TGIL
YONSEI UNIVERSITY
TRANSLATIONAL GENOME INFORMATICS LABORATORY

A robust benchmark for evaluating and improving mosaic variant detection

Yoo-Jin Ha, Jisoo Kim, Seungseok Kang, Junhan Kim, Myung Joon Oh, Se-Young Jo, Hyun Seok Kim, and Sangwoo Kim
Severance Biomedical Science Institute, Brain Korea 21 PLUS Project for Medical Sciences, Yonsei University College of Medicine, Korea

ABSTRACT

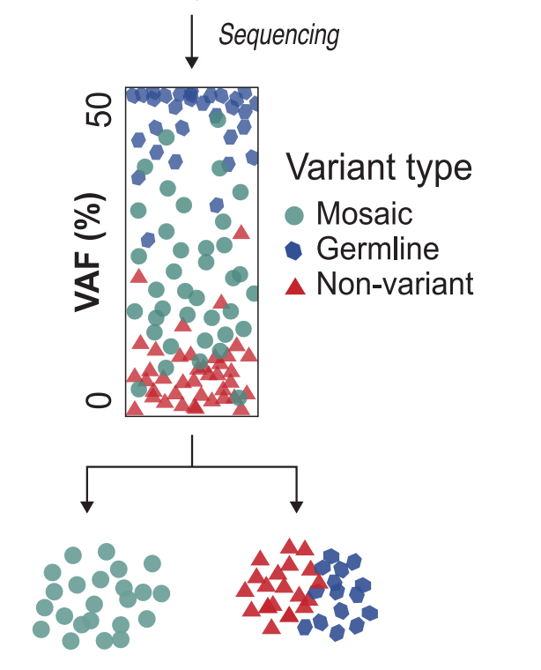
The rapid advances in sequencing and analysis technologies have enabled the accurate detection of diverse forms of genomic variants, including germline, somatic, and mosaic mutations. However, unlike for the former two mutations, the best practices for mosaic variant calling still remain chaotic due to the technical and conceptual difficulties faced in evaluation. Here, we present our benchmark of nine feasible strategies for mosaic variant detection based on a systematically designed reference standard that mimics mosaic samples, with 390,153 control positive and 35,208,888 negative single-nucleotide variants and insertion–deletion mutations. We identified the condition-dependent strengths and weaknesses of the current strategies, instead of a single winner, regarding variant allele frequencies, variant sharing, and the usage of control samples. Moreover, feature-level investigation directs the way for immediate to prolonged improvements in mosaic variant detection. Our results will guide researchers in selecting suitable calling algorithms and suggest future strategies for developers.

#mosaicism #mosaic variant #variant calling

INTRODUCTION

After conception, postzygotic mutations continuously occur throughout life in humans, causing **somatic mosaicism** in an individual. The variant type, time of origination, and locations of the mosaic mutations result in unique mosaic patterns in a combinatorial manner and further affect phenotypes, including various noncancerous diseases.

The detection of mosaic mutations is an intricate process both **conceptually** and **technically**. For example, mutations occured in the developmental process lead to a complex relationship among the affected and unaffected tissues; mutations may or may not be shared between a pair of samples.



Technical challenges

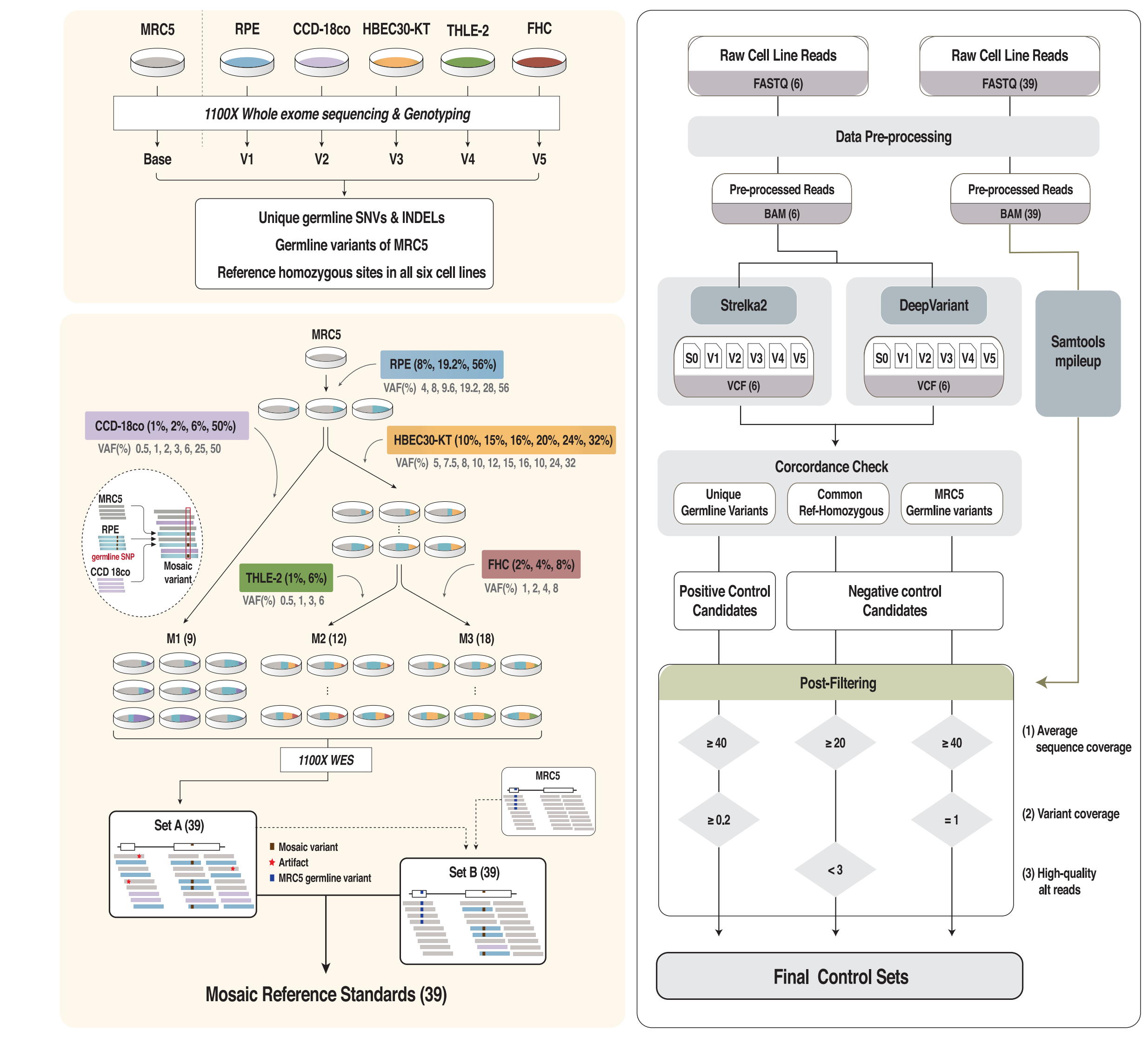
- **Variant allele frequencies (VAFs) vary widely** from extremely low (< 1%) to the level of germline variants (approximately 50%)
- **No clear reference sample exists** as the mutations can be shared. They also can be largely unbalanced among shared tissues.

