# Somatic mutation detection workflow validity distinctly influences clinical decision

Pei-Miao Chien[1,2], Chinyi Cheng[1], Tzu-Hang Yuan[2,3], Yu-Bin Wang[2], Pei-Lung Chen[1,3,5], Chien-Yu Chen[3,4], Jia-Hsin Huang[2], Jacob Shujui Hsu[1,3]

[1]Graduate Institute of Medical Genomics and Proteomics, College of Medicine, National Taiwan University, Taipei, Taiwan
[2]Tawain AI Labs, Taipei, Taiwan
[3]Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan
[4]Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan
[5]Department of Biomechatronics Engineering, National Taiwan University, Taipei, Taiwan

## Abstract

Identifying somatic mutations from tumor tissues holds substantial clinical consequences for making informed medical decisions. Evaluating the accuracy and robustness of somatic mutation analysis workflows has become essential when employing whole exome sequencing (WES) analysis in clinical settings. In the study, we utilized a set of tumor WES data the Sequencing and Quality Control Phase 2 (SEQC2) project to systematically benchmark the workflow analytical validity, including various combinations of read aligners and mutation callers. The read aligners included BWA; Bowtie2; built-in DRAGEN-Aligner; DRAGMAP; and HISAT2 as well as the callers Mutect2; TNscope; built-in DRAGEN-Caller; and DeepVariant. Among all combinations, DRAGEN showed the best performance with mean F1-score of 0.9659 in SNV detection, while the combination of BWA and Mutect2 showed the second highest mean F1-score of 0.9485. Notably, our results suggested that the mutation callers had a significantly higher impact on the overall sensitivity than the aligners. For drug-related biomarkers, Sentieon TNscope tended to underestimate tumor mutation burden and missed many drug-resistance mutations such as FLT3(c.G1879A:p.A627T) and MAP2K1(c.G199A:p.D67N). Our investigation provides a valuable guide for cancer genomic researchers on tumor mutation identification, accomplished through an in-depth performance comparison among diverse tool combinations.

## 1 Introduction

Somatic mutations are pivotal in understanding tumor cell development and provide molecular information for clinical oncology. Many cancer treatments, such as targeted therapies, heavily rely on pinpointing specific somatic mutations, such as BRAF, VEGF, EGFR, and BRCA mutations, across different types of cancers. Moreover, many somatic mutations, such as EGFR, FLT3, and MAP2K1 genes, are associated with drug resistance, indicating the importance of mutation detection capability. The conflicting results can dramatically affect the downstream clinical decision. Somatic mutations can also be detected in the peripheral blood as the tumor DNA will be released through the circulating system as Circulating tumor DNA (ctDNA). More and more studies have shown that the somatic mutations surveillance in circulating tumor DNA (ctDNA) can indicate the in-situ tumor status. Identifying clinical informative mutations in ctDNA relies more heavily on robust analytical workflow. However, compared to the mutations in the tumor origins, the variant allele frequency (VAF) of the mutation in ctDNA is mixed and often very low, reflecting the difficulty of detecting the mutation signal from the background noise. Next-generation sequencing holds the promise of detecting mutation with low VAF by increasing the sequence depth. However, because of the technical limitations, short read alignment can introduce systemic errors. In the difficult-to-map or repetitive genomic regions, read mapping can be ambiguous and insertions/deletions (INDELs) can be challenging to find the accurate positions (Betschart *et al.* 2022, Olson *et al*. 2022, Wagner *et al.* 2022). Despite both aligners and callers with considerable methods available, the detecting ability and the reproducibility remain challenging. Therefore, a systematic validation of analytic workflows is inevitable.

Nevertheless, the lack of a standardizing biological specimen often leads to inconclusive information in the field. Unlike the germline genomic variation based on several individual samples in the genome in a bottle (GIAB) consortium, the benchmarking materials for somatic mutation detection often rely on synthetic or simulated datasets. In contrast to germline variant detection, the caller must be sensitive to somatic variants, which often have relatively low allele frequencies. Until recently, the FDA-led Sequencing and Quality Control Phase 2 (SEQC2) consortium has released a benchmarking dataset for cancer variant detection in tumor-only mode (Jones *et al.* 2021). It can enhance the data quality and provide a golden standard for further benchmarking investigation.

Consequently, the evaluation of different somatic variant detection tools has become feasible. In this study, we focused on the Comparison of different combinations of aligners and callers. Many reputable aligners and callers were selected to evaluate the performances. The aligners of BWA (Li *et al.* 2021), Bowtie2 (Langmead *et al.*

2012), an Illumina DRAGMAP (https://github.com/Illumina/DRAGMAP), the build in aligner of the Illumina commercialize DRAGEN Bio-IT Platform (abbreviated as DRAGEN-A) (https://www.illumina.com/products/by-type/informatics-products/dragen-bioit-platform.html), and a graph-based aligner HISAT2 (Kim *et al.* 2019) were selected. The callers for the cancer variant include GATK Mutect2 (Cibulskis *et al.* 2013), Sentieon TNscope (Kendig *et al.* 2019), Illumina DRAGEN Bio-IT Platform build-in caller (abbreviated as DRAGEN-C), and Google DeepVariant (Poplin *et al.* 2018) were selected. Detailed information of all aligners and callers is given in Supplementary Table 2. We reanalyzed the SEQC2 benchmarking dataset and compared 18 combinations of aligners and callers for their performances on tumor-only mode in cancer single nucleotide variants (SNVs) and small insertions/deletions (INDELs) detection. Our results provide critical information and evidence that no single combination produced superior results in detecting all the variants from the samples. In turn, the merged or ensemble-based approach, which combines results from multiple combinations, is most likely to improve the performance of variant detection in tumor samples.

