Improving the performance of ClairS and ClairS-TO with new real cancer cell-line datasets and PoN

Ruibang Luo, Zhenxian Zheng, Lei Chen
School of Computing and Data Science, HKU

Contents

- Overview
- Results
    - How could ClairS make the best use of real cancer cell-line datasets?
    - Could synthetic samples compensate for missing a cancer type in the real samples?
    - BQ jittering in model training to accommodate different BQ distributions in multi-center generated datasets
    - Will the 'SS+RS model' improve the performance at lower tumor purities or higher levels of normal contamination?
    - Compare the performance of ClairS v0.4.0 and DeepSomatic v1.7.0 for ONT somatic variant calling at different AF ranges and coverages
    - Improve ClairS-TO performance by "tagging indels at sequence with low entropy" and "using CoLoRSdb as PoN"
    - How would ClairS-TO benefit from using real samples for model training?
    - Compare the performance of ClairS-TO v0.3.0 and DeepSomatic v1.7.0 for ONT tumor-only somatic variant calling at different AF ranges and coverages
- References
- Supplementary Materials
    - Supp Table 1. The number of somatic/germline/artifact variants derived from four real cancer samples H1437, HCC1937, HCC1954, and H2009, for the SS and SS+RS model training

Overview

ClairS was first released in January 2023 [1]. Back then, only one pair of tumor/normal cell-line HCC1395/BL sequenced by the SEQC2 consortium was available. About 40k SNVs and Indels were found in the pair. There were too few variants to train a deep neural network for general-purpose somatic variant calling. Alternatively, we trained models with synthetic tumor samples generated from the real data of normal samples, considering that the germline variant specific to a sample can be considered as a somatic variant to another sample if the data of two samples are combined. The method can theoretically generate unlimited synthetic tumor samples at any tumor purity, normal contamination, or allelic fraction spectrum [2]. The use of synthetic tumor samples was successful. Nevertheless, there was no question that synthetic tumor samples would fail to cover some characteristics that can only exist in real tumor samples, such as cancer-specific indel length distribution and cancer-type specific mutational signatures.

In August 2024, Park et al. from UCSC and Google released the real data of five pairs of cell-lines, including HCC1395/BL, HCC1937/BL, HCC1954/BL, H1437/BL, and H2009/BL, together with a new somatic variant calling named DeepSomatic [3]. The work is significant, and it has enabled the Clair team to study how real cancer cell-line datasets can further improve the performance of ClairS and ClairS-TO. Our questions asked and answers are shown in detail below.

As a quick summary, both ClairS and ClairS-TO have improved performance using a model initially trained with synthetic samples and augmented with real samples. The improvements are especially noticeable in somatic Indel calling. When comparing the variants called using a model trained with real samples only against a model trained initially with synthetic samples and then real samples augmented, we found a few true variants that were missed by the model trained with real samples only. The result suggests that synthetic samples can effectively mitigate the risk of missing real sample of a specific cancer-type in model training. According to our findings, ClairS and ClairS-TO will start to offer two types of models: 1) real samples augmented based on a model trained from synthetic samples, and 2) trained with synthetic samples only. The first type provides better performance in most of the usage scenarios. The second type is free from biases led by the exclusion of a cancer-type in model training.

Some more techniques are also introduced to both ClairS and ClairS-TO to ramp up the performance, including the use of a new public panel of normal (PoN), tagging indels based on sequence entropy, and base-quality (BQ) jittering in model training. The BQ jittering idea was inspired by our observation that the BQ distribution in Park et al. (SRR28305167, HCC1395 peak at 38, avg 33.05) is lower than the ONT offered GIAB datasets (use HG002 for example, peak at 40, avg 36.04). Without BQ jittering, ClairS and ClairS-TO risk from either overconfidence or the contrary when handling sequencing data from different centers that have different BQ distributions.

Results

How could ClairS make the best use of real cancer cell-line datasets?

To answer the question, we need to compare the performance of the following three models: 1) trained from synthetic samples only (SS), 2) trained from real samples only (RS), and 3) initially trained from synthetic samples and augmented with real samples (SS+RS). We followed the steps in Zheng et al. [2] for preparing synthetic samples from GIAB samples and known germline variants. We used the truth variants of HCC1395 from SEQC2 [4], and the truth variants of COLO829 from New York Genome Center [5]. The HCC1395 R10.4.1 ONT dataset is from SRR28305167. The COLO829 R10.4.1 ONT dataset is from Nanopore EPI2ME Labs [6].

In terms of the real samples used for training, different from Park et al. [3], we used only four real sample pairs (HCC1937/BL, HCC1954/BL, H1437/BL, and H2009/BL) for training. To facilitate benchmarks in downstream, we excluded the entire HCC1395, and chr1 of all samples from training, considering that 1) the already small number of known somatic

variants in HCC1395, and what's worse when looking into only chr1 (chr1 has only ~100 indels), could lead to insufficient statistical power and unreliable performance comparisons, 2) HCC1395, HCC1937, and HCC1954 are all lung cancer cell-lines, thus the exclusion of the entire HCC1395 would not remove a cancer-type (lung cancer) from the training samples completely. In the future release of ClairS and ClairS-TO models, we will keep excluding 1) the entire HCC1395, and 2) chr1 of all real samples. Meanwhile, ClairS and ClairS-TO have not included COLO829/BL for training. Algorithms that have included HCC1395 in model training might use COLO829 for benchmarking against ClairS and ClairS-TO. Noteworthy, COLO829 is a melanoma sample, the cancer type of which is not included in the four real samples for training, thus we can use it to conclude how synthetic samples could compensate for model training when a cancer type is missing in the real samples.