This ambiguity is reflected in the disparate set of approaches applied in recent studies, and these circumstances urgently demand a rigorous cataloging and assessment of mosaic detection algorithms. Above all, the construction of robust and biologically compatible reference standards is a prerequisite.

METHODS

Construction of mosaic reference standards

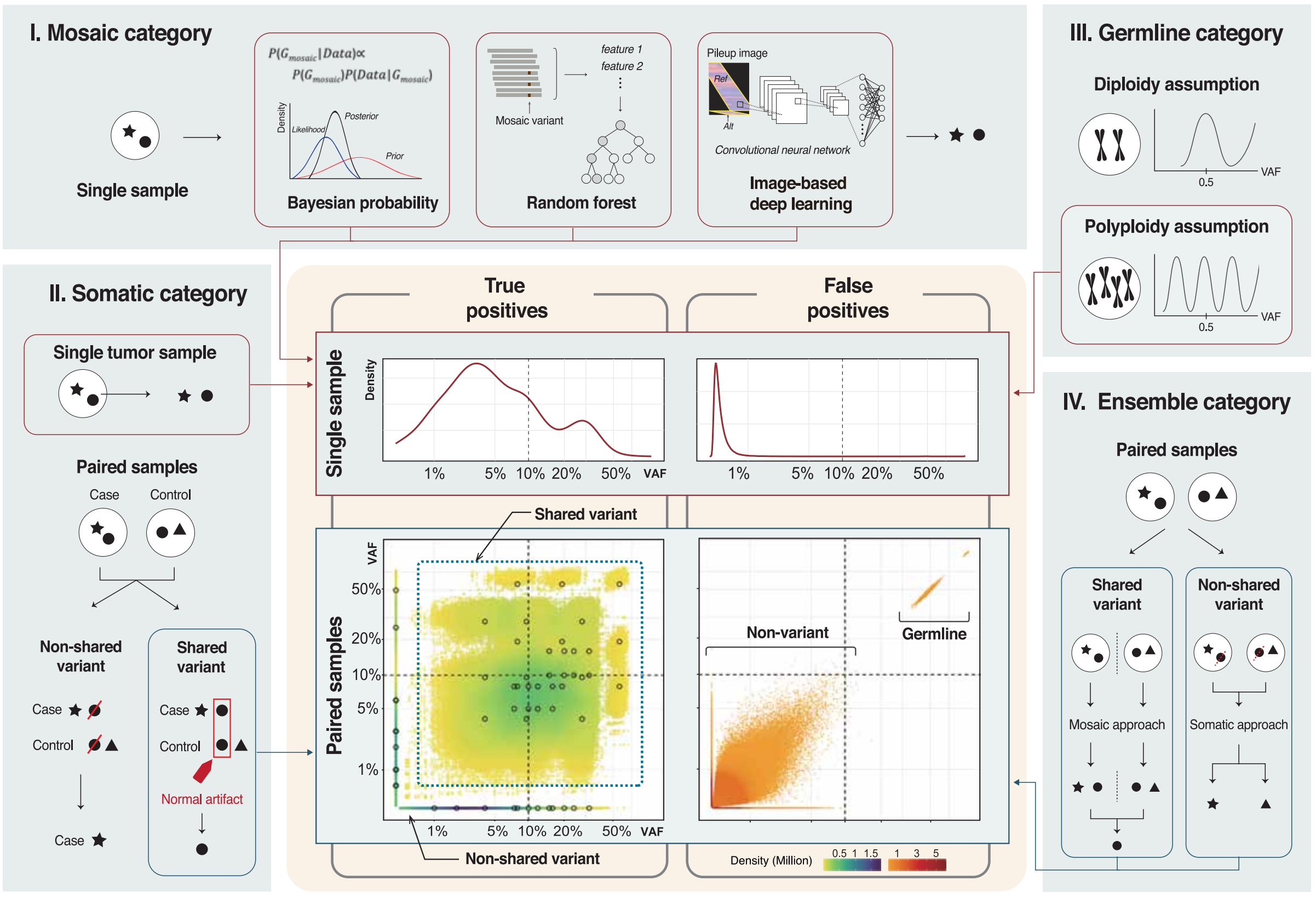
We generated robust, large-scale, and cell line-based reference standards using **386,613 single-nucleotide variants (SNVs) and insertion-deletion mutations (INDELs)** as positive controls and **35,133,353 negative control** positions. By mixing the six normal cell lines (MRC5, RPE, CCD-18co, HBEC30-KT, THLE-2, FHC) cumulatively, the unique germline variants represented mosaic variants in desired allele frequencies. The uniqueness of the reference standard lies in the internal hierarchical structure that mimics the mutation acquisition process under cellular differentiation, such as during the early embryonic development.



