A robust benchmark for evaluating and improving mosaic variant calling strategies

Yoo-Jin Ha

Yonsei University College of Medicine <https://orcid.org/0000-0003-1667-8054>

Jisoo Kim

Yonsei University College of Medicine

Seungseok Kang

Yonsei University College of Medicine

Junhan Kim

Yonsei University College of Medicine <https://orcid.org/0000-0002-5437-6033>

Se-Young Jo

Yonsei University College of Medicine

Sangwoo Kim (✉ swkim@yuhs.ac)

Yonsei University College of Medicine <https://orcid.org/0000-0001-5356-0827>

Article

Keywords:

Posted Date: January 11th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-871399/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A robust benchmark for evaluating and improving mosaic variant calling strategies

Yoo-Jin Ha¹, Jisoo Kim¹, Seungseok Kang¹, Junhan Kim¹, Se-Young Jo¹, and Sangwoo Kim^{1,*}

¹Department of Biomedical Systems Informatics, Brain Korea 21 PLUS Project for Medical Science, Yonsei University College of Medicine, Seoul 03722, Republic of Korea

*Correspondence should be addressed to: swkim@yuhs.ac

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 calling. Our results will guide researchers in selecting suitable calling algorithms and suggest future strategies for developers.

Post-zygotic mutations continuously occur along the zygote-to-adult trajectory, resulting in genetic mosaicism. Recently, the capabilities to directly detect mosaic mutations at the genome level have led to a series of discoveries, including the mutational processes and landscapes involved in human development^{1,2} and aging^{3,4}, the causes of neurological disorders⁵, and cancer predispositions⁶. As more research questions are being answered, there is growing attention on the complete investigation of mosaicism, warranted by the accurate detection of mosaic mutations.

The detection of mosaic mutations is an intricate process not only technically but also conceptually. Given the definition of genetic mosaicism (the presence of two or more genotypes), mosaic mutation is only loosely defined. Conventional somatic mutations fall within the boundary of mosaic mutations, as they lead to genetic differences between tissues. Mutations in the developmental process lead to a complex relationship among the affected and unaffected tissues; in other words, mutations may or may not be shared between a pair of samples. Variant allele frequencies (VAFs) vary widely ranging from extremely low ($< 1\%$) to the level of germline variants (approximately 50%) depending on the time and location of occurrence and can also be largely unbalanced among shared tissues because of the asymmetry in proliferation and selective pressure. This ambiguity is reflected in the disparate set of approaches applied in recent studies, such as targeting variants with unlikely VAFs for normal zygosity in a single sample^{7,8}, searching for shared variants in a pair of samples⁹, and machine-learning algorithms^{10,11}. These circumstances urgently demand a rigorous cataloging and assessment of mosaic detection algorithms, as conducted for germline and somatic variants¹²⁻¹⁷, but should be in a more sophisticated manner to cover the full extent of scenarios that mosaic variants can represent. Above all, the construction of robust and biologically compatible reference standards is a prerequisite.

Here, we present our benchmark for nine practically used mosaic variant detection approaches based on a newly constructed reference standard. The benchmark and reference standard have been systematically designed to evaluate the performance of the algorithms in multiple different conditions, including VAFs, sequencing depth, variant types, variant sharing, VAF balances, and the use of matched controls. For every analysis, we focus on drawing the strengths and weaknesses of the approaches and deducing the basis for the properties while reporting the accuracy of each algorithm. Finally, we investigate the effectiveness of internal features and their combinations and the potential use of multiple samples, which may improve accuracy.

Results

Benchmark setting

We constructed a new reference standard for evaluation. The detailed information about the procedure, quality control, and the final specifications is described in a separate data description paper¹⁸. Briefly, the reference standard material is a collection of 39 mixtures of six pre-genotyped normal cell lines, each one harboring 7,566 to 11,606 known single-nucleotide variants (SNVs) and insertion–deletion (INDEL) mutations in a wide spectrum of VAFs (0.5%–56%). In total, there were 374,565 and 12,048 control positive SNVs and INDELs, respectively, approximately 70% of which presented a low VAF (<10%)¹⁸. Likewise, 35,113,417 and 19,936 non-variant and germline variant sites are listed as negative control in each sample.

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 (**Fig. 1a**). The 39 mixtures belong to one of the three mixture categories (M1–M3) that symbolize three distinct descendants, each one harboring a common and lineage-specific set of variants acquired from cumulative cell line mixing. Thus, the comparison of the mixtures from different categories enables testing for both shared and non-shared variants. Moreover, different mixtures in a same category share the same set of variants, but with different VAFs, which are determined by the mixing ratios. Overall, there are 741 possible combinations of selecting two mixtures, which can represent nearly complete scenarios in the mosaic mutation detection, regarding variant sharing and VAF composition. The final dataset was prepared by conducting deep whole-exome sequencing (1,100×) of the mixtures, and multi-step down-samplings (125×, 250×, and 500×).