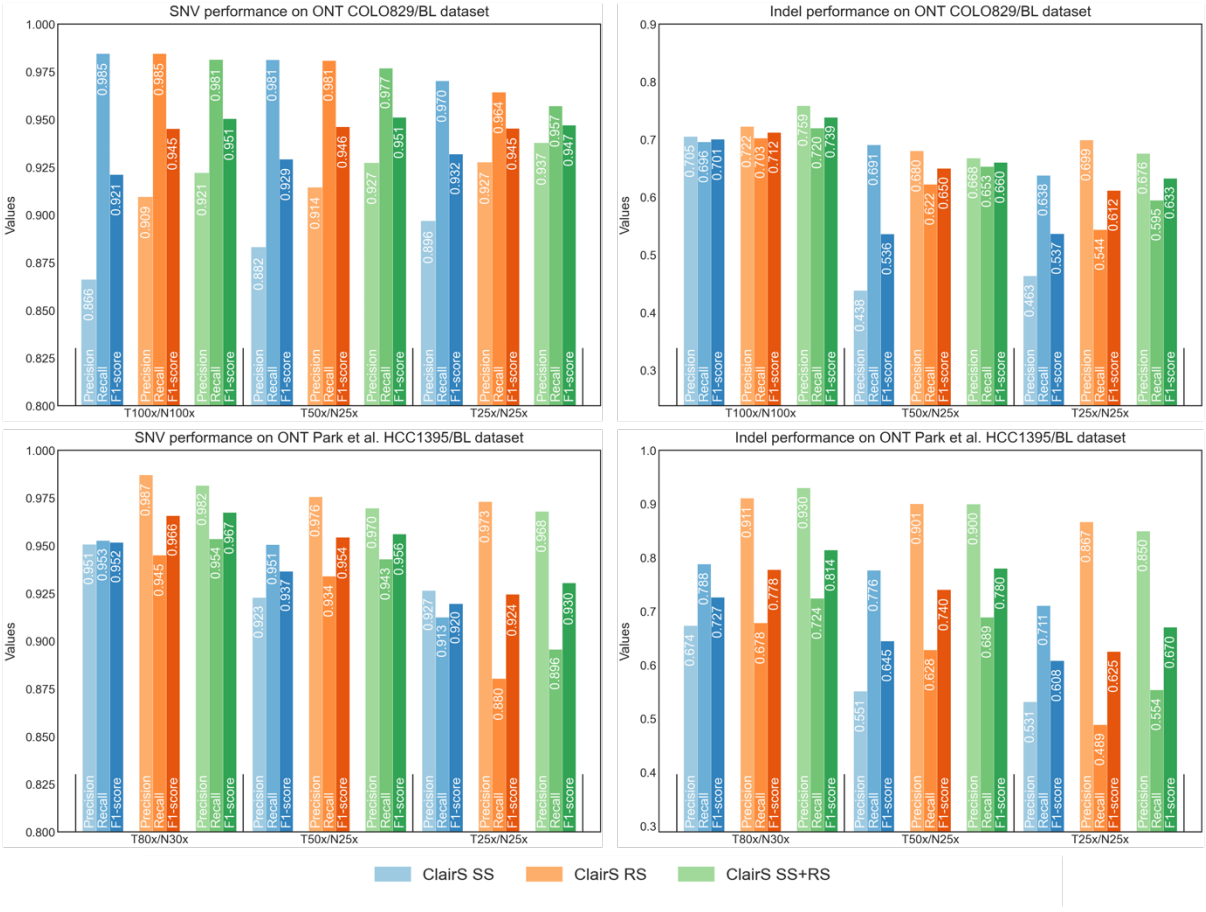

Figure 1. The SNV and Indel performance of ClairS using ONT COLO829/BL and ONT HCC1395/BL for benchmarks. High, medium, and low tumor coverage settings were benchmarked. SS: trained from synthetic samples only; RS: trained from real samples only; SS+RS: initially trained from synthetic samples and augmented with real samples.

The SNV and Indel performance of ClairS using both COLO829/BL and HCC1395/BL at different tumor coverage settings are shown in Figure 1. At all settings in both samples, using the F1-score, the RS model performed better than the SS model, and the SS+RS model performed better than the RS model. It is worth noticing that in COLO829/BL, the precisions are generally higher than the recalls, while the opposite is observed in HCC1395/BL (highlighting the differences between the two cancer samples). Nonetheless, the SS+RS model has always balanced precision and recall better than the SS or RS model. We

conclude that SS+RS is the best way for ClairS to use real cancer cell-line datasets. It is likely that ClairS model has learned the mutational signatures that are shared among between cancer types (for example, the signature 1 in the COSMIC mutational signatures version 2) [7] through the inclusion of real samples in model training, but the extent remains to be further studied.

Could synthetic samples compensate for missing a cancer type in the real samples?

There is a risk of missing a cancer type in the real samples for model training because of the existence of various mutational signatures that were found only in specific cancer types [4]. Missing a cancer type in model training might lead to lower variant calling performance and risk distorting the mutational signatures of the variants called. The risk can be mitigated if the cancer-type agnostic synthetic samples are used together with real samples for model training.

In COLO829/BL, we found 18 true variants that were called by the SS and SS+RS models, but missed by both the RS model and DeepSomatic. Four examples are shown in Figure 2 with an IGV screenshot. We speculate that these variants might contribute to mutational signatures that are absent or rare in the real samples used for training.

