*Bioinformatics*, 35(17), 2019, 3157–3159
doi: 10.1093/bioinformatics/btz018
Advance Access Publication Date: 12 January 2019
Applications Note

OXFORD

Sequence analysis

# SMuRF: portable and accurate ensemble prediction of somatic mutations

**Weitai Huang[1,2], Yu Amanda Guo[1], Karthik Muthukumar[1], Probhonjon Baruah[1], Mei Mei Chang[1] and Anders Jacobsen Skanderup[1,*]**

[1]Department of Computational and Systems Biology, Agency for Science Technology and Research, Genome Institute of Singapore, Singapore 138672, Singapore and [2]Graduate School of Integrative Sciences and Engineering, National University of Singapore, Singapore 117456, Singapore

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on August 1, 2018; revised on November 26, 2018; editorial decision on December 31, 2018; accepted on January 7, 2019

## Abstract

**Summary:** Somatic Mutation calling method using a Random Forest (SMuRF) integrates predictions and auxiliary features from multiple somatic mutation callers using a supervised machine learning approach. SMuRF is trained on community-curated matched tumor and normal whole genome sequencing data. SMuRF predicts both SNVs and indels with high accuracy in genome or exome-level sequencing data. Furthermore, the method is robust across multiple tested cancer types and predicts low allele frequency variants with high accuracy. In contrast to existing ensemble-based somatic mutation calling approaches, SMuRF works out-of-the-box and is orders of magnitudes faster.

**Availability and implementation:** The method is implemented in R and available at https://github.com/skandlab/SMuRF. SMuRF operates as an add-on to the community-developed bcbio-nextgen somatic variant calling pipeline.

**Contact:** skanderupamj@gis.a-star.edu.sg

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Identification of somatic mutations from matched tumor and normal samples is challenged by sequencing noise and alignment ambiguities as well as the heterogeneous composition of tumors. Recent studies have revealed low concordance between existing methods for somatic variant calling (Hwang *et al.*, 2015; Kroigard *et al.*, 2016; O'Rawe *et al.*, 2013; Roberts *et al.*, 2013). Additionally, a benchmark study demonstrated that the accuracy of a given somatic mutation calling algorithm can vary extensively across different workflows and pipelines (Alioto *et al.*, 2015). Parameters influencing this variation may be choice of alignment algorithm, use of local re-alignment, as well as configuration of a multitude of post-processing filters. The consensus of multiple callers have been used to improve the accuracy of somatic variant calling (Callari *et al.*, 2017; Ellrott *et al.*, 2018; Rashid *et al.*, 2013). Taking this one step further, a machine learning based ensemble method may combine multiple mutation callers with auxiliary sequence and alignment features to improve mutation calling accuracy (Ding *et al.*, 2012; Fang *et al.*, 2015; Wood *et al.*, 2018). While such approaches may improve accuracy, they are generally not portable: The end-user must obtain suitable training and testing datasets and need to have knowledge of machine learning (Supplementary Fig. S1A ). There is therefore a need for accurate and pre-trained ensemble approaches for somatic mutation calling that can be ported between research groups. Here, we developed a Somatic Mutation calling method using a Random Forest (SMuRF), which combines predictions from four mutation callers with auxiliary alignment and mutation features using supervised machine learning (Supplementary Fig. S1B).

## 2 Implementation

SMuRF is available as an R package. Briefly, the bcbio-nextgen framework (https://github.com/chapmanb/bcbio-nextgen) is used to generate somatic variant calls from 4 different methods: MuTect2 (Cibulskis *et al.*, 2013), Freebayes somatic (ArXiv: https://arxiv.org/abs/1207.3907), VarDict (Lai *et al.*, 2016) and VarScan (Koboldt *et al.*, 2012). Variant and auxiliary features are extracted from the VCF files. The SMuRF random forest model is pre-trained on a gold standard set of mutation calls curated by the International Cancer Genome Consortium (ICGC) community using deep (>100×) whole genome sequencing (WGS) of two tumors (Alioto *et al.*, 2015). Feature extraction and prediction of somatic variants takes ~10 min for tumor-normal WGS data on a standard computer (4 CPUs, 16GB RAM).

## 3 Overview

SMuRF SNV and indel models were trained on matched tumor-normal WGS data from a chronic lymphocytic leukemia (CLL) patient and a medulloblastoma (MB) patient, where the true somatic mutations have been identified and curated by the International Cancer Genome Consortium (ICGC) (Alioto *et al.*, 2015). The training data was augmented to expose the model to additional variation in sequencing coverage, tumor purity and tumor/normal coverage

