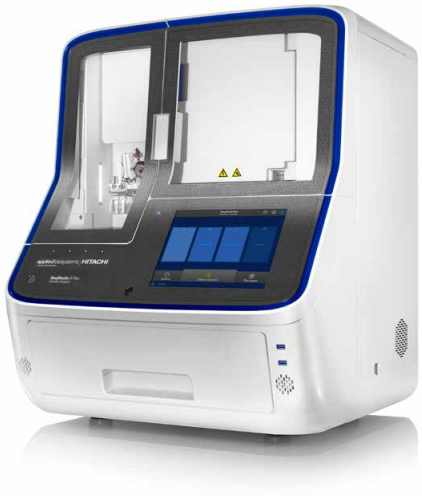
Sanger sequencing and fragment analysis

SeqStudio Flex system—the latest-generation medium-throughput genetic analyzer



In this application guide, we demonstrate:

- High data quality and comparability to other Applied Biosystems™ genetic analyzers
- The intuitive run setup and secondary analysis software
- Highlights of Sanger sequencing and fragment analysis applications, including:
  - Cell line authentication
  - Microsatellite instability
  - dsDNA QC
  - Multiplex PCR analysis
  - Gene editing confirmation
  - SARS-CoV-2 sequencing
  - Minor allelic variant sequencing
  - Next-generation sequencing confirmation
  - Plasmid sequencing
  - Difficult template sequencing

Introduction

Fluorescent capillary electrophoresis (CE) is a genomic analysis method that separates fluorescently labeled DNA fragments based on size, and is the foundation of Sanger sequencing and fragment analysis. The simple workflow, single-base resolution, rapid analysis time, small sample volume, and flexibility have resulted in widespread adoption for a variety of applications used in basic, translational, and clinical research as well as drug discovery and the development of novel therapies and vaccines. The Applied Biosystems™ brand is a leader in CE reagents, instrumentation, and software, and continues to provide innovative solutions that meet modern research needs.

The Applied Biosystems™ SeqStudio™ Flex system, the latest addition to our CE portfolio, is the most advanced medium-throughput genetic analyzer that produces high-quality Sanger sequencing and fragment analysis results. The SeqStudio Flex instrument is founded on well-proven fluorescent CE technology (Figure 1A) while offering new capabilities that improve ease of use, flexibility, serviceability, and connectivity for improved laboratory efficiency and the freedom to focus on your science. New features include continual plate loading and 4-plate capacity (Figure 1B), an on-board computer with an improved, intuitive user interface (Figure 1C), remote run and data monitoring (Figure 1D–F), remote troubleshooting, and dynamic spectral calibration. These innovations in usability and advanced communication simplify the CE workflow, bringing the flexibility of Applied Biosystems CE instruments to bear on new and exciting genetic analysis applications.

In this guide, we introduce a selection of the broad spectrum of genetic analysis applications that can be run on the SeqStudio Flex instrument, and show that the quality of the data produced is equivalent to that of other Applied Biosystems™ genetic analyzers, including the 4-capillary SeqStudio™ and 24-capillary 3500xL systems.

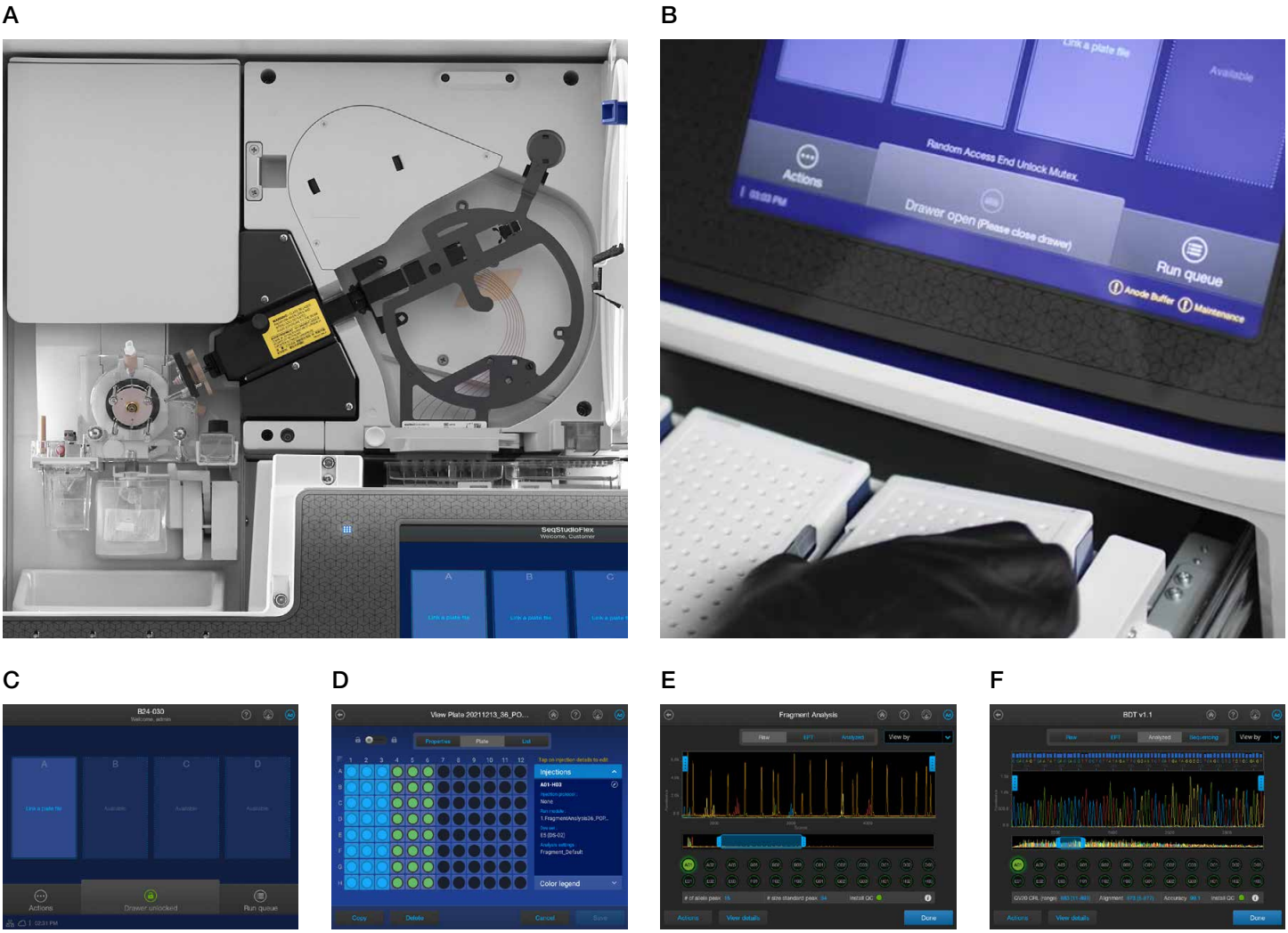


Figure 1. Introducing the SeqStudio Flex Genetic Analyzer. (A) The SeqStudio Flex system is built on the same foundational platform and uses the same buffer and polymer consumables as the 3500 Series Genetic Analyzers. **(B)** Up to 4 plates can be loaded on the instrument, and the software allows users to define and load a plate during an ongoing run. **(C)** On-board touchscreen computer with intuitive user interface. **(D)** Streamlined plate addition or modification of plates. Fragment analysis **(E)** or sequencing data **(F)** can be visualized in real time or once a run is completed. Data can be transferred from the instrument by USB drive, a local area network (LAN), or Wi-Fi–based remote access.

Results

Fragment analysis applications

Fragment analysis is a highly flexible method that involves the separation of different-sized and differentially labeled DNA fragments by CE. One of the most common methods of generating fragments for analysis is polymerase chain reaction (PCR); because of the flexibility afforded with the choice of PCR primers, a specifically sized fragment corresponding to a PCR target sequence is straightforward to generate. Along with the ability to label fragments with up to four different fluorophores, researchers have a large degree of flexibility in experimental design.

Cell line authentication

Disease research relies on the analysis of samples obtained and manipulated *ex vivo*. These include human cell lines, induced pluripotent stem cells (iPSCs), and chimeric antigen receptor (CAR) T cells. Contamination and human error can occur and impact research results or put lives in danger.

For example, the International Cell Line Authentication Committee (ICLAC) has found that there are at least 451 cell lines being used by researchers that are misidentified with no known authentic stock [1]. Furthermore, a study from 2019 found that that 32,755 articles reported results obtained with misidentified cells—these were in turn found to be cited in an estimated half a million other publications [2].

Finally, as research progresses into *ex vivo* cell therapies—cells that are removed from a donor, manipulated in a lab, and returned to a host—it is critical to confirm that donor cells have the expected genotype and match the intended recipient. Therefore, it is crucial to know the provenance of human cells and confirm that they have the desired identity.

An array of Applied Biosystems™ solutions for identifying human samples are based on fragment analysis of short tandem repeats (STRs). STRs are microsatellite sequences that are highly variable and provide a unique molecular fingerprint for a human sample. The Applied Biosystems™ GlobalFiler™ cell line authentication (CLA) kit generates a molecular fingerprint for 24 different STR loci, while the Applied Biosystems™ Identifiler™ Plus CLA kit analyzes 16 STR loci. Both of these easy-to-use kits are optimized for use on the Applied Biosystems family of CE-based genetic analyzers. For more information, see reference 3.

To illustrate the performance of the SeqStudio Flex instrument relative to other Applied Biosystems genetic analyzers, DNA from six commonly used human cell lines (A549, Jurkat, U2OS, HEK293, M4A4GFP, and HeLa) was purified using Invitrogen™ RecoverAll™ kits. For STR analysis, 1 ng of genomic DNA was used with either the GlobalFiler CLA or Identifiler CLA kit according to the supplied protocols. Each sample was run in triplicate. Samples were run on the SeqStudio Flex, 3500xL, and SeqStudio genetic analyzers. Allelic calls were made using the Applied Biosystems™ GeneMapper™ 6.1 software. The authenticity of the cell lines was confirmed by checking the results against databases for STRs of common cell lines [4].

The STR peak profiles at all loci for all five cell lines were comparable across all instruments (Figure 2A). Using the GlobalFiler CLA kit, we found a 100% match to the expected STR allelic profile for all cells using the SeqStudio Flex and 3500xL genetic analyzers, and a 98.7% match with the SeqStudio genetic analyzer (Tables 1 and 2). Similarly, using the Identifiler CLA kit, we obtained a 100% match with the SeqStudio Flex system, a 99.6% match with the 3500xL system, and a 98.8% match with the SeqStudio system.

To examine the ability to detect cell line contamination, we mixed M4A4GFP and HeLa gDNA at 50%, 25%, 10%, and 5% HeLa DNA. The Identifiler CLA kit was used to analyze 1 ng of each gDNA mixture. HeLa cells have 16 alleles that are not found in the M4A4GFP profile. We found that even when the contaminating HeLa cells make up 5% of a gDNA sample, the genetic analyzers can detect between 9% and 23% of the unique HeLa alleles (Figure 2B). Together, these results demonstrate that the SeqStudio Flex instrument generates cell line authentication results of similar or better quality compared to the rest of the genetic analyzer family.

