Journal of Advanced Research 44 (2023) 161–172

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Article

Consistency and reproducibility of large panel next-generation sequencing: Multi-laboratory assessment of somatic mutation detection on reference materials with mismatch repair and proofreading deficiency

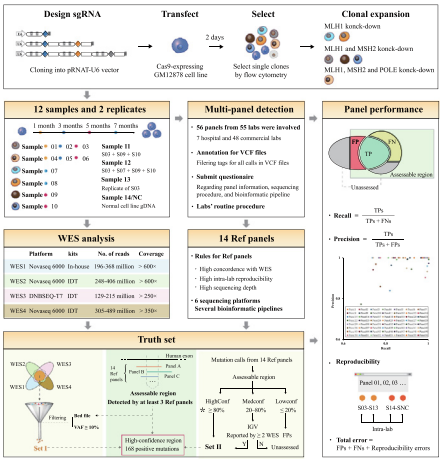
Duo Wang [a,b,c,1], Yuanfeng Zhang [a,b,c,1], Rui li [a,b,c], Jinming Li [a,b,c,*], Rui Zhang [a,b,c,*]

[a] National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital/National Center of Gerontology, P. R. China
[b] Graduate School of Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, P. R. China
[c] Beijing Engineering Research Center of Laboratory Medicine, Beijing Hospital, Beijing, P. R. China

HIGHLIGHTS

- Cell lines with MMR and proofreading deficiency were constructed using the CRISPR technology for reference samples.
- Paired tumor–normal reference samples close to clinical specimens were prepared.
- A reliable process to determine high-confidence region and positive variants was designed.
- A multi-laboratory study was conducted to evaluate reproducibility and accuracy of participating panels.
- Potential sources of false-discovery were explored for assay optimization.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 December 2021
Revised 16 March 2022
Accepted 27 March 2022
Available online 31 March 2022

Keywords:
Precision oncology
Somatic mutation

ABSTRACT

Introduction: Clinical precision oncology increasingly relies on accurate genome-wide profiling using large panel next generation sequencing; however, difficulties in accurate and consistent detection of somatic mutation from individual platforms and pipelines remain an open question.
Objectives: To obtain paired tumor–normal reference materials that can be effectively constructed and interchangeable with clinical samples, and evaluate the performance of 56 panels under routine testing conditions based on the reference samples.
Methods: Genes involved in mismatch repair and DNA proofreading were knocked down using the CRISPR-Cas9 technology to accumulate somatic mutations in a defined GM12878 cell line. They were

Abbreviations: AF, allele frequency; FN, false negatives; FP, false positives; LOD, limit of detection; MMR, mismatch repair; NGS, next-generation sequencing; POLE, DNA polymerase epsilon; TMB, tumor mutation burden; WES, whole-exome sequencing; WGS, whole-genome sequencing; qRT-PCR, quantitative reverse transcription polymerase chain reaction; EQA, external quality assessment; SNV, single-nucleotide variants; IGV, Integrative Genomics Viewer.

Peer review under responsibility of Cairo University.

* Corresponding authors at: National Center for Clinical Laboratories, Beijing Hospital, No.1 Dahua Road, Dongdan, Beijing 100730, P. R. China.
E-mail addresses: jmli@nccl.org.cn (J. Li), ruizhang@nccl.org.cn (R. Zhang).

[1] These authors contribute equally to the article.

https://doi.org/10.1016/j.jare.2022.03.016
2090-1232/

Targeted panel sequencing
Reference material

used as reference materials to comprehensively evaluate the reproducibility and accuracy of detection results of oncopanels and explore the potential influencing factors.
Results: In total, 14 paired tumor–normal reference DNA samples from engineered cell lines were prepared, and a reference dataset comprising 168 somatic mutations in a high-confidence region of 1.8 Mb were generated. For mutations with an allele frequency (AF) of more than 5% in reference samples, 56 panels collectively reported 1306 errors, including 729 false negatives (FNs), 179 false positives (FPs) and 398 reproducibility errors. The performance metric varied among panels with precision and recall ranging from 0.773 to 1 and 0.683 to 1, respectively. Incorrect and inadequate filtering accounted for a large proportion of false discovery (including FNs and FPs), while low-quality detection, cross-contamination and other sequencing errors during the wet bench process were other sources of FNs and FPs. In addition, low AF (<5%) considerably influenced the reproducibility and comparability among panels.
Conclusions: This study provided an integrated practice for developing reference standard to assess oncopanels in detecting somatic mutations and quantitatively revealed the source of detection errors. It will promote optimization, validation, and quality control among laboratories with potential applicability in clinical use.

Introduction

Large panel (usually more than 1 Mb) next-generation sequencing (NGS) is a finely balanced approach between the whole exome and hotspot detection for oncology. It concentrates on genomic regions where pathogenic mutations appear and reduce the turnaround time and cost. Furthermore, it is now increasing used for therapeutic target identification [1,2], treatment decision [3], and prognosis judgment [3] by detecting genome-wide mutations in various malignant cancers. Moreover, oncopanels allow the estimation of tumor mutation burden (TMB) with a high correlation with whole exome sequencing (WES), which is an important biomarker for prescribing immune checkpoint inhibitors [4,5]. However, clinically oriented targeted assays are error prone at every step, leading to inaccurate and inconsistent results among laboratories [6–8]. To promote the application and standardization of targeted sequencing for clinical use, validation and quality control using well-characterized and paired tumor–normal reference samples should be conducted.

Currently available reference materials are not perfect but have been identified and utilized in different ways for standardization studies. First, the spike-in method to synthesize variant-containing fragments and gene-editing method to introduce mutation point by point are both uneconomic and not interchangeable with patient samples [9]. Second, the germline method, which entails using the existing and easily identified mutations with high allele frequencies (most of them are germline mutations) in tumor cell lines, is used to mimic somatic mutations and could offer abundant variants. However, it is not associated with the occurrence and development of cancers [10], and has questionable applicability in clinical somatic mutation research. Third, the available and certified tumor cell lines either have no paired normal cell lines or have limited mutations and allele frequency (AF) range [11]. More importantly, as most laboratories develop their methodologies based on these cell lines, it is familiar and biased to launch a blinded external quality assessment (EQA) with them. In addition, reliable methods for establishing reference datasets on other kinds of reference materials are challenging, because some somatic mutations with low AF are difficult to validate using conventional droplet digital PCR (ddPCR) and Sanger sequencing, as well as low-coverage WES and whole-genome sequencing (WGS).

In the present study, we developed a series of paired tumor–normal reference materials using effective building and feasible validating methods, which were based on easily available normal cell lines that were close to real-world clinical tumors. To increase somatic variants in normal cell lines, we knocked down MutL homolog 1(*MLH1*) and MutS homolog 2*(MLH2)*, two key genes in the mismatch repair (MMR) pathway, and the proofreading-associated DNA polymerase epsilon (*POLE*) gene using the CRISPR-Cas9 technology. The highly conserved MMR system exists in prokaryotes and eukaryotes, playing a critical role in repairing mismatched nucleotides caused by replicative errors and physical and chemical damage [12]. Pol ε, the protein product of the *POLE* gene, possesses 3 → 5 exonuclease activity, which recognizes and removes nucleotides erroneously incorporated during DNA replication [13]. Deficiencies in the MMR and proofreading systems have been reported to lead to genome instability and somatic hypermutation, which have been observed in several types of cancer [14,15]. Thus, it is practical to generate hundreds of somatic mutations and set up tumor–normal pairs by knocking down the *MLH1*, *MLH2*, and *POLE* genes in normal cell lines. We then practiced an integrated method to easily determine somatic variants [16], and evaluated the performance of oncopanels in somatic mutation detection simultaneously. WES results and reliable oncopanel results selected from those with high sequencing depth, perfect reproducibility, and superior concordance with WES results, were used to determine positive somatic variants in our reference samples. Our study aimed at developing a reference standard to assess diverse oncopanel products with respect to performance metrics such as precision, recall, and F1-score, and tracing the source of detection errors for assay optimization, which will aid in delivering reliable NGS testing for clinical application.

Materials and methods

Sample panel design and preparation

Six DNA repair or proofreading gene knockdown clonal cell lines constructed using the CRISPR-Cas9 technology and validated using quantitative reverse transcription polymerase chain reaction (qRT-PCR) and western blot were used for reference sample preparation (Supplementary Methods). All cells were passaged regularly and expanded until expected stocks were generated. Genomic DNA from six clones and the original Cas9-expressing GM12878 cell line were extracted, and a panel of 15 DNA samples was prepared by the NCCL (Beijing, China). The 15 reference samples and corresponding cells are as follows: S01–S03 were from double-gene (*MLH1* and *MSH2*) knockdown clonal cells after culturing over a defined period of time (1, 3, and 5 months); S04–S06 were from

three-gene (*MLH1*, *MSH2* and *POLE*) knockdown clonal cells after culturing over a defined period of time (1, 3, and 5 months); S07 and S08 were from two different *MLH1* knockdown clonal cells after culturing for 7 months; S09 and S10 were from two different double-gene (*MLH1* and *MSH2*) knockdown clonal cells after culturing for 5 months; S11 was a mixture of S03, S09, and S10, whereas S12 was a mixture of S03, S07, S09, and S10 at equal proportions (Table S1). SNC was prepared using the original Cas9-expressing GM12878 cell line as a normal matched sample to filter out irrelevant mutations. S13 and S14 were the replicates of S03 and SNC, respectively, which were designed to assess the intra-reproducibility of panels with twice assays on the same sample. All the samples dispensed as 30 μL aliquots into 200 μL thin-wall polypropylene PCR tubes with a concentration of 20–30 ng/μL and stored at −20 °C.

