Towards best practice in cancer mutation detection with whole-genome and whole-exome sequencing

The Somatic Mutation Working Group of the SEQC-II Consortium

Abstract

Clinical applications of precision oncology require accurate tests that can distinguish cancer-specific mutations from errors introduced at each step of next generation sequencing (NGS). For NGS to successfully improve patient lives, discriminating between true mutations and artifacts is crucial. To date, no study has addressed the effects of cross site reproducibility together with the potentially influential interactions between biological, technical, and computational factors on the accurate identification of variants.

Here we systematically interrogated somatic mutations in paired tumor-normal cell lines to identify factors affecting detection reproducibility and accuracy. Different types of samples with varying input amount and tumor purity were processed using multiple library construction protocols. Whole-genome (WGS) and whole-exome sequencing (WES) were carried out at six sequencing centers followed by processing with nine bioinformatics pipelines to evaluate reproducibility. We identified artifacts of C>A mutations in WES due to sample and library processing and highlighted limitations of bioinformatics tools for artifact detection and removal.

By examining the interactions and effects of NGS platforms, library preparation protocol, tumor content, read coverage, and bioinformatics processes concomitantly, here we recommend

actionable practices for cancer mutation detection experiments and analysis using NGS technologies.

Introduction

Precision oncology seeks to identify whether a patient carries specific genetic aberrations relevant to a treatment or disease in order to guide pharmaceutical target identification, inform patient therapy, dictate dosing, suggest probability of adverse events[1], and/or determine membership of an individual into a specific patient subset. With ever-decreasing cost, more researchers and clinicians are using next-generation sequencing (NGS) to profile clinical samples. Currently, a panoply of sample processing protocols, library preparation methods, sequencing technologies, and bioinformatics pipelines exists to detect mutations relevant to cancer. Furthermore, samples can arrive at the testing lab in different states (fresh vs FFPE), the amount of input DNA can be variable, and tumor purity is rarely consistent across clinical samples, presenting significant challenges to sequencing assays, instruments, and analytical tools. All of these technical challenges likely underlie the estimated irreproducibility rates of 51%-89% in pre-clinical research [2–4].

Previous studies have successfully addressed individual components of somatic variant calling in isolation; for example, bioinformatics pipelines[5,6] or sample factors[7] have been studied individually. Many reports have compared bioinformatics pipelines/callers based on the accuracy and consistency of mutation detection[8–10]; others have compared other components, such as assay development[11], library preparation[12], and biosample resources[13,14]. The majority of these studies used either *in-silico* approaches with simulated ground truth or real tumor samples[11,15–17] for benchmarking. The data sets or samples from previous studies either do not accurately represent real human tumor biopsies (*in-silico*) or are not sustainable (real tumor)

for multiple benchmark studies. Although benchmarking with tumor tissue sounds more realistic for cancer mutation detection, the existence of spatial heterogenicity of tumor tissue[18,19], i.e. different slices of tissue may have different mutation profiles, and the limited sample quantity cannot support comprehensive studies required for reliable assessment of detection reproducibility.

Here we develop systematic methods to evaluate performance using realistic reference samples and datasets. We profile the previously characterized[20–22] and commercially available cell lines (breast cancer and matched normal). We outlined an experimental design where we sought to test the influence of each variable within a typical NGS-based tumor profiling workflow. Our design included different biospecimen types (fresh vs FFPE), input amounts, library preparation methods, different Illumina sequencing instruments, sequencing centers, and bioinformatics pipelines, which allowed us to investigate how these experimental and analytical elements may interact to affect mutation detection. We find that the whole process of cancer mutation detection, from experiment to informatics, greatly benefits from a well-integrated protocol. While reproducibility of WGS is better than WES, artifacts could be raised from biospecimen source, laboratory protocols, or bioinformatics tools.

Results

Study design

To pinpoint factors affecting somatic variant calling, a matched pair of breast cancer cell lines (HCC1395 and HCC1395BL) was selected for profiling (detailed in our "reference samples" manuscript, DOI:10.1101/625624). Although practical constraints precluded an exhaustive examination of all possible combinations of all possible variables, those chosen for study here reflect our real-world assessment of commonly-encountered factors exerting significant effects on the final result (**Fig. 1a, Suppl. Fig. 1**).

Our experimental design covered many real-world scenarios that occur in research or clinical laboratories, such as FFPE, heterogeneity of tumor biopsy, limited specimen DNA, ever-changing NGS machine models, and analytical tools. We generated 1,015 call sets to evaluate the cross-center and cross-platform reproducibility of mutation detection, as well as the performance of mutation calling under various conditions.

Survey of read quality

WGS was performed at six sequencing centers, generating a total of 42 sequencing results from the standard TruSeq PCR-free libraries prepared from 1000 ng of input DNA. The cross-platform comparison included three different platforms: HiSeq4000, HiSeq X10, and NovaSeq S6000. All sequencing centers and platforms produced high quality data as demonstrated by base call Phred quality scores above Q30, and greater than 99.8% of reads mapped to the reference genome (GRCh38). Variation was observed in the quantity of reads per sample generated, with some centers consistently delivering much higher coverages (100X) than others (50X), which was driven by sequencing platform yield differences as well as run pooling schemes and target coverage (**Fig. 1b**). Among the WGS libraries prepared using fresh cells, insert size distribution

and G/C content were uniform (40 – 43% G/C). Moreover, all of the WGS libraries had very low adapter contamination (<0.5%). Less than 10% of reads mapped redundantly for most libraries, indicating high complexity of the WGS libraries. Similar mapping statistics were observed in NovaSeq WGS runs (**Suppl. Table 1**).

Similarly, WES performed across six sequencing centers using three different HiSeq models (HiSeq1500, HiSeq2500, and HiSeq4000), generated sequencing results where 99% of reads were mapped successfully (**Suppl. Table 2**). The largest variations in sequencing yield and coverage on target were seen between sequencing centers and even sometimes between different replicates of the same cell line at the same sequencing center. These variations were largely due to uneven library pooling and sequencing yield differences between platforms. This confounding factor was easily identified by statistical analysis and could be removed with downsampling to produce equivalent coverage for each run. WES libraries, demonstrated much higher adapter contamination among all six centers, compared with WGS data. More than 20% of sequences in WES from two sequencing centers came from adapters, while the remaining centers had less than 12% adapter contamination (**Suppl. Fig. 2a)**. Higher adapter contamination was related to longer sequencing read length in base pairs (bp, 2x 150 bp vs. 2x125 bp). The WES libraries had higher G/C content (44 - 54%) compared with WGS libraries (40 – 43%). The libraries with more serious adapter contamination also had much higher G/C content compared other WES libraries. Generally, regions with more reads on target region had higher G/C content (**Suppl. Fig. 2b)**.

When comparing library preparation kits across different DNA inputs, both similarities and differences were noted. The average percentage of mapped reads ranged from 96% to 99.9% across TruSeq PCR-free, TruSeq-Nano, and Nextera Flex libraries prepared with 250, 100, 10, or 1 ng of DNA input. However, the percentage of non-redundant reads was very low (<20%) for TruSeq-Nano with 1 ng input, presumably due to PCR amplification (**Suppl. Fig. 2c**). Nevertheless, the percentage of non-redundant reads for 1 ng with Nextera Flex (also a PCR-based protocol) was reasonably good (~70%; comparable to the performance of the TruSeq-Nano 100 ng protocol). In addition, overall G/C content was not affected by DNA input amount or library preparation kit. Thus, the Nextera Flex Library prep may be superior to the TruSeq-Nano for lower input DNA amounts. The FFPE libraries were prepared from cells fixed with formaldehyde at four time intervals. These samples yielded results with a high percentage of mapped reads and non-redundant read frequencies that were comparable to the results generated from WGS libraries prepared with fresh cells (**Suppl. Table 1).**

Evaluation of DNA quality

The Global Imbalance Value (GIV) is a commonly used indicator of DNA damage[23] and was thus used to monitor DNA quality in NGS runs. We found high GIV scores for the G>T/C>A mutation pair in WES of HCC1395 and HCC1395BL cell lines. The GIV for G>T/C>A scores was inversely correlated with insert fragment size (**Suppl. Table 3)**, which is also found to be associated with DNA shearing time. Longer shearing time produces shorter DNA fragments. Insert fragment size and G/C content were also inversely correlated, suggesting increased off-target (non-exome region) reads when larger DNA fragments were sequenced. We observed high G>T/C>A scores

(>1.5) when insert fragment sizes were between 160-180 bp. When insert fragment sizes were larger than 200 bp, we observed little or no imbalance (**Fig. 1c**). In contrast, we did not observe this imbalance in WGS runs **(Suppl. Fig. 2d)**, for which the insert sizes were normally larger than 300 bp (**Suppl. Table 1, Suppl. Fig. 3a)**. We did not observe such imbalance in other mutation pairs, such as T>G/A>C (**Fig. 1c**). Previous reports demonstrate that excessive acoustic shearing results in 8-oxoG damage [24]. Therefore, the high ratio of G>T/C>A observed in some WES runs was likely an artifact of oxidative DNA damage during the fragmentation process.

Formaldehyde also causes the deamination of guanine. Thus, the GIV of G>T/C>A is a good indicator of FFPE induced DNA damage[25]. Consistent with this, we observed "dose" dependent GIV imbalance in FFPE sample WGS runs (**Fig. 1d**). Taken together, these results indicate that WES is more sensitive to site-to-site library preparation variation than WGS. We propose that WGS rather than WES is best suitable to FFPE.

Reproducibility of cancer mutation detection

To assess the reproducibility of cancer mutation detection with WES and WGS, we performed a total of twelve repeats of WES and WGS at six sequencing centers (**Fig. 1a**, **Suppl. Fig. 1).** We used three mutation callers (MuTect2[26], Strelka2[27], and SomaticSniper[28]) on alignments from three aligners (Bowtie2[29] , BWA[30], and NovoAlign) to generate a total of 108 Variant Call Format (VCF) files from WES and WGS analyses separately.

