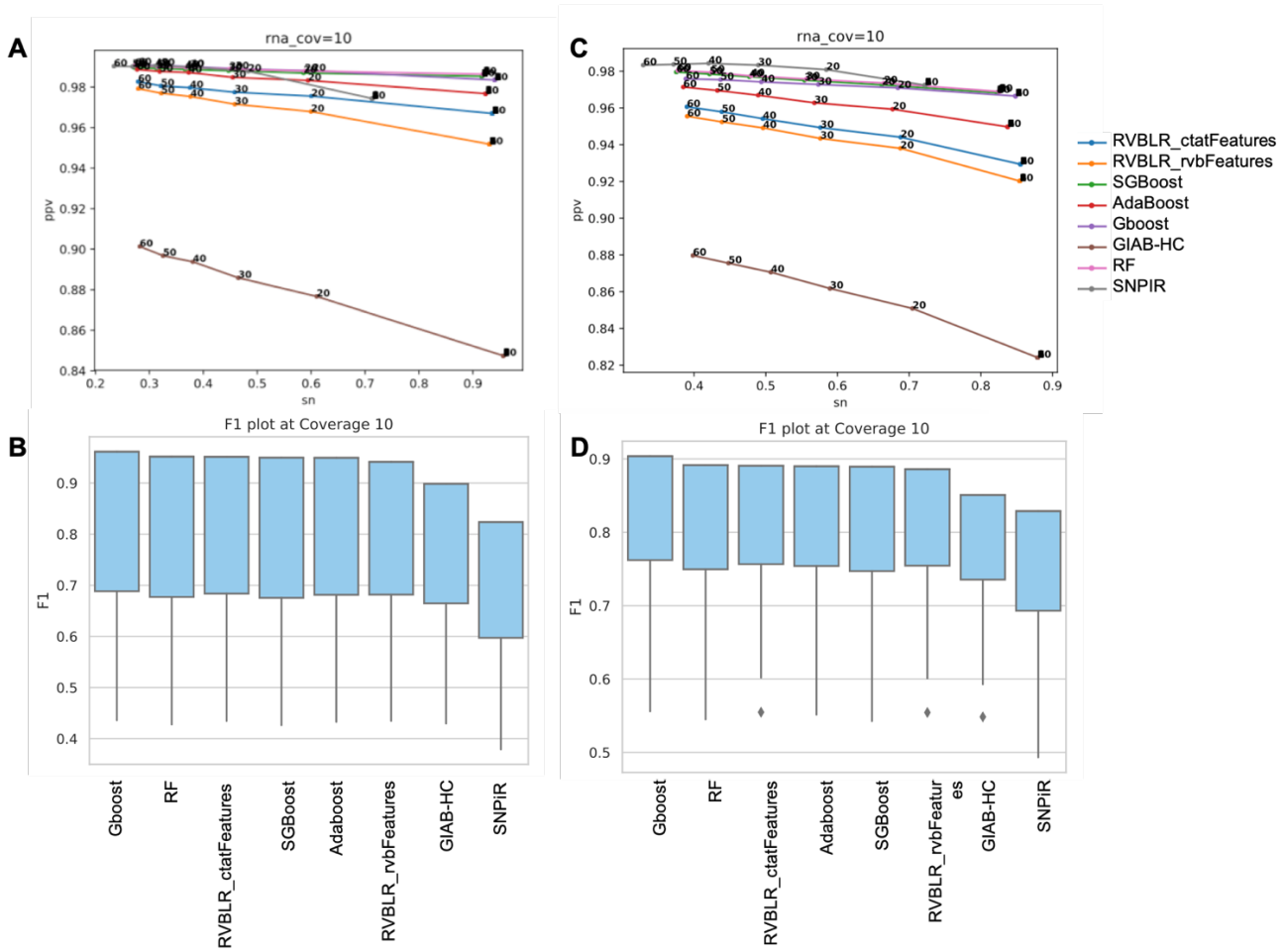
# Results

Our primary goal was to produce a robust pipeline that could differentiate true variants i.e. both SNPs and Indels from false positive artifacts in RNA-seq. To access the veracity of our CTAT Mutations pipeline, we compared the performance of the machine learning models in the pipeline with other popular tools used for variant filtration on GIAB gold standard data. Additionally, to test the generalizability of our model, we compared the performance of our pipeline with other tools on five types of cancers with diverse tissue of origin and consequently, diverse heterogeneity. As demonstrated below, CTAT Mutations can provide an accurate and reliable prediction of variants in clinical settings and enable the use of SNP arrays for personalized medicine.

## CTAT Mutations achieves the best accuracy on GIAB Gold Standard dataset

First, we evaluated the performance of the four CTAT Boosting models against four variant filtration methods - RVBoost (Variant Ranking - with original features prescribed by the tool RVBLR_rvbFeatures), RVBoost (Variant Ranking - with CTAT features - RVBLR_ctatFeatures), SNPiR (hard-filtering) and baseline GATK Best Practices workflow on the GIAB dataset. Since our models are trained on general definition of variants which includes both SNPs and INDELs, we evaluated CTAT Boosting models on both variants combined and later evaluated the performance only SNPs by excluding INDELs.