We selected nine mosaic detection strategies for evaluation. The inclusion criteria were: (1) algorithms that explicitly aim to detect mosaic mutations^{10,19}, (2) procedures that have been used previously to discover mosaic variants^{8,20}, and (3) algorithms that can be applied for mosaic mutation detection via simple modifications⁹ (**Methods**). As not all strategies are primarily developed to target mosaic mutations, some results do not represent the intrinsic performance of the baseline algorithm. The nine strategies were classified into four major categories based on their baseline algorithms: mosaic, somatic, germline, and ensemble (**Fig. 1b**). The mosaic category includes three algorithms that specifically target mosaic mutations: MosaicHunter (MH)¹⁹, MosaicForecast (MF)¹⁰, and DeepMosaic (DM)¹¹, which exploit Bayesian, Random-Forest, and deep-learning algorithms, respectively (**Fig. 1b upper**). Algorithms in the somatic and germline categories are modified procedures based on the somatic and germline variant callers. Somatic callers basically take a pair of samples to find mutually exclusive variants, but can also be used to find the shared

Figure 1. An overview of benchmark based on mosaic reference standards.

(a) Cumulative mixing of six pre-genotyped normal cell lines to construct reference standards. It mimics the mosaic variant acquisition process during the early development. A total of 39 reference standards were established depending on different mixing ratio and combination of the sources within three different mixture types (9 M1, 12 M2, and 18 M3). When selecting two samples, variants are shared or not, and the variant allele frequencies (VAF) of shared variants are balanced or unbalanced in respect to the selection among 741 possible combinations. (b) Landscape of the true and false positives in the reference standards (center) and the applied detection approaches are shown (outer). Density of the variants along the VAFs in single and paired-sample analysis is shown. The dashed line denotes VAF 10%, the data on the plane and axis refer to the shared (blue dashed rectangle) or non-shared variants, respectively. Possible false positives from two types of negative controls (non-variant and germline variants) are shown after 1/1000 down sampling. Four categories of detection approaches based on their baseline algorithms, mosaic, somatic, germline, and ensemble were applied to the datasets. Red and blue boxes and followed by arrows depict analysis type (single or paired) of each.

mutations by altering the filter usage (**Fig. 1b left**). Likewise, germline callers can be used to detect low- to medium-level mosaic mutations modifying the ploidy assumptions⁸ (**Fig. 1b right**). Although many callers can be applied to these categories, Mutect2 (MT2)²¹ and HaplotypeCaller²² (HC) with ploidy 20 and 200 (HC20 and HC200) were selected as representatives of the somatic and germline categories, respectively. The ensemble category includes one procedure (M2S2MH) that consists of the combined use of three different callers (MosaicHunter, Mutect2, and Strelka2²³) and filtration²⁰.

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 aim of this task is to sort out true mosaic variants from wild-type and germline sites, based on the sequence alignment and VAF. As wild-type sites are generally presented in low VAFs derived from sequencing artifacts (< 1%–5%) and the VAFs at germline sites follow binomial-like distributions centered at 50%, most approaches primarily target the variants with intermediate VAFs (**Fig. 2a**). Moreover, inherent uncertainty exists in distinguishing high-VAF mosaic variants from germline variants, which requires external information, such as the population-level allele frequency⁸. Six approaches based on five algorithms (MH, MF, DM, MT2, HC20, and HC200) support SNV calling, whereas four approaches (MF, MT2, HC20, and HC200) support INDEL calling and, therefore, were evaluated (**Methods**).

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 (4%–20%); MT2 showed higher sensitivity and lower precision than MF (**Fig. 2b**). 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 ($\geq 20\%$), other approaches showed their own strengths; for example, HC showed high sensitivity, and mosaic callers (MH, MF, and DM) showed high precision. For most approaches, increases in sequencing depth resulted in an improvement in sensitivity and precision, especially at low VAFs (<10%) (**Fig. 2c and Supplementary Fig. 1a**). The only exception was DM, which showed the best performance at 250×, at which the algorithm was trained¹¹. This implies that image-based deep-learning approaches can be further improved by diversifying the models for various read depths. For INDELs, the best performing approaches differed by VAF range: MF was the best approach at very-low (<5%) and high ($\geq 15\%$) VAFs, whereas HC20 was the best at medium VAFs (5%–15%) (**Fig. 2d**). Similar to SNVs, a gain in F1-score was observed at a higher read depth (**Fig. 2e and Supplementary Fig. 1b**). Nevertheless, the overall performance was far lower than that of SNV detection, demanding improvements.

Figure 2. Evaluation of single sample-based calling

(a) A schematic overview of mosaic variant detection in a single sample. True mosaic variants with varied allele frequencies (VAF) should be distinguished from the germline variants and diverse non-variant sites. (b and c) Sensitivity, precision, and F1-score in different VAF categories are shown, where each contains the same number of positive controls. (b) SNV detection performance of six applied approaches in the eight VAF categories in 1,100×. (c) INDEL detection performance of four applied approaches in the seven categories in 1,100×. (d and e) Comparison of the F1-scores of each approach among the three different sequencing depths: 125×, 250×, and 500× in (d) SNVs, (e) INDELs. (f) Diagram illustrating the consistency of the variant call sets within each approach towards four different four depths: 125×, 250×, 500×, and 1,100×. The consistency of true positives and two different types of false positives (non-variant and germline) are shown, with the relative sizes and relationships between the call sets of four read depths. Colored parts represent the call set specific to each depth (nVennR) (g) Similarity of call sets between different approaches. The similarities were calculated using the Jaccard index. MH, MosaicHunter; MF, MosaicForecast; DM, DeepMosaic; MT, Mutect2; HC20, HaplotypeCaller with ploidy option 20; HC200, HaplotypeCaller with ploidy option 200.