Calling of mosaic variants is susceptible to two different types of errors: (1) calling non-variant sites (e.g., reference allele) and (2) calling germline variants. Therefore, we provided two different versions of the final sets—set A and set B, containing non-variant sites and germline variants as negative controls respectively.

Evaluation of mosaic variant calling strategies

We selected nine mosaic detection strategies for evaluation on the criteria of (1) algorithms that explicitly aim to detect mosaic mutations (2) procedures that have been used previously to discover mosaic variants and (3) algorithms that can be applied for mosaic mutation detection via simple modifications. The nine strategies were classified into four major categories based on their baseline algorithms: **mosaic**, **somatic**, **germline**, and **ensemble**.

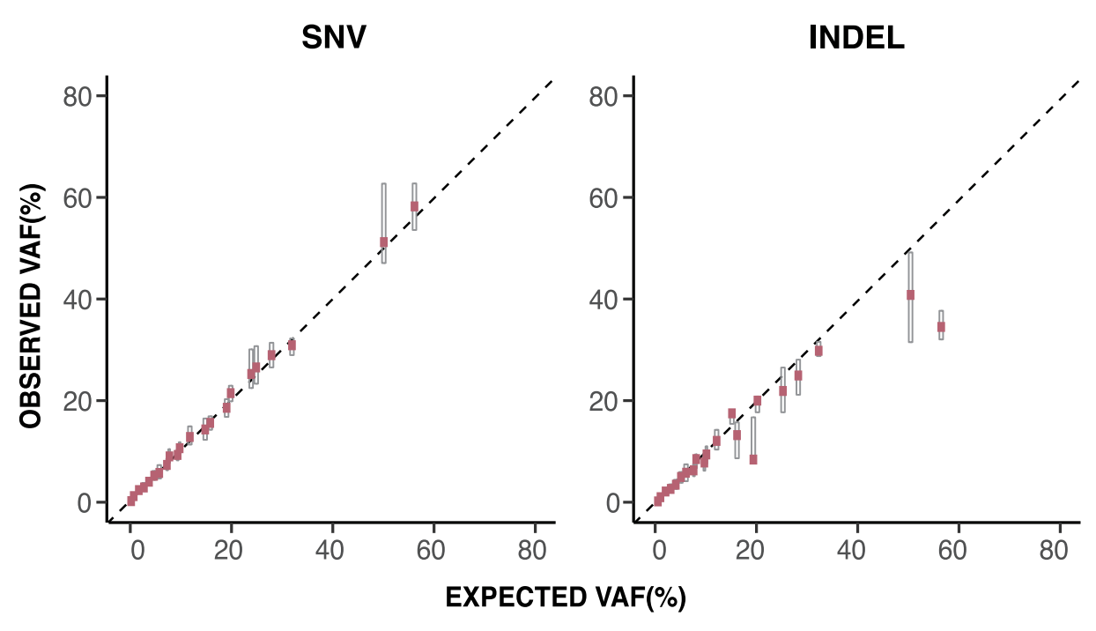


Evaluated strategies

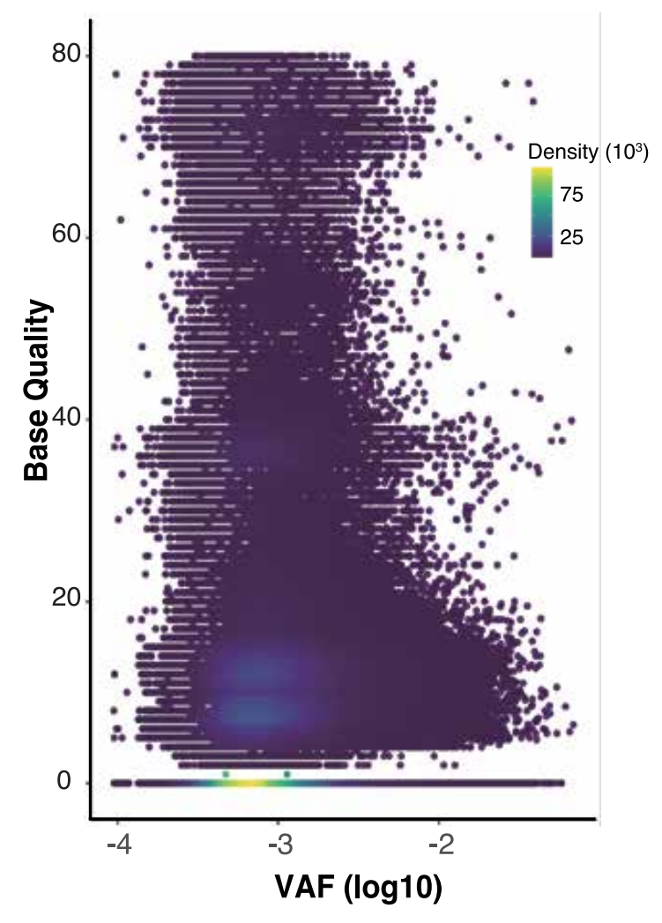
Mosaic category: MosaicHunter (MH), MosaicForecast (MF), and DeepMosaic (DM), which exploit Bayesian, Random-Forest, and deep-learning algorithms, respectively.
Somatic category: Mutect2 (MT2)
Germline category: HaplotypeCaller (HC) with ploidy 20 and 200 (HC20, HC200)
Ensemble category: M2S2MH that consists of the combined use of three different callers (MosaicHunter, Mutect2, and Strelka2)

