Clinical Chemistry 68:11
1425–1435 (2022)

UMIErrorCorrect and UMIAnalyzer: Software for Consensus Read Generation, Error Correction, and Visualization Using Unique Molecular Identifiers

Tobias Österlund [a,b,c] Stefan Filges [c] Gustav Johansson,[b,c,d] and Anders Ståhlberg [a,b,c,*]

BACKGROUND: Targeted sequencing using unique molecular identifiers (UMIs) enables detection of rare variant alleles in challenging applications, such as cell-free DNA analysis from liquid biopsies. Standard bioinformatics pipelines for data processing and variant calling are not adapted for deep-sequencing data containing UMIs, are inflexible, and require multistep workflows or dedicated computing resources.

METHODS: We developed a bioinformatics pipeline using Python and an R package for data analysis and visualization. To validate our pipeline, we analyzed cell-free DNA reference material with known mutant allele frequencies (0%, 0.125%, 0.25%, and 1%) and public data sets.

RESULTS: We developed UMIErrorCorrect, a bioinformatics pipeline for analyzing sequencing data containing UMIs. UMIErrorCorrect only requires fastq files as inputs and performs alignment, UMI clustering, error correction, and variant calling. We also provide UMIAnalyzer, a graphical user interface, for data mining, visualization, variant interpretation, and report generation. UMIAnalyzer allows the user to adjust analysis parameters and study their effect on variant calling. We demonstrated the flexibility of UMIErrorCorrect by analyzing data from 4 different targeted sequencing protocols. We also show its ability to detect different mutant allele frequencies in standardized cell-free DNA reference material. UMIErrorCorrect outperformed existing pipelines for targeted UMI sequencing data in terms of variant detection sensitivity.

CONCLUSIONS: UMIErrorCorrect and UMIAnalyzer are comprehensive and customizable bioinformatics tools that can be applied to any type of library preparation protocol and enrichment chemistry using UMIs. Access to simple, generic, and open-source bioinformatics tools will facilitate the implementation of UMI-based sequencing approaches in basic research and clinical applications.

Introduction

The rapid development of massive parallel sequencing techniques allows investigation of most types of DNA samples and sequences to gain insight into genomic alterations (1, 2). In healthcare, DNA sequencing is used in a wide range of applications, such as cancer (3–5), genetic mosaicism (2, 6, 7), immunology (8, 9), forensics (10), metagenomics (11), and infectious diseases (12). Cancer applications include diagnostics, prognostics, and precision medicine (13, 14). Tumor biopsies are commonly analyzed with either targeted or whole-genome sequencing. Despite constantly improving therapies and diagnostics, there is an emerging need to monitor patients over time to determine treatment efficacy as well as to detect treatment resistance and relapse as early as possible. Normally, tissue biopsies cannot be collected repeatedly. Instead, minimally invasive liquid biopsies can be longitudinally collected from body fluids, such as blood and urine (13–15). Circulating tumor-DNA (ctDNA) is released from the tumor into the circulation through apoptosis, necrosis, and active secretion and is thus considered a useful biomarker in cancer management (16, 17). For most clinically relevant applications the concentration of ctDNA in a typical blood plasma sample is low, even down to the range of single ctDNA molecules (18).

Standard sequencing with conventional methods for variant calling only allows detection of mutant allele frequencies (MAFs) down to 1% to 5% (19–21), which

[a]Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden; [b]Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden; [c]Sahlgrenska Center for Cancer Research, Department of Laboratory Medicine, Institute of Biomedicine, University of Gothenburg, Gothenburg, Sweden; [d]SiMSen Diagnostics AB, Gothenburg, Sweden.

*Address correspondence to this author at: Sahlgrenska Center for Cancer Research, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Medicinaregatan 1F, 405 30 Gothenburg, Sweden. e-mail: anders.stahlberg@gu.se.
Received May 3, 2022; accepted July 8, 2022.
https://doi.org/10.1093/clinchem/hvac136

is not sufficient for several clinical ctDNA applications, requiring a sensitivity well below 1% MAF using low DNA concentrations (13, 14, 21). There are several sources of sequencing errors, but most errors are polymerase-induced, generated during library construction and sequencing (22, 23). To improve sequencing sensitivity, unique molecular identifiers (UMIs), also referred to as molecular barcodes, can be used. The UMI typically consists of a randomized 8 to 12 nucleotides long sequence, which is attached to the DNA fragments in the initial library preparation step (5, 24). During library construction and sequencing, each target DNA molecule will be amplified with its specific UMI, enabling sequence reads with the same UMI to be traced back to the same original DNA molecule (Fig. 1, A). By generating consensus reads for the DNA molecules with the same UMI, sequencing errors are efficiently reduced. In some approaches, the UMIs also allow for accurate quantification, since each original molecule is only counted once (i.e., PCR duplicates are removed) (25). Several experimental methods have been developed for UMI-based sequencing including PCR-based protocols, such as Safe-SeqS (5), QiaSeq (26), and simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection using next-generation sequencing (SiMSen-Seq) (27), as well as hybridization capture-based-approaches, such as duplex sequencing (24) and circle sequencing (28). Despite the common use of UMIs, downstream data analysis is performed

