The Journal of Molecular Diagnostics, Vol. 23, No. 2, February 2021

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

TECHNICAL ADVANCE

SnackVar

Check for updates

An Open-Source Software for Sanger Sequencing Analysis Optimized for Clinical Use

Young-gon Kim, Man Jin Kim, Jee-Soo Lee, Jung Ae Lee, Ji Yun Song, Sung Im Cho, Sung-Sup Park, and Moon-Woo Seong

From the Department of Laboratory Medicine, Seoul National University Hospital, Seoul, Republic of Korea

Accepted for publication November 10, 2020.

Address correspondence to Moon-Woo Seong, M.D., Ph.D., Department of Laboratory Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea. E-mail: mwseong@snu.ac.kr.

Despite the wide application of next-generation sequencing, Sanger sequencing still plays a necessary role in clinical laboratories. However, recent developments in the field of bioinformatics have focused mostly on next-generation sequencing, while tools for Sanger sequencing have shown little progress. In this study, SnackVar (*https://github.com/Young-gonKim/SnackVar*, last accessed June 22, 2020), a novel graphical user interface–based software for Sanger sequencing, was developed. All types of variants, including heterozygous insertion/deletion variants, can be identified by SnackVar with minimal user effort. The featured reference sequences of all of the genes are prestored in SnackVar, allowing for detected variants to be precisely described based on coding DNA references according to the nomenclature of the Human Genome Variation Society. Among 88 previously reported variants from four insertion/deletion–rich genes (*BRCA1*, *APC*, *CALR*, and *CEBPA*), the result of SnackVar agreed with reported results in 87 variants [98.9% (93.0%; 99.9%)]. The cause of one incorrect variant calling was proven to be erroneous base callings from poor-quality trace files. Compared with commercial software, SnackVar required less than one-half of the time taken for the analysis of a selected set of test cases. We expect SnackVar to be a cost-effective option for clinical laboratories performing Sanger sequencing. *(J Mol Diagn 2021, 23: 140–148; https://doi.org/10.1016/j.jmoldx.2020.11.001)*

Before the widespread application of next-generation sequencing (NGS), Sanger sequencing was the mainstream technology for DNA sequencing in clinical and research laboratories. Although NGS has replaced Sanger sequencing in many applications,[1–3] Sanger sequencing is still the method of choice for specific kinds of diagnostic tests performed in clinical laboratories. In addition to tests that target small genetic regions, for which Sanger sequencing is the most cost-effective, secondary tests derived from NGS (eg, cross-validation of NGS variant calls, family member studies) have newly emerged as Sanger sequencing applications. Although the general consensus is that Sanger validation is not required for all NGS variant calls,[4–6] NGS does produce a certain fraction of low-quality variant calls, which require an orthogonal method. Sanger sequencing is still the first choice for the validation of less confident NGS variant calls. Family member tests are required not only for the risk assessment and management of family members but also for the correct classification of proband variants according to the American College of Medical Genetics and Genomics guidelines.[7] Except for planned tests, such as trio tests, family member tests are usually ordered after potentially pathogenic variants are found in a proband. For family member tests, Sanger sequencing is a reasonable choice because only the region of interest, where the proband variant is found, needs to be targeted. For these reasons, as the volume of NGS tests increases, Sanger sequencing is also expected to retain its own role in clinical laboratories.

Although the vast majority of research and development in the field of bioinformatics has focused on NGS, software packages for Sanger sequencing have shown little progress

The authors have no funding to report.
Disclosures: None declared.

https://doi.org/10.1016/j.jmoldx.2020.11.001

over time. In clinical laboratories, the most widely used packages are commercial software options such as Sequencher (Gene Codes, Ann Arbor, MI), Mutation Surveyor (SoftGenetics, State College, PA), and SeqScape (Thermo Fisher Scientific, Waltham, MA). Most free software packages are out of date and have significant limitations in terms of their functions and the platforms they run on.[8–10] Even with commercial packages, the application of coding DNA reference sequences and Human Genome Variation Society (HGVS) nomenclature[11] is not straightforward and requires preanalysis configuration steps. In particular, for heterozygous insertion/deletion (indel) variants, only a few tools provide variant calling based on the deconvolution of mutant and wild-type traces.[3,8,12] To the best of our knowledge, no existing software provides a comprehensive application of HGVS nomenclature based on coding DNA references along with specifics such as exact description of deletion/insertion variants, application of the 3′ rule for indel variants,[13] and the description of predicted amino acid changes.

SnackVar (*https://github.com/Young-gonKim/SnackVar*, last accessed June 22, 2020), a graphical user interface–based software, was developed for the easy detection of variants with Sanger sequencing in clinical laboratories. SnackVar enables users to detect variants from trace files with minimum user interaction. A coding DNA reference is applied simply by typing in a Reference Sequence (RefSeq) NM number or by searching RefSeq from a gene name. All kinds of sequence variants, including complex indel variants (eg, deletion/insertion variant with hundreds of indel sizes), can be reported in the correct format of the HGVS nomenclature. When multiple equivalent descriptions exist for an indel variant, the one at the most 3′ position is reported, following the 3′ rule of HGVS. When both forward and reverse traces are available (which they commonly are), the two traces are compared with each other to filter out false variant calls. SnackVar can run on all types of major platforms and operating systems because it is written in Java, which is known for its platform independence. In validation tests using previously reported cases, SnackVar correctly identified all types of variants when trace files of appropriate quality were given, requiring a shorter time for analysis than a commercial software package.

