Computational and Structural Biotechnology Journal 17 (2019) 1348–1359

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Mini Review

Applications and analysis of targeted genomic sequencing in cancer studies

Findlay Bewicke-Copley [a,*], Emil Arjun Kumar [a,b], Giuseppe Palladino [b], Koorosh Korfi [a], Jun Wang [a,*]

[a] Centre for Cancer Genomics and Computational Biology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK
[b] Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

ARTICLE INFO

Article history:
Received 31 March 2019
Received in revised form 18 October 2019
Accepted 22 October 2019
Available online 7 November 2019

Keywords:
Targeted sequencing
Variant calling
PCR duplicates
Background error
Cancer genomics
Clinical samples

ABSTRACT

Next Generation Sequencing (NGS) has dramatically improved the flexibility and outcomes of cancer research and clinical trials, providing highly sensitive and accurate high-throughput platforms for large-scale genomic testing. In contrast to whole-genome (WGS) or whole-exome sequencing (WES), targeted genomic sequencing (TS) focuses on a panel of genes or targets known to have strong associations with pathogenesis of disease and/or clinical relevance, offering greater sequencing depth with reduced costs and data burden. This allows targeted sequencing to identify low frequency variants in targeted regions with high confidence, thus suitable for profiling low-quality and fragmented clinical DNA samples. As a result, TS has been widely used in clinical research and trials for patient stratification and the development of targeted therapeutics. However, its transition to routine clinical use has been slow. Many technical and analytical obstacles still remain and need to be discussed and addressed before large-scale and cross-centre implementation. Gold-standard and state-of-the-art procedures and pipelines are urgently needed to accelerate this transition. In this review we first present how TS is conducted in cancer research, including various target enrichment platforms, the construction of target panels, and selected research and clinical studies utilising TS to profile clinical samples. We then present a generalised analytical workflow for TS data discussing important parameters and filters in detail, aiming to provide the best practices of TS usage and analyses.

Contents

Abbreviations: BWA, Burrows-Wheeler Aligner; FF, Fresh Frozen; FFPE, Formalin Fixed Paraffin Embedded; CLL, Chronic Lymphocytic Leukaemia; FL, Follicular Lymphoma; tFL, Transformed Follicular Lymphoma; NSCLC, Non-Small Cell Lung Carcinoma; NHL, Non-Hodgkin Lymphoma; GATK, Genome Analysis Toolkit; COSMIC, Catalogue of Somatic Mutations in Cancer; ESP, Exome Sequencing Project; ICGC, International Cancer Genome Consortium; NCCN, the National Comprehensive Cancer Network®; TCGA, The Cancer Genome Atlas; QC, Quality Control; BAM, Binary Alignment Map; SAM, Sequence Alignment Map; VAF, Variant Allele Frequency; NGS, Next Generation Sequencing; WES, Whole Exome Sequencing; WGS, Whole Genome Sequencing; TS, Targeted Sequencing; UMI, Unique Molecular Identifiers; MBC, Molecular Barcode.

* Corresponding authors.
E-mail addresses: f.copley@qmul.ac.uk (F. Bewicke-Copley), j.a.wang@qmul.ac.uk (J. Wang).

https://doi.org/10.1016/j.csbj.2019.10.004
2001-0370/

1. Introduction

Recent advances in next-generation sequencing technology have revolutionised our understanding of cancer biology and clinical research. It is now more affordable than before to carry out large-scale NGS experiments with a reasonable turnaround time. This has led to a rapidly expanding body of pioneering research exploring the genomic landscape and molecular mechanisms of various cancer types, as well as the discovery of genetic drivers (i.e., mutations that confer a selective growth advantage, thus promoting cancer development), exemplified by the effort from large international sequencing initiatives, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). This has generated vast amounts of data and identified numerous biomarkers and targets for patient stratification and therapeutics. Although the translation of these findings into the clinic has been slow, in certain settings, NGS is becoming a complementary diagnostic tool, guiding the decision making to achieve personalised and/or precision medicine in a number of cancers [1–6]. With the magnitude of sequencing data generated, the continuing development of advanced bioinformatics tools capable of handling these data efficiently in a timely manner is vital for NGS-centred research and clinical implementation. Researchers and clinicians are now faced with a wide range of NGS techniques and platforms with no clear consensus guidelines, where the trade-offs between costs, accuracy, power and technical difficulties must be considered.

There are three main types of NGS sequencing of DNA that can be used for the identification of genomic mutations: whole-genome sequencing, whole-exome sequencing and targeted sequencing (Fig. 1). We summarise and compare the key information of these three platforms in Table 1. Compared to WGS and WES, TS, is a powerful approach that can fulfil the best balance between the accurate identification of targeted events with great sensitivity, and the overall cost and data burden for large-scale executions. For the data analysis, many existing methods and pipelines designed for WGS/WES can be applied to TS. However, due to the high depth of TS, extra care needs to be taken during the analysis to ensure only high-quality variant calls are retained, especially for data generated from low quality or fragmented DNA and/or without matched normal control. Currently, as with other types of NGS, there is still a lack of gold-standard pipelines

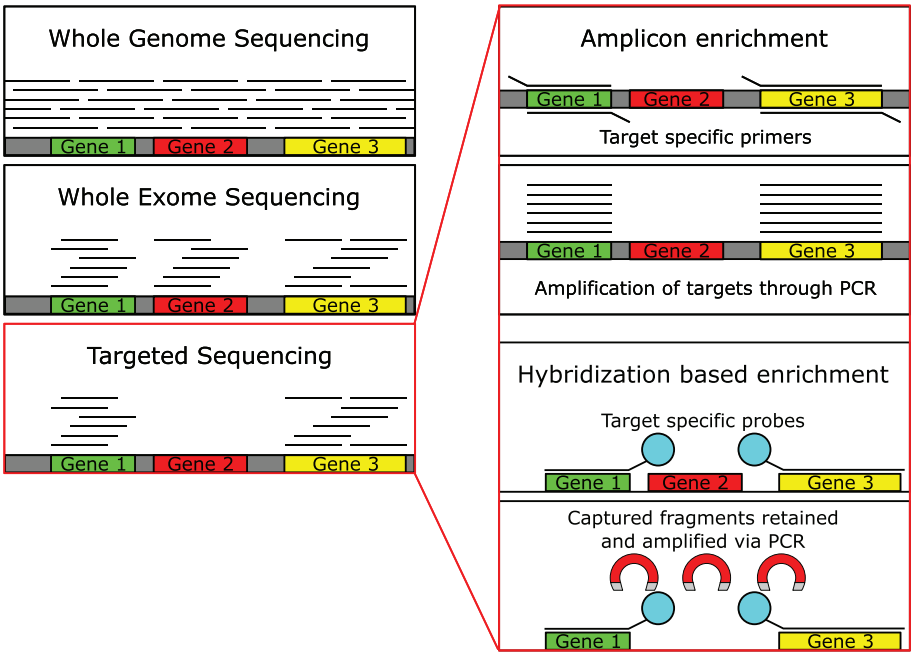

Fig. 1. Methods of DNA-seq. Whole genome sequencing, whole exome sequencing and targeted sequencing are illustrated. For the latter, the two library preparation approaches are shown.

Table 1
Different types of Next Generation Sequencing for genomics.

Platform	Cost (per sample, USD)	Sites	Region Size (bp)	Depth	Data size (Processed Bam)
WGS	$1000–$3000	All coding and non-coding regions	~3×10^9	30–60×	Depending on coverage ~60 GB–350 GB
WES	$500–$2000	Exonic regions	~6×10^7	150–200×	Depending on coverage ~5 GB–20 GB
TS	$300–$1000	Specifically targeted regions	Varies by panel size ~1×10^5–1×10^7	200–1000× +	Varies by panel size and coverage ~100 MB–5 GB

Table 2
A comparison of targeted methods of genomic analysis.

Platform	Target size	Cost (per sample, USD)	Massively Parallel	Minimum allele frequency	Purpose in Research
TS	~1×10^5–1×10^7 bp	$300–$1000	True	1% (without error suppression)	Discovery/Validation
Sanger Sequencing	300–1000 bp	<$30	False	~15%	Validation
Digital PCR	1–80 bp	<$10	False	<0.001%	Validation

for TS analysis, which can lead to poor reproducibility between laboratories, even for the same data sets. This greatly affects the accuracy and efficacy of TS in calling variants for prognostic and therapeutic signatures, as different labs working with different pipelines may not reliably call these variants.