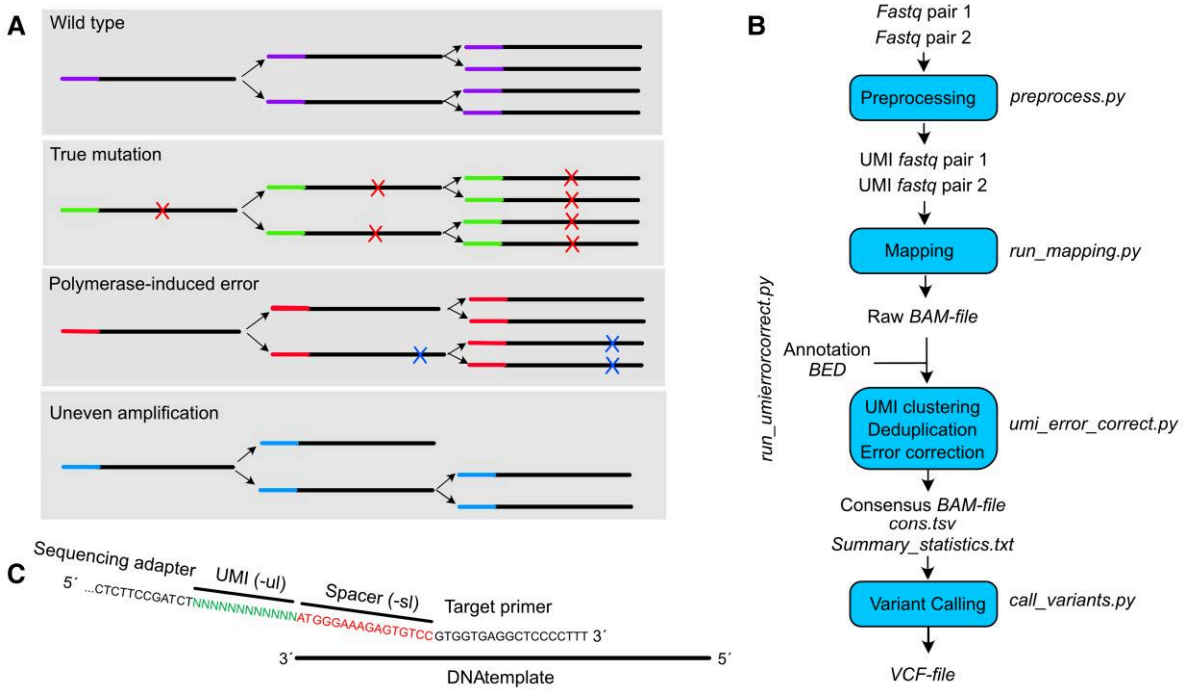

Fig. 1. UMIErrorCorrect pipeline. (A), Principle of error correction using UMIs. Each template DNA molecule is barcoded with a UMI. UMIs are then used to correct for polymerase-induced errors and uneven amplification. The red crosses represent a true mutation in the original DNA molecule, while blue crosses represent polymerase-induced errors. The use of UMIs enables recovery of original molecules through the formation of consensus reads; (B), The UMIErrorCorrect workflow consists of four subprocesses that can be executed either by a single command or by running the 4 scripts sequentially. The input files are either single- or paired-end fastq files, with an optional BED file containing target DNA coordinates used for annotation. UMIErrorCorrect output data and files are shown in Table 1; (C), Example of library sequence context using SiMSen-Seq. The UMI (*-ul*) and spacer lengths (*-sl*) are specified in UMIErrorCorrect to describe the sequence context, based on the library construction protocol used. For SiMSen-Seq the UMI and spacer lengths are 12 and 16 nucleotides, respectively. Examples of other UMI-based approaches are shown in Supplemental Fig. 1.

with method-specific bioinformatics tools that are adjusted to specific constraints, limiting the possibilities to compare data sets. Several software tools are also not available as open-source, minimizing the possibilities to compare workflows and applied algorithms. In addition, existing tools often require multistep workflows and dedicated computing resources.

Here, we developed Unique Molecular Identifier Error Corrector (UMIErrorCorrect), an open-source analysis pipeline for targeted sequencing data with UMIs, designed to be easy and straightforward to install and use. The pipeline includes a variant caller to assess low-frequent variants. We also provide UMIAnalyzer and UMIVisualizer, an R package, and a graphical user interface implemented as an R Shiny app, for data analysis and visualization of sequencing data processed with UMIErrorCorrect. We demonstrate the use of UMIErrorCorrect and UMIVisualizer by analyzing standardized reference material with known MAFs using SiMSen-Seq and compare the performance with other pipelines. We also analyzed publicly available data sets generated by 3 different experimental approaches: Roche Avenio, QiaSeq, and Archer. We show that UMIErrorCorrect can be applied to any data set with UMIs and provide a generic and open-source bioinformatics workflow that enables users to compare and evaluate different data sets. This will facilitate the standardization and clinical implementation of UMI-based sequencing approaches.

Materials and Methods

UMIERRORCORRECT PIPELINE

The UMIErrorCorrect pipeline is open source and implemented in Python. The pipeline was implemented and tested using Python version 3.6.9. Details regarding the implementation of UMIErrorCorrect can be found in the online Supplemental Material and Methods.

LOW-FREQUENCY ALLELE VARIANT CALLER

A low-frequency variant caller was developed, where the probability of the maximum nonreference allele not being background noise was calculated. For mathematical derivation of the variant caller, see Supplemental Material and Methods. UMIErrorCorrect will provide

Table 1. UMIErrorCorrect output files.

Output file	Example and file type	Description	Input to UMIAnalyzer
Consensus reads	Sample1_consensus_reads.bam Sample1_consensus_reads.bam.bai	Alignment of consensus reads and indexes	Optional
Consensus reads tab-separated values	Sample1_cons.tsv	Table with consensus allele counts per nucleotide position	Yes
VCF[a]	Sample1.vcf	Called variants	No
Summary statistics	Sample1_summary_statistics.txt	Coverage of raw and consensus reads per region	Yes
Target coverage	Sample1_target_coverage.txt	Number of raw and consensus reads on target	No
Histogram of variants	Sample1_histogram.png	Histogram showing probability distribution of called variants	No
UMIs in header fastqs	Sample1_umis_in_header_R1.fastq.gz Sample1_umis_in_header_R2.fastq.gz[b]	Raw reads with UMIs and spacers removed	No
Raw reads	Sample1.sorted.bam Sample1.sorted.bam.bai	Alignment of raw reads and indexes	No

[a]Variant call format.
[b]For single-end sequencing samples, this file is missing.

information about count, frequency, and allele type of the maximum nonreference allele. The positions with nonreference allele counts higher than zero will be subjected to variant calling. The probability of the nonreference allele count being a background error was modeled using the beta-binomial distribution. The beta prior distribution was fitted to background errors of a typical SiMSen-Seq data set (for details, see Supplemental Materials and Methods). The default Q-value cutoff for variant calling is set to $Q \geq 20$ corresponding to a P value cutoff ≤ 0.01.