In this study, as shown in **Fig. 2a**, both BWA and NovoAlign demonstrated a substantial pool of calls which were agreed upon under every repeated WGS or WES run called by three callers (MuTect2, Strelka2, and SomaticSniper). We did not observe significant differences among results from the three aligners on WES. However, calling results from WGS with Bowtie2 as the aligner tended to have fewer consistent SNV calls, indicating that mutation calling from Bowtie2 alignment was conservative **(Fig. 2a)**.

We then fixed the alignment to the widely used Burrows-Wheeler Aligner (BWA), and compared the results from our three callers, observing the differences in WES versus WGS run performance. While SomaticSniper displayed results that showed more consistent SNV calls in WES than in WGS, those derived by MuTect2 and Strelka2 showed greater divergence. Both MuTect2 and Strelka2 were more consistent when used for WGS rather than WES (**Fig. 2b)**. Further examination of Strelka2 results from BWA alignments of twelve repeated WES and WGS runs confirmed this observation (**Fig. 2c)**. Here we also introduced the O_Score, a metric to measure reproducibility of repeated analyses (see Supplemental Methods). O_Scores for Strelka2 and MuTect2 for WES runs were not only significantly lower than for WGS runs, but also more variable. Unexpectedly, even though its overall O_Score was much lower than Strelka2 and MuTect2 in WGS runs, SomaticSniper showed better consistency in calling results in WES than in WGS (**Suppl. Fig. 4a)**. We also compared WGS runs with HiSeq vs. WGS runs with NovaSeq using BWA for alignment and Strelka2 for calling. Both platforms were remarkably similar in terms of reproducibility, indicating that results from HiSeq and NovaSeq were comparable (**Suppl. Fig. 4b**). Taken together, Strelka2 had the best reproducibility in WGS

repeated runs but the worst in WES repeated runs, whereas MuTect2 had the best reproducibility in WES repeated runs.

Factors influencing reproducibility of cancer mutation detection

After establishing the O_Score to measure the reproducibility of NGS platform mutation detection, we also used it to determine what variables contribute most to variation between different repeated WGS or WES analyses separately. We included not only callers but machine model, read coverage, non-duplicated reads, G/C content, insert size, and two GIV scores (G>T/C>A and T>G/A>C). The combination of these parameters represented most of the variation in WGS and WES runs, as a very high proportion of variance in the O_Score could be predicted by these variables. For WGS, the combination of these eight individual variables accounted for >99% of the O_Score variance ($R^2 > 0.99$) (**Suppl. Fig. 5a**). On the other hand, individual variables and five interaction terms (callers*coverage, callers*percent GC, caller*machine model, callers*GIV(G>T/C>A), and callers*non-duplicated reads) were significant for O_Score variance in WES runs (**Suppl. Fig. 5b**).

As all twelve WGS runs were done with TruSeq PCR-free libraries with the same amount of DNA input, the percentage of non-duplicated reads did not affect reproducibility. While individual variables may influence WES run reproducibility, their effect levels were dependent on caller selection (**Fig. 2d**). Taken together, only a few factors (read coverage and callers) affected WGS reproducibility, whereas several factors, such as caller, read coverage, insert fragment size, GC content, GIV (G>T/C>A) score, and their interactions influenced WES run reproducibility. It is

noteworthy that since an older machine model (HiSeq1500) was used in only one WES sequencing center, which also had low insert fragment size and high GIV (G>T/C>A) scores in sequencing reads, the impact of machine model on the WES run reproducibility could be a confounding factor. In contrast, the most influential factor for the performance of WGS was the caller, next followed by read coverage.

Effect of non-analytical and analytical factors on mutation calling

To thoroughly investigate the influence of non-analytical and analytical factors on cancer mutation calling, we set out to define the truth set of somatic mutation in HCC1395 for our benchmarking study (see our "reference samples" manuscript, DOI:10.1101/625624).

With three different library preparation protocols and varying DNA input amounts across multiple library preparations, we analyzed outcomes using combinations of the three callers and three aligners. MuTect2 was reliable, except for calling of the 1 ng TruSeq-Nano libraries (**Fig. 3a**). Strelka2's performance suffered by more than 50% from the 1 ng TruSeq-Nano libraries, as did SomaticSniper. We conclude that Nextera Flex library preparation might be a better option for a low input DNA quantity.

FFPE processing can have an effect on variant calling results[31]. In our study, samples of fresh cells and cells processed with FFPE were called with Strelka2, SomaticSniper, and MuTect2. Both MuTect2 and Strelka2's precision and recall were greatly reduced when samples were subjected to FFPE processing (**Fig. 3b**). On the other hand, SomaticSniper demonstrated only a

small decrease in both metrics, but otherwise underperformed significantly compared with the other two callers.

Sometimes FFPE or low tumor purity samples are all that is available for study. Bioinformatics can potentially reduce sample related biases. To evaluate bioinformatics pipelines, reads were pre-processed using Trimmomatic[32] or BFC[33] to assess whether read trimming and error correction could affect WES variant call recall or precision (see Online Methods). Applying error correction with BFC to FFPE samples increased precision but did not improve the recall rate (**Suppl. Fig. 6a**). However, precision was improved for fresh DNA samples that have undergone BFC processing, but with a lower recall rate than results from Trimmomatic processing. Taken together, these results indicate that BFC is appropriate in cases of severe DNA damage but may not be worthwhile if there is only mild damage from a process such as sonication.

FFPE is known to cause G>T/C>A artifacts as well[31]. Trimmomatic and BFC were investigated for their ability to detect these errors. Since the DNA damage causing the G>T/C>A mutation may not be isolated to only low quality base calls at the end of reads, Trimmomatic is not designed to remove this type of artifact. Trimmomatic processed data was more skewed toward C>A artifacts than was BFC processed data, which showed changes more broadly across nucleotide transitions (**Fig. 4a**). BFC reduced C>A artifacts, but introduced a few artifacts of other types, such as T>C mutations, indicating that caution should be exercised when using bioinformatics tools to correct FFPE artifacts.

Calling accuracy was dependent on the choice of caller and aligner, as well as how the caller and aligner interact. Strelka2 results with BWA-aligned reads were balanced, whereas results with Bowtie2-aligned reads seemed conservative. In contrast, Strelka2 results with NovoAlign-aligned reads seemed aggressive (**Fig. 4b**). When we examined the mapping quality scores for the three alignments, BWA's mapping quality scores were usually between 50 and 60, Bowtie2's scores were between 40 and 50, and NovoAlign's scores were between 60 and 70. Strelka2 was trained and modeled on the BWA alignment and thus works best in the bioinformatics context where it was developed[27]. Taken together, these results indicate that there might be a joint effect between aligner and caller, depending on how callers were developed and on which aligners' dataset they were trained on.

Next, the effect of the Genome Analysis Toolkit (GATK) local indel realignment and base quality score recalibration (BQSR) was queried. MuTect2 and Strelka2 identified a very similar number of SNVs regardless of whether the process was employed (**Fig. 4c**). Conversely, SomaticSniper was highly sensitive to post alignment processing, with some SNVs gained and far more lost. Relatedly, the precision and recall rate changed dramatically when this process was applied for SomaticSniper calling, but not for calling by MuTect2 or Strelka2 (**Suppl. Fig. 6b, c**). Clearly, our study confirms, with real-world experimental data, the importance of a full understanding of how the various components of mutation analysis by NGS methods work and interact with each other.

Tumor purity and coverage also play a role in caller performance. As expected, higher read coverage yielded more SNV calls (**Fig. 4d**). When tumor purity was high (>50%), 50X performed very similarly to 100X coverage across all callers tested. When tumor purity was low (<50%), calling was much more sensitive to sequencing depth. To test the performance of callers on low tumor purity samples, we pooled reads from WGS triplicate runs on samples that were sequenced at 100X coverage to generate coverage of either 200X or 300X on each cell line. In addition to the three main callers used in this study, we also included two new tools, TNscope[34] and Lancet[35], to compare their capabilities for detecting mutations from a mix with tumor DNA as low as 5%. With high tumor purity (>50%), we again observed that the accuracy of all the callers, with the exception of SomaticSniper, declined slightly with higher read coverage, indicating that our truth set lacked <5% variant allele frequency (VAF) mutations. However, when tumor purity was 20% or lower, the benefit of higher coverage for mutation detection was apparent. For a sample with 20% tumor, Lancet, Strelka2, and TNscope performed similarly well with 300X coverage (**Fig. 4d)**. On the other hand, SomaticSniper performed poorly at any tumor purity level, and increased read coverage did not rescue performance. These results indicate that tumor purity is much more influential than coverage in the ranges tested here.

Performance of WGS and WES across multiple sequencing centers

As this study leveraged six sequencing centers performing twelve WES and WGS experiments simultaneously, we were able to assess inter- and intra-center reproducibility of the two sequencing platforms. Using our established call set as a reference, we defined resulting SNV calls from any of the NGS runs with the pair of cell lines into three categories: 1) Repeatable

(SNVs in the truth set defined in “HighConf” and “MedConf” categories); 2) Gray zone (SNVs defined in “LowConf” and “Unclassified” categories); and 3) Non-Repeatable (SNVs not found in the truth set) **(Suppl. Fig. 7)**. Cross-center and cross-platform variations were very small for “Repeatable” SNVs, indicating that all individual NGS runs, regardless of sequencing centers or NGS platforms, detected most “true” mutations consistently. This “consistency” dropped dramatically for SNVs in the “Gray zone”, and further down to nearly zero, if SNVs were “Non-Repeatable” (**Table 1)**.