Whole-exome sequencing

The 13 samples, namely, S01–S12 and SNC, were sent to four centers for WES analysis. Sequencing platforms, protocols, and bioinformatic pipelines are detailed in the Supplementary Methods. All assays were performed according to the manufacturer's protocol.

Participating laboratories

The prepared samples were shipped on ice to 55 clinical laboratories. All the laboratories were assigned the same coded samples, and detailed instructions for storage conditions and assay procedures were provided. Laboratories were required to perform the detection using their routine procedures, and the raw reads were mapped to the human reference genome (hg19). Laboratories were required to submit their results, including the variants and corresponding allele frequencies. All variants were reported according to the Human Genome Variation Society guidelines. Because a variant might have different descriptions across different transcripts, we recommended that the participants use the mane-selected transcripts in the ClinVar database. Questionnaires were sent to obtain information regarding their panel parameters (such as size, number of genes, detectable ranges, minimum limit of detection), sequencing procedures (such as the platforms and kits for library construction and sequencing), databases and bioinformatics tools employed, and assay-specific quality metrics (such as minimum coverage thresholds, mapping qualities, and Q scores). In addition, laboratories were required to submit BAM and VCF files for all 15 samples. Because of the different bioinformatic filtering rules in somatic mutation calling, nine defined filtering tags were required to annotate all mutation calls in the column "FILTER" in the VCF files, using vcftools [17] or any other similar software: "PASSALL," "LOWQUAL," "NONCODING," "MATCHED," "GERMLINEDB," "SNP," "OUTSCOPE," "NOCS," and "OTHER." [18]. The details of these filtering tags are provided in the Supplementary Methods.

Establishment of the truth set

WES usually performs worse than high-depth panel products in detecting somatic variants with low AF. For 12 non-repetitive samples, S01–S12, the truth set consists of two parts: set I comprised mutations with VAF $\geq$ 10% and set Ⅱ included mutations with VAF < 10%. WES results were used to determine set Ⅰ, and 14 panels (termed Ref panels) with high sequencing depth were selected to determine set II based on the rules listed in Supplementary Methods.

Data analysis

To assess participant performance effectively, a set of scoring rules was established previously. Mutation results in the assessable region were compared to the truth set to identify false negatives (FNs) or false positives (FPs). Considering the variability of mutation detection around the limit of detection (LOD), the mutation results with AF $\geq$ 5% differed from the truth set were defined as false FPs and FNs. To verify our findings, all false mutation results, including FPs and FNs, were visually reviewed using the Integrative Genomics Viewer (IGV). In addition, discordant results between any pair of replicates (S03–S13 and S14–SNC) were considered as reproducibility errors. The performance of the panels was classified as either acceptable or improvable by penalty, which was determined using the following formula: penalty = FPs + FNs + 2 × Reproducibility Errors. The performance of panel with the penalty smaller than the median of all panels was considered acceptable, otherwise it was considered improvable.

Precision (positive predictive value) was calculated as follows: mutations observed in results of both the panels and truth sets (true positives; TPs) divided by total mutations called by panels (TPs and FNs).

$$\text{Precision} = \frac{\text{TPs}}{\text{TPs} + \text{FPs}} = \frac{\text{concordant mutation between panel and truth sets}}{\text{mutation called by panel}}$$

Recall (sensitivity) was calculated as follows: mutations observed in the results of both panel and truth sets (TPs) divided by total mutations in the truth set (TPs and FNs).

$$\text{Recall} = \frac{\text{TPs}}{\text{TPs} + \text{FNs}} = \frac{\text{concordant mutation between panel and truth sets}}{\text{mutation in truth set}}$$

Reproducibility was determined as the ratio of concordant calls to the unique calls of both duplicate samples.

$$\text{Reproducibility} = \frac{\text{concordant}}{\text{concordant} + \text{discordant}}$$

Bedtools were used for the overlap calculation. The scripts for truth set analysis, panel performance assessment, and other analysis pipelines were written in Python.

Results

Generation of DNA repair and proofreading deficient cell lines to increase mutations

To achieve a spontaneous increase of variants in cell lines, we introduced loss-of-function mutations in the *POLE*, *MLH1*, and *MSH2* genes using the CRISPR-Cas9 technology. Briefly, three sgRNA expressing vectors (Table S2) were transfected into Cas9-expressing GM12878 cell line to elicit DNA proofreading and repair deficiency. Six stable single clones were selected for continuous subculture, including two *MLH1* knockdown clones, three double-gene (*MLH1* and *MSH2*) knockdown clones, and one three-gene (*MLH1, MSH2* and *POLE*) knockdown clone (Fig. 1A). Finally, gene inactivation was verified using qRT-PCR, revealing a significant reduction of *MLH1, MSH2* and *POLE* mRNA expression in the six clonal cells (Fig. 2A). Western blot analysis demonstrated the loss or decrease in the protein expression of MLH1, MSH2 and POLE (Fig. 2B).

Furthermore, for sample preparation, we passaged the six clones for up to seven months to allow cells to accumulate