UMIANALYZER R PACKAGE AND UMIVISUALIZER APP

The code for the R package UMIAnalyzer is available for R versions $\geq 4.1.0$. UMIAnalyzer can be installed from the comprehensive R archive network using the command *install.packages("umiAnalyzer").* The visualization app, UMIVisualizer, is built with Shiny, a framework for making interactive web-applications in R. UMIVisualizer can be started with the command *umiAnalyzer::runUmiVisualizer()*. UMIAnalyzer provides functions and classes for reading, analyzing, and plotting data processed with UMIErrorCorrect. UMIAnalyzer functions also serve as the backend for the UMIVisualizer app.

SIMSEN-SEQ

We analyzed Seraseq ctDNA Mutation Mix v2 (Seracare), with known MAFs using the SiMSen-Seq protocol as described earlier (27). For details on library construction and sequencing, see Supplemental Material and Methods. The sequencing data were uploaded to the NCBI Sequence Read Archive repository (PRJNA788522 and PRJNA507366).

PUBLIC DATA SETS

Two publicly available data sets were also used and analyzed from the NCBI Sequence Read Archive, PRJNA577992 (Roche Avenio and QiaSeq) and PRJEB31811 (Archer).

Results

DEVELOPMENT OF UMIERRORCORRECT: A GENERIC AND EASY-TO-USE BIOINFORMATICS TOOL FOR SEQUENCING READS WITH UMIS

Here, we developed UMIErrorCorrect, a generic and easy-to-use bioinformatics pipeline that processes sequencing data containing UMIs, requiring only fastq files as input. UMIErrorCorrect is implemented in Python, and the schematic workflow is shown in Fig. 1, B. The complete pipeline includes preprocessing, mapping, error correction, allele counting, and variant calling. The pipeline starts by a single command (*run_umierrorcorrect.py*). In the first step (*preprocess.py*), the UMI and spacer sequences are trimmed off from all reads and the UMI is added to the header of the fastq file. Examples of UMI setups are shown in Fig. 1, C and Supplemental Fig. 1. In the second step (*run_mapping.py*), all reads are aligned to an indexed reference genome of choice using the Burrows–Wheeler alignment tool, bwa mem (29). In the third step (*umi_error_correct.py*), reads are grouped into UMI families based on target DNA region (i.e., chromosomal position and UMI sequence). To allow for sequencing errors in the UMI sequence, an edit-distance threshold is introduced (default edit distance ≤ 1), and UMI clustering is performed similarly to the UMI clustering in UMI tools (30). Next, error-corrected consensus reads are generated based on all reads in each UMI family. The consensus read generation provides one consensus read per UMI family. This step corrects for both polymerase-induced errors and uneven amplifications (Fig. 1, A). For annotation purposes, an optional Browser Extensible Data (BED) file, containing chromosomal coordinates and names for all targeted sequencing regions can be provided by the user. UMIErrorCorrect generates the same results without the BED file, but without annotation of the targeted regions.

UMIERRORCORRECT OUTPUT FILES

The UMIErrorCorrect pipeline produces 11 output data files per sample (Table 1). Supplemental Table 1 contains an example of the output file content for a sample analyzed with SiMSen-Seq. A consensus read BAM file is generated, representing the alignment of all consensus reads. The consensus group size cutoff parameter (i.e., the minimal number of raw reads required to generate a consensus read) is set to 2 when generating the consensus read BAM file.

The *cons.tsv* file is a tab-separated table with each row representing one position in the reference genome, after error correction and consensus read generation. The columns show information about the nucleotide counts (A, C, G, and T), insertions, deletions, and unknown bases for each nucleotide position. The maximum nonreference allele is presented in the last 3 columns of the *cons.tsv* file (genotype change, variant allele count, and variant allele frequency). A variant call format file is provided that includes the nucleotide positions with high probability of being true variants compared to the background. The developed variant caller is designed for UMI sequencing data, enabling detection of low variant allele frequencies. The background error distribution is modeled using a beta-binomial distribution (26, 31). The default Q-value cutoff for the variant caller is set to $Q \geq 20$, representing a probability $P <$

0.01 of the detected variant allele not being a background error. For additional details about the developed variant caller, see Supplemental Material and Methods.

The summary statistics file is a tab-separated text file showing the number of raw reads and consensus reads that overlap with each targeted region. The target coverage file shows the percentage of all raw reads and consensus reads that cover the target regions specified in the input BED file. A low percentage of target coverage indicates off-target and unspecific library amplification and/or enrichment. The histogram.png file shows an image of the distribution of Q-values (i.e., probability scores for variants present in the variant call format file). The fastq files with UMIs in header and the raw reads BAM file are intermediate files, including all reads before error correction and consensus read generation.