Taken together, these results indicated that there were two major sources for discordant SNV calls between any two different library preparations: 1) stochastic effects of sequence coverage on SNVs with low VAF, which are sensitive to read coverage and mainly represented by SNVs in the “Gray zone” group; 2) artifacts due to library preparation, mainly represented by SNVs in the “Non-Repeatable” group. The benefit of high read coverage not only empowers the detection of mutations with low VAF, but also increases result reproducibility (for both WES and WGS), likely due to reduction of stochastic effects.

Using multi-variate analysis, we were able to further dissect the source of variations driving reproducibility of mutation detection by WES and WGS. Consistent with our results from O_Score analysis (**Fig. 2d**), callers, read coverage, and platforms were the major factors influencing the reproducibility of mutation detection. However, the subset of SNVs/indels (Repeatable, Non-Repeatable, and Gray zone) was the dominant source of inconsistent mutation calls (**Suppl. Fig. 8 a)**. In addition, we observed that the difference between inter-

/intra-center variations was subjected to callers (**Suppl. Fig. 8b)** and SNV/indels subsets (**Suppl. Fig. 8c**). Overall, inter-center variations for WES were larger than inter-center variations for WGS, whereas the difference of intra-center variation between WES and WGS was observed to be insignificant (**Suppl. Fig. 8d**).

Moreover, results from the Jaccard index score analysis confirmed our conclusions: for WES, Strelka2 was less reproducible than MuTect2 (**Fig. 2b, Suppl. Fig. 4a, and Suppl. Fig. 6a**). As shown in **Fig. 4a**, Strelka2 was very sensitive in detecting C->A mutation artifacts from excessive sonication. Thus, it was not surprising to see that many "Non-Repeatable" SNVs detected by Strelka2 in WES were C->A mutations. At a less extensive level, MuTect2 also detected C->A mutation artifacts; however, this artifact was not observed by SomaticSniper (**Suppl. Fig. 9).**

Precision and recall rates from WES and WGS were also compared across all three callers and all twelve replicates directly. Mutations shared by two replicates generally had higher precision. Moreover, almost all mutations called by both MuTect2 and SomaticSniper were true. Leveraging additional callers increased precision; however, this was at the cost of recall (**Fig. 5**). Taken together, by the precision metric, WGS clearly out-performed WES across replicates, callers, and sequencing centers. These results demonstrate the importance of using enough library replicates during study design, rather than trying to compensate by using multiple callers.

Finally, we compared precision, recall, and F-score (as defined in Supplemental Methods) across a range of variant allele frequencies with three callers in both WES and WGS. Interestingly, Strelka2 exhibited the best performance on WGS samples but the worst performance on WES samples (**Suppl. Fig. 10**), which was consistent with the results from the reproducibility study (**Fig. 2**). On the other hand, even though SomaticSniper did not perform well overall, it was not affected by C>A artifacts. However, for both WES and WGS runs, we did observe that the limit of VAF calling by SomaticSniper was ~12%, and many false positives were misidentified by SomaticSniper at a higher VAF (**Suppl. Fig. 11a, b**). Therefore, the "better" performance of SomaticSniper compared to Strelka2 on WES was driven by SomaticSniper's insensitivity to artifacts from the DNA fragmentation process (**Fig. 1c and Suppl. Fig. 11a**).

Discussion

We observed that each component of the sequencing and analysis process can affect the final outcome. Overall concordance and correlation of WES and WGS results were good. WES had a better coverage/cost ratio than WGS. However, sequencing coverage of the WES target regions was not even (**Suppl. Fig. 11c**). In addition, WES showed more batch effects/artifacts due to laboratory processing and thus had larger variation between runs, laboratories, and likely between researchers preparing the libraries. As a result, WES was less reproducible than WGS. WGS had more uniform coverage and was less sensitive to different runs and laboratories. Our experimental design also allowed us to estimate inter-/intra-center variation for both WES and WGS platforms. WES had much larger inter-center variation than WGS; however, intra-center

variation for both platforms was quite comparable (**Suppl. Fig. 8**). Biological (library) repeats removed some artifacts due to random events ("Non-Repeatable" calls) and thus offered much better calling precision than did a single test. Analytical repeats (two bioinformatics pipelines) also increased calling precision at the cost of increased false negatives (**Fig. 5**). We found that biological replicates are more important than bioinformatics replicates in cases where high specificity and sensitivity are needed.

Detection of cancer mutations is an integrated process. No individual component can be singled out in isolation as more important than any other, and specific components affect and interact with each other. Every component and every combination of components can lead to false discovery (false positives or false negatives). Individual components of cancer mutation calling should never be considered "plug and play". Although the initial steps of NGS test validation may be performed in three separate stages (platform, test-specific, and informatics)[36], our study demonstrates the complex interdependency of these stages on overall NGS test performance. Thus, final validation and verification studies should be performed using the entire sample-to-result pipeline. Detailed recommendations of cancer mutation detection experiments and analysis are provided in **Suppl. Table 6**.

We cannot conclude that results from our study are generalizable to other cancer types. However, high complexity of chromosome loss/gains (see our "reference samples" manuscript, DOI:10.1101/625624) and a large number of somatic mutations, cell line HCC1395 resembles the abnormalities commonly seen in a hyper-diploid cancer genome.

In summary, the pair of tumor/normal cell lines, data sets with FFPE DNA, tumor/normal DNA mix, and performance of bioinformatics components established in this study may serve as a reference for the NGS research community to perform benchmarking studies for the development of new NGS products, assays, and informatics tools. We also provide recommendations for DNA fragmentation for WES runs, and a selection of NGS platforms, and bioinformatics tools based on the nature of available biosamples and study objectives.

Online Methods

Data availability

All raw data (FASTQ files) are available on NCBI's SRA database (SRP162370). The truth set for somatic mutations in HCC1395, VCF files derived from individual WES and WGS runs, and source codes for "tornado" plot are available on NCBI's ftp site (ftp://ftp-trace.ncbi.nlm.nih.gov/seqc/ftp/release/Somatic_Mutation_WG/). Alignment files (BAM) are also available on Seven Bridges' s Cancer Genomics Cloud (CGC) platform and Digicon's BioGenLink™.

Acknowledgements

We thank Sivakumar Gowrisankar of Novartis, Susan Chacko of the Center for Information Technology, the National Institute of Health for their assistance with data transfer, Dr. Jun Ye of Sentieon for providing the Sentieon software package. We also appreciate Dr. Laufey Amundadottir of the Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), National Institutes of Health (NIH), for the sponsorship and the usage of the NIH Biowulf cluster; Dr. Reena Phillip of the Center for Devices and Radiological Health, U.S. Food and Drug Administration, for her advices on study design; and Seven Bridges for providing storage and computational support on the Cancer Genomic Cloud (CGC). The CGC has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I. Research done in Dr. Charles Wang's lab was partially supported by the Ardmore Institute of Health grant 2150141 (CW) and Dr. Charles A. Sims' gift. Chunlin Xiao and Steve Sherry were supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. This work also used the computational resources of the

NIH Biowulf cluster (http://hpc.nih.gov). Original data was also backed up on the servers provided by Center for Biomedical Informatics and Information Technology (CBIIT), NCI.

In addition, we thank following people for their participation of working group discussions: Meredith Ashby, Ozan Aygun, Xiaopeng Bian, Pierre Bushel, Fabien Campagne, Tao Chen, Han-Yu Chuang, Youping Deng, Don Freed, Paul Giresi, Ping Gong, Yan Guo, Christos Hatzis, Susan Hester, Jonathan Keats, Eunice Lee, You Li, Sharon Liang, Tim McDianiel, Jai Pandey, Anand Pathak, Tieliu Shi, Jeff Trent, Mingyi Wang, Xiaojian Xu, Chaoyang Zhang. The following people from the University of Toledo Medical Center helped to troubleshoot or set up the FFPE protocol: Abdelraof Al-Agha, Theresa Cummins, Carol Freeman, Connie Nowak, Amy Smigelski, Jiyoun Yeo, and Veronika Kholodovych.

Disclaimer

This is a research study, not intended to guide clinical applications. The views presented in this article do not necessarily reflect current or future opinion or policy of the US Food and Drug Administration. Any mention of commercial products is for clarification and not intended as endorsement.

Author contributions

Study conceived and designed by: Wenming Xiao[1,36], Leming Shi[2], Eric Donaldson[35], Samir Lababidi[30], Louis M. Staudt[37], and Zivana Tezak[36]

Biosample preparation: Thomas Blomquist[6], Jennifer A. Hipp[6], Liz Kerrigan[5], Kurt Langenbach[5], Maryellen de Mars[5], and James Willey[6]

NGS library preparation and sequencing: Wanqiu Chen[7], Xin Chen[7], Zhong Chen[7], Jiri Drabek[25,40], Ana Granat[15], Tiffany Hung[9], Kenneth Idler[14], Howard Jacob[14], Erich Jaeger[15], Rasika Kalamegham[9], Bindu Kanakamedala[33], Sulev Kõks[31,40,44], Yuliya Kriga[23], Ulrika Liljedahl[34,40], Roberta Maestro[16,40], Christopher Mason[18], Timothy Mercer[22], Chris Miller[14], Malcolm Moos[41], Ali Moshrefi[15], Aparna Natarajan[15], Baitang Ning[1], Jessica Nordlund[34,40], Eric Peters[9], Virginie Petitjean[11], Maurizio Polano[16,40], Arati Raziuddin[23], Ene Reimann[40,44], Andreas Scherer[26,40], Gary Schroth[15], Jyoti Shetty[23], Leming Shi[2], Marc Sultan[11], Weida Tong[1], Bao Tran[23], Tiffany Truong[15], Petr Vojta[25,40], Charles Wang[7,44], Jing Wang[27], and Yuanting Zheng[2]