**SNPs + INDELs**


A
rna_cov=10
ppv
sn
C
rna_cov=10
ppv
sn
RVBLR_ctatFeatures
RVBLR_rvbFeatures
SGBoost
AdaBoost
Gboost
GIAB-HC
RF
SNPIR
B
F1 plot at Coverage 10
F1
Gboost
RF
RVBLR_ctatFeatures
SGBoost
Adaboost
RVBLR_rvbFeatures
GIAB-HC
SNPIR
D
F1 plot at Coverage 10
F1
Gboost
RF
RVBLR_ctatFeatures
Adaboost
SGBoost
RVBLR_rvbFeatures
GIAB-HC
SNPIR

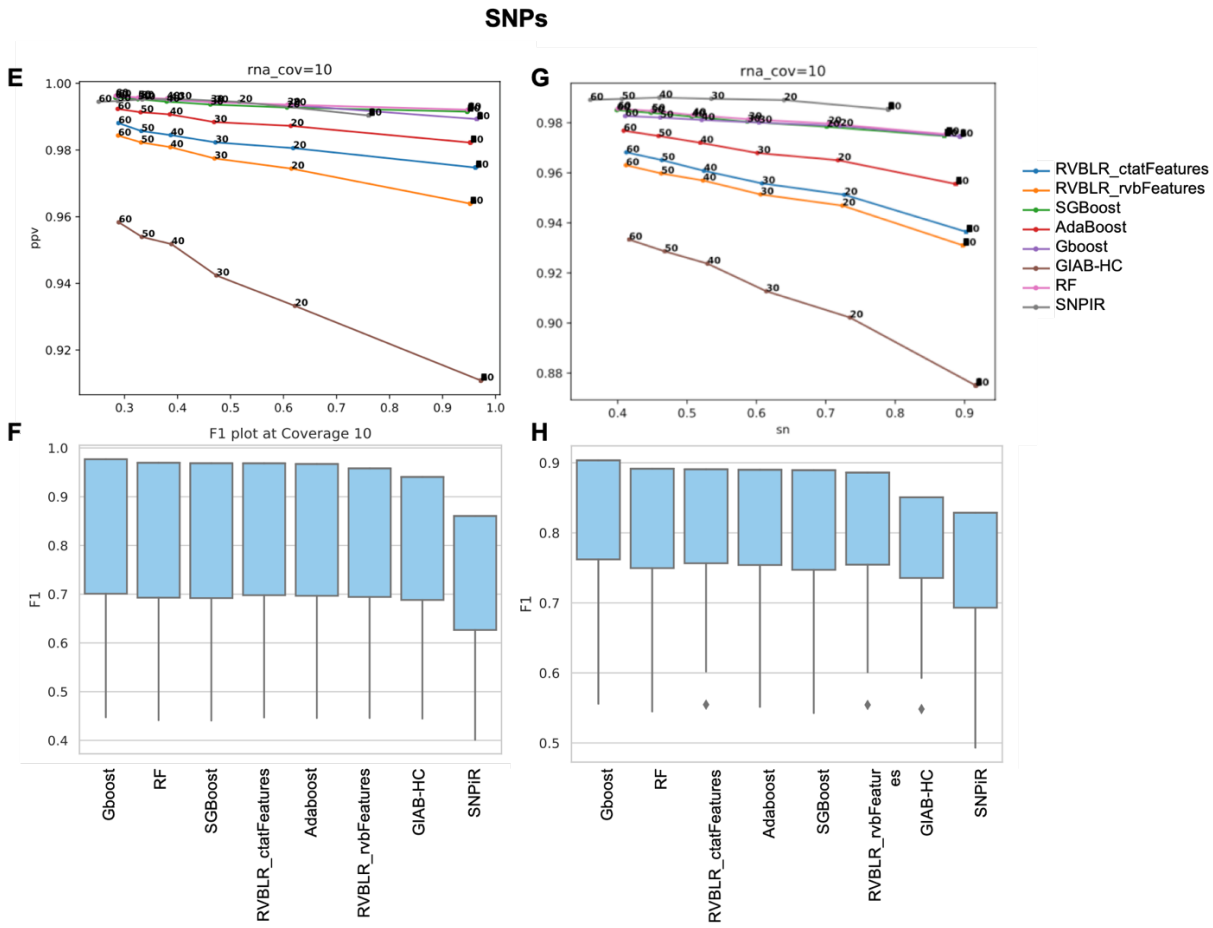



Figure 5: Model Comparison A) PR Curve of GIAB variants benchmarked against high confidence variants B) F1 scores of benchmarking variants against high confidence variants. C) PR Curve of GIAB variants benchmarked against matched WES. D) F1 scores of benchmarking against matched WES E) PR curve of benchmarking only SNPs against high confidence variants. F) F1 scores of benchmarking SNPs against matched WES. G) PR curve of benchmarking only SNPs against matched WES. H) F1 scores of benchmarking SNPs against matched WES. The lowest F1 value corresponds to minimum read depth of 60 and the highest value corresponds to minimum read depth of 10.

Essentially, the baseline GATK best practices workflow provided a considerably good performance on the GIAB dataset with precision (PPV) of 84.74% and recall (sensitivity) of 95.63% (Figure 5A). Using GIAB high confidence SNPs for benchmarking, Gradient Boosting outperformed all the methods by providing precision of 98.09% and recall of 93.33% and a resulting F1 score of 95.67% followed by RF (Precision = 98.63%, Recall = 91.9%, F1= 95.17%). RVBLR_ctatFeatures had a comparable performance with RF (Precision = 96.69%, Recall = 93.61%, F1= 95.13%), followed by SGBoost (Precision = 98.53%, Recall = 91.62%, F1= 94.95%) and Adaboost (Precision = 97.68%, Recall = 92.33%, F1= 94.93%) (Figure 5B). All four CTAT Boosting algorithms provided similar performances and showed an improvement over the baseline model suggesting that our models have trained well on the dbSNP labels. The performance of CTAT models was followed by RVBLR_rvbFeatures (Precision = 95.18%, Recall = 93.11%, F1= 94.14%) showing an improvement over the baseline model but slightly less than CTAT Boosting models. Compared to baseline performance, SNPiR greatly improved the precision by 13% (97.43%) however, with a decrease in sensitivity by ~13% (71.31%). Overall, this leads to a reduced F1 score of SNPiR (82.35%) when compared to baseline GATK performance (89.85%). This shows that even though SNPiR is able to provide a substantial improvement in the precision value, it comes at the cost of sensitivity of prediction. Hard-Filtering methods focus on removing candidate sites based on manually chosen thresholds by an expert or commonly seen behavior in the field, but it is a difficult task to manually comprehend the complexity of importance of each feature and corresponding threshold in defining a variant.

