An ensemble approach to accurately detect somatic mutations via adaptive boosting

Abstract
LB-306

Li Tai Fang[1,*], Pegah T. Afshar[2,*], Aparna Chhibber[1], John C. Mu[1], Marghoob Mohiyuddin[1], Greg Gibeling[1], Sharon Barr[1], Narges Bani Asadi[1], Mark Gerstein[3], Daniel C. Koboldt[4], Wing H. Wong[5,6], and Hugo Y. K. Lam[1]

INTRODUCTION

Identifying somatic mutations is a key analysis in cancer research. The samples are often impure, and tumors are heterogeneous. Oftentimes, an algorithm works well for one tumor but not for another. SomaticSeq is an accurate somatic mutation detection pipeline that implements a stochastic boosting algorithm with an ensemble approach. It integrates multiple algorithms, i.e., MuTect, SomaticSniper, VarScan2, JointSNVMix2, and VarDict. Individual call sets are combined, and over 70 sequencing and genomic features are extracted, which are then provided to an adaptively boosted decision tree learner to build a classifier. The learner is trained with sophisticated simulated data to discriminate true mutations from very noisy data of tumor samples. Fig. 1 presents the schematic of the workflow.

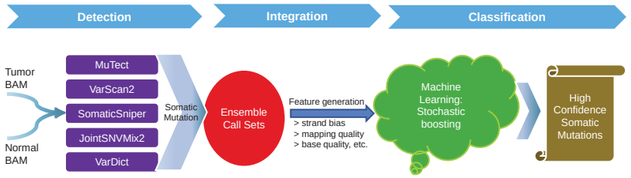

Fig 1) Schematic of the SomaticSeq Workflow: mutation calls from multiple tools are merged into an ensemble, and up to 72 features for each of the calls are extracted from the BAM files. They are provided to the machine learning model, which calculates the probability for each call, yielding a high-confidence somatic mutation call set.

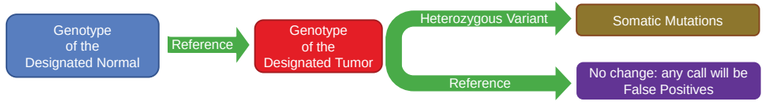

Fig 2) Building the ground truth for *in silico* titration.

VALIDATION

We have validated SomaticSeq's with the DREAM Challenge data (Fig. 4 & 5), *in silico* titration of two genomes (Fig. 6), as well as real tumor data (Table 1). Fig. 2 & 3 depict the method of *in silico* titration.

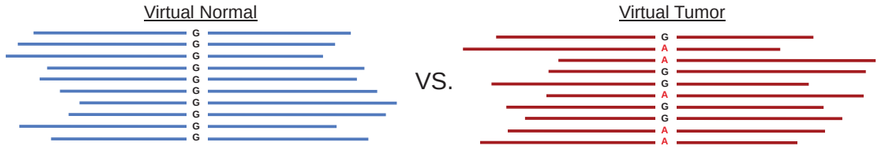

Fig 3) *in silico* Titration of two human genomes. Blue represents reads from a designated normal genome. Red represents reads from a designated tumor genome. Different allele frequencies are archived by mixing in different proportions of the two genomes.

RESULTS

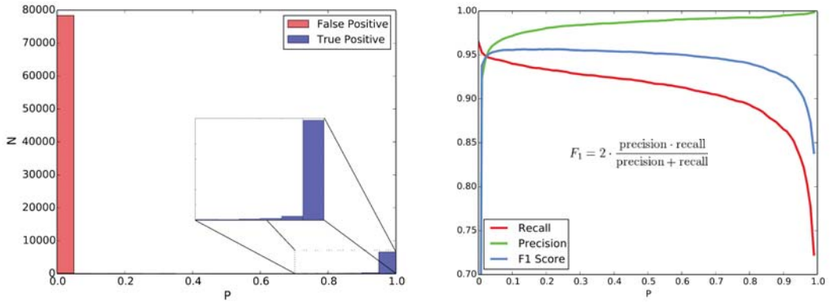

Fig 4 a) Histogram of probability values (P) of all the mutation candidates in DREAM challenge stage 3. Higher values (closer to 1) imply that calls are more likely true somatic mutations. b) An accuracy plot showing recall, precision, and F1 scores vs probability cut offs. The probability values are calculated based on the Stage 2 trained classifier.

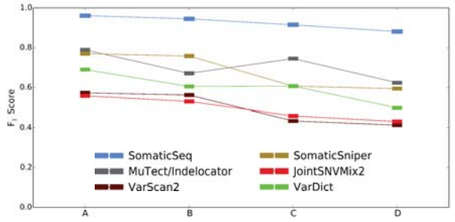

Fig 5) Tumor and Normal from DREAM Challenge is mixed at different ratios to simulate challenging data sets. The x-axis label, A, B, C, and D denotes the 4 different settings described as follows:

A) Pure Normal / Pure Tumor (VAF = 50%, 33%, and 20%)
B) Normal contaminated with 5% Tumor / Pure Tumor
C) Pure Normal / Tumor contaminated with 30% Normal
D) Contaminated Normal / Contaminated Tumor

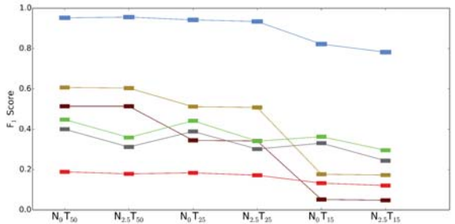

Fig 6) *in silico* titration. The x-axis label describes the expected VAF (%) in Normal and Tumor, e.g., $N_{2.5}T_{15}$ means expected VAF of 2.5% in the normal, and 15% in the tumor.

SNV	COLO-829		CLL1		TCGA-AZ-6601	
	# calls	Recall	# calls	Recall	# calls	Recall
MuTect	46,831	0.996	8,361	0.895	6,492	0.992
VarScan2	64,927	0.987	19,797	0.888	254,105	0.705
SomaticSniper	53,077	0.996	13,690	0.907	8,325	0.677
JointSNVMix2	85,983	0.996	22,534	0.899	9,058	0.960
VarDict	53,076	0.857	5,748	0.883	471,404	0.939
Ensemble	191,696	0.998	55,196	0.935	851,040	1.000
SomaticSeq	37,054	0.989	2,575	0.891	5,563	0.879

Table 1) Three sets of publicly available real tumor-normal sequencing data. The classifier used was trained on the DREAM data. COLO-829: melanoma cell line. CLL1: chronic lymphoblastic leukemia. TCGA-AZ-6601: colon adenocarcinoma.

1. Bina Technologies, Roche Sequencing, Redwood City, CA 94065
2. Department of Electrical Engineering, Stanford University, Stanford, CA 94035
3. Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520
4. The Genome Institute, Washington University in St. Louis, St. Louis, MO 63108
5. Department of Statistics, Stanford University, Stanford, CA 94035
6. Department of Health Research and Policy, Stanford University, Stanford, CA 94035

* Authors contributed equally