Materials and Methods

Software Development

The implementation of SnackVar is summarized in Figure 1. Java Development Kit 8 (Oracle, Redwood Shores, CA) and its component JavaFX were used to develop the graphical user interface–based software. BioJava Legacy (*https://github.com/biojava/biojava-legacy*, last accessed June 22, 2020) was used to read trace files (.ab1 files). Featured reference sequences were created for all genes in hg19 using coding region and exon information from the ncbiRefSeqCurated table in the University of California Santa Cruz Genome Browser (*http://genome.ucsc.edu*, last accessed June 22, 2020) and they were prepared in SnackVar.

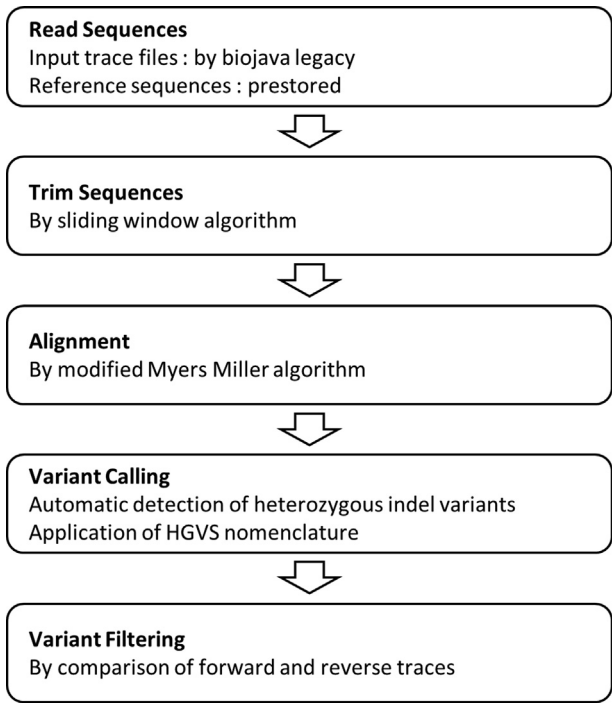

Figure 1 The implementation of SnackVar. HGVS, Human Genome Variation Society; indel, insertion/deletion.

SnackVar has its own trimming criteria based on a sliding window algorithm that uses quality scores. Both the 5′ end and 3′ end are scanned until a window of average quality score of 25 is met. Determination of the 25 threshold was based on repeated tests using development data during development, and the threshold of minimum quality score was not used. The resultant trimming of traces can be modified by users.

For the alignment of sequences, the Myers and Miller algorithm was chosen considering factors such as license, availability, and known performance. Although published in 1988, the Myers and Miller algorithm is one of the most updated versions of the Needleman-Wunsch algorithm–based dynamic programming algorithms. This algorithm was implemented based on the pseudocode presented in the original publication.[14] The Myers and Miller algorithm has improved performance in terms of time and space complexity compared with the Needleman-Wunsch algorithm. Although both of these algorithms are global alignment algorithms, in SnackVar, a local alignment algorithm is needed because the alignment of an input sequence in a relatively longer reference sequence should be found. The Myers and Miller algorithm was modified for local alignment in the same way that the Needleman-Wunsch algorithm was modified to its local alignment version, the Smith-Waterman algorithm. The steps required for transition from the Needleman-Wunsch algorithm to the Smith-Waterman algorithm are simple and