We then evaluated the performance of CTAT Boosting models by comparing with matched-exome sequence as reference and observed a similar trend with Gradient Boosting and RF outperforming all the other models (Figure 5C, 5D). The PPV of Gradient Boosting improved by ~14% with a minimal tradeoff of sensitivity (~3%). RVBLR_ctatFeatures showed better performance than Adaboost and SGBoost, followed by RVBLR_rvbFeatures. SNPiR still showed worse performance than baseline GATK performance. Although, the performance of RVBoost is similar to the four CTAT Boosting classification models. A consistent improved performance of RVBoost with CTAT features showed that the RNA context specific features used by CTAT Mutations have added value to the performance of RVBoost.

Finally, we replicated the analysis by removing INDELs to access the performance exclusively on SNPs. Both Gradient boosting and RF showed best performance compared to other algorithms. Compared to all variants, we observed a decrease in the performance of RVBLR_ctatFeatures compared to SGBoost when benchmarked against high confidence regions (Figure 5E, 5F). Similarly, the performance of RVBLR_ctatFeatures decreased further when benchmarked against the matched WES variants (Figure 5G, 5H). SNPiR continued to perform worse than baseline performance. Nevertheless, the four CTAT Boosting models consistently provided improved precision at a minimal tradeoff of sensitivity.

CTAT Mutations exhibits robust performance on cancer samples with diverse tissue-of-origin

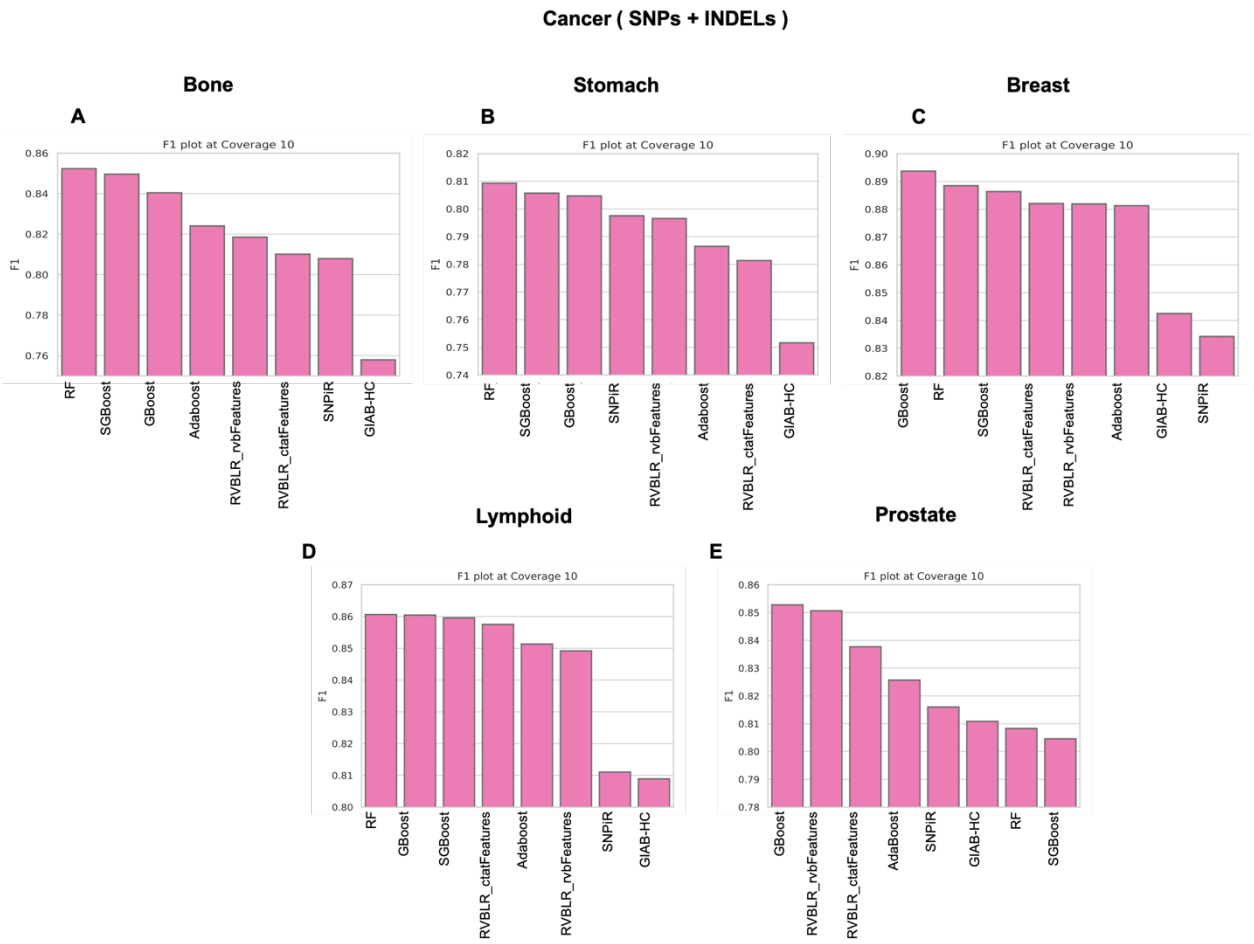



**Figure 6.** A-E) F1 score of 8 algorithms across 5 cancer tissue types.

Following the analysis of GIAB dataset, we assessed the performance of CTAT Mutations on five different cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) - representing bone (CADOES1), breast (HS274T), hematopoietic and lymphoid tissue (K562), prostate (LNCAPCLONEFGC) and stomach (SNU520) cancer. These cancer cell lines have different intrinsic gene expression patterns and harbor different mutational profiles relevant to the corresponding cancer types and hence reflect different challenges to the CTAT mutation pipeline

for variant detection based on RNA-seq data. Since, unlike the GIAB data, there are no truth or gold standard variant sets available for these samples, we used GATK HaplotypeCaller variant calls from matched WES samples as the reference truth set for benchmarking as demonstrated with GIAB dataset.

