Title

Guidelines for Sanger sequencing and molecular assay monitoring

Journal

Journal of Veterinary Diagnostic Investigation, 32(6)

ISSN

1040-6387

Authors

Crossley, Beate M

Bai, Jianfa

Glaser, Amy

et al.

Publication Date

2020-11-01

DOI

10.1177/1040638720905833

Peer reviewed

Journal of Veterinary Diagnostic Investigation
2020, Vol. 32(6) 767–775

DOI: 10.1177/1040638720905833
jvdi.sagepub.com

Guidelines for Sanger sequencing and molecular assay monitoring

Beate M. Crossley,[1] Jianfa Bai, Amy Glaser, Roger Maes, Elizabeth Porter, Mary Lea Killian, Travis Clement, Kathy Toohey-Kurth

Abstract. Genetic sequencing, or DNA sequencing, using the Sanger technique has become widely used in the veterinary diagnostic community. This technology plays a role in verification of PCR results and is used to provide the genetic sequence data needed for phylogenetic analysis, epidemiologic studies, and forensic investigations. The Laboratory Technology Committee of the American Association of Veterinary Laboratory Diagnosticians has prepared guidelines for sample preparation, submission to sequencing facilities or instrumentation, quality assessment of nucleic acid sequence data performed, and for generating basic sequencing data and phylogenetic analysis for diagnostic applications. This guidance is aimed at assisting laboratories in providing consistent, high-quality, and reliable sequence data when using Sanger-based genetic sequencing as a component of their laboratory services.

Key words: quality score; Sanger sequencing; validation.

Introduction

It is a critical function of veterinary diagnostic laboratories (VDLs) to continually update testing methodologies in order to provide the best, most accurate, and cost-effective testing for their clients. The trend in molecular approaches in diagnostic medicine has evolved from selective use of standard PCR to routine and widespread use of real-time PCR (rtPCR), and most recently to bench-level next-generation, or high-throughput, sequencing. Throughout this evolution, and continuing today, Sanger DNA sequencing has served as the gold standard for determination of nucleic acid (NA) sequences, whether occurring naturally or produced synthetically. Sequence analysis is critical in laboratory medicine for the identification of emerging pathogens or new genotypes of existing pathogens, for identifying important evolutionary changes in the genomes of recognized pathogens, and critically, for verification of unusual laboratory findings such as recovering a specific pathogen from a new species or a new geographic area.

In addition to serving as a confirmatory assay of high diagnostic specificity, high-quality NA sequence analysis is critical for initial molecular-assay development, monitoring the efficacy of molecular-based assays and their key components, forensic investigations into the source of an agent or disease outbreak, and for genotyping or differentiation between field and vaccine strains. Given the complexity and the relatively high cost of sequencing, sequence analysis has until recently remained a supplemental tool in most VDLs. Logistically, for most VDLs, NA Sanger sequencing was routinely outsourced to specialized facilities and suppliers. Regardless of whether outsourced or in-lab, sample preparation, analysis, and interpretation of sequencing results remain the responsibility of the submitting diagnostic laboratory. This includes the responsibility for quality assessment of the sequencing results, manual editing of the generated sequence data as required, and final interpretation of the findings.

As with any laboratory technique, the steps associated with sequence analysis, from sample preparation through final analysis of the results, require protocols and guidance describing the process and ensuring that the work is executed consistently. Recognizing the critical importance of sequence analysis in current laboratory medicine, the Laboratory Technology Committee of the American Association of Veterinary Laboratory Diagnosticians (AAVLD) compiled the experience and utilized the expertise currently existing in accredited VDLs to provide consensus information and to propose national guidelines for sequence analysis, specifically targeting VDLs. The focus of the guidance is on traditional Sanger sequencing, a technological approach that forms the basis for both manual and automated NA-sequencing approaches.

California Animal Health and Food Safety Laboratory, University of California–Davis, Davis, CA (Crossley, Toohey-Kurth); Department of Population Medicine and Diagnostic Sciences, Cornell University, Ithaca, NY (Glaser); Kansas State Veterinary Diagnostic Laboratory, Kansas State University, Manhattan, KS (Bai, Porter); National Veterinary Services Laboratories, Ames, IA (Killian); Animal Disease Research and Diagnostic Laboratory, South Dakota State University, Brookings, SD (Clement); Veterinary Diagnostic Laboratory, Michigan State University, Lansing, MI (Maes).

[1]Corresponding author: Beate M. Crossley, California Animal Health and Food Safety Laboratory, University of California, Davis, PO Box 1770, Davis, CA 95616. bcrossley@ucdavis.edu

Sanger sequencing

Numerous modifications and advances in methodology have been published and adapted for use since sequence analysis was introduced in 1975,[16] including de novo sequencing and large-scale parallel sequencing (next-generation sequencing).[7,10] Traditional Sanger sequencing not only forms the basis for the newer and automated approaches, but continues to be the most common sequencing approach used in VDLs for sequence verification, assay monitoring, and as the foundation for many phylogenetic analyses.