UMIVISUALIZER ENABLES DATA INTERPRETATION AND VISUALIZATION

After UMIErrorCorrect, the directories with the output files for all analyzed samples can be loaded into R using the R-package UMIAnalyzer and the R shiny app UMIVisualizer. Both allow the user to interactively explore and visualize the experimental data, simplifying the interpretation of results generated by UMIErrorCorrect. UMIAnalyzer generates an UMI experiment class object by loading and merging the cons.tsv files and the summary statistics files for all samples. The consensus reads BAM file is also used to visualize UMI family size distribution with a histogram. A workflow describing the graphical user interface of UMIVisualizer is shown in Fig. 2. By default, UMIVisualizer will show both the variants and the background positions with background errors. The user can then vary different parameters to investigate their effects on variant calling and background noise. The next step is to select which samples and targets (i.e., assays) are of interest to analyze. Next, the user can select the consensus group size cutoff, the minimum alternative allele consensus read count, or the false discovery rate cutoff (Fig. 2, A). Downstream results are shown both in table format and in the data viewer and plots section (Supplemental Fig. 2). There are also functions for quality controls, such as plots showing the number of consensus reads for each target, the effects of consensus group size cutoffs on the number of consensus reads, and a histogram of the UMI group size distribution for each sample (Supplemental Fig. 2). If the user is only interested in predefined nucleotide positions (i.e., hotspot nucleotide positions for mutations), another BED file with positions may be imported into UMIVisualizer, resulting in a simplified data result table. UMIVisualizer contains several functions to combine samples and to plot longitudinal data (Fig. 2, B). A descriptive metadata file can be uploaded for each sample to add categorical data used for generating more user-customized plots. UMIVisualizer can also generate an html report, summarizing variants and plots.

COMPREHENSIVE SIMSEN-SEQ DATA ANALYSIS USING UMIERRORCORRECT AND UMIVISUALIZER

To demonstrate the functionality of UMIErrorCorrect and UMIVisualizer, we generated a SiMSen-Seq data set using Seracare ctDNA reference material, targeting 5 hot-spot mutation sequences (Figs. 3, A–D). Four different MAFs (0%, 0.125%, 0.25, and 1%) were sequenced in triplicates using 20 ng DNA to an estimated depth of 3.3 reads per UMI. Figures 3, A and B show summary sequencing statistics. The mean background error for all analyzed nucleotide positions was 0.013%, and the maximum detected background error for any specific nucleotide position was 0.244% found in the *TP53_a* assay. Figure 3, C shows that the predefined mutations were detectable at expected MAFs for all targeted sequences. All 5 mutations were significantly above the background error rate for each specific nucleotide position at 0.125% MAF when comparing the mutated nucleotide positions with the same position in wild-type (2-sample *t*-test, $P \leq 0.05$). However, if we instead compared the allele frequencies of the mutated positions with all nucleotide positions within the 5 amplicons, only the mutations in *PIK3CA_b* and *TP53_a* were significantly above the background error levels at 0.125% MAF, while the mutations in *KIT* and *TP53_b* were significant at ≥0.25% MAF (2-sample *t*-test, $P \leq 0.05$).

An alternative approach is to use the variant caller in UMIErrorCorrect. The output Q-value represents the probability that the observed MAF is not a background error. Figure 3, D shows the sensitivity (i.e., the ability to detect mutations) and the specificity (i.e., the ability to avoid false positives) to detect variants. The sensitivity was calculated as the true positive rate (i.e., the percentage of the 5 predefined mutations that were identified in all replicates). The specificity was calculated as 1 − the false-positive rate (i.e., the number of background errors divided by the number of background positions). The sensitivity and specificity are shown as a function of 3 Q-value cutoffs ($Q \geq 10$, ≥ 15, and ≥ 20) and 2 consensus group size cutoffs (consensus group size ≥ 3 and ≥ 10). For the consensus group size cutoff ≥ 3, at the lowest Q-value cutoff ($Q \geq 10$), the sensitivities were 46.7%, 93.3%, and 100% to detect 0.125%, 0.25%, and 1% MAF, respectively. The corresponding specificities were 100%, 100%, and 99.7%. Increased Q-value cutoff to ≥15 decreased the sensitivity to 26.7% and 73.3% at 0.125% and 0.25% MAF, respectively, while the sensitivity at