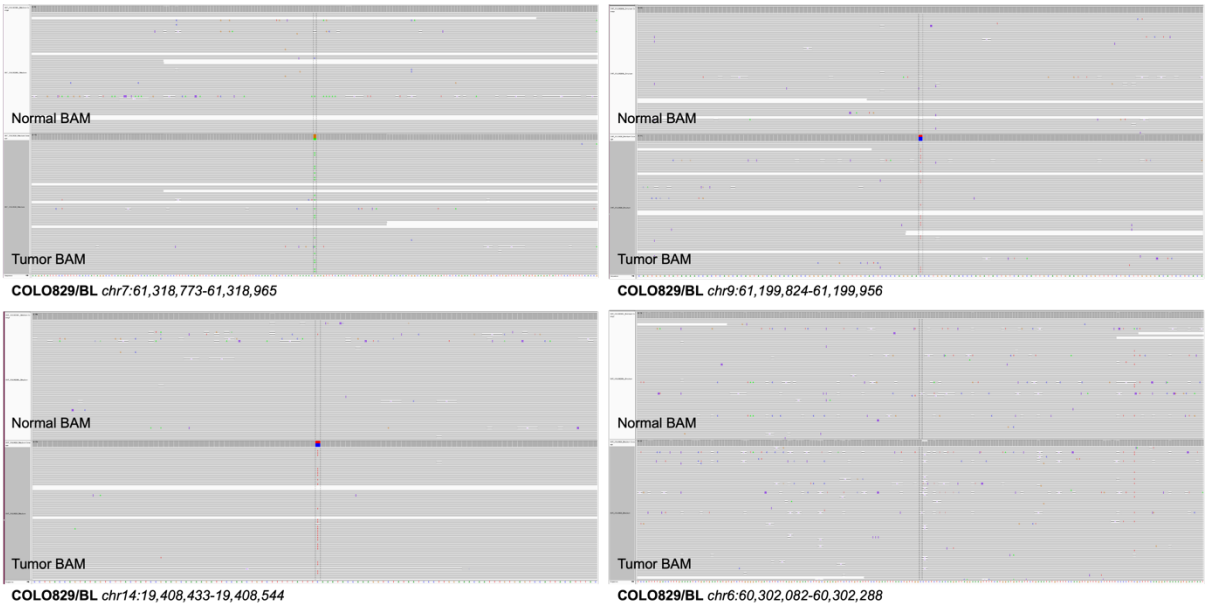

Figure 2. Four examples of variants called by the SS and SS+RS models, but missed by both the RS model and DeepSomatic.

Starting from ClairS v0.4.0 and ClairS-TO v0.3.0, we will provide both the SS model and SS+RS model for the latest chemistry and basecaller combinations. The SS+RS model provides better performance and fits most usage scenarios. SS model can be used when missing a cancer-type in model training is a concern.

BQ jittering in model training to accommodate different BQ distributions in multi-center generated datasets

In the datasets released by Park et al. [3], we observed a BQ distribution lower than the GIAB sample datasets released by ONT [8] (Figure 3a). The BQ distribution in Park et al. (HCC1395 peak at 38, avg 33.05) is lower than the ONT GIAB datasets (use HG002 for example, peak at 40, avg 36.04). The lower BQ distribution has led to an extra amount of input being ignored, which is one of the factors contributing to a lower performance of ClairS and ClairS-TO shown in Park et al.

BQ distribution difference is not uncommon and can be caused by many reasons, including different sample preparation protocols, different labs, different basecaller settings used, etc. A variant caller may underperform if the samples for model training do not have a comprehensive coverage of different BQ distributions.

In the new version of ClairS and ClairS-TO, we introduced BQ jittering to mitigate the problem. BQ jittering works as follows. For each variant generated during synthetic sample generation, we make a duplication. For each duplication, all the base qualities in the duplication are added with *d*, where *d* is a non-zero integer randomly sampled from a Gaussian distribution with mean equals to 0 and variance equals to 5. The original variants and the BQ jittered duplications are all used for model training. BQ jittering only applies to synthetic samples.

The performance of BQ jittering is shown in Figure 3b. When using the SS model (trained from synthetic samples generated using the ONT HG001 and HG002 datasets) as the basis, and Park et al. HCC1395/BL for benchmarking, BQ jittering has offered up to 0.4% F1-score increase in SNV performance, and up to 3.2% increase in Indel performance.