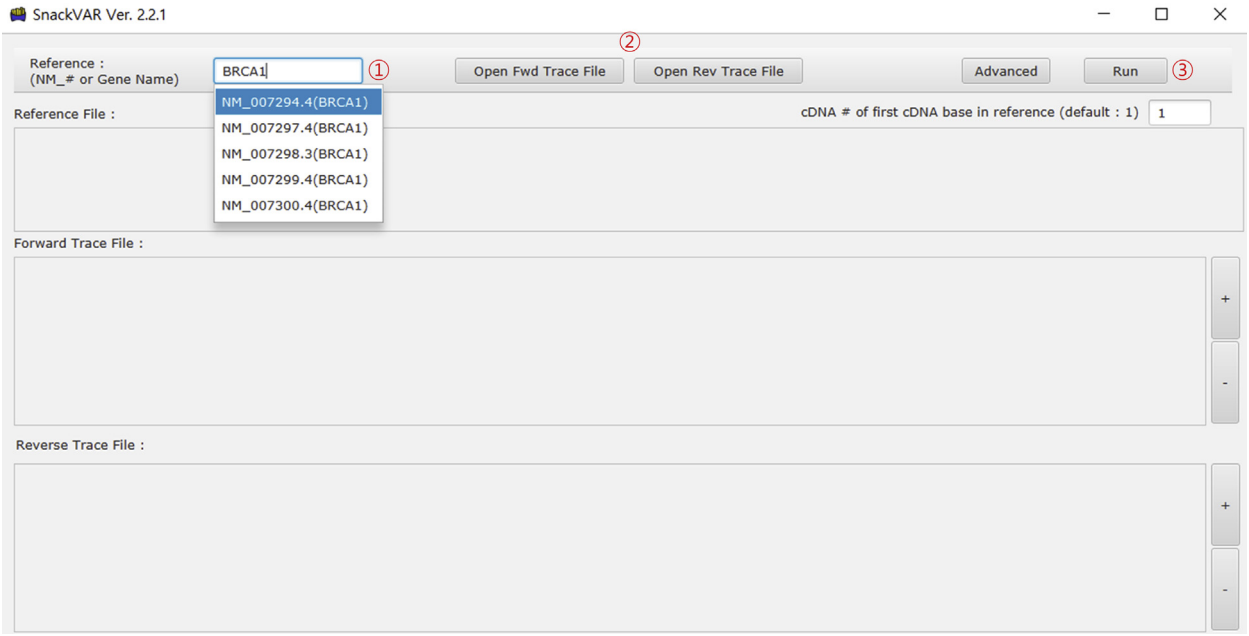

Figure 2 The reference sequence is chosen by searching for a RefSeq transcript with the gene name *BRCA1* (1). After choosing a reference sequence, the analysis is performed by loading the forward/reverse trace files (2) and clicking the Run button (3).

have been described in the literature (*http://biochem218.stanford.edu/Projects%202004/Chan.pdf*, last accessed September 17, 2020). Because the Myers and Miller algorithm is a pairwise alignment algorithm, when both forward and reverse traces are available, the alignment between reference sequence and forward trace sequence and the alignment between reference sequence and reverse trace sequence are independently performed. Subsequently, the results of the two alignments are linearly scanned and combined to make the alignment of three sequences.

The subtraction-based method was used for the detection of heterozygous indel variants from superimposed traces; this method subtracts the reference sequence from the genotype trace to infer the mutated trace, as in most other software packages.[8,15,16] Our approach is most similar to the method described in a work published in 2007,[16] in the sense that ambiguity sequences are first created from the mixed bases and the alignment is performed by using the substitution matrix (EDNAFULL); this method provides a partial alignment score to the match between a reference base and the ambiguity base. Based on this alignment, subtraction is performed to produce the mutated trace, and the inferred mutated trace is compared with the reference sequence base-by-base for the exact calling of an indel variant.

Because simply calling all signals with a peak height higher than the threshold as variants yields too many false-positive variant calls, several variant-filtering heuristics were implemented. Along with the rules based on the peak shape description, the most important part of the variant-filtering algorithm of SnackVar is the consensus of the forward and reverse traces. When both forward and reverse traces are available, a variant detected from only one trace is filtered out if the quality of the corresponding position in the other trace is good.

Data Used for Software Testing and Performance Validation

This study was approved by the institutional review board at Seoul National University Hospital (SNUH), Seoul, Republic of Korea. The stored trace files from the Sanger sequencing analysis of four indel-rich genes (*BRCA1*, *APC*, *CALR*, and *CEBPA*) performed in SNUH were used for the software testing and performance validation of SnackVar. In SNUH, tests of *BRCA1* and *APC* were performed to confirm the NGS results, and all variants detected were germline variants. Tests for *CALR* and *CEBPA*, however, were performed only by Sanger sequencing, and the variants were presumed to be somatic variants from hematologic malignancies. Cases containing at least one positive variant (pathogenic or likely pathogenic for *BRCA1* and *APC*) were included, and when multiple amplicons were tested from a case, only trace files from variant-containing amplicons were used. Variants detected from deep intronic regions (>10 bp apart from the coding region) were excluded from the analysis. Trace files from 54 positive cases tested from August 2018 to October 2019 were chosen and used as development data for software development and primary testing. For development data, previously reported variant descriptions were used from the beginning; each time unexpected results were obtained from SnackVar during development, the corresponding part of the software was modified. After the development of SnackVar was completed, an additional 28 positive cases tested from January 2018 to July 2018 were selected as validation data.