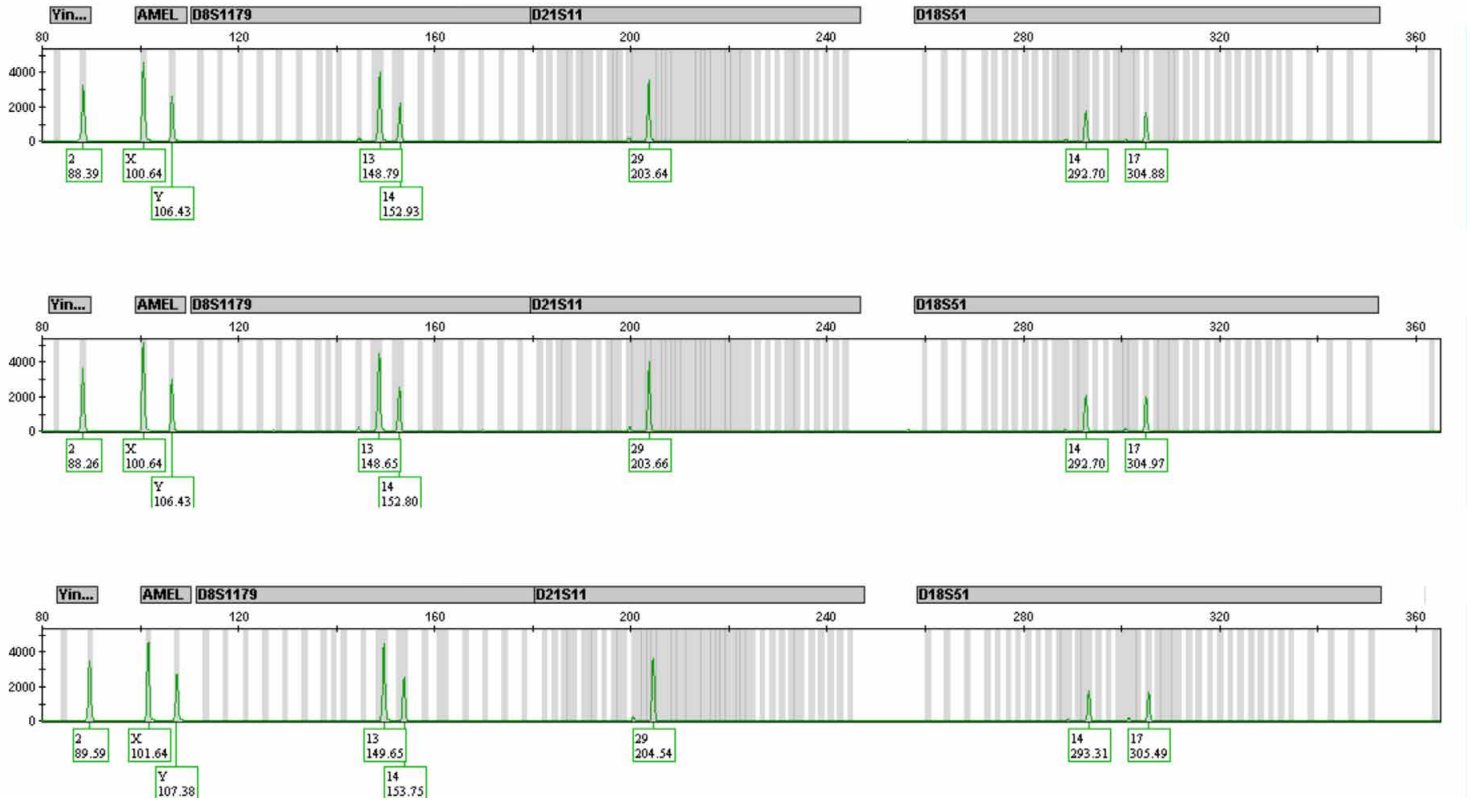


Figure 2A. CLA peak morphologies on three genetic analyzer platforms. The SeqStudio Flex (top), 3500xL (middle), and SeqStudio (bottom) genetic analyzers produce similar peak morphologies and allele calls using the GlobalFiler CLA kit. A single channel is shown; similar results were seen in all five colored channels and with additional samples.

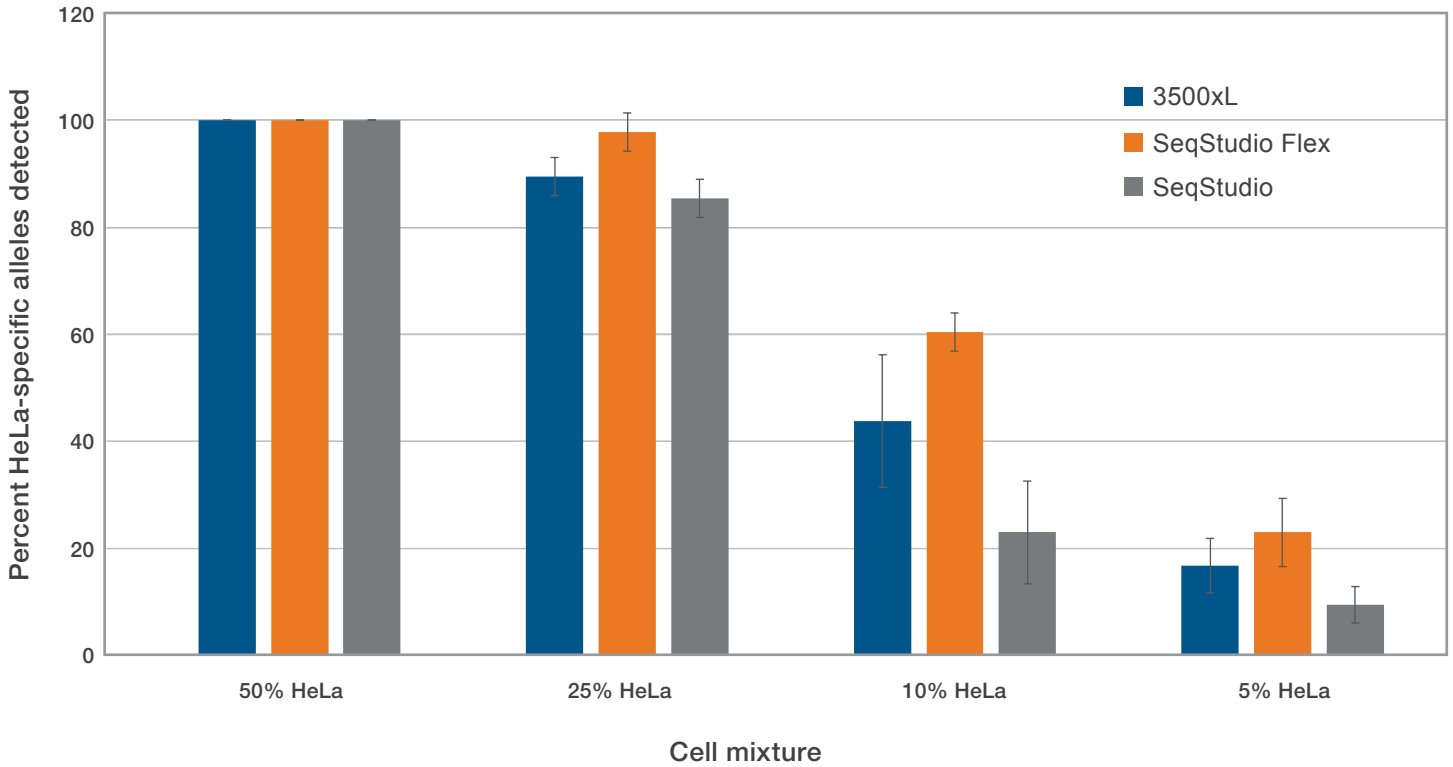


Figure 2B. Detecting contaminating alleles at low concentrations. Decreasing amounts of HeLa gDNA were added to M4A4GFP gDNA. The total amount of DNA was kept to 1 ng. The mixture was analyzed in triplicate using the IdentiFiler Plus CLA kit. The number of HeLa-specific alleles detected in each titration versus number expected was determined. Note that HeLa-specific alleles were detected in the 5% sample on all three instruments.

Table 1. Allelic profiles were collected on SeqStudio Flex, 3500xL, and SeqStudio genetic analyzers. The example shown is from human A549 cells analyzed using the GlobalFiler CLA kit.

Locus	Channel	SeqStudio Flex		3500xL		SeqStudio	
AMEL	G	X	Y	X	Y	X	Y
CSF1PO	B	10	12	10	12	10	12
D10S1248	P	13	16	13	16	13	16
D12S391	P	18		18		18	
D13S317	R	11		11		11	
D16S539	B	11	12	11	12	11	12
D18S51	G	14		14		14	
D198433	Y	13		13		13	
D1S1656	P	17	18.3	17	18.3	17	18.3
D21S11	G	29		29		29	
D22S1045	R	15		15		15	
D2S1338	P	24		24		21	
D25441	Y	10	13	10	13	10	13
D3S1358	B	16		16		16	
D5S818	R	11		11		11	
D7S820	R	8	11	8	11	8	11
D8S1179	G	13	14	13	14	13	14
DYS391	G	10		10		10	
FGA	Y	23		23		23	
SE33	R	19	25.2	19	25.2	19	25.2
TH01	Y	8		8		8	
TPOX	B	8	11	8	11	8	11
VWA	B	14		14		14	
Yindel	G	2		2		2	

Table 2. The total number of alleles that were correctly called using the GlobalFiler CLA kit was totaled across three replicates for six cell lines. Each cell line has a slightly different combination of homozygous and heterozygous allele numbers, so the total number of alleles will vary between cell lines.

	SeqStudio Flex	3500xL	SeqStudio
A549	102/102	102/102	102/102
Jurkat	153/153	153/153	153/153
U2OS	105/105	105/105	105/105
HEK293	111/111	111/111	104/111*
M4A4GEP	96/96	96/96	96/96
HeLa	117/117	117/117	115/117*

* Technical issues during one of three injections resulted in some anomalous calls in these samples.

Microsatellite instability

Many types of cancer display deficiencies in DNA mismatch repair (MMR), producing an overall higher mutation rate across the genome [5]. A higher mutation rate often means a higher rate of neoantigen production, providing opportunities for immune therapy research [6].

There have been at least 11 different loci implicated in MMR [7]. Looking for an inactivating event in the sequence of all these loci can be complicated, time consuming, and expensive. Examining the outcome of perturbations in the MMR pathway by microsatellite instability (MSI) analysis provides a practical alternative. MSI analysis can be difficult on NGS systems due to the highly repetitive nature of microsatellite sequences. Thus, fragment analysis of microsatellite loci length using CE is a widely used method for detecting MSI.