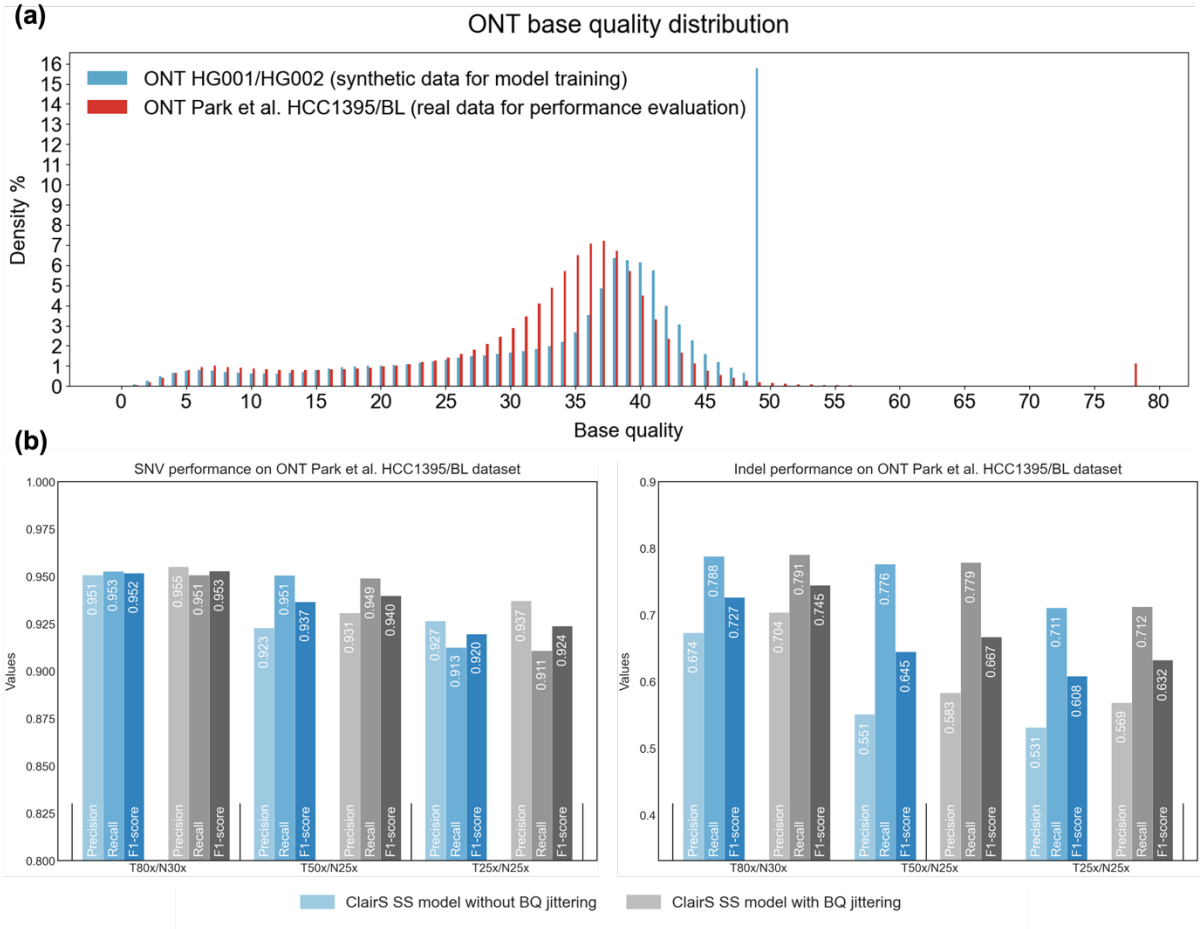

Figure 3. (a) The BQ distribution difference between ONT HG001/HG002 and Park et al. HCC1395/BL datasets. (b) The SNV and Indel performance of ClairS using the SS model as the basis and Park et al. HCC1395/BL for benchmark, with and without BQ jittering for model training.

Will the 'SS+RS model' improve the performance at lower tumor purities or higher levels of normal contamination?

The cancer cell-line samples usually have high tumor purities and low levels of normal contamination, which are not often the case with cancer samples from biopsy or FFPE that are more commonly used for diagnosis. We wondered whether the performance improvements led by using real samples in model training would persist at lower tumor purities or higher levels of normal contamination.

As shown in Figure 4, SS+RS has performed better than the RS model at all tumor purities and levels of normal contamination. While the real samples used for ClairS model training have fixed tumor purity and level of normal contamination, the improvement in SS+RS over RS is likely led by the availability of rich combinations of tumor purity and level of normal contamination in the synthetic samples [2]. Park et al. [3] use an idea similar to the synthetic samples in ClairS to create multiple tumor purities and levels of normal contamination, but is by mixing the real tumor cell-line data with normal cell-line data, while ClairS is mixing normal cell-line data of different samples. The former generates synthetic samples that

have an error profile and mutational signatures closer to real cancer samples, while the latter is much easier to generate more combinations of tumor purity and level of normal contamination.

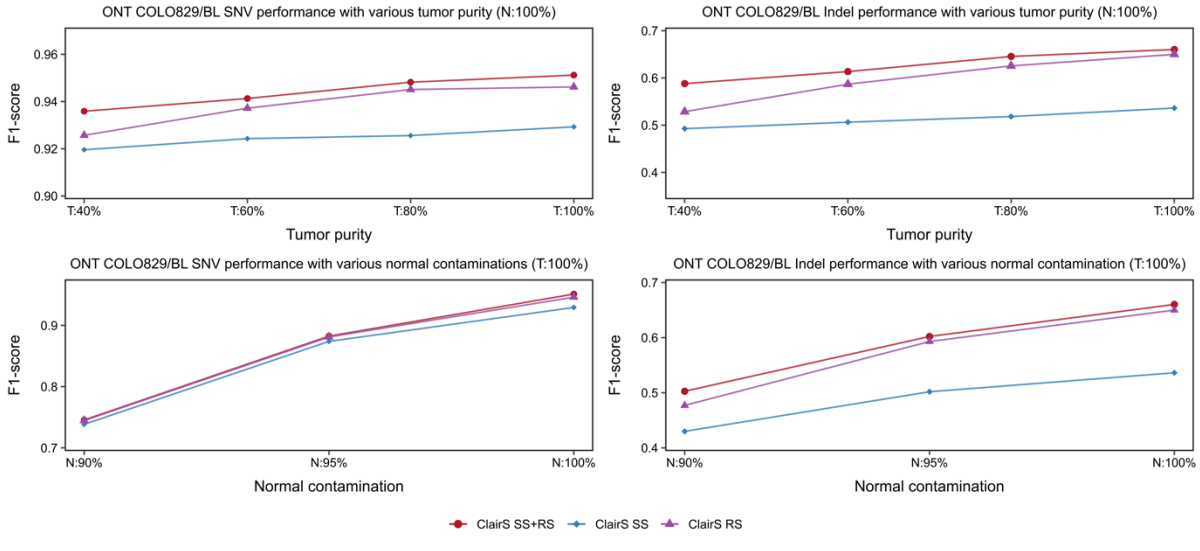

Figure 4. (a) SNV and Indel performance at different tumor purities using 50x/25x of ONT COLO829/BL for benchmarking. (b) SNV and Indel performance at different level of normal contamination using 50x/25x of ONT COLO829/BL for benchmarking.