RESULTS

Quality validation of positive and negative controls



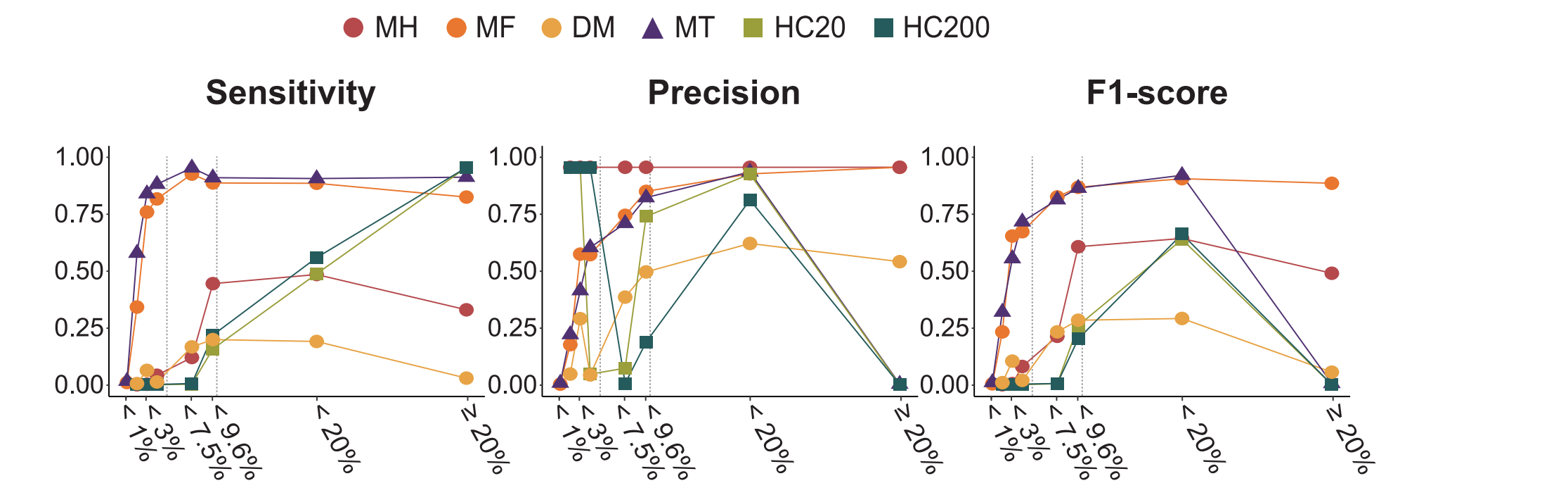
To validate the quality of positive controls, we inspected the correlations between expected VAFs (variant allele frequencies) of the design and observed VAFs. Both SNVs and INDELs in the entire range of VAFs had a high coefficient of Pearson correlation (r = 0.97, p < 2.2e-16 and r = 0.91, p < 2.2e-16, respectively).



A median of 10,202,428 positions with unexpected alternative alleles was found in set A, which was approximately 30% of the total targeted regions in ultra-high depth data (1100×). They had a wide range of base quality (0 to 80), and artifacts were concentrated at VAF near 0.001, with a base quality of zero. However, **a notable number of artifacts was found with high base quality**, which are destructive in mosaic variant detection.

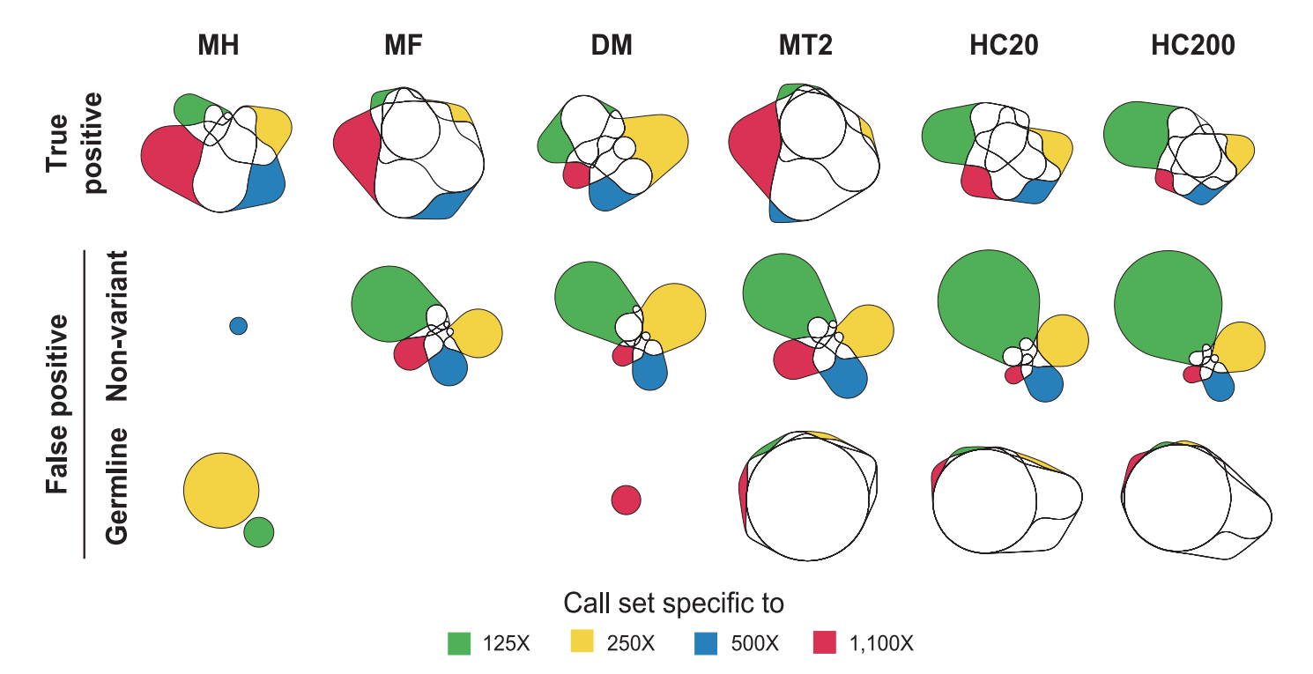
Evaluation of single sample-based calling