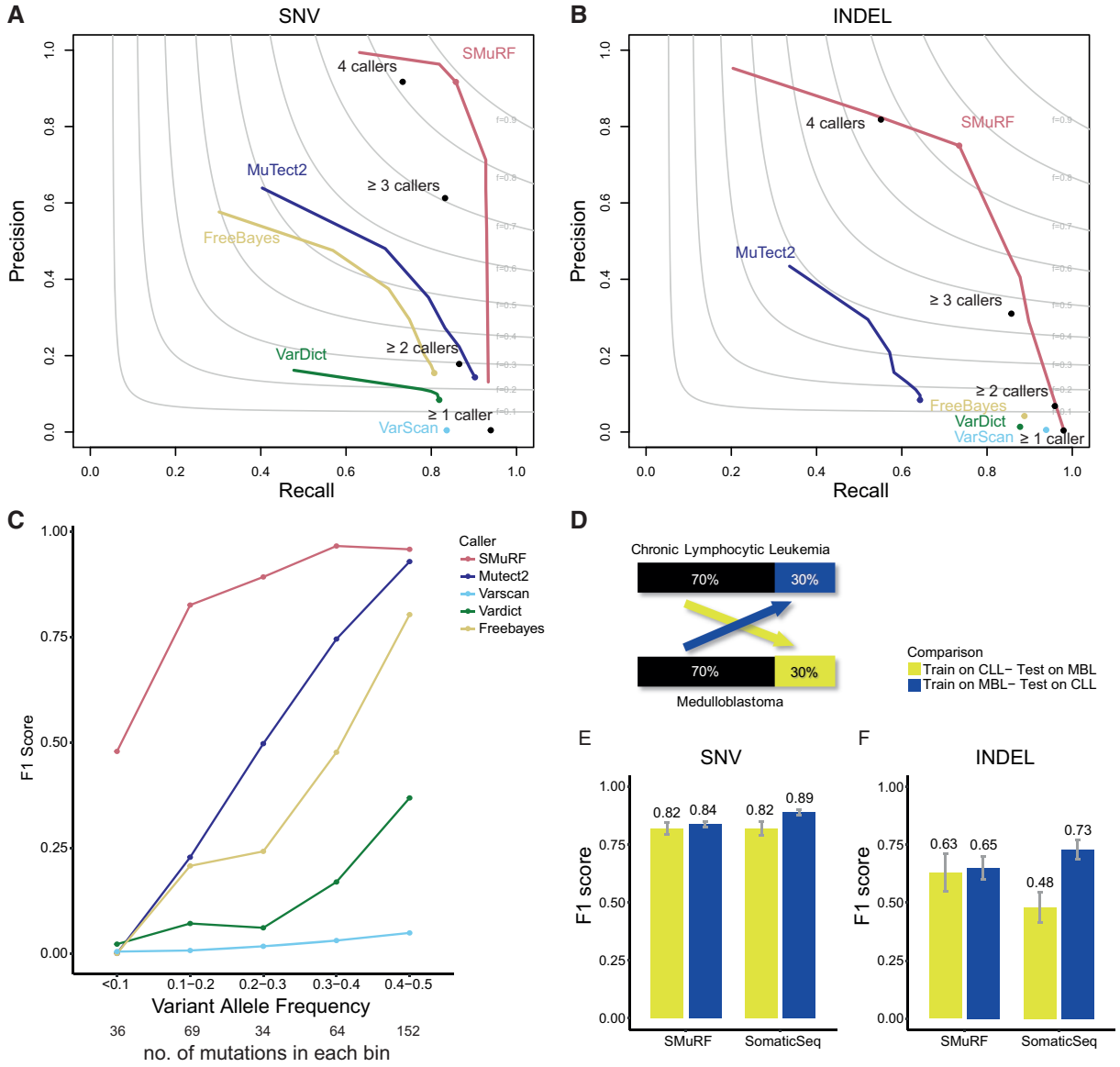


**Fig. 1.** Performance of SMuRF. Precision-recall profiles for individual somatic mutation callers and SMuRF evaluated on (**A**) SNV and (**B**) indels using 20% withheld test data. Curves show the performance of the individual algorithms under different variant score thresholds (MuTect2 tumor log-odds score, Freebayes log-odds score, VarDict SSF score, VarScan SSC score and SMuRF confidence score). Solid points refer to the default performance of the caller in the bcbio-nextgen workflow. Black solid points denote the accuracy of calls identified by the majority-voting scheme in bcbio-nextgen (at least 1, 2, 3 or 4 callers). The grey contours indicate F1 scores as a function of recall and precision. (**C**) Accuracy of SMuRF and individual callers as a function of somatic variant allele frequency in the test set; F1 scores evaluated for each variant allele frequency bin. (**D–F**) Evaluation of SMuRF and SomaticSeq performance when trained and tested across different cancer types. (D) Models were trained on 70% of CLL data and tested on 30% of MB data (and vice versa). F1 scores were recorded for SMuRF and SomaticSeq SNV (E) and indel (F) predictions. Error bars represent the standard deviation of the mean across 10 random training/test data splits (same splits for both methods)