Data analysis: Wenming Xiao[1,36], Li Tai Fang[3], Yongmei Zhao[4], Luyao Ren[2], Justin Lack[4], Ogan Abaan[39], Margaret Cam[12], Qingrong Chen[10], Xin Chen[7], Yun-Ching Chen[8], Ben Ernest[19], Meijian Guan[42], Parthav Jailwala[12], Roderick V Jensen[17], Wendell Jones[21], Jian-Liang Li[28], Zhipan Li[32], Xuelu Liu[43], Charles Lu[14], Daoud Meerzaman[10], Urvashi Mehra[19], Corey Miles[1], Cu Nguyen[10], Brian N. Papas[28], Mehdi Pirooznia[8], Wolfgang Resch[13], Fayaz Seifuddin[8], Tsai-Wei Shen[4], Keyur Talsania[4],

Cristobal Juan Vera[4], Jingya Wang[29], Ashley Walton[4], Leihong Wu[1], Chunlin Xiao[38], Chunhua Yan[10], Gokhan Yavas[1], Ying Yu[2], and Bin Zhu[20]

Data management: Wenming Xiao[1,36], Chunlin Xiao[38], Steve Sherry[38], Cu Nguyen[10], and Daoud Meerzaman[10]

Manuscript writing: Wenming Xiao[1,36], Rebecca Kusko[24], Luyao Ren[2], Li Tai Fang[3], Malcolm Moos[41], Yongmei Zhao[4], and Charles Wang[7]

Project management: Wenming Xiao[1,36], Leming Shi[2], Charles Wang[7], Chunlin Xiao[38], and Huixiao Hong[1]

Affiliations

[1]Bioinformatics branch, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079; [2]State Key Laboratory of Genetic Engineering, School of Life Sciences and Shanghai Cancer Center, Fudan University, 2005 SongHu Road, Shanghai, China 200438; [3]Bioinformatics Research & Early Development, Roche Sequencing Solutions Inc. , 1301 Shoreway Road, Suite #300, Belmont, CA 94002; [4]Advanced Biomedical and Computational Sciences, Biomedical Informatics and Data Science Directorate, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD21701; [5]ATCC (American Type Culture Collection), 10801 University Blvd, Manassas, VA 20110; [6]Department of Pathology, University of Toledo Medical Center, 3000 Arlington Road, Toledo, Ohio, 43614; [7]Center for Genomics, Loma Linda University School of Medicine, 11021 Campus St., Loma Linda, CA 92350; [8]Bioinformatics and Computational Biology Core, National Heart Lung and Blood Institute, National Institutes of Health, 12 SOUTH DR. Bethesda MD 20892; [9]Companion Diagnostics Development, Oncology Biomarker Development, Genentech, 1 DNA Way, South San Francsisco, CA 94080; [10]Computational Genomics Research, Center for Biomedical Informatics and Information Technology (CBIIT), National Cancer Institute, 9609 Medical Center Drive, Rockville, MD 20850; [11]Biomarker development, Novartis Institutes for Biomedical Research, Fabrikstrasse 10, CH-4056 Basel, Switzerland; [12]CCR Collaborative Bioinformatics Resource (CCBR), Office of Science and Technology Resources, Center for Cancer Research, NCI, Bldg 37, Rm 3041C, 37 Convent Drive, Bethesda, MD 20892; [13]Center for Information Technology, National Institutes of Health, 12 SOUTH Dr. Rm 2N207E, Bethesda, MD 20892; [14]Computational Genomics, Genomics Research Center (GRC), AbbVie, 1 North Waukegan Road, North Chicago, IL 60064; [15]Core applications group, Product development, Illumina Inc , 200 Lincoln Centre Dr. Foster City, CA 94404; [16]Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, National Cancer Institute, Unit of Oncogenetics and Functional Oncogenomics, Via Gallini 2, 33081 Aviano (PN), Italy; [17]Department of Biological Sciences, Virginia Tech, Life Sciences 1, 970 Washington St., Blacksburg, VA 24061; [18]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA, 1305 York Ave., Y13-05, New York, NY 10065; [19]Digicon, 7926 Jones Branch Drive, Suite 615, McLean, VA 22102; [20]Division of Cancer Epidemiology and Genetics, National

Cancer Institute, National Institutes of Health, 9609 Medical Center Drive, Bethesda, Maryland 20892, USA.; [21]EA Genomics, a business unit of Q2 Solutions, Suite 100, 5927 South Miami Blvd Morrisville, NC USA 27560; [22]Genomics and Epigenetics Division, Garvan Institute of Medical Research, 384 Victoria St, Darlinghurst NSW 2010, Australia; [23]Genomics Laboratory, Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD21701; [24]Immuneering Corporation, One Broadway 14th Fl Cambridge MA 02142 USA (Current address: Illumina Inc, 5200 Illumina Way, San Diego, CA); [25]IMTM, Faculty of Medicine and Dentistry, Palacky University Olomouc, Hnevotinska 5, 77900 Olomouc, the Czech Republic; [26]Institute for Molecular Medicine Finland (FIMM), HiLIFE, P.O.Box 20, FI-00014 University of Helsinki, Finland; [27]National Institute of Metrology, No.18, BeiSan Huan Dong Lu, Beijing, China 100029; [28]Integrative Bioinformatics, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA., 111 T W Alexander Dr, Durham, NC 27709 ; [29]MedImmune, LLC, 1 MedImmune Way, Gaithersburg, MD, 20878; [30]Office of Health Informatics, Office of the Chief Scientist, Office of the Commissioner, Food and Drug Information, 10903 New Hampshire Avenue, Silver Spring, Maryland 20993; [31]Perron Institute for Neurological and Translational Science, Verdun St, Nedlands, Western Australia, 6009, Australia; [32]Sentieon Inc., 465 Fairchild Drive, Suite 135, Mountain View CA 94043; [33]Transcriptomic Research, Division of Genomics and Epigenetics, Garvan Institute of Medical Research , The Kinghorn Cancer Centre, 370 Victoria Street, Darlinghurst, NSW 2010; [34]Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Molekylär Medicin, Box 1432, BMC, Uppsala, 75144 Sweden; [35]Division of Antiviral Products, Office of Antimicrobial, Center for Drug Evaluation and Research, FDA, Silver Spring, Maryland; [36]The Center for Devices and Radiological Health, U.S. Food and Drug Administration, FDA, Silver Spring, Maryland; [37] Lymphoid Malignancies Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland; [38]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,45 Center Drive, Bethesda, Maryland 20894; [39]Seven Bridges Genomics Inc.,1 Main St, 5th Floor, Suite 500 Cambridge, MA 02142; [40]member of EATRIS ERIC- European Infrastructure for Translational Medicine; [41]The Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, FDA, Silver Spring, Maryland; [42]SAS Institute Inc., Cary, NC 27513, USA; [43]Biomedical Imaging Research Services Section, Center for Information Technology, National Institutes of Health, 12 Center Drive, Bethesda, Maryland 20894; [44]University of Tartu, Department of Pathophysiology, 19 Ravila Street, Tartu, 50411, Estonia

References:

1. Zhang, C., Hong, H., Mendrick, D. L., Tang, Y. & Cheng, F. Biomarker-based drug safety assessment in the age of systems pharmacology: from foundational to regulatory science. *Biomarkers in Medicine* **9**, 1241–1252 (2015).

2. Glasziou, P., Meats, E., Heneghan, C. & Shepperd, S. What is missing from descriptions of treatment in trials and reviews? *BMJ* **336**, 1472–1474 (2008).

3. Vasilevsky, N. A. *et al*. On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* **1**, e148 (2013).

4. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).

5. Alioto, T. S. *et al*. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications* **6**, 10001 (2015).

6. Griffith, M. *et al*. Genome Modeling System: A Knowledge Management Platform for Genomics. *PLOS Computational Biology* **11**, e1004274 (2015).

7. Chalmers, Z. R. *et al*. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine* **9**, (2017).

8. Xu, H., DiCarlo, J., Satya, R. V., Peng, Q. & Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **15**, 244 (2014).

9. Ghoneim, D. H., Myers, J. R., Tuttle, E. & Paciorkowski, A. R. Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res Notes* **7**, 864 (2014).

10. Wang, Q. *et al*. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med* **5**, 91 (2013).

11. Simen, B. B. *et al*. Validation of a next-generation-sequencing cancer panel for use in the clinical laboratory. *Arch Pathol Lab Med* **139**, 508–517 (2015).

12. Linderman, M. D. *et al*. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med Genomics* **7**, 20 (2014).

13. Zook, J. M. *et al*. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology* **32**, 246–251 (2014).

14. Zook, J. M. *et al*. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data* **3**, 160025 (2016).

15. Lin, M.-T. *et al*. Clinical Validation of KRAS, BRAF, and EGFR Mutation Detection Using Next-Generation Sequencing. *Am J Clin Pathol* **141**, 856–866 (2014).

16. Singh, R. R. *et al*. Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. *J Mol Diagn* **15**, 607–622 (2013).

17. Griffith, M. *et al*. Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems* **1**, 210–223 (2015).

18. Morrissy, A. S. *et al*. Spatial heterogeneity in medulloblastoma. *Nature Genetics* **49**, 780–788 (2017).

19. Araf, S. *et al*. Genomic profiling reveals spatial intra-tumor heterogeneity in follicular lymphoma. *Leukemia* **32**, 1261 (2018).

20. Stephens, P. J. *et al*. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).

21. Kalyana-Sundaram, S. *et al*. Gene Fusions Associated with Recurrent Amplicons Represent a Class of Passenger Aberrations in Breast Cancer. *Neoplasia* **14(8)**, 702–708 (2012).

22. Zhang, J. *et al*. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. *Genome Res*. **26**, 108–118 (2016).

23. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**, 752–756 (2017).

24. Costello, M. *et al*. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* **41**, e67 (2013).

25. Do, H. & Dobrovic, A. Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization. *Clinical Chemistry* **61 (1)**, 64–71 (2015).

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

27. Saunders, C. T. *et al*. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

28. Larson, D. E. *et al*. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).

29. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10**, R25 (2009).

30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

31. Ivanov, M. *et al*. Towards standardization of next-generation sequencing of FFPE samples for clinical oncology: intrinsic obstacles and possible solutions. *Journal of Translational Medicine* **15**, 22 (2017).

32. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

33. Li, H. BFC: correcting Illumina sequencing errors. *Bioinformatics* **31**, 2885–2887 (2015).

34. Freed, D., Pan, R. & Aldana, R. TNscope: Accurate Detection of Somatic Mutations with Haplotype-based Variant Candidate Detection and Machine Learning Filtering. *bioRxiv* 250647 (2018). doi:10.1101/250647

35. Narzisi, G. *et al*. Lancet: genome-wide somatic variant calling using localized colored DeBruijn graphs. *bioRxiv* 196311 (2017). doi:10.1101/196311

36. Gargis, A. S. *et al*. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology* **30**, 1033–1036 (2012).

Concordance of SNVs (measured by Jaccard index)		**MuTect2**		**Strelka2**		**SomaticSniper**	
		WES	**WGS**	**WES**	**WGS**	**WES**	**WGS**
Overall SNVs	**overall**	**0.54**	**0.72**	**0.36**	**0.74**	**0.45**	**0.34**
	intra-center	**0.57**	**0.73**	**0.44**	**0.76**	**0.52**	**0.36**
	inter-center	**0.54**	**0.72**	**0.35**	**0.74**	**0.44**	**0.34**
Repeatable	**overall**	**0.83**	**0.92**	**0.9**	**0.91**	**0.87**	**0.90**
	intra-center	**0.87**	**0.93**	**0.94**	**0.93**	**0.89**	**0.91**
	inter-center	**0.82**	**0.91**	**0.89**	**0.91**	**0.86**	**0.90**
Gray zone	**overall**	**0.29**	**0.28**	**0.33**	**0.21**	**0.30**	**0.22**
	intra-center	**0.43**	**0.32**	**0.44**	**0.25**	**0.45**	**0.26**
	inter-center	**0.26**	**0.28**	**0.31**	**0.21**	**0.27**	**0.21**
Non-Repeatable	**overall**	**0.01**	**0.02**	**0.03**	**0.07**	**0.13**	**0.08**
	intra-center	**0.02**	**0.03**	**0.06**	**0.1**	**0.21**	**0.10**
	inter-center	**0.01**	**0.02**	**0.02**	**0.07**	**0.12**	**0.08**

Table 1: Cross-center/cross-platform mutation detection reproducibility with MuTect2, Strelka2, and SomaticSniper: Overall SNVs: all SNVs in a pair of NGS runs; Repeatable: SNVs defined in “HighConf” and “MedConf” categories; Gray zone: SNVs defined in “LowConf” and “Unclassified” categories; Non-Repeatable: SNVs not defined in above four categories. WES: Average Jaccard scores of WES pairs; WGS: Average Jaccard scores of WGS pairs.

Figure legends

Figure 1. Study design and read quality. **(a)** Study design to capture non-analytical and analytical factors affecting cancer mutation detection. DNA was extracted from either fresh cells or FFPE processed cells and fragmented at three intended sizes. Libraries with various amounts of DNA input (either from random shot gun or exome-capture) were generated with three different library preparation kits and run on WGS and WES in parallel following recommended protocols (see Online Methods). Twelve replicates were performed at six sequencing centers. Three centers prepared WGS and WES libraries in triplicate, while the other three centers prepared a single WGS and WES library (3x3+3). 144 libraries were sequenced on either a HiSeq or NovaSeq instrument. 2 pre-alignments: BFC or Trimmomatic; 3 alignments: BWA, Bowtie, and NovoAlign; 2 post-alignments: GATK or no-GATK were evaluated. Numbers in parentheses represent possible combinations at that level. Further detail of experiment design is depicted in **Suppl. Fig. 1**. **(b)** Read yields (blue), mapping statistics (red), and genome coverage (yellow line) from twelve repeated WGS runs. **(c)** Global Imbalance Value (GIV) of G>T/C>A and T>/A>C mutation pairs in WES and WGS runs. Six centers used a range of time (80 to 300 seconds) for DNA shearing. As a result, insert DNA fragment sizes ranged from 161 bp to 274 bp. **(d)** Distribution of GIV scores of FFPE DNA with four different fixation times (1, 2, 6, and 24 hours), analyzed with WES or WGS. FFX: FFPE on WES platform, FFG: FFPE on WGS platform.

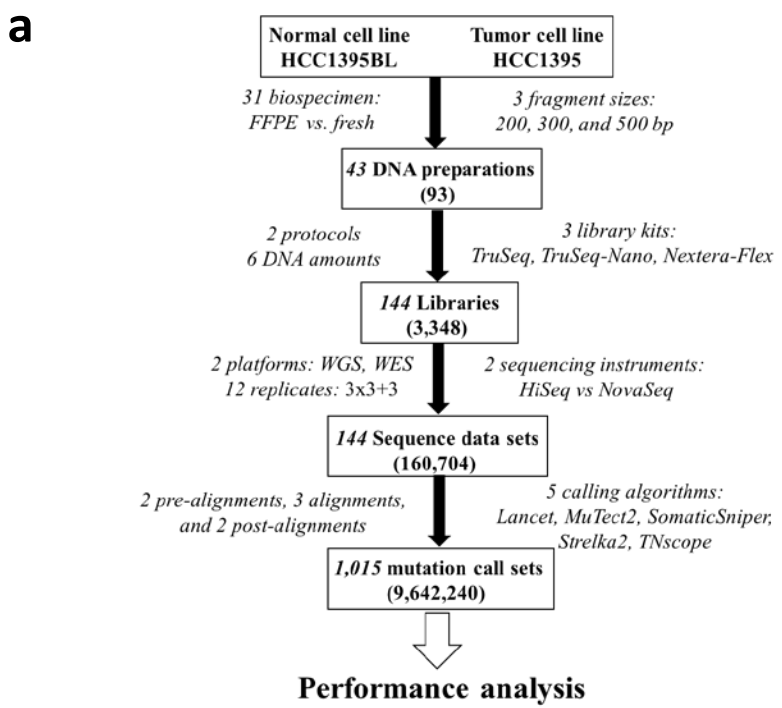
Figure 2. Mutation calling reproducibility. For a-c, the Y-axis: cumulative percentage of overlapping SNVs in a call set; the X-axis: Percent of SNVs in all call sets. The solid lines are the results from WGS and the dotted lines are the results from WES. (**a**) SNV overlaps across 108

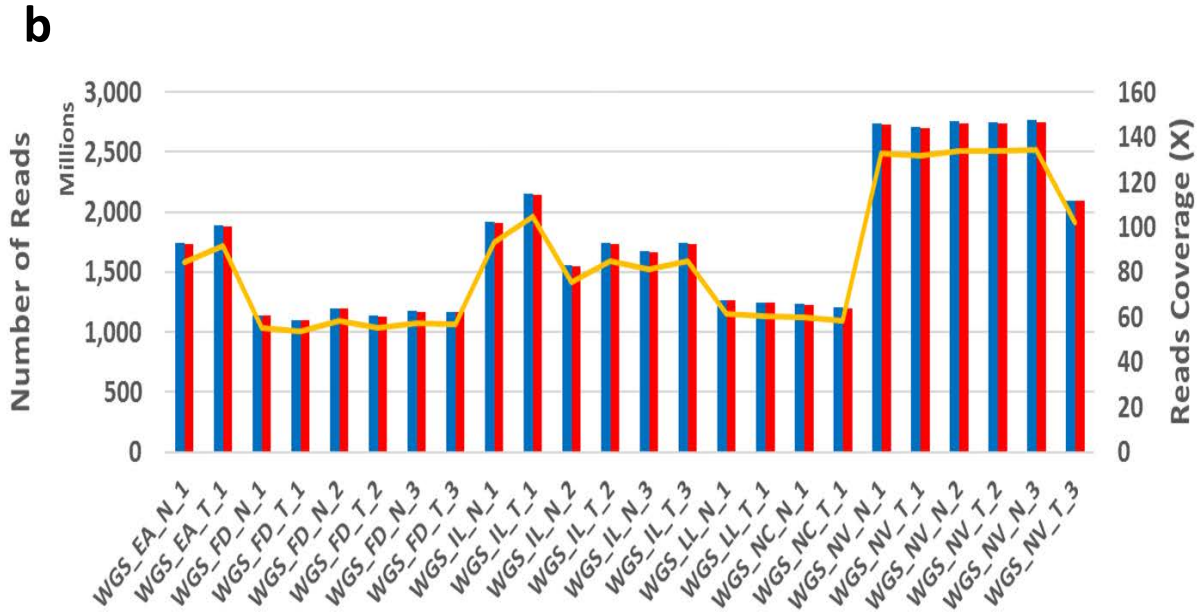
VCF results from **twelve** repeated WGS and WES runs that were analyzed with **three** aligners: BWA, Bowtie2, and NovoAlign, and **three** callers: MuTect2, Strelka2, and SomaticSniper. (**b**) SNV overlaps across **36** VCF results from **twelve** repeated WGS and WES runs analyzed by MuTect2, Strelka2, and SomaticSniper from only the BWA alignments. **(c)** SNV overlaps in **each of the twelve** repeated WES and WGS runs, **respectively,** analyzed by Strelka2 from BWA alignments. The y-axis shows the probability of SNVs that missed in equal to or less than the percentage of the twelve call sets that is depicted at the x-axis. The solid lines are the results from WGS and the dotted lines are the results from WES. Sequencing sites are depicted by colors: magenta for EA, green for FD, blue for IL, yellow for LL, cyan for NC and red for NV. Repeated sequencing runs from a sequencing site are differentiated by the corresponding color gradients. Sequencing sites are depicted by colors: magenta for EA, green for FD, blue for IL, yellow for LL, cyan for NC and red for NV. Repeated sequencing runs from a sequencing site are differentiated by the corresponding color gradients. (**d**) Effect summary of the model for WES or WGS. Effect tests were performed to evaluate the importance of each independent variable in a fixed effect linear model fitted for WES or WGS. F statistics and the corresponding P-value were calculated for the variables. Both P-value and LogWorth (-log10 (P value)) are plotted.

