A Machine-Learning Framework for Accurate Classification and Quantification of Oncogenic Variants Using the QuantideX® NGS DNA Hotspot 21 Kit

Lando Ringel, Blake Printy, Joseph Kaplan, Brian C Haynes, and Jessica L Larson

Department of Research and Development, Asuragen, Inc., Austin, TX

Summary

- In tumorigenic tissue, differentiating somatic variants from non-reference background noise is an ongoing challenge.
- A machine-learning model was developed to classify variants covered by the QuantideX® NGS DNA Hotspot 21 Kit*,**.
- Training and testing cohorts consisted of FFPE tumor samples, cell-line admixtures, and internally generated *in silico* libraries.
- During training, the model classified both SNVs and Indels with ≥99.2% sensitivity and PPV under 5-fold cross-validation.
- When validated on an independently verified testing cohort, the model attained 98.8% sensitivity and 99.6% PPV.

Introduction

Targeted amplicon-based next-generation sequencing (NGS) is a common and critical tool for profiling somatic mutations within tumor genomes in both clinical and research settings. Accurately classifying variants remains an ongoing challenge due to limited DNA quantities and compromised integrity, as well as chemical modifications of formalin-fixed paraffin-embedded (FFPE) tumor DNA specimens. A comprehensive machine-learning approach was developed to accurately differentiate true biological variants from process-related artifacts as part of the companion automated analysis software for the QuantideX NGS DNA Hotspot 21 Kit*,**.

Materials and Methods

DNA was extracted from FFPE tumor specimens and cell-line admixtures. Libraries for each sample were prepared and quantified using QuantideX NGS DNA Hotspot 21 Kit*,** (Asuragen Inc.), a 46 amplicon-based panel covering >1500 COSMIC mutations across 21 oncogenes. Sequencing was performed with the Illumina MiSeq. Libraries with at least 200 functional copies were used. Due to a lack of diversity in variant representation, an *in silico* based method was developed whereby SNVs and Indels were injected into to the sequencing output of mutation-negative libraries. These *in silico* SNVs and Indels augmented the training and testing cohorts at varying percentages throughout the covered domain of the panel. Read sequence quality, frequency, and variability were characterized with numerical features to capture the complex batch and locus-specific error profiles for all non-reference base-calls in each library. A machine-learning model was trained on these features to accurately identify both SNVs and Indels. The model was validated on an independent testing cohort with SNVs and Indels verified by the Oncomine Focus Assay (Thermo Fisher Scientific Inc.). The feature extraction methods, background error profiles, and trained machine-learning model were incorporated into the QuantideX NGS Reporter software.

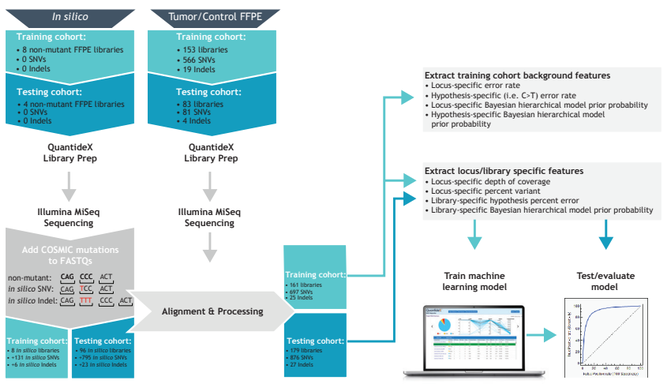

Figure 1. Overview of QuantideX NGS Model Development. The training cohort contained 153 libraries with 566 SNVs and 19 Indels. An additional 131 SNVs and 6 Indels (*in silico*) were generated using 8 mutation-free FFPE FASTQs and added to this training cohort. The testing cohort contained 83 libraries with 81 SNVs and 4 Indels. An additional 795 SNVs and 23 Indels were added to the testing cohort with the *in silico* method. Numerical features were extracted from aligned BAM files to train and test the variant calling model.

*CE-IVD for US export only. **Research Use Only. Not for use in diagnostic procedures.
Presented at AMP 2018

Results

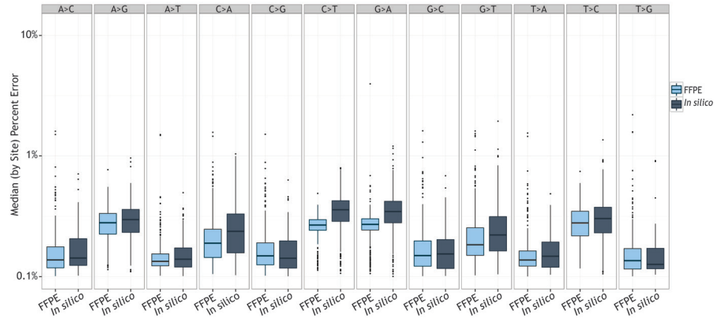

Figure 2. FFPE and *In Silico* Libraries Have Comparable Hypothesis-Specific Background Error Rates. For each possible nucleotide substitution, the median percent error rate across all loci covered by the training cohort is shown. The medians are stratified by FFPE (light blue) and *in silico* (dark gray) library types. Because the *in silico* libraries were generated by modified mutation-free FFPE FASTQ files, the background error rates do not meaningfully differ (mean difference: 1.00%, after accounting for substitution type). The background error rates for C>T and G>A are relatively higher for both library types due to the documented effect of deamination in FFPE samples.