Recent studies (Dietlein *et al.* 2022, Pleasance *et al.* 2022) have broadly demonstrated that using WES or WGS data can lead to precise cancer treatment. However, previous studies have shown that the aligner and the variant caller could heavily influence the results of variant detection (O'Rawe *et al.* 2013, Pabinger *et al.* 2014). Additionally, several studies indicated that different tool combinations often generate discordant variant detection results (Cornish *et al.* 2015, Kumaran *et al.* 2019, Liang *et al.* 2019, Liu *et al.* 2022, Wang *et al.* 2019, Zhao *et al.* 2020).

## 2 Material and methods

### 2.1 Benchmark

The benchmarking dataset was downloaded from SEQC2 released in 2021 (SRA accession: SRR13076390 to SRR13076398) and renamed as S1-S9 in this study. The dataset consists of nine sets of paired-end WES reads from three different exome enrichment kits (Agilent, Roche, IDT), each with three technical replicates conducted by three different laboratories. Nine sample details are listed in Supplementary Table 1.

Benchmark and confidence regions were obtained from Jones *et al* 2021. For the truth-positive variants, there are 41,395 SNVs and 537 INDELs in the confidence regions, 52.17% of which have variant allele frequency (VAF) lower than 0.2.

### 2.2 Aligner and variant caller combination implementations

In this study, the whole analytic workflow was illustrated in Supplementary Fig 1, and the tools version were listed in Supplementary Table 2. We adopted five aligners and four callers comprising up to 20 different combinations. The first phase involved processing the raw reads to produce analysis-ready BAM files. This included read alignment against GATK human_g1k_v37_decoy.fasta reference genome followed by read deduplication. Forty-five (five aligners × nine samples) text-based Sequence Alignment Map (SAM) files were obtained and converted into the analysis-ready binary version format files (BAM). The second phase was a variant discovery from analysis-ready BAM files and left-normalized by bcftools in VCF format. However, HISAT2_DRAGEN-C and HISAT2_DeepVariant encountered data format problems, resulting in their exclusion from the analysis. As a result, we evaluated 18 different combinations in total.

### 2.3 Performance evaluation

We used GATK SelectVariants to filter out the variants that were not in confidence regions and compared with the benchmark by using Haplotype Comparison Tools (https://github.com/Illumina/hap.py) and in-house developed programs to evaluate the performances of the combinations. The performance of each tool combination was evaluated by comparing the VCF file against the SEQC2 truth set benchmark VCF file and calculating the F1 score, recall, precision.

### 2.4 Computing environment and resources

All analyses were executed by the high-performance computing (HPC) cluster computing environment in Taiwania 3 at the National Center for High-performance Computing, Taiwan. Custom-developed programs were employed to evaluate the performance of each combination. In the case of DeepVariant, the Docker image was obtained from https://github.com/google/deepvariant and installed on a Google Kubernetes Engine (GKE)-based HPC cluster computing environment in collaboration with Taiwan AI Labs.

### 2.5 Analysis of COSMIC gene census and cancer-related biomarkers

We obtained curated cancer gene census (N=738) and drug-associated genes (N=15) from the Catalogue Of Somatic Mutations In Cancer (COSMICv3.3). The SEQC2 truth set contains 512 genes (512/738=69.3%). Also, we analyzed the tumor mutational burden (TMB) and mutational signature. Tumor mutational burden is calculated by

$$TMB = \frac{Nonsynonymous\ SNV\ counts}{Exon\ region\ (22{,}519{,}561bp)}$$

Exon regions are defined by SEQC2 which contains 22,519.561 base regions .

Mutational Signature analyses were done by SigProfilerExtractor v1.1.4 (https://github.com/AlexandrovLab/SigProfilerExtractor), using the latest COSMIC v3.3 signature database.

## 3 Results

### 3.1 Comparison of different WES library preparation kits

We evaluated the average mapping depth of nine samples to analyze the deduplicated alignments by Picard (2.27.4). Regardless of the aligners, we observed distinct patterns across library preparation methods. Samples with the IDT WES enrichment kit exhibited lower mapping depth and higher duplication rates (Supplementary Table 3), whereas those treated with the Agilent showed higher mapping depth.

This indicates that the choice of aligner had limited impact on the observed mapping depth. Overall, these findings emphasize the significance of library preparation methods and sample quality in influencing the mapping depth and duplication rates, while suggesting that the choice of aligner may have moderated influences.

### 3.2 Comparison of the overall SNV and INDEL calling abilities

To obtain an overview of the performance of all variant calling and aligning combinations, we focused on the F1-score of 18 different combinations. The results were separated into SNVs (Figure 1A, Supplementary Figure 3-4, Table 4) and INDELs (Figure 1B, Supplementary Figure 3-4, Table 5). For the SNV detection, the built-in DRAGEN, the combination of DRAGEN-A and DRAGEN-C, had the best performance, achieving a mean F1-score of 0.966, a mean recall rate of 0.954 and a mean precision of 0.979. In general, each sample shown similar pattern among 14 analytic workflows, suggesting the importance of the library preparation step. Among all free resource combinations, BWA_Mutect2 has the best performance in SNV, with mean recall rate 0.953, mean precision 0.944 and mean F1-score 0.949. We observed that sample 3 from the laboratory 3 shown the highest F1 scores followed by sample 2 from the laboratory 2 and sample 1 from the laboratory 1, indicating the variant call inconsistencies introduced by the laboratories. Among all combinations, the sample 3 was always the best while sample 4 was always the worst one. Overall, Agilent WES library preparation kit performed higher F1 scores among other kits. However, even though the library preparation kit is identical, the F1 scores can be varied, indicating

the importance of benchmarking works for the variant calling analyses. It is worth to note that all of the combinations with DeepVariant as the variant caller had the relatively poor performance with F1/Recall/Precision of 0.4/0.3/0.6 in SNV and 0.3/0.2/0.5 in INDEL. This indicates that DeepVariant is not suitable for somatic variant calling. Therefore, we excluded DeepVariant for any further comparisons.