In the Sanger-sequencing approach, amplified DNA or complementary DNA (cDNA) is annealed to an oligonucleotide primer and then extended by the DNA polymerase enzyme that incorporates either a mixture of the 4 deoxynucleotide triphosphates (dNTPs: dATP, dGTP, dCTP, dTTP) or chain-terminating dideoxynucleotide triphosphates (ddNTPs: ddATP, ddGTP, ddCTP, ddTTP). The inclusion of rate-limiting concentrations of the ddNTPs stops the elongation reaction as the ddNTPs are incorporated, resulting in distinguishable DNA fragments of various lengths.[16,17]

Current Sanger sequencing automation generally supports the generation of NA sequences up to 800–1,000 bp.[6,17,21] Typically, the most important limitations of the approach are low-quality sequences within the first 15–40 bp because of primer binding, and an inability to distinguish single base pair differences in longer segments (e.g., > 900 bp). Taking note of those limitations, both commercial and non-commercial sequence analysis software continues evolving to help the user assess and trim low-quality data automatically.

The VDL guidelines presented herein target the Sanger sequencing approach performed on capillary gel electrophoresis equipment. The guidance covers 1) sample preparation, 2) submission of DNA samples for sequencing, and 3) interpretation of chromatograms including software-specific guidance for monitoring and maintaining quality control. Suggestions are included for the sequence analysis of short segments (< 100 bp) given that these are frequently generated in VDLs during assay development, monitoring of PCR-based assays and related assay components, and for verification of rtPCR amplicons.

Sample handling and amplicon generation

In order for NA sequencing to be successful, long, non-degraded strands of amplicon (target, template) DNA are needed. The handling of diagnostic case materials and subsequent extraction of the target DNA, or extraction of target RNA and conversion to cDNA used in the PCR amplification step, is critical in providing the high-quality DNA template that is mandatory for sequence analysis. Therefore, throughout the process, sample handling must target preservation and extraction of intact DNA and RNA. Storage duration and temperature, fixatives, and other treatments (e.g., demineralization) of diagnostic case materials have an important impact on the stability and quality of the DNA and RNA that can be recovered. Diagnostic case materials that have decayed or been exposed to extreme heat, or preservation methods known to degrade NAs or crosslink the matrix thus making retrieval impossible (e.g., formalin-fixed preparations),[4,13,26] should be avoided, especially for RNA targets, which are more prone to degradation.

Numerous publications and commercial product guidance documents direct laboratorians to extraction protocols and reagents aimed at recovery of high-quality DNA and RNA from specified tissue types and conditions. This is the critical first step in generating the quality of DNA target needed for sequence analysis.[2,3] Commercial NA-extraction kits are continuously improving in their ability to recover both DNA and RNA in a single extraction procedure. As such, they have become essential for rtPCR assays that amplify both DNA and RNA targets, such as syndromic PCR assay panels. For optimal results, sequencing procedures require DNA or RNA amplicons that are longer than those used for rtPCR (i.e., 150 bp), requiring that extraction protocols are selected accordingly (e.g., Trizol extraction or extraction kits designed to provide strands of intact NA > 1,500 bp).[15] Similarly, it is important to keep in mind that the quality of the PCR assay used to generate the sequencing target, including primer design and optimization of the assay reagents and performance conditions, will critically impact the quality and quantity of the target (DNA or cDNA amplicon) generated for sequence analysis.

Online open-access primer design tools (e.g., NCBI, https://www.ncbi.nlm.nih.gov/tools/primer-blast/; IDT, idtdna.com; Primer3, http://bioinfo.ut.ee/primer3-0.4.0/; Primer3plus, http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/) and commercial primer design tools (e.g., Geneious, https://geneious.com; DNASTAR, https://dnastar.com; CLC Main Workbench, https://www.qiagenbioinformatics.com/products/clc-main-workbench/) are widely available and can be used to assist with primer design ensuring appropriate length and melting temperature, as well as the avoidance of sequences with the potential for secondary primer binding, mispriming, and hairpin or dimer formation. Degenerate primers sometimes cause problems in priming the sequencing reaction, and therefore should be avoided if possible. Primers are part of the DNA template submission package to sequencing laboratories. Most sequencing centers also provide common primers such as M13 forward/M13 reverse as part of their service at no charge. For targets that are cloned into some of the more common cloning vectors (e.g., specific plasmids), it is possible to use or purchase the primers at their optimal concentration directly from commercial sequencing facilities.