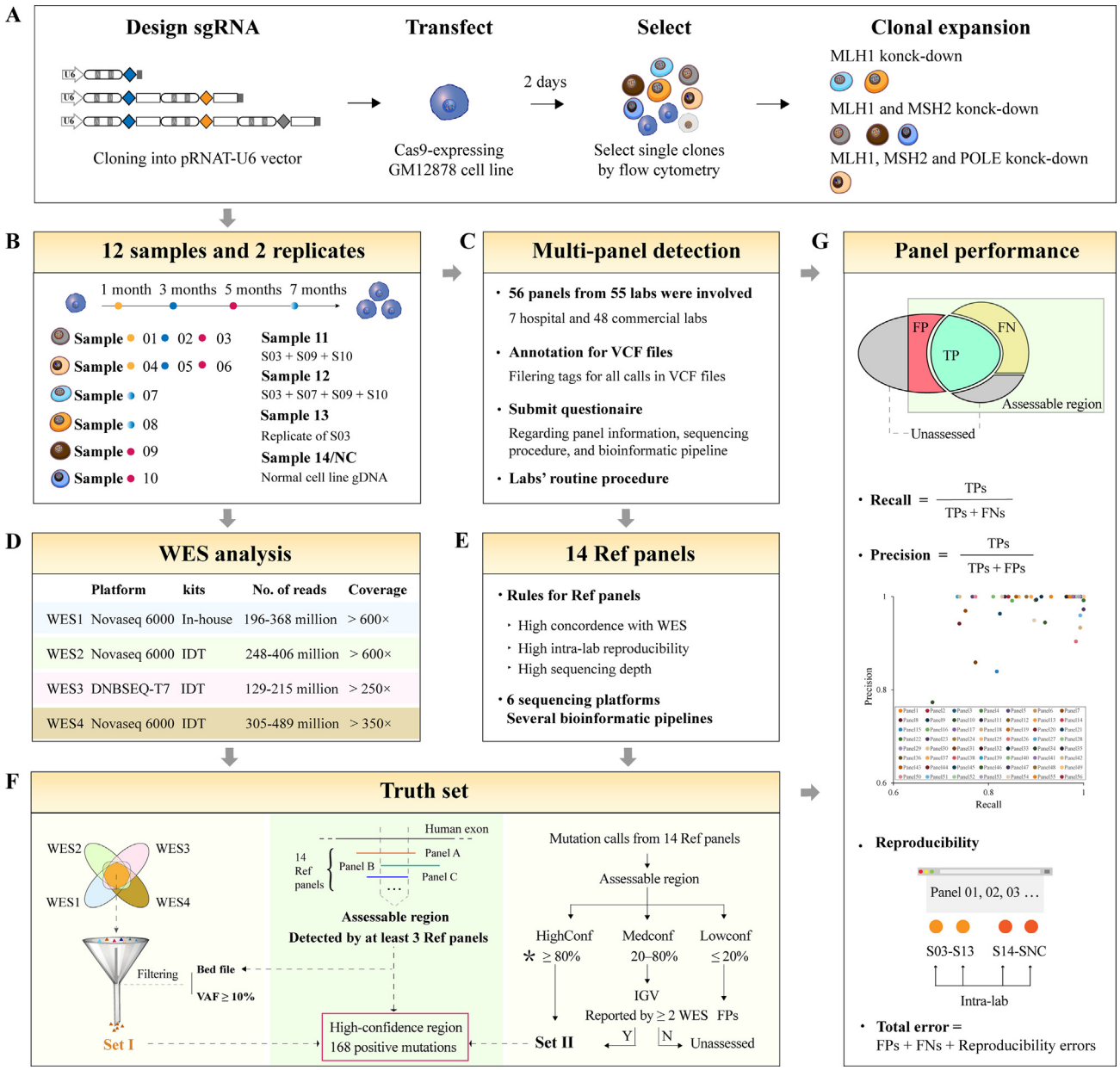


Fig. 1. Comprehensive study design for assessing the analytical performance of multiple targeted panel sequencing technologies. (A) Three tRNA-sgRNA plasmids were used for targeting DNA repair and proofreading genes, and six clones with *MLH1, MSH2* or *POLE* gene knock-down were selected for sample preparation. The ellipse represents the human tRNA sequence, diamonds with different color represent different gRNA sequence, the white rectangle represents sgRNA scaffold sequence, and the gray rectangle represents Pol III terminator sequence. The cells with different colors indicate different single clones. (B) A total of 14 reference samples and one paired normal sample were prepared as shown in Methods. (C) A total of 56 panels were involved, and were required to perform the detection using routine procedure and submit mutation results and questionnaire regarding panel information. (D) Basic information of four WES analysis is listed in embedded table. See Methods for detailed information. (E) Fourteen panels termed "Ref panels" were selected according to the detailed rules in Methods, the mutation results of which were used to determine the positive mutations for truth set. (F) The rules for generation of a high-confidence region and the truth set. The truth set Ⅰ comprised mutations with AF ≥ 10% reported by all four WES, and truth set Ⅱ were determined by 14 Ref panels. * indicates the concordance of reported mutation results among detectable Ref panels. The lines with different color represent detection ranges of the Ref panel or human exon region. (G) Mutation results were subjected for performance assessment with respect to recall, precision, and reproducibility.

sufficient mutations. We then selected cells frozen after different periods of culturing for DNA isolation to prepare samples S1–S10 (Table S1). S11 and S12 were mixed, as indicated in the Materials and Methods section. All the samples were subjected to WES analysis to identify the somatic mutations that accumulated during culture (Fig. 1D). In general, the mutation increased significantly in the first month. Cells containing approximately 70–240 non-synonymous somatic mutations were used for reference samples preparation to assess the performance of the targeted NGS panels (Fig. 2C).

The 56 NGS panels from 55 laboratories were included in the study

In total, 57 reports were obtained from 56 laboratories before the cutoff date. Among these responses, one laboratory reported that their panel size was 32 Mb, which was considered as WES and therefore not included. Consequently, 7 hospital laboratories and 48 commercial laboratories with 56 targeted gene panels at different stages of development were included in this study (Fig. 1B).

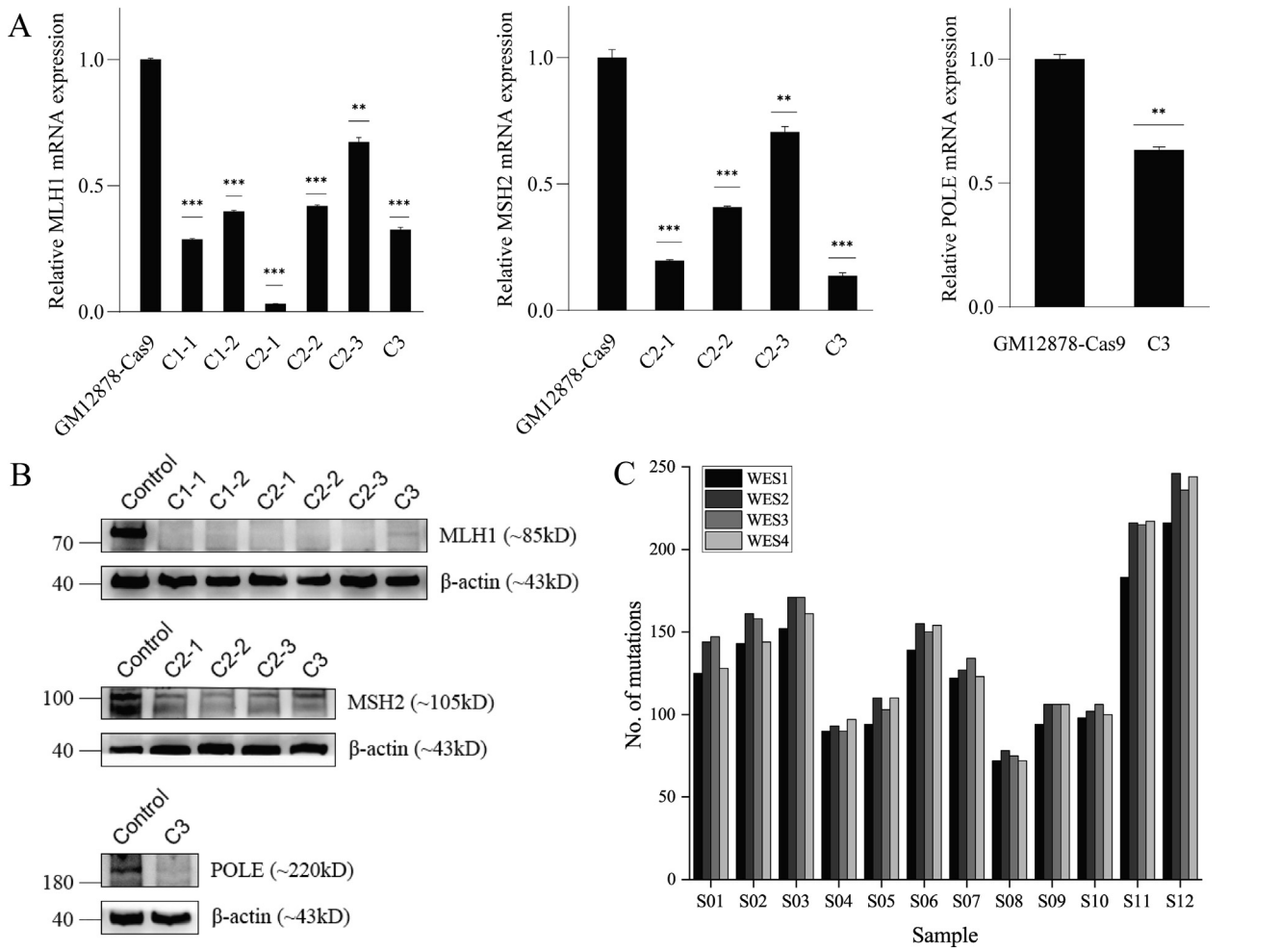


Fig. 2. Generation of DNA repair and proofreading gene knock-downs in human Cas-expressing GM12878 cell line. (A) qRT-PCR for *MLH1*, *MSH2* and *POLE* in the selected clones and control cell line (indicative of original GM12878-Cas9 cell line). Expression was normalized to GAPDH. Mean and SD (error bars) of n = 3 independent experiments are indicated. Asterisks indicate significant differences in mRNA expression between six knock-down cells and GM12878-Cas9 cell line. (two-sided *t*-test; ** for P < 0.01, *** for P < 0.001) (B) Western blot analysis of MLH1, MSH2 and POLE expression in selected clones and control cell line (indicative of original GM12878-Cas9 cell line). β-actin was used as a loading control. (C) The number of non-synonymous somatic mutations from 4 WES analysis in 12 samples from 6 selected clones.

All laboratories submitted the questionnaires (regarding panel information, sequencing procedure, and mutation analysis tools), mutation results and detectable range as required. The panels information was summarized in and detailed in Table S3. The sequencing quality control metrics of panels in 14 tumor samples and a matched normal sample were summarized in Table S4. The exon region size of these gene panels ranged from 0.84 to 2.95 Mb, and the number of genes detected by the gene panels ranged from 162 to 1460. The panel of participants used different NGS approaches. Most panels (42/56, 75%) used Illumina sequencing platforms, including the NovaSeq 6000 System (34/56, 60.7%), NextSeq 550 (7/56, 12.5%), NextSeq 2000 (1/56, 1.8%), HiSeq (1/56, 1.8%) and NextSeq 500 (1/56, 1.8%). Other platforms shared the remaining panels, such as MGISEQ 2000 (MGI Tech, Shenzhen, China) (7/56, 17.9%), NextSeq CN500 (Berry Genomics, Hangzhou, China) (3/56, 5.4%), DNBSEQ-T7 (MGI Tech) (1/56, 1.8%), and GENETRON S2000 (GENETRON, Beijing, China) (1/56, 1.8%). For target enrichment, all the panels used the hybrid capture method. The LOD of the participating panels ranged from 0.5% to 5%, and the most used LOD were 1% (19/56, 33.9%) and 5% (13/56, 23.2%). Seven panels used a lower cutoff for mutations occurring in hotspot genes compared to those in non-hotspot genes. For example, three of them used 1% for mutations occurring in hotspot genes, 2.5% for single-nucleotide variants (SNVs) and 2% for insertion and deletion variants (indels) occurring in non-hotspot genes. In addition, bioinformatic pipelines varied notably among panels, as indicated by the fact that several different software packages, including those developed in-house, were used for read mapping, variant detection, filtering, and annotation (Table 1). All laboratories claimed that the reported results met the internal quality control standards. The mean Q30 of all panels ranged from 85.7% to 99.9% for all samples except for one panel with a Q30 of 77.2%. The average effective sequencing depth of all panels ranged from 100x to 3415x for SNC, and ranged from 552x to 6173x for all tumor samples (Table S4).