For the small insertion/deletion (INDEL) identification, BWA_DRAGEN-C had the best performance, slightly better than the built-in DRAGEN, with a mean F1-score of 0.792, recall rate of 0.954 and precision of 0.678. Among all open-source combinations, HISAT2_Mutect2 combinations had highest mean F1 score 0.72 with mean recall 0.88 and mean precision 0.62. Also, DRAMAP_Mutect2 had highest mean recall rate at 0.95 with mean precision 0.53 and mean F1 score 0.68, indicating HISAT2_Mutect2. Unlike the patterns in SNV detection, we observed that the commercial variant callers tended to perform better as DRAGEN-C was the best among others followed by TNscope. The sample 3 was not always the best dataset in INDEL detection. Agilent WES library preparation kits had better performance for DRAGEN-C, and IDT WES kits had better performance for other callers. Still, even with the identical library preparation kits, we can still observe varied performances among laboratories.

Furthermore, we found that analysis tool combination selection influences the variant calling results as suitable analysis tools may compensate low mapping depth sample accuracy. For example, IDT WES kits sample 5 and 6 had relatively low depth of coverage, but the overall SNV F1 scores in DRAGEN were higher than many other sample results. It is even outperforming all sample results in TNscope groups. Also, even sample 4 had the lowest depth of coverage, but the INDEL calling performance can outperform other high depth of coverage samples. In Mutect2 calling INDEL group, sample 4 is the best among other samples.

We also dissected the variant calls among aligners and callers. We calculated the variant call for same aligner with different callers and vice versa. Our analysis revealed that the variant callers played a more critical role in the overall process compared to the aligners. Specifically, when using the same caller, approximately 94% of the variants were observed in the intersection of five aligners. However, when using the same aligner, only about 40% of the variants were presented in the intersection of four callers (Supplementary Fig5-8). Of all the callers analyzed, Mutect2 had the highest intersection with DRAGEN-C, indicating that its results were more similar to those produced by DRAGEN-C.

### 3.3 Combining different combinations can improve overall detection capability

In order to achieve higher F1 score, we systematically combined the results from various combinations configurations, and identify intersecting variants within the sample 3 (SRR13076392) sample. Our approach involves the retention of free-resource tools and the exclusion of DeepVariant, enabling a direct comparison with DRAGEN. Performance evaluations are conducted using the F1-score. Our findings underscore the value of combining multiple combination results, demonstrating enhanced overall performance compared to the BWA_Mutect2 for SNV and HISAT2_Mutect2 for INDEL. In the context of SNV detection, combining Mutect2 with different aligners leads to slight improvements in F1-scores, aligning with conclusions drawn in previous sections. However, DRAGEN as a single tool is still the best among others. However, in the context of INDEL detection, the combination results exhibit improved precision and achieve superior F1-scores compared to DRAGEN.

### 3.4 The limitation of detecting the lowest variant allele frequency (VAF)

To investigate the performance of different variant calling tools in detecting somatic variants across various variant allele frequency (VAF) regions, we divide VAF range into 100 bins. In known set, most of the variants distribute in VAF < 0.2 region. Using best quality sample SRR13076392 to exclude the concern of data quality problem, all three variant calling tools DRAGEN-C, Mutect2, TNscope share same VAF detect limitation of 0.0063 while the lowest VAF in known set is 0.0058.

Most of the mutation callers' detection were limited by the variant allele frequency (VAF) at 0.0063 whereas the lowest VAF is at 0.0058. Overall, DeepVariant is particularly not suitable for somatic mutation analysis.

Subsequently, we conducted a comparison of true positives, false negatives, and false positives across different VAF regions(Fig 3.). DRAGEN-C demonstrated a high sensitivity in detecting the majority of variants, albeit with a small proportion of extremely low VAF variants going undetected. Conversely, the BWA_Mutect2 combination exhibited a tendency to miss variants across all VAF ranges. DRAMAP_Tnscope, on the other hand, displayed a higher rate of false negatives, particularly in detecting low VAF variants, with more than half of the variants remaining undetected. These findings underscore the distinct capabilities of each variant calling tool in detecting somatic variants within specific VAF ranges. Regarding false positives (FP), our observations indicate that all variant calling tools produce false positive calls across the entire range of variant allele frequencies (VAF), with a higher concentration of false positives in the low VAF regions.

### 3.5 False negative detection

Combining all the 9 samples all combinations results, it can found 42195 out of 42212 known positive variants.

Combining the results of nine samples and all combinations, it is possible to identify 42,195 correct variants from a total of 42,212 known set correct variants.

### 3.6 Impact of different variant calling results in TMB

Tumor Mutational Burden (TMB) plays a crucial role in clinical use due to its association with tumor biology and response to immunotherapy. TMB refers to the total number of somatic mutations present in a tumor genome, typically measured as the number of mutations per megabase of the sequenced DNA.

In the TMB assay, we observed different tendencies among different tools and the impact of different sample processing methods on the final result interpretation (Supplementary Fig9.).

The result points out that Mutect2 tends to overestimate TMB, particularly in samples processed by Roche WES enrichment kit. On the other hand, DRAGEN-C results are closer to the truth set, while TNscope tends to underestimate TMB.

These observations highlight the importance of selecting appropriate tools and sample processing methods in TMB assays. Choosing the right tools can aid in obtaining more accurate and reliable TMB results, thereby improving the Accuracy of clinical diagnosis and treatment decisions.