Compare the performance of ClairS v0.4.0 and DeepSomatic v1.7.0 for ONT somatic variant calling at different AF ranges and coverages

We compared the performance between “ClairS v0.4.0 with the SS model” (ClairS SS), “ClairS v0.4.0 with the SS+RS model” (ClairS SS+RS), and “DeepSomatic v1.7.0” for tumor-normal pair somatic variant calling. The results are shown in Figure 5. Both ClairS and DeepSomatic have excluded COLO829/BL in their model training, thus, we have used all chromosomes in COLO829/BL for benchmarking, which includes 42,991 SNVs and 984 Indels. Overall, ClairS SS+RS outperformed DeepSomatic, and DeepSomatic outperformed ClairS SS. At high, medium and low coverages (T100x/N100x, T50x/25x, T25x/N25x), the SNV F1-score differences between ClairS SS+RS and DeepSomatic are 0.024, 0.026, and 0.008, the Indel F1-score differences are 0.047, 0.015, and 0.004. At five different AF ranges from low to high (0.05-0.15, 0.15-0.3, 0.3-0.45, 0.45-0.6, 0.6-1.0), the SNV F1-score differences between ClairS SS+RS and DeepSomatic are 0.039, 0.005, -0.001, -0.001, -0.007, the Indel F1-score differences are 0.081, 0.079, 0.040, 0.043, -0.020.

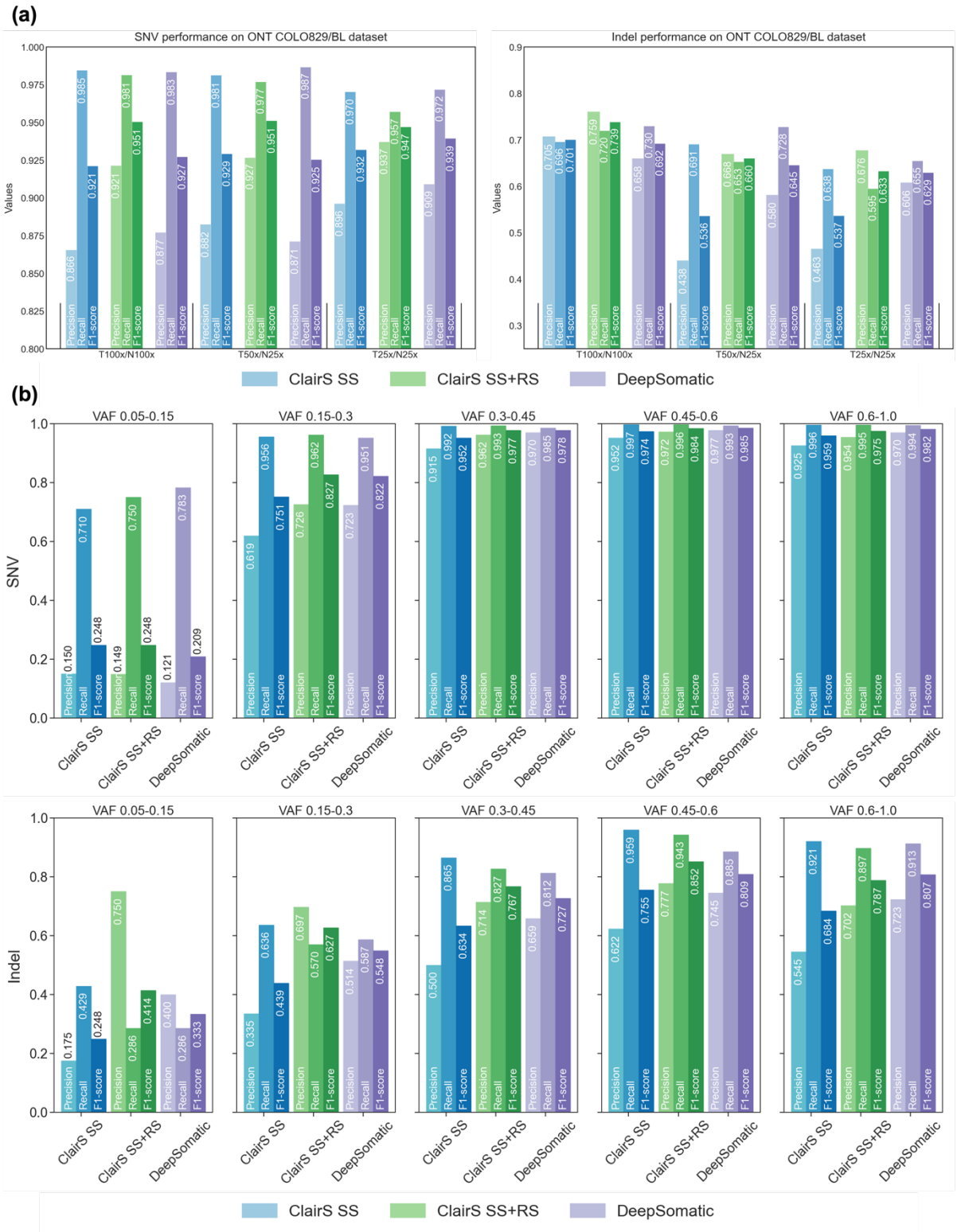

Figure 5. Performance comparison between "ClairS v0.4.0 with the SS model", "ClairS v0.4.0 with the SS+RS model", and "DeepSomatic v1.7.0", at (a) different coverages, and (b) at different AF ranges for SNV and Indel, respectively.