Bearing all these in mind, in this review we first present a general overview of TS, associated platforms and their implementations in various cancer studies. We show that TS provides a powerful and versatile tool to profile clinical samples in cancer research and clinical trials. We then present a generalised analytical workflow for TS data, with commonly used software, and important parameters and filters discussed in each step. We aim to provide guidance on how to analyse the data in a more standardised manner.

2. Targeted genomic sequencing

TS focuses on a number of targeted regions often including many known drivers or clinically-actionable genes of interest and identifies sequence variants with high confidence and accuracy. For example, the genes *KRAS* and *TP53* are often targeted across a range of cancer types, as they are commonly found to be mutated with a number of hotspots. *BRAF* and *EGFR* are also screened in many solid tumours, as they contain clinically relevant mutations [7–13].

The great sequencing depth utilised in TS (e.g., ultra-deep sequencing at a depth of 10,000x and higher) makes it very powerful for profiling clinical samples, such as formalin fixed paraffin embedded (FFPE) and circulating tumour DNA (ctDNA) where DNA quality and/or tumour content is low. Greater depth of coverage also allows TS to pick out mutations that are present only in a small fraction of malignant cells (i.e., sub-clonal), and in the setting of detecting minimal residual disease, with variant allele frequency (VAF) sufficiently detected as low as 0.1–0.2% [14–16]. All these attributes ensure that TS is superior to non-NGS based techniques (e.g., Sanger sequencing and digital PCR) and WGS/WES in large-scale genomic testing and clinical trial setting (see comparison details in Table 2).

TS has been widely used in cancer studies and clinical trials to stratify patients into risk groups based on the mutational status of key genes [17–23]. In clinical practice, Foundation Medicine has launched the first FDA-approved broad companion diagnostic (CDx) that is clinically and analytically validated for solid tumours. This platform identifies genomic alterations and biomarkers across 300+ genes with a median depth of coverage of 500x. It is suitable for processing FFPE samples with quick turnaround (<2 weeks), offering invaluable information for therapeutic targets and immunotherapy biomarkers.

2.1. Methods of targeted sequencing

Targeted sequencing comes in two main forms, amplicon or capture-based (Fig. 1). Amplicon-based enrichment utilises specifically designed primers to amplify only the regions of interest prior to library preparation [24]. Alternatively, in capture-based approaches, the DNA is fragmented and targeted regions are enriched via hybridization oligonucleotide bait sequences attached to biotinylated probes, allowing for isolation from the remaining genetic material [24,25]. Amplicon-based enrichment is the cheaper of the two technologies and shows a greater number of on target reads; however, the coverage of these regions is more uniform with hybrid sequencing [24,26]. Some commercially available amplicon platforms attempt to address the coverage issues by using specific primers that are able to amplify overlapping fragments in a single PCR reaction [27]. Amplicon based sequencing requires much less starting material than hybrid-capture, making it ideal if there is little DNA available for TS.

Hybrid-capture has been shown to produce fewer PCR duplicates than amplicon enrichment (<40% and up to ~80%, respectively) [24]. These duplicates are also more trivial to remove computationally, as the random shearing of the DNA in hybrid-capture platforms reduces the likelihood of two unique fragments aligning to the same genomic coordinates compared with the identical amplicons generated by amplicon enrichment platforms. This makes hybrid-capture especially useful for samples where these PCR artefacts are more likely to occur, such as FFPE and ctDNA samples. Further, certain regions of the genome make primer design for amplicon enrichment difficult (e.g. regions with a high number of repeated sequences). The long bait sequences used in hybrid-capture, however, allow a greater level of specificity in region selection. Overall hybrid-capture based platforms provide more accurate and uniform target selection, whilst amplicon-based platforms are often used in small scale experiments where sample quantity or cost are a factor.

2.2. Platforms for targeted sequencing

There are several commercially available platforms for these two approaches. Many of these platforms are also used for WES. An outline of these platforms is shown in Table 3. Despite the differences between the various platforms, they have been shown to lead to relatively concordant variant calling [24].

2.3. Use and design of panels for targeted sequencing

2.3.1. Targeted panel construction

The term targeted panel is used here to refer to the collection of genomic coordinates that are of interest to the user. An important difference between WES panels and targeted panels, is that TS is not constrained to canonical gene targets and can target other regions, such as promoters [28] or breakpoints [29].

There are commercially available targeted gene panels, usually designed for research [30,31] or clinical purposes [32,33]. They are designed to amplify genomic regions that are known to be of interest within cancer, or specific cancer subtypes. Using these panels greatly speeds up the process of the sequencing as they have already been designed, tested and validated.

Commonly, however, users design their own customised panels dependent on their research questions, although thorough target validation of these panels is needed before use. Customised panels are often generated by a thorough review of the current literature and cross referencing publicly available cancer mutation resources such as TCGA, ICGC, CbioPortal, and Catalogue of Somatic Mutations in Cancer (COSMIC) (http://cancer.sanger.ac.uk) databases [34–38], selecting genes that are frequently mutated, and targets that have been functionally validated in that cancer. In many cancer studies, an initial discovery cohort has been initially profiled with WGS or WES to the identify significantly mutated genes (via algorithms like MutSigCV [39], dNdScv [40], oncodriveFM [41]). These genes are then selected for TS with higher depth in the validation cohort(s) to establish their validity and frequencies [42–45]. Examples of the applications of these panels are included in the next section.

2.3.2. Applications of targeted gene panels in cancer studies

There are a large body of clinical studies that utilise genomic TS for research on clinical samples. Some recent examples have been listed in Table 4 [17,43–50], with targeted panels ranging from as few as 25 genes [44] to 122 genes [49]. These studies illustrate that a wide range of TS platforms, sequencing depths, data processing and variant calling methods were used.

Table 3
An overview of some commercially available TS platforms.

Platform	Company	Enrichment	Protocol overview
Ion AmpliSeq™	Thermo Fisher Scientific	Amplicon	Targeted regions are amplified through target specific primers. These primers are removed, the sequencing adapters are added and the amplicons are amplified again to generate the library. Needs to be sequenced using Ion Torrent™ Sequencer.
Access Array	Fluidigm	Amplicon	Amplifies target regions, adding an overhanging universal adapter. The universal adapter is then bound by the sequencing adapters. Can be sequenced on both Ion Torrent™ and Illumina platforms.
Haloplex^{HS}	Agilent	Amplicon	Circularises restriction enzyme fragmented gDNA using biotinylated probes. Probes captured using magnetic streptavidin beads. Circular molecules are then amplified to generate a linear library.
GeneRead DNAseq Targeted Panels V2	Qiagen	Amplicon	Targeted regions amplified via multiplexed PCR-based enrichment. The samples are pooled, and the amplicons are purified using AMPure XP beads. Sequencing library can then be created using a platform specific kit.
TruSeq Amplicon	Illumina	Amplicon	Probes are bound at either end of a targeted region. The region is amplified via PCR, leaving an amplicon of the region with probes either end. Indices and sequencing adapters are then bound to the overhanging ends of the probes.
SureSelect^{XT}	Agilent	Hybridization	Fragmented gDNA is amplified and the targeted regions are captured using target specific biotinylated probes. These probe-bound fragments are isolated and amplified to create the library.
SeqCap EZ	Roche Nimblegen	Hybridization	Fragmented gDNA is amplified, the sequencing adapters are added, and these fragments are then amplified. Target specific probes are added and probe bound fragments are isolated to generate the library.
Cell3™Target	Nonacus	Hybridization	DNA is enzymatically fragmented and Illumina Unique Molecular Identifier (UMI) containing adapters are ligated. The fragments are amplified prior to target enrichment using biotin-labelled probes and streptavidin coated beads. The enriched fragments are amplified again and then sequenced on an Illumina sequencer.

3. Guidance for analysis of targeted genomic sequencing

In this section we provide detailed guidance for the analysis of TS, from initial quality control (QC) and data pre-processing, to variant calling, annotation and filtering (Fig. 2). Commonly used methods and software in each step and important parameters/filters are discussed, aiming to provide readers a comprehensive overview of the whole analytical process from raw reads to high-confidence annotated calls. We further focus on PCR duplication marking/removal and variant filtering in greater depth, as these are crucial steps to ensure the best quality variant calls. Key steps of TS data analysis and commonly used software are listed in Table 5.