The detection of mosaic variants in a single sample is analogous to conventional somatic variant calling without a matched control. The evaluation revealed that the detection methods exhibit benefits and pitfalls of their performance in **category-specific manner**. In a high-depth (1,100×) setting, MF and MT2 showed the best F1-score in detecting mosaic SNVs, with robust sensitivity and precision in a wide VAF range; MT2 showed higher sensitivity and lower precision than MF.

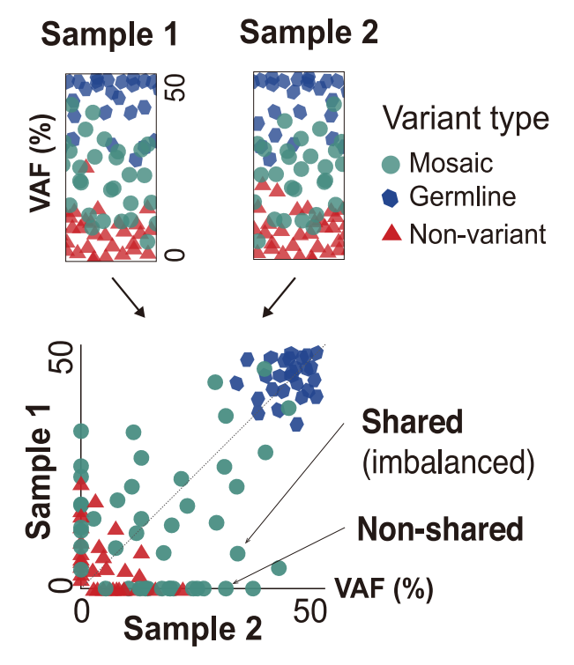


We found that most of the performance gain in MF and MT2 was achieved at low VAFs (<10%), benefiting from the strength of somatic variant calling; MF takes the raw calls of MT2 as input. In the high-VAF area (≥20%), other approaches showed their own strengths; for example, HaplotypeCaller (HC) with additional ploidy setting showed high sensitivity, and mosaic callers (MH, MF, and DM) showed high precision.

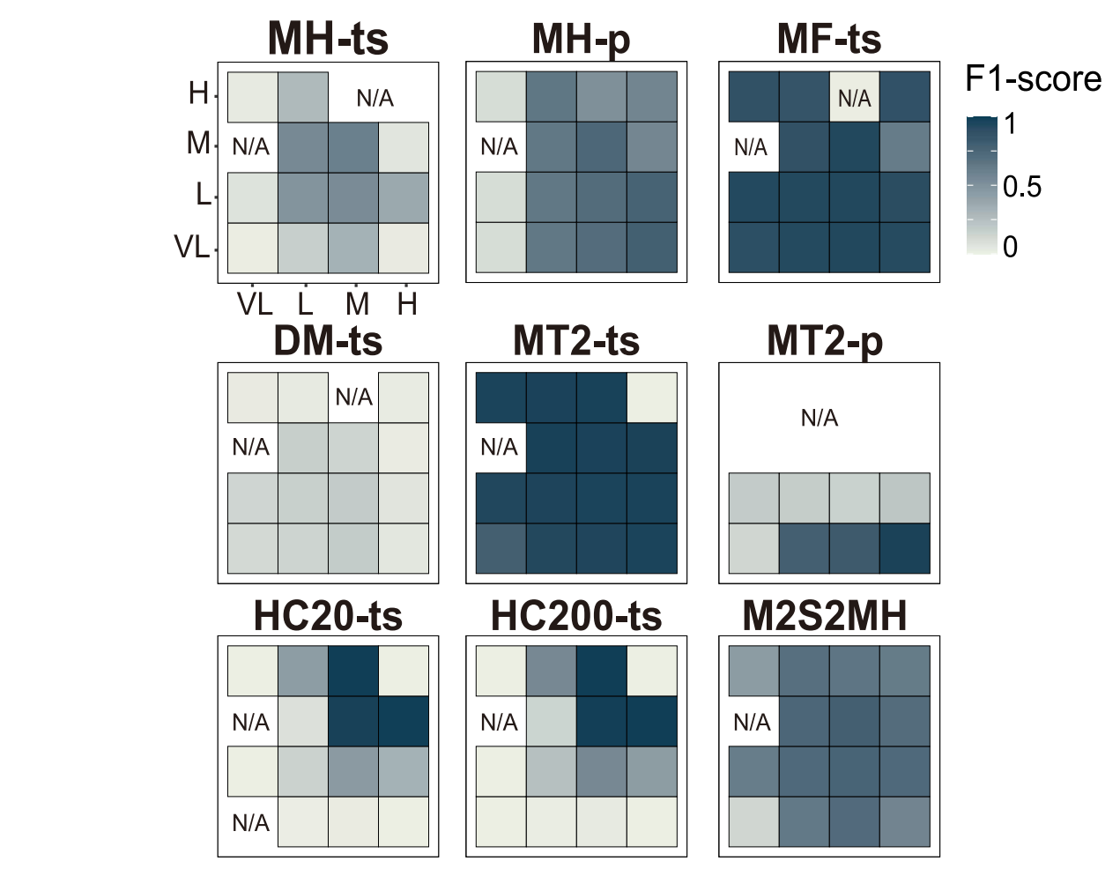
Further assessment of the **consistency of the call sets** towards read depth (125×, 250×, 500×, and 1,100×) showed an unexpected behavior. We observed a substantial loss of true positives and an extra gain of false positives at higher sequencing depths in indicates that the current approaches should consider various sequencing depths in their model construction.