imbalance (Supplementary Fig. S1C and Supplementary Methods). SMuRF was trained on 80% of the data, with 20% of the data withheld as a test set. Highly predictive features were mostly somatic variant scores provided by individual methods as well as mapping and base quality estimates (Supplementary Table S1). SMuRF achieved F1-scores of 0.88 and 0.74 for SNVs and indels, respectively (Fig. 1A and B, Supplementary Tables S2 and S3 and Supplementary Fig. S3 for SNV coding regions). Importantly, SMuRF showed improved accuracy over the best mutation calling submissions reported in the benchmark by Alioto et al. using the same dataset (best reported F1-scores 0.79 and 0.65 for SNV and indels, respectively) (Alioto *et al.*, 2015). In our analysis, while individual methods could recover most of the true SNVs, this came at the cost of very low precision (<40% precision at 80% recall) (Supplementary Table S2). In contrast, SMuRF could recover 86% of the true SNVs (recall) at 92% precision on the withheld test set. While a simple consensus approach using the intersection of individual methods performed well (F1 = 0.82), SMuRF achieved markedly higher recall (86% versus 74%) at a similar level of precision. All methods, including SMuRF, were mostly robust when tested under different levels of tumor purity (Supplementary Fig. S4). However, SMuRF showed substantially improved accuracy at low somatic variant allele frequencies (VAFs) as compared to individual methods (Fig. 1C), which is particularly important in the setting of tumor heterogeneity inference (Shi *et al.*, 2018). We further benchmarked SMuRF SNV calling using independent data from the DREAM Somatic Mutation Challenge where artificial tumor data has been generated using an in-silico approach (Ewing *et al.*, 2015). While the performance of individual methods varied across these datasets, SMuRF was highly accurate across all synthetic tumors (F1 > 0.8) (Supplementary Fig. S5). Overall, these results support that SMuRF is robust and can generalize to unseen data.

Analysis of indel prediction accuracy showed that individual mutation callers could recover most of the true indels (64–94% recall), but only at the cost of very low precision (<8%). Interestingly, simple consensus approaches performed well for indel prediction (F1 0.46 and 0.66 for 3 and 4-caller consensus, respectively). However, while consensus methods suffered from either low recall (0.55) or precision (0.31), SMuRF obtained high indel prediction accuracy (F1 = 0.74) with both high recall (0.74) and precision (0.75) (Fig. 1B, Supplementary Table S3). We also analyzed the extent that SMuRF predicts the same somatic mutations in tumor samples profiled with both (>200× coverage) WES and (<100× coverage) WGS. When restricting analysis to variants in coding regions, SMuRF predicted somatic SNVs and indels with comparable or higher concordance than individual methods (Supplementary Figs S6 and S7).

Finally, we compared SMuRF to two existing machine learning-based methods. The first was MutationSeq, a pre-trained ensemble SNV caller (Ding *et al.*, 2012), which achieved an F1-score of 0.68, similar to the other individual SNV callers in our analysis (Supplementary Fig. S8). Next, we compared the performance of SomaticSeq (Fang *et al.*, 2015), a method that required users to train their own predictive model (see Supplementary Methods). The trained SomaticSeq model had slightly increased test set prediction accuracy over SMuRF for both SNV (0.90 versus 0.88) and indels (0.78 versus 0.75) (Supplementary Fig. S9B and C). We further evaluated how the methods generalized when models were trained and tested across different tumor datasets and found that SomaticSeq showed greater test accuracy variation (Fig. 1D–F). This was especially pronounced for indel prediction, where the F1 accuracy of SomaticSeq varied from 0.48 to 0.73 (SMuRF 0.63–0.65) when tested on the MBL or CLL sample, respectively. Furthermore, SomaticSeq used ~24 h to predict both SNVs and indels since it also computes auxiliary features from the raw alignment data. In contrast, SMuRF depends only on VCF files and predicts both SNVs and indels in ~10 min (Supplementary Fig. S9A). Overall, these results support that SMuRF is both accurate and computationally efficient.

In summary, SMuRF is an accurate, portable and user-friendly ensemble-based somatic mutation caller, which should benefit both cancer genomics studies as well as clinical applications.

## Acknowledgements

We thank I. Kassam, N. Rohatgi, K. Krishnamachari, M. N. Mojtabavi, G. Zhu, U. Ghoshdastider and T. Kulshrestha for their support and discussion during the development and testing of SMuRF.

## Funding

This work was supported by an Open Fund Individual Research Grant from the Singapore National Medical Research Council (OFIRG15nov072).

*Conflict of Interest*: none declared.

## References

Alioto,T.S. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, **6**, 10001.

Callari,M. *et al.* (2017) Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Med.*, **9**, 35.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213.

Ding,J. *et al.* (2012) Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics (Oxford, England)*, **28**, 167–175.

Ellrott,K. *et al.* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, **6**, 271–281.e277.

Ewing,A.D. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, **12**, 623.

Fang,L.T. *et al.* (2015) An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.*, **16**, 197.

Hwang,S. *et al.* (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5**, 17875.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Kroigard,A.B. *et al.* (2016) Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS One*, **11**, e0151664.

Lai,Z. *et al.* (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.*, **44**, e108.

O'Rawe,J. *et al.* (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.*, **5**, 28.

Rashid,M. *et al.* (2013) Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics (Oxford, England)*, **29**, 2208–2210.

Roberts,N.D. *et al.* (2013) A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics (Oxford, England)*, **29**, 2223–2230.

Shi,W. *et al.* (2018) Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep.*, **25**, 1446–1457.

Wood,D.E. *et al.* (2018) A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.*, **10**, eaar7939.