Preparing DNA template for sequence analysis submission

The steps involved in submitting a DNA template for sequence analysis include identification, isolation, purification, and quantification of the target DNA template. The target DNA

submitted for Sanger sequencing is often an amplicon from a conventional or rtPCR assay used for diagnostic purposes. However, the guidance provided herein applies also to plasmid inserts and synthesized products that are used as positive amplification controls in PCR-based testing. For optimal sequencing results, the target submitted for Sanger sequencing should exhibit a single product, or band, as confirmed by capillary electrophoresis or gel electrophoresis procedures. If one "band" is present, indicating a homogeneous product, the amplicon can be purified and used for sequencing. For PCR reactions in which more than one product is formed, the unique band of interest should first be isolated. Several gel purification methods are available to isolate the single band of interest. Examples include physical methods such as electrophoresis into a preformed trough, enzymatic methods such as agarose digestion, or purification methods using columns or magnetic beads.[5,18] Although in theory it is possible, given the specificity of the sequencing primers, to generate NA sequence information from a single amplicon included in a multiplex PCR reaction, this is an approach that is prone to primer interference, lower product yields, and increased analysis complexity, and therefore is not recommended for diagnostic applications. In most cases, repeating the PCR reaction using one specific pair of primers will resolve the problem. When the original template concentration is very low, reamplifying the single target from diluted multiplex PCR amplicons can often yield the desired concentration for sequencing.

Prior to sequence analysis, purification of the DNA or cDNA amplicon should be performed in order to eliminate unincorporated dNTPs, polymerase enzymes, unbound primers, salts, and other impurities that were part of the PCR reaction generating the DNA or cDNA sequencing target. Numerous options are available for this PCR product purification step, including bead-based, column-based, and enzymatic methods, all of which are available commercially.[5,18] For VDLs that outsource their sequencing, commercial sequencing laboratories typically provide, via their websites, a list of recommendations or preferences for specific cleanup kits and procedures to be used prior to submitting samples to their facility for sequencing. Many sequencing facilities also provide PCR purification services for an additional fee.

Once purified, the DNA or cDNA amplicon, as well as the sequencing primers, must be quantified to ensure the appropriate concentration and ratio between primers and sequencing target, which is important for the success of the sequencing reaction (Fig. 1A). An important factor in determining the target quantity is knowledge of the length of the PCR product. The amplicon length is determined by the position of the PCR primers and can be calculated by subtracting the forward primer nucleotide position from the reverse nucleotide primer position, using nucleotide sequence maps (e.g., NCBI, https://www.ncbi.nlm.nih.gov/nuccore) or other tools. Additionally, the size of a PCR amplicon can be directly visualized on electrophoresis gels and is most easily assessed when the amplicon is loaded immediately adjacent to a molecular weight ladder. All 3 factors (i.e., concentration of DNA/DNA amplicon, length of amplicon, and correct amount of primers) are essential for a successful sequencing reaction.

Quantification and purity of NA are typically assessed by spectrophotometric analysis or ultraviolet fluorescence tagging. Spectrophotometric equipment (e.g., NanoDrop; Thermo Fisher Scientific, Wilmington, DE) or fluorometers relying on the intercalation of an indicator dye (e.g., Qubit; Thermo Fisher Scientific) are commonly available in VDLs. Ultraviolet wavelength light (260 nm) absorption is directly related to NA concentration in the sample. Purity assessments are additionally employed to detect the presence of contaminating proteins. Peak light absorption by proteins, in particular the aromatic amino acids, occurs at a wavelength of 280 nm[15,24]; calculating the 260 nm/280 nm absorbance ratio of the sample will provide the purity of the sequencing target. It is suggested that a ratio of 1.8 or above is optimal for DNA samples. Low calculated ratios in samples are an indication of residual protein remaining after the extraction process, or of a suboptimal NA concentration. Although calculating the 260 nm/280 nm ratio is a valuable approach for measuring NA contamination in a protein solution, it must be noted that the approach is not sensitive enough to measure protein contamination in NA solutions.[24] It is additionally suggested that measurements be taken at a wavelength of 230 nm to check for other contaminants, such as carbohydrates, phenol, guanidine, and glycogen. The optimal values for 260 nm/230 nm ratio calculations are 2.0–2.2.[25] The guidance on DNA concentration values provided herein is based on a single amplicon in the sample; samples containing multiple amplicons (e.g., from multiplex PCR reactions) add to the total weight of the measured DNA and will provide false assumptions when calculating the necessary primer mix for the sequencing reaction.