Truth set comprising 168 somatic mutations in a high-confidence region was generated

According to the rules for generating the truth set listed in the Materials and Methods, including high consistency with WES results with AF greater than 10% and reproducibility across two pairs of replicates, 14 panels (termed the "Ref panels") from 13 laboratories were selected for determining positive variants (Fig. 1E). A variety of sequencing platforms and bioinformatic tools were used by these 14 panels (Table S3). In brief, 14 panels used sequencing platforms, including NovaSeq 6000, NextSeq 550, HiSeq 4000, NextSeq CN500, MGISEQ 2000, and GENETRON

Table 1
Methodological variance among participating panels.

56 panels from 55 laboratories			
Methodological variance	N	Bioinformatic pipeline variance	N
1. Sequencing platform		1. Quality Control	
NovaSeq 6000	34	Fastp	26
NextSeq 550	7	FastQC	16
MGISEQ 2000	7	BRQC	3
NextSeq CN500	3	FQstat	1
NextSeq 2000	1	Trimmomatic	1
HiSeq 4000	1	In-house	8
Nextseq 500	1	2. Reads Mapping ‡	
DNBSEQ-T7	1	BWA	54
GENETRON S2000	1	Speedseq	1
2. Panel size (Mb)		Sentieon	1
< 1.00	1	3. Variant detection ‡	
1.00–2.00	24	GATK	2
2.00–3.00	34	GATK Mutect2	20
≥ 3.00	7	VarDict	11
3. Targeted genes		VarScan	10
< 300	1	Strelka	4
300–600	28	Lofreq	2
600–900	24	Sentieon	2
≥ 900	3	SomaticSniper	1
4. Limit of detection (LOD)		Scalpel	1
0.5%	4	realDcaller	1
0.6%	1	Freebayes	1
1%	19	Lancet	1
2%	9	Pisces	1
3%	3	In-house	9
5%	13	10. Variant filtering ‡	
Other †	7	GATK FilterMutectCalls	15
5. Sequencing Mode		VarDict	4
Paired End	56	VarScan	4
6. Sequencing read length (bp)		GATK Mutect2	1
100	8	GATK	1
150	48	Lofreq	1
7. DNA input (ng)		Sentieon	1
10–99	17	In-house	29
100–299	22	11. Variant annotation ‡	
300–499	4	Annovar	40
≥ 500	13	snpEff	15
8. Fragment DNA		VEP	13
Ultrasonic shearing:		Transvar	1
Covaris	33	In-house	7
Bioruptor	7		
Qsonica	3		
NA	1		
Enzyme digestion:			
KAPA (Roche) enzyme	4		
TIAEN enzyme	1		
NEB enzyme	1		
Other	9		

† Different cutoff for mutations occurred in hotspot genes and non-hotspot genes.
‡ More than one software may be used for mutation analysis.

S2000, which had an average depth ranging from 500x to 2000x. Variant calling tools, which including VarScan [19], VarDict [20] and GATK Mutect1 and Mutect2 [21], were combined with different variant filtering method (GATK FilterMutectCalls [22], Vardict and in-house developed software) and alignment strategy such as BWA [23] to create reliable set of positive variants.

The detectable region varied significantly across laboratories, and an assessable region with a size of 1.8 Mb was defined (details provided in Supplementary Methods). For the candidate calls in the assessable region, there were results from at least three Ref panels to determine whether they were positive or negative. Within the assessable region, the first part of truth set (termed set Ⅰ) was defined as mutations with AF ≥ 10% called by all four WES analyses. To further fully enrich the truth set, the mutation results with AF ≥ 5% of 12 non-repetitive samples from the 14 Ref panels were considered as candidate mutation calls for the truth set (termed set Ⅱ). According to the consistency among the Ref panels, the candidate mutation calls were classified into three confidence levels, including high-confidence (HighConf) calls reported by more than 80% of detectable Ref panels, medium-confidence (MedConf) calls reported by 20%–80% of detectable Ref panels, and low-confidence (LowConf) calls reported by less than 20% of detectable Ref panels. The HighConf and LowConf calls reviewed by IGV were considered as TPs (set Ⅱ) and FPs, respectively. MedConf calls were accepted into truth set Ⅱ if reported by at least two WES and passed IGV review, whereas the remaining mutation calls in MedConf were classified into an "Unassessed" list and not involved for assessing the performance of panels (Fig. 1F). More details are presented in the "Establishment of the truth set" in Materials and Methods section.

Based on the above rules, we identified 168 positive mutations in the high-confidence region (referred to as assessable region) (Table S6) and 32 mutations in the unassessed list (Table S7). Table 2 summarizes the characteristics of mutations in the truth set from 12 samples. The types of mutations included missense SNV (69/168, 41.1%), nonsense SNV (8/168, 4.8%), synonymous SNV (38/168, 22.6%), splice site SNV (1/168, 0.6%), frameshift insertion (16/168, 9.5%), frameshift deletion (35/168, 20.8%), and complex mutations (1/168, 0.6%). Indels accounted for 30.3% of the truth set (51/168), among which 41 indels that occurred on the *MLH1*, *MSH2*, and *POLE* genes were introduced by the CRISPR technology. There were 10 indels larger than 10 bp, which were a challenge for the participating panels. Notably, 38 mutations were found in the COSMIC database (Table 2).

Performance of somatic mutations detection in the assessable region

We evaluated the performance of the panels in detecting somatic mutations in the assessable region by comparing them with the truth set, and the total error of 56 panels was calculated (Fig. 1G). Considering the reproducibility errors occurred in duplicate samples, the penalty was calculated to judge the performance rating in assessment report (Materials and Methods). Accordingly, the performance of 28 panels from 27 laboratories was considered acceptable and they had a penalty less than 10. The remaining panels were classified as improvable with a penalty ranging from 10 to 293.

Across the 56 panels, a total of 1306 errors were found, including 179 FPs (range from 0 to 39), 729 FNs (range from 0 to 62) and 398 reproducibility errors (range from 0 to 126). Only 2 panels (2/56, 3.6%) from 2 laboratories correctly reported all the results, and 90.1% of the total error (1177/1306) came from 28 panels judged as improvable. No FPs was found in the results of 42.9% of the panels (24/56); no FNs were found in the results of 14.3% of the panels (8/56); no reproducibility errors was found in the results of 26.8% of the panels (15/56). The errors in all 14 samples (except SNC) of the 56 panels are summarized in Table S5.

Precision and recall (sensitivity) of the 56 panels were calculated for 12 samples against the truth set. The F1-score was the trade-off between precision and recall, which was ranked based on the value (Fig. 3). Performance metrics varied substantially across panels: precision ranged from 0.773 to 1, recall ranged from 0.683 to 1, and F1-score ranged from 0.725 to 1. Only one panels had a precision of less than 0.8, whereas the recall of 14.3% of panels (8/56) were less than 0.8. Consistently, false-negatives were more common in results from 56 panels, indicating that it is more difficult to achieve a high sensitivity for somatic mutation detection. The F1-score of 45 panels (45/56, 80.3%) was greater than 0.9, which indicated both high precision and recall. Notably, two panels, namely OncoScreen Plus and EpiCGP, were used by more than one laboratory. Three laboratories using the former panel and two laboratories using the latter panel all had quite high recall (100%, ≥ 97%) and precision (≥97%, 100%) (Fig. S1A).

Table 2
Characteristic of mutation in truth set of 12 samples.