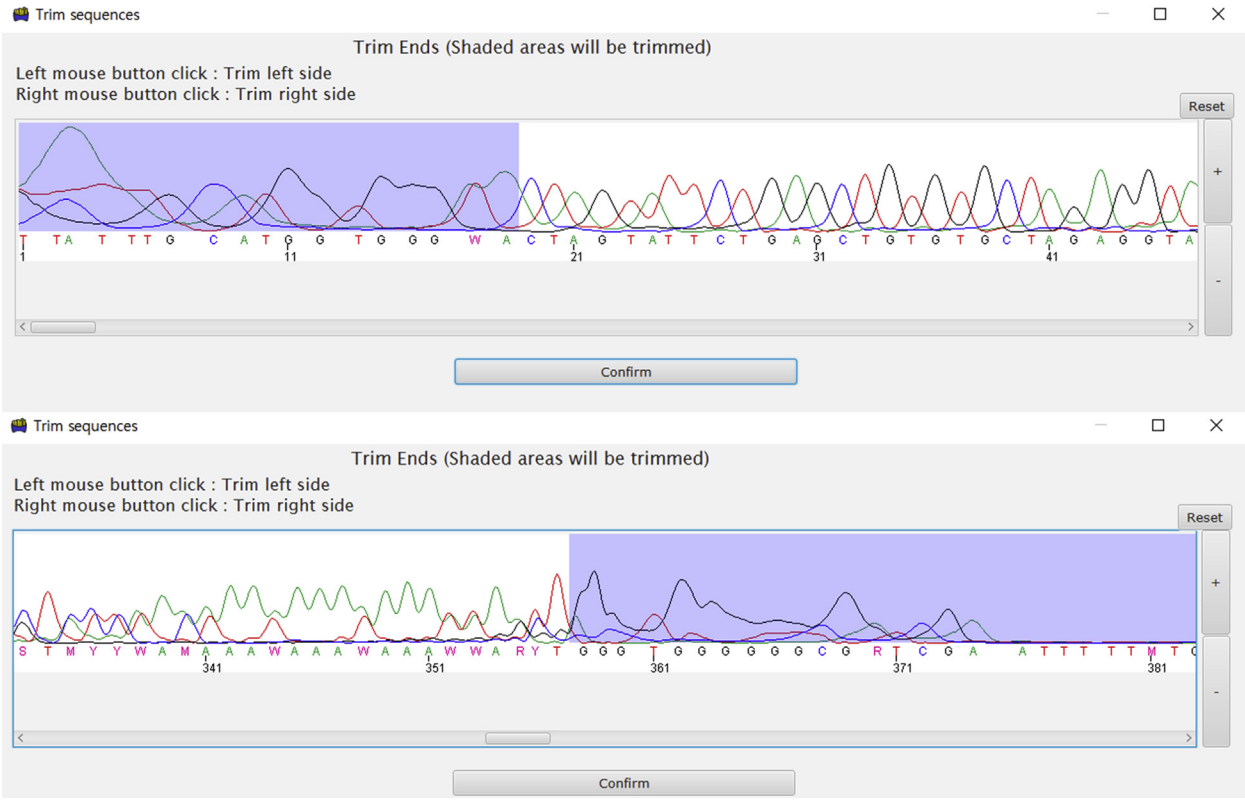

Figure 3 When the length of the trace to be trimmed is longer than the predefined threshold (35 bp), users are suggested to manually adjust the trimming of the traces. The **upper** and **lower portions** of the figure show the 5′ end and 3′ end of an input trace, respectively. This trace has a superimposed portion caused by a heterozygous insertion/deletion. The trimming position of the 3′ end is adjusted to contain the maximal number of valid double-peak bases and a minimal number of noisy signals.

The cases of validation data were analyzed by using SnackVar without information of reported variant descriptions, and the results were compared thereafter. Among the 28 cases of validation data, 17 representative cases with unique variants were chosen and used for an additional analysis to compare SnackVar versus a commercial software.

The detection of homozygous variants in Sanger sequencing was easier than that of heterozygous variants for both single-nucleotide variants (SNVs) and indel variants. They can be called directly from the alignment of sequences. Because pathogenic homozygous variants are rare and were not included in the collected cases, the validation of the detection of homozygous variants was not separately described in this study. However, the detection of homozygous SNVs could be validated through many benign homozygous SNVs in the cases included. The detection of homozygous indel variants was validated through the modification of reference sequences, which resulted in the creation of artificial homozygous indel variants from wild-type traces.

Comparison with Commercial Software

Seventeen cases with unique variants were chosen from validation data and were used for comparison analysis. Selected cases were analyzed by using both SnackVar and a commercial software used in the SNUH laboratory, SeqScape (version 2.7). Two laboratory technicians experienced in Sanger sequencing (10 and 15 years' experience, respectively) participated in the experiment. Both technicians were familiar with SeqScape and were given a brief training session on using SnackVar before performing the test. Both participants analyzed the cases using both SnackVar and SeqScape. The amount of time spent for the analysis and the sensitivity of variant detection were compared between the two tools. For fairness of comparison, test cases were divided into two sets of similar number of cases, and one set was tested by SnackVar first and then by SeqScape, and the other set was tested by SeqScape first and then by SnackVar. The order of the sets tested was opposite for the two participants. The time required for constructing featured reference sequences in SeqScape was not included in the comparison of time. Because SeqScape often generates multiple variant calls from a single indel variant and does not provide exact application of HGVS nomenclature, description of heterozygous indel variants was performed manually with the assistance of handwriting on printed traces, which is a legacy process in the SNUH laboratory.

Evaluation of Limit of Detection

To evaluate the limit of detection (LOD) of SnackVar in terms of variant allele frequency (VAF), a commercially available control material with known VAF, OncoSpan gDNA (Horizon Discovery, Cambridge, UK) was used.

Figure 4 The result of alignment is shown and can be browsed at the top. Chromatograms below are focused according to the mouse click on the alignment. In this example, a variant NM_007294.4 (*https://www.ncbi.nlm.nih.gov/nuccore*; *BRCA1*) :c.5030_5033del is detected and is described in the Variant list table at the bottom. A list of the equivalent expressions describing the same variant is shown in the Equivalent Expressions column. The one using the most 3′ coordinate, c.5030_5033del, is given in the Variant column. The expected amino acid change, p.(Thr1677Ilefs*2), is also described.