**Impact of different variant calling results in Cancer genes**

To investigate the performance of different variant callers in cancer-related genes, we obtained a gene list from the Catalogue Of Somatic Mutations In Cancer (COSMIC) curated cancer gene database. This gene list serves as a reference for our analysis and comparison of variant callers in the context of cancer-related genetic variation. Out of the curated gene list comprising a total of 736 genes, only 380 genes are found to be in consensus with the known set, excluding synonymous SNVs, non-frameshift deletions, and non-frameshift insertions. We excluded variants that do not have an impact on protein coding. Then we filter out genes that only found by two third results and calculate each samples found variant number divided by the number of known set variants for each gene. 23 genes are left after the filter and result are shown in Fig4. If all different pipeline results fail to identify variants in samples with the same treatment, it is considered a sample-related issue. This suggests that there may be a lack of mapped reads in the gene region, resulting in no variant calls within that gene. DeepVariant

can't successfully found these genes variants at all which makes it more clearly that DeepVariant is not suitable for somatic variant calling and significantly influences the outcome. In the comparison between DRAGEN-C, Mutect2, and TNscope, TNscope tends to exhibit a higher rate of variant loss in its variant calling. On the other hand, different sample treatments also have an impact on the results. Specifically, sample SRR13076393, SRR13076394 and SRR13076395, which were treated by IDT, show a higher rate of false negatives compared to the results from Roche or Agilent treatments.

**Impact of different variant calling results in Drug associated genes**

We also obtained a gene list from the COSMIC curated drug-associated gene database. These genes may influence the decision in clinical used and personal medical treatment decision. This gene list obtains 15 genes, ABL1 and BTK aren't exist in the Known set and in Known set variants in JAK1 JAK2 SMO are all synonymous SNV, hence these 5 genes are not put into Comparison. In the result (Fig 5.), DeepVariant loss over half variants.

In the context of variants, our focus lies on those that have been documented to be associated with drug resistance or play a role in treatment selection. These specific variants have garnered significant attention due to their implications for personalized medicine and optimizing therapeutic strategies for patients. Among all variants, Fms-like tyrosine kinase 3, FLT3(c.G1879A:p.A627T) which is reported that patients with this mutation may exhibit increased resistance to conventional chemotherapy agents, such as cytarabine and anthracyclines and alternative treatment is needed for them. However, DeepVariant and TNscope tends to loss this important variant. In same treatment sample, DRAGEN-C and Mutect2 are able to correctly detect this variant. In another case, Mitogen-Activated Protein Kinase Kinase 1, MAP2K1(c.G199A:p.D67N) which is a key player in the mitogen-activated protein kinase (MAPK) signaling pathway, which regulates cell growth, proliferation, and survival. Variants in the MAP2K1 gene have been identified in various cancers, including melanoma, colorectal cancer, and lung cancer. The mutation has garnered attention due to its potential impact on treatment selection and drug response. Studies have shown that patients harboring this mutation may exhibit resistance to certain targeted therapies, such as BRAF inhibitors, that are commonly used in the treatment of melanoma. This variant has been associated with activation of alternative signaling pathways that bypass the MAPK pathway, rendering the tumor cells less responsive to BRAF-targeted treatments. All samples using DeepVariant as caller loss this variant and 36out of 45 of sample of using TNscope caller failed to detect this variant.

**Impact of different variant calling results in mutational signature**

In overall mutational signature single base substitution(SBS) analysis, all results despite the sample treatment or used pipeline share same results as known set result, except those using DeepVariant as caller(Table 2). Those results with using DeepVariant as caller loss SBS6 signature which is associated with defective DNA mismatch repair and is found in microsatellite unstable tumors. Also in double base signature (DBS), DeepVariant will miss DBS11 signature.

## Discussion

In this study, we evaluated the performances of 18 combinations of aligner and variant callers for somatic variants detection by using the benchmarking dataset released by SEQC2. A major difference with other previous benchmarking studies was that we focused on the usage of tumor-only model for variant detection. In fact, because for the clinical applications, the tumor and normal paired samples were always difficult to be obtained. As the results shown in…different practices of wet lab techniques may influence the results of variant detection, even though the same pipeline was utilized. Some callers performed the poor results of variant detection in low VAF like DeepVariant. The callers in the combinations may influence the results of variant detection in different VAFs. The results of shared variants in different combinations shown that even if some combinations have demonstrated superiority in the F1-scores, but there were still a number of variants that which did not detect. Those number of the variants might be some important variants for some diseases. That's one of the reason some studies applied the merged or ensemble-based approach which majorly combines the results of variant calling from multiple variant caller or combinations to improve the Accuracy of variant detection. The results of this study provide supportive evidence that the merging the results of multiple variant caller or combinations could improve the results of variant detection by using single or few variant caller or combinations.

## Reference


Betschart R.O., Thiéry A., Aguilera-Garcia D. et al. Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment.*Sci Rep* 2022; 12:21502.

Cibulskis K., et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* 2013*;* **31***(*3*): p.* 213–219.

Cornish A. *and* C. Guda, A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *Biomed Research International* 2015*;* **2015**.

Dietlein F., et al., Genome-wide analysis of somatic noncoding mutation patterns in cancer. *Science* 2022*;* **376***(*6589*): p.* 152–+.

Jones W., et al., A verified genomic reference sample for assessing performance of cancer panels detecting small variants of low allele frequency. Genome Biology 2021; **22**(1).

Kendig K.I., et al., Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. *Frontiers in Genetics* 2019*;* **10**.

Kim D., et al., Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 2019*;* **37***(*8*): p.* 907–+.

Kumaran M., U. Subramanian, *and* B. Devarajan, Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *Bmc Bioinformatics,* 2019*;* **20**.

Langmead B. *and* S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012 *;***9***(*4*): p.* 357–U54.

Liang, Y, et al., Comparative Analysis for the Performance of Variant Calling Pipelines on Detecting the de novo Mutations in Humans. *Frontiers in Pharmacology* 2019*;* **10**.

Li H. and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; **25**(14): p. 1754–1760.

