APPLICATION NOTE

Applied Biosystems 3500/3500xL genetic analyzers

Using Sanger capillary electrophoresis sequencing to confirm variants discovered by next-generation sequencing (NGS)

Key findings

- For critical studies, sequencing data obtained from next-generation sequencing systems should be confirmed
- Sanger sequencing by capillary electrophoresis (CE) is an ideal orthogonal technology for verification of NGS base calls
- We have developed an easy-to-use, integrated workflow for NGS-to-CE confirmation

Next-generation sequencing analyses have revolutionized our understanding of biological processes. In many basic science or clinical studies, substantive insights have been made by comparing the primary DNA sequence of genes in different groups of subjects. In such studies, researchers attempt to identify the role that variations of nucleotide sequence between the groups may play in disease susceptibility, disease progression, or phenotypic variation. The accurate identification of sequence variants is therefore instrumental in ensuring successful experimental outcomes. However, the chemistries used in NGS technologies are prone to low, but detectable, levels of sequencing errors that may lead to confusing or incorrect interpretations. For example, a recent study found that up to 2% of the variants detected by NGS were not reproducible and required additional confirmation by Sanger sequencing [1]. Therefore, before any firm conclusions are drawn from an NGS study, potential variants identified by NGS should be confirmed by an orthogonal method.

This application note presents information to help researchers using next-generation sequencing–based data to confirm their results. We show why confirmation is an important part of any analysis of NGS-based data. In addition, we show how the reagents and systems available under the Applied Biosystems™ brand portfolio facilitate the orthogonal verification of NGS results.

ThermoFisher SCIENTIFIC

Why is orthogonal verification of NGS-based data important?

System-dependent sequencing biases

Confirming NGS results requires more than just repeating the sequencing on the same platform. Repeat sequencing will not eliminate all errors, because there could be sequencing biases that are specific to the platform being used. For example, GC-rich regions can be particularly hard to read through. They therefore can introduce misincorporation errors, the specific type of which can be influenced by the sequencing system used [2]. Similarly, errors can be introduced when sequencing through homopolymeric sequences [2]. Further, the library preparation method used can also potentially introduce systematic biases (e.g., strand bias) [3]. Lastly, different .bam file alignment and variant caller software can yield different (erroneous) results even for "relatively simple" single nucleotide variants (SNVs) [4,5]. These different types of errors have different frequencies and modalities that vary from system to system. Thus, they are unlikely to be resolved by simply resequencing on the same platform. Verification of any sequence variant therefore depends on reanalysis using a different platform that does not have the same systemic biases.

Uneven coverage may produce uneven accuracy

Another source of uncertainty can arise from uneven coverage of NGS sequencing reads. Uneven coverage issues arise in NGS analyses because not all sequences can be sequenced with the same efficiency. Sequences that are more difficult for the polymerase to read through due to secondary structure, highly repetitive sequences, or other factors, will be underrepresented in the subsequent data files. In addition, some NGS targeted sequencing approaches depend on capture by hybridization before library preparation [6]. This hybridization, too, can be highly sequence-dependent and can affect the representation of some sequences in the final libraries. And because confidence in the results is highly dependent on the number of instances of sequences detected, any regions that are underrepresented in the final data output could be more suspect, reducing confidence that variants detected are real.

Sanger-based capillary electrophoresis sequencing solutions

One of the workhorses of the genomic community for the last quarter century has been Sanger-based sequencing: polymerase termination with fluorescent dideoxynucleotides followed by sequence collection on automated capillary electrophoresis (CE) instruments. This is a robust and inexpensive method with chemistries and analysis tools that are easy to use and well understood—and because of its high accuracy and ease of data interpretation, it is considered to be the gold standard technology for DNA sequence analysis. It is therefore an ideal system for confirming variant calls made on NGS platforms. We have been at the forefront of improving CE sequencing technologies, and offer a complete solution for CE sequencing needs. From finding and resourcing predesigned amplification primers and PCR reagents, through BigDye™ chain-termination sequencing reagent mixes, capillary electrophoresis, and finally, data analysis software, Applied Biosystems™ products cover the entire CE workflow (Figure 1). In the sections below, we provide detailed information on the reagent choices, CE systems, and workflows that have been optimized for confirming NGS data by Sanger CE sequencing.