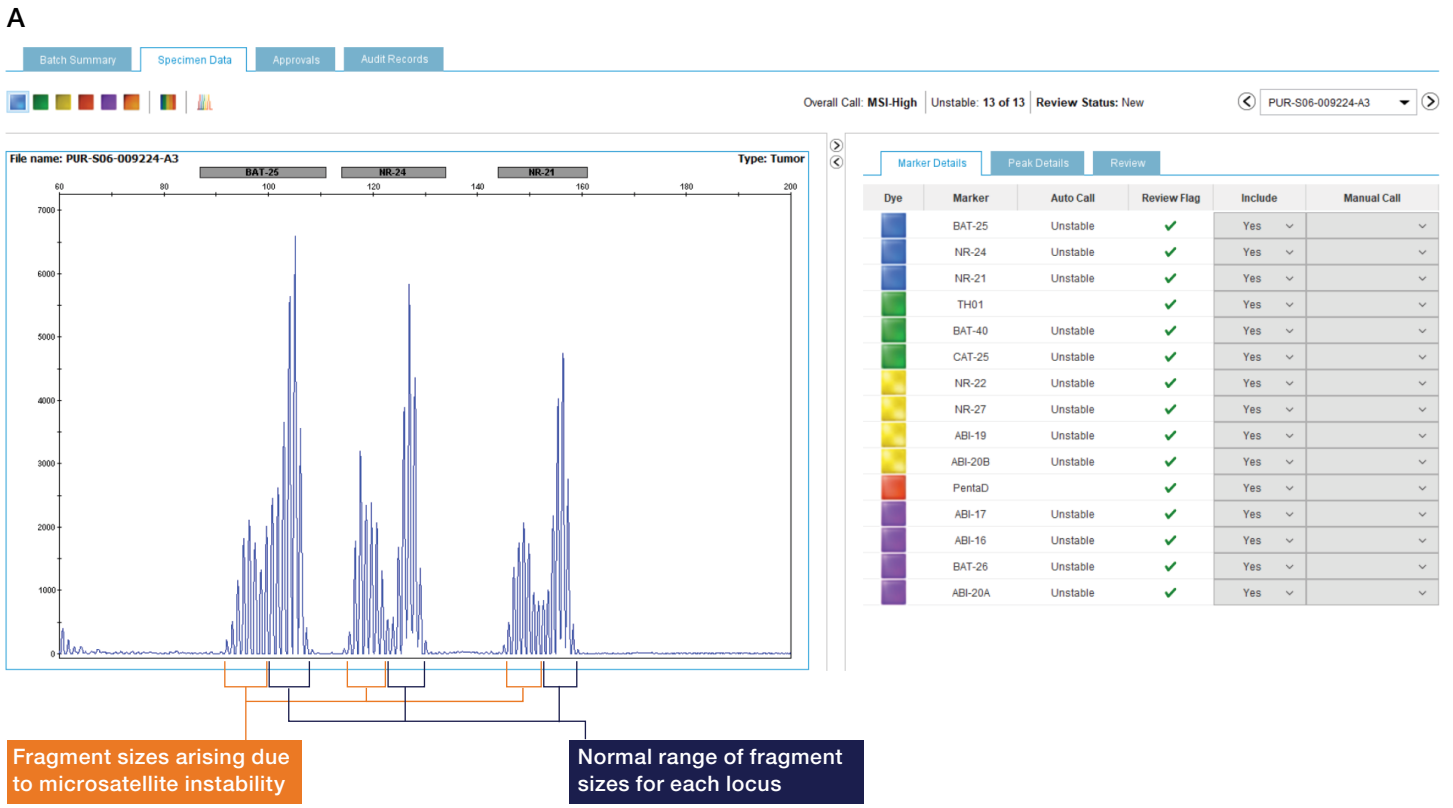
The Applied Biosystems™ TrueMark™ MSI analysis system analyzes a panel of 13 microsatellite loci, including the Bethesda set of markers [8], as a measure of microsatellite instability. TrueMark MSI assays also provide STR information at two highly variable loci, allowing for sample identity confirmation.

To ease the analysis of MSI raw data, we developed Applied Biosystems™ TrueMark™ MSI Analysis Software to simplify the calls at each locus (Figure 3A). Furthermore, we incorporated analysis algorithms that do not require side-by-side analysis of normal, non-tumor tissue, reducing the number of samples needed and the expense of an MSI analysis. For more information, see reference 9.

To illustrate the performance of the SeqStudio Flex instrument relative to other genetic analyzers, nine tumor/normal adjacent pairs and one tumor-only FFPE sample in blocks were obtained. Sections (10 µm) were cut from the FFPE blocks and gDNA was extracted from the sections, and 1–2 ng of genomic DNA was used in the TrueMark assay according to the published protocol. Fragments were analyzed on the SeqStudio Flex and 3500xL genetic analyzers, and results were interpreted using the TrueMark MSI Analysis Software (Figure 3A).

To compare the performance between genetic analyzers, we tallied the number of loci that could be called either stable, unstable, or no call (Figure 3B). The total number of loci that fell into each class was highly concordant. One sample (S07-001886-A5) had very little gDNA recovered, so using 2 ng of the recommended input amount was not possible. Nevertheless, the two instruments were able to make similar calls for the loci that were detectable.

Samples from the same FFPE blocks were used for standard immunostaining for MMR proteins, and the overall MSI determinations using results from the SeqStudio Flex and 3500xL systems were completely concordant with the immunostaining results. These results demonstrate that the SeqStudio Flex system generates MSI analysis data that are equivalent to those obtained with the 3500xL system.



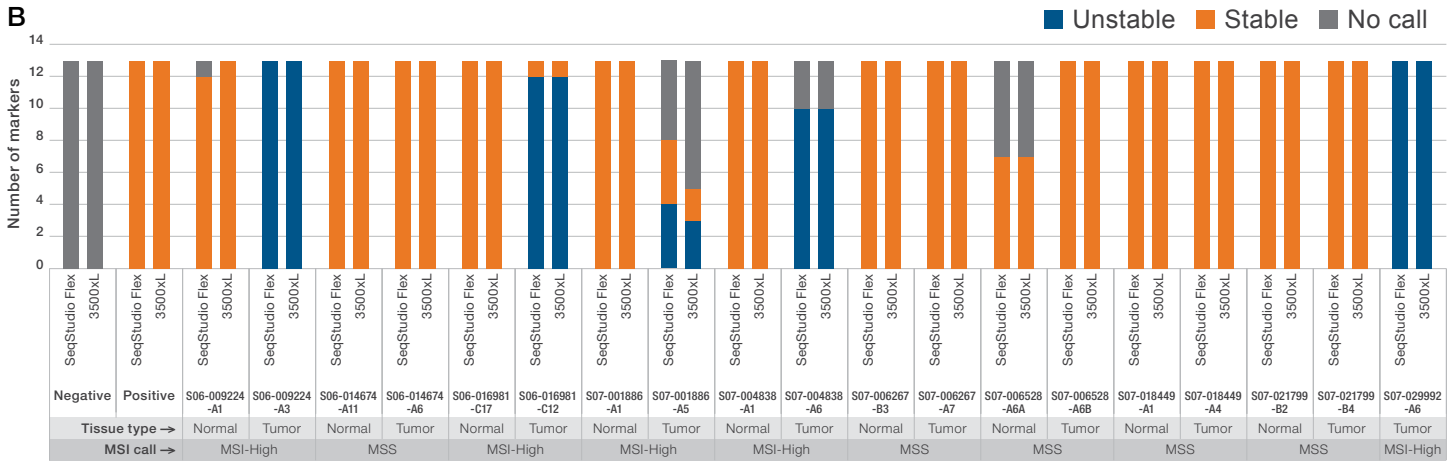


Figure 3. The SeqStudio Flex and 3500xL instruments produced similar data in MSI analysis. (A) The TrueMark MSI assay analyzes 13 microsatellite loci for instability, including the widely used Bethesda standards. Loci that are determined to be unstable can be autocalled; the software will then use the totality of the calls to make an overall call for the sample. The assay includes two highly variable short tandem repeat (STR) sequences (THO1 and PentaD) that can be used to confirm sample identity. The proprietary algorithms used by the software do not require side-by-side analysis of normal, non-tumor tissue in order to make stable/unstable calls. **(B)** Nine tumor/normal adjacent pairs and one tumor-only sample were analyzed using the TrueMark MSI assay. The number of loci called by the software was highly similar using data from both instruments. Sample S07-001886-A5 had suboptimal gDNA recovered; not all of the loci amplified equivalently and thus produced slightly different results on the two instruments.

dsDNA QC

Many protocols in genomic analyses start from a pool of DNA. Frequently the size and amount of DNA in that pool are confirmed prior to initiating further studies. For example, extracting DNA from FFPE-preserved tissue can result in fragmented DNA that may be too small to provide useful information. Thus, it is beneficial to know the overall size distribution before moving experiments forward.

In next-generation sequencing (NGS) experiments, libraries of pooled fragments are often generated before loading onto the sequence-reading chip. Knowing the size and abundance of DNA fragments in that library is important for predicting the success of the NGS result.

Finally, it is beneficial to confirm whether a specific amplicon was generated from a PCR reaction. Agarose gels are often useful for this, but while they are inexpensive, they are not extremely precise or high in throughput.

The size and relative abundance of double-stranded DNA fragments can be analyzed on Applied Biosystems genetic analyzers. For this application, Invitrogen™ TOTO™-1 DNA-intercalating fluorescent dye is incubated with a sample. The sample is subsequently run using Applied Biosystems™ POP-7™ polymer with the capillary heater turned off. This helps maintain the DNA in a mostly double-stranded, nondenatured state. By comparing the resulting migration of a test fragment with the migration of known dsDNA standards (e.g., phage lambda DNA digested with HindIII, phage phiX174 DNA digested with BsuRI), the size of the test fragment(s) can be determined.

To assess the performance of the SeqStudio Flex system relative to the 3500xL system for analyzing dsDNA, we obtained samples of extracted genomic DNA from cell lines, extracted genomic DNA from FFPE slides, a *BRAF* amplicon for cycle sequencing, or lambda HindIII and phiX174 BsuRI digested DNA. dsDNA samples were mixed with a 50 nM solution of TOTO-1 dye and single-stranded (ss) Applied Biosystems™ LIZ™ size standards in water. Samples were incubated at 37°C for 10 minutes and separated on the genetic analyzer calibrated with the E5 dye set and capillary heating off. Fragments were analyzed using GeneMapper 6.1 software.

When lambda HindIII standards are analyzed using this method, the well-known pattern of eight fragments, ranging from 23,130 bp to 125 bp, can be seen (Figure 4A). In the same injection but in a different capillary, we analyzed 50 ng of genomic DNA purified from a cell line. Comparing the migration of the gDNA to the lambda standard shows that the majority of the intact DNA ranges between approximately 9 kb and 20 kb.

Similarly, we ran a mixture of lambda HindIII and phiX174 BsuRI standards in another capillary. In addition to the eight lambda fragments, 10 more fragments that range from 1,353 to 72 bp can be seen (Figure 4B). DNA extracted from an FFPE slide runs from about less than 70 bp to about 200 bp, plus some large DNA fragments in the 4 kb–10 kb range.

Finally, we checked the quality of a *BRAF* amplicon before cycle sequencing (Figure 4C). According to the lambda and phiX174 standards, the amplicon migrates in the range of around 200–230 bp. The expected size of this amplicon is 208 bp, matching well with the observed size. Note that in each example, the profiles obtained on the SeqStudio Flex and 3500xL analyzers were very similar, demonstrating that the SeqStudio Flex and 3500xL systems generate equivalent dsDNA fragment data.

This method provides a path to fast and simple QC of dsDNA samples. There are several enhancements that would increase the information gained from these runs. For example, the ss LIZ internal standards can be recalibrated to the lambda/phiX marker set. Once they have been redefined, the LIZ standards can act as an internal size standard when run in the same capillary as the unknown. Thus, it may be possible to quantify the amount in a peak, provided a standard curve is run at the same time. Finally, we have found that this application also works on the 4-capillary SeqStudio instrument (data not shown).