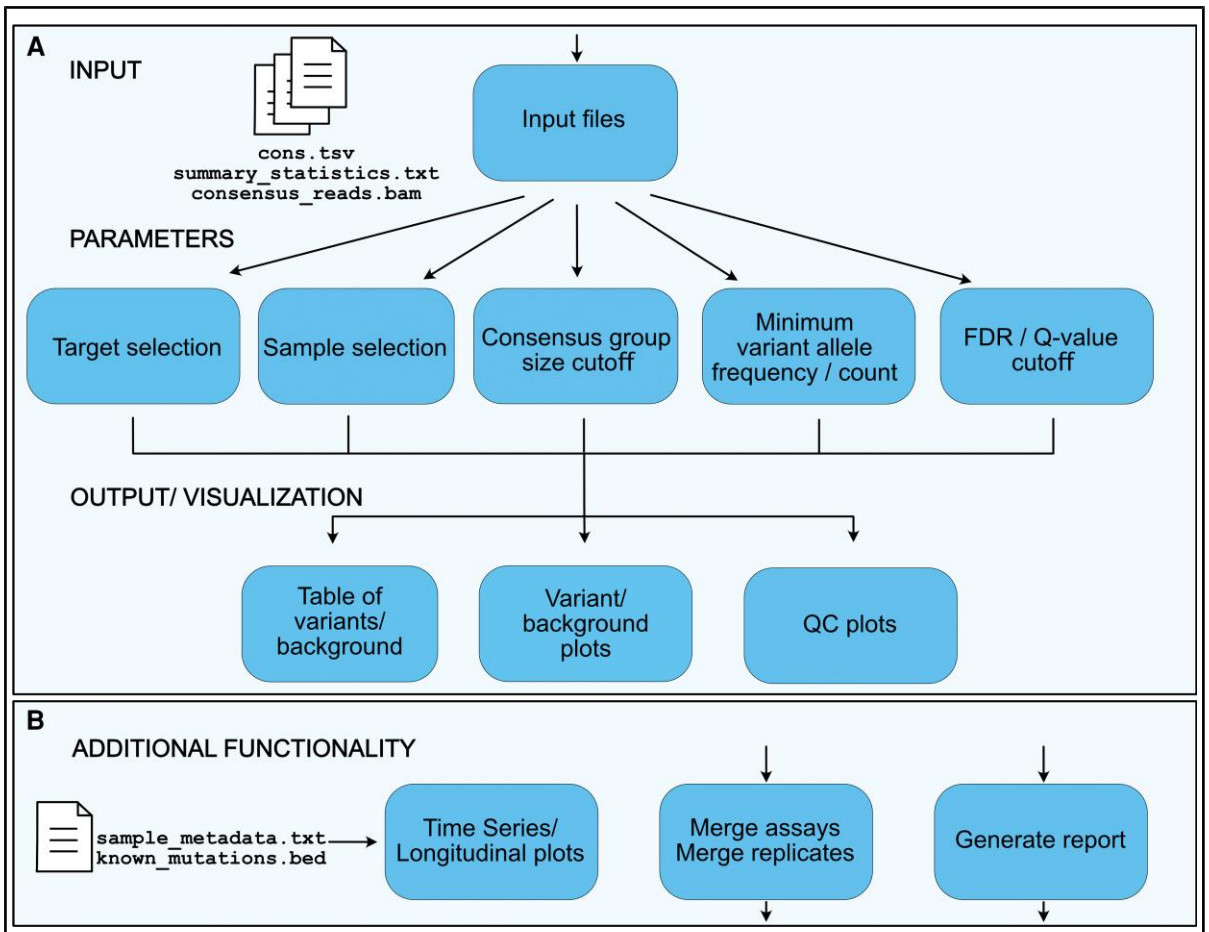

Fig. 2. Graphical user interface of UMIVisualizer. (A), For each sample, 3 of the output files generated by UMIErrorCorrect (cons.tsv, summary_statistics.txt, and consenus_reads.bam) are loaded into R Shiny App UMIVisualizer. The user can define which data to analyze and visualize them using different parameter settings. The output consists of tables and plots summarizing all selected results. Examples from the UMIVisualizer are presented in Supplemental Fig. 2; (B), Optional functions in UMIVisualizer.

1% MAF remained at 100%. The specificities at Q ≥ 15 were 100% for all MAFs. A Q-value cutoff ≥ 20 resulted in sensitivities at 13.3%, 40.0%, and 100%, respectively, and the specificities remained at 100%. Changes in the second parameter, consensus group size cutoff, decreased the sensitivity for all 3 MAFs and Q-values when the consensus group size cutoff increased from ≥3 to ≥10, while the specificity remained the same.

To study different sequencing parameters using UMIVisualizer further, we sequenced 20 ng SeraCare ctDNA to an estimated depth of 10 reads per UMI. We also analyzed 5 ng SeraCare ctDNA with a sequencing depth of 3.3 and 10 reads per UMI, respectively. The results are presented in Supplemental Fig. 3.

UMIERRORCORRECT AND UMIVISUALIZER ARE GENERIC TOOLS TO ANALYZE AND VISUALIZE UMI-CONTAINING DATA

UMIErrorCorrect and UMIVisualizer are generic tools, suitable for any experimental protocol or sequencing platform. To demonstrate their use, we analyzed 3 publicly available data sets, utilizing UMIs generated by 3 different methods: Roche Avenio (32), QiaSeq (32), and Archer (7) (Supplemental Table 2). The configuration of the UMI and spacer sequences for the 3 methods are shown in Supplemental Fig. 1. The Roche Avenio panel is a target capture-based panel that targets 15 genes, including 322 sequence regions. The median coverage for the Roche Avenio panel was 206, using consensus group size cutoff ≥ 3 (i.e., the median number of