Figure 1. NGS-to-CE workflow. A complete workflow for verifying variants discovered by NGS systems is shown. Applied Biosystems products have been designed to be optimized to work together. NGS variants that are marked for verification can be directly imported into the Primer Designer Tool for picking and ordering appropriate PCR primer pairs that can be directly sequenced with the BigDye™ Direct sequencing kit and the 3500 Genetic Analyzer. The resulting data (.vcf file from NGS and .ab1 trace files from CE) are then aligned and compared in the Next-Generation Confirmation (NGC) tool in Thermo Fisher Cloud.

The Primer Designer Tool: the right source for predesigned PCR primer pairs for CE sequencing

The Primer Designer™ Tool is a web-based application (**thermofisher.com/primerdesigner**) that greatly facilitates the search and selection of optimal PCR primer pairs that help meet your needs for resequencing any exon of the human exome or sections of the mitochondrial genome (Figure 2).

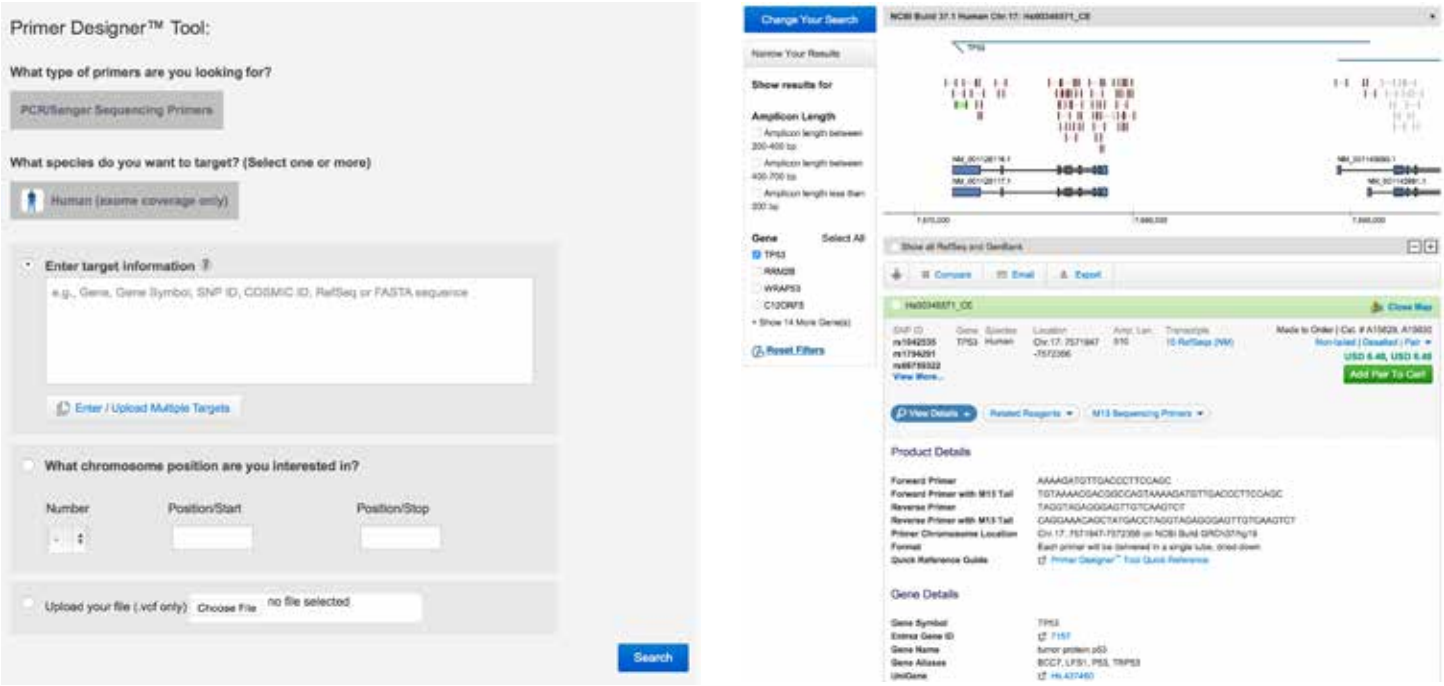

Figure 2. The Primer Designer Tool simplifies choosing sequences for CE sequencing. The user interface clearly indicates where sequence information is to be entered (left). The software will then present primers optimal for PCR amplification followed by cycle sequencing. Details about primer position on a locus map, amplicon length, and transcripts recognized can be visualized once the locus is chosen (right)