Figure 4. Analysis of double-stranded DNA by CE. (A) Different DNA preparations were electrophoresed through capillaries containing POP-7 polymer at reduced temperature. Single-stranded LIZ-labeled fragments were included in the capillary. Lambda HindIII digest fragments and genomic DNA from A549 cells were analyzed in different capillaries. Note the typical pattern of lambda fragments (top panels). These can be used to estimate the size of the gDNA population of about 9–23 kb (bottom panels). Equivalent performance was seen on both instruments. **(B)** A mixture of lambda HindIII and phiX174 BsuRI digests and DNA extracted from an FFPE tumor sample were analyzed in different capillaries. Note the typical pattern of phage fragments (top panels); these can be used to estimate the size of the fragmented DNA (75–190 bp) and the large genomic DNA (bottom panels). Equivalent performance was seen on both instruments. **(C)** An amplicon containing the *BRAF* V600E region was analyzed before cycle sequencing, along with a mixture of lambda HindIII and phiX174 BsuRI digests in different capillaries. Note the typical pattern of phage fragments (top panels); these can be used to estimate the size of the *BRAF* amplicon (194–230 bp). This is very similar to the expected size of 208 bp. Equivalent results were seen on both instruments.

Multiplex PCR analysis

Amplification of DNA sequences by PCR remains a workhorse of almost all aspects of molecular biology and genomic research. As the depth of knowledge and scientific questions become more complex, there is an increasing need to analyze many targets in a single sample. However, singleplex PCR can be tedious, involving setting up individual reactions for each of the desired targets.

Methods that can analyze multiple PCR amplicons from a single sample would streamline workflows and conserve precious samples. One method to address this is fragment analysis by CE, using multiple fluorescent dyes for multiplex capabilities. Here, a very large number of targets can be analyzed from a single sample, since it can separate unique and specific amplicons based on both size and fluorescent dye in a single capillary. In fact, the CLA and MSI analysis kits described above are examples of multiplex PCR analysis of fragments.

We demonstrated how fragment analysis by multiplex PCR works using a custom panel for detecting respiratory viral pathogens. We designed a set of PCR primers that could detect 12 different respiratory RNA viruses. The panel was designed such that each organism generated a different sized amplicon. Oligos contained 6-FAM™ labeling at the 5' end, and the set of primers was pooled so the panel could be mixed with sample in a single step. The resulting amplicons could be separated and examined using any of the CE genetic analyzers.

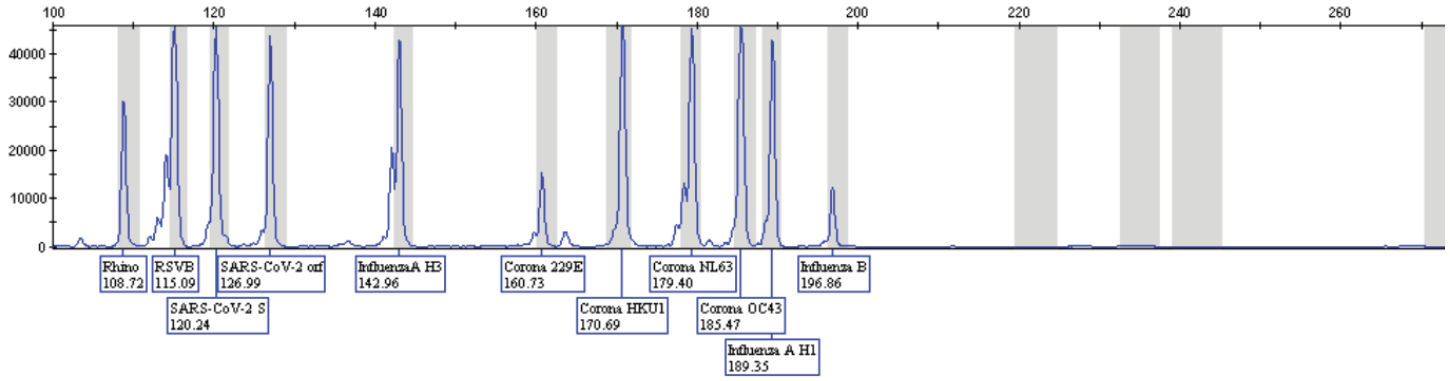
To illustrate the performance of the SeqStudio Flex instrument compared to the other genetic analyzers, we obtained Invitrogen™ GeneArt™ synthetic DNA targets or purified genomic RNA targets (ATCC) for common respiratory viruses. These targets, either single or mixed, were amplified using Applied Biosystems™ TaqMan® Multiplex Master Mix and the custom panel of pooled primers described above. Following PCR, amplicons were mixed with Applied Biosystems™ Hi-Di™ formamide and LIZ standards and separated on SeqStudio Flex and 3500xL genetic analyzers. The resulting data were analyzed using GeneMapper v6.1 software.

The functionality of the primers was confirmed by testing each primer pair individually against its cognate target (data not shown), and against the mixture of all DNA targets with the pooled set of primers (Figure 5A). Each of the primer pairs was able to recognize its target if present, in either the singleplex or multiplex assay. We confirmed that the approach could work with RNA genomes by first synthesizing cDNA from either single viral genomes or a mixture of all viral genomes, then PCR amplifying with either individual primers or the pooled primer set. Each of the organisms was detected if the appropriate target was in the sample (Figure 5B).

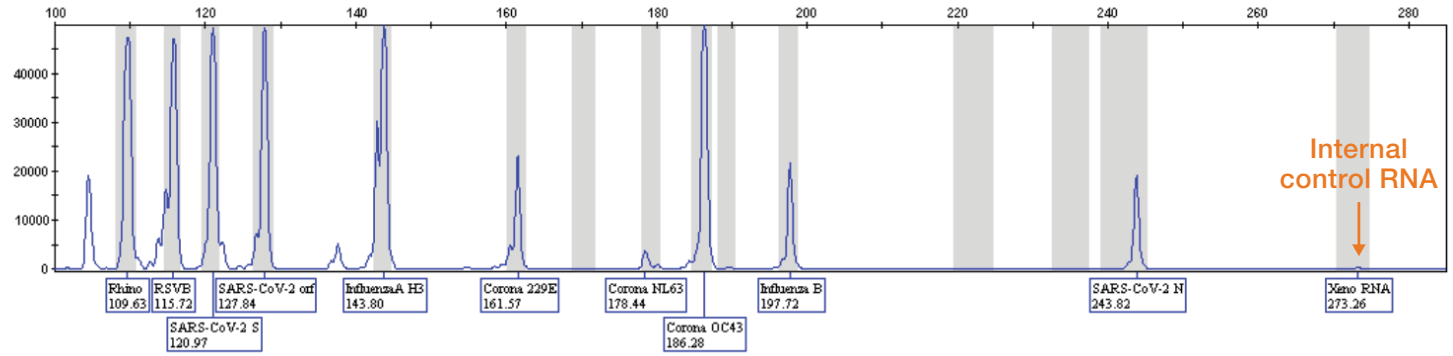
Finally, we compared the results obtained with the SeqStudio Flex, 3500xL, and SeqStudio genetic analyzers. Each of the instruments produced comparable results (Figure 5C). These results therefore confirm that the SeqStudio Flex system generates multiplex PCR data that is equivalent to that obtained with the other members of the CE family of instruments.

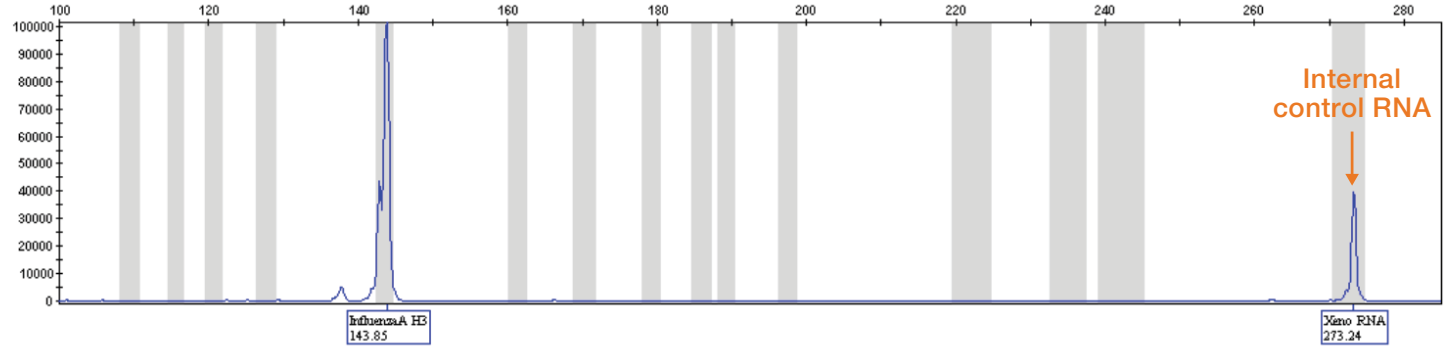
Multiplex PCR coupled with fragment analysis presents a tremendous opportunity to perform complex analyses with minimal effort. Designs of a panel are limited only by the ability to design primer sets against the targets of interest. It can therefore be adapted and utilized in nearly any scenario where detection of many target sequences in a single sample is desired.

A


Rhino 108.72
RSVB 115.09
SARS-CoV-2 orf 126.99
SARS-CoV-2 S 120.24
InfluenzaA H3 142.96
Corona 229E 160.73
Corona HKU1 170.69
Corona NL63 179.40
Corona OC43 185.47
Influenza A H1 189.35
Influenza B 196.86

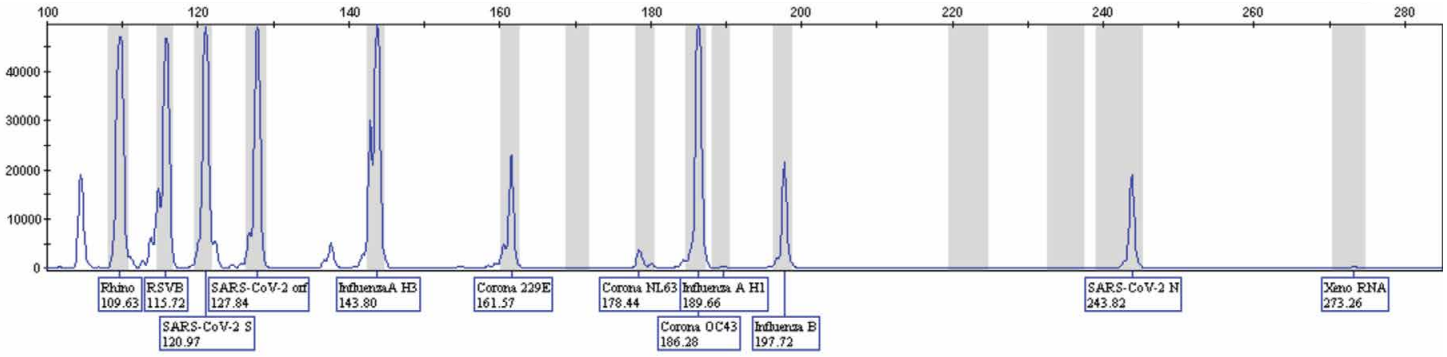
B


Rhino 109.63
RSVB 115.72
SARS-CoV-2 orf 127.84
SARS-CoV-2 S 120.97
InfluenzaA H3 143.80
Corona 229E 161.57
Corona NL63 178.44
Corona OC43 186.28
Influenza B 197.72
SARS-CoV-2 N 243.82
Xeno RNA 273.26
Internal control RNA