Further assessment of the consistency of the call sets towards read depth showed an unexpected behavior (Fig. 2f, Supplementary Fig. 1c for INDEL). With an increase in the sequencing depth, a gain in the number of true positives is commonly expected, without losing previous calls. Likewise, a proper reduction of the number of false positives is also predicted. However, we observed a substantial loss of true positives and an extra gain of false positives at higher sequencing depths in all approaches, which indicates that the current approaches should consider various sequencing depths in their model construction. There were also low overlaps among the call sets from different approaches, both in true positives and false positives (Fig. 2g, Supplementary Fig. 1d and 1e). In this case, however, the low consistency may indicate a possibility of further improvement by referring to the true and false calls from other algorithms, or by composing ensemble approaches.

Evaluation of paired sample-based calling

Mosaic variants generally affect multiple parts of an individual; thus, sampling more than a single site is a better option. In a sample pair, mosaic variants can exist either in one or both samples, comprising a non-shared or shared form (**Fig. 3a upper**). Detection of a non-shared mosaic variant is equivalent to the conventional somatic variant detection problem and can be handled using relevant callers; therefore, it was not prioritized here. By contrast, shared variants are harder to call as they should be distinguished from wild-type and germline variants simultaneously. VAFs comparably deviating from 50% (expected in heterozygous) in both samples are useful, but inter-sample imbalance can be present¹. Nevertheless, there are few tools that directly detect shared mosaic variants. Instead, the modified use of existing algorithms has been used alternatively⁷.

We evaluated nine and five strategies that could be applied to detect mosaic SNVs and INDELs, respectively, in paired samples. These strategies are divided into two major categories: two-single and paired (**Fig. 3a lower**). Strategies in the two-single category attempt to call mosaic variants in each sample and report their intersections. We applied MH, MF, DM, MT2, HC20, and HC200 to two-single approaches for the SNVs, and MF, MT2, HC20, and HC200 for INDELs (we will use the suffix ‘-ts’ to represent two-single). Strategies in the paired category take both samples together to call shared mosaic variants. We found that MH, MT2, and M2S2MH were applicable to SNVs, and MT2 was also applicable to INDELs (the suffix ‘-p’ will be used). Among them, M2S2MH is the only strategy that directly targets shared mosaic variants, whereas MH and MT2 required further modifications (**Methods**).

Our evaluations revealed even more complex relationships among algorithms, strategies, and VAFs (**Fig. 3b-3c, Supplementary Fig 2a-2d for all read depths**). With the lack of optimal models, the paired approach did not outperform the two-single approaches. Instead, the benefit was algorithm-specific. In particular, MT2-p showed lower sensitivity, especially at low (<5%) and high (>25%) VAFs, than MT2-ts. Conversely, MH-p showed better sensitivity than MH-ts, without an increase in the number of false positives. We assume that the joint genotyping model of MH led to the better utilization of the sample pairs. Regarding other callers, M2S2MH showed a robust performance by reinforcing the sensitivity of MH with a read-level rescuing procedure. Germline approaches (HC20-ts and HC200-ts) showed high sensitivity at medium to high VAFs (>10%), but called many false positives at VAFs > 25%, rendering the best performing area at 10%–25%. MF-ts showed a comparable performance to that of MT2-ts, with higher precision at very-low (<5%) and high (>25%) VAFs. Although the performance was specific to the baseline algorithms and VAF ranges, we noted that the paired approaches were less susceptible to VAF imbalances within samples (**Fig. 3d**), indicating the fact that tightly considering sample pairs would potentially be more beneficial.

203

218

221

The algorithm and VAF specificity of the performance suggests an instant way to improve the mosaic variant detection on an application level. We partitioned the answer set into 16 ($= 4 \times 4$) 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 (**Fig. 3e**). Although MF-ts and MT2-ts marked the best F1-score, in general, other algorithms showed a better performance within particular VAF areas, especially for INDEL detection (**Supplementary Fig. 3a**). Mapping these “local winners” into the VAF space rendered the current best practice for integrating multiple strategies (**Fig. 3f**). Compared with the single best performing strategy, an ensemble of the five strategies increased the overall F-score from 0.89 to 0.96 and from 0.52 to 0.60 for SNVs and INDELs, respectively.