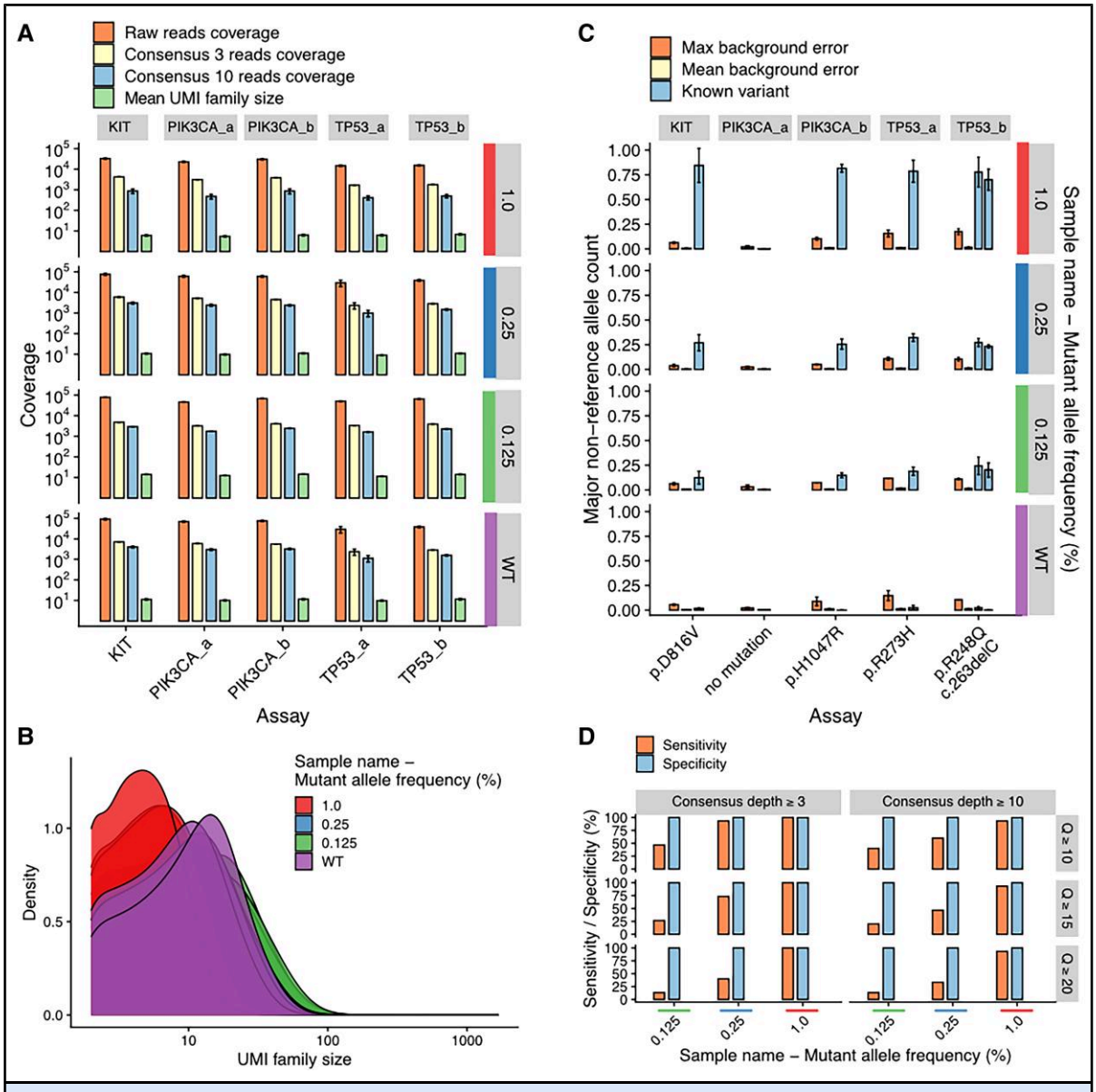

Fig. 3. Validation of UMIErrorCorrect and UMIVisualizer using SiMSen-Seq and standardized reference materials. (A), Summary sequencing statistics. Raw reads coverage represents the number of reads mapped to the target before consensus sequence generation. Consensus 3 coverage and consensus 10 coverage are the coverage of consensus reads when using consensus group size cutoff ≥3 and ≥10, respectively; (B), Density plot showing the UMI family size distribution. The *ggplot geom_density()* function was used to smoothen the curves for values < 10. n = 3; (C), Detection of MAFs in Seracare ctDNA reference materials. The expected variants are shown in blue together with maximum (orange) background error for any nucleotide position and mean (yellow) background error for all nucleotide positions. The SiMSen-Seq panel used in this experiment consists of 5 assays targeting sequences in *KIT*, *PIK3CA*, and *TP53*. The assay PIK3CA_a contained no predefined mutation, while the assay TP53_b contained 2 predefined mutations. Mean ± SE is shown for consensus group size ≥ 3. n = 3; (D), Sensitivity and specificity to detect mutations using different cutoff values for different consensus group sizes and Q-values. n = 3.

Table 2. Target coverage per analyzed data sets.

Data Set	Method	Median consensus coverage per target[a]	CV (%)[b]	Maximum consensus coverage per target[a]	CV (%)[b]
2	Roche Avenio	206 ± 27	18.2	333 ± 83	37.5
3	QiaSeq	2010 ± 39	13.3	472 ± 116	25.1
4	Archer	157 ± 29	18.6	269 ± 101	24.5

[a]Mean ± SE is shown for median and max consensus coverage per target using consensus group size cutoff ≥ 3.
[b]Coefficient of variation for median and maximum consensus coverage per target.

Table 3. Comparison between UMIErrorCorrect and 2 other UMI-based variant callers.

Method[a]	Sensitivity (%)			Specificity (%)			Time per sample(s)	Peak memory size (GB)
	0.125% MAF	0.25% MAF	1.0% MAF	0.125% MAF	0.25% MAF	1.0% MAF		
UMIErrorCorrect[b]	46.7	93.3	100	100	100	99.7	32.5	2.05
MAGERI[c]	33.3	40.0	40.0	98.5	99.6	98.7	11.1	257.8
SmCounter2[d]	0	0	53.3	100	100	99.7	73.2	0.44

[a]UMI-VarCal and DeepSNVMiner were not included in the comparison, since they require paired-end sequencing data.
[b]Using default parameters and Q ≥ 10.
[c]Using default parameters and Q ≥ 20 (36).
[d]Using default parameters (26).

UMIs per target sequence was 206) (Table 2). The QiaSeq panel is an amplicon-based panel that targets 275 genes, including 4427 amplicons. The median coverage for the QiaSeq panel was 2010. Roche Avenio and QiaSeq data sets also contained spiked-in variants (32), and the variants identified after removing variants found in healthy control samples are shown in Supplemental Table 3 using a Q-value cutoff ≥ 15. The Archer panel was designed to target variants causing genetic disorders (7) and is a target capture-based panel that covers 185 genes and 270 targets. The median coverage for the Archer panel was 156 (Table 2), and the detected genetic variations are shown in Supplemental Table 3.

Discussion

UMI-based error correction in sequencing data enables ultrasensitive detection and quantification of variants. This ability provides the basis for the development of novel applications in cancer management, such as screening, diagnostics, monitoring of treatment efficacy, detection of treatment resistance, minimal residual disease, and relapse at an early stage (13, 14, 33, 34). Here, we developed a Python pipeline for handling sequencing data with UMIs, including alignment, UMI clustering, error correction, and variant calling. Further, we developed an accompanying R package for analyzing and visualizing UMI-corrected data.