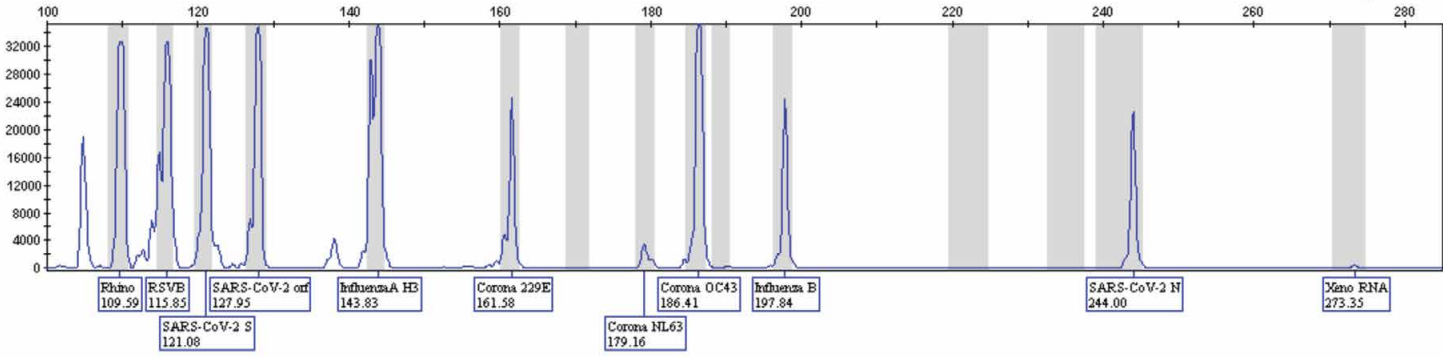

InfluenzaA H3 143.85
Xeno RNA 273.24
Internal control RNA

C

SeqStudio Flex



3500xL



SeqStudio

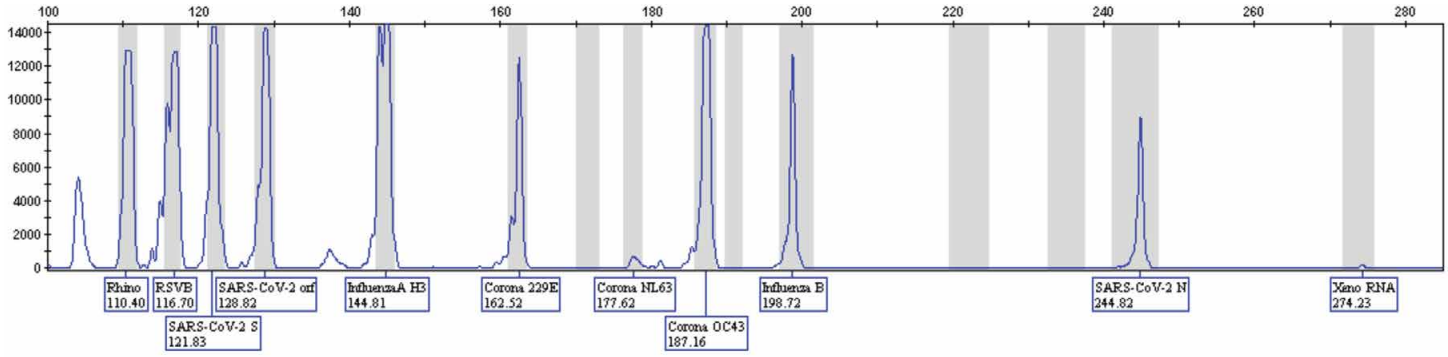


Figure 5. Analysis of multiple targets by multiplex PCR and fragment analysis. (A) DNA targets of the RNA viruses shown above were amplified in a single reaction using a panel of 6-FAM–labeled PCR primers. Each primer pair will produce a virus-specific amplicon size that can be separated and defined by fragment analysis. The grey bins were defined to detect the targets of this particular panel. **(B)** Purified RNA genomes of the viruses shown above were either pooled (top) or left single (bottom). cDNA synthesis followed by amplification using the pooled primers confirmed that multiple targets or a single target can be picked up in a single reaction. As an internal control for cDNA synthesis and amplification, an exogenous RNA (Applied Biosystems™ Xeno™ RNA) was spiked into the RNA pool and the primer pool contained primers for the exogenous RNA. **(C)** Equivalent ability to detect the multiplex amplicons was seen on all instruments; an example from the RNA genomic pool is shown here.

Sequencing applications

Sanger sequencing is the trusted standard for obtaining DNA sequence information. It powered the Human Genome Project and investigators continue to rely on this method to generate highly accurate, targeted sequencing results. Applied Biosystems™ products support fast and straightforward Sanger sequencing workflows that provide a high degree of accuracy, long-read capabilities, and simple data analysis. An entire workflow can be completed in less than one workday, from sample to answer, providing the flexibility to support a diverse range of applications in many research areas. We describe below several popular applications that commonly use Sanger sequencing methods.

Plasmid sequencing (long read length)

Modern genetic analyses often require extensive manipulation of DNA sequences. Pieces of DNA are cut, amplified, joined, propagated, and purified in tasks ranging from subcloning inserts into plasmids, through generating libraries, to constructing viral vectors for gene therapy.

Because many of these steps involve DNA polymerases, misincorporation of bases and genetic drift can lead to undesired mutations in the target sequences. Confirmation of manipulated DNA sequence is therefore necessary to ensure the intended sequences in the final product are correct.

Sequence confirmation is also advisable for any genetic engineering experiment, such as AAV viral vector construction or gene synthesis experiments. The ability to have long Sanger sequencing reads simplifies the chemistry and analysis workflow needs.

Applied Biosystems™ BigDye™ Terminator v1.1 and v3.1 cycle sequencing chemistries are the gold standard for Sanger sequencing by CE. After cycle sequencing, various options exist for clean-up before electrophoresis, including Centri-Sep™ purification columns and plates and Applied Biosystems™ ExoSAP-IT™ enzyme mix. Finally, the genetic analyzers have running modules that have been optimized for long reads.

To illustrate the performance of the SeqStudio Flex system against other genetic analyzers, we performed a standard sequencing experiment. Plasmid pGEM™-3Zf(+) was sequenced using both v1.1 and v3.1 BigDye Terminator chemistries in forward and reverse directions. Sequencing reactions for each direction were performed with 8 replicate reactions in each direction. Reactions were purified using Centri-Sep plates, dried, and resuspended in Hi-Di formamide.

The resulting reactions were run on the SeqStudio Flex, 3500xL, and SeqStudio genetic analyzers. Sequencing traces were analyzed using Applied Biosystems™ Sequence Scanner and Biomatters Geneious™ software. Using BigDye Terminator v1.1 chemistry, we were able to obtain over 1,000 bp of high-quality sequence on the SeqStudio Flex and 3500xL systems, and slightly less on the SeqStudio system (Figure 6A).

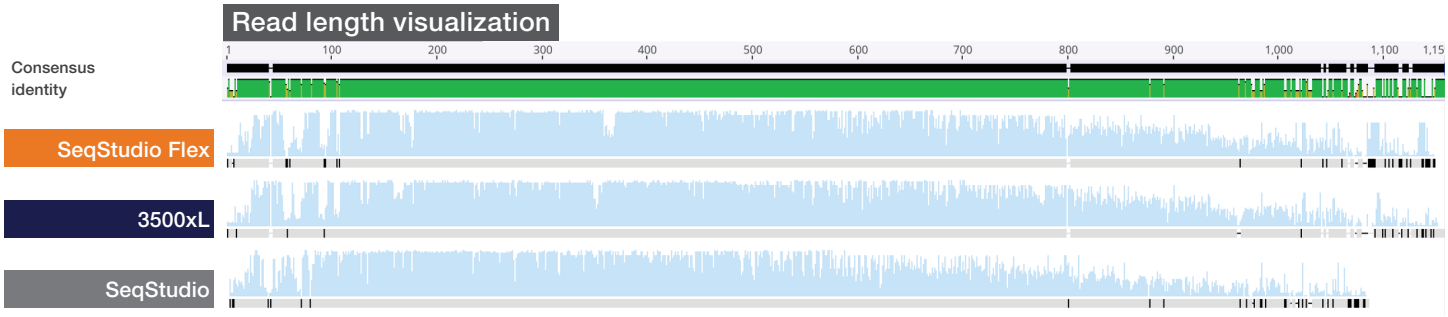
There are several quality control metrics that are generated by the Sequence Scanner Software. One measurement is the basecalling quality value (QV); it represents the shape and signal-to-noise ratio of a peak in the chromatogram. High-quality peaks typically have a QV greater than 20. The QV provides the most objective metric to evaluate the confidence of the base call and is used to calculate the trace score.

The trace score is the average of basecalling QVs for bases in the legible range of the chromatogram. The contiguous read length (CRL) is the longest read generated without poor-quality breaks in the chromatogram. The QV20+ value is the number of bases that had a quality value greater than 20 (high-quality reads) in the entire chromatogram. We used these metrics to compare the performance (Figure 6B).

The values for each of these QC metrics was nearly identical across the three instruments. Similarly, we sequenced the pGEM-3Zf(+) cloning vector using BigDye Terminator v3.1 chemistry, and obtained very similar read lengths on the SeqStudio Flex and 3500xL systems (Figure 7A). Furthermore, the trace score (Figure 7B), CRL, and QV20+ (Figure 7C) values were very similar.

Finally, we aligned the total read, which may include some regions that were homologous but not contiguous, to the pGEM-3Zf(+) reference sequence. We found an excellent degree of homology for the entire long read obtained on all instruments (Figure 7C). These results confirm that the SeqStudio Flex system generates high-quality, long reads that are equivalent to the quality obtained with the other CE instruments.