Four *BRCA2* variants, including two SNVs and two short indels, were chosen; their VAFs ranged from 25.0% to 47.0% (Supplemental Table S1). Considering that the LOD of Sanger sequencing is generally recognized as 15% to 20%,[17] three independent dilutions were made comprising one-half, one-third, and one-quarter dilution from the original sample to cover approximately 10% of VAF. The details of variants and their dilutions are shown in Supplemental Table S2.

Statistical Analysis

Calculation of percent agreements and their 95% CIs were performed by using an online statistical software program

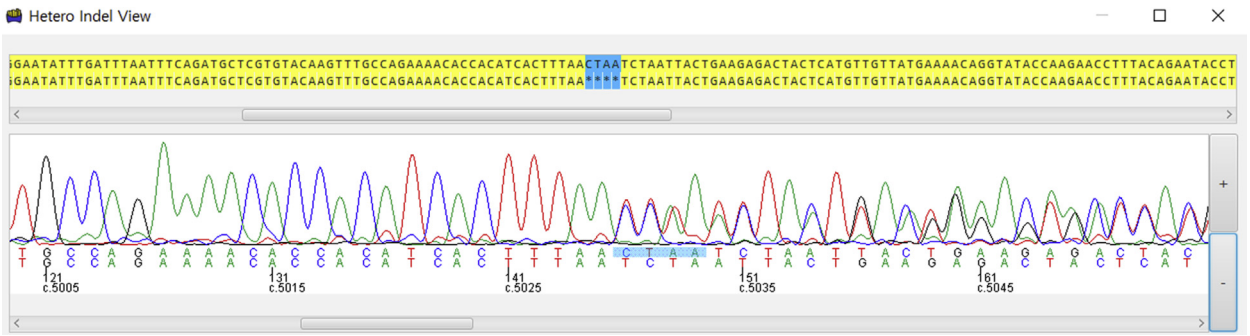

Figure 5 Heterozygous insertion/deletion (hetero indel) view of a variant NM_007294.4 (*https://www.ncbi.nlm.nih.gov/nuccore*, *BRCA1*):c.5030_5033del. The detected hetero indel variants can be reviewed by using the hetero indel view function. The mutant trace and the wild-type trace are deconvoluted, and the sequences are displayed separately. The differences between the wild-type and mutant sequences are shaded in **blue**. Bases with a **yellow background** indicate matches between the wild-type and mutant traces after shifting the mutant trace.

Table 1 Summary of Variants Detected

Variant type	No. of variants (correctly detected by SnackVar)		
	Development data (54 cases)	Validation data (28 cases)	Overall (82 cases)
SNV	19 (19, 100.0%)	5 (5, 100.0%)	24 (24, 100.0%)
indel	39 (38, 97.4%)	25 (25, 100.0%)	64 (63, 98.4%)
Overall	58 (57, 98.3%)	30 (30, 100.0%)	88 (87, 98.9%)

All variants are heterozygous variants.
indel, insertion/deletion; SNV, single-nucleotide variant.

(VassarStats Kappa, *http://vassarstats.net/kappa.html*, last accessed October 15, 2020).

Results

SnackVar Usage

The initial status of SnackVar after launching is depicted in Figure 2. The first step in the analysis is the selection of a reference sequence. Because the featured reference sequences of all genes are prestored in SnackVar, users can simply choose one by searching a RefSeq transcript ID (Figure 2). Forward and/or reverse trace files can be loaded by using the Open Fwd Trace File button and the Open Rev Trace File button. When the trace to be trimmed is shorter than the predefined threshold (default, 35 bp; can be adjusted in the Advanced menu), automatic trimming is performed without user confirmation. When the trace to be trimmed is not shorter than the threshold, SnackVar requires the user to review the trace, confirm the trimming position, and modify the trimming position if needed (Figure 3).

After loading the trace files, a user can click the Run button to perform the analysis. The detected variants are listed in a Variant List table in the form of HGVS nomenclature with some additional information such as expected amino acid changes (Figure 4). When an indel variant is detected, the variant is checked to see if it has equivalent descriptions. If these exist, the variant with the most 3′ coordinate is chosen as a representative, and all other expressions are listed in the