3.1. Quality control and data pre-processing

3.1.1. QC and alignment

The first step of all NGS pipelines is to assess the quality of the sequenced reads, using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). It summarises and visualises base quality score for every base pair sequenced, which allows users to have an overview of the read quality and decide whether a trimming step is needed, especially at the 3′ end where the base quality is often lower. FastQC also produces summarised information of adapter fragment contamination and GC content within all reads. This analysis determines whether adapter fragments have been incorporated into the reads and need to removed using software such as CutAdapt [51]. The GC content of the reads is useful to indicate whether the sample is contaminated with DNA from another

Table 4
Selected example of studies that analysed mutations using targeted DNA sequencing in human samples.

Disease	Tissue Origin	Authors	Journal	Genes	Depth	Platform	Target capture mode	Machine	FFPE/ fresh frozen (FF)	Duplicate handling	Variant Calling
Acute Myeloid Leukaemia	Tumour	Ivey et al. 2016	New England Journal of Medicine	51	1280x	Agilent Haloplex^{HS}	Amplification	HiSeq 2000	Not Reported	Not Reported	VarScan2
	Normal Peripheral blood	Abelson et al. 2018	Nature	111	Not Reported	Roche NimbleGen Agilent SureSelect	Hybrid Capture	HiSeq 2000	FF	MBC	Varscan2 Shearwater ML Pindel
Breast Cancer	Tumour	Ellis et al. 2012	Nature	Variable	Not Reported	Roche NimbleGen	Hybrid Capture	Not Reported	FFPE	Picard	VarScan2 BreakDancer
	Germline	Couch et al. 2015	Journal of Clinical Oncology	122	300x	Illumina TruSeq Amplicon	Amplification	HiSeq™ 2000	Not Reported	Not Reported	GATK Unified Genotyper/ SAMtools
FL	Tumour	Okosun et al. 2014	Nature Genetics	28	840x	Fluidigm Access Array™	Amplification	Miseq	FF	Not Reported	VarScan2
	Tumour	Pastore et al. 2015	The Lancet Oncology	74	Not Reported	Agilent SureSelect	Hybrid Capture	HiSeq 2500	FFPE	Picard	MuTect Indel Locator
	Tumour	Araf et al. 2018	Leukaemia	25	8000x	Agilent Haloplex^{HS}	Amplification	MiSeq	FFPE	UMI	VarScan2
Pancreas	Tumour	Sausen et al. 2015	Nature Communications	116	754x	Agilent SureSelect	Hybrid Capture	HiSeq 2000/ 25000 & MiSeq	Both	CASAVA	VariantDx
Skin Cancer	Normal Skin	Martincorena et al. 2015	Science	74	500x	Agilent SureDesign	Hybrid Capture	HiSeq 2000/ 25000	FF	Picard	Shearwater ML

organism, as this would likely lead to a secondary peak due to the different GC content of that genome [63].

Next, raw or trimmed reads are aligned to the reference genome to generate Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) files for each sample. Commonly used aligners include the Burrows-Wheeler Aligner (BWA) [53] and Bowtie2 [54]. Ion Torrent™ also have their own customised aligner specifically for working on data generated from their platform. Within alignment files the mapping quality score (i.e., the likelihood of a read mapping to multiple locations in the genome) is recorded for each read, in addition to their mapped coordinates.

It should be noted that the experimental and web-lab quality of TS experiments is also a key determinant of the sequencing data quality, such as how fragmented the DNA is, and the amount of input DNA. Low quantity of input DNA will require more PCR cycles, leading to a high level of PCR duplicates and limiting the achievable depth of coverage of the experiment. Monitoring the experimental quality of TS is always part of good laboratory practice, ensuring the highest quality of sequencing data in the downstream analyses. It is also important to check for germline/tumour mix-ups and contamination whilst running the pipeline. Whilst these errors are very difficult to determine from the FASTQ files alone, they may become more apparent in the later analytic stages, such as variant calling and VAFs, e.g. a large number of variants called in the germline that are absent in the tumour sample.

3.1.2. Assessment of off-target reads

Various QC steps should always take place to ensure the best quality of TS data. As TS focuses on regions of interest in the design panel, we expect the majority of reads generated should come from targeted regions, however, off-target reads are a common occurrence. After alignment, the percentage of reads that cover targeted regions can be assessed using software such as bedtools [52], and the GATK coverage module. A high proportion of off-target reads may indicate that the TS experiment has failed, or the targeted regions contain too many repeat sequences. This could be possibly adjusted by making the capture or library preparation process more efficient, e.g., adjust input DNA to beads ratio, and wash more stringently. With a large panel of hundreds of targeted genes, roughly >70% of the reads aligning to the targeted regions is a positive indicator of a good quality TS data set [26].

3.1.3. Marking and removal of PCR duplicates

PCR duplicates are sequence reads that align to the same genomic coordinates and typically arise during PCR steps in the library preparation. The duplication rate tends to be much higher for fragmented DNA of low quality, e.g. FFPE and ctDNA, reaching ~50–60% for some cases, while for FF DNA, this rate is usually less than 20%. These PCR duplicates need to be marked and removed before any downstream analysis, as including them will lead to overestimation of coverage in targeted regions, and more importantly result in incorrect allele frequency estimation.

A number of software are used to search for PCR duplicates within aligned NGS data. A commonly used program is the MarkDuplicates function within Picard Tools (http://broadinstitute.github.io/picard/). This tool looks for reads with the same start and end coordinates and then add tags to the bam files that mark these reads as duplicates. Another tool, SAMtools rmdup, simply outright removes the duplicate reads retaining the read with the highest mapping quality [55]. However, these software based attempts cannot discriminate between two unique reads that happen to align in the same position by chance and actual duplicates [64]. There are additional molecular techniques, such as Unique Molecular Identifiers or Molecular Barcodes (MBC), available to ensure only unique reads are measured in the downstream analysis. These are exemplified by the Nonacus Cell3™ Target, Agilent Haloplex^{HS} and SureSelect^{XT} platforms.

UMIs or MBCs are random short nucleotide sequences that are ligated to the DNA fragments during the library preparation. These sequences act as barcodes that mark each read as coming from the amplification of a single fragment, providing a more accurate mechanism for determining PCR duplicates. The different methods

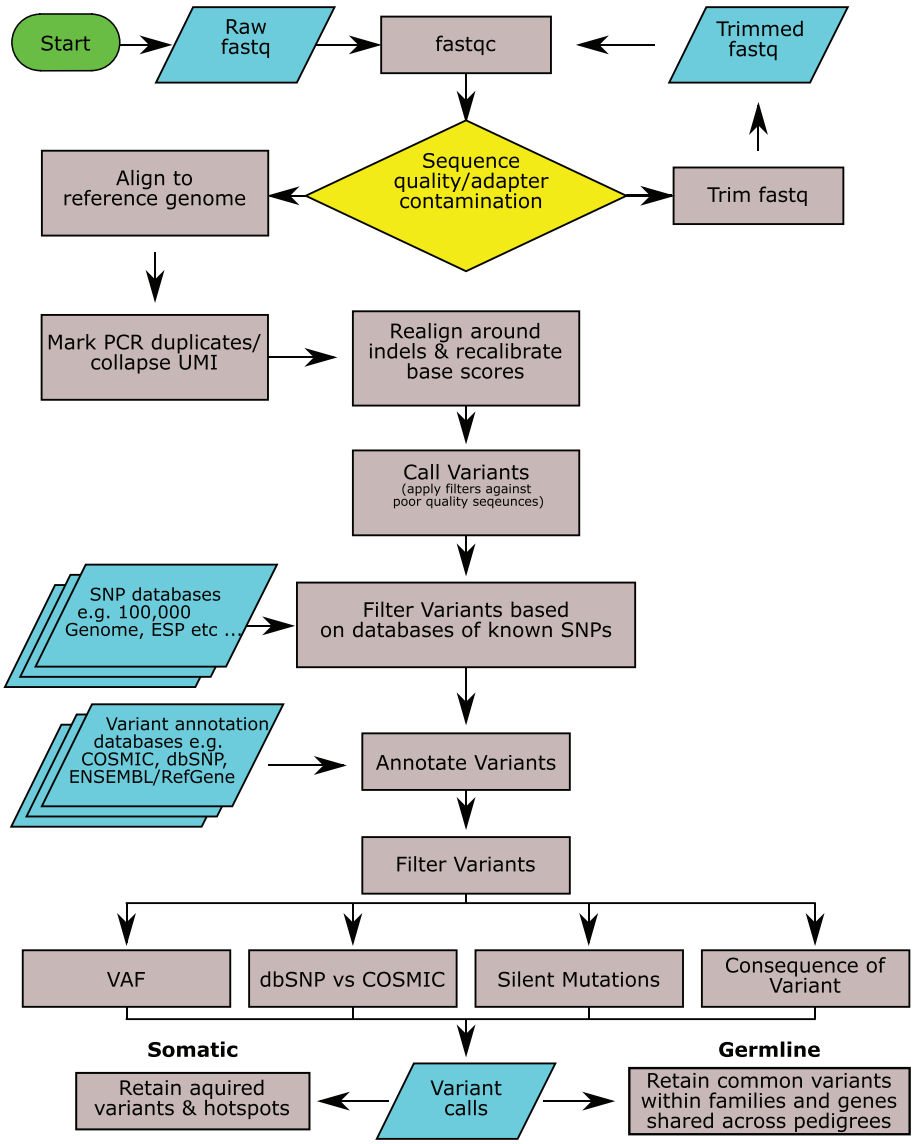

Fig. 2. A generalised workflow for calling variants in clinical samples. This workflow includes quality check, sequence alignment and further processing, variant calling, annotation and filtering.

