Improved Tumor-Only Somatic Variant Calling Using a Gradient Boosted Machine Learning Algorithm

#6651

Nick A. Phillips, Patrick Jongeneel, John West, Richard Chen, Jason Harris
Personalis, Inc. | 1330 O'Brien Dr., Menlo Park, CA 94025

Introduction

Accurate identification of somatic variants in a tumor sample is often accomplished by utilizing a paired normal tissue sample from the same patient to enable the separation of private germline mutations from somatic variants. However, a paired normal sample is not always available, making accurate somatic variant calling more challenging. Composite proxy normals and other filtering approaches can be used in lieu of a paired normal sample, but the resulting somatic call set may suffer from incomplete germline filtering and reduced sensitivity compared to paired tumor-normal analysis. To address these limitations, we developed a novel, machine learning based, tumor-only somatic small variant classifier, which leverages gradient boosted decision trees to significantly increase somatic variant specificity from a tumor-only analysis without reducing overall sensitivity.

Methods

Data Preprocessing

We produced a ground truth set of somatic SNVs and indels from 350 whole exome-sequenced tumor-normal pairs using a validated cancer bioinformatics pipeline. We then generated a feature set from each tumor sample by aggregating attributes including: allelic frequency and read depth, tumor cellularity estimations, germline variant calls from HaplotypeCaller, tumor-only somatic variant calls from Mutect and Mutect2 using a proxy-normal, copy-number alterations, annotations from databases such as GnomAD and COSMIC, and problematic-region annotations including homopolymers. Somatic variant truth labels were assigned using filtered Mutect2 output from the tumor-normal analysis. The samples were randomly split into training and testing sets in a 90-10 ratio.

Machine Learning Workflow

The problem of somatic variant calling was framed as two independent subtasks: filtering of non-somatic "germline bleedthrough" variants from the tumor-only callsets, and "rescuing" somatic variants missed by variant callers from the pileup data. For each subtask, we trained a gradient-boosted decision tree to predict the somatic likelihood of each candidate variant. Model hyperparameters were optimized using a random search during stratified cross-validation, and model performance was evaluated on a hold-out test set.

Results

Baseline Tumor-Only Calling

Baseline tumor-only somatic variant calling performance was evaluated by comparing the tumor-only Mutect output to our identified ground truth callsets for all 350 samples in the training and testing set.

Baseline Tumor-Only Variant Calling Performance

Training Data			
Precision Mean	Precision Std.	Recall Mean	Recall Std.
0.195	0.209	0.676	0.157
Test Data			
Precision Mean	Precision Std.	Recall Mean	Recall Std.
0.161	0.147	0.685	0.156

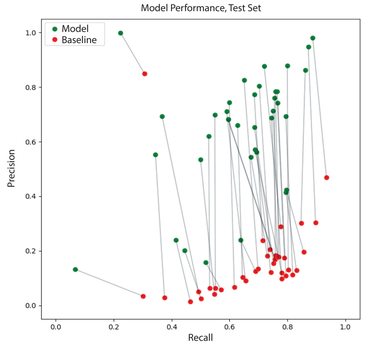
Somatic Classification Results

For each classification problem, we trained a gradient boosted decision tree model using Microsoft's LightGBM framework. We optimized model hyperparameters using a random search over maximum depth, minimum data in leaf, and number of leaves. Models were trained using stratified 5-fold cross validation.

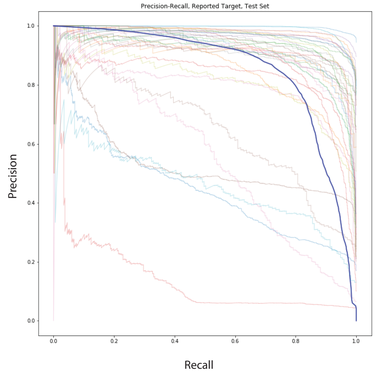
Using an operating point that optimized Youden's J statistic on the validation set, we observed a significant increase in model precision on the test set, with only a minor reduction in sensitivity compared to tumor-only somatic variant calling.

Average Performance with Optimal Model

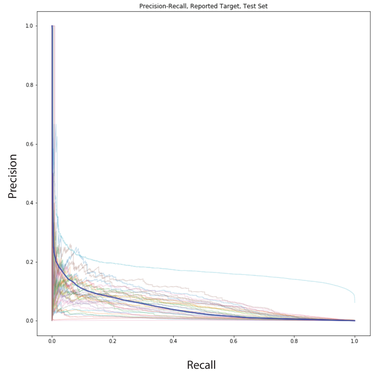
Test Data	
Precision	0.644
Recall	0.634



Filtration Model



Rescue Model

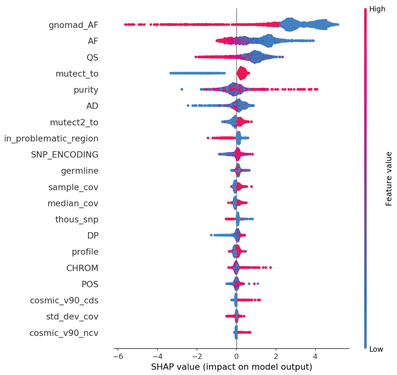


Contact:
nick.phillips@personalis.com

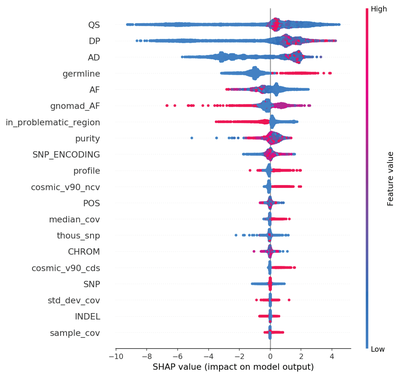
Model Interpretation

To improve interpretability of our model, we employed Shapely additive explanations (SHAP) to obtain feature importance values. Our analysis revealed that, for filtration, the most important features are cancer database annotations, allelic fraction, and quality scores. For the rescue model, the most important features are quality scores, read depth, and allelic depth. SHAP summary plots are shown below.

Filtration Model SHAP



Rescue Model SHAP



Conclusion

Our machine learning approach greatly enhances the performance of somatic variant calling when a paired normal sample is not available. On a held out test set, we demonstrated a significant improvement to average precision of somatic small variant calls, with only a minor reduction in average recall. Depending on the use-case, model operating points can be adjusted to fine-tune the tradeoff between variant recall and the precision of the resulting callset. Finally, model interpretation has revealed a subset of highly discriminative features, which may prove useful for variant interpretation, future feature engineering, or model tuning.

Personalis