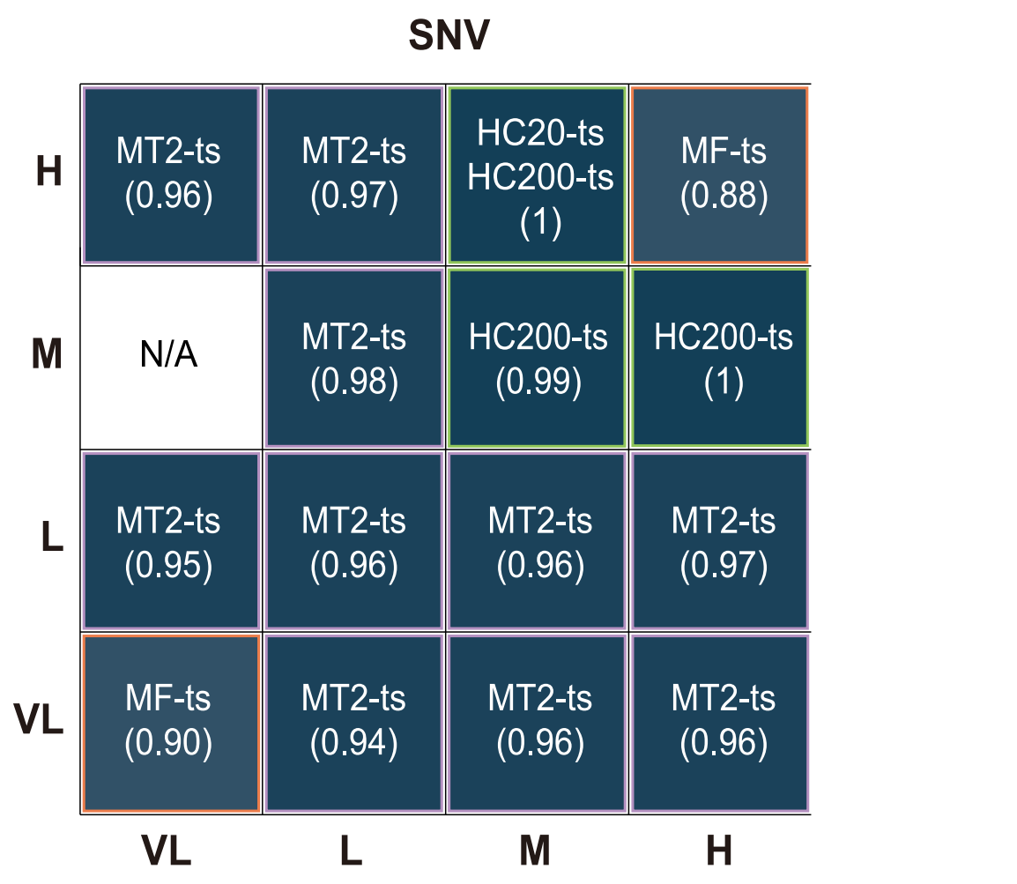


Evaluation of variant detection in paired samples



In a sample pair, mosaic variants can exist either in one or both samples, comprising a non-shared or shared form. We evaluated nine strategies that could be applied to detect **shared mosaic variants** in paired samples. Our evaluation revealed complex relationships among algorithms, strategies, and VAFs. With the lack of optimal models, instead of a single winner, the detection strength varied highly among the strategies. To quantify the accuracy of the shared variant detection, we partitioned the answer set into 16 (= 4 × 4) VAF areas: four ranges (very-low: < 5%; low: 5%–10%; medium: 10%–25%; and high: > 25%) for each sample, to describe the landscape of the best performing ranges.