Table 5
Steps and commonly used software for the processing of targeted sequencing data.

Step	Software
QC	FastQC, CutAdapt [51], bedtools [52]
Alignment	BWA [53], Bowtie2 [54], Torrent Suite
PCR Duplicates Handling or Unique Molecular Identifier /Molecular Barcode (MBC) deconvolution	Duplicates - Picard, SAMtools [55], Torrent Suite UMI/MBC – fgbio, Agilent Genomics NextGen Toolkit, Gencore, Connor
Realignment and base score recalibration	Genome Analysis Tool Kit (GATK) [56]
Variant calling	MuTect2 (GATK), Strelka2 [57], VarScan2 [58], HaplotypeCaller (GATK), Torrent Suite, Pindel [59], Ion Reporter Software, VariantDX
Annotation	Annovar [60], snpEFF [61], Variant Effect Predictor [62]

of PCR duplicate handling are outlined in Fig. 3. An outline of the additional steps required for UMI/MBC workflow are as follows:

- Align the reads to the reference genome first as usual, with all the barcodes contained in a separate file.
- The reads are then grouped by their barcodes to ensure the duplicate reads are found next to one another.
- The reads with the same UMIs are then collapsed to create consensus reads, with all duplicated reads removed. Available programs to deal with UMIs or MBCs include 'fgbio' (http://fulcrumgenomics.github.io/fgbio), The Agilent Genomics NextGen Toolkit (AGeNT) (Agilent Technologies, http://www.genomics.agilent.com), Gencore (https://github.com/OpenGene/gencore) and Connor (https://github.com/umich-brcf-bioinf/Connor).

Fig. 4 demonstrates the effect of UMI duplicates from FFPE and FF samples from follicular lymphoma biopsies (unpublished in-house data, used here for demonstration purpose only). These samples were processed using a custom hybrid capture panel from Nonacus run in-house at BCI. The number of duplicate reads found to share the same UMI were counted for each sample. While FF samples had an average of 2–3 duplicated reads per consensus sequence, FFPE samples had a far greater number of PCR dupli-

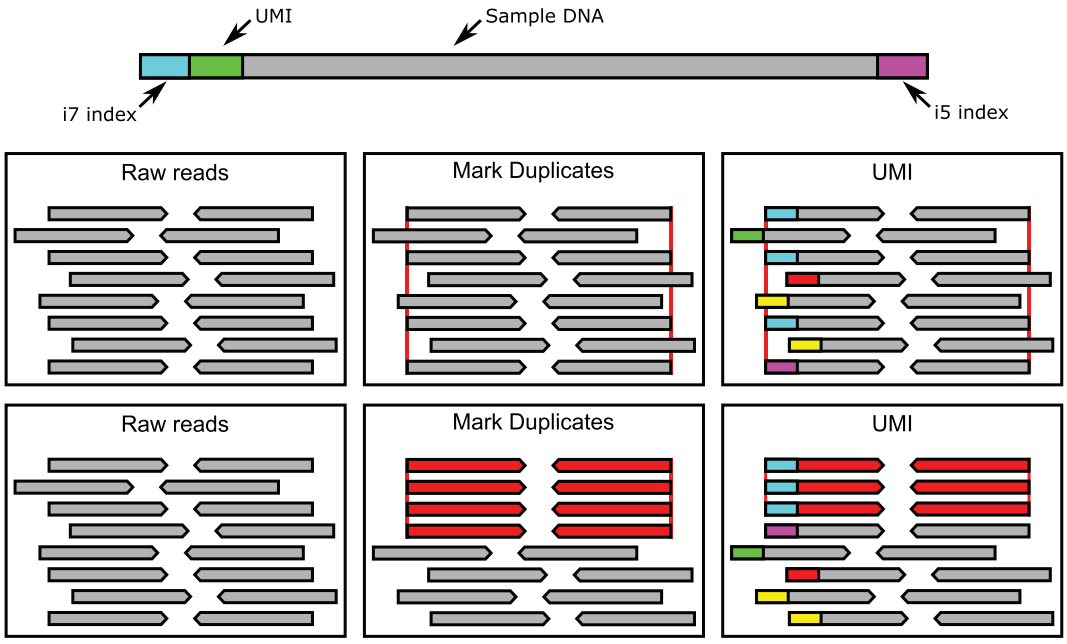

Fig. 3. Considering duplicates in next generation sequencing. PCR duplicates can occur during the course of NGS. Whilst duplicates will appear to be separate reads, they are actually technical noise due to errors during PCR and sequencing. The two methods of correcting these errors are detailed above. The red lines indicate reads the start and end coordinates of the duplicates. Reads are coloured based on whether they are considered individual reads (grey) or duplicates (red), the coloured bars at the start of each read in the UMI panel represent different UMI sequences. In the above situation marking duplicates would cause 4 reads to be combined into a single read whereas the UMI based duplicate method is able to distinguish between true duplicates and unrelated reads with the same coordinates.

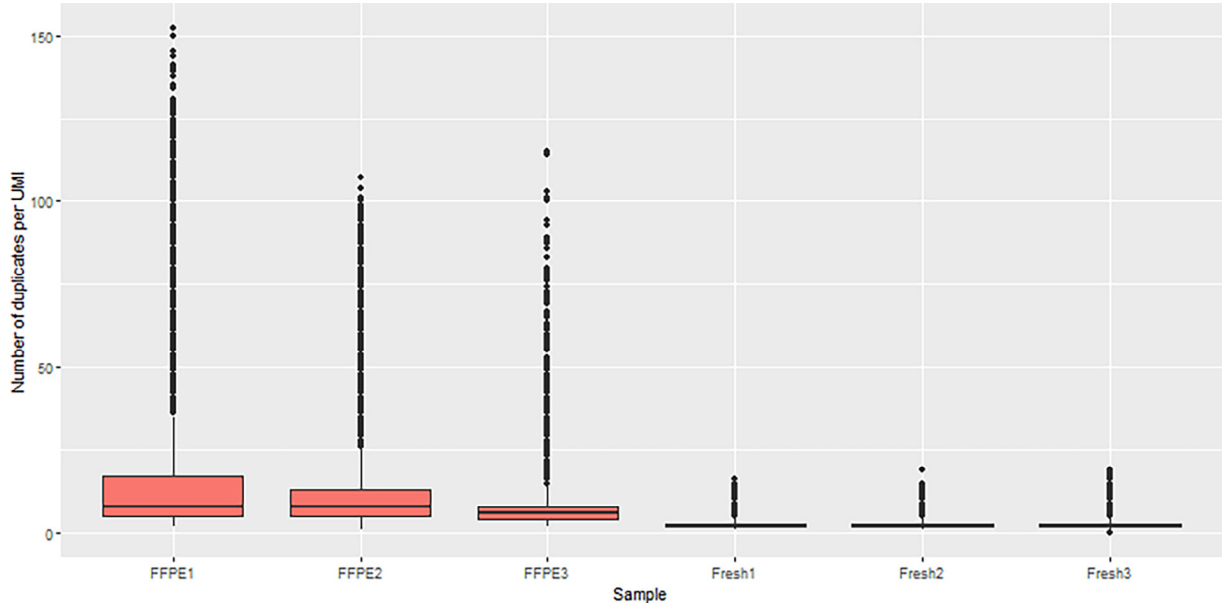

Fig. 4. The number of UMIs found in common across FFPE and FF clinical samples. FF and FFPE follicular lymphoma biopsies were sequenced using the Nonacus hybrid capture platform (unpublished in-house data for demonstration purposes). The number of UMI tagged duplicates that were found in these samples were counted. Only consensus reads with at least two duplicate reads were considered in this analysis.

cates, with some consensus sequences having >50–100 duplicate reads. This is likely due to the increased amplification needed to produce enough DNA from FFPE samples combined with the reduced quality of DNA extracted from FFPE samples. Here our case clearly shows that the usage of UMI or MBC can greatly increase the accuracy of detection of low-quality DNA with much improved allelic quantifications, e.g., for FFPE and ctDNA.