Improve ClairS-TO performance by "tagging indels at sequence with low entropy" and "using CoLoRSdb as PoN"

In ClairS-TO v0.3.0, we applied two new techniques to improve performance. The results are shown in Figure 6.

The first technique is to tag indels called at sequence with low entropy. For each Indel candidate, the entropy of the flanking 32bp in the reference genome is calculated using 5-mer. If the entropy is lower than a threshold (default at 0.9), an Indel is tagged as "LowSeqEntropy". Excluding the LowSeqEntropy variants increased the F1-score by 0.060 and 0.222, in ONT 50x COLO829 and ONT 50x HCC1395, respectively.

The second technique is the use of CoLoRSdb (Consortium of Long Read Sequencing Database) [9] as PoN for non-somatic variant tagging. The idea was inspired by Park et al., who have used CoLoRSdb as a PoN in DeepSomatic [3]. Variants with AF>=0.001 in CoLoRSdb are chosen, and we required only positional matching between a variant candidate and PoN (the same as how 1000G PoN was used, but different from gnomAD and dbSNP in which allele matching is required). As shown in Figure 6, the use of CoLoRSdb has led to 0.217 (SNV) and 0.075 (Indel) F1-score increase when using ONT 50x COLO829, as well as 0.128 (SNV) and 0.112 (Indel) increase when using ONT 50x HCC1395, for tumor-only somatic variant calling.

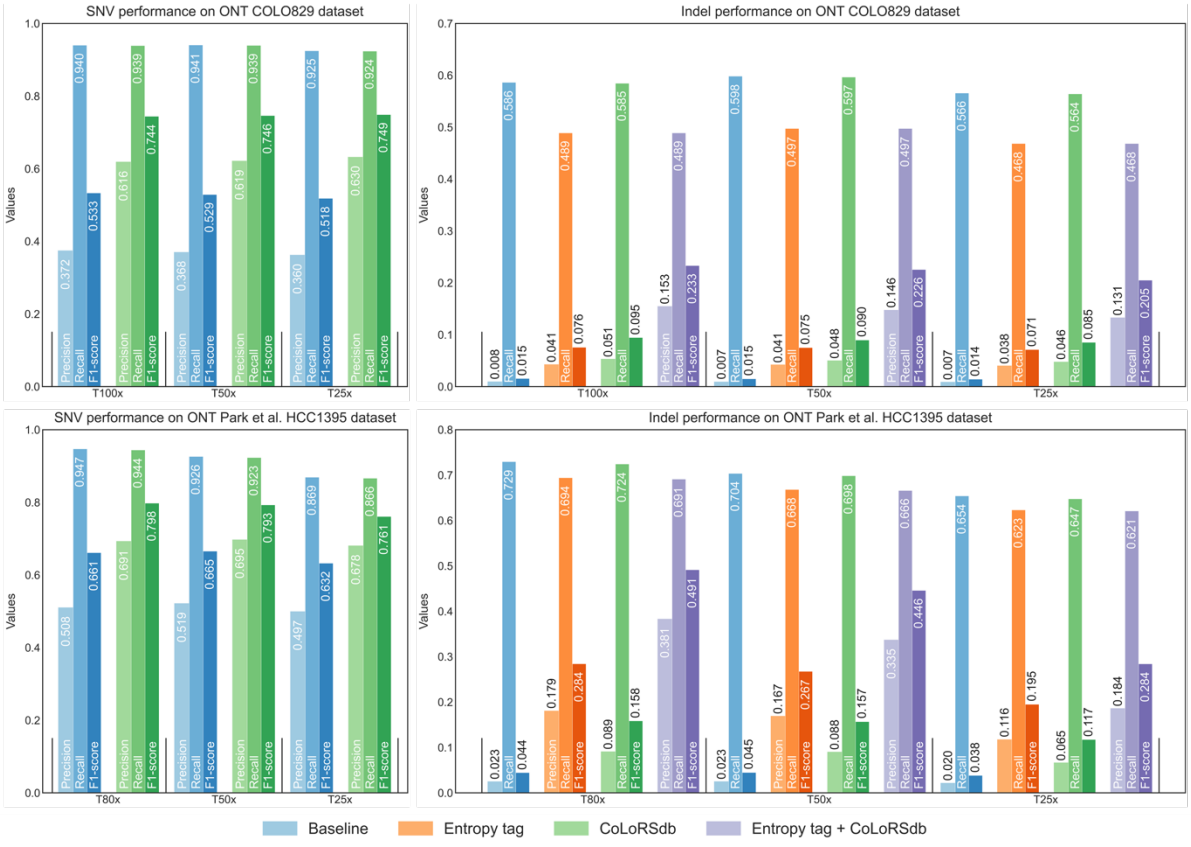

Figure 6. The SNV and Indel performance of ClairS-TO at baseline and after enabling "tagging indels at sequence with low entropy" and/or "using CoLoRSdb as PoN" enabled. High, medium, and low tumor coverage of either COLO829 or HCC1395 were used for benchmarking.

How would ClairS-TO benefit from using real samples for model training?

Similar to the way we explored how ClairS could make use of the real samples to improve performance, we also explored how three models perform differently in ClairS-TO, including models 1) trained from synthetic samples only (SS), 2) trained from real samples only (RS), and 3) initially trained from synthetic samples and augmented with real samples (SS+RS). Both new techniques “tagging indels at sequence with low entropy” and “using CoLoRSdb as PoN” are enabled in this experiment. As shown in Figure 7, the performance of the SS+RS model is unanimously better than the SS and RS models. However, the overall improvement is not as significant as observed in ClairS. The improvement led by using real samples for model training is more observable when calling Indels at lower coverage.