Liu J., Q.M. Shen, *and* H.G. Bao, Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. *Plos One* 2022*;* **17***(*1*).*

Olson ND, Wagner J, McDaniel J, Stephens SH et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics 2022;* 2(5), 100129

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

Pabinger S., et al., A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* 2014*;* **15***(*2*): p.* 256–278.

Pleasance E., et al., Whole-genome and transcriptome analysis enhances precision cancer treatment options. *Ann Oncol* 2022*;* **33***(*9*): p.* 939–949.

Poplin R., et al., A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology 2018; **36**(10): p. 983–+

Wang Q., et al., Comparison of somatic variant detection algorithms using Ion Torrent targeted deep sequencing data. *Bmc Medical Genomics* 2019*;***12**.

Wagner J., Olson N.D., Harris L. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol 2022;40, 672–680

Zhao S., et al., Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci Rep* 2020*;* **10***(*1*): p.* 20222.

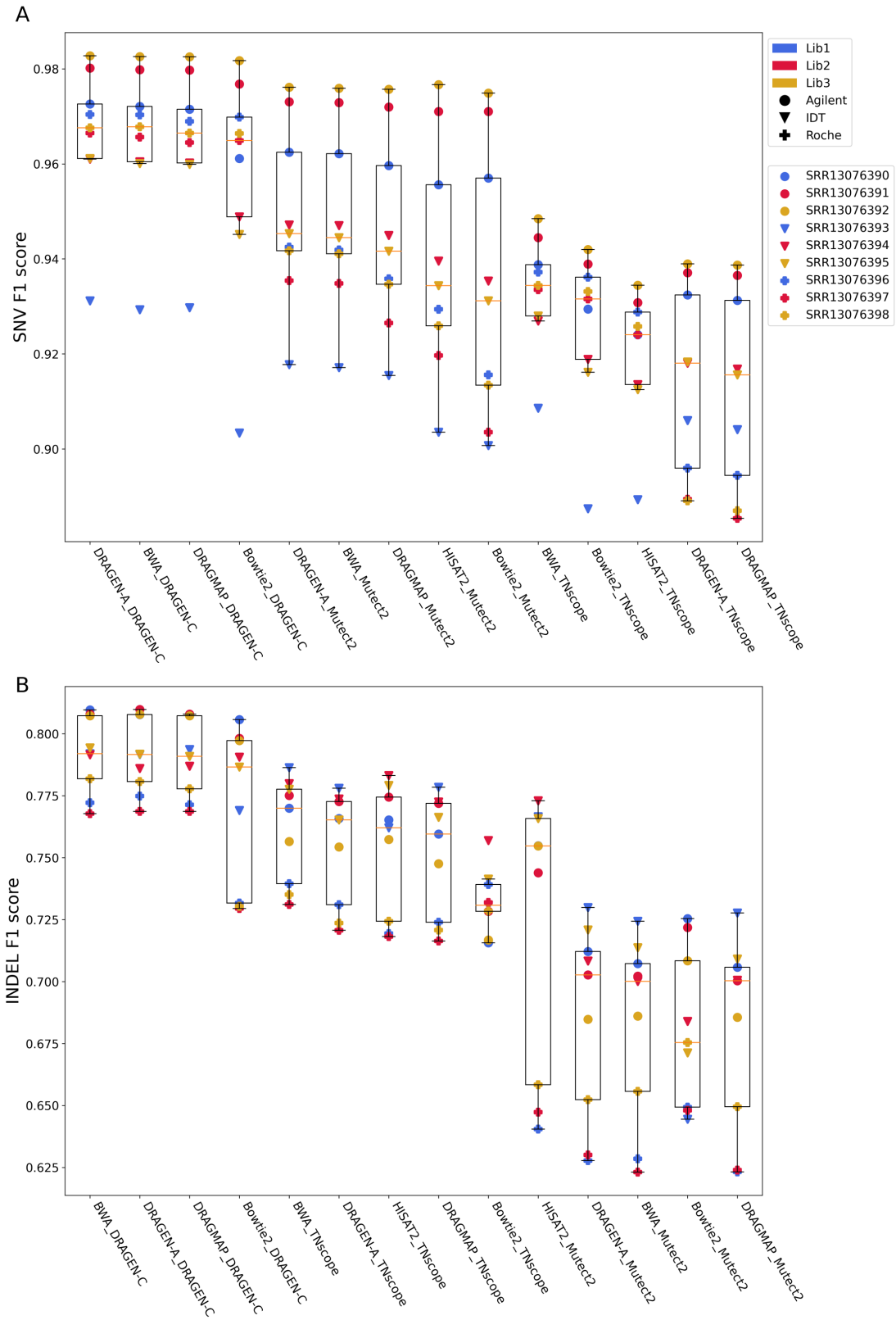


**Figure 1. Comparison the F1 scores of somatic mutation calls among workflows with SEQC2 WES datasets as truth set.**

Comparison of overall F1 scores for nine WES samples among three different labs and three library preparation kits (Agilent/IDT/Roche).