Evaluation of the building blocks: features and filters

Variant calling utilizes multiple information as features, such as simple field values in raw data (e.g., the number and frequency of B-alleles or the average base-call quality), or specifically formulated measures (e.g., proportion of clipped alt-reads with ≥ 10 bp. Thus, a variant calling algorithm is a decision process of selecting, calculating, and organizing such features, and a feature-level evaluation provides 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; see **Supplementary Table 2** for the full list of features). The area under the curve (AUC) of the 48 features widely ranged from 0.5 to 1 (**Fig. 4a**), among which the features with an AUC closer to 1 are potentially informative in further classifying the current false and positive calls. Eleven features had an AUC > 0.9 and their efficiency differed by the variant type (SNV or INDEL) and error type (vs. wild-type and germline). For example, “*refhom-likelihood*” (RLH; AUC = 0.99) and “*mosaic-likelihood*” (MLH; AUC = 0.98), the likelihoods for wide-type and mosaic genotypes were calculated using MF, and “*QUAL*” (AUC = 0.99 and 0.96 in HC200 and HC20, respectively), a well-known confidence feature for genotyping in HC, were shown to further distinguish wild-type SNVs. Likewise, “*score 2*” (potential mosaic possibility; AUC = 0.99) in DM and “*QD*” (quality by depth; AUC = 0.98) in HC could be utilized for filtering germline SNVs. Although no prominent features are shown for wild-type INDELs, “*GERMQ*” (Phred-scaled quality that alternative alleles are not germline variants, AUC=0.99) in MT2, and “*QD*” (AUC=0.98) in HC20 were shown to be useful in distinguishing germline INDELs (**Fig. 4a right**). Overall, a more active use of these features would increase the overall performance in the short-term.

Figure 4. Evaluation of the features and filters.

(a) Receiver operating characteristics (ROC) of forty-eight features used in the four approaches (MF, MT2, DM, and HC) for mosaic SNV (left) and INDEL (right) detection are shown. Each feature was investigated on two types of false positives (non-variant and germline) if it can further classify the final variant set of own approach into true and false positives (Supplementary Table 2 for full names and definitions of the evaluated features). Features with AUC over 0.9 are presented in red lines. (b) The tested features were categorized into three categories, sequencing, alignment, and genotype level based on the original intention of their usage. Genotype-level features had significantly larger AUCs than the other two (Wilcoxon's rank sum test, $p = 0.00003$ with sequencing-level and $p = 0.00002$ with alignment-level), and the median of each category is shown in red. (c) The efficiencies of 16 independently adjustable post-filters from MT2 and HC200 were tested by obtaining the additional true and false positives when each of them was disabled. The number of true and false positive counts are shown in log10 scale (upper) and the resultant differences in F1-scores are shown (lower). Delta F1 score was calculated by F1-score in default settings subtracted by the F1-score with disabled-filter. RLH Refhom likelihood, MLH Mosaic likelihood, NMRS Mismatches p, PoN Panel of normals, SB Strand bias, WE Weak evidence, BQ Base qual, MQ Map qual, Frg Fragment, Pos Position, HT Haplotype, CE Clustered events, GL Germline, Slp Slippage, MtA Multiallelic, LoC LowCoverage, RPR ReadPosRankSum

We inspected the generalized properties of the informative features (**Fig. 4b**). The sources, scales, and complexity of the 48 features were highly diverse, but could be categorized into three different levels: sequencing, alignment, and genotype. Sequencing level features are the raw values and their derivatives regarding sequencing reads, base-call, and quality. Alignment-level features are values that annotate the patterns and noises in read mapping and genotype-level features are intermediate scores in the genotyping of positions, such as likelihoods and confidence values. We found that most of the informative features (12/13, 92%) were of the genotype level. Likewise, category-level

grouping of AUCs showed that the genotype-level features were substantially more informative than the other categories (average AUC = 0.77 vs. 0.59, $p = 3 \times 10^{-5}$ and 2×10^{-5} for the sequencing- and alignment-level features, respectively) (**Fig. 4b**). Therefore, we suggest that opportunities for improvement lie in the better use of the genotype-level features.

Next, we evaluated another usage of such information: a filter with which the call sets are post-processed and refined usually by using a single threshold value. A good filter is expected to remove the false calls, while leaving the true calls, and increase the overall accuracy (e.g., F1-score). We tested the efficiency of the 16 independently adjustable filters used in MT2 and HC200 by disabling and comparing the changes in the call sets (**Fig. 4c upper**). We found that most of the filters removed a substantial number of false positives (0 to 233,874); however, it was also accompanied by the corresponding number of lost true positives (0 to 160,967). Overall, the contribution of the filters to the overall performance (F1-score) was limited (-0.002 to 0.038, mean = 0.003) (**Fig. 4c lower**). These results imply that the naïve, single threshold-based filtration is not an effective strategy for solving the mosaic variant calling problem, and should not be prioritized in the development of new algorithms.

Stepping forward: additional strategies for mosaic variant calling

We have shown that there is no single algorithm that fits all mosaic variant types and more major breakthroughs are evidently needed. Here, we propose and test two strategies that can direct further advances for reinforcing the current approaches or for developing new algorithms.