In these five samples, we observed that Random Forest exhibited best performance in 3 samples (Bone, Stomach and Lymphoid tissue) and Gradient Boosting exhibited best performance in the other two samples (Breast and Prostate) (Figure 6A-E). Stochastic gradient boosting was among top three performances in all but Prostate cancer. Adaboost was the weakest of the four CTAT Boosting algorithms but always performed better than baseline. SNPiR performed better than baseline GATK HaplotypeCaller in all the samples except Breast cancer. Similarly, both RVboost algorithms (RVBLR_rvbFeatures and RVBLR_ctatFeatures), performed better than baseline in all samples and better than SNPiR in all but Stomach cancer.

Out of all the samples, Prostate cancer was the most striking in that two of our best performers- Random Forest and Stochastic Gradient Boosting showed an F1 score less than baseline GATK HaplotypeCaller. Overall, we see a consistent performance by gradient boosting algorithm across all tissue types by performing better than both RVBoost algorithms, SNPiR hard filtering and baseline GATK variant caller. Even when Gradient Boosting was not the best algorithm, it displayed an F1 score within 2% of the best performance. We see a similar trend with Random Forest, Gradient Boosting and Stochastic Gradient Boosting performing being the top three algorithms in predicting SNPs (after removing INDELs), except in Prostate cancer where

only Gradient Boosting retains its accuracy (Supplementary Figure 2). Thus, Gradient boosting is robust to changes in tissue types.

# Chapter IV

## Discussion

Here, we developed CTAT Mutations, an end-to-end variant detection pipeline that performs variant calling, variant refinement and benchmarking to reduce the false positive load in predicting cancer variants. To encapsulate the complexity associated with a true SNP in RNA-seq data, we designed features that neutralize errors introduced due to alignment artifacts near splice junctions, reverse transcription, RNA editing, low variant allele frequencies, and repeat regions. Owing to the non-linear and non-obvious nature of these complex features, we developed CTAT Boosting, a collection of non-parametric ensemble tree-based models - Gradient Boosting, Stochastic Gradient Boosting, Random Forest, and Adaptive Boosting that discriminate true variants from likely false positives. To validate our results, we compared the output variants to reference high-confidence data, as well as tumor-matched exome samples under three different accuracy statistics that measure the tradeoff between precision and recall of prediction.

We designed the CTAT Mutations pipeline such that the predictions are robust to RNA-seq data generation and processing errors, tissue types and intra-tumor heterogeneity. Machine learning algorithms are adept at learning complex patterns of features that define a variant, provided that the features themselves capture RNA-seq relevant information in a non-redundant manner. Considering this, we judiciously chose 37 features, 18 were GATK-specific and remaining 19 were newly introduced to embed the context of mutational patterns in diverse RNA-seq samples. Choosing a smaller set of 17 non-redundant features through correlation-based feature selection made our models less prone to overfitting.

The variant refinement problem is particularly challenging to fit in a machine learning framework due to lack of availability of large datasets for training. Even if large datasets are available, the intra-tumor heterogeneity and tissue specificity makes it difficult to generalize models. For example, certain genes are expressed in specific tissues and a model trained on one tissue type may not predict accurately on another tissue type. Moreover, the negative samples represent an "Unknown" label rather than a "False" label in biology and limited "Unknown" labels usually lead to class imbalance. We tackled these issues by using dbSNP annotations as labels for training which renders the entire prediction set as training data, prevents sample bias as the labels are not specific to the tissue and reduces the risk of overfitting. Additionally, this labeling generates a large training dataset and eliminates the need to generate specialized data for model training. The dbSNP annotations are provided as a part of CTAT Mutations annotation process with no additional effort from the user. To address class imbalance, we penalized the wrong prediction of 'Unknown' class which ensures that the positive class is not predicted inordinately. Additionally, hyperparameters lie at the core of any model and can make or break the model. Complex models are required for understanding the complex relationships between variant features, which often come at the cost of a large number of incongruous parameters. We used Bayesian hyperparameter optimization with standard 10-fold cross validation and F1 optimization function so that the recall of prediction is not compromised for increased precision.

On comparing 7 different machine learning models, we found that ensemble tree based models were robust to the mixed data types used in CTAT Mutations. The ensemble tree models are a collection of decision trees that combine majority votes of weak decision trees to predict an

output class. Bagging models like Random Forest use strong learners independently to reduce variance whereas boosting models (Gradient Boosting and Adaboosting) use weak learners sequentially to reduce bias. Thus, if there is a latent bias in the data, which is usually the case with cancer data, the prediction will be poor. This is exactly what we observed in case of the prostate cancer sample. Similarly, if boosting models are not properly optimized, they can lead to overfitting as the base decision trees are optimized locally. The sequential learning process can amplify errors arising from noisy data. Since our boosting models were optimized with Bayesian optimization, we observed robust performance for Gradient boosting and Adaboost, whereas Stochastic Gradient Boosting might have shown reduced F1 score due to high noise in the prostate cancer sample.