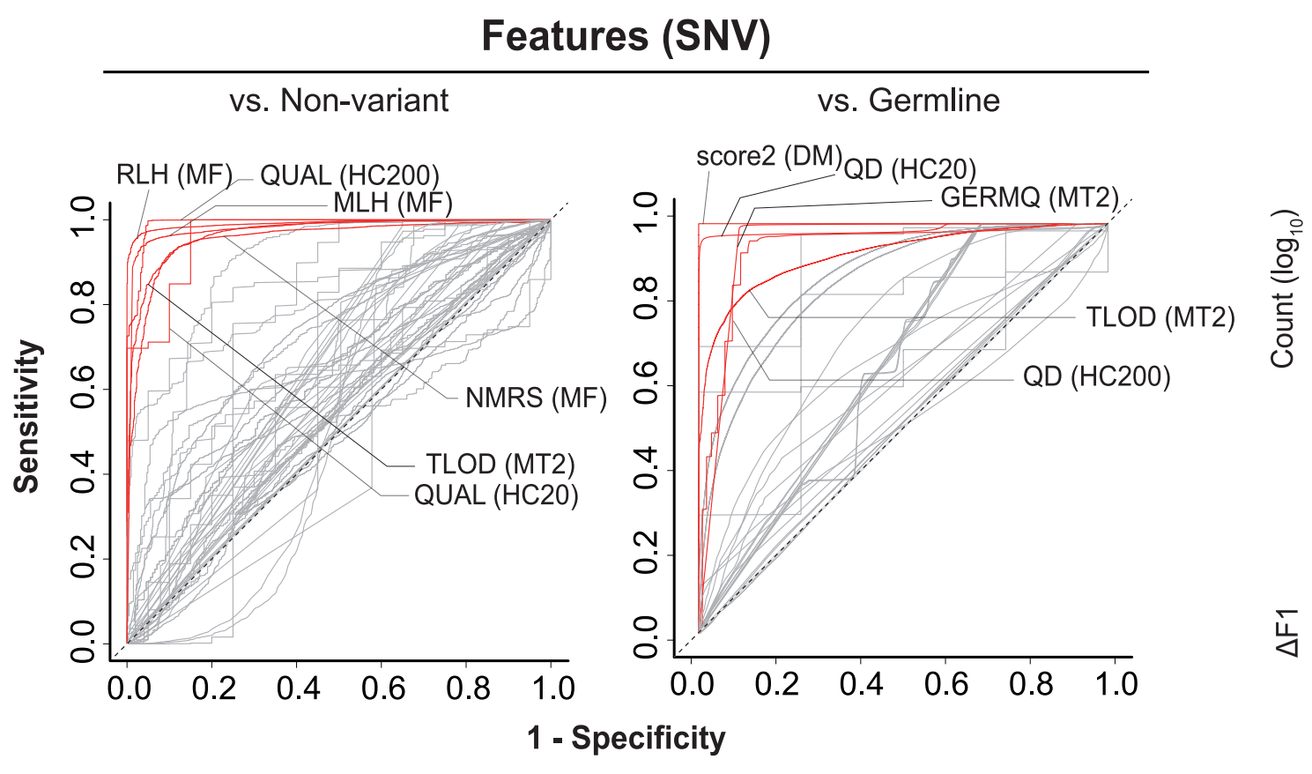


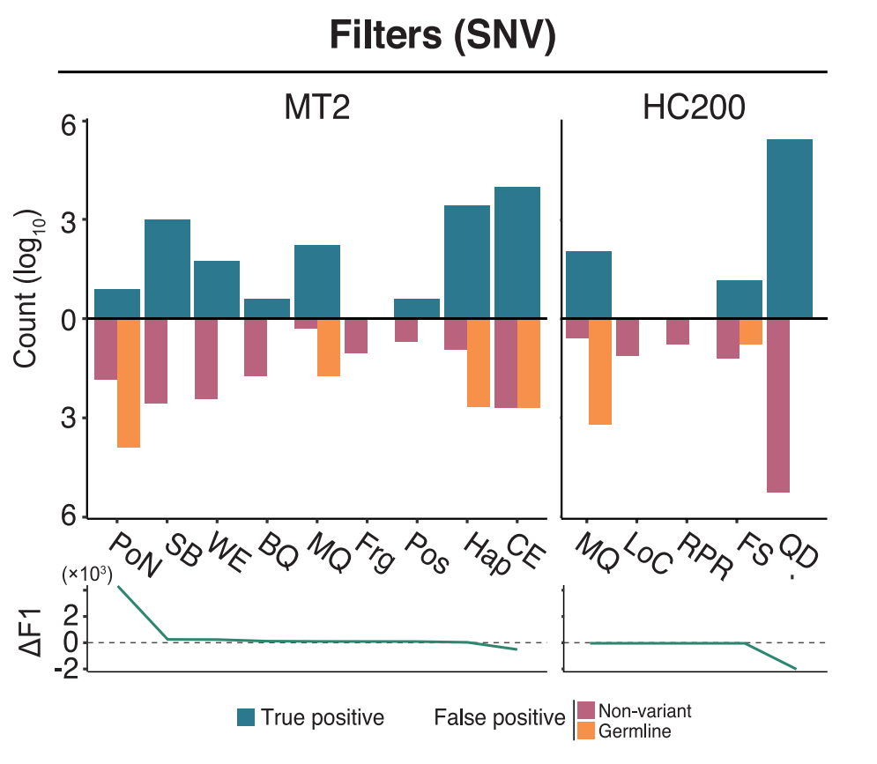


Although MF-ts and MT2-ts marked the best F1-score, in general, other algorithms showed a better performance within particular VAF areas (left). Mapping these "local winners" into the VAF space rendered the **current best practice for integrating multiple strategies.** Compared with the single best performing strategy, an ensemble of the five strategies increased the overall F-score from 0.89 to 0.96 (right).

Evaluation of the building blocks: features and filters

As variant calling algorithm is a decision process of selecting, calculating, and organizing such features, feature-level evaluation provided fundamental resources for developers. Using positive and negative calls, we evaluated forty-eight features that have been used in four different mosaic detection algorithms (MF, MT2, DM, and HC).