The Primer Designer Tool comprises a virtual design collection of over 650,000 primer pairs yielding amplicons ranging from 125 to 600 bp that cover 95.9% of human coding exons and 74.9% of human noncoding exons, as well as the human mitochondrial genome. Complete sequence coverage is achieved for 94% of coding exons. Sophisticated bioinformatics rules and filters were applied to help ensure high target specificity, and minimize the formation of primer dimers. The tool also contains the primer pairs used in the Ion AmpliSeq™ exome panel (n = 283,961 amplicons ranging from 200 to 400 bp) and the Ion AmpliSeq™ Cancer Hotspot Panel v2 (n = 207 amplicons with amplicon lengths shorter than 200 bp). The use of these Ion AmpliSeq™ primer pairs for resequencing individual targets by Sanger sequencing is described in detail in a different application note [7].

The Primer Designer Tool offers the choice of designing primers with or without M13 universal sequencing primers. However, having M13-tailed primers facilitates the sequencing workflow and we therefore recommend incorporating them into the primer designs.

To make it easier to validate Ion Torrent™ NGS data with CE, we provide direct links from the Torrent Suite™ Variant Caller software into the Primer Designer Tool (Figure 3). This feature allows users to enter sequences from the Torrent Suite Variant Caller data page directly into the Primer Designer Tool, where they can be used to design and order primers. Alternatively, the Primer Designer Tool accepts other common sequence file formats, such as .vcf files exported from NGS platforms, or direct entry of genomic coordinates flanking the region of interest.