Interestingly, our models were trained on dbSNP labels and still achieved good accuracy on GIAB data when benchmarked against high confidence SNPs and matched exome sequence variants. The GIAB dataset consists of samples from healthy donors, but the models trained on GIAB transferred with high accuracy to cancer transcriptomes with diverse tissue of origin. Framing variant refinement as a classification problem has advantages over ranking and hard filtering methods as shown by robust Gradient Boosting performance on different tissue types (GIAB and Cancer samples), variant types (SNPs and INDELs) and reference samples (Exome and High Confidence references). However, current workflows focus on hard filtering, variant ranking and aggregating output from multiple variant callers. Both hard filtering and aggregation methods reduce candidate sites. Aggregation methods are prone to compounding artifacts from the tools and hard filtering methods compromise severely on the sensitivity of prediction, mostly leading to lower F1 score than the baseline method. Variant Ranking models like RVboost lack

robust performance across tissue types because they are not using optimized parameters. Manually optimized models often lead to overfitting on training data and become hard to generalize. Additionally, every time a new feature is introduced, the model needs to be tuned again to fit those features in RVboost. Gradient Boosting model in CTAT mutations consistently performed better than RVboost, SNPiR and the baseline GATK variant caller. To visualize the cancer variants, we provide an IGV report that uses CRAVAT and COSMIC resources to annotate cancer variants (Supplementary Figure 3).

Although RNA sequencing data introduces artifacts that increase the false positive burden in RNA-seq based variant calling, the benefits of using RNA-seq in detecting novel post-transcriptional variants and tissue specific variants in cancer far outweigh the drawbacks. A robust and reliable end-to-end pipeline like CTAT Mutations will contribute to explaining the critical determinants of cancer and developing reliable clinical diagnostic tools in precision medicine.

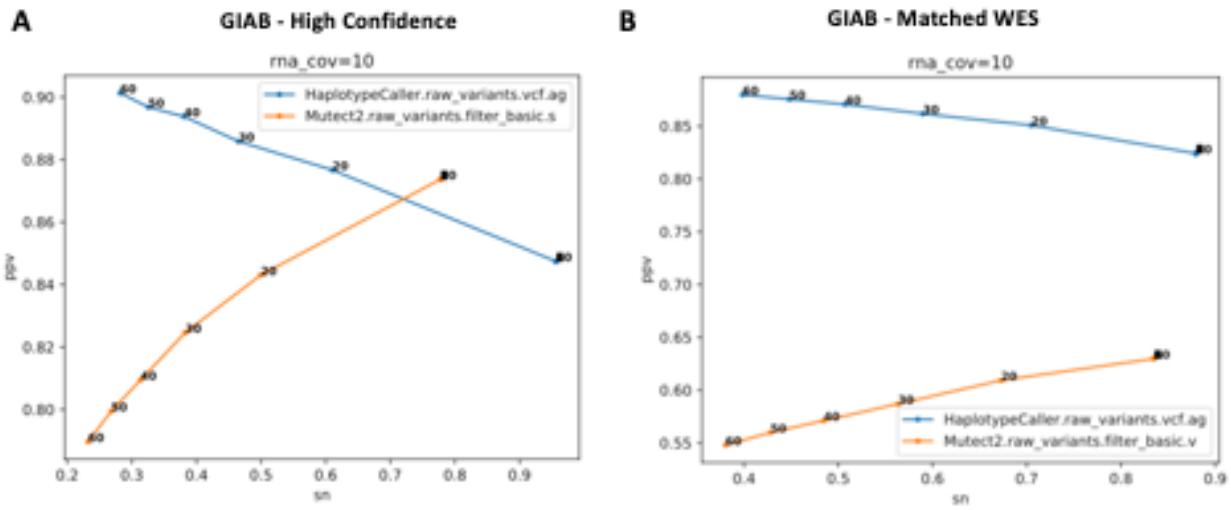



**Figure 1:** Comparison of GATK variant callers - HaplotypeCaller and Mutect2 A) PR curve of benchmarking against high confidence GIAB variants. Best Haplotype performance - Precision = 84.74%, Recall = 95.63%, F1= 89.85%. Best Mutect2 performance - Precision = 87.36%, Recall = 77.74%, F1= 82.27%) B) PR curve of benchmarking against matched WES variants. Best Haplotype performance - Precision = 82.41%, Recall = 87.93%, F1= 85.08%. Best Mutect2 performance - Precision = 81.38%, Recall = 70.31%, F1= 75.44%. In both cases, the performance of Mutect2 declines rapidly with increase in the minimum read depth from 10 to 60.

| Name | Source | Description |
|---|---|---|
| AC | GATK | ALT allele count in genotypes |
| AF | GATK | ALT allele frequency |
| AN | GATK | Genotype specific allele counts |
| ALT | GATK | Alternate Allele |
| BaseQRankSum | GATK | Base quality calculated using Z-score of Wilcoxon rank sum test between REF and ALT alleles |
| DJ (Wang et al. 2014) | CTAT Mutations* | Distance from the variant position to the closest exon-exon junction. |
| DP | GATK | Approximate read depth |
| ED (Wang et al. 2014) | CTAT Mutations* | Number of times the 100bp flanking region of a variant maps to the reference genome with more than 90% sequence similarity as reported by BLAT. |
| Entropy | CTAT Mutations | Entropy for sequence in window of length 7 centered at the variant position |
| ExcessHet | GATK | Excess heterozygosity calculated using Phred-scaled p-value |
| FS | GATK | Fisher's Exact Test |
| GT | GATK | Genotype |
| Homopolymer(Piskol et al. 2013) | CTAT Mutations** | Variant is located in or near a homopolymer sequence |
| MLEAC | GATK | Allele counts measured as Maximum likelihood expectation (MLE) |
| MLEAF | GATK | Allele frequency measured as Maximum likelihood expectation (MLE) |
| MMF | GATK | Multi-mapped read fraction (TMMR / TPR) |