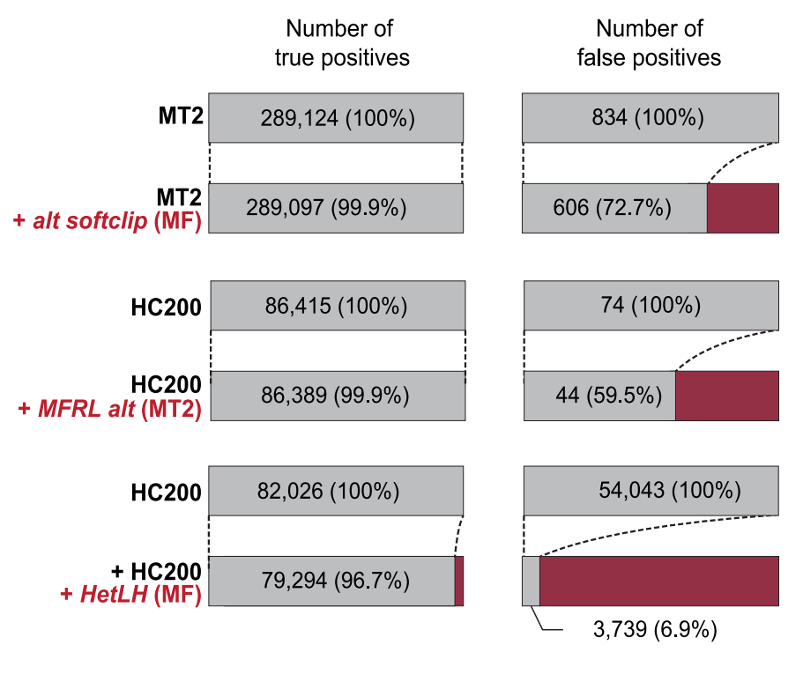




Eleven features had an AUC > 0.9 and AUC closer to 1 are potentially informative in further classifying the current false and positive calls (left). We also tested the efficiency of the 16 independently adjustable filters used in MT2 and HC200 by disabling and comparing the changes in the call sets (right). The contribution of the filters to the overall performance (F1-score) was limited (-0.002 to 0.038, mean = 0.003), implying that naïve, single threshold-based filtration is not an effective strategy for solving the mosaic variant calling problem.

Additional strategies for mosaic variant calling

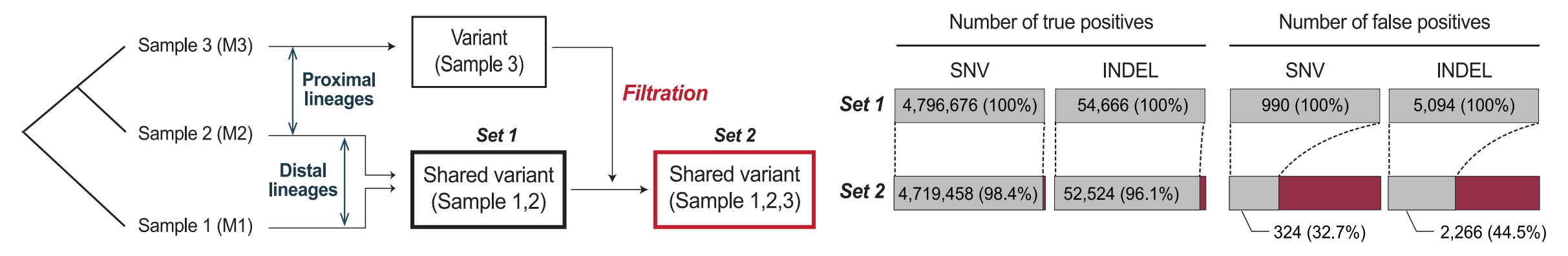
Call set and feature-level recombinations



Cross-reference of diverse features applied in multiple algorithms would lead to a more fundamental improvement in the short-term. For example, in a single sample setting, MT2 showed high sensitivity, but was accompanied by many false positives. We found that the MT2 call set could be efficiently improved by applying the foreign feature "alt softclip" developed for MF, which removed 27.3% (228/834) of false calls from wild-type sites and lost only 0.009% (27/289,124) of the true answers.

Exploiting multiple samples and their lineages

For improving the accuracy of shared variant detection, we found that utilizing ⩾3 samples can be useful, particularly for precision, as variants shared in two samples of distal lineages are likely to be present in the third sample within the lineages.



A pilot application of this idea was conducted on the 1,944 possible combinations, and 67.3% and 55.5% of the shared false positives of SNVs, and INDELs could be removed while losing only a small fraction (1.6% and 3.9%) of true positives, thereby further increasing the F1-score from 0.94 to 0.95 and from 0.18 to 0.29 for SNVs and INDELs, respectively.

REFERENCES

1. Thorpe, J., Osei-Owusu, I. A., Avigdor, B. E., Tupler, R. & Pevsner, J. Mosaicism in Human Health and Disease. Annu Rev Genet 54, 487-510, doi:10.1146/annurev-genet-041720-093403 (2020).
2. Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting Somatic Mutations in Normal Cells. Trends Genet 34, 545-557, doi:10.1016/j.tig.2018.04.003 (2018).
3. Ha, Y.-J. et al. Establishment of reference standards for multifaceted mosaic variant analysis. bioRxiv, 2021.2008.2031.458343 (2021).

ACKNOWLEGEMENTS

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2008050).

YONSEI UNIVERSITY
COLLEGE OF MEDICINE