Table 2 Results from the Validation Data

Case no.	Gene	Reported result	SnackVar output
1†	*BRCA1*	c.3627dupA, p.Glu1210Argfs*9	c.3627dup, p.(Glu1210Argfs*9)
2†	*BRCA1*	c.5030_5033del, p.Thr1677Ilefs*2	c.5030_5033del, p.(Thr1677Ilefs*2)
3†	*BRCA1*	c.5496_5506del11insA, p.Val1833Serfs*7	c.5496_5506delinsA, p.(Val1833Serfs*7)
4†	*BRCA1*	c.5080G>T, p.Glu1694*	c.5080G>T, p.(Glu1694*)
5	*BRCA1*	c.5030_5033delCTAA, p.Thr1677Ilefs*2	c.5030_5033del, p.(Thr1677Ilefs*2)
6†	*BRCA1*	c.3296delC, p.Pro1099Leufs*10	c.3296del, p.(Pro1099Leufs*10)
7†	*BRCA1*	c.390C>A, p.Tyr130*	c.390C>A, p.(Tyr130*)
8†	*BRCA1*	c.3331_3334delCAAG, p.Gln1111Asnfs*5	c.3331_3334del, p.(Gln1111Asnfs*5)
9	*BRCA1*	c.390C>A, p.Tyr130*	c.390C>A, p.(Tyr130*)
10†	*BRCA1*	c.5074+1G>T, p?	c.5074+1G>T
11†	*BRCA1*	c.1511dupG, p.Lys505*	c.1511dup, p.(Lys505*)
12	*BRCA1*	c.5496_5506del11insA, p.Val1833Serfs*7	c.5496_5506delinsA, p.(Val1833Serfs*7)
13	*BRCA1*	c.3157del, p.Glu1053LysfsTer9	c.3157del, p.(Glu1053Lysfs*9)
14	*BRCA1*	c.3991C>T, p.Gln1331Ter	c.3991C>T, p.(Gln1331*)
15†	*BRCA1*	c.922_924delinsT, p.Ser308*	c.922_924delinsT, p.(Ser308*)
16†	*APC*	c.2216_2219dup, p.Asn741Cysfs*16	c.2216_2219dup, p.(Asn741Cysfs*16)
17†	*APC*	c.3317dupG, p.Ala1107Serfs*12	c.3317dup, p.(Ala1107Serfs*12)
18	*CALR*	c.1099_1150del, p.Leu367Thrfs*46	c.1099_1150del, p.(Leu367Thrfs*46)
19	*CALR*	c.1099_1150del, p.Leu367Thrfs*46	c.1099_1150del, p.(Leu367Thrfs*46)
20	*CALR*	c.1154_1155insTTGTC, p.Lys385Asnfs*47	c.1154_c.1155insTTGTC, p.(Lys385Asnfs*47)
21	*CALR*	c.1154_1155insTTGTC, p.Lys385Asnfs*47	c.1154_c.1155insTTGTC, p.(Lys385Asnfs*47)
22	*CALR*	c.1099_1150del, p.Leu367Thrfs*46	c.1099_1150del, p.(Leu367Thrfs*46)
23	*CALR*	c.1154_1155insTTGTC, p.Lys385Asnfs*47	c.1154_c.1155insTTGTC, p.(Lys385Asnfs*47)
24†	*CALR*	c.1099_1150del, p.Leu367Thrfs*46	c.1099_1150del, p.(Leu367Thrfs*46)
25†	*CALR*	c.1154_1155insTTGTC, p.Lys385Asnfs*47	c.1154_c.1155insTTGTC, p.(Lys385Asnfs*47)
26†	*CEBPA*	c.247delC, p.Gln83Serfs*77	c.247del, p.(Gln83Serfs*77)
		c.899_961dup, p.Asp320_Asp321ins21	c.899_961dup, p.(Asp320_Asn321insSerAspLysAlaLysGlnArgAsnValGluThrGlnGlnLysValLeuGluLeuThrSerAsp)
27†	*CEBPA*	c.199dup, p.Tyr67Leufs*41	c.199dup, p.(Tyr67Leufs*41)
28†	*CEBPA*	c.97_100del, p.Phe33Profs*126	c.97_100del, p.(Phe33Profs*126)
		c.928_930dup, p.Thr310dup	c.928_930dup, p.(Thr310dup)

All variants are heterozygous variants.
Transcripts can be found on Nuccore (*https://www.ncbi.nlm.nih.gov/nuccore*, last accessed August 9, 2020) *BRCA1*: NM_007294.4, *APC*: NM_000038.6), *CALR*: NM_004343.3, and *CEBPA*: NM_004364.4.
†Cases used in the comparison analysis.

Table 3 Results of Time Comparison Analysis

Tester	Case set	SeqScape (minutes)	SnackVar (minutes)	SeqScape—SnackVar (minutes)
Tester 1	Set 1	45	17.5	27.5
	Set 2	55	27.5	27.5
	Total	100	45	55
Tester 2	Set 1	45	22.5	22.5
	Set 2	60	22.5	37.5
	Total	105	45	60

Set 1: Cases 1, 2, 3, 4, 6, 7, 8, 10, 11, and 15 from the validation data (10 cases, 10 variants). Set 2: Cases 16, 17, 24, 25, 26, 27, and 28 from the validation data (7 cases, 9 variants). Order of tests performed by Tester 1: Set 1 (SnackVar), Set 2 (SeqScape), Set 1 (SeqScape), and Set 2 (SnackVar). Order of tests performed by Tester 2: Set 1 (SeqScape), Set 2 (SnackVar), Set 1 (SnackVar), and Set 2 (SeqScape).

Equivalent Expressions column of the Variant list table. Variant calling of heterozygous indel variants can be reviewed by using the Hetero Indel View button (Figure 5).