Type	Number	VAF (%)				Cosmic database
		5–10	10–20	20–30	30–50	
Total SNV						
Synonymous substitution	38	10	11	8	9	10
Missense substitution	69	17	19	10	23	18
Nonsense substitution	8	3	2	0	3	0
Total Indel						
Frameshift insertion	16	0	3	2	11	1
Frameshift deletion	35	7	5	3	15	9
Splice_site	1	0	0	0	1	0
Complex	1	0	0	0	1	0
Total	168	37	40	23	67	38

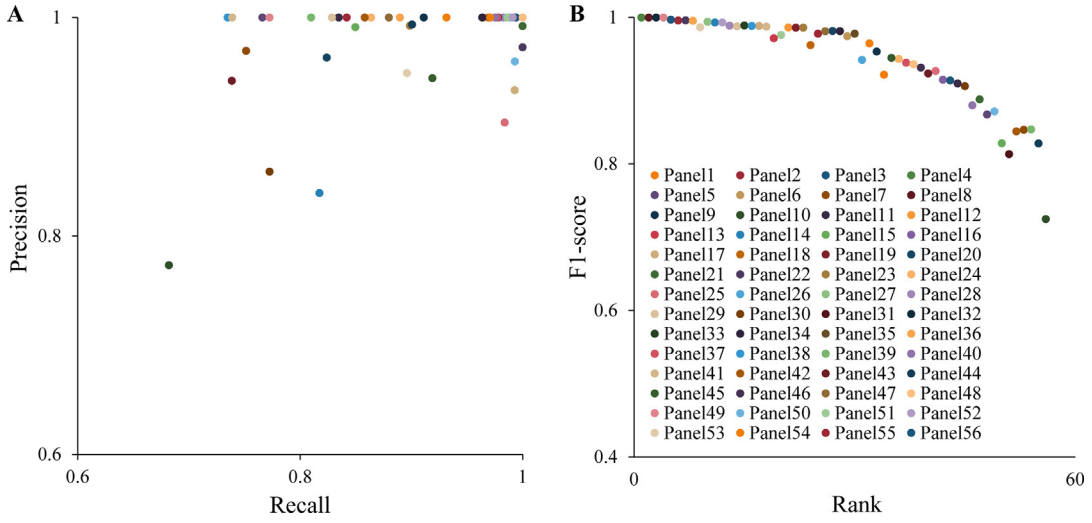


Fig. 3. Overview of the performance in somatic mutation detection among 56 panels. (A) Precision-recall plot for pooled sample. (B) The F1-score calculated using precision and recall were ranked according to the value. Colors represent individual panels. These two figures share the same legend in (B).

Potential sources of errors

Consistency among panels

The truth set only included mutations with AF $\geq$ 5%, and assessment of the performance of panels was not conducted in lower AF mutations. A pooled analysis of the mutation results of 12 samples were performed to explore the impact of low AF on mutation detection. All mutations in the assessable region reported by the 56 panels were classified into high consistency mutations (detected or not detected by more than 80% of detectable panels) and low consistency mutations (detected by 20%–80% of detectable panels). Allele frequencies of the latter were considerably lower than those of the former, and most of them were close to the LOD (0%–5%). Similarly, mutations detected by less than 20% of the panels also generally had low AFs, which were more likely to be FPs (Fig. 5A). Thus, low AF was a main source of variation in mutation detection and contributed to the inconsistency among the results from 56 panels.

Reproducibility

S13 and S14 were designed as replicates of S03 and SNC, which should have the same mutation results under the same testing procedure. A total of 398 reproducibility errors were reported by 40 panels, with the number ranging from 0 to 126. More than five reproducibility errors were observed in 28.6% of the panels (16/56), among which 13 panels had a LOD of less than 3%. Mutation results of two pairs of replicates from these 13 panels were extracted and classified into three subgroups according to VAF: mutations with AF $\leq$ 3%, mutations with AF ranging from 3% to 5%, and mutations with AF $\geq$ 5%. Reproducibility was defined as the fraction of mutations shared between any pair of replicates, and the intra-lab reproducibility of the 13 panels was calculated (Materials and Methods). The intra-lab reproducibility was generally high (0.625–1) for mutations with AF $\geq$ 5% among 13 panels, but was relatively low (0–0.57) and varied widely among panels for mutations with AF $\leq$ 3% (Fig. 4). Importantly, the number of mutations with AFs ranging from 3% to 5% was relatively small, which resulted in the reproducibility of few panels being relatively high for these mutations. Despite this, compared to the mutations with AF $\geq$ 5%, the reproducibility for mutations with AF ranging from 3% to 5% was relatively low (0.14–1) and varied among panels. The cross-lab reproducibility between laboratories using the same panels was assessed, and a high reproducibility (>0.8) among all samples was observed, with cross-lab reproducibility being slightly worse than intra-lab reproducibility in a few samples (Fig. S1B).

Source of FPs

Previous studies have shown that low VAF is a main source of FPs [24], but the present study focused on non-negligible FPs with an AF $\geq$ 5% reported by all panels. There were eight FPs from one panel that did not submit BAM files as required, and the remaining 171 FPs from 31 panels were analyzed.

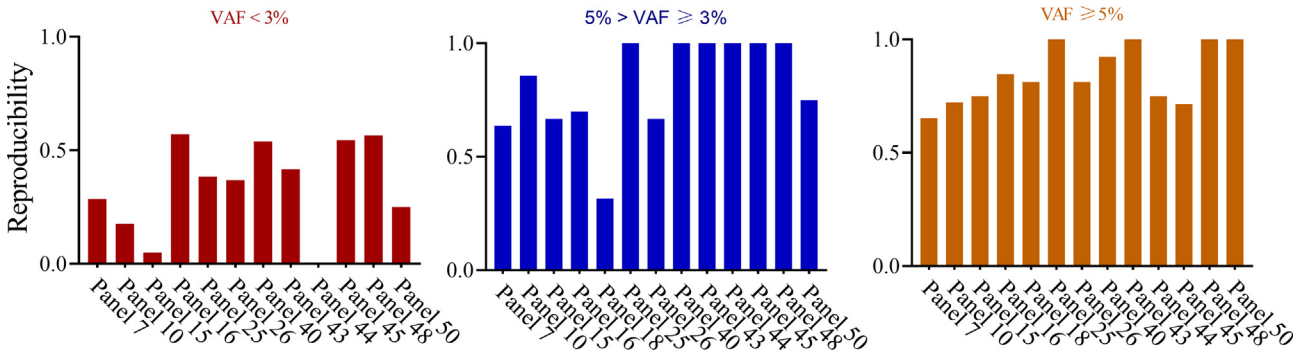


Fig. 4. Reproducibility is reported separately for variant candidates at high (AF ≥ 5%), mid (AF range from 3% to 5%) and low frequency(AF < 3%) (as above). P < 0.0001 for reproducibility between mutations with AF < 3% and AF ≥ 5%. The mutations for reproducibility analysis were from 14 panels with more than five reproducibility errors and limit of detection (LOD) of less than 3%.

All FPs were divided into three clusters based on different sources (Table 3). FPs that were likely to occur in the wet bench process were classified as cluster 1. Sample cross-contamination during library preparation led to 27 FPs from a single panel (panel 10). For example, nine positive variants in truth set of S1 but not S4, were reported in S4 and supported by its raw data. There is another severe but not entirely clear type of error in the wet bench process. Specifically, 32 unique FPs reported by 8 panels, which had common characteristics were considered from sequencing artifacts, strand bias or similar sequencing errors magnified by technicians. First, they are all A(T) > G(C) SNVs with a similar AF range from 5% to 10% except one TC deletion and four with AF ≥ 10%. Second, several base alterations were near the end of the reads supporting them. Third, these base alterations were supported by one or two panels, but not by raw data from the vast majority of the other panels. In addition, there were six FPs from panel 10 classified into this cluster, because they were labeled as LOWQUAL by the other panels and demonstrated poor performance in the wet bench process.

FPs that occurred in the dry bench process (bioinformatic pipeline) were classified as cluster 2. In total, 44 FPs from 12 panels were detected as a result of software faults. Each of the 12 panels reported the same FPs in different samples but their raw data did not support these variants. For example, panel 15 reported chrX:66766406_66766408del in nine samples, and panel 18 reported chrX:66765210_66765227del in 3 samples, but no reads in BAM files of these tumor samples supported them. As at least one in-house software for mapping, variant calling or filtering is involved in all these panels, the reason of these FPs may be software issues or wrong reference sequences. Except the software faults, five panels reported eight FPs because they did not remove non-human DNA fragments from the sequencing data. Total of 31 FPs from 13 panels should be filtered out by tumor–normal pairs. They did not recognize these variants in SNC because their poor sequencing depth on SNC did not meet stringent filtering criteria. In addition, 14 panels reported 15 indel FPs which were rejected by other panels because variants were gathered around positive indels.

FPs related to other reasons were classified as Cluster 3. Seven FPs from six panels were labeled as LOWQUAL by others because of insufficient depth or low AF, which might be weak variants that suffer from variation in AF among multiple assays. Besides, one panel reported one FP as the result of a bad format.

Table 3
The source of false-positives.

Source	No. of FPs	No. of panels
Cluster 1 Wet Bench Process		
cross-contamination	27	1
artifacts / other sequencing errors	38	9
Cluster 2 Dry Bench Process		
software faults	44	12
fail in filtering mutation in normal	31	13
fail in filtering non-human DNA	8	5
gathered around positive variants	15	14
Cluster 3 Other Reason		
weak variants	7	6
bad data format	1	1
Total	171	31

Source of FNs

There were 42 of the 56 panels that correctly annotated the VCF files using tags in the Materials and Methods, and 571 FNs with AF ≥ 5% were found in these panels. We tracked all FN mutations in the corresponding VCF files and summarized the tags annotated for each FN mutation. We then searched for mutations in the BAM files if there was no record of them in the VCF files.