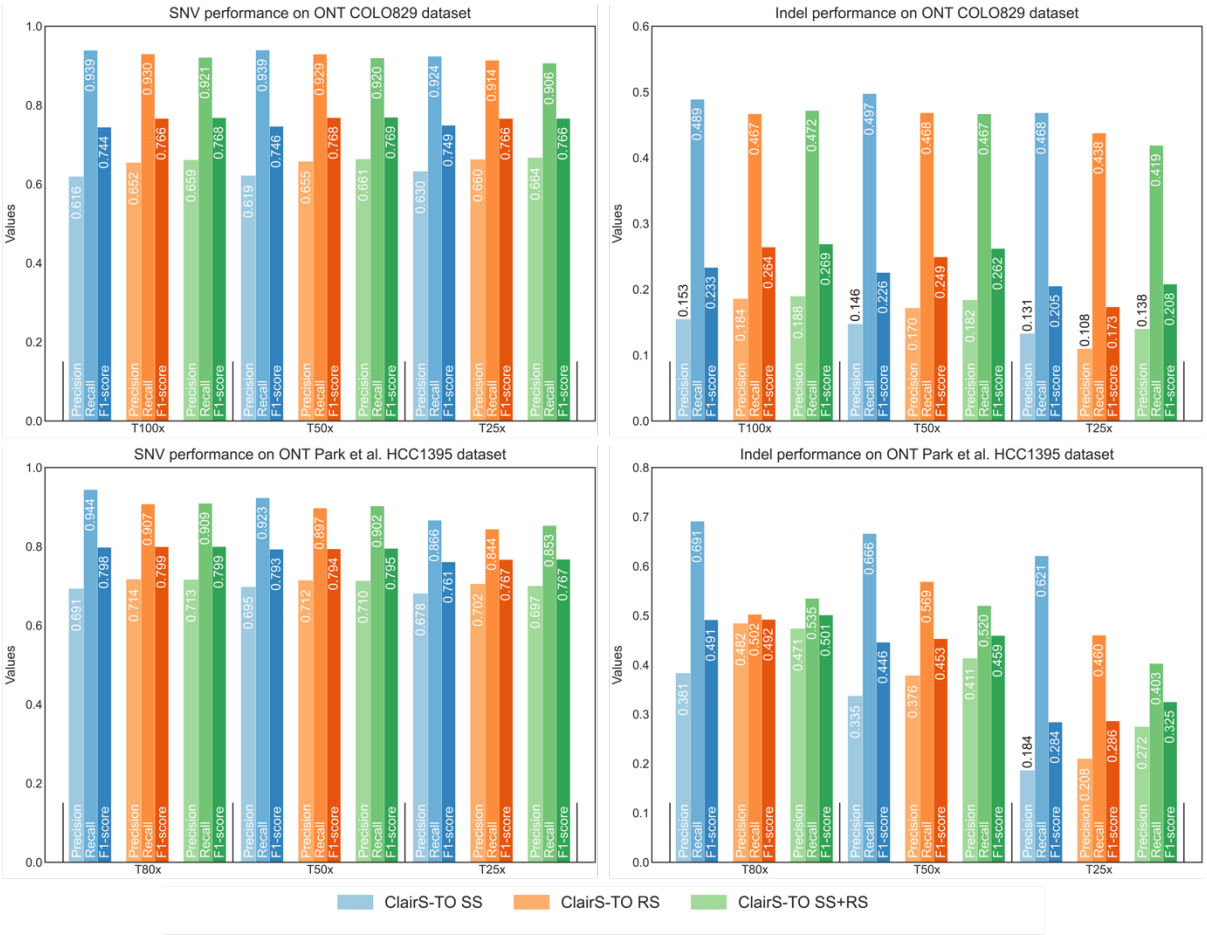

Figure 7. The SNV and Indel performance of ClairS-TO using ONT COLO829 and ONT HCC1395 for benchmarks. High, medium, and low coverages were benchmarked. SS: trained from synthetic samples only; RS: trained from real samples only; SS+RS: initially trained from synthetic samples and augmented with real samples.

Compare the performance of ClairS-TO v0.3.0 and DeepSomatic v1.7.0 for ONT tumor-only somatic variant calling at different AF ranges and coverages

We compared the performance between “ClairS-TO v0.3.0 with the SS model” (ClairS-TO SS), “ClairS-TO v0.3.0 with the SS+RS model” (ClairS-TO SS+RS), and “DeepSomatic v1.7.0” for tumor-only somatic variant calling. The results are shown in Figure 8. Similar to the benchmarks between ClairS and DeepSomatic, we used all chromosomes in COLO829 for benchmarking (including 42,991 SNVs and 984 Indels). Overall, both ClairS-TO SS and ClairS-TO SS+RS have outperformed DeepSomatic. At high, medium and low coverages (T100x, T50x, T25x), the SNV F1-score differences between ClairS-TO SS+RS and DeepSomatic are 0.154, 0.162, and 0.162, the Indel F1-score differences are 0.190, 0.190, and 0.152. At five different AF ranges from low to high (0.05-0.15, 0.15-0.3, 0.3-0.45, 0.45-0.6, 0.6-1.0), the SNV F1-score differences between ClairS-TO SS+RS and DeepSomatic are 0.283, 0.267, 0.099, 0.060, 0.082, the Indel F1-score differences are 0.002, 0.267, 0.189, 0.184, 0.145.