| | | |
|---|---|---|
| MQ | GATK | Root Mean Square of Mapping Quality of the reads |
| MQRankSum | GATK | Mapping quality calculated using Z-score of Wilcoxon rank sum test between REF and ALT allele |
| PctExtPos (Wang et al. 2014) | CTAT Mutations* | Fraction of reads that support variant in the first six bases of the read |
| QD | GATK | Normalized confidence score of variants |
| REF | GATK | Reference Allele |
| RNAEDIT | CTAT Mutations | RNA-editing sites from REDIPortal |
| RPT(Piskol et al. 2013) | CTAT Mutations** | Repeat family from UCSC Genome Browser Repeatmasker Annotations |
| RS | CTAT Mutations | Reference SNP number is a locus accession for a variant type assigned by dbSNP |
| ReadPosRankSum | GATK | Read position bias calculated using Z-score of Wilcoxon rank sum test between REF and ALT allele |
| SAO | GATK | Variant Allele Origin |
| SNP | GATK | SNP types - T:A, T:C, T:G, G:A, G:C, G:T, C:A, C:G, C:T, A:C, A:T, A:G |
| SOR | GATK | Strand bias calculated as 2*2 Symmetric Odds Ratio |
| SPLICEADJ(Piskol et al. 2013) | CTAT Mutations** | Variant is within specified distance of a reference exon splice boundary |
| TCR | CTAT Mutations | Total Covered Reads, meet quality requirements, variant position anywhere in read |
| TDM | CTAT Mutations | Total duplicate marked number of reads that are duplicate marked |
| TMMR | CTAT Mutations | Total multi-mapped reads at site |

| | | |
|---|---|---|
| TPR(Piskol et al. 2013) | CTAT Mutations** | Total Passed Reads, reads that PASS filtering |
| VAF | CTAT Mutations | Variant allele fraction (VPR / TPR) |
| VMMF | CTAT Mutations | Variant-supporting multi-mapped read fraction (VMMR / VPR) |
| VMMR | CTAT Mutations | Total variant-supporting multi-mapped reads |
| VPR(Piskol et al. 2013) | CTAT Mutations** | Reads that PASS filtering that contain the variation |

**Table 1**: List of features used for variant refinement. * RVBoost specific features reimplemented in CTAT Mutations. ** are SNPiR specific features reimplemented in CTAT Mutations. GATK (Poplin et al. 2017, McKenna et al. 2010, DePristo et al., 2011; Van der Auwera et al., 2013)

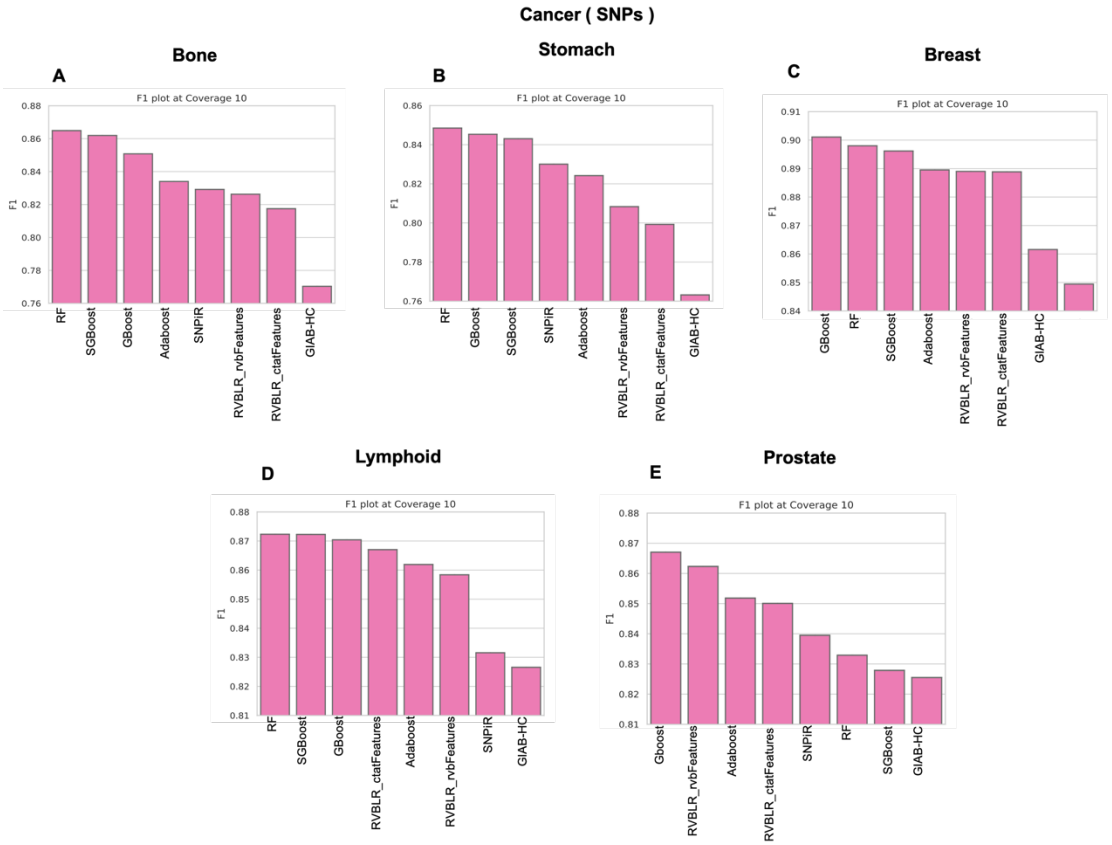



**Figure 2:** F1 performance comparison of variant refinement methods on SNPs in five different cancer tissues. Gradient Boosting and Random Forest were best performers on GIAB gold standard data. Here, we noticed exceptional performance from Random Forest on Bone, Stomach, Breast and Lymphoid tissue. However, Random Forest performance declined on Prostate tissue. Gradient Boosting is robust across tissue types. Gradient boosting was always one of the top three performing algorithms and consistently performed better than RVBoost, SNPiR and baseline GATK.