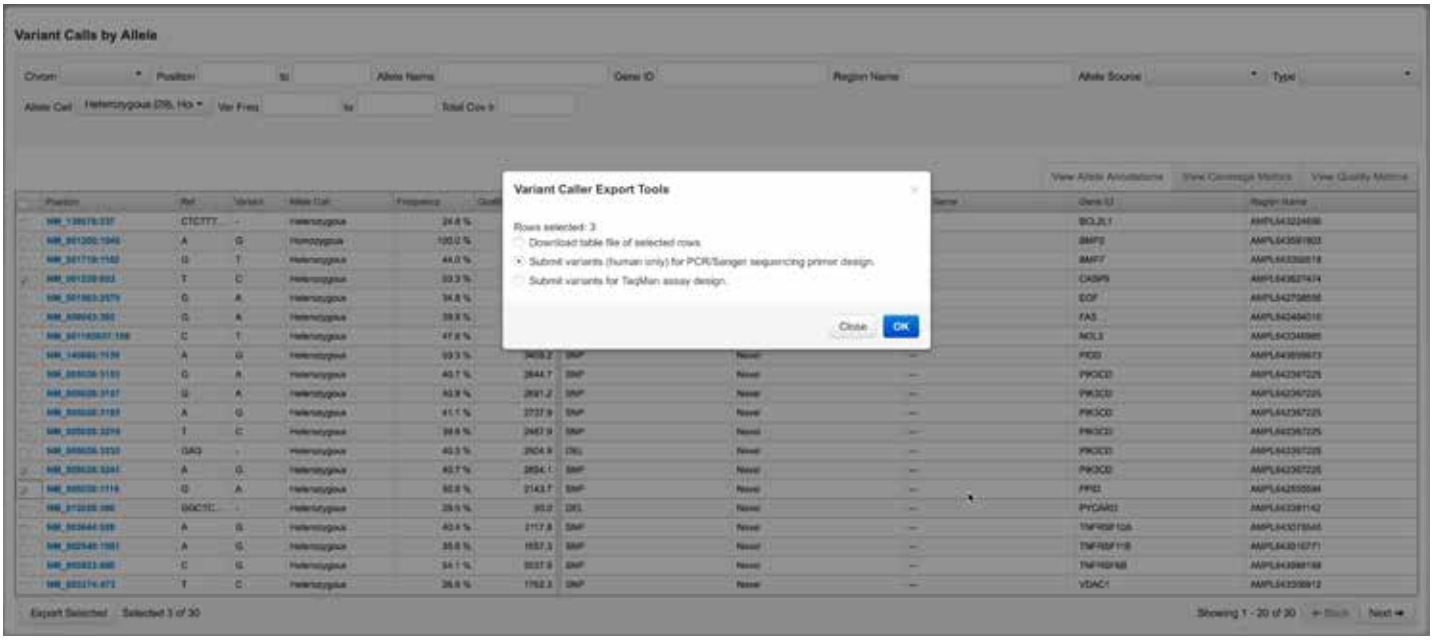

Figure 3. Output from Torrent Suite Variant Caller software can be entered into the Primer Designer Tool. Variants of interest detected by semiconductor sequencing can be selected by check mark and exported directly into the Primer Designer Tool. This facilitates the ordering of primers needed to confirm the Ion Torrent sequencing results by CE sequencing.

BigDye Direct Cycle Sequencing Kit simplifies workflow

The primers designed using the Primer Designer Tool can be used to generate amplicons for sequencing by ordinary PCR from the starting template. However, to facilitate generating and sequencing these amplicons, we developed the Applied Biosystems™ BigDye™ Direct Cycle Sequencing Kit (Cat. No. 4458688), a single-tube solution that is easy to use and significantly reduces hands-on time (up to 40%) at the bench. This kit provides the PCR reagent to amplify the target of interest, the sequencing reagent, and special M13-sequencing forward and reverse primers to sequence both strands. The only reagents that need to be supplied by the user are the DNA template (typically 5–10 ng) and the PCR primer pair for the amplicon to be sequenced. Note that the PCR forward and reverse target-specific primers need to be tagged with respective M13 sequences to allow the sequencing primers to anneal in the sequencing reaction. A significant advantage of the BigDye Direct workflow is that the whole workflow can be completed in a single tube (per sample)—that is, there is no need for a sample transfer into a different tube or plate. A single-tube solution that eliminates reagent transfers therefore eliminates a critical source of error. This also minimizes experimental variability caused by multiple handling steps, and in the end helps save researchers up to 3 hours of processing time (Figure 4). To use the kit, template is mixed with the amplification primers and reagents, and