3.1.4. Realignment, base score recalibration and estimation of sequencing coverage

Next, filtered alignments are further processed to improve the alignment quality, including local realignment around indels and base quality score recalibration using GATK. The step of local realignment is to improve the alignment quality for bases around known and suspected indel positions to reduced false positive

calls. Base score recalibration is carried out to recalculate base quality scores for all sequenced reads based on known polymorphisms (e.g., SNPs from 1000G Project). The base and mapping quality scores are used to filter reads during variant calling and the fine-tuning that occurs in this step is important to ensure only high-confidence variants are called.

Base coverage information is another important parameter to assess the overall quality of TS data. Using recalibrated BAM files, one can further calculate the coverage/depth for bases within the targeted regions, using Bedtools or GATK coverage. Depending on the quality of DNA and total number of reads generated, several hundred times depth per base is often expected, although some regions may have much higher coverage or targeted rates than others. However, for ultra-deep sequencing, the depth of tens of thousands of reads is often required to detect very low frequency clones.

3.2. Variant calling

Once all TS pre-processing steps are completed, these high-quality alignment data are ready for variant calling. Variant calling is the process of comparing the aligned reads to a reference genome or matched normal DNA sequences to identify base pair variations. Here we describe the procedure for samples with matched normal and without matched normal separately. We then focus on variant calling parameters and filters which can be tuned accordingly to achieve the best outcome.

3.2.1. A note on paired end sequencing

Whilst paired end sequencing improves the accuracy of read alignment, it can lead to overlapped regions within read pairs. This problem is especially prevalent with shorter DNA fragments. As these paired reads are from the same DNA fragment, the overlapping sections are duplicates of one another and will lead to the bases in this region being counted twice. To combat this source of error most variant callers have built in methods to handle overlaps. For example, SAMtools mpileup will set the mapping quality of one of the overlapped reads to 0 (i.e., not mapped) within this region, to ensure it is not included in the variant calling.

3.2.2. Samples with matched normal

Normal refers to DNA extracted from the non-tumour tissue of the cancerous organ, however it is often that "blood derived DNA" is used as a germline control when normal tissue is unsuitable or unavailable. Five of the studies shown in Table 4 included matched tumour/normal samples [17,22,34,46,48]. This allows for the patient specific germline SNPs to be identified and disregarded in the tumour sample [57,65]. Commonly used software for somatic variant calling includes MuTect2 [56], Strelka2 [57] and VarScan2 [58] somatic mode. These methods process normal and tumour alignment BAM files together, first by calling variants against a reference genome before determining somatic variants based on sophisticated models (e.g., mixture model), taking into consideration of factors like depth, error rate and haplotype to call high-confidence variants. However, one can always further filter produced variants using the total coverage and number of supporting reads for the sites [66–68].

Studies comparing multiple variant callers found poor overlap results between different methods [66,67]. MuTect2 and Strelka2 seemed to perform well compared to their contemporaries, and also display a good level of concordance with one another in their SNV calls (~90%), although their indel calls were much less concordant (~55%) [68]. Thus, in some studies, variants were called using multiple callers and only these supported by at least two methods were selected [68,69].

3.2.3. Samples without matched normal

However, matched normal samples are often not available in the clinical setting, especially for retrospective FFPE or FF samples. Consequently, it is much more challenging to call reliable somatic variants in this scenario, confounded by the presence of a large number of germline SNPs. As with matched samples, variant calling starts with comparison to the reference genome identifying all possible variant positions, including germline and potential somatic calls, using VarSan2 or SAMtools [55]. Both methods can work on one or multiple sample alignments in a 'pileup' format generated by SAMtools, and call variants against the reference genome with filters, such as the minimum base and mapping quality, the number of supporting reads and total coverage for called sites. Another commonly used method is GATK, where the HaplotypeCaller module can be used to call variants for multiple samples effectively. HaplotypeCaller is able to call SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active region. In regions with many variations detected, HaplotypeCaller reassembles the reads in that region without the use of existing mapping information. This makes the calls much more accurate, especially for different types of variants close to each other. Various variant annotation and filtering steps are then applied, outlined below, to remove low-confidence and non-significant calls, which we will expand in the following sections. Without the matched normal samples, germline variants cannot be removed as above and must be thoroughly processed to remove commonly or benign variants that are less likely to originate from the tumour. In many studies, although the matched normal samples were not available, a panel of normal DNA samples were included to help significantly remove SNPs and sequencing artefacts [15,17,70]. In order to increase the validity of normal samples, they should be as closely matched as possible to the study cohort e.g. sex and ethnicity.

3.2.4. Calling inherited germline variants

The pipeline presented in Fig. 2 is versatile, and can also be applied for germline variant calling. However, when calling somatic variants, positions that are called in common between samples are highly likely to be SNPs/artefacts and should be removed, unless they are mutational hotspots. In contrast when calling germline variants these recurring calls should be kept if they are found within members of the same pedigree as these are likely to be inherited variants. An additional consideration with germline calls is that VAFs for these variants should be at ~50% (heterozygous, using a VAF range of 30–70% or 40–60% depending on the sequencing depth) or ~100% (homozygous). Any 'sub-clonal' variants should be ignored in germline variant calling. Furthermore, any genes shared across different pedigrees should be regarded as important familial gene candidates.

3.2.5. Variant calling parameters and filtering

A set of important parameters need to be considered for variant calling and filtering for high-quality calls. These include,

- Number of total reads: this parameter can be used to ensure there is sufficient coverage over the position for variants to be called. Often a minimum of 20-30x depth is required for TS [71–75].
- Number of variant supporting reads: this parameter should be set in order to limit variants with very few supporting reads being considered. The value can be tuned based on the average coverage of the samples. Usually the minimum value ranges from 4 to 10 reads [26,47,76].
- Minimum base and mapping quality score: Setting a threshold for base and mapping quality scores stops poorly sequenced or aligned reads from being considered in the variant calling.

The default minimum values of many programmes are set as 20–30 as these correspond to an accuracy of 99% and 99.9% respectively.

- Minimum allele frequency for called variants: Like the number of variant supporting reads, this can be used to eliminate variant positions with low levels of support. Often, a relatively low threshold (e.g., 3% with a depth of 200x) is initially used to include most of the variants, and further filtering and refinement are performed via testing a range of threshold values to choose the best cutoff value for VAF. For FFPE samples, the final threshold is set as at least 10% or even 20% across many studies [77,78]. For FF samples this threshold can be much lower depending on overall sequencing depth [46,50]. One should note that the tumour purity of clinical samples is often highly heterogeneous. Thus, filtering simply based on an observed VAF cutoff may not provide the most accurate way to include high-quality or exclude low-quality calls. One way to overcome this is to further adjust VAF values based on the estimates of tumour purity of clinical samples, and apply the threshold on these adjusted VAFs to filter calls for the downstream analyses. When an accurate measurement of tumour purity is not available, VAFs of mutations in many known clonal driver genes (e.g., *KRAS* and *TP53* for many solid tumours) could be used to derive a rough estimate.

Additional parameters also include:

- Strandedness of variant supporting reads: If a variant occurs within a sample, paired sequencing should show evidence of this variant on both strands. Therefore if the majority of the reads for a variant occur on only one strand (i.e., strand bias), it could suggest that variant reads are artefacts [58,76]. In many programmes, at least one supporting read is required to be present on each strand for the called variants. In VarScan2, it is possible to require that a maximum of 90% of all reads (across reference and alternative alleles) are found on one strand, meaning positions that have a strand bias will be ignored.
- Significance score for a statistical test: Many variant callers will calculate a statistical evaluation of the likelihood of a variant differing from the reference allele [47,76]. VarScan2 for example provides the user with a *p* value for a Fisher's Exact Test on the observed and expected variant reads. This can be used to further eliminate low-quality calls.

These parameters can be fine-tuned based on the aims of the project and the data that is generated. Among the publications reviewed in Table 4, for example, the high depth of sequencing in Araf et al. (estimated at ~8000×) combined with error suppression allowed variants to be called with VAFs as low as 0.1% [44]. Elis et al. initially called variants of VAFs as low as 2% in matched FFPE samples, utilising the validation mode within VarScan2, followed by further customised filtering to retain high quality calls [48]. In defining the m7-FLIPI index a VAF cut-off of >10% was used. These data were generated from FFPE samples and many samples had no matched germline to filter out SNPs, meaning a robust cut-off was necessary to ensure high quality calls [17].