First, we considered a call set- or feature-level recombination of multiple algorithms. We have already shown that an ensemble of multiple algorithms can instantly improve the performance of paired sample-based detection. Similarly, 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 at the VAF areas, mostly germline variants. 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 (**Fig. 5a top**). Likewise, applying the MT2 feature “*MFRL alt*” (a median fragment length of reads with alternative alleles) and the MF feature “*Het Likelihood*” (genotype likelihood of the variant being germline heterozygous) to the HC200 call set removed 40.54% (30/74) and 93.08% (50,304/54,043) of the false calls from the wild-type and germline sites, respectively, losing 0.03% (26/86,415) and 3.3% (2,732/82,026), respectively, of the true calls (**Fig. 5a middle and bottom**). In the share-variant detection, we noted that the “rescue” procedure (**Method, Fig. 5b**) of M2S2MH was effective,

Figure 5. Additional strategies for mosaic variant calling

(a) Accuracy enhancement by the call set- and feature-level recombination of multiple approaches. Three examples with the “alt softclip” in MF, “MFRL alt” in MT2, and “HetLH” in MF, applied to the call sets of MT2 with false positives from non-variant sites, HC200 with false positives from non-variant sites, and HC200 with germline false positives, respectively. Removed percentage of true or false positives with the applications are shown in red. (b) A schematic overview of the “rescue” procedure of M2S2MH approach. A shared variant that detected by only in the either of two samples could be rescued with read-level information. (c) The improvement in F1-scores of before and after applying the “rescue” procedure to MosaicHunter (MH), MosaicForecast (MF), and DeepMosaic (DM), F1-scores shown in partitioned variant allele frequency (VAF) areas as in Figure 3e. (d) Strategy for precision enhancement by utilizing multiple samples (≥ 3) and their distances in the developmental lineage (left). Variants not shared in Sample 2 and Sample 3 (proximal lineages) are unlikely to be shared in a sample in distal lineages (Sample 1 and Sample 3). Like so, the shared variants among Sample 1 and Sample 2 (named as Set 1) were filtered out if they were not present in Sample 3 (Set 2). The removed true and false positives when applied to shared SNVs and INDELs are shown (right). In total of 1,944 combinations among 39 reference standards in three categories (M1, M2, M3) were tested with MT2 call sets as an example.

increasing the F1-score by 2.5-fold (0.26 to 0.64) from the original MH calls. A similar approach of using different baseline callers (MF and DM) also achieved an increase in the F1-score by 0.83 to 0.93 and 0.1 to 0.36, respectively (Fig. 5b). We presumed that testing full combinations of call sets and features, preferably in a VAF-specific manner, may bring instant benefits without developing a new algorithm *ab initio*. Second, the usage of multiple samples was tested. Although no such algorithms have been developed yet, considering ≥ 3 samples can be useful, particularly for improving precision. For example, variants shared in two samples of distal lineages (e.g., ectoderm and mesoderm) are likely to be present in the third sample within the lineages (e.g., another ectoderm). Similarly, variants not shared in the same lineage are unlikely to exist in a distal lineage. A pilot

application of this idea was conducted on the 1,944 possible combinations of three different mosaic samples (**Fig. 5c**). We found that 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 (**Fig. 5d**). Thereby, we anticipate that a generalized algorithm for using multiple tissues would increase the sensitivity and precision of mosaic variant calling.

Discussion

In recent decades, the technologies used for detecting germline and somatic mutations have been greatly improved. The F1-scores of germline variant calling have exceeded 0.99¹⁷. Clonal somatic mutation calling (e.g., cancer) has gained credence on a clinical level, by lowering the limit of detection down to around 1%²⁴. A number of benchmarks and competitions have led to the best trustworthy practices^{13,25}. Contrarily, this is not the case for mosaic mutations. Compared with its biological importance, the techniques used for mosaic variant calling still remain in their infancy. The lack of robust benchmark studies perplexes researchers. Moreover, the complex presentation patterns and restricted definition of mosaic mutations aggravate the problem. We presumed that, currently, the resolution of this uncertainty is a more acute problem than developing new algorithms.

The construction of comprehensive and versatile reference standards is the first step in achieving a robust benchmark. So far, several different approaches have been used for germline and somatic variants. Setting up a cooperative, public standard model has provided a good reference standard (e.g., Genome in a Bottle¹⁴) for germline variants, and similar efforts are being made for somatic variants¹⁶. However, deriving a robust standard reference is yet to be accomplished for these variants due to the intrinsic difficulties in finalizing confident true calls. The direct engineering of the genome using gene editing technologies (e.g., CRISPR-cas9) is another approach and has been used to produce commercial products²⁶. Nonetheless, the small number of true answers confines its usage on an application level (e.g., validation of clinical panels). Generating *in silico* simulated datasets (e.g., mixing BAM files) is a simple but powerful method¹⁶. Although this can popularize the following benchmark studies, we noted that the error profiles, especially the sequencing errors, are restricted in the source data and cannot represent the real-world level artifacts (**Supplementary Fig. 3b**). Overall, we believe that our approach to construct a standard reference is costly and time-consuming, but the most proper way for securing both the scale and robustness.