Two parameters of SnackVar, namely gap opening penalty and the cutoff for double peak detection, can be adjusted in the Advanced menu. The default values are 30 for the gap opening penalty and 0.3 for the cutoff for double peak detection. The default value of the gap opening penalty gives the expected alignment in most cases, but some heterozygous indel variants from poor-quality traces with erroneous base callings can only be detected with a higher gap opening penalty. In such cases, SnackVar automatically applies a higher gap opening penalty and notifies the user with a pop-up that the heterozygous indel optimization mode is activated. There were no tested cases in which manual adjustment of the gap opening penalty was necessary. In somatic cases with a low fraction of mutations, using a lower cutoff for double peak detection such as 0.2 or 0.1 is preferable.

Validation Using Reported Data

A summary of the test cases and variants detected is presented in Table 1. There were 58 previously reported variants in development data, and the result of SnackVar showed agreement with reported results in 57 variants [98.3% (89.5%; 99.9%)] (Supplemental Table S2). One somatic variant from the *CEBPA* gene, c.922_1058dup, was incorrectly called. This duplication of 137 bp was called the insertion of 136 bp between c.920 and c.921 by SnackVar based on erroneous base callings (which was proven as erroneous by human review) from poor-quality trace files. All variant calls of SnackVar from the validation data [$N = 30$; 100.0% (85.9%; 100.0%)] showed agreement with reported variants (Table 2). The total number of variants contained in the development data and validation data was 88. Disregarding recurrently appearing variants, there were a total of 51 unique variants, including 36 different heterozygous indel variants. Among the 88 variants, 87 [98.9% (93.0%; 99.9%)] were correctly identified by the final version of SnackVar. The longest heterozygous indel variant correctly detected by SnackVar was c.899_961dup from the *CEBPA* gene.

Comparison Analysis with Commercial Software

The 17 cases chosen for the comparison analysis are shown in Table 2 (cases marked with a dagger). Both participants correctly identified all 19 variants using both SnackVar and SeqScape. The results of the time comparison analysis are shown in Table 3. Using SeqScape, the overall amount of time required for the analysis of all cases was 100 minutes and 105 minutes for the two participants, respectively; using SnackVar, it took each participant 45 minutes. SnackVar consistently required a shorter time for the analysis for both case sets and for both testers. The effect of exposure of cases to the participants during the repeated tests was not observed because the differences in time between SeqScape and SnackVar were consistently higher in the SnackVar-prior and SeqScape-later configurations, which are Set 1 by Tester 1 and Set 2 by Tester 2.

Evaluation of LOD

For two SNVs and two short indel variants, all dilutions from one-half to one-quarter dilution were correctly detected by SnackVar. To determine a heterozygous peak of low VAF, however, we necessarily had to adjust the cutoff value for double peak detection from the default value 0.3 to lower values such as 0.2 or 0.1, especially in case of SNVs.

Discussion

The current study describes the development and validation of SnackVar, a platform-independent software for Sanger sequencing. Based on a simple and intuitive user interface, SnackVar is easy to use without specific user instructions. Prestored reference sequences make the process simpler. Compared with a commercial software package, SnackVar required a much shorter time for the analysis of test cases. The variant that was incorrectly identified by SnackVar had poor-quality trace files containing base calling errors in the indel region and could only be identified via thorough human inspection by an experienced examiner. To the best of our knowledge, SnackVar is the first Sanger sequencing analysis tool that allows for the precise and comprehensive application of HGVS nomenclature based on coding DNA reference and the 3′ rule for indel description. Predicted amino acid changes are also presented by SnackVar according to the HGVS nomenclature.

In a recent publication, a new command line—based software for Sanger sequencing, Tracy,[3] was described. The Tracy results are provided in Variant Call Format/Binary Variant Call Format, to be viewed by using graphical user interface—based web services provided by the authors; they can also be used in downstream bioinformatics pipelines such as annotation tools. Although this method may be

convenient for integration purposes, it requires additional tools or steps for stand-alone Sanger sequencing analysis. Being a command line−based tool that runs only on Linux and Mac, Tracy may not be easy to use for all operators. The performance of variant detection was not validated in this study, and their trimming policy, which is based on the fixed, user-defined number of bases, may limit the wide applicability to heterogeneous traces with varying quality.

There are situations in which mismatches in the alignment can be represented as either gaps or substitutions. For an extreme example, one substitution can be represented as one insertion and one deletion. The rationale of using the gap opening penalty is to give penalties to creation of gaps in the alignment because substitutions are more common events in nature. Usually, multiples of the substitution penalty (eg, four times in a previous study[18]) are set as the gap opening penalty. There is no consensus on its optimal value as it is related to the likelihood of encountering a gap instead of a substitution in a given data set.[19] Because a single substitution has a penalty of four in an EDNAFULL substitution matrix used in SnackVar, the default value of a gap opening penalty of 30 is 7.5 times the substitution penalty. A smaller gap opening penalty than this value occasionally resulted in incorrect homozygous indel variant calls in SnackVar, usually at low-quality ends of traces. The gap opening penalty of 30 removed most of the inappropriate gap creation at the lower quality ends and gave an expected alignment in most cases. As described in the Results, a higher value of the gap opening penalty (up to 200) is automatically applied when the heterozygous indel variant is called misleadingly as a homozygous indel variant due to an inappropriate gap creation in the alignment. As far as we tested, users do not need to manually adjust the gap opening penalty option.