All the FNs were analyzed, and the sources of them were divided into three clusters (Fig. 5B). Cluster I included the main source of the FN results, which was related to low sequencing quality. There were 281 FNs originating from cluster I, of which 241 FNs from 31 panels were labeled as LOWQUAL and filtered out because the quality value or allele frequency were below the threshold. The other 40 FNs from one panel were not supported by any reads in the corresponding raw sequencing data. Cluster II referred to some redundant filtering that was irrelevant to sequencing quality. Specifically, 9 panels filtered out 96 mutation calls of no clinical significance or synonymous mutation, which were unnecessary in our analysis. Cluster III concerned incorrect filtering. Therein, 106 FNs labeled MATCHED were reported by 15 panels because of incorrect filtering by principle of tumor–normal pairs; and these FNs involved 17 mutations with a low AF (range from 0.2% to 5%) and mutant reads (range from 2 to 144) in matched normal sample, which further revealed the challenge of low AF mutations in matched normal sample to somatic analysis. In addition, 62 mutation calls were incorrectly judged as germline mutations, mutations in noncoding regions, or clonal hematopoiesis-derived mutations and filtered out, indicating inappropriate databases used in mutation analysis. The seven FNs labeled OUTSCOPE from five panels were a consequence of the discordance between the declared and real detection ranges. Noticeably, 14 positive mutations were rejected by manual review for unknown reason before submission. In total, FNs of 40.5% of the panels (17/42) were from single source, such as LOWQUAL or MATCHED, and FNs of 26.2% of the panels (11/42) were from two sources, which indicated common errors in the mutation analysis process (Fig. 5C).

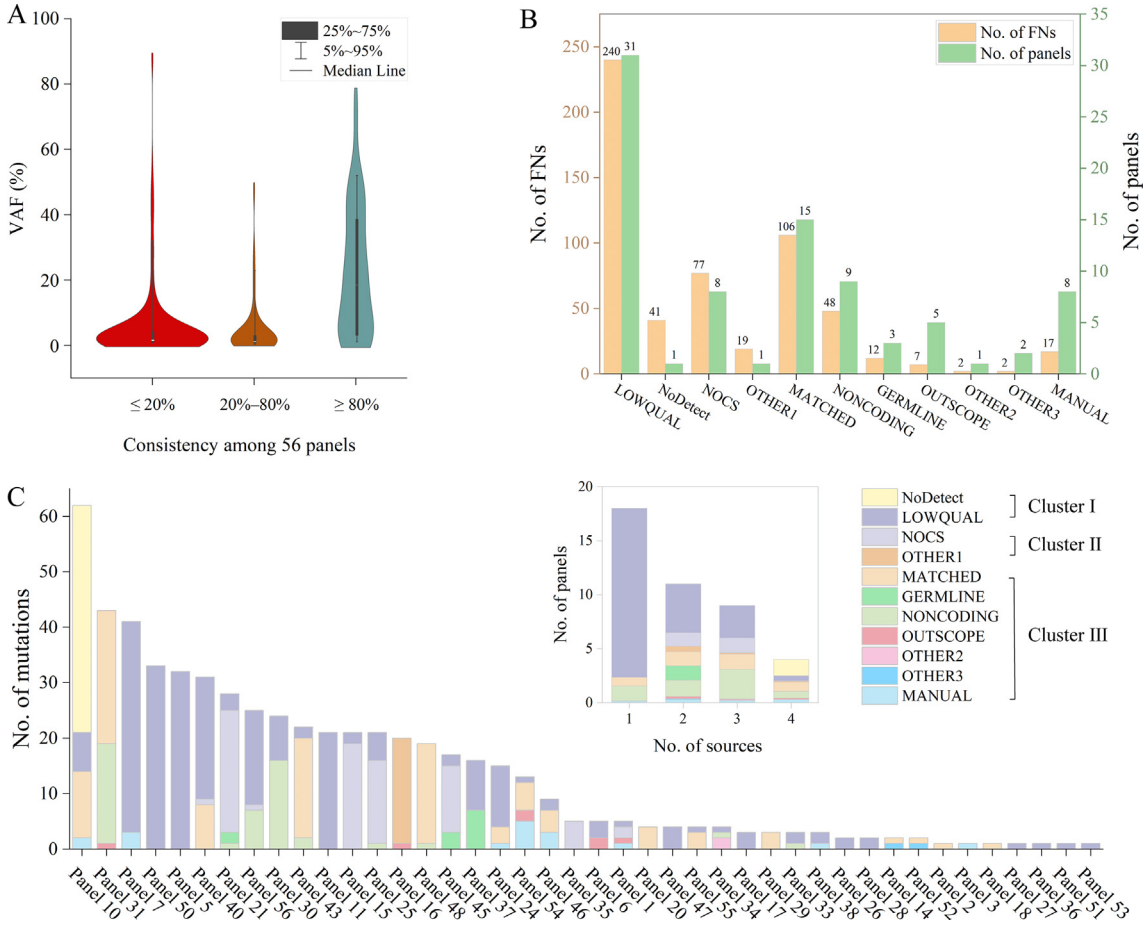


Fig. 5. Source of variations among results from 56 panels. (A) Distribution of VAFs in dependence of mutations with three consistency level among 56 panels. The consistency level was defined as the ratio of the number of panels reporting a mutation and the number of detectable panels in 12 samples. (B) The number of false negatives and panels corresponding to each source of error. OTHER1 indicates that mutation calls are incorrectly filtered because of synonymous mutations. OTHER2 indicates that mutation calls are filtered out because of clonal hematopoiesis-derived mutations. Mutation calls labeled OTHER3 are filtered out because laboratories considered that many mutations gather in this position. (C) The number of false-negatives and proportion of source of errors for 42 panels that correctly submitted the VCF files. The inset highlights the number of panels where all false-negatives originate from a different number of sources. NoDetect refers to positive mutations not detected by the panel, supporting by no difference in mutant reads between the tumor and normal samples in the raw data.

Impact of variant type on performance

While non-SNV (indels and multi-nucleotide variants (MNVs)) variant detection was more challenging than SNV variant detection in previous researches [24,25], the performance of SNV and non-SNV variant detection varied among specific panels in our study. We observed similar or better SNV detection performance compared with that of non-SNV in a substantial part of panels. However, the recall, precision, and reproducibility of SNV variant detection were worse than non-SNV variants in other 39, 16 and 29 panels, respectively (Fig. S2A). To explain this phenomenon, the source of SNV and non-SNV false discovery from these panels was compared. Importantly, problems related to the bioinformatic filtering process (Cluster 2 or Ⅱ, and Cluster 3 or III) contributed to 51.7% of SNV FNs versus 13.8% of non-SNV FNs (Fig. S2B), while also caused 43.2% and 48.1% of SNV FPs and non-SNV FPs, respectively (Fig. S2C). Thus, other processes rather than variant calling had a great impact on low recall and precision of somatic SNV variant detection. For panels with a better reproducibility on non-SNV variant detection than SNV variant detection, a vast majority of reproducibility errors, including both SNV and non-SNV variants, were clustered in an AF range below 10% (Fig. S2D).

Impact of sequencing platform on performance

The average recall, precision and reproducibility of panels using each sequencing platform were all acceptable, but the dispersion of that in panels using Illumina NovaSeq 6000, MGISEQ-2000 (including GENETRON 2000) and Illumina NextSeq 500/CN500 was quite notable (Fig.S3). The outliers of Illumina NovaSeq 6000 were composed of 12 panels with similar problems. Each of these panels performed average or top on 1 or 2 metrics but bottom at the other, suggesting that the cut-off and filter settings of these panels were highly imbalanced. For instance, Panel 50 had a perfect precision (100%), but its recall (77.24%) and reproducibility score (68.42%) were far below others. As its cut-off setting of somatic variants was "AF ≥ 1% & Alt Allele ≥ 6 reads" and the AFs of its most reproducibility errors (5/6) were between 1% and 2%, it can be seen from its data that such a low cut-off incurred many intractable weak positives and fragile reproducibility. Meanwhile, in order to achieve a sky-high precision at this extreme cut-off, the filter criteria of variant calling were made very stringent (data not shown), leading to excessive false negatives and poor recall. The only outlier of Illumina Next 500/CN500 and the only DNBSEQ-T7 result shared the same problem with these 12 panels above. The only

outlier of MGISEQ-2000 was Panel 10 with unusual cross-sample contamination, which could not reflect the real performance of its sequencing platform either. In sum, the apparent performance difference among sequencing platforms was actually driven by problems of laboratories, instead of intrinsic error of sequencing platforms. Also, taking the limited number of panels using each sequencing platform into account, we cannot conclude that there was a significant performance gap among sequencing platforms.

Discussion

This study comprehensively assessed performance of somatic mutation detection using large panel NGS, including preparation of reference samples using DNA repair gene and proofreading gene knock-down cell lines, inclusion of multiple panels from hospital or commercial laboratories, generation of the truth set based on results from WES and panels, and evaluation of the reproducibility, FPs, and FNs of the panels.