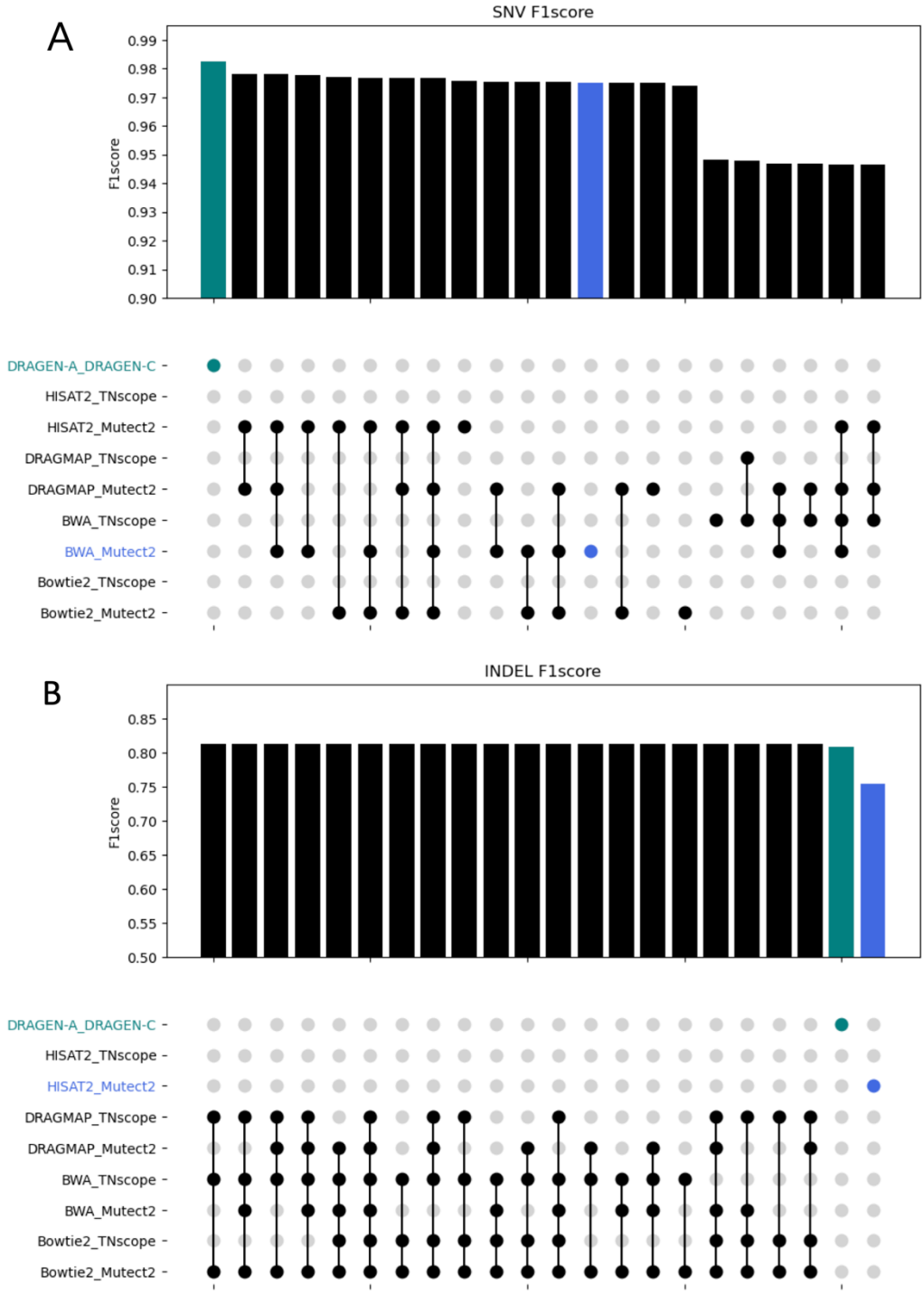


**Figure 2. Top 20 F1 scores of different workflow coalitions (A) SNV (B) INDEL.** Green bar indicates the performance of built-in DRAGEN (DRAGEN_A_GRAGEN_C), and the blue bar represents the best open-source combination BWA_Mutect2 for SNV as well as HISAT2_Mutect2 for INDEL