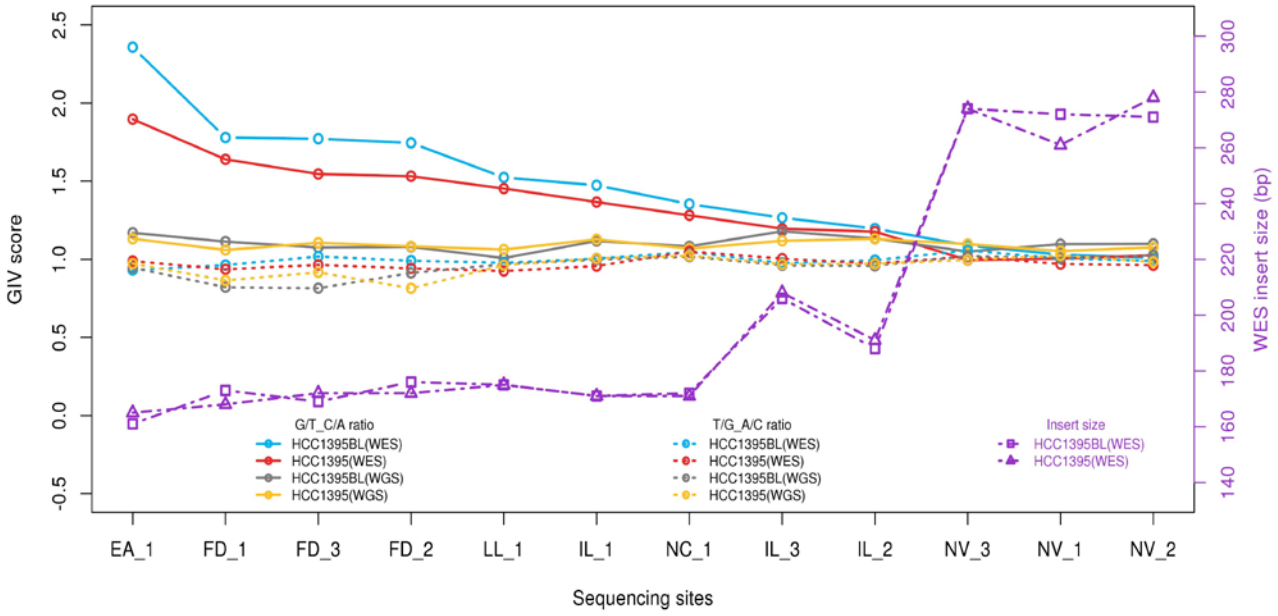
Figure 3. Non-analytical factors affecting mutation calling. **(a)** Caller performance on three library preparation protocols with different DNA inputs: 1, 10, 100, 250, and 1000 ng. **(b)** Performance of MuTect2, Strelka2, and SomaticSniper on fresh DNA and FFPE DNA.

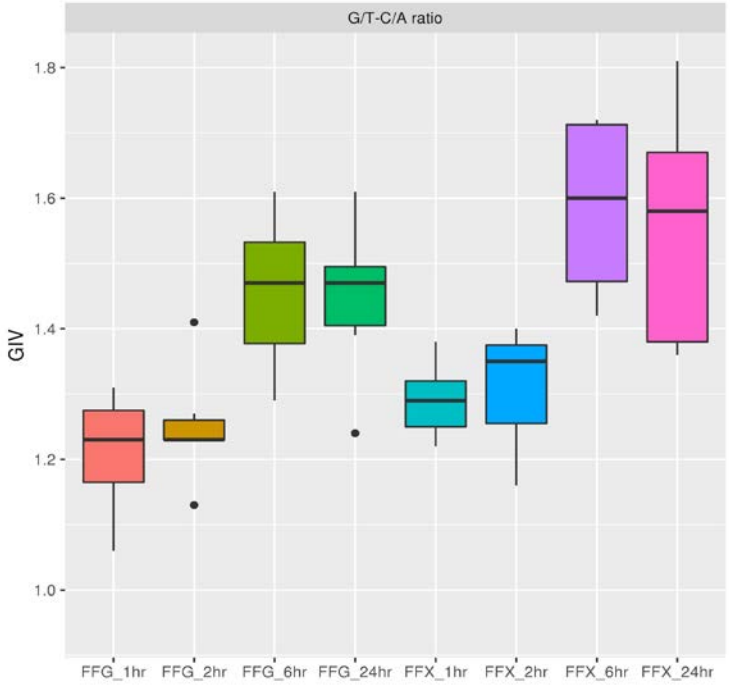
Figure 4. Bioinformatics for enhanced calling. **(a)** Distribution of mutation types called with Strelka2 on BWA alignments of four WES runs preprocessed by Trimmomatic or BFC. FFX: WES run on FFPE DNA, EA_1, FD_1, and NV_1, represented three WES runs with fresh DNA. **(b)** Performance of mutation calling by Strelka2 on three alignments, Bowtie2, BWA, and NovoAlign. The insert figure is a violin plot of MAPping Quality (Mapq) scores from three alignments for an example WGS run. **(c)** Effect of post-alignment processing (indel-realignment + Base Quality Score Recalibration [BQSR]) on mutation calling by MuTect2, Strelka2, and SomaticSniper. **(d)** Effect of tumor purity (20% vs 50%) on five callers: Lancet, MuTect2, Strelka2, TNscope, and SomaticSniper with read coverage from 10X to 300X.

Figure 5. Precision and recall of overlapping SNVs/indels that were supported by biological repeats (library repeats) or analytical repeats (two different callers). Each row or column represents calling results from a WES or WGS run called by one of the three callers from a BWA alignment.

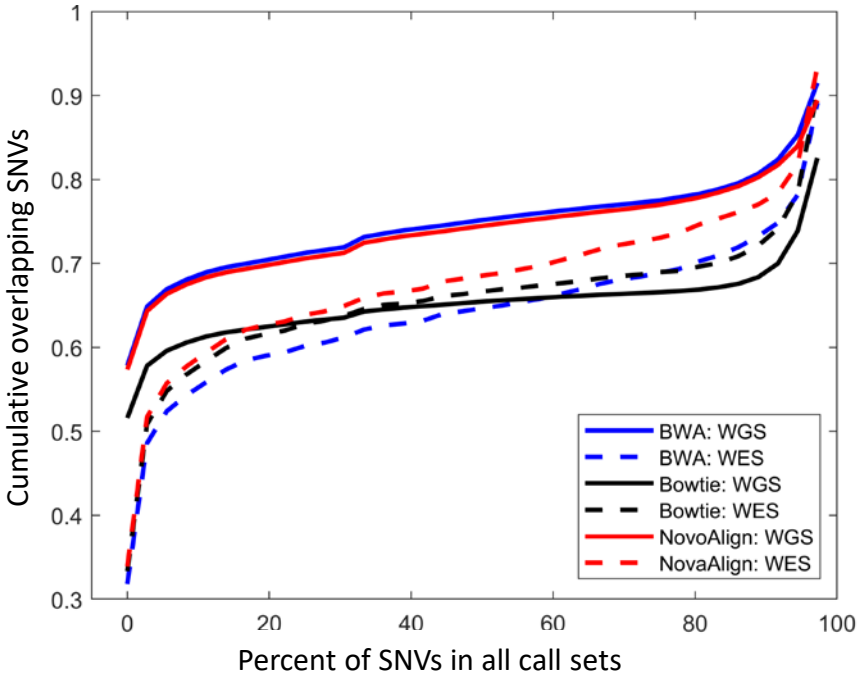

a
Normal cell line HCC1395BL
Tumor cell line HCC1395
31 biospecimen: FFPE vs. fresh
3 fragment sizes: 200, 300, and 500 bp
43 DNA preparations (93)
2 protocols 6 DNA amounts
3 library kits: TruSeq, TruSeq-Nano, Nextera-Flex
144 Libraries (3,348)
2 platforms: WGS, WES 12 replicates: 3x3+3
2 sequencing instruments: HiSeq vs NovaSeq
144 Sequence data sets (160,704)
2 pre-alignments, 3 alignments, and 2 post-alignments
5 calling algorithms: Lancet, MuTect2, SomaticSniper, Strelka2, TNscope
1,015 mutation call sets (9,642,240)
Performance analysis


b
Number of Reads
Millions
Reads Coverage (X)
3,000
2,500
2,000
1,500
1,000
500
0
160
140
120
100
80
60
40
20
0
WGS_EA_N_1
WGS_EA_T_1
WGS_FD_N_1
WGS_FD_T_1
WGS_FD_N_2
WGS_FD_T_2
WGS_FD_N_3
WGS_FD_T_3
WGS_IL_N_1
WGS_IL_T_1
WGS_IL_N_2
WGS_IL_T_2
WGS_IL_N_3
WGS_IL_T_3
WGS_LL_N_1
WGS_LL_T_1
WGS_NC_N_1
WGS_NC_T_1
WGS_NV_N_1
WGS_NV_T_1
WGS_NV_N_2
WGS_NV_T_2
WGS_NV_N_3
WGS_NV_T_3