It is always a primary challenge to prepare reference materials that are commutable with authentic clinical samples and suitable for tumor–normal detection approach for standardization studies. Clinical FFPE specimen are not reproducible and replaceable, and the mutational profile is heterogenous [26]. Moreover, the quality and quantity of DNA extracted from them are crucial for performing genomic profiling, which generally depend on pre-analytical factors such as fixation time, within block position, and DNA extraction methods [27,28]. Noting this, the pre-analytical workflow has been widely evaluated by previous multicenter studies using FFPE cell line samples that could closely mimic the real FFPE samples [29–32]. To evaluate the analytical performance of all participating panels exclusively, DNA samples from engineered cell lines were distributed to laboratories. Only three MMR and proofreading genes were knocked down using the CRISPR-Cas9 technology, which was more effective to obtain mutations than a previous study wherein 36 genes were edited to generate reference cell lines with 125 mutations [33]. Given that MMR and proofreading systems are critical for high fidelity of genome replication and stability [34], defects in them would play a key role in spontaneous mutagenesis and tumor development [35]. Mutations in genes involved in MMR and proofreading systems have been observed in several tumor types, and cause a corresponding dMMR mutational signature and high burden of genomic mutations [36]. Likewise, $MLH1^{KO}$ human stem cell organoids had a mutational signature similar to that of MMR-deficient cancers [37,38]. It could be inferred that our reference cell lines are close to clinical tumor samples with respect to the extent of mutation and mutation signature, allowing for benchmarking sequencing pipeline in the whole genome. Eleven loss-of-function deletion or insertion mutations in *MLH1*, *MSH2* and *POLE* gene were finally introduced in six clones (Table S6), leading to an obvious increase in the number of mutations, especially after initial culturing. Significantly, clone C1-1 had approximately 130 non-synonymous mutations in exon after culturing of 7 months, which increased at a rate close to about 100 mutations per genome per day of an $MLH1^{KO}$ organoid [37]. The accumulation speed of mutations in six clonal cells was different, implying that the different introduced mutations would cause genome instability at different levels. Considering the changes of mutations during successive passage caused by genome instability, we frozen sufficient cell for WES validation and laboratory use at once. Importantly, mutational changes allow us to prepare samples with different mutational profile, and different mutation burden, which is also meaningful for control material of TMB detection. Besides, deficiency in MMR system could cause microsatellite instability (MSI) which is also a promising biomarker for immunotherapy. We will focus on this significant field in the future research based on our reference materials which present prospects for multicenter assessment of MSI detection [39].

Another challenge for standardization studies is reliable methods to generating the truth set for the reference samples. Considering Sanger sequencing and ddPCR are limited in confirming mutations with a large number and low AF, orthogonal targeted sequencing and WES/WGS have been widely used in previous studies on establishing reference datasets and panel validation [18,40]. Multiple sequencing platforms, centers, and various bioinformatic pipelines are applied on reference samples to minimize the biases specific to each process [41,42]. Similarly, inspired by experience of consecutive three years of EQA by the European Molecular Genetics Quality Network (EMQN) and the Genomics Quality Assessment (GenQA), it is effective to determine the truth set as consensus mutations submitted from participating laboratories [16]. We included four WES, as well as 14 participating panels including six sequencing platforms and multiple combinations of variants mapping, calling, and filtering tools, to establish the high-confidence region and HighConf mutations, and set a higher consistency threshold for truth set compared to the aforementioned study (80% vs. 75%) [16]. Moreover, all the laboratories were required to annotate all the mutation calls in VCF files, so that we could correct the MedConf mutations if they were incorrectly filtered out. In this way, we could also explore the sources of FPs and FNs. It should be noted that we did not validate the truth set using Sanger sequencing or ddPCR because of convincing rules for the truth set. First, the Ref panels selected for generating truth set have high sequencing depth that was far more reliable than WES/WGS and demonstrated perfect performance on two pairs of replicates. Second, eighteen libraries on each sample were constructed compared to only four WES libraries and one WGS library in previous study [10], which was more important than focusing on multiple callers upon limited library number in the study design [32]. Finally, we successfully traced the source of all false mutation results against the truth set, and there was no dispute from laboratories on the truth set and reported errors. Our reference material containing abundant mutations and the corresponding workflow are inseparable. It is recommended that the workflow should be applied to guarantee the truth set is generated accurately and effectively.

We distributed homogeneous reference samples to laboratories, and required them to apply their own routine testing procedures. Generally, NGS tests need to be validated before clinical use [43]. Therefore, the results from their own routine testing procedures reflect the real detection proficiency in clinical use, which our multi-laboratory study emphasize. We had a comprehensive understanding of the methodological variance and detection performance of domestic large panel NGS. The methodology of somatic mutation detection varied among laboratories, including sequencing platforms, library preparation protocols and bioinformatic tools (Table 1). Particularly, most panels (30/56, 53.6%) involved in-house bioinformatic tools for reads mapping, variants calling, filtering and annotation. The low accuracy and comparability of the panels were revealed, and the specific impact factors have been explored in several previous researches [24,32,44,45]. In our study, we analyzed the false discovery of panels in detecting somatic mutations with AF $\geq$ 5%, which theoretically influences clinical application of not only actionable mutations but TMB [46,47]. FNs are more common than FPs. We tracked each FNs and found that low sequencing quality and inappropriate filtering settings were the two sources of FNs. Mutations labeled LOWQUAL are either true low-quality mutations or observed because of too stringent filtering rules [48]. However, 40 mutations were not detected by one panel, which originated from the problems in sequencing reactions in wet bench process, with evidence of no difference of mutant reads between tumor and normal sample

under a high coverage depth. Previous research has demonstrated that the choice of bioinformatic pipeline influences variant calling results [32]. In this study, 49.6% of FNs (283/571) originated from problems during variant filtering, and we focused on the content of incorrect filtering beyond the influence of specific bioinformatic tools. For example, several laboratories only reported mutations related to TMB analysis or targeted therapy and filtering synonymous mutations and mutations of no clinical significance, which were considered as errors beyond detection proficiency and had no impact on clinical decisions. In addition, FNs labeled MATCHED indicated the importance of accurately distinguishing somatic variants from matched tumor–normal sample by different algorithms [49]. Mutations with AF much lower than germline mutations (50% or 100%) in matched normal sample (SNC) were challenging for variants callers to isolate somatic mutations, which prompted further optimization of the algorithm. A few laboratories judged positive mutations as germline mutations, mutations in noncoding region or clonal hematopoiesis-derived mutations and filtered them, revealing that bioinformatic tools or databases should been optimized. Finally, a few FNs were from incorrectly manual filtering, which could be solved by improving the training professional analysts.

On the FPs, it is evident that errors during the wet bench process appeared solely in 9 panels, but the consequence could be serious once occurred. For example, sample cross-contamination resulted in 32 FPs from one panel. This cluster of errors could be reduced by building standard operation procedure (SOP), training laboratory technicians, and other management strategies. Errors during the dry bench process in 24 panels were relatively common, but had limited influence for each assay. Specifically, software faults could be eliminated by revising the database and in-house software. Inappropriate filtering settings and similar errors could be reduced by adequate regular performance validation. In addition, FPs caused by weak variant, the inherent problem of NGS, may be gradually solved by novel methods, like machine learning, which is currently beyond the scope of our research area.

It has been elucidated that SNPs were easier to identify than indels in germline mutation detection [25,50]. However, SNV and non-SNV somatic variants detection demonstrated a similar sensitivity and reproducibility in previous research involving limited indels and MNVs [24]. In our study, a higher recall, precision, and reproducibility in detecting non-SNV variants among several panels was found to originate from processes, such as wet-bench operation, and variant filtering. Noting this, every section of somatic mutation detection should be highly-regarded during the development of workflow. Additional reference material with abundant somatic mutations, especially indels and MNVs, is required to validate and benchmark somatic mutation detection workflow. In addition, different sequencing platform have intrinsic error rate, and we compared the performance of panels using different sequencing platform. We found that the real capabilities of sequencing platforms were covered by improper and imbalanced settings. Laboratories developing and using panels are supposed to conduct rounds of validation and verification to ensure acceptable recall, precision, and reproducibility at a proper cut-off.

Conclusion

In conclusion, this multi-laboratory study promoted the standardization of somatic mutation detection using large panel next-generation sequencing. A naturally suitable approach for generating reference materials for future research on standardization was provided. In addition, comprehensive evaluation revealed that the variable detection performance of somatic variants with AF $\geq$ 5% by oncopanel sequencing was not unsolvable, and could be overcome by method optimization, validation, and quality control.

Compliance with Ethics Requirements

This article does not contain any studies with human or animal subjects.

CRediT authorship contribution statement

Duo Wang: Investigation, Software, Data curation, Visualization, Writing – original draft. **Yuanfeng Zhang:** Investigation, Software, Data curation, Visualization, Writing – original draft. **Rui li:** Investigation, Software. **Jinming Li:** Conceptualization, Methodology, Writing – review & editing. **Rui Zhang:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the 55 laboratories for participating actively in our research and for reporting the test results to NCCL on time. We also thank Genetron Health (Beijing) Co. Ltd. (Beijing, China) and Nanjing Geneseeq Technology Inc. (Nanjing, Jiangsu, China) for providing whole exome sequencing support.

This work was supported by the grants from National Key Research and Development Program of China (2018YFE0201600).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2022.03.016.