Despite all efforts, this benchmark has potential pitfalls, particularly in the interpretation of the analysis results, limitation in the search space, and data dependency. First, the performance of the mosaic calling “strategies” should not be confused with their baseline algorithms, especially when they were used in an unintended way. For example, the performance of MT2-p does not directly indicate the somatic mutation calling performance of MT2, with the modified use of normal filters. Likewise, MH-p originally reported variants with inconsistent genotypes in two samples (e.g., germline in one and mosaic in the other) and has been modified to call the shared variants by referring to the internal genotype probability matrix. The composition of all the two-single approaches for shared variant detection (MH-ts, MF-ts, DM-ts, MT2-ts, and HC20/200-ts) are generally acceptable; however, the usages were not explicitly declared in the original algorithms. Again, these

modifications were conducted to test the potentially applicable strategies in the absence of specifically developed algorithms. Moreover, the complexity of the use of parameters limits the benchmark. There are at least 4 to 110 parameters whose use can be adjusted and the number of combinations of which reaches up to tens of millions. Because of the intractability, evaluations have been conducted using default parameters, assuming that the empirical suggestions from developers is close to optimal. Finally, the composition of the standard reference can affect the results, such as the distribution of VAFs in the datasets, sequencing platform, read length, and error profiles. For example, the cumulative cell line mixing (**Fig. 1a**) produced a large set of robust mosaic variants; however, it can also alter the composition of germline variant sites. We resolved the problem of the loss of pure germline sites by preparing a separated set of reference standard that reserved germline sites¹⁸. Notably, there are remaining unavoidable noises at the germline variant sites of the original cell lines. Although the effect is assumed as ignorable (**see Supplementary Notes**), we should be aware of it when evaluating tools that consider the alignment patterns of the flanking regions.

In summary, we present the first systematic benchmark of mosaic variant calling strategies. Our analysis revealed the sequencing depth-, VAF-, and error type-specific strengths and diversity of the current algorithms, feasible strategies for ensemble approaches, and directions for future development. We anticipate that our study will be a good starting point for the technical advances in mosaic variant calling at the germline and somatic variant levels.

Online Methods

Datasets

We obtained a set of mosaic reference standards (BAM) based on the accumulative mixing of pre-genotyped normal cell lines from the Sequence Read Archive [PRJNA758606]. Diverse combinations and ratios of the mixtures generated 39 reference materials, harboring both abundant mosaic variants of varied allele frequencies (**Supplementary Table 1**) and two types of negative controls, non-variant sites (Set A) and germline variants (Set B)¹⁸. Moreover, three mixture categories (9 M1, 12 M2, and 18 M3) that represent the genotypes of distinct lineages were highly useful for evaluating the detection performance of shared and non-shared variants among different tissues in 741 possible combinations. Among the full sets of controls in every reference material, we took only high-confident regions into account by excluding the simple repeats and segmental duplications supplied by UCSC. Thus, 345,552 positive SNVs (92%) and 8,706 INDELs (72%) were adjustable, and each of the 39 reference materials contains two types of negative controls, 33,111,725 non-variant (94%) and 18,151 germline sites (91%). Information on the positive and negative controls used here is shown in **Supplementary Table 1**. We also performed multi-step down-samplings (125×, 250×, and 500×) of original reference materials (1,100× on average) for a comprehensive assessment of performance under different sequencing depths.

Variant calling

Nine detection approaches were applied to the 39 pairs of reference standards. For the single sample and two-single analyses, MosaicHunter (v.1.0, single mode); MosaicForecast (v.0.0.1, 250× trained models for SNVs); DeepMosaic (v.0.0, efficientnet-b4_epoch_6.pt); Mutect2 (4.1.9.0, tumor only mode and applied FilterMutectCalls); GATK HaplotypeCaller (4.1.8.0) with a ploidy option of 20 and 200, with the quality filters adjusted according to the criteria based on GATK Variant Quality Score Recalibration (VQSR, $QD \geq 2$, $FS \leq 60$, $DP \geq 20$, $MQ \geq 40$, $ReadPosRankSum \geq -8$, and $-2.5 \leq MQRankSum \leq 2.5$). Raw variant calls from Mutect2 were used as input for MosaicForecast and DeepMosaic^{10,11}. Variants tagged as “mosaic” and “PASS” were only kept for downstream analysis. The population frequency data, dbSNP(b154) for Mosaic Hunter and panel of normal (1000g_pon.hg38.vcf.gz) from GATK resource bundle (<https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0>) for Mutect2 were applied after the removal of positive controls, since they were derived from mutually exclusive germline variants¹⁸. Segmental duplication and simple repeats from UCSC were

used whenever it was recommended to remove repeats to filter out confounding regions^{10,11,19}. For paired-sample analysis, MT2-p, and MH-p were applied with simple modifications to detect the shared mosaic variants. Variants of Mutect2 paired mode tagged as “normal artifacts” were selected for MT2-p to be exploited as an alternative filtering strategy for shared variant detection. In MH-p, paired naïve mode, if (1) the joint probability of two samples with “mosaic” variants was over 0.05 and (2) if it was larger than that of any other genotype combination, the variants were considered as shared, whereas the remaining variants remained as sample-specific. In the M2S2MH approach, Mutect2 (paired mode) and Strelka2 (v.2.9.10, somatic) with Manta (v.1.6.0) which was used for removing small INDELs,²³ were applied to the sample-specific variants. For shared variant detection, MosaicHunter (single mode) was applied to each sample, followed by a comprehensive filtering process using the read counts, depths, and VAFs of both samples, as previously mentioned²⁰. The detailed pipelines of the approaches used for all nine variant callings are shown in **Supplementary Fig. 4**.