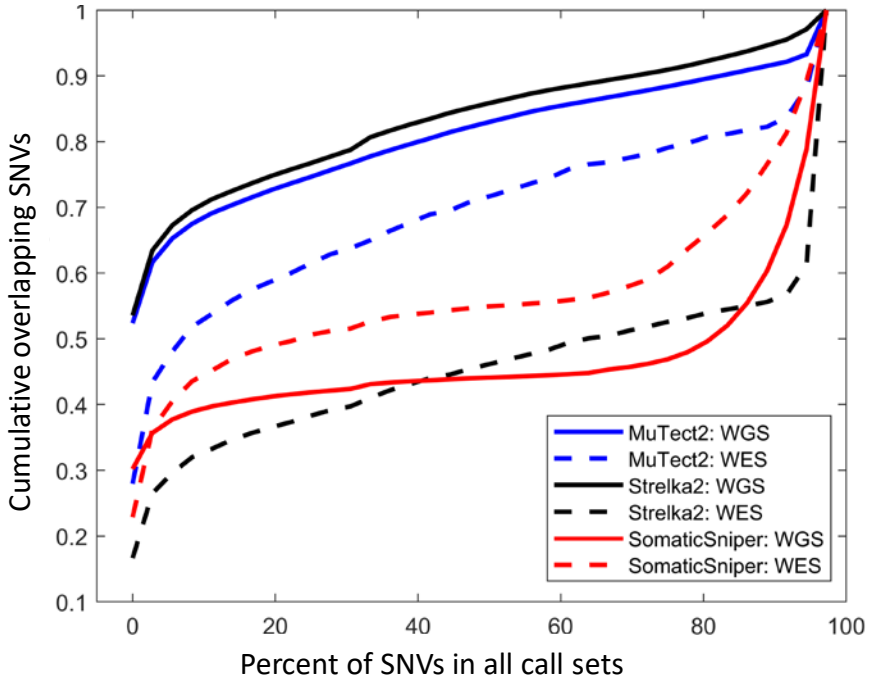

c
GIV score
WES insert size (bp)
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
300
280
260
240
220
200
180
160
140
G/T_C/A ratio
HCC1395BL(WES)
HCC1395(WES)
HCC1395BL(WGS)
HCC1395(WGS)
T/G_A/C ratio
HCC1395BL(WES)
HCC1395(WES)
HCC1395BL(WGS)
HCC1395(WGS)
Insert size
HCC1395BL(WES)
HCC1395(WES)
EA_1
FD_1
FD_3
FD_2
LL_1
IL_1
NC_1
IL_3
IL_2
NV_3
NV_1
NV_2
Sequencing sites


d
G/T-C/A ratio
GIV
1.8
1.6
1.4
1.2
1.0
FFG_1hr
FFG_2hr
FFG_6hr
FFG_24hr
FFX_1hr
FFX_2hr
FFX_6hr
FFX_24hr

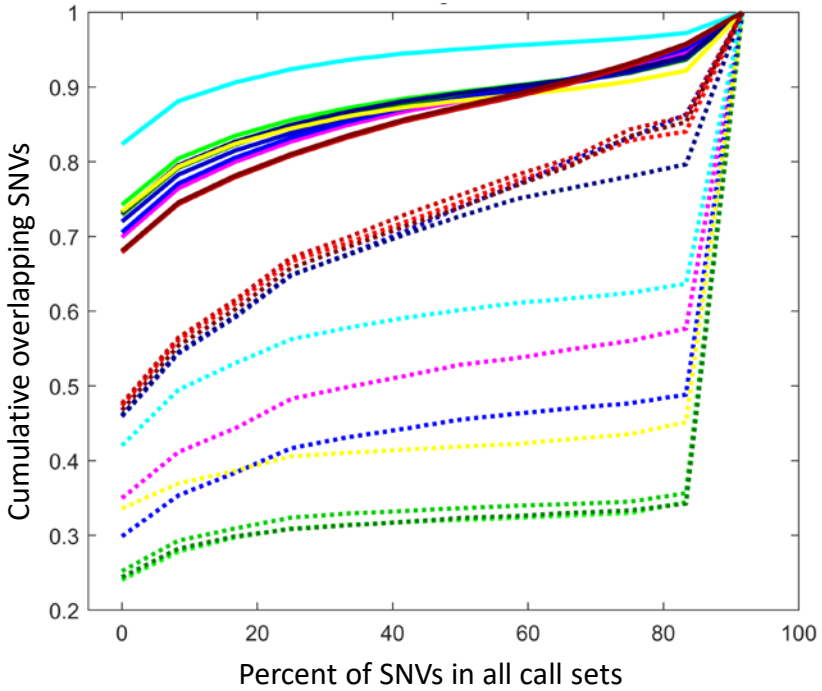
a



b



c



d

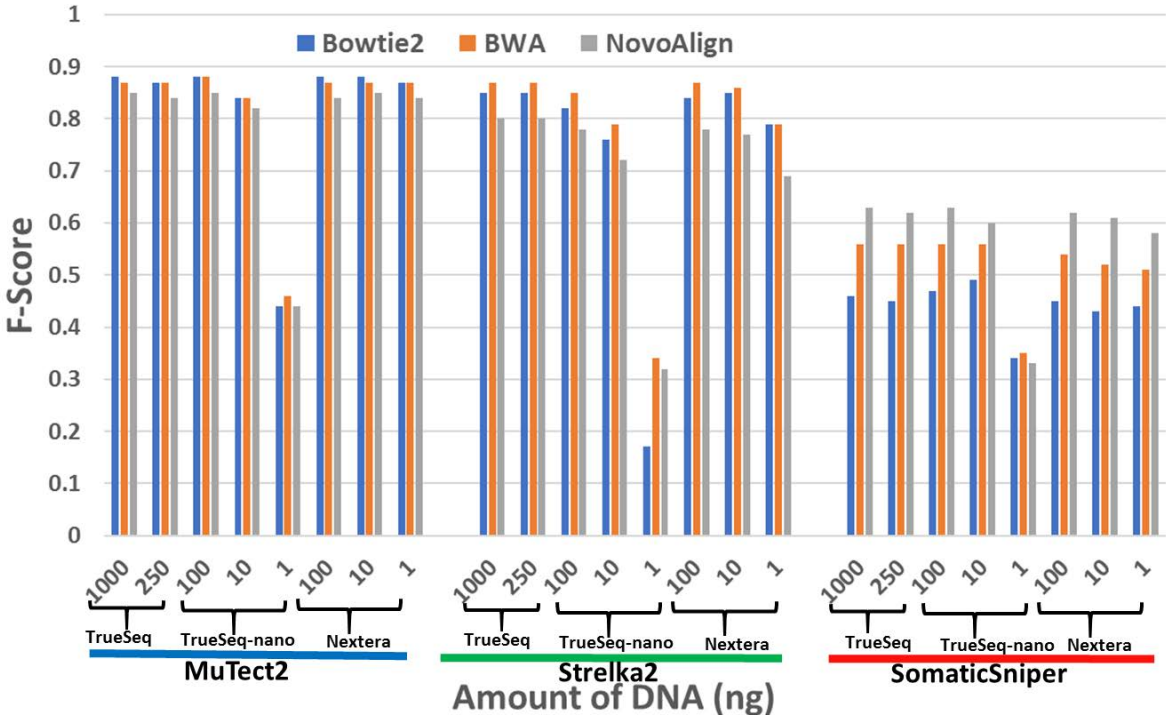
Effect Summary (WES)

Source	LogWorth	PValue
Callers*Mean Coverage Depth	8.044	0.00000
Callers*Percent GC	5.469	0.00000
Machine model*Callers	4.521	0.00003
Machine model	4.476	0.00003 ^
Callers*GIV(G>T)	3.639	0.00023
GIV(G>T)	3.619	0.00024 ^
Callers*Percent Non-duplicated Reads (Mapped Trimmed)	3.183	0.00066
Callers	3.131	0.00074 ^
Percent Non-duplicated Reads (Mapped Trimmed)	2.896	0.00127 ^
Percent GC	2.254	0.00557 ^
Median Insert Size	1.927	0.01183
Mean Coverage Depth	0.682	0.20805 ^
GIV(T>G)	0.494	0.32054

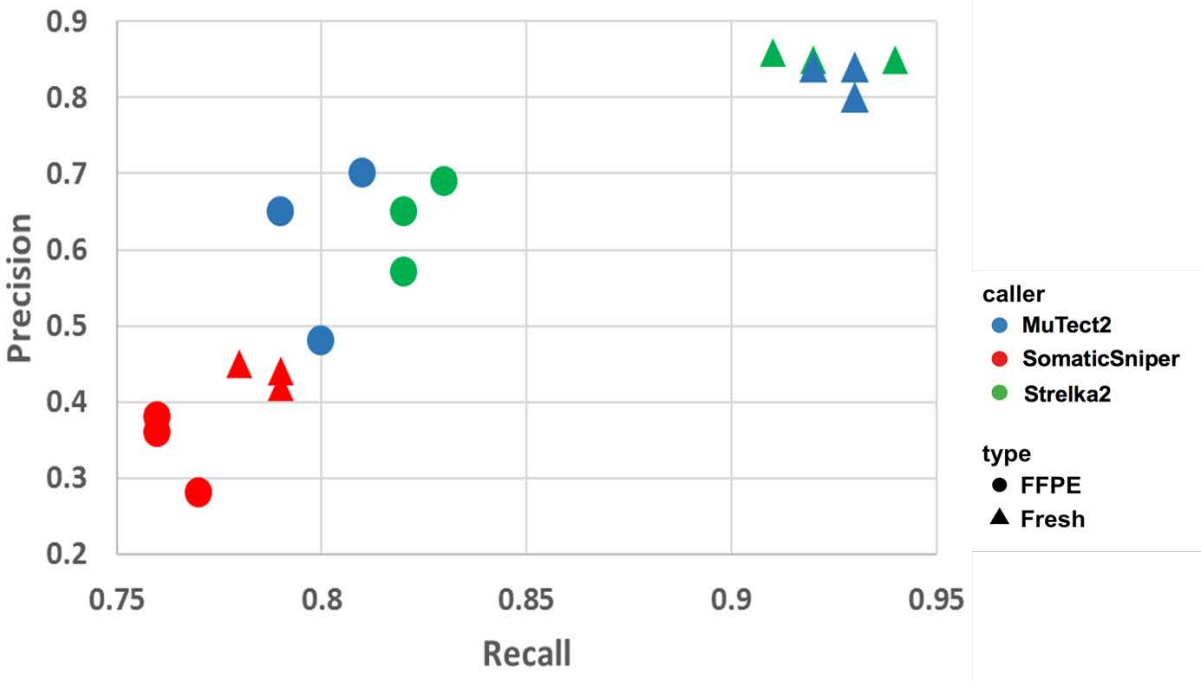
Effect Summary (WGS)