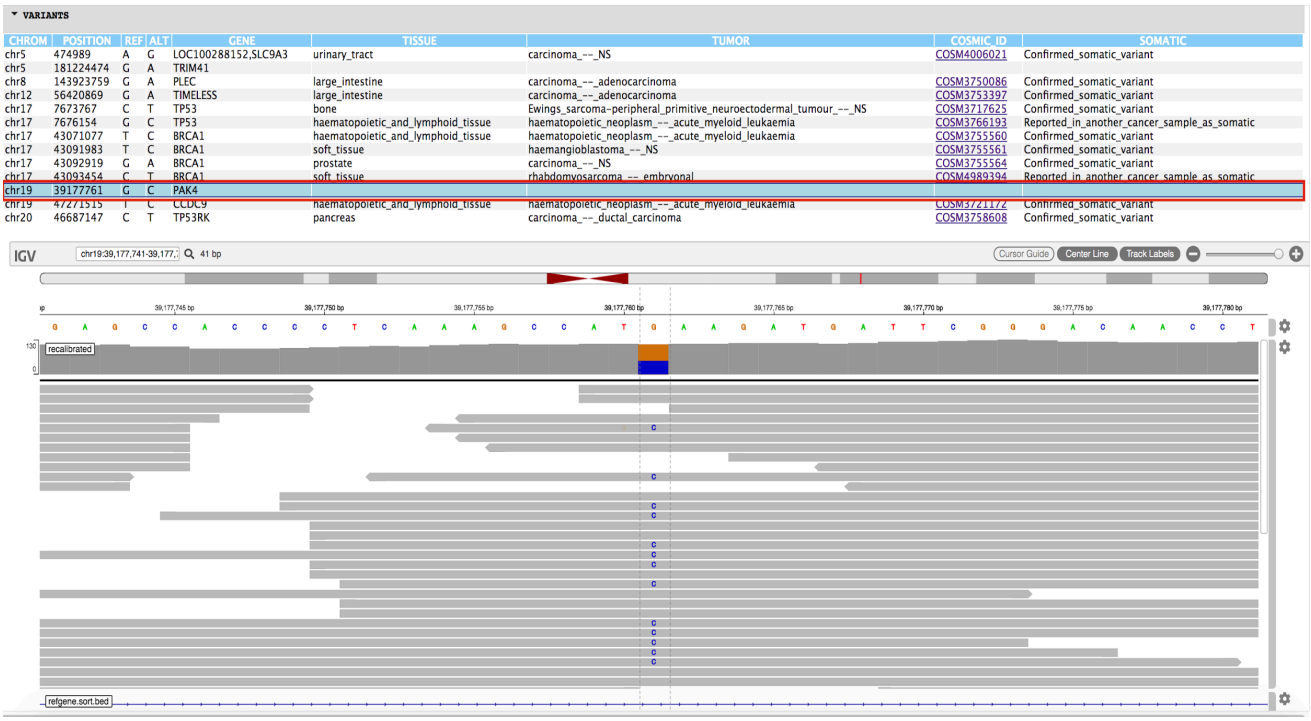



**Figure 3.** IGV visualization with CTAT Mutations: The report displays cancer specific information through COSMIC and CRAVAT databases for SNPs detected by CTAT mutations. The lower part of the report provides information about the exact location of the variant in the context of splice junctions.

Code and Data Availability:

CTAT Mutations: https://github.com/NCIP/ctat-mutations/wiki

GIAB data:

1. Reads:

- Left reads: ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R1_001.fastq.gz
- Right reads: ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R2_001.fastq.gz

2. GIAB bam file : ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/project.NIST_NIST7035_H7AP8ADXX_TAAGGCGA_1_NA12878.bwa.markDuplicates.bam
3. GIAB vcf output file:

   ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/analysis/GARVAN_snps_indels_12172013/project.NIST.hc.snps.indels.vcf
4. GIAB RNAseq: https://www.ncbi.nlm.nih.gov/sra/?term=SRR5665260
5. Reference genome:

   https://data.broadinstitute.org/Trinity/CTAT_RESOURCE_LIB/GRCh37_v19_CTAT_lib_Feb092018.source_data.tar.gz

6. High Confidence Regions (Grch37):

   https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_nosomaticdel.bed

7. Reference vcf : https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_PGandRTGphasetransfer.vcf.gz

References

Adetunji M. O., Schmidt C. J., Lamont S. J., Abasht B. (2019). Variant Analysis Pipeline for Accurate Detection of Genomic Variants from Transcriptome Sequencing Data. *bioRxiv*, 625020. doi:10.1101/625020

Ainscough B. J., Barnell E. K., Ronning P., Campbell K. M., Wagner A. H., Fehniger, T. A., Griffith, O. L. (2018). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nature Genetics*, 50(12), 1735-1743. doi:10.1038/s41588-018-0257-y

Alioto T. S., Buchhalter I., Derdak S., Hutter B., Eldridge M. D., Hovig E., Gut I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, 6, 10001. doi:10.1038/ncomms10001 https://www.nature.com/articles/ncomms10001#supplementary-information

Bosio M., Valencia A., & Capella-Gutierrez S. (2019). SmartRNASeqCaller: improving germline variant calling from RNAseq. *bioRxiv*, 684993. doi:10.1101/684993