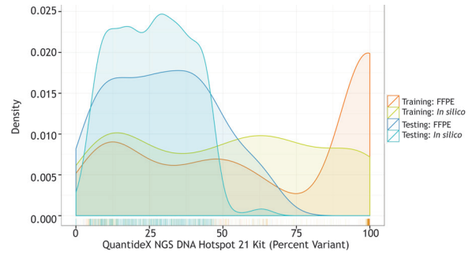

Figure 3. Variant Allele Frequencies (VAFs) for Training and Testing Cohorts. The variant caller was trained (orange) on allele frequencies spanning 0% to 100% to cover all hypothetically observed VAFs, and validated (blue) on allele frequencies largely below 50% to capture more challenging VAFs often observed in low tumor purity settings.

A

Training Cohort	Number of libraries	TP	FN	FP	Sn	PPV
FFPE SNVs	153	563	3	4	99.5%	99.3%
FFPE Indels		19	0	0	100.0%	100.0%
In Silico SNVs	8	128	3	0	97.8%	100.0%
In Silico Indels		6	0	0	100.0%	100.0%
Overall	**161**	**716**	**6**	**4**	**99.2%**	**99.4%**

B

Testing Cohort	Number of libraries	TP	FN	FP	Sn	PPV
FFPE SNVs	83	77	4	1	95.1%	98.7
FFPE Indels		4	0	2	100.0%	50.0%
In Silico SNVs	96	788	7	0	99.1%	100.0%
In Silico Indels		23	0	0	100.0%	100.0%
Overall	**179**	**892**	**11**	**3**	**98.8%**	**99.6%**

Table 1. Model Performance on A) Training (5-Fold Cross-Validation) and B) Independent Testing Cohorts. Column headings indicate TP (true positives), FN (false negatives), FP (false positives), Sn (sensitivity), and PPV (positive predictive value). There were two FP Indels in the testing cohort. The first Indel was near the limit of detection (VAF=5.7%) and came from a contrived specimen; the second Indel (VAF=56.6%) is believed to be a TP, but was annotated as two adjacent SNVs by the Oncomine Focus Assay.

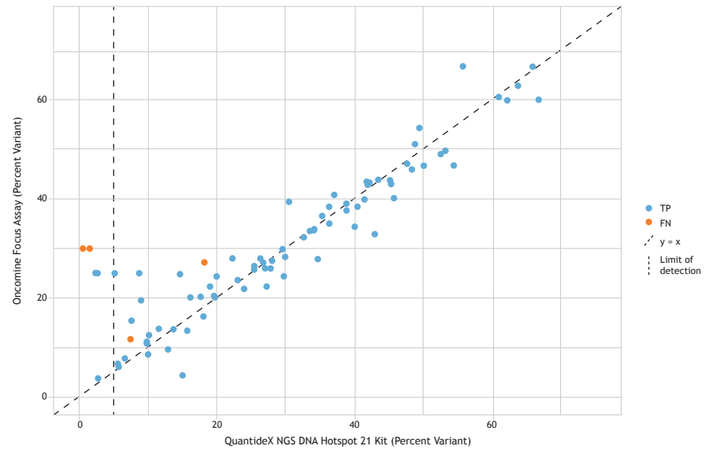

Figure 4. Observed VAFs for Validation FFPE Libraries are Significantly Correlated with the Expected VAFs of the Oncomine Focus Assay (N=85, Spearman rho = 0.904, p-value < 2.2 x 10-16). Two false negative SNVs (orange) with expected VAF close to 30% are located on neighboring sites, and were combined and identified as a (false) positive indel by the QuantideX NGS DNA Hotspot 21 Kit*,** variant caller.

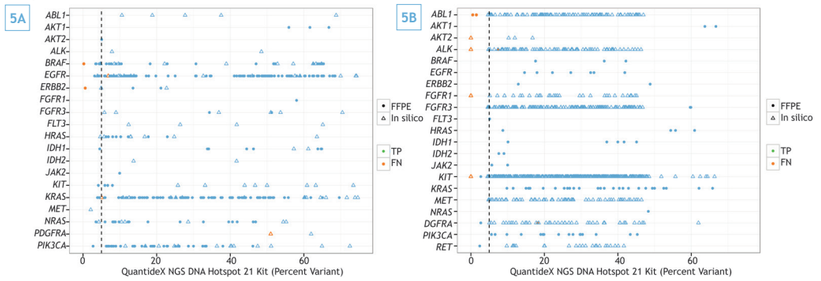

Figure 5. Observed VAFs of Expected SNVs and Indels Used for A) Training and B) Testing the QuantideX NGS DNA Hotspot 21 Kit*, Variant Caller.** The y-axis represents unique genes covered by the Hotspot 21 panel; x-axis shows the observed VAF of each variant. The vertical line represents 5% VAF. **A)** Under 5-fold cross-validation, the variant caller accurately classified 716 out of 722 expected variants in our training cohort. The mean and median observed percent variant for false negatives were 11.3% and 4.7%, respectively. **B)** The variant caller was able to detect 892 out of 903 expected variants in the testing cohort. The mean and median observed percent variant for false negatives were 2.5% and 0.0%, respectively. All expected variants with observed VAF equal to 0.0% were the result of *in silico* artifacts. Additional titration studies of probit analysis revealed limit of detection values of 3.7% to 6.4% for both SNVs and Indels (data not shown, see poster ST130).

Conclusions

- The presented variant caller, included as part of the QuantideX NGS DNA Hotspot 21 Kit*,**, addresses the challenges of calling low frequency somatic variants by directly accounting for both sample and locus-specific error profiles.
- FFPE libraries with computationally augmented COSMIC variants enabled training and validation across all gene regions targeted by the assay.
- This machine-learning approach offers a robust and sensitive framework for oncology diagnostics and clinical trial research, and is readily adaptable to other targeted NGS assays and platforms.

Asuragen®