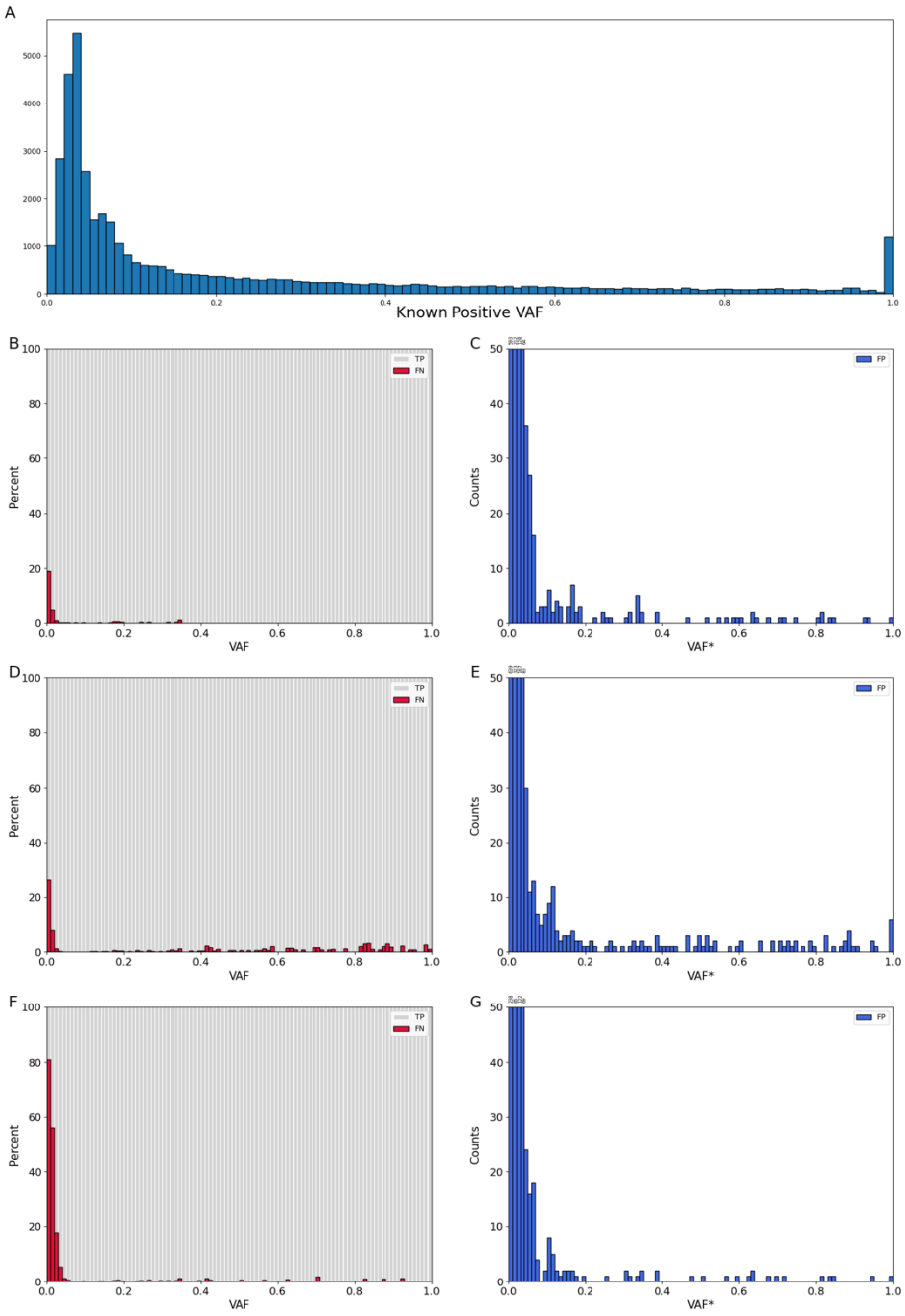


Figure 3. Variant allele frequency (VAF) spectrums for truth set and the calling results from selected combinations of S3 (SRR13076392) sample.

(A) VAF distribution in SEQC2 benchmark truth set. (B) For the DRAGEN-A_DRAGEN-C combination, red bar represents the percentage of false negative (FN) calls and the grey bar is for true positive (TP). (C) The number of false positive (FP) calls across different VAF. (D)(E) BWA-Mutect2 combination (F)(G) DRAGMAP-TNscope combination.

Fig4. True positive discovery rate of 9 samples under different pipeline in cancer related genes.

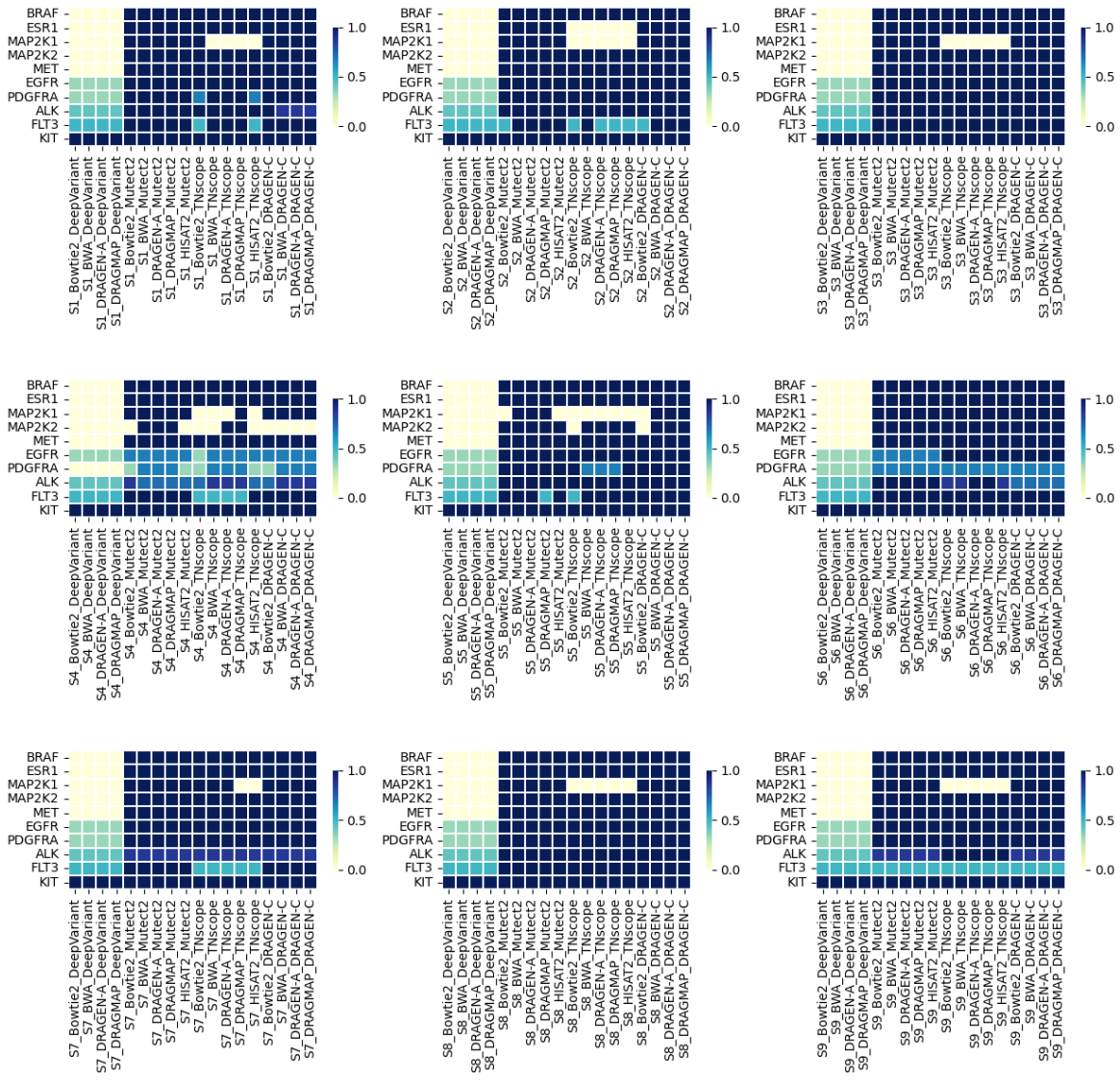


Fig5. True positive discovery rate of 9 samples under different pipeline in COSMIC provided drug resistance variants.