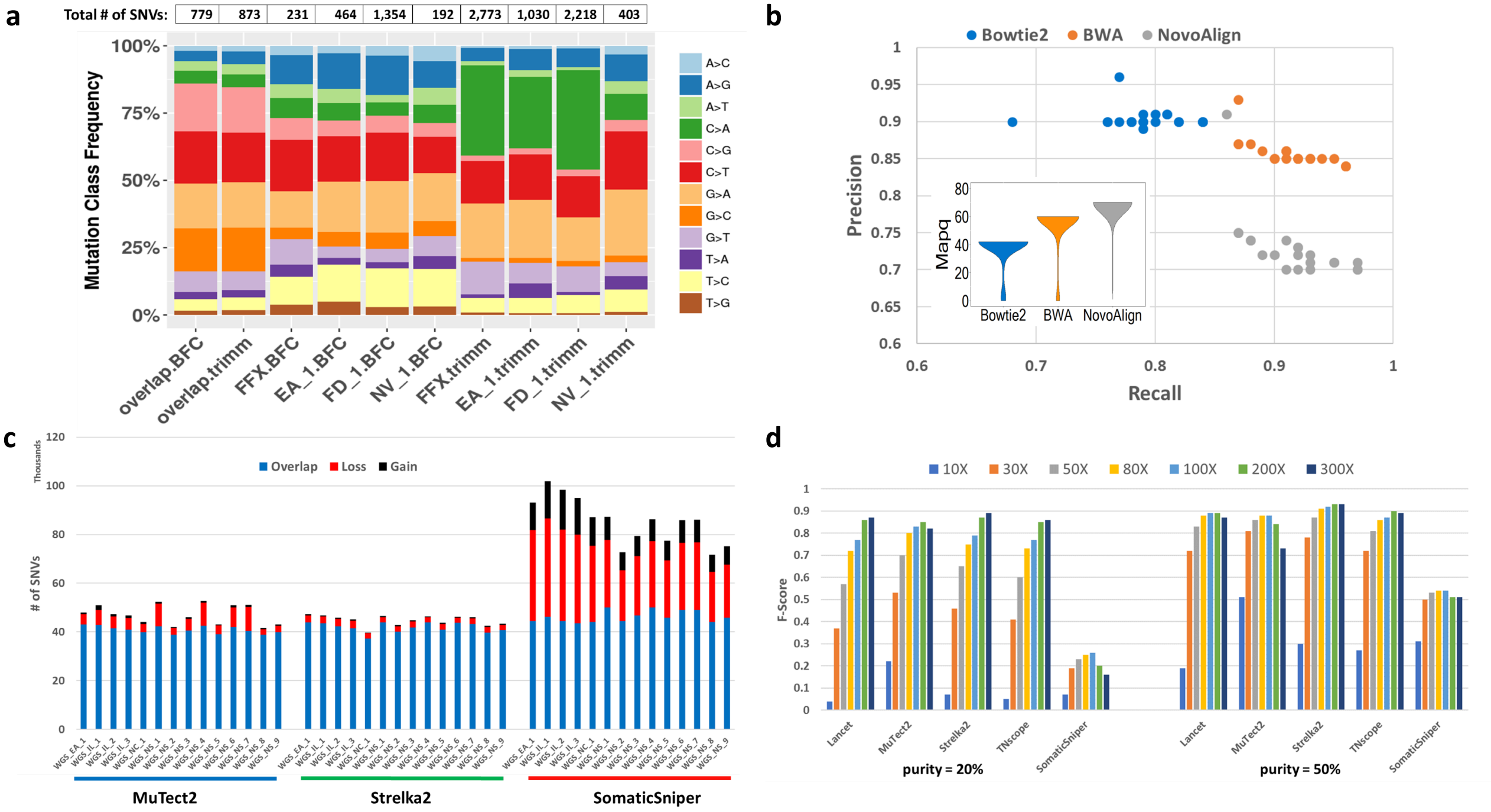
Source	LogWorth	PValue
Callers	27.424	0.00000
Mean Coverage Depth	2.685	0.00206
Percent Non-duplicated Reads (Mapped Trimmed)	1.430	0.03719
Percent GC	0.587	0.25903
GIV(G>T)	0.472	0.33711
Median Insert Size	0.092	0.80852
Machine model	0.036	0.92058
GIV(T>G)	0.020	0.95563

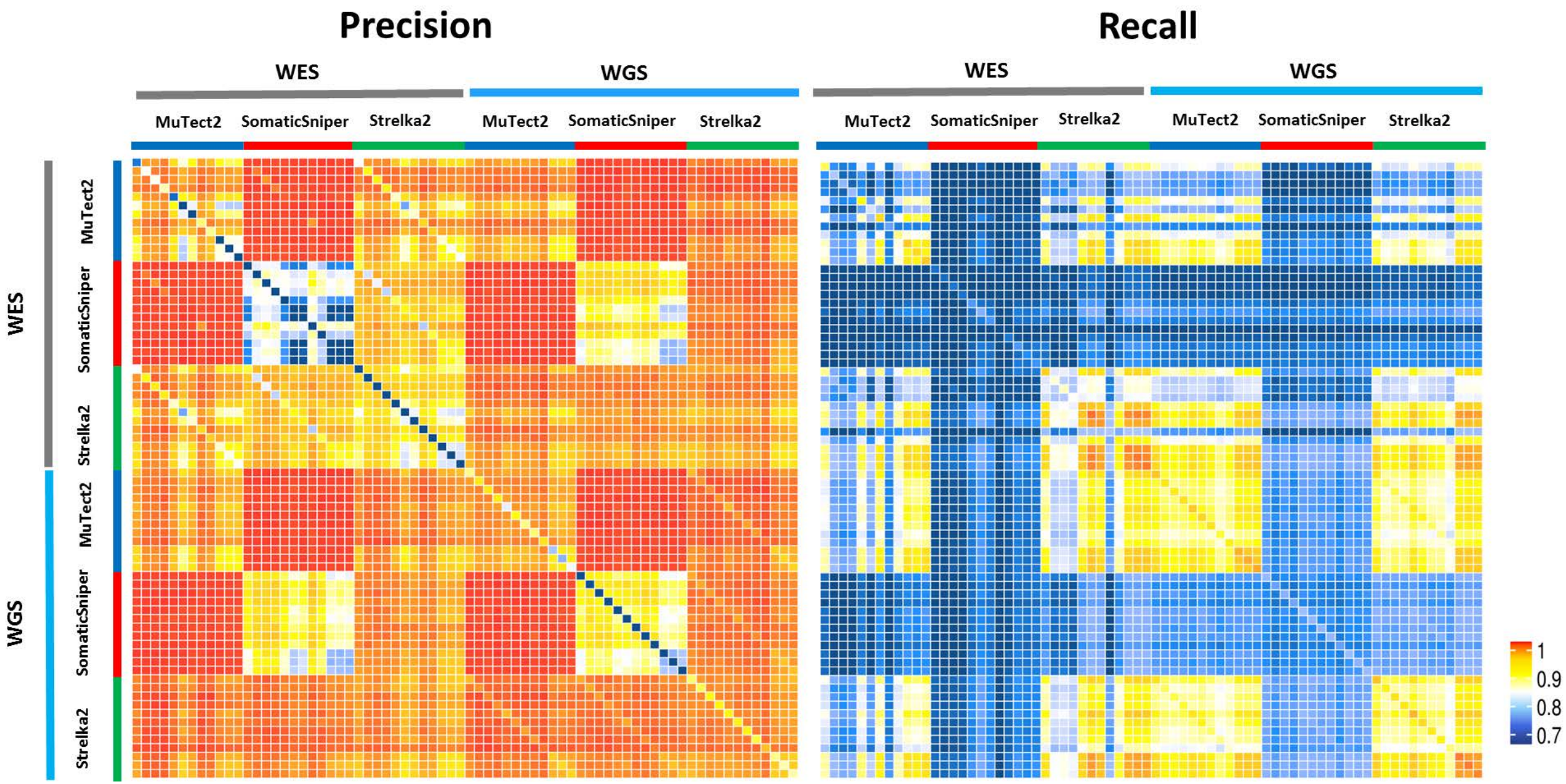
a


Bowtie2
BWA
NovoAlign
F-Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1000
250
100
10
1
100
10
1
TrueSeq
TrueSeq-nano
Nextera
MuTect2
Strelka2
SomaticSniper
Amount of DNA (ng)

b


Precision
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.75
0.8
0.85
0.9
0.95
Recall
caller
MuTect2
SomaticSniper
Strelka2
type
FFPE
Fresh


a
Total # of SNVs: 779 873 231 464 1,354 192 2,773 1,030 2,218 403
Mutation Class Frequency
100%
75%
50%
25%
0%
overlap.BFC
overlap.trimm
FFX.BFC
EA_1.BFC
FD_1.BFC
NV_1.BFC
FFX.trimm
EA_1.trimm
FD_1.trimm
NV_1.trimm
A>C
A>G
A>T
C>A
C>G
C>T
G>A
G>C
G>T
T>A
T>C
T>G
b
Bowtie2
BWA
NovoAlign
Precision
Recall
Mapq
c
Thousands
Overlap
Loss
Gain
# of SNVs
MuTect2
Strelka2
SomaticSniper
d
10X
30X
50X
80X
100X
200X
300X
F-Score
Lancet
MuTect2
Strelka2
TNscope
SomaticSniper
purity = 20%
purity = 50%


Precision
Recall
WES
WGS
MuTect2
SomaticSniper
Strelka2
1
0.9
0.8
0.7