A



B

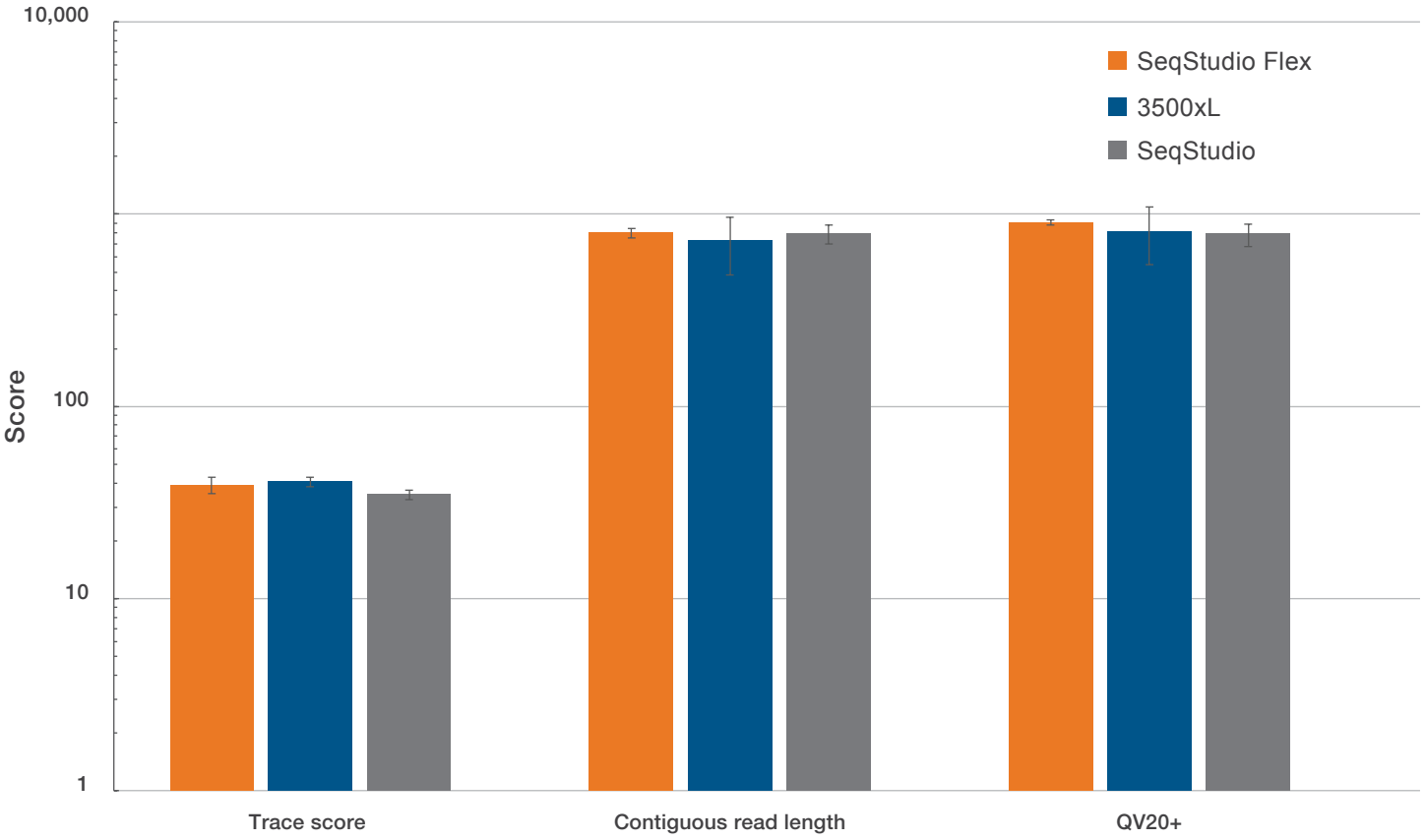


Figure 6. Analysis of plasmid sequencing results of long reads using BigDye Terminator v1.1 chemistry. (A) The read lengths are depicted in the Manhattan plot; the alignment with the consensus pGEM-3Zf(+) sequence is shown on the green bar on the top. The light blue bars indicate the quality of the read at that base; taller bars are better-quality reads. The SeqStudio Flex and 3500xL genetic analyzers produced almost identical results and were able to read more than 1,000 bp; the SeqStudio instrument had a slightly lower read length (~1,000 bp). **(B)** Quality metrics for reads were obtained on all three instruments. The sequence quality metrics produced on all three were nearly identical.

Figure 7. Analysis of plasmid sequencing results of long reads using BigDye Terminator v3.1 chemistry. (A) The read lengths are depicted in the Manhattan plot. The light blue bars indicate the quality of the read at each base; taller bars are better-quality reads. The SeqStudio Flex and 3500xL genetic analyzers produced almost identical results and were able to read more than 1,100 bp; the SeqStudio instrument had a slightly lower read length (~950 bp). Trace score **(B)**, contiguous read length, and QV20+ values **(C)** were obtained on all three instruments. In addition, the number of homologous bases in the reads was also obtained. The SeqStudio Flex and 3500xL systems gave similar results, and the SeqStudio system had slightly lower values (data not shown).

SARS-CoV-2 sequencing (medium read-length)

Some investigations require focused sequence information for a specific region of a genome. While genome-wide discovery tools such as NGS provide vast amounts of data over very large regions, such complex approaches are not always needed in focused-based research. Sanger sequencing provides a simple, inexpensive, and easy-to-interpret solution when focusing on a region.

The Applied Biosystems™ BigDye™ Terminator v1.1 and v3.1 chemistries are also widely used for medium read-length sequencing. Another option, Applied Biosystems™ BigDye™ Direct chemistry, makes use of universal M13 primers to further simplify the cycle sequencing workflow.

When coupled with clean-up using the Applied Biosystems™ BigDye XTerminator™ Purification Kit, the entire cycle sequencing workflow can be performed in a single well of a 96-well plate. The genetic analyzer family has sequencing modules that have been optimized for BigDye Direct and BigDye Terminator sequencing reactions.

To compare targeted sequencing results, we sequenced the region of the SARS-CoV-2 genome that encompasses the Spike gene and adjacent sequences. Overlapping M13-tailed sequencing primers (12 pairs) for the entirety of the SARS-CoV-2 Spike gene have been previously described [10]. One thousand copies of SARS-CoV-2 genomic RNA were reverse-transcribed using Invitrogen™ SuperScript™ IV VILO™ Master Mix. Amplicons (500–600 bp) were generated using 1 µL of cDNA and amplicon pairs, cycle sequenced using the BigDye Direct kit, and purified using the BigDye XTerminator kit. The resulting reactions were run on SeqStudio Flex, 3500xL, and SeqStudio genetic analyzers. Chromatograms were analyzed using Sequence Scanner and Geneious software.

The Spike gene region of SARS-CoV-2 is covered by the 12 overlapping amplicons spanning over 5 kb. High-quality Sanger sequencing results were obtained from nearly all 12 amplicons in both directions (Figure 8A). These results could be assembled into a single contiguous read (contig) that was confirmed to be the SARS-CoV-2 Spike gene by BLAST alignment. The read quality metrics were nearly identical on all three genetic analyzers (Figure 8B). Consistent basecalling quality, as demonstrated by the data obtained on the three instruments, facilitates using Sanger sequencing to assemble the sequence of focused regions using overlapping mid-length amplicons, generating contigs that cover a locus of interest.

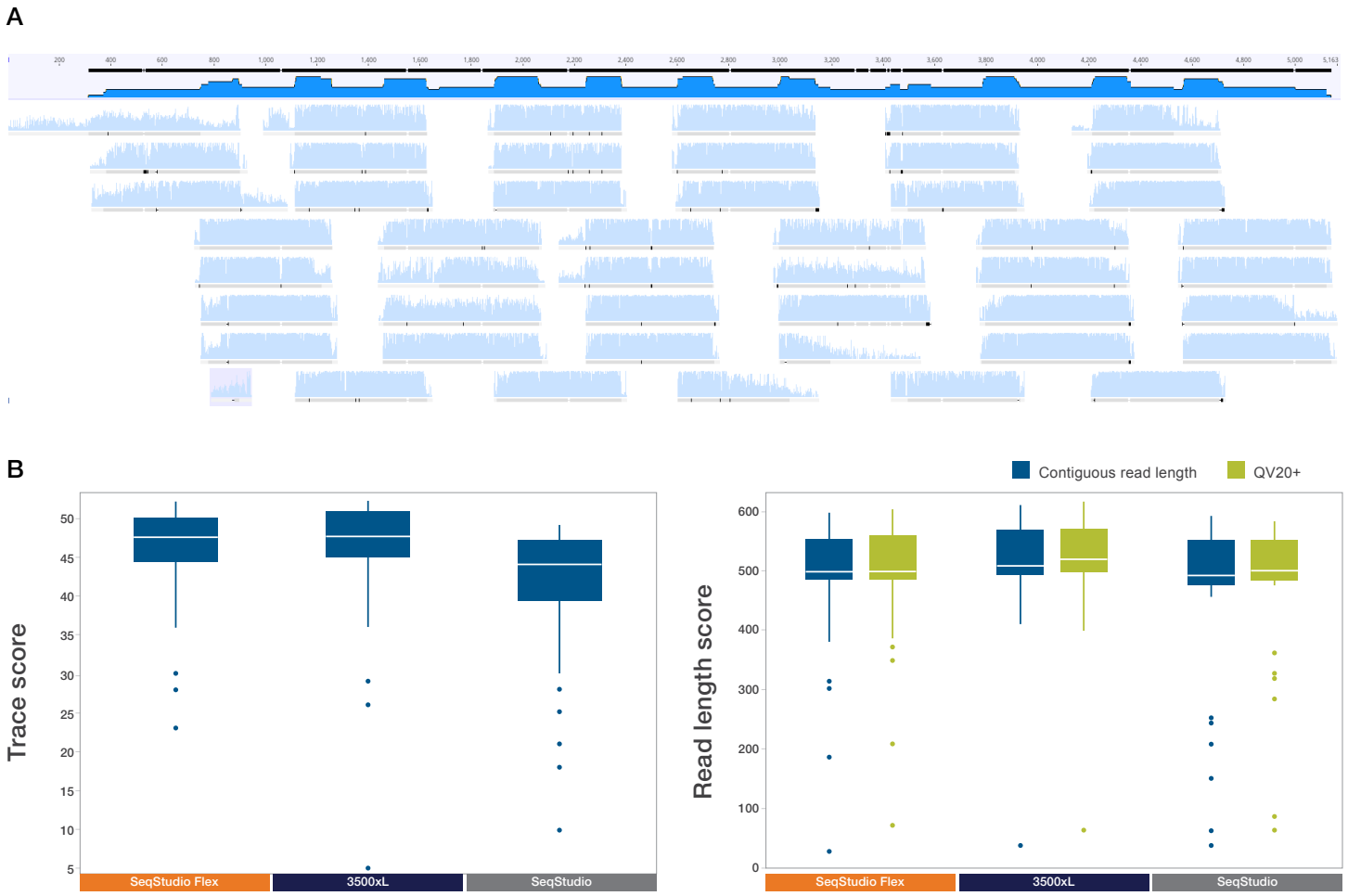


Figure 8. Sequencing the SARS-CoV-2 Spike gene by overlapping amplicons. (A) Twelve amplicons were used to sequence SARS-CoV-2 RNA in duplicate. Reads produced by the amplicons were mapped to each other to recapitulate the Spike gene sequence. The read quality in each amplicon is indicated by the height of the light blue bars. Data were collected on the SeqStudio Flex system; similar alignments were performed with data from the 3500xL and SeqStudio systems, with highly similar results. Note that some of the amplicons had some quality drop-off; this was also observed with those runs on the other instruments. **(B)** Trace score (left) and contiguous read length and QV20+ values (right) were compared on all three instruments. Sequencing quality results were almost identical.

Minor allelic variant sequencing and NGS confirmation

Discovery-based genomic research, such as next-generation sequencing (NGS), often uncovers novel or unexpected variants or other sequence anomalies. Investigators look for ways to verify these new discoveries using orthogonal methods. Sanger sequencing is the method of choice for confirming NGS results studies because of its simple workflow and unambiguous results.

For these confirmatory studies, short amplicons, usually covering only the region to be confirmed, need to be sequenced. Moreover, minor allelic variants, present in a heterogeneous sample, can be identified by Sanger sequencing and software processing from an electropherogram that contains mixed base calls.

For these types of NGS confirmation studies, BigDye Terminator or BigDye Direct cycle sequencing chemistries and the BigDye XTerminator purification system are frequently used. The genetic analyzer instruments have run modules optimized for short amplicons, so that confirmation data can be rapidly obtained. Finally, we have developed the specialized Applied Biosystems™ Minor Variant Finder software that can detect allelic variants that are present down to 5% in Sanger sequencing reactions.

To illustrate the performance of the SeqStudio Flex instrument for minor allele confirmation experiments, we focused on the *BRAF* gene. Standardized samples with *BRAF* V600E variants at 50%, 10%, 5%, and 2.5% minor allele frequencies were used as templates. Primers covering 172 nt surrounding the *BRAF* V600E mutation were used to generate sequencing amplicons from these samples in duplicate. The resulting amplicons were processed using BigDye Direct cycle sequencing and BigDye XTerminator purification chemistries. The purified reactions were run on SeqStudio Flex, 3500xL, and SeqStudio genetic analyzers, and analyzed using Sequence Scanner and Minor Variant Finder software.

We successfully detected the *BRAF* V600E variants at each of the concentrations tested (Figure 9A). In the results view of Minor Variant Finder, the frequency of the minor variant calculated by the software can be seen in each of the samples. In the chromatogram view of Minor Variant Finder, the forward and reverse traces for the control and test sample, as well as the software-processed traces, can be used to verify the software call of a minor allele.