**Table 1: Recall, precision, F1 score of SNV and INDELs results under different combinations.** The table only included the best combination for commercial tool, open-source combination and the worst combination foe SNV and INDEL, respectively. The numbers are recall/precision/F1 scores.

| | SNV | | | INDEL | | |
|---|---|---|---|---|---|---|
| | Commercial | Open-source | | Commercial | Open-source | |
| Combination/ | DRAGEN-A | BWA | Bowtie2 | DRAGEN-A | HISAT2 | Bowtie2 |
| SampleNo. | DRAGEN-C | Mutect2 | DeepVariant | DRAGEN-C | Mutect2 | DeepVariant |
| SRR13076390 | 0.964/0.982/0.973 | 0.956/0.968/0.962 | 0.272/0.811/0.407 | 0.970/0.692/0.808 | 0.879/0.661/0.755 | 0.189/0.560/0.283 |
| SRR13076391 | 0.982/0.978/0.980 | 0.976/**0.970**/0.973 | 0.274/0.810/0.410 | 0.983/0.688/**0.810** | 0.910/0.629/0.744 | 0.189/0.554/0.282 |
| SRR13076392 | **0.990**/0.976/**0.983** | **0.983**/0.969/**0.976** | 0.272/0.812/0.407 | **0.991**/0.682/0.808 | **0.922**/0.639/0.754 | 0.195/0.559/0.289 |
| SRR13076393 | 0.883/**0.985**/0.931 | 0.889/0.947/0.917 | 0.284/0.819/0.422 | 0.878/**0.721**/0.792 | 0.807/**0.730**/0.767 | 0.199/0.566/0.294 |
| SRR13076394 | 0.940/0.983/0.961 | 0.940/0.954/0.947 | 0.282/0.816/0.419 | 0.941/0.675/0.786 | 0.881/0.688/**0.773** | 0.200/0.565/0.296 |
| SRR13076395 | 0.942/0.982/0.961 | 0.942/0.947/0.944 | 0.282/0.816/0.419 | 0.931/0.689/0.792 | 0.855/0.693/0.766 | 0.217/0.588/0.317 |
| SRR13076396 | 0.967/0.974/0.970 | 0.967/0.918/0.942 | 0.273/0.822/0.410 | 0.970/0.645/0.775 | 0.892/0.499/0.640 | 0.178/0.539/0.268 |
| SRR13076397 | 0.958/0.975/0.966 | 0.964/0.907/0.935 | 0.275/0.821/0.412 | 0.965/0.639/0.769 | 0.891/0.508/0.647 | 0.200/0.578/0.298 |
| SRR13076398 | 0.960/0.975/0.968 | 0.964/0.920/0.941 | 0.272/0.819/0.409 | 0.961/0.657/0.781 | 0.885/0.524/0.658 | 0.189/0.554/0.282 |
| Average | 0.954/0.979/0.966 | 0.953/0.944/0.949 | 0.276/0.816/0.413 | 0.954/0.676/0.791 | 0.880/0.619/0.723 | 0.195/0.563/0.290 |

**Table 2. Mutational signature results for S3/S4 samples in the top and the last performance combinations.**

| | Aligner | Caller | SBS(N=96) | DBS(N=11) | ID(N=18) |
|---|---|---|---|---|---|
| Known true set | | | SBS1 (17.58%) | DBS10 (64.42%) | ID4 (30.46%) |
| | | | SBS5 (51.32%) | DBS11 (35.58%) | ID15 (32.7%) |
| | | | SBS6 (18.16%) | | ID16 (36.84%) |
| | | | SBS54 (12.94%) | | |
| S3 | DRAGEN-A | DRAGEN-C | SBS1 (17.38%) | DBS10 (64.04%) | ID7 (46.04%) |
| | | | SBS5 (51.42%) | DBS11 (35.96%) | ID10 (27.64%) |
| | | | SBS6 (18.36%) | | ID16 (26.32%) |
| | | | SBS54 (12.84%) | | |
| | BWA | Mutect2 | SBS1 (17.20%) | DBS10 (60.94%) | ID7 (57.82%) |
| | | | SBS5 (51.42%) | DBS11 (39.06%) | ID10 (42.18%) |
| | | | SBS6 (18.42%) | | |
| | | | SBS54 (12.96%) | | |
| | BOWTIE2 | DeepVariant | SBS1 (19.28%) | DBS10 (100%) | ID4 (29.5%) |
| | | | SBS5 (58.32%) | | ID15 (29.32%) |
| | | | SBS54 (22.40%) | | ID16 (45.18%) |
| S4 | DRAGEN-A | DRAGEN-C | SBS1 (17.56%) | DBS10 (63.42%) | ID7 (41.96%) |
| | | | SBS5 (51.40%) | DBS11 (36.58%) | ID10 (31.02%) |
| | | | SBS6 (17.06%) | | ID16 (27.02%) |
| | | | SBS54 (13.98%) | | |
| | BWA | Mutect2 | SBS1 (17.38%) | DBS10 (60.86%) | ID7 (48.18%) |
| | | | SBS5 (52.48%) | DBS11 (39.14%) | ID10 (51.82%) |
| | | | SBS6 (16.66%) | | |
| | | | SBS54(13.48%) | | |
| | BOWTIE2 | DeepVariant | SBS1 (19.50%) | DBS10 (100%) | ID7 (33.0%) |
| | | | SBS5 (58.58%) | | ID10 (29.12%) |
| | | | SBS54 (21.92%) | | ID16 (37.88%) |