DETECT SOMATIC MUTATIONS

SomaticSeq

An ensemble and machine learning approach to accurately detect somatic mutations

Accurate detection of somatic mutations has been a challenge in cancer NGS analysis. Somatic mutations are much rarer on the genomic scale than germline variants, and detection is further confounded by tumor heterogeneity and cross-contamination between tumor and normal samples. Oftentimes, a somatic mutation caller performs well for one tumor but not for another.[1]

To address this challenge, Roche Sequencing has developed SomaticSeq,[1] an open-source, ensemble somatic mutation detection pipeline that integrates various somatic mutation callers and utilizes a machine learning algorithm to yield a high-confidence call set.

Methodology

SomaticSeq takes an integrative approach and combines six state-of-the-art somatic mutation callers for SNVs and indels: MuTect,[2] SomaticSniper,[3] VarScan2,[4] JointSNVMix2,[5] VarDict,[6] and SomaticIndelDetector.[7] The SomaticSeq workflow, depicted in Figure 1, involves the following steps:

1. Preprocessing

Sequences are first preprocessed according to GATK best practices.

2. Detection

The BAM files are run through each somatic caller, and a combined call set is obtained.

3. Integration

Over 70 sequencing features are extracted for each variant call in the combined call set.

Some top features include:

- Caller classification
- Mapping quality
- Base call quality
- Strand bias
- Read counts
- dbSNPs membership

4. Classification

A machine learning algorithm generates an ensemble of weighted decision trees of these features based on the training set, then scores each of the consensus variant calls as either false positive or true positive.

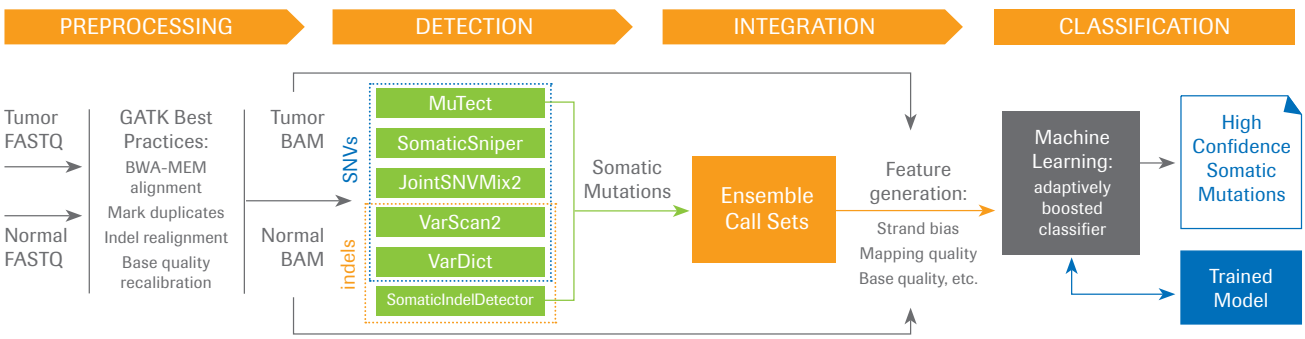

Figure 1. Workflow of SomaticSeq

Application

The SomaticSeq pipeline has been validated with data from the DREAM Challenge,[8] *in silico* titration of two genomes, as well as real tumor data. Figure 2 depicts cross validation results of applying SomaticSeq to data from Stage 3 of the DREAM challenge, where the ensemble method is shown to be more accurate than the individual callers.

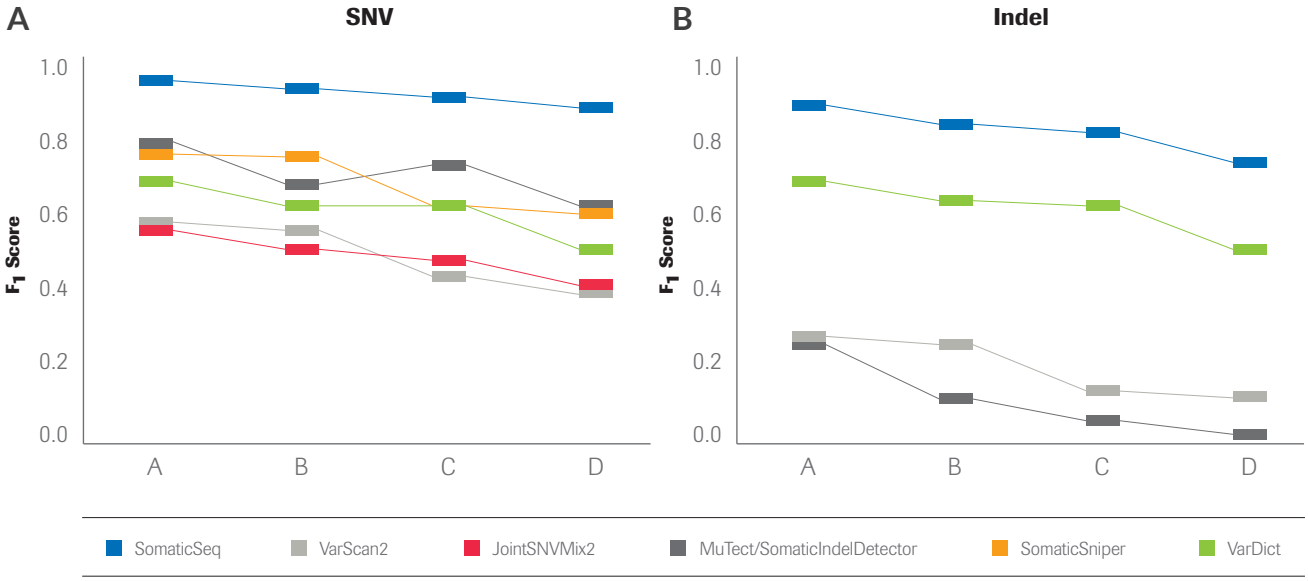

Figure 2. F1 scores of SomaticSeq and the individual tools for the DREAM Challenge Stage 3 cross validation, demonstrating SomaticSeq's superior performance for SNV and INDEL detection A: pure normal / pure tumor. B: contaminated normal / pure tumor. C: pure normal / contaminated tumor. D: contaminated normal / contaminated tumor. Stage 2 data was used as the training set.

Get SomaticSeq

SomaticSeq is currently available at github (github.com/bioinform/somaticseq). Users have the flexibility to customize SomaticSeq by submitting their own relevant training set, and by specifying any combination of up to 10 somatic callers.

References

1. Fang LT, Afshar PT, Chhibber A, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 2015;16:197-210.
2. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9.
3. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28(3):311-7.
4. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76.
5. Roth A, Ding J, Morin R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics.2012;28(7):907-13.
6. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016.
7. Banerji S, Cibulskis K, Rangel-Escareno C, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012; 486(7403):405-9.
8. Ewing AD, Houlahan KE, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat. Methods. 2015;12(7)623-30.

Published by:

Roche Sequencing
4300 Hacienda Drive
Pleasanton, CA 94588

sequencing.roche.com