amplified by PCR. Next, the sequencing primers and reagents are added to the same tube and PCR cycled for sequencing. Finally, unincorporated dyes from the sequencing reaction are removed using the Applied Biosystems™ BigDye™ XTerminator™ Purification Kit (Cat. No. 4376484) without the need to transfer the sequencing products to a new tube. Alternatively, the reaction is passed through a spin column. The cleaned-up, purified sequencing fragments are now ready to be applied to the analyzer.

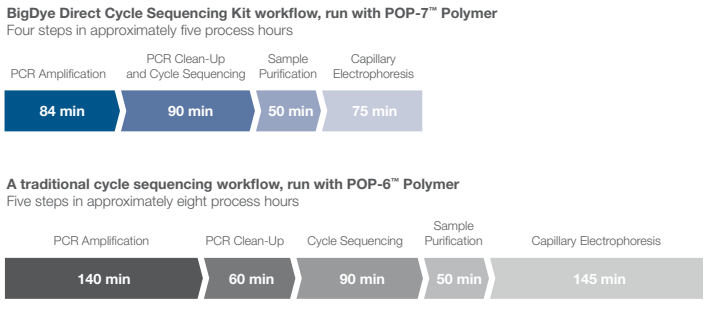

Figure 4. BigDye Direct Cycle Sequencing Kit increases speed and accuracy of cycle sequencing. Combining PCR clean-up and cycle sequencing into a single step reduces the traditional CE sequencing workflow by 40% and reduces the potential for errors.

The 3500/3500xL family of genetic analyzers

For reading the sequence information, we developed the easy-to-use 3500 family of capillary electrophoresis genetic analyzers. Depending on throughput needed, the analyzers are available in 8-capillary (3500) and 24-capillary (3500xL) formats. The system features an advanced long-life solid-state laser, operates using standard power outlets, and has a small footprint, facilitating easy setup and operation. Hands-on time is reduced by the availability of ready-to-use, load-and-run consumables. The consumables are preformulated and packaged for single use, so the possibility of mixing and handling errors is minimized. The consumables also include integrated radio frequency identification (RFID) tags on the product labels. These enable viewing, tracking, and reporting of critical information about reagents and consumables, including usage, lot number, part number, expiry date, and on-instrument lifetime within the 3500 Series Data Collection Software. These features help streamline critical daily administrative tasks, to help save time and effort when tracking system performance.