3.3. Annotation and further filtration of variants.

Following variant calling, the next step is to annotate the variants in relation to genes (e.g., within or outside a gene), codon and amino acid positions, and classify types of variants, such as nonsense, missense, exonic deletions and synonymous variants. This allows for greater understanding of their functional consequences on genes they relate to. In many TS studies, only non-silent exonic or splicing mutations are selected for further analysis, focusing on functional coding variants and mutations only. However, these criteria may vary depending on the region of interest or the purpose of the study, e.g. variants in promoter regions or UTRs of the genome.

Commonly used variant annotation methods include ANNOVAR, SNPeff, VEP and Oncotator [60–62,79]. These methods provide rich resources of gene and regulatory annotation, functional prediction, sequence conservation and frequencies in the population level. Here we describe a general annotation and filtering workflow for variants called in a cancer TS experiment without matched normal tissue. The workflow follows as below,

1. Gene annotation: annotate variants against Ensembl or RefGene gene models, to retain all non-silent variants including those affecting splice sites and exonic indels.
2. SNP and cancer variant identification and filtering: Find variants that are overrepresented in the general population. Datasets such as dbSNP, 1000 Genome Project, NHLBI GO Exome Sequencing Project (ESP), The Genome Aggregation Database (gnomAD) [80] and ExAC [81] include the estimated frequency of variants. Any variants with minor allele frequency >1% are excluded, as these more common variants are less likely to have any oncogenic implications. Filtered variants are then annotated against the COSMIC database (a cancer mutation catalogue), allowing those variants present in dbSNP but also previously identified as cancer mutations to be retained.
3. Variant recurrence filtering: the remaining non-silent variants are still likely to contain many SNPs and sequencing artefacts. Specifically, variants that occur in many samples (e.g., >15/20% of samples) but are not known COSMIC hotspots are likely these candidates for removal. When VAFs of those variants are consistently low (e.g., <5% when UMIs are not used) across all samples, these typically represent sequencing artefacts. When recurrent variants have consistently high VAFs (over 30/40%) across all samples, this suggests that they are likely SNPs. A panel of normal samples (unmatched) sequenced alongside the tumour samples can significantly aid in reducing these recurrent variants if they also occur in the normal controls.
4. Variant and gene prioritisation: functional consequences of variants are predicted using databases such as SIFT [79], PolyPhen [80], and MutationTaster [82]. Highly scored variants are likely to have strong deleterious effects on the targeted genes, warranting further investigation. Genes with deleterious variants that are over-represented across the cohort, are potentially strongly involved in the biology of that cancer. However, care must be taken when selecting candidates for further study as confounding factors can also cause a high level of mutations in individual genes, e.g. gene length. Commonly used programmes to detect significantly mutated genes (e.g., MutSigCV and dNdScv) can still be applied to TS data to prioritise candidate genes.

3.4. Estimation of background error rate

The sensitivity of NGS is in the regions of VAF 1% [83,84]. However, there is a need in some studies to identify variants with much lower VAFs, e.g., to detect very small subclonal and minimal residual disease (MRD) mutations. To achieve this, higher depth of sequencing is usually required, and a comprehensive strategy is needed to differentiate between genuine calls and background sequencing artefacts or the background noise rate at VAFs < 1%. Tawana et al. applied ultra-deep TS (depth of 10,000–100,000×) to investigate pre-existing leukaemic clones and disease evolution in sequential acute myeloid leukaemia biopsies [16]. Two independent strategies were used to account for the noise level: first, the

reference and variant allele supporting reads of targeted variants were compared among sequential samples and also with a panel of non-related DNA, to ensure these were not recurrent sequencing artefacts; second, the reads for the variants of interest were also compared with those of variants detected within surrounding base pairs, to exclude false positive calls due to background noise, with the background noise rate also calculated at around 0.20%. This successfully led to the discovery of a small clone (3% of cells) harbouring a *TET2* nonsense mutation, which expanded and became the dominant clone at a later stage.

There are also software packages available that try to control for background mutation rate in non-matched samples using a panel of normals [85,86]. Integrated digital error suppression (iDES) is one such method that utilises a combination of CAPP-Seq molecular barcodes and background 'polishing' that is able to reduce the error rate further than either method used in isolation [86]. The molecular barcodes allowed an *in silico* reassembly of the original DNA duplex reducing sequencing artefacts, whilst the polishing was carried out using a novel method, which removed variants that were statistically indistinguishable from background levels found in a panel of normals. Whilst combining the methods resulted in the best improvement in background error rate reduction (~15 fold) the polishing alone was shown to improve the error by ~3 fold, similar to the effect of the molecular barcodes. Therefore, the iDES polishing alone could be easily included in existing variant calling pipelines to reduce the error rate. The iDES software can be found at: http://cappseq.stanford.edu/ides/.

4. Summary and outlook

TS is a powerful and invaluable tool for mutational detection, and it has been widely applied in cancer research and clinical studies across many cancer types. Compared to its counterparts WGS and WES, TS can screen a large number of samples at much reduced costs and computational burden. This makes it extremely attractive for clinical research with fast turnaround. Until the WGS/WES cost drops to an affordable rate for large-scale applications, TS will continue to be the main genomic tool in disease genotyping. The capability of TS to detect subclonal mutations, sequencing ctDNA and for minimal residual disease monitoring also makes it a useful genetic tool to track disease evolution and study drug resistance.

However, the use of TS in routine clinical practice is still in its infancy. Whilst these data demonstrate TS can generate clinically relevant results, the key question remains whether TS can be used as a stand-alone genomic diagnostic tool. We believe that this depends on the clinical questions under investigation. When clonal mutations are explored for diagnosis and targeted therapies, TS is accurate with a normal depth of 300-500x. When subclonal events and/or MRD serve as the focus, we recommend that TS should be validated and interpreted with other approaches (e.g., digital PCR). As shown above rarer events can be observed with very high depth sequencing, but the levels of sequencing artefacts will also increase. In these cases, the estimation of the background noise level is crucial in determining an appropriate cutoff for acceptable variants. However, for FFPE samples, we argue that a cutoff of VAF value of 10% should be implemented in cases where these samples are investigated for diagnosis and prognosis due to the poor DNA quality of these samples. Note that tumour purity should also be accounted for if necessary.

One should also be aware of the limitations of TS. Due to its nature, targeting pre-defined genes, it is less useful or efficient for the detection of large-scale rearrangements (e.g., structural variants) and copy number changes, compared to the whole-genome profiling (e.g., WGS/WES). However, for known common translocation events (e.g., t(14;18) in follicular lymphoma), one can still design primers or probes to capture regions spanning the breakpoints, and TS should be able to detect these events by identifying reads that cleanly span the breakpoints. For copy number changes, normalised sequencing depth of coverage can still be used to infer copy number status, using software such as CONTRA [87] and SeqCNV [88]. However, this still remains challenging, strongly determined by the TS quality and uniformity of coverage across genes. Although not suitable for the detection of novel genomic events, TS remains as a powerful and economical tool to identify known events in patients.

The current challenge and bottleneck for large-scale cross-centre TS applications is the lack of gold-standard methods for identifying cancer-associated mutations. Individual laboratories tend to develop their own pipelines with different parameters used, often leading to a poor level of overlapped results. Thus, there is an urgent need for reliable and standard data processing and mining methods that can bring TS into routine clinical practice. We argue that benchmarking studies are urgently needed to address this issue. Initiatives such as the ICGC-TCGA DREAM Genomic Mutation Calling Challenge on WGS data are first steps in this direction [89]. Once we have the standard off-the-shelf TS analysis methods and pipelines accepted by the community, these can then be widely used across many research and clinical settings.

Authors contributions

F.B-C and J.W. conceived and designed the study. F.B.C performed the study and collated the data. F.B.C and J.W. wrote the manuscript. All authors contributed to the revision of the manuscript.

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.

Acknowledgement

All Authors acknowledge support from Cancer Research UK Centre of Excellence Award to Barts Cancer Centre [C16420/A18066].