Several variant calling approaches for targeted sequencing with UMIs are published, including DeepSNVMiner (35), MAGERI (36), smCounter2 (26), and UMI-VarCal (37). These bioinformatics tools are publicly available, except smCounter2, which is only available upon request from Qiagen. DeepSNVMiner uses a simple heuristic consensus read count threshold and does not include any model for estimating background errors. Simple count thresholds often have substantially higher false-positive rates than approaches that explicitly model background noise (26). Therefore, MAGERI and smCounter2 use a beta-binomial model for variant calling, by estimating each type of nucleotide change individually. However, smCounter2 could not reach reliable estimates for rare nucleotide variants and is instead approximating them (26). MAGERI requires a dedicated computing environment and uses a single reference data set to construct its background error model (36), which may be problematic when comparing different types of technologies and data sets (26) since, for instance, the choice of DNA polymerase may affect the error profile (23, 38). UMIErrorCorrect also uses a beta-

binomial distribution to model the background error. Instead of modeling different nucleotide substitution types, UMIErrorCorrect models the maximum nonreference allele count in relation to the coverage at each position regardless the type of nucleotide change. An advantage of smCounter2 and UMIErrorCorrect is that both approaches allow the users to train the background error model with their own reference data set. UMI-VarCal uses a custom pileup algorithm and then uses sequencing quality scores to model the background error rates. However, most errors that occur during amplification and sequencing should be corrected using the UMI clustering alone, whereas errors introduced during the first cycle of barcoding or errors from chemical modification of the original DNA template can have high base quality scores that are not accounted for in this approach (36). Furthermore, UMI-VarCal does not have the functionality to interface with an alignment tool and therefore requires the user to perform this step separately, whereas MAGERI does not require any external tools for mapping. UMIErrorCorrect and smCounter2 require a preinstallation of bwa although the containerized version of UMIErrorCorrect works out of the box using either Docker or Singularity.

To validate our pipeline, we sequenced ctDNA reference material using a targeted SiMSen-Seq panel. At 0.25% and 1% MAF, the sensitivity was 93.3% and 100%, respectively, using a Q-value ≥10 and consensus group size ≥ 3, while the sensitivity was 46.7% for the most challenging variant calling scenario (i.e., 0.125% MAF). To evaluate the overall performance of UMIErrorCorrect, we compared our results with the outputs of MAGERI (36) and SmCounter2 (26) (Table 3). UMIErrorCorrect displayed higher sensitivity compared to both MAGERI and SmCounter2 for all different MAFs. In the most challenging scenario with MAF 0.125%, the sensitivity for UMIErrorCorrect was 46.7%, compared to 33.3% and 0% for MAGERI and SmCounter2, respectively. The specificity was equal for UMIErrorCorrect and smCounter, and both showed higher specificity than MAGERI.

The amplicon coverage ranged from 3139 to 4922 consensus reads in our SiMSen-Seq data set using 20 ng DNA sequenced to a depth of 3.3 reads per UMI. We observed a significant drop in sensitivity for the 0.125% MAF samples. This was expected due to sampling and DNA fragmentation (39). At the lowest coverage, this corresponds to less than 4 mutated molecules at 0.125% MAF and 6 mutated molecules at the highest coverage. At 0.125% MAF, the number of mutated reads is at the same magnitude as the background noise, making variant calling challenging. In line with these data, the sensitivity to detect few mutated reads was clearly different when comparing data using different DNA concentrations, 5 and 20 ng DNA, respectively. One approach to detect low-frequency variants is to lower the Q-value cutoff. However, lowering the Q-value cutoff increases the risk of including false positives. The user can therefore also include a minimum allele count threshold specifying the minimal number of consensus reads with the variant allele. Another strategy to ensure that called variants are true is to increase the consensus group size cutoff. Compared to other relevant variant callers, UMIErrorCorrect and UMIVisualizer provide a flexible framework to adjust multiple variables and visualize their effect on variant calling as well as to evaluate and optimize variant calling settings for specific constraints and applications.

In summary, UMIErrorCorrect, UMIAnalyzer, and UMIVisualizer provide an all-in-one pipeline that requires only raw sequencing data as input and enables the user to choose their own mode of analysis. They provide a variant-calling algorithm with a customizable background error model as well as facilitate data exploration, interpretation, and visualization for nonspecialist users. The pipeline is agnostic toward library preparation protocol and enrichment chemistry and can be applied to data generated by any commercial panels, such as Roche Avenio, QiaSeq, and Archer.

Supplementary Material

Supplementary material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: BED, browser extensible data; ctDNA, circulating tumor-DNA; MAF, mutant allele frequency; SiMSen-Seq, simple multiplexed PCR based barcoding of DNA for ultrasensitive mutation detection using next-generation sequencing; UMI, unique molecular identifier.

Human Genes: *KIT*, KIT proto-oncogene receptor tyrosine kinase; *PIK3CA*, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha; *TP53*, tumor protein p53.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: A. Ståhlberg is board member for Tulebovaasta, Iscaff Pharma, and SiMSen Diagnostics. G. Johansson declares employment with SiMSen Diagnostics.
Consultant or Advisory Role: None declared.
Stock Ownership: A. Ståhlberg declares stock ownership in Tulebovaasta, Iscaff Pharma, and SiMSen Diagnostics. G. Johansson declares stock ownership in SiMSen Diagnostics.

Honoraria: None declared.
Research Funding: T. Österlund, Assar Gabrielssons Research Foundation, Johan Jansson Foundation for Cancer Research, Sahlgrenska University Hospital; S. Filges, Assar Gabrielssons Research, Foundation, Johan Jansson Foundation for Cancer Research, Lion's Cancer Research Fund of Western Sweden, Wilhelm and Martina Lundgren Foundation, University of Gothenburg; G. Johansson, Assar Gabrielssons Research Foundation, Johan Jansson Foundation for Cancer Research, Anna-Lisa och Bror Björnsson stiftelse, University of Gothenburg, AstraZeneca, Simsen Diagnostics; A. Ståhlberg, Region Västra Götaland, Sweden, Swedish Research Council (2020-01008), the Swedish state under the agreement between the Swedish government and the county councils, the ALF-agreement (965065), Sweden's Innovation Agency (2018-00421, 2020-04141), Swedish Cancer Society (19-0306), Swedish Childhood Cancer Foundation (2020-007, MTI2019-0008), The Sjöberg Foundation, Region Västra Götaland, Sahlgrenska University Hospital, University of Gothenburg.
Expert Testimony: None declared.
Patents: A. Ståhlberg is co-inventor of SiMSen-Seq that is patent protected (US serial no. 15/552,618).