The 3500 genetic analyzer family was primarily developed for ease of use. Sample and consumable loading are intuitive and, once loaded, operation of the instrument is fully automated. Because the data is obtained by progressive electrophoretic reading of single nucleotide–length differences of fragments, sequence data can be obtained in as little as 1 hour.

Analysis of CE data and confirmation of NGS variants

The data analysis software on the 3500 genetic analyzer family provides user-friendly navigation with its intuitive dashboard design. It also offers preconfigured plate templates to further support rapid and efficient sequence analysis. Thermo Fisher Cloud (**http://www.lifetechnologies.com/us/en/home/cloud.html**) contains a convenient and free-to-use tool: the QC module in the Sanger section allows scientists to view, quality check, manage, and share sequencing data. To conduct variant analysis of CE data and directly compare variants from the relevant NGS .vcf file with the sequence output from the 3500 Genetic Analyzer, a subscription to use the Variant Analysis (VA) and Next Generation Conformation

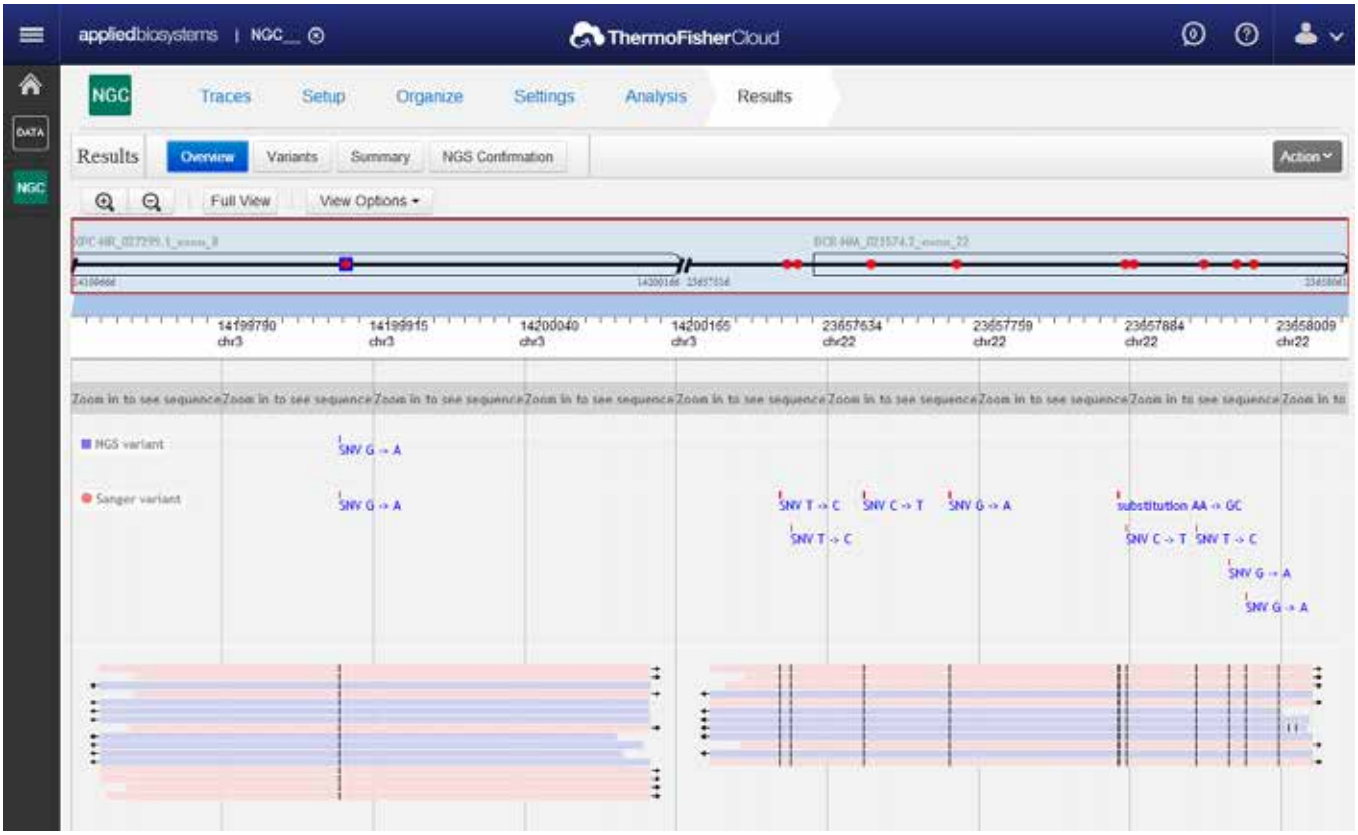

Figure 5. Alignment of Sanger CE trace files with NGS .vcf file data. The alleles identified and imported in the NGS .vcf file are shown with a blue bar. Alleles identified by CE sequencing are shown with a red bar. Note the G→A allele identified by NGS was confirmed by CE. Several alleles were identified by CE only; these were not present in the NGS .vcf file due to shorter read length.

(NGC) module can be purchased (Figure 5). The NGC module is a powerful tool that can bring NGS and Sanger data together to generate a comprehensive side-by-side comparative report. To that end a .vcf file from an NGS run is uploaded together with Sanger sequencing files (.ab1) generated using the Primer Designer Tool assays described above. The software displays a visual alignment of both file types and the user can examine the trace profiles to verify the accuracy of the base call (Figure 6). Together, these features make the 3500 Genetic Analyzer one of the easiest capillary electrophoresis systems to use.

CE (Conformité Européenne) mark. This designation is required for clinical research throughout the EU. An RFID tagging technology has been incorporated into all key consumables, allowing the laboratory to track consumable usage for the administrative records required in regulated environments.

Next-generation sequencing for discovery, capillary electrophoresis for confirmation

NGS platforms offer unprecedented scales of sequence determination, but also introduce the need to make sure the information obtained in