References

[1] Beigh M. Next-generation sequencing: the translational medicine approach from "Bench to Bedside to Population". Medicines 2016;3:14. https://doi.org/10.3390/medicines3020014.
[2] Mestan KK, Ilkhanoff L, Mouli S, Lin S. Genomic sequencing in clinical trials. J Transl Med 2011;9:1–10. https://doi.org/10.1186/1479-5876-9-222.
[3] Masunaga N, Kagara N, Motooka D, Nakamura S, Miyake T, Tanei T, et al. Highly sensitive detection of ESR1 mutations in cell-free DNA from patients with metastatic breast cancer using molecular barcode sequencing. Breast Cancer Res Treat 2018;167:49–58. https://doi.org/10.1007/s10549-017-4487-y.
[4] Kaderbhai CG, Boidot R, Beltjens F, Chevrier S, Arnould L, Favier L, et al. Use of dedicated gene panel sequencing using next generation sequencing to improve the personalized care of lung cancer e20523–e20523. Oncotarget 2016;7. https://doi.org/10.18632/oncotarget.8391.
[5] Tsimberidou A, Iskander NG, Hong DS, Wheler JJ, Falchook GS, Fu S, et al. Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative apostolia-maria. Clin Cancer Res 2012;18:6373–83. https://doi.org/10.1158/1078-0432.CCR-12-1627.
[6] The AARC Project GENIE Consortium, Abstract. AACR Project GENIE: powering precision medicine through an international consortium. Cancer Discov 2017;7:818–31. https://doi.org/10.1158/2159-8290.CD-17-0151.
[7] Cardarella S, Ogino A, Nishino M, Butaney M, Shen J, Lydon C, et al. Clinical, pathologic, and biologic features associated with BRAF mutations in non-small cell lung cancer. Clin Cancer Res 2013;19:4532–40. https://doi.org/10.1158/1078-0432.CCR-13-0657.

[8] Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat J-P, et al. A landscape of driver mutations in melanoma. Cell 2012;150:251–63. https://doi.org/10.1016/j.cell.2012.06.024.

[9] Lochhead P, Kuchiba A, Imamura Y, Liao X, Yamauchi M, Nishihara R, et al. Microsatellite instability and braf mutation testing in colorectal cancer prognostication. J Natl Cancer Inst 2013;105:1151–6. https://doi.org/10.1093/jnci/djt173.

[10] Lovly CM, Dahlman KB, Fohn LE, Su Z, Dias-Santagata D, Hicks DJ, et al. Routine multiplex mutational profiling of melanomas enables enrollment in genotype-driven therapeutic trials. PLoS ONE 2012;7. https://doi.org/10.1371/journal.pone.0035309.

[11] Wala J, Bandopadhayay P, Greenwald N, O'Rourke R, Sharpe T, Stewart C, et al. Genome-wide detection of structural variants and indels by local assembly. BioRxiv 2017;105080. https://doi.org/10.1101/105080.

[12] Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. N Engl J Med 2004;350:2129–39. https://doi.org/10.1056/NEJMoa1402685.

[13] Lindeman NI, Cagle PT, Beasley MB, Chitale DA, Dacic S, Giaccone G, et al. Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors guideline from the college of American pathologists. Int J Thorac Oncol 2013;8:823–59. https://doi.org/10.1097/JTO.0b013e318290868f.

[14] Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med 2015;21:751–9. https://doi.org/10.1038/nm.3886.

[15] Shin H, Choi Y, Yun JW, Kim NKD, Kim S, Jeon HJ, et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nat Commun 2017;8:1–10. https://doi.org/10.1038/s41467-017-01470-y.

[16] Tawana K, Wang J, Renneville A, Csaba B, Van Delft FW, Treleaven J, et al. Disease evolution and outcomes in familial AML with germline CEBPA mutations. Blood 2015;126:1214–24. https://doi.org/10.1182/blood-2015-05-647172.

[17] Pastore A, Jurinovic V, Kridel R, Hoster E, Staiger AM, Szczepanowski M, et al. Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. Lancet Oncol 2015;16:1111–22. https://doi.org/10.1016/S1470-2045(15)00169-2.

[18] Oh BY, Shin H-T, Yun JW, Kim K-T, Kim J, Bae JS, et al. Intratumor heterogeneity inferred from targeted deep sequencing as a prognostic indicator. Sci Rep 2019;9:4542. https://doi.org/10.1038/s41598-019-41098-0.

[19] Li J, Meeks H, Feng BJ, Healey S, Thorne H, Makunin I, et al. Targeted massively parallel sequencing of a panel of putative breast cancer susceptibility genes in a large cohort of multiple-case breast and ovarian cancer families. J Med Genet 2016;53:34–42. https://doi.org/10.1136/jmedgenet-2015-103452.

[20] Xu F, Wu LY, Chang CK, He Q, Zhang Z, Liu L, et al. Whole-exome and targeted sequencing identify ROBO1 and ROBO2 mutations as progression-related drivers in myelodysplastic syndromes. Nat Commun 2015;6.. https://doi.org/10.1038/ncomms9806.

[21] Wang L, Jing Y, Yu L, Li YY, Lv N, Bo J, et al. Implications of mutational spectrum in myelodysplastic syndromes based on targeted next-generation sequencing. Oncotarget 2017;8:82475–90. https://doi.org/10.18632/oncotarget.19628.

[22] Parry M, Rose-Zerilli MJJ, Ljungström V, Gibson J, Wang J, Walewska R, et al. Genetics and prognostication in splenic marginal zone lymphoma: Revelations from deep sequencing. Clin Cancer Res 2015;21:4174–83. https://doi.org/10.1158/1078-0432.CCR-14-2759.

[23] Wu S, Ou T, Xing N, Lu J, Wan S, Wang C, et al. Whole-genome sequencing identifies ADGRG6 enhancer mutations and FRS2 duplications as angiogenesis-related drivers in bladder cancer. Nat Commun 2019;10:720. https://doi.org/10.1038/s41467-019-08576-5.

[24] Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, et al. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. Hum Mutat 2015;36:903–14. https://doi.org/10.1002/humu.22825.

[25] Mertes F, ElSharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 2011;10:374–86. https://doi.org/10.1093/bfgp/elr033.

[26] Hung SS, Meissner B, Chavez EA, Ben-Neriah S, Ennishi D, Jones MR, et al. Assessment of capture and amplicon-based approaches for the development of a targeted next-generation sequencing pipeline to personalize lymphoma management. J Mol Diagnostics 2018;20:203–14. https://doi.org/10.1016/j.jmoldx.2017.11.010.

[27] Schenk D, Song G, Ke Y, Wang Z. Amplification of overlapping DNA amplicons in a single-tube multiplex PCR for targeted next-generation sequencing of BRCA1 and BRCA2. PLoS ONE 2017;12:1–16. https://doi.org/10.1371/journal.pone.0181062.

[28] Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 2017;23:703–13. https://doi.org/10.1038/nm.4333.

[29] Kim HK, Park WC, Lee KM, Hwang HL, Park SY, Sorn S, et al. Targeted next-generation sequencing at copy-number breakpoints for personalized analysis of rearranged ends in solid tumors. PLoS ONE 2014;9.. https://doi.org/10.1371/journal.pone.0100089.

[30] Barrio S, Stühmer T, Da-Viá M, Barrio-Garcia C, Lehners N, Besse A, et al. Spectrum and functional validation of PSMB5 mutations in multiple myeloma. Leukemia 2018:8–9. https://doi.org/10.1038/s41375-018-0216-8.

[31] White BS, Lanc I, O'Neal J, Gupta H, Fulton RS, Schmidt H, et al. A multiple myeloma-specific capture sequencing platform discovers novel translocations and frequent, risk-associated point mutations in IGLL5. Blood Cancer J 2018;8. https://doi.org/10.1038/s41408-018-0062-y.

[32] Steward DL, Carty SE, Sippel RS, Yang SP, Sosa JA, Sipos JA, et al. Performance of a multigene genomic classifier in thyroid nodules with indeterminate cytology. JAMA Oncol 2018;15213:204–12. https://doi.org/10.1001/jamaoncol.2018.4616.

[33] Nikiforova MN, Mercurio S, Wald AI, Barbi de Moura M, Callenberg K, Santana-Santos L, et al. Analytical performance of the ThyroSeq v3 genomic classifier for cancer diagnosis in thyroid nodules. Cancer 2018;124:1682–90. https://doi.org/10.1002/cncr.31245.

[34] Chen K, Meric-Bernstam F, Zhao H, Zhang Q, Ezzeddine N, Tang LY, et al. Clinical actionability enhanced through deep targeted sequencing of solid tumors. Clin Chem 2015;61:544–53. https://doi.org/10.1373/clinchem.2014.231100.