Performance evaluation of single sample analysis

For each approach, the precision, sensitivity, and F1-score were calculated based on the call set. To investigate the detection performance, the variants were divided into eight categories based on their VAF (<1%, 1%–2%, 2%–3%, 3%–4%, 4%–7.5%, 7.5%–9.6%, 9.6%–20%, and >35%) for the SNVs, and into seven categories (<1%, 1%–2%, 2%–4%, 4%–5%, 5%–9.6%, 9.6%–15%, and 15%–22%) for the INDELs, so that equal numbers of variants were contained in each VAF category. Precision was recalibrated based on the density of positive controls (the number of positive controls per megabase), because a high positive control density overemphasizes the true positives over the false positives. The recalibration was done as follows with w , the weight of the overrepresented density of the positive controls:

$$i) w = \frac{\# \text{ expected positive controls}}{\# \text{ positive controls in the data}}$$

$$ii) \text{ Precision} = \frac{TP \times w}{(TP \times w) + FP}$$

The number of expected positive control was estimated based on the known prior densities of SNVs (1 per 1 Mb) and INDELs (0.1 per 1 Mb)^{4,6}, yielding w values of 252.75 and 81.63, respectively.

The variant call consistency within an approach was investigated in three categories; the true positives and false positives from two different types of negative controls revealed an interesting relationship within each variant set of varied sequencing depth: 125×, 250×, 500×, and 1,100×. We used the R

package nVennR²⁷ (0.2.3) for the set analysis. We also observed the similarity of variant calls among the approaches using the Jaccard index (**Fig. 2f** and **Supplementary Fig. 1d-1e** for all sequencing depths). The detailed relationships of the variant calls with the VAFs in each intersection are shown in **Supplementary Fig. 1f**.

Performance evaluation of paired-sample analysis.

We analyzed the paired samples for shared variant detection in the two possible ways, namely, two-single and paired analyses. Six (MH-ts, MF-ts, DM-ts, MT2-ts, HC20-ts, and HC200-ts) and four two-single approaches (MF-ts, MT2-ts, HC20-ts and HC200-ts) were used for SNV and INDEL detection, respectively. In the two-single mode, the intersected variant call sets from two samples were collected and considered shared variants. Concurrently, MT2-p (SNV and INDEL), MH-p (SNV), and M2S2MH (SNV) could be evaluated for shared mosaic variant callings, in which the variant sites of the evaluated samples were joint-genotyped. We evaluated the performance of 378 combinations from M1 + M2 (21) and M3 (18), which generated 34 different VAF relationships of 1,481,274 shared variants, in total (162 combinations with M1 and M3, 216 with M2 and M3, and 2,697 and 4,835 shared variants, respectively). The calculated sensitivity in all possible VAF combinations and the distributions of false positives of shared SNVs and INDELs are shown, clearly presenting the distinctive characteristics of each approach (**Fig. 3a and 3b**). The full performance comparisons within different sequencing depths (125×, 250×, and 500×) are shown in **Supplementary Fig. 2a-2d**. We also analyzed the sensitivity limitations of shared variant detection in two categories—balanced and unbalanced—with the latter included variants whose difference in variant allele frequency between two samples was greater than 2-fold.

Then, we divided the VAFs into four ranges, very-low (VL), low (L), medium (M), and high (H), based on a VAF <5%, ≥5% and <10%, ≥10% and <25%, and ≥ 25%, respectively, for generating 16 shared VAF combinations to quantify the detection performances according to the VAFs. Among the 16 possible combinations, we could assign the shared variants of the reference standards to 15 groups, and by gathering the best F1-score for each combination, we could suggest an instant enhancement of shared variant detection using an ensemble of different detection approaches. The F1-scores were quantified within the each grouped VAF level for SNV and INDEL (**Fig. 3e and Supplementary Fig. 3a**) detection. The overall F1-score was calculated after normalization of the positive control counts in each VAF bin, to ensure that computed F1-score is not limited to this study.

Evaluation of features and filters

To calculate the potential power of the building blocks and features and to further classify the true and false positives for an improved accuracy, the AUCs of 48 features of MF, MT2, DM, HC20, and HC200 were calculated. True and false positives within the approaches were applied to pROC (v.1.17.01) and the AUC of each feature was determined. In total, 13 features among 48 were observed to have an AUC > 0.9 in SNV and INDEL detection. Then, they were divided into three levels based on their intended use: 20 features into the sequencing level, 15 into the alignment-level, and 13 into the genotype level. The full sets of evaluated features with the definitions, AUC, and categorizations are listed in **Supplementary Table 2**. We observed that the genotype-level features had significantly higher AUCs than those at the sequencing ($p = 3 \times 10^{-5}$, Wilcoxon's rank sum test) and alignment levels ($p = 2 \times 10^{-5}$, Wilcoxon's rank sum test).

To further investigate the efficiency of the filters that could be assessed independently, 16 post-filters from MT2 and HC were tested. By disabling each filter, we compared the new F1-score to the original F1-score under default settings by collecting additional true and false positives. The difference in F1-score for each filter was calculated by subtracting the newly calculated F1-scores from the original F1-score. The detailed numbers of the additional true and false positives are shown in **Supplementary Table 3**. The post-filters of MH were excluded as they were utilized in series; in other words, calculating their efficacy was highly dependent on the order of their application.