References

[1] Lowery MA, Ptashkin R, Jordan E, Berger MF, Zehir A, Capanu M, et al. Comprehensive molecular profiling of intrahepatic and extrahepatic cholangiocarcinomas: potential targets for intervention. Clin Cancer Res 2018;24(17):4154–61.
[2] Pozdeyev N, Gay LM, Sokol ES, Hartmaier R, Deaver KE, Davis S, et al. Genetic analysis of 779 advanced differentiated and anaplastic thyroid cancers. Clin Cancer Res 2018;24(13):3059–68.
[3] Hyman DM, Taylor BS, Baselga J. Implementing genome-driven oncology. Cell 2017;168(4):584–99. doi: https://doi.org/10.1016/j.cell.2016.12.015.
[4] Devarakonda S, Rotolo F, Tsao M-S, Lanc I, Brambilla E, Masood A, et al. Tumor mutation burden as a biomarker in resected non-small-cell lung cancer. J Clin Oncol 2018;36(30):2995–3006.
[5] Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet 2019;51(2):202–6.
[6] Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 2014;15(1):56–62. doi: https://doi.org/10.1038/nrg3655.
[7] Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet 2018;19(5):269–85. doi: https://doi.org/10.1038/nrg.2017.117.
[8] Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol 2019;20(1). doi: https://doi.org/10.1186/s13059-019-1659-6.
[9] Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet 2017;18(8):473–84. doi: https://doi.org/10.1038/nrg.2017.44.
[10] Jones W, Gong B, Novoradovskaya N, Li D, Kusko R, Richmond TA, et al. A verified genomic reference sample for assessing performance of cancer panels detecting small variants of low allele frequency. Genome Biol 2021;22(1). doi: https://doi.org/10.1186/s13059-021-02316-z.

[11] Min YK, Park KS. The application of control materials for ongoing quality management of next-generation sequencing in a clinical genetic laboratory. Medicina (Kaunas) 2021;57(6). doi: https://doi.org/10.3390/medicina57060543.
[12] Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature 2009;461(7267):1071–8. doi: https://doi.org/10.1038/nature08467.
[13] Pursell ZF, Isoz I, Lundström EB, Johansson E, Kunkel TA. Yeast DNA polymerase epsilon participates in leading-strand DNA replication. Science 2007;317(5834):127–30. doi: https://doi.org/10.1126/science.1144067.
[14] Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science 2017;357(6349):409–13.
[15] Park VS, Pursell ZF. POLE proofreading defects: Contributions to mutagenesis and cancer. DNA Repair (Amst) 2019;76:50–9. doi: https://doi.org/10.1016/j.dnarep.2019.02.007.
[16] Gutowska-Ding MW, Deans ZC, Roos C, Matilainen J, Khawaja F, Brügger K, et al. One byte at a time: evidencing the quality of clinical service next-generation sequencing for germline and somatic variants. Eur J Hum Genet 2020;28(2):202–12.
[17] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27(15):2156–8.
[18] Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat Biotechnol 2021;39(9):1151–60.
[19] Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 2009;25(17):2283–5.
[20] Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016; 44(11): e108. doi: 10.1093/nar/gkw227.
[21] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31(3):213–9.
[22] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–303.
[23] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754–60. doi: https://doi.org/10.1093/bioinformatics/btp324.
[24] Gong B, Li D, Kusko R, Novoradovskaya N, Zhang Y, Wang S, et al. Cross-oncopanel study reveals high sensitivity and accuracy with overall analytical performance depending on genomic regions. Genome Biol 2021;22(1). doi: https://doi.org/10.1186/s13059-021-02315-0.
[25] Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol 2019;37(5):555–60.
[26] Malapelle U, Mayo-de-Las-Casas C, Molina-Vila MA, Rosell R, Savic S, Bihl M, et al. Consistency and reproducibility of next-generation sequencing and other multigene mutational assays: A worldwide ring trial study on quantitative cytological molecular reference specimens. Cancer Cytopathol 2017;125(8):615–26.
[27] Al Zoughbi W, Kim D, Alperstein SA, Ohara K, Manohar J, Greco N, et al. Incorporating cytologic adequacy assessment into precision oncology workflow using telepathology: An institutional experience. Cancer Cytopathol 2021;129(11):874–83.
[28] Group SOSW, Zhang Y, Blomquist TM, Kusko R, Stetson D, Zhang Z, et al. Deep oncopanel sequencing reveals fixation time- and within block position-dependent quality degradation in FFPE processed samples. bioRxiv. 2021:2021.04.06.438687. doi: 10.1101/2021.04.06.438687.
[29] Patton S, Normanno N, Blackhall F, Murray S, Kerr KM, Dietel M, et al. Assessing standardization of molecular testing for non-small-cell lung cancer: results of a worldwide external quality assessment (EQA) scheme for EGFR mutation testing. Br J Cancer 2014;111(2):413–20.
[30] Kapp JR, Diss T, Spicer J, Gandy M, Schrijver I, Jennings LJ, et al. Variation in pre-PCR processing of FFPE samples leads to discrepancies in BRAF and EGFR mutation detection: a diagnostic RING trial. J Clin Pathol 2015;68(2):111–8.
[31] Dijkstra JR, Opdam FJM, Boonyaratanakornkit J, Schönbrunner ER, Shahbazian M, Edsjö A, et al. Implementation of formalin-fixed, paraffin-embedded cell line pellets as high-quality process controls in quality assessment programs for KRAS mutation analysis. J Mol Diagn 2012;14(3):187–91.
[32] Xiao W, Ren L, Chen Z, Fang LT, Zhao Y, Lack J, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat Biotechnol 2021;39(9):1141–50.
[33] Suzuki T, Tsukumo Y, Furihata C, Naito M, Kohara A. Preparation of the standard cell lines for reference mutations in cancer gene-panels by genome editing in HEK 293 T/17 cells. Genes and Environ 2020;42(1).
[34] Barbari SR, Shcherbakova PV. Replicative DNA polymerase defects in human cancers: Consequences, mechanisms, and implications for therapy. DNA Repair (Amst) 2017;56:16–25. doi: https://doi.org/10.1016/j.dnarep.2017.06.003.
[35] Ijsselsteijn R, Jansen JG, de Wind N. DNA mismatch repair-dependent DNA damage responses and cancer. DNA Repair (Amst) 2020;93:. doi: https://doi.org/10.1016/j.dnarep.2020.102923102923.
[36] Pozdeyev N, Fishbein L, Gay LM, Sokol ES, Hartmaier R, Ross JS, et al. Targeted genomic analysis of 364 adrenocortical carcinomas. Endocr Relat Cancer 2021;28(10):671–81.
[37] Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science 2017;358(6360):234–8.
[38] Hodel KP, Sun MJS, Ungerleider N, Park VS, Williams LG, Bauer DL, et al. POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. Mol Cell 2020;78(6):1166–1177.e6.
[39] Ganesh K, Stadler ZK, Cercek A, Mendelsohn RB, Shia J, Segal NH, et al. Immunotherapy in colorectal cancer: rationale, challenges and potential. Nat Rev Gastroenterol Hepatol 2019;16(6):361–75.
[40] Tomlins SA, Hovelson DH, Harms P, Drewery S, Falkner J, Fischer A, et al. Development and Validation of StrataNGS, a Multiplex PCR, Semiconductor Sequencing-Based Comprehensive Genomic Profiling Test. J Mol Diagn 2021;23(11):1515–33.
[41] Craig DW, Nasser S, Corbett R, Chan SK, Murray L, Legendre C, et al. A somatic reference standard for cancer genome sequencing. Sci Rep 2016;6(1). doi: https://doi.org/10.1038/srep24607.
[42] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 2014;32(3):246–51.
[43] Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, et al. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. J Mol Diagn 2017;19(3):341–65.
[44] Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods 2015;12(7):623–30.
[45] Krøigård AB, Thomassen M, Lænkholm AV, Kruse TA, Larsen MJ. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. PLoS One. 2016; 11(3): e0151664. doi: 10.1371/journal.pone.0151664.
[46] Stenzinger A, Endris V, Budczies J, Merkelbach-Bruse S, Kazdal D, Dietmaier W, et al. Harmonization and Standardization of Panel-Based Tumor Mutational Burden Measurement: Real-World Results and Recommendations of the Quality in Pathology Study. J Thorac Oncol 2020;15(7):1177–89.
[47] Wang D, Ma K, Deng W, Li J, Xiang S, Zhang Y, et al. Development and Analytical Validation of a Targeted Next-Generation Sequencing Panel to Detect Actionable Mutations for Targeted Therapy. Onco Targets Ther 2021;14:2423–31.
[48] Singh RR. Next-Generation Sequencing in High-Sensitive Detection of Mutations in Tumors: Challenges, Advances, and Applications. J Mol Diagn 2020;22(8):994–1007. doi: https://doi.org/10.1016/j.jmoldx.2020.04.213.
[49] Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J 2018;16:15–24. doi: https://doi.org/10.1016/j.csbj.2018.01.003.
[50] Chen J, Li X, Zhong H, Meng Y, Du H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep 2019;9(1):9345. doi: https://doi.org/10.1038/s41598-019-45835-3.