Each of the genetic analyzers produced very similar results in the frequency of the minor allele calculated (Table 3), even below 5% minor allele frequency. We also compared the sequencing quality metrics from these reactions and found that the three instruments produced highly similar quality run data (Figure 9B). These results demonstrate that the SeqStudio Flex instrument can confirm the presence of minor allelic variants by Sanger sequencing with a performance equivalent to 3500xL and SeqStudio genetic analyzers.

Table 3. Quantification of minor allele determinations on the three genetic analyzer platforms. Note that although the measured frequency was higher than expected, it was equivalently high on all instruments, suggesting an anomaly of the sample rather than the instrument.

Sample	Expected	Measured: SeqStudio Flex instrument	Measured: 3500xL instrument	Measured: SeqStudio instrument
50% *BRAF* V600E	50	62.4	62.6	62.8
10% *BRAF* V600E	10	14.8	13.9	15.1
5% *BRAF* V600E	5	7.8	6.9	7.0
2.5% *BRAF* V600E	2.5	4.3*	4.8*	4.2*

* Although a 2.5% concentration was tested and detected in this experiment, the ability to call at this low frequency may vary in other samples and with other alleles.

Figure 9. Detecting minor alleles by short amplicon sequencing. (A) Minor Variant Finder software can analyze Sanger sequencing traces and detect sequence variants that are present at 5% of the total. Templates containing *BRAF* V600E variants at 50%, 10%, 5%, and 2.5% were analyzed on the SeqStudio Flex, 3500xL, and SeqStudio instruments. In each case, the software was able to detect the 5% variant reproducibly (yellow highlights, left). The Sanger traces of control, test, and software-processed samples can be viewed in the chromatogram view (right). The traces looked identical on the three platforms; the allele quantity determined by the software in each set is indicated with the yellow highlight. **(B)** Trace scores (left) and contiguous read length and QV20+ values (right) were compared on all three instruments. Sequencing quality results for these small amplicons were almost identical.

Sequencing of difficult templates

Some organisms and regions of the human genome are difficult to sequence due to a high GC content. These regions cause polymerases to stutter, stop, or disengage entirely from a sequencing template, precluding high-quality results. Furthermore, some regions of genomes contain a large amount of highly repetitive sequences.

NGS platforms can have difficulty aligning such sequences to a reference genome. Because of its robust chemistry and unambiguous results without the need of a reference genome, Sanger sequencing can overcome many of these difficulties.

To assist with the PCR amplification of GC-rich sequences in a genome, we developed a GC enhancer reagent. This add-in reagent helps DNA polymerases generate templates from GC-rich sequences that can be used for cycle sequencing. BigDye Terminator or BigDye Direct sequencing chemistries are subsequently used, followed by purification and capillary electrophoresis on a genetic analyzer. To analyze GC-rich or highly repetitive sequences, the genetic analyzers require no special run modules.

To determine the quality of sequence obtained from difficult templates, we analyzed the human *CEBPA* gene, a gene with regions on average 66% GC. Primers were chosen to amplify eight different regions in the human *CEBPA* gene. These regions were amplified using Applied Biosystems™ AmpliTaq Gold™ polymerase and the GC enhancer reagent. Amplicons were sequenced using BigDye Direct cycle sequencing and BigDye XTerminator purification chemistries. The resulting reactions were run on SeqStudio Flex, 3500xL, and SeqStudio genetic analyzers.

All of the instruments were able to produce high-quality results from these GC-rich amplicons. Comparing the electropherograms obtained on SeqStudio Flex and 3500xL instruments revealed nearly identical traces (Figure 10A). Although only one representative sequencing trace is shown, similar results were seen with other amplicons and in both sequencing directions.

In support of these results, the QC analysis revealed nearly identical values for each of the metrics (Figure 10B). Together, these results show that even with difficult templates, the SeqStudio Flex instrument produces data equivalent to that generated on other Applied Biosystems genetic analyzers.

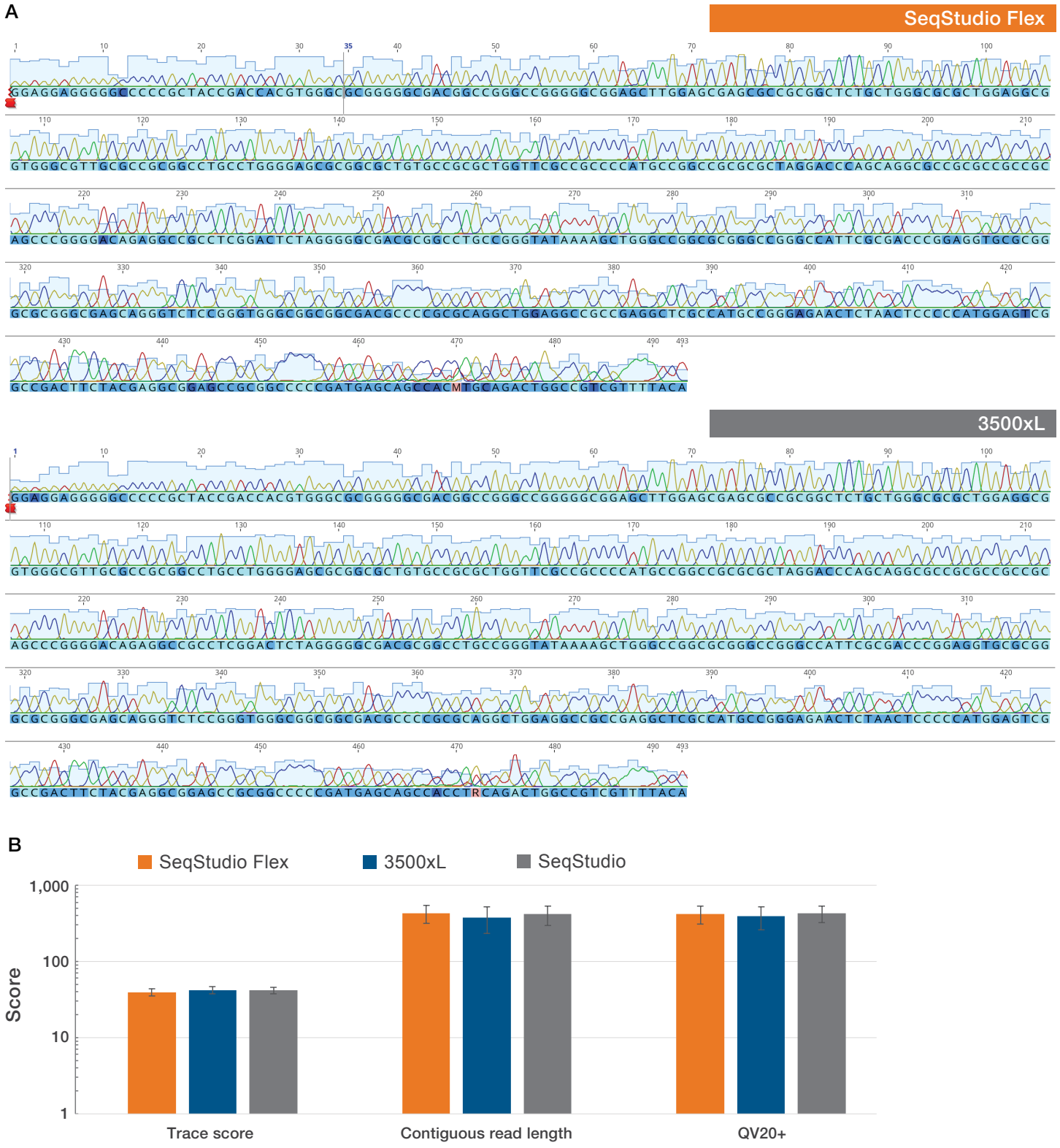


Figure 10. Sequencing through difficult regions. Portions of the human *CEBPA* gene, which averages 66% GC content, were amplified with the GC enhancer reagent before sequencing. **(A)** Chromatograms from one amplicon, obtained on the SeqStudio Flex and 3500xL genetic analyzers, are shown. This particular region is 77.8% GC. Note that the chromatograms are nearly identical. **(B)** Trace scores, contiguous read length, and QV20+ values were compared for all amplicons on all three instruments. Sequencing quality results for these difficult amplicons were almost identical.

Sanger sequencing for genome edit efficiency analysis

Genome editing has revolutionized the way biological research is performed. Because of its simplicity and efficiency, an investigator can make any type of a change to a genome, from engineering a specific SNP, to knocking out the function of a gene, to making large-scale genomic rearrangements.

Most researchers describe a continuous genome editing process that starts with design of a guide RNA and ends with a stable cell line containing the desired edit. However, throughout the process, there are steps where knowing the efficiency of the modification is necessary. Often, this involves monitoring a cell culture for the presence and frequency of a desired edit. In many cases, the efficiency challenge is linked to the efficiency of gRNA and CRISPR-related enzymes. Researchers therefore need unambiguous and straightforward methods for analyzing the efficiency of their genome editing reactions.

Analysis of the steps in a genome editing experiment often make use of cycle sequencing using BigDye Terminator or BigDye Direct chemistries, coupled with purification and electrophoresis on genetic analyzers. To facilitate the analysis of genome editing experiments, we developed the Applied Biosystems™ SeqScreener Gene Edit Confirmation App. This groundbreaking software is used to analyze the efficiency of edits and visualize subcultures that contain the optimal edit.

To characterize the genome editing analysis on the SeqStudio Flex instrument, we generated 96 CRISPR-directed edits in the human presenilin gene (*PSEN1*). For this experiment, we used different guide RNAs, transfection methods, and homology-directed repair templates, resulting in a diversity of editing efficiencies. gDNA from transfected cultures was isolated and sequenced using *PSEN1* primers, BigDye Direct cycle sequencing, BigDye XTerminator purification, and electrophoresis on SeqStudio Flex, 3500xL, and SeqStudio instruments. The resulting sequencing traces were analyzed using Sequence Scanner and SeqScreener Gene Edit Confirmation App software.