Feature-level recombination

We tested whether a foreign feature of a distinct approach could be utilized in an independent variant call set from other approaches. We validated this hypothesis on three cases by observing a significant removal of false positives (27.3%–97.08%), and a small proportion of true positives (0.0009%–3.3%) were lost. Variants in the original call set could be tested when they were found in another approach accompanying the foreign features. For example, adjustable MT2 calls (99% of true and 92% of false positives in non-variant sites were adjustable) using the “*alt softclip*” of MF (which was removed if the value was greater than 0.05) could filter out 27.3% of false positives with a 0.009% loss of true calls. Likewise, HC200 variant calls could be filtered using the “*MFRL alt*” of MT2 (<150, to 0.99% of true and 0.11% of adjustable false positives from non-variant sites) and the “*Het likelihood*” of MF (>0.25, 95% of adjustable true and 8% of germline false positives) could filter out 40.54% and 93.08% of false positives with an extremely small loss of true calls (0.03% and 3.3%, respectively).

Lineage distance-based filtering

To speculate on the advantages of using multi-samples (three, in here) for an enhanced shared mosaic variant detection, we applied the shared variant call sets of MT2, of 1,944 combinations generated from 9 M1, 12 M2, and 18 M3, with each mixture type denoting different parts or tissues in an organism. We first collected the shared SNVs and INDELs among M1 and M2, which were more distal in lineage than M2 and M3. The variants were filtered out if they were not present in M3, given that the variants were unlikely to be shared in a more proximal lineage. The resultant shared variants were compared to the original variant sets.

Code availability

The scripts used for evaluation is available in a public repository GitHub²⁸.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2008050). The authors thank MID (Medical Illustration & Design) for helping to design the figures.

Author Contributions

Sa.K. and Y-J.H. designed and initiated the study. Y-J.H., Ji.K., Se.K., and Ju.K. conducted the main analysis. Y-J.H. and Sa.K. wrote and edited the manuscript with input from coauthors. S-Y.J. worked on data visualization. Sa.K. led the project.

Competing Interests

The authors declare that they have no competing interests.

References

1. Bae, T. *et al.* Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550-555 (2018).
2. Ju, Y.S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714-718 (2017).
3. Vijg, J. & Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* **182**, 12-23 (2020).
4. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911-917 (2018).
5. Lim, E.T. *et al.* Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat Neurosci* **20**, 1217-1224 (2017).
6. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
7. Freed, D. & Pevsner, J. The Contribution of Mosaic Variants to Autism Spectrum Disorder. *PLoS Genet* **12**, e1006245 (2016).
8. Wang, Y. *et al.* Comprehensive identification of somatic nucleotide variants in human brain tissue. *Genome Biol* **22**, 92 (2021).
9. Son, H. *et al.* Multi-organ analysis of low-level somatic mosaicism reveals stage- and tissue-specific mutational features in human development. *bioRxiv*, 2021.2008.2023.457440 (2021).
10. Dou, Y. *et al.* Accurate detection of mosaic variants in sequencing data without matched controls. *Nat Biotechnol* **38**, 314-319 (2020).
11. Yang, X. *et al.* DeepMosaic: Control-independent mosaic single nucleotide variant detection using deep convolutional neural networks. *bioRxiv*, 2020.2011.2014.382473 (2021).
12. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**, 555-560 (2019).
13. Chen, Z. *et al.* Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep* **10**, 3501 (2020).
14. Zook, J.M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-251 (2014).
15. Zook, J.M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38**, 1347-1355 (2020).
16. Ewing, A.D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623-630 (2015).
17. Zhao, S., Agafonov, O., Azab, A., Stokowy, T. & Hovig, E. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci Rep* **10**, 20222 (2020).
18. Ha, Y.-J. *et al.* Establishment of reference standards for multifaceted mosaic variant analysis. *bioRxiv*, 2021.2008.2031.458343 (2021).
19. Huang, A.Y. *et al.* MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Res* **45**, e76 (2017).
20. Breuss, M.W. *et al.* Autism risk in offspring can be assessed through quantification of male sperm mosaicism. *Nat Med* **26**, 143-150 (2020).
21. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219 (2013).

598 22. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
599 generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

600 23. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**,
601 591-594 (2018).

602 24. Shin, H.T. *et al.* Prevalence and detection of low-allele-fraction variants in clinical cancer samples.
603 *Nat Commun* **8**, 1377 (2017).

604 25. Koboldt, D.C. Best practices for variant calling in clinical sequencing. *Genome Med* **12**, 91 (2020).

605 26. Suzuki, T., Tsukumo, Y., Furihata, C., Naito, M. & Kohara, A. Preparation of the standard cell lines
606 for reference mutations in cancer gene-panels by genome editing in HEK 293 T/17 cells. *Genes*
607 *Environ* **42**, 8 (2020).

608 27. Perez-Silva, J.G., Araujo-Voces, M. & Quesada, V. nVenn: generalized, quasi-proportional Venn and
609 Euler diagrams. *Bioinformatics* **34**, 2322-2324 (2018).

610 28. Yoo-Jin Ha, J.K., Seungseok Kang, Junhan Kim, Se-Young Jo, and Sangwoo Kim* Benchmark for
611 improved accuracy in mosaic variant calling. *GitHub* (2021).

612

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NBTBMsubsupple.docx](#)
- [NBTBMSupTable.xlsx](#)