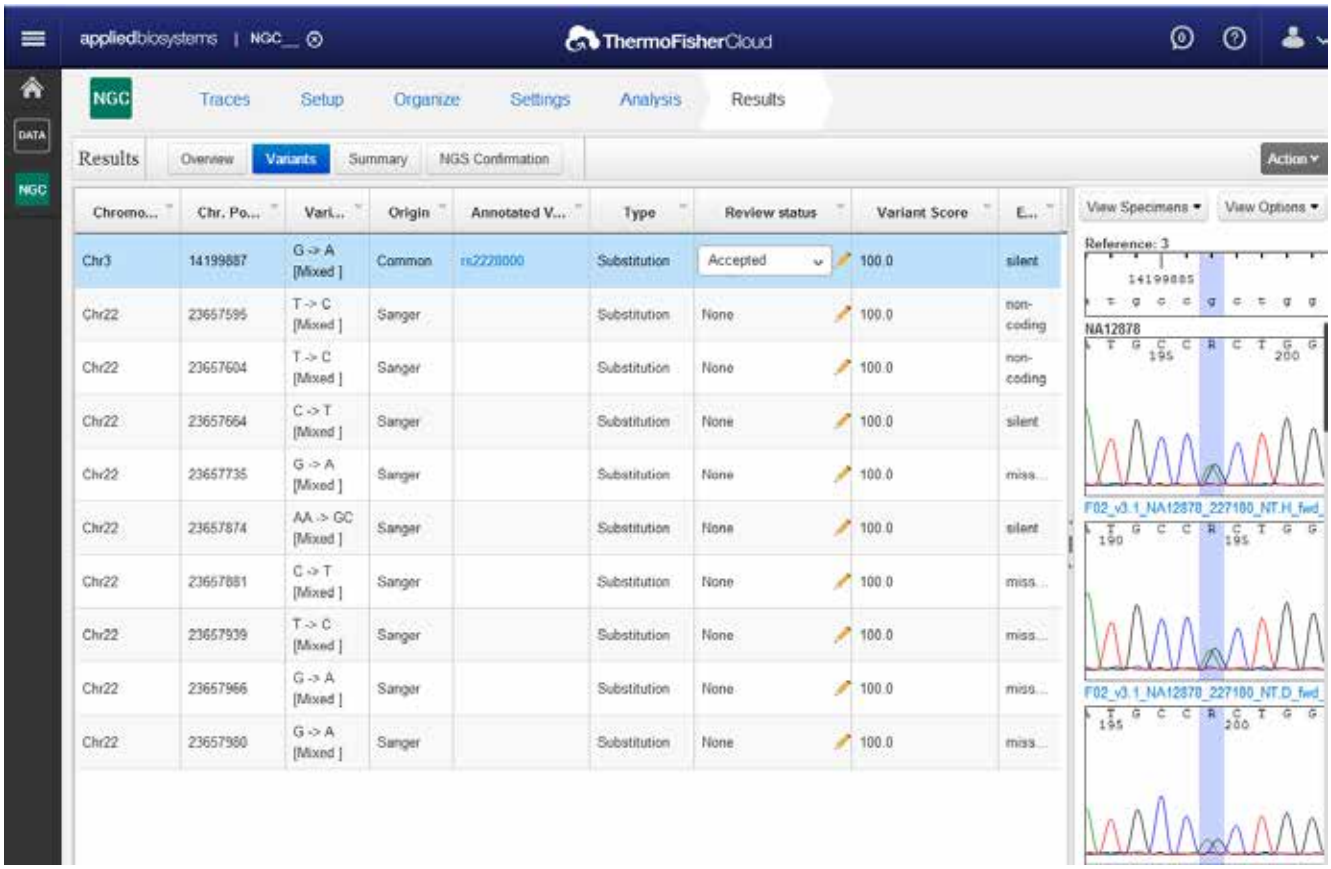

Figure 6. Review of variants detected by Sanger sequencing traces in the NGC module of Thermo Fisher Cloud. The NGC module generates a combined NGS and Sanger .vcf file and an annotated variant report.

Solutions for clinical science researchers

The promise of personalized medicine means that more sequencing solutions are required in clinical research laboratories. These laboratories need extra layers of controls and safeguards to ensure that precious samples are properly utilized and not wasted. In addition, the instrumentation and reagents must be robust so that the correct answer is returned to the clinical science researchers and subjects. The 3500 Genetic Analyzer systems have received the

these discovery-based experiments is correct. CE sequencing offers an ideal mechanism to confirm the presence of variants uncovered by NGS systems. First, the CE sequencing workflow is less labor intensive than NGS, minimizing time to results, minimizing costs, and minimizing the possibilities of user errors. Second, because CE sequencing focuses on longer reads of a single-target region, analysis of the data is more straightforward than NGS data analysis. Finally, because NGS and CE sequencing use different chemistries for sequence determination, systemic

errors are less likely to occur in both systems. Any variants that are detected by both NGS and CE sequencing are therefore likely to be genuine.

Summary

In this application note, we described a complete workflow for confirming next-generation sequencing data by capillary electrophoresis. We showed how the Primer Designer Tool facilitates selection and ordering of primers needed for CE sequencing. We showed the simplicity of ordering CE primers directly from Ion Torrent Variant Caller software. We described how the BigDye Direct kit provides a one-tube solution for up-front PCR and sequencing work, and how the 3500 family of CE sequencers offer ease of use and control necessary for confirmation of variant calls. Finally, we described how the CE software makes the results easily interpretable. Together, these tools allow scientists performing basic or clinical research to easily and quickly generate high-quality confirmation data of sequences that are critical to make their studies successful.

References

1. Gudbjartsson DF, Helgason H, Gudjonsson SA et al. (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 47(5):435–444.
2. Ross MG, Russ C, Costello M et al. (2013) Characterizing and measuring bias in sequence data. *Genome Biol* 14(5):R51.
3. van Dijk EL, Jaszczyszyn Y, Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* 322(1):12–20.
4. Ding L, Wendl MC, McMichael JF, Raphael BJ (2014) Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* 15(8):556–570.
5. Kim SY, Speed TP (2013) Comparing somatic mutation-callers: beyond Venn diagrams. *BMC Bioinformatics* 14:189.
6. Wall JD, Tang LF, Zerbe B et al. (2014) Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res* 24(11):1734–1739.
7. Sanger sequencing of FFPE genomic DNA and Ion Ampliseq library pools using the enhanced Primer Designer Tool. Available at **thermofisher.com/ceapplications**

applied biosystems

To learn more, go to
thermofisher.com/primerdesigner

ThermoFisher SCIENTIFIC