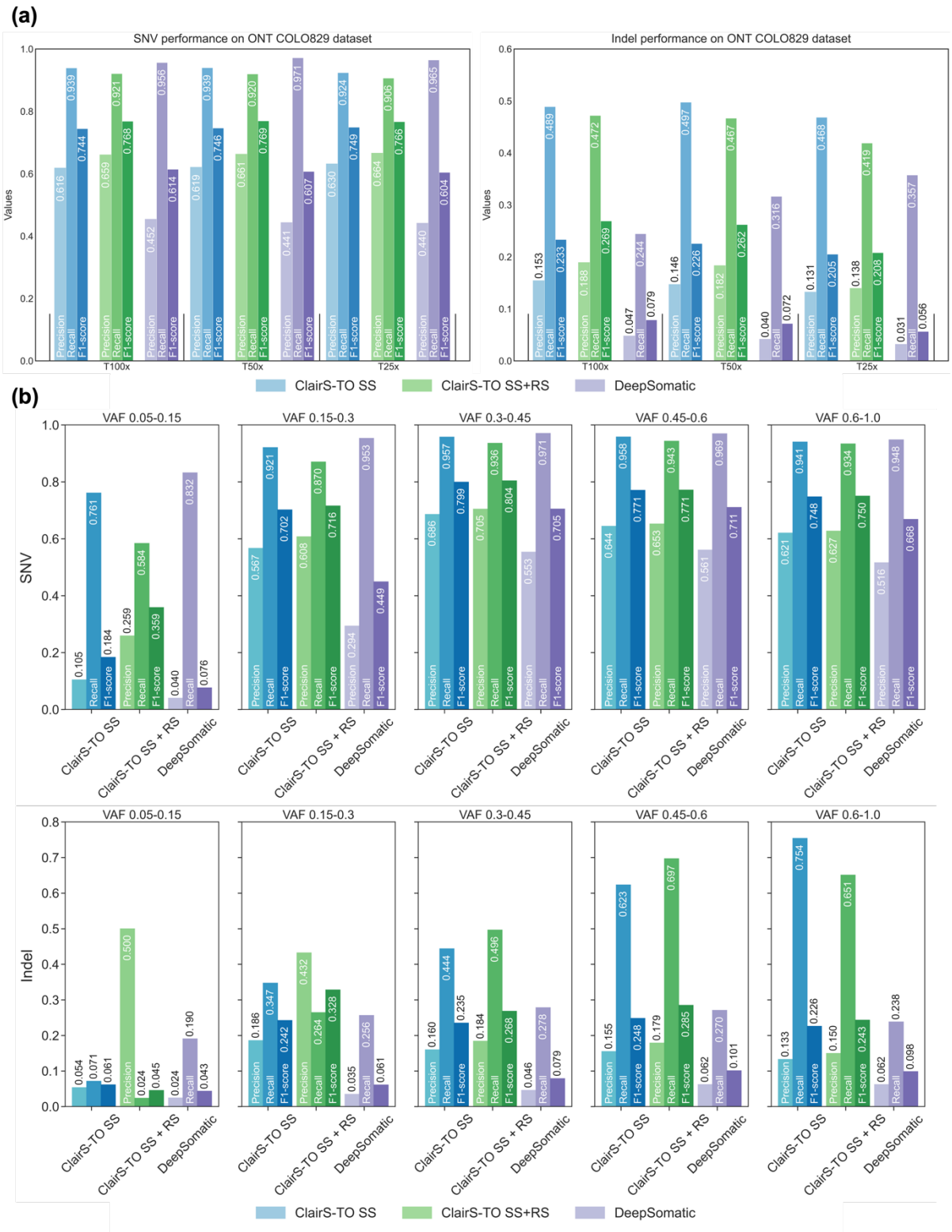

Figure 8. Performance comparison between “ClairS-TO v0.3.0 with the SS model”, “ClairS-TO v0.3.0 with the SS+RS model”, and “DeepSomatic v1.7.0”, at (a) different coverages, and (b) at different AF ranges for SNV and Indel, respectively.

References

[1] https://github.com/HKU-BAL/ClairS?tab=readme-ov-file#latest-updates
[2] Zheng Z X, Su J, Chen L, et al. ClairS: a deep-learning method for long-read somatic small variant calling[J]. bioRxiv, 2023: 2023.08. 17.553778.
[3] Park J, Cook D E, Chang P C, et al. DeepSomatic: Accurate somatic small variant discovery for multiple sequencing technologies[J]. bioRxiv, 2024.
[4] Fang L T, Zhu B, Zhao Y, et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing[J]. Nature biotechnology, 2021, 39(9): 1151-1160.
[5] Arora K, Shah M, Johnson M, et al. Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms[J]. Scientific reports, 2019, 9(1): 19123.
[6] Nanopore EPI2ME Labs, https://labs.epi2me.io/colo-2024.03, 2024.
[7] https://cancer.sanger.ac.uk/signatures/signatures_v2/
[8] Nanopore EPI2ME Labs, https://labs.epi2me.io/giab-2023.05, 2023.
[9] https://zenodo.org/records/13145123

Supplementary materials

Supp Table 1. The number of somatic/germline/artifact variants derived from four real cancer samples H1437, HCC1937, HCC1954, and H2009, for the RS and SS+RS model training.

Variant type	Category/ Sample	H1437	HCC1937	HCC1954	H2009	Total
SNV	Somatic	205,198	41,429	54,396	350,641	651,664
	Germline	784,043	207,145	271,980	1,753,205	3,016,373
	Artifact	734,213	1,299,395	736,714	739,084	3,509,406
Indel	Somatic	11,647	2,622	8,085	13,505	35,859
	Germline	58,235	13,110	40,425	67,525	179,295
	Artifact	3,319,681	3,647,502	3,350,636	3,323,585	13,641,404