The LOD of Sanger sequencing is generally recognized as 15% to 20%, and the evaluation of LOD using commercialized material revealed that SnackVar can effectively detect somatic variants with VAFs as low as this limit. Using a lowered cutoff for second peak detection, SnackVar could detect variants with a lower VAF than this limit. However, due to the lack of test material, this LOD evaluation was not performed for long indel variants.

One limitation of SnackVar is that it is capable of processing only one amplicon at a time. Although this method can maximize simplicity, it may be viewed as inconvenient when users have many cases to process from the same region at one time or have to analyze cases with many amplicons. Although cases with many amplicons are mostly analyzed by using NGS, there are situations in which batch processes are desirable. Advantages of SnackVar in terms of speed compared with SeqScape revealed in our comparison analysis might have been exaggerated because only variant-containing amplicons were included in the analysis. For large-scale laboratories with extensive cases of Sanger sequencing analysis, SnackVar could be adjunctively used for indel detection, in addition to their commercially licensed software packages such as Sequencher or SeqScape. For smaller laboratories with limited cases of Sanger sequencing, SnackVar could be a cost-effective alternative to any type of tool they are using. Another limitation is that, as with all other automated systems used in the clinical environment, the results of SnackVar should not be directly incorporated into the report. The results of SnackVar should be confirmed by human users before reporting, especially when the quality of trace files seems suboptimal. Finally, in the comparison analysis, even if we reversed the order of software packages tested by the two participants, repeated tests using the same data set might have resulted in improved performance of both software packages in terms of both accuracy and speed.

In conclusion, SnackVar, a novel software for Sanger sequencing analysis, was developed. SnackVar is expected to identify all kinds of variants, including heterozygous indel variants, from trace files of appropriate quality. The detected variants are reported in the correct format of HGVS nomenclature and can be directly used in clinical reports without transformation. SnackVar may be useful for many clinical laboratories performing Sanger sequencing.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2020.11.001*.

References

1. Fogel BL, Lee H, Strom SP, Deignan JL, Nelson SF: Clinical exome sequencing in neurogenetic and neuropsychiatric disorders. Ann N Y Acad Sci 2016, 1366:49−60
2. Totomoch-Serra A, Marquez MF, Cervantes-Barragán DE: Sanger sequencing as a first-line approach for molecular diagnosis of Andersen-Tawil syndrome. F1000Res 2017, 6:1016
3. Rausch T, Fritz MH, Untergasser A, Benes V: Tracy: basecalling, alignment, assembly and deconvolution of Sanger chromatogram trace files. BMC Genomics 2020, 21:230
4. Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ: Confirming variants in next-generation sequencing panel testing by Sanger sequencing. J Mol Diagn 2015, 17:456−461
5. Beck TF, Mullikin JC, NISC Comparative Squencing Program, Biesecker LG: Systematic evaluation of Sanger validation of next-generation sequencing variants. Clin Chem 2016, 62:647−654
6. Mu W, Lu HM, Chen J, Li S, Elliott AM: Sanger confirmation is required to achieve optimal sensitivity and specificity in next-generation sequencing panel testing. J Mol Diagn 2016, 18:923−932
7. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015, 17:405−424
8. Carr IM, Camm N, Taylor GR, Charlton R, Ellard S, Sheridan EG, Markham AF, Bonthron DT: GeneScreen: a program for high-throughput mutation detection in DNA sequence electropherograms. J Med Genet 2011, 48:123−130

9. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Van Broeckhoven C, De Rijk P: novoSNP, a novel computational tool for sequence variation discovery. Genome Res 2005, 15:436–442
10. Treves DS: Review of three DNA analysis applications for use in the microbiology or genetics classroom. J Microbiol Biol Educ 2010, 11: 186–187
11. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE: HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat 2016, 37:564–569
12. Dong C, Yu B: Mutation surveyor: an in silico tool for sequencing analysis. Edited by Silico Tools for Gene Discovery. Springer, 2011. pp. 223–237
13. Deans ZC, Fairley JA, den Dunnen JT, Clark CJ: HGVS nomenclature in practice: an example from the United Kingdom National External Quality Assessment Scheme. Hum Mutat 2016, 37:576–578
14. Myers EW, Miller W: Optimal alignments in linear space. Comput Appl Biosci 1988, 4:11–17
15. Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, Wilson RK, Mardis ER: PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. Genome Res 2007, 17: 659–666
16. Tenney AE, Wu JQ, Langton L, Klueh P, Quatrano R, Brent MR: A tale of two templates: automatically resolving double traces has many applications, including efficient PCR-based elucidation of alternative splices. Genome Res 2007, 17:212–218
17. Tsiatis AC, Norris-Kirby A, Rich RG, Hafez MJ, Gocke CD, Eshleman JR, Murphy KM: Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. J Mol Diagn 2010, 12:425–432
18. Chao KM, Miller W: Linear-space algorithms that build local alignments from fragments. Algorithmica 1995, 13:106–134
19. Carroll H, Clement MJ, Ridge P, Snell QO: Effects of gap open and gap extension penalties. In Proceedings of the Biotechnology and Bioinformatics Symposium (BIOT), Provo, Utah, 20-21; October 2006. pp. 19–23