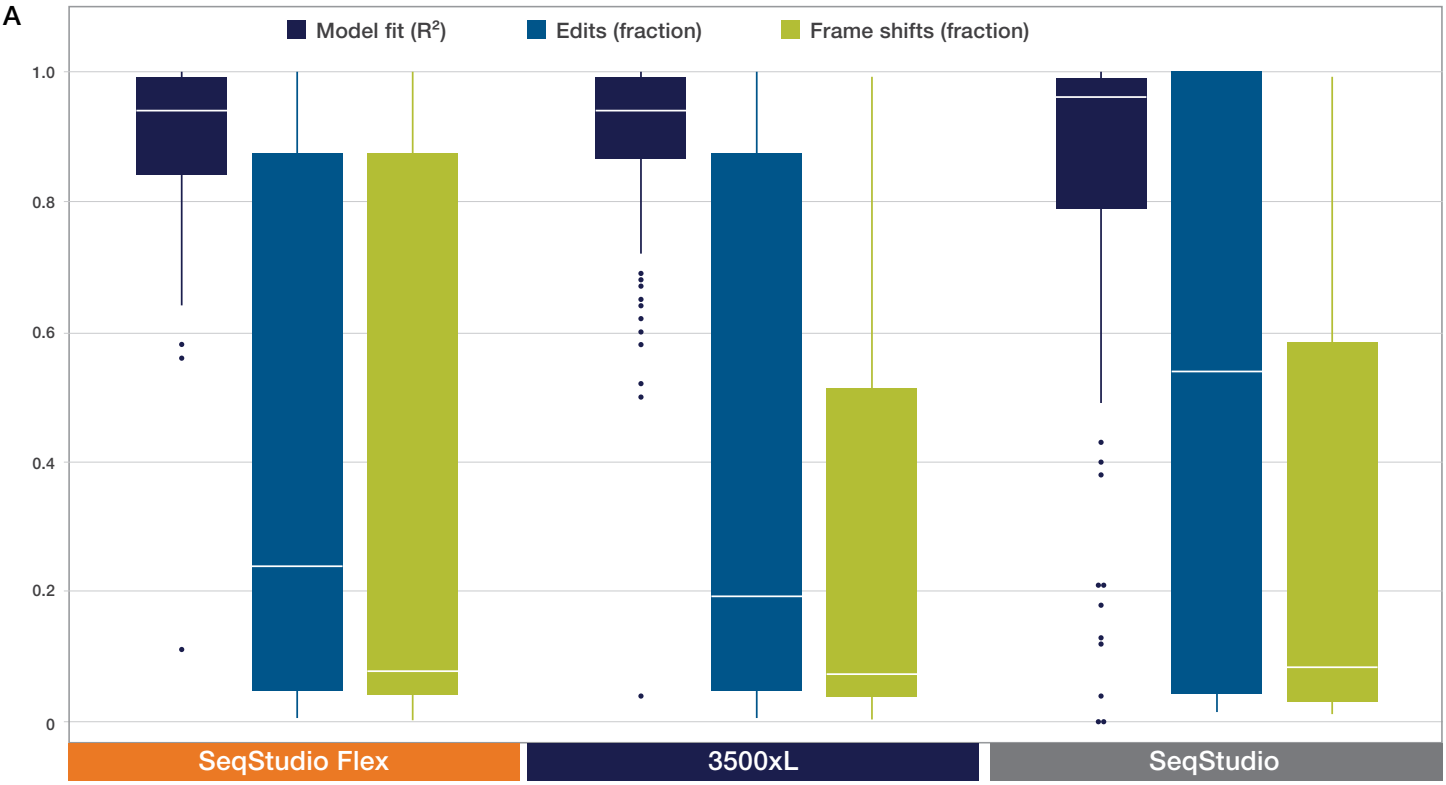
The SeqScreener Gene Edit Confirmation software computes and displays results in several different ways. One way is by generating numerical metrics for the results (Figure 11A). The Model Fit (R^2) is the overall confidence in the results. Typically, good-quality results have R^2 values of 0.8 or greater; however, the threshold can vary for each sample. The Edits value is the cumulative fraction of all successful edits in the measured sample. The wide range of values indicates the wide range of efficiencies across all the different editing conditions tested.

The Frame Shift value is the fraction of gene edits that are predicted to eliminate the expression of the encoded protein. This information is useful when screening for loss-of-function gene knockouts. The three genetic analyzers produced very similar genome editing efficiency results.

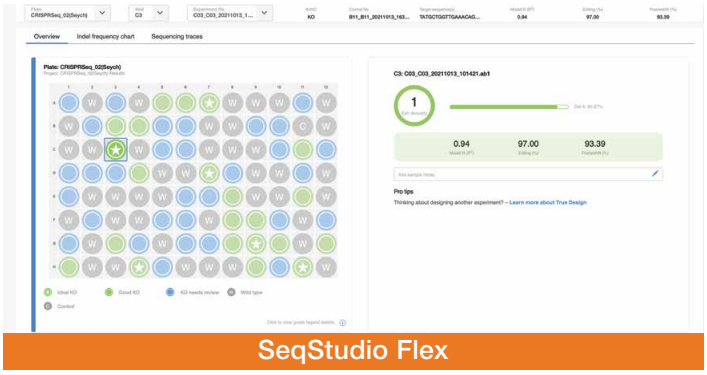
Another way SeqScreener software displays results is in an infographic showing the status of analysis in wells of a 96-well plate, modeling the results contained within the test plate (Figure 11B). In this representation, green wells contain successful edits, blue wells indicate visual inspection of sequencing traces is recommended, and the grey wells indicate wild-type (W), control (C), or problem (!). Starred wells indicate the wells containing the cells with the highest-efficiency, idealized edit. Clicking on a well brings up more information about the success in that well.

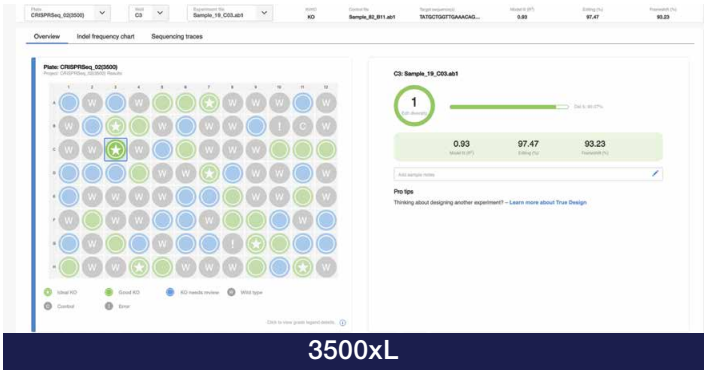
The plate infographic produced by the three instruments was very similar. There were a few differences, mostly in the visual inspection vs. problem categories in the SeqStudio system compared to the SeqStudio Flex and 3500xL systems. This was likely due to the degradation in sequence read quality, since the plate was used multiple times. Nevertheless, these results demonstrate that the performance of the SeqStudio Flex Genetic Analyzer is equivalent to the 3500xL and SeqStudio systems for monitoring genome editing efficiencies.

A



B





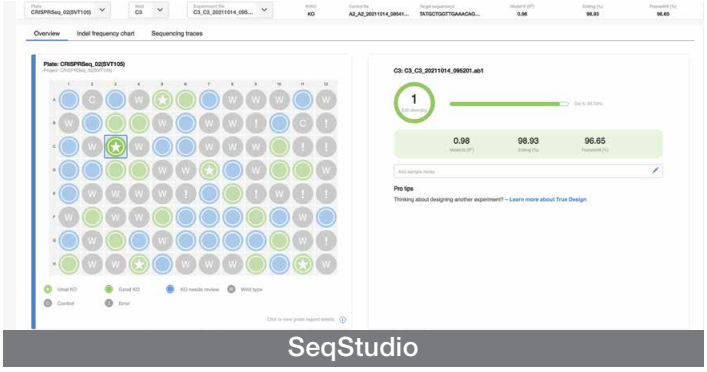


Figure 11. Sanger sequencing for genome editing analysis. **(A)** 96 samples from a *PSEN1* editing experiment were analyzed by Sanger sequencing and SeqScreener Gene Edit Confirmation App software. The software calculates the efficiency of the edits in each well and produces three metrics: model fit R^2 value, the fraction of DNA in a well that has an edit, and the fraction of edits in a well that are frame shifts. Sequencing data produced by all three instruments produced identical editing efficiency metrics. **(B)** SeqScreener Gene Edit Confirmation App software also generates plate maps with editing efficiencies shown for each well. Green wells contain successful edits, blue wells indicate visual inspection of sequencing traces is recommended, and the grey wells indicate wild-type (W), control (C), or problem (!). Starred wells indicate the wells containing the cells with the highest-efficiency, idealized edit. Clicking on a well will bring up more information (to the right of the plate map) about the success in that well. The plate infographics produced by the three instruments were very similar; most of the differences related to the need for visual inspection rather than outright discrepancies.

Conclusions

Capillary electrophoresis is a powerful method for genetic analysis. Whether it is fragment analysis or Sanger sequencing, the flexibility of CE can open doors to many types of applications. The SeqStudio Flex system, a start-of-the-art CE-based genetic analyzer, facilitates contributions to new discoveries and insights. Notably, the SeqStudio Flex system generates data quality that is equivalent to that of other genetic analyzer instruments, making it easy to transition research to the new platform. To learn more about the features that make the new SeqStudio Flex system the most advanced Applied Biosystems genetic analyzer, visit **thermofisher.com/seqstudioflex** or contact your local Thermo Fisher partner.

References

1. **https://iclac.org/databases/cross-contaminations/**
2. Horbach SPJM, Halffman W (2017) The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. *PLoS ONE* 12(10): e0186281. **https://doi.org/10.1371/journal.pone.0186281**
3. Authenticating human cell lines using CLA IdentiFiler and CLA GlobalFiler kits on capillary electrophoresis platforms. **http://assets.thermofisher.com/TFS-Assets/GSD/Application-Notes/authenticating-human-cell-lines-str-kits-capillary-electrophoresis-application-note.pdf**
4. ATCC STR database. **https://www.atcc.org/en/search-str-database**
5. Gologan A, Sepulveda AR (2005) Microsatellite instability and DNA mismatch repair deficiency testing in hereditary and sporadic gastrointestinal cancers. *Clin Lab Med* 25(1):179–196. doi: 10.1016/j.cll.2004.12.001. PMID: 15749237.
6. Zhao P, Li L, Jiang X et al. (2019) Mismatch repair deficiency/microsatellite instability-high as a predictor for anti-PD-1/PD-L1 immunotherapy efficacy. *J Hematol Oncol* 12:54. https://doi.org/10.1186/s13045-019-0738-1
7. Reyes GX et al. (2015) New insights into the mechanism of DNA mismatch repair. *Chromosoma* 124:443–462.
8. Boland CR et al. (1998) A National Cancer Institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* 58:5248–5257.
9. TrueMark MSI Assay For Microsatellite Instability Testing User Guide. **https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2FMAN0018868_MSI%20Assay_UG.pdf**
10. Protocol for Sanger sequencing of the SARS-CoV-2 spike (S) gene. **https://assets.thermofisher.com/TFS-Assets/GSD/brochures/sequencing-sars-cov-2-spike-gene-protocol.pdf**

Ordering information

Product	Quantity	Cat. No.
Instruments		
SeqStudio 8 Flex System Includes: SeqStudio 8 Flex Genetic Analyzer with data collection software, 1-year warranty, 1-day onsite SmartStart orientation, and DNA sequencing and fragment analysis reagent kits for system performance check		A53627
SeqStudio 8 Flex Genetic Analyzer System plus 1-year extended warranty Includes all items from A53627 plus additional 1-year warranty		A53789
SeqStudio 24 Flex System Includes: SeqStudio 24 Flex Genetic Analyzer with data collection software, 1-year warranty, 1-day onsite SmartStart orientation, and DNA sequencing and fragment analysis reagent kits for system performance check		A53630
SeqStudio 24 Flex Genetic Analyzer System plus 1-year extended warranty Includes all items from A53630 plus additional 1-year warranty		A53792
Reagents		
CLA Identifiler Plus PCR Amplification Kit	200 reactions	A44660
CLA Identifiler Direct PCR Amplification Kit	200 reactions	A44661
CLA GlobalFiler PCR Amplification Kit	200 reactions	A44662
TrueMark MSI Assay	100 reactions	A45295
BigDye Terminator v1.1 Cycle Sequencing Kit	100 reactions	4337450
BigDye Terminator v3.1 Cycle Sequencing Kit	100 reactions	4337455
BigDye Direct Cycle Sequencing Kit	100 reactions	4458687
BigDye XTerminator Purification Kit	100 preps	4376486
ExoSAP-IT PCR Product Cleanup Reagent	100 reactions	78200.200.UL
RecoverAll Total Nucleic Acid Isolation Kit for FFPE	40 reactions	AM1975

Learn more at **thermofisher.com/seqstudioflex**