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

References

1. Kennedy SR, Loeb LA, Herr AJ. Somatic mutations in aging, cancer and neurodegeneration. Mech Ageing Dev 2012;133:118–26.
2. Martincorena I, Roshan A, Gerstung M, Ellis P, van Loo P, McLaren S, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 2015;348:880–6.
3. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A 2008;105:13081–6.
4. Schmitt MW, Loeb LA, Salk JJ. The influence of subclonal resistance mutations on targeted cancer therapy. Nat Rev Clin Oncol 2016;13:335–47.
5. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 2011;108:9530–5.
6. Acuna-Hidalgo R, Sengul H, Steehouwer M, van de Vorst M, Vermeulen SH, Kiemeney LALM, et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. Am J Hum Genet 2017;101:50–64.
7. Crowgey EL, Washburn MC, Kolb EA, Puffenberger EG. Development of a novel next-generation sequencing assay for carrier screening in old order Amish and Mennonite populations of Pennsylvania. J Mol Diagn 2019;21:687–94.
8. Johansson G, Kaltak M, Rîmniceanu C, Singh AK, Lycke J, Malmeström C, et al. Ultrasensitive DNA immune repertoire sequencing using unique molecular identifiers. Clin Chem 2020;66:1228–37.
9. Boyd SD, Marshall EL, Merker JD, Maniar JM, Lyndon N, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med 2010;1:1–16.
10. de la Puente M, Phillips C, Xavier C, Amigo J, Carracedo A, Parson W, et al. Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. Forensic Sci Int Genet 2020;45:102213.
11. Milani C, Ticinesi A, Gerritsen J, Nouvenne A, Andrea Lugli G, Mancabelli L, et al. Gut microbiota composition and Clostridium difficile infection in hospitalized elderly individuals: a metagenomic study. Sci Rep 2016;6:1–12.
12. Nasu A, Marusawa H, Ueda Y, Nishijima N, Takahashi K, Osaki Y, et al. Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. PLoS One 2011;6:e24907.
13. Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet 2019;20:71–88.
14. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic—implementation issues and future challenges. Nat Rev Clin Oncol 2021;18:297–312.
15. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer 2017;17:223–38.
16. Hu Z, Chen H, Long Y, Li P, Gu Y. The main sources of circulating cell-free DNA: apoptosis, necrosis and active secretion. Crit Rev Oncol Hematol 2021;157:103166.
17. Bronkhorst AJ, Wentzel JF, Aucamp J, van Dyk E, du Plessis L, Pretorius PJ. Characterization of the cell-free DNA released by cultured cancer cells. Biochim Biophys Acta 2016;1863:157–65.
18. Abbosh C, Birkbak NJ, Swanton C. Early stage NSCLC—challenges to implementing ctDNA-based screening and MRD detection. Nat Rev Clin Oncol 2018;15:577–86.
19. Xu H, DiCarlo J, Satya VR, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genom 2014;15:1–10.
20. Stead LF, Sutton KM, Taylor GR, Quirke P, Rabbitts P. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution. Hum Mutat 2013;34:1432–8.
21. Andersson D, Kristiansson H, Kubista M, Ståhlberg A. Ultrasensitive circulating tumor DNA analysis enables precision medicine: experimental workflow considerations. Expert Rev Mol Diagn 2021;21:299–310.
22. 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;20:1–15.
23. Filges S, Yamada E, Ståhlberg A, Godfrey TE. Impact of polymerase fidelity on background error rates in next- generation sequencing with unique molecular identifiers/barcodes. Sci Rep 2019;9:3503.
24. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 2012;109:14508–13.
25. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 2012;9:72–4.
26. Xu C, Gu X, Padmanabhan R, Wu Z, Peng Q, DiCarlo J, et al. Smcounter2: an accurate low-frequency variant caller for targeted sequencing data with unique molecular identifiers. Bioinformatics 2019;35:1299–309.
27. Ståhlberg A, Krzyzanowski PM, Egyud M, Filges S, Stein L, Godfrey TE. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. Nat Protoc 2017;12:664–82.
28. Liu H, Li L, Duan L, Wang X, Xie Y, Tong L, et al. High specific and ultrasensitive isothermal detection of microRNA by padlock probe-based exponential rolling circle amplification. Anal Chem 2013;85:7941–7.
29. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009;25:1754–60.
30. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res 2017;27:491–9.
31. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Res 2013;41:e89.

32. Lam SN, Zhou YC, Chan YM, Foo CM, Lee PY, Mok WY, et al. Comparison of target enrichment platforms for circulating tumor DNA detection. Sci Rep 2020;10:4124.
33. Sánchez R, Ayala R, Martínez-López J. Minimal residual disease monitoring with next-generation sequencing methodologies in hematological malignancies. Int J Mol Sci 2019;20:2832.
34. Tie J, Wang Y, Tomasetti C, Li LL, Springer S, Kinde I, et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Sci Transl Med 2016; 8:346ra92.
35. Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. PeerJ 2016;4:e2074.
36. Shugay M, Zaretsky AR, Shagin DA, Shagina IA, Volchenkov IA, Shelenkov AA, et al. MAGERI: computational pipeline for molecular-barcoded targeted resequencing. PLoS Comput Biol 2017;13: 13–7.
37. Sater V, Viailly PJ, Viailly PJ, Lecroq T, Prieur-Gaston É, Bohers É, et al. UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. Bioinformatics 2020; 36:2718–24.
38. 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.
39. Deveson IW, Gong B, Lai K, LoCoco JS, Richmond TA, Schageman J, et al. Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology. Nat Biotechnol 2021;39: 1115–28.