Boudellioua I., Kulmanov M., Schofield P. N., Gkoutos G. V., & Hoehndorf R. (2019). DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinformatics*, 20(1), 65. doi:10.1186/s12859-019-2633-8

Chen J., Li X., Zhong H., Meng Y., & Du H. (2019). Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Scientific Reports*, 9, 9345. doi:10.1038/s41598-019-45835-3

Cibulskis K., Lawrence M., L Carter S., Sivachenko, A., Jaffe D., Sougnez C., Getz G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31. doi:10.1038/nbt.2514

Coudray A., Battenhouse A. M., Bucher P., & Iyer V. R. (2018). Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ*, 6, e5362-e5362. doi:10.7717/peerj.5362

Deng N., Zhou H., Fan H., & Yuan Y. (2017). Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget*, 8(66), 110635-110649. doi:10.18632/oncotarget.22372

DePristo M. A. *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491-498. doi:10.1038/ng.806

Ding J., Bashashati A., Roth A., Oloumi A., Tse K., Zeng T., Shah S. P. (2012). Feature-based

classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* (Oxford, England), 28(2), 167-175. doi:10.1093/bioinformatics/btr629

Gonorazky H. D., Naumenko S., Ramani A. K., Nelakuditi V., Mashouri P., Wang P., Dowling J. J. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *The American Journal of Human Genetics*, 104(3), 466-483. doi:https://doi.org/10.1016/j.ajhg.2019.01.012

Hwang S., Kim E., Lee I., & Marcotte E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5, 17875. doi:10.1038/srep17875 https://www.nature.com/articles/srep17875#supplementary-information

Lin M., Whitmire S., Chen J., Farrel A., Shi X., & Guo J. T. (2017). Effects of short indels on protein structure and function in human genomes. *Scientific Reports*, 7(1), 9313. doi:10.1038/s41598-017-09287-x

Liu F., Zhang Y., Zhang L., Li Z., Fang Q., Gao R., & Zhang Z. (2019). Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome biology*, 20(1), 242-242. doi:10.1186/s13059-019-1863-4

Luo R., Sedlazeck F. J., Lam T. W., & Schatz M. C. (2019). A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nature Communications*, 10(1), 998. doi:10.1038/s41467-019-09025-z

McKenna A *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 20(9), 1297-1303. doi:10.1101/gr.107524.110

Neums L., Suenaga S., Beyerlein P., Anders S., Koestler D., Mariani A., & Chien J. (2017). VaDiR: an integrated approach to Variant Detection in RNA. *GigaScience*, 7(2). doi:10.1093/gigascience/gix122

Picardi E., D'Erchia A. M., Lo Giudice C., & Pesole G. (2017). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic acids research*, 45(D1), D750-D757. doi:10.1093/nar/gkw767

Piskol R., Ramaswami G., & Li J. B. (2013). Reliable identification of genomic variants from RNA-seq data. *American journal of human genetics*, 93(4), 641-651. doi:10.1016/j.ajhg.2013.08.008

Poplin R. *et al.* (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv, 201178. doi:10.1101/201178

Radenbaugh A. J., Ma S. Ewing A., Stuart J. M., Collisson E. A., Zhu J., & Haussler D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PloS one*, 9(11), e111516-e111516. doi:10.1371/journal.pone.0111516

Romagnoni A., Jégou S., Van Steen K., Wainrib G., Hugot J. P., Peyrin-Biroulet L., International Inflammatory Bowel Disease Genetics, C. (2019). Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Scientific Reports*, 9(1), 10351. doi:10.1038/s41598-019-46649-z

Rusch M., Nakitandwe J., Shurtleff S., Newman S., Zhang Z., Edmonson M. N., Zhang J. (2018). Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. *Nature Communications*, 9(1), 3962. doi:10.1038/s41467-018-06485-7

Sandmann S., Karimi M., de Graaf A. O., Rohde C., Göllner S., Varghese J., Dugas M. (2018). appreci8: a pipeline for precise variant calling integrating 8 tools. *Bioinformatics*, 34(24), 4205-4212. doi:10.1093/bioinformatics/bty518

Sheng Q., Zhao S., Li C.-I., Shyr Y., & Guo Y. (2016). Practicability of detecting somatic point mutation from RNA high throughput sequencing data. *Genomics*, 107. doi:10.1016/j.ygeno.2016.03.006

Treangen T. J., & Salzberg S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13, 146. doi:10.1038/nrg3164

Van der Auwera G. A. *et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics, 43(1110), 11.10.11-11.10.33. doi:10.1002/0471250953.bi1110s43

Wang C., Davila J. I., Baheti S., Bhagwate A. V., Wang X., Kocher J. P. A., Asmann Y. W. (2014). RVboost: RNA-seq variants prioritization using a boosting method. *Bioinformatics*, 30(23), 3414-3416. doi:10.1093/bioinformatics/btu577

Wu C., Zhao X., Welsh M., Costello K., Cao K., Abou Tayoun A., Sarmady M. (2019). Using

Machine Learning to Identify True Somatic Variants from Next-Generation Sequencing. *Clinical Chemistry*, 66(1), 239-246. doi:10.1373/clinchem.2019.308213

Xu C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16, 15-24. doi:https://10.1016/j.csbj.2018.01.003

Zhao Y., Wang K., Wang W. l., Yin T.T., Dong W. Q., & Xu C. J. (2019). A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics*, 20(1), 160. doi:10.1186/s12864-019-5533-4

Zook J. M., Chapman B., Wang J., Mittelman D., Hofmann O., Hide W., & Salit M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology*, 32(3), 246-251. doi:10.1038/nbt.2835

Zook J. M., McDaniel J., Olson N. D., Wagner J., Parikh H., Heaton H., Salit M. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature biotechnology*, 37(5), 561-566. doi:10.1038/s41587-019-0074-6