[35] Kortüm KM, Langer C, Monge J, Bruins L, Egan JB, Zhu YX, et al. Targeted sequencing using a 47 gene multiple myeloma mutation panel (M3P) in -17p high risk disease. Br J Haematol 2015;168:507–10. https://doi.org/10.1111/bjh.13171.

[36] Nikiforova MN, Wald AI, Roy S, Durso MB, Nikiforov YE. Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. J Clin Endocrinol Metab 2013;98:1852–60. https://doi.org/10.1210/jc.2013-2292.

[37] Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: Somatic cancer genetics at high-resolution. Nucleic Acids Res 2017;45:D777–83. https://doi.org/10.1093/nar/gkw1121.

[38] Zhang K, Wang H. Cancer Genome Atlas Pan-cancer analysis project. Chinese J Lung Cancer 2015;18:219–23. https://doi.org/10.3779/j.issn.1009-3419.2015.04.02.

[39] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8. doi:10.1038/nature12213.

[40] Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal patterns of selection in cancer and somatic tissues. Cell 2017;171(1029–1041):. https://doi.org/10.1016/j.cell.2017.09.042e21.

[41] Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res 2012;40:1–10. https://doi.org/10.1093/nar/gks743.

[42] Okosun J, Wolfson RL, Wang J, Araf S, Wilkins L, Castellano BM, et al. Recurrent mTORC1-activating RRAGC mutations in follicular lymphoma. Nat Genet 2016;48:183–8. https://doi.org/10.1038/ng.3473.

[43] Okosun J, Bödör C, Wang J, Araf S, Yang C-Y, Pan C, et al. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. Nat Genet 2014;46:176–81. https://doi.org/10.1038/ng.2856.

[44] Araf S, Wang J, Korfi K, Pangault C, Kotsiou E, Rio-Machin A, et al. Genomic profiling reveals spatial intra-tumor heterogeneity in follicular lymphoma. Leukemia 2018;32:1258–63. https://doi.org/10.1038/s41375-018-0043-y.

[45] Ivey A, Grech A, Vyas P, Patel Y, Freeman SD, Griffiths M, et al. Assessment of minimal residual disease in standard-risk AML. N Engl J Med 2016;374:422–33. https://doi.org/10.1056/nejmoa1507471.

[46] Sausen M, Phallen J, Adleff V, Jones S, Leary RJ, Barrett MT, et al. Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. Nat Commun 2015;6:1–6. https://doi.org/10.1038/ncomms8686.

[47] Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, Stebbings L, Menzies A, Widaa S, Stratton MR, Jones PH, Campbell PJ. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 2015;348(6237):880–6. https://doi.org/10.1126/science.aaa6806.

[48] Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature 2012;486:353–60. https://doi.org/10.1038/nature11143.

[49] Couch FJ, Hart SN, Sharma P, Toland AE, Wang X, Miron P, et al. Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. J Clin Oncol 2015;33:304–11. https://doi.org/10.1200/JCO.2014.57.1414.

[50] Abelson S, Collord G, Ng SWKK, Weissbrod O, Mendelson Cohen N, Niemeyer E, et al. Prediction of acute myeloid leukaemia risk in healthy individuals. Nature 2018;559:400–4. https://doi.org/10.1038/s41586-018-0317-6.

[51] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads kenkyuhi hojokin gan rinsho kenkyu jigyo. EMBnet J 2013;17:10–2. https://doi.org/10.14806/ej.17.1.200.

[52] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.

[53] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–95. https://doi.org/10.1093/bioinformatics/btp698.

[54] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.

[55] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.

[56] 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 Aaron. Genome Res 2010;25:2078–9. https://doi.org/10.1101/gr.107524.110.20.
[57] Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 2018. https://doi.org/10.1038/s41592-018-0051-x.
[58] Wilson RK, Mardis ER, McLellan MD, Koboldt DC, Shen D, Zhang Q, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–76. https://doi.org/10.1101/gr.129684.111.
[59] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865–71. https://doi.org/10.1093/bioinformatics/btp394.
[60] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:1–7. https://doi.org/10.1093/nar/gkq603.
[61] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80–92. https://doi.org/10.4161/fly.19695.
[62] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol 2016;17:1–14. https://doi.org/10.1186/s13059-016-0974-4.
[63] Nederbragt AJ, Rounge TB, Kausrud KL, Jakobsen KS. Identification and quantification of genomic repeats and sample contamination in assemblies of 454 pyrosequencing reads. Sequencing 2010;2010:1–12. https://doi.org/10.1155/2010/782465.
[64] Li H, Wren J. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 2014;30:2843–51. https://doi.org/10.1093/bioinformatics/btu356.
[65] Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 2012;28:907–13. https://doi.org/10.1093/bioinformatics/bts053.
[66] Bian X, Zhu B, Wang M, Hu Y, Chen Q, Nguyen C, et al. Comparing the performance of selected variant callers using synthetic data and genome segmentation. BMC Bioinf 2018;19:1–11. https://doi.org/10.1186/s12859-018-2440-7.
[67] 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;11:1–15. https://doi.org/10.1371/journal.pone.0151664.
[68] Chin S-F, Bruna A, Callari M, De Mattos-Arruda L, Caldas C, Sammut S-J, et al. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. Genome Med 2017;9:1–11. https://doi.org/10.1186/s13073-017-0425-1.
[69] Echeverria GV, Powell E, Seth S, Ge Z, Carugo A, Bristow C, et al. High-resolution clonal mapping of multi-organ metastasis in triple negative breast cancer. Nat Commun 2018;9. https://doi.org/10.1038/s41467-018-07406-4.
[70] Teer JK, Zhang Y, Chen L, Welsh EA, Cress WD, Eschrich SA, et al. Evaluating somatic tumor mutation detection without matched normal samples. Hum Genomics 2017;11:1–13. https://doi.org/10.1186/s40246-017-0118-2.
[71] Cheng AY, Teo YY, Ong RTH. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics 2014;30:1707–13. https://doi.org/10.1093/bioinformatics/btu067.
[72] Kim K, Seong M-W, Chung W-H, Park SS, Leem S, Park W, et al. Effect of next-generation exome sequencing depth for discovery of diagnostic variants. Genomics Inform 2015;13:31. https://doi.org/10.5808/GI.2015.13.2.31.
[73] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–9. https://doi.org/10.1038/nature07517.
[74] Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. Brief Bioinform 2010;11:484–98. https://doi.org/10.1093/bib/bbq016.
[75] Ajay SS, Parker SCJ, Abaan HO, Fuentes Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Res 2011;21:1498–505. https://doi.org/10.1101/gr.123638.111.
[76] Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for germline variant calling and somatic mutation detection. Curr Protoc Bioinforma 2013;44:233–45. https://doi.org/10.1002/0471250953.bi1504s44.
[77] Jiang Y, Redmond D, Nie K, Eng KW, Clozel T, Martin P, et al. Deep sequencing reveals clonal evolution patterns and mutation events associated with relapse in B-cell lymphomas. Genome Biol 2014;15:1–17. https://doi.org/10.1186/s13059-014-0432-0.
[78] Wong SQ, Li J, Tan AYC, Vedururu R, Pang JMB, Do H, et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. BMC Med Genomics 2014;7:1–10. https://doi.org/10.1186/1755-8794-7-23.
[79] Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, et al. Oncotator: cancer variant annotation tool. Hum Mutat 2015;36:E2423–9. https://doi.org/10.1002/humu.22771.
[80] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 2019;531210. https://doi.org/10.1101/531210.
[81] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–91. https://doi.org/10.1038/nature19057.
[82] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014;11:361–2. https://doi.org/10.1038/nmeth.2890.
[83] 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 Diagnostics 2017;19:341–65. https://doi.org/10.1016/j.jmoldx.2017.01.011.
[84] Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet 2018;19:269–85. https://doi.org/10.1038/nrg.2017.117.
[85] Deng S, Lira M, Huang S, Wang K, Valdez C, Kinong J, et al. TNER : a novel background error suppression method for mutation detection in circulating tumor DNA 2017:1–12. doi:10.1101/214379.
[86] Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 2016;34:547–55. https://doi.org/10.1038/nbt.3520.
[87] Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, et al. CONTRA: Copy number analysis for targeted resequencing. Bioinformatics 2012;28:1307–13. https://doi.org/10.1093/bioinformatics/bts146.
[88] Chen Y, Zhao L, Wang Y, Cao M, Gelowani V, Xu M, et al. SeqCNV: A novel method for identification of copy number variations in targeted next-generation sequencing data. BMC Bioinf 2017;18:1–9. https://doi.org/10.1186/s12859-017-1566-3.
[89] 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 2016;12:623–30. https://doi.org/10.1038/nmeth.3407.