As with the purification protocol, individual commercial sequencing facilities will provide recommendations (Table 1) for appropriate concentrations and ratios per their equipment and procedural preferences, and in many instances, commercial sequencing facilities will perform both the purification and quantification steps for a fee. Different technical approaches for quantification and verifying purity have been well documented in the published literature.[15,23,25]

It is important to note that although purity ratios provide information on amplicon quantity and quality, they do not guarantee successful sequencing, which may additionally be influenced by primer design and location, among other technical sequencing issues.

Submitting samples to a sequencing facility

The choice of sequencing services to be used, whether internal to the diagnostic laboratory or outsourced, should be guided by a quality assessment, customer service, and cost-efficiency.

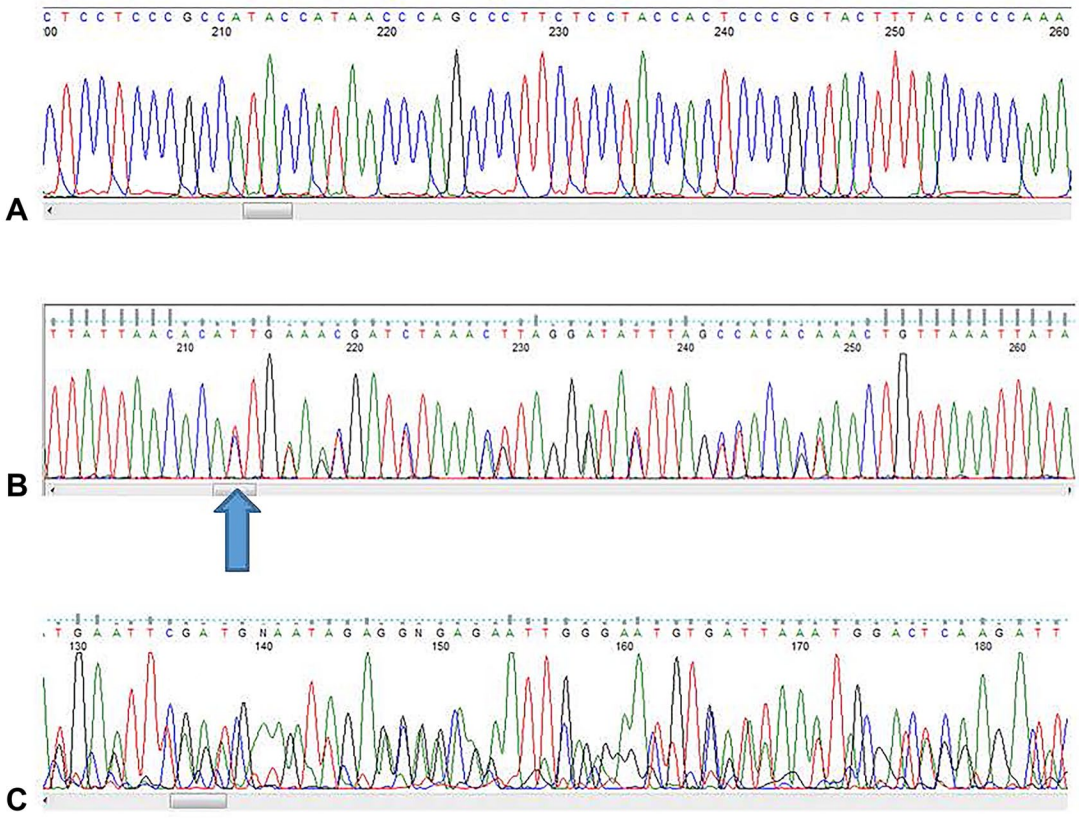

Figure 1. Examples of sequences. **A.** A good nucleic acid sequence (porcine circovirus 2). **B.** Dual infections are indicated by overlapping peaks, example pointed out by arrow (infectious bronchitis virus). **C.** High background noise (avian influenza virus), using FinchTV software v.1.4.0.

Table 1. An example showing appropriate premixed template length and weight (adapted from https://www.elimbio.com/Sample_Preparation.htm) for a PCR premix utilizing a specified 8 pmol of primer concentration.

Amplicon type	Length (bp)	Weight (ng)
Linear PCR amplicons	100–200	2–3
	200–500	6–8
	500–1,000	10–15
Circular DNA (e.g., plasmids)		500

It is a good management practice for laboratories that outsource to an outside vendor to verify the quality of the outside laboratory. Many commercial laboratories have Good Laboratory Practices or similar certifications (e.g., ISO/IEC 17025:2017, https://www.iso.org/obp/ui/#iso:std:iso-iec:17025:ed-3:v1:en) documenting the consistency, quality, reliability, and reproducibility of results. It is recommended that VDLs obtain the appropriate documentation of certification or ask to audit the commercial laboratory to ensure that it meets the required quality standard of the laboratory providing sequencing as a component of their own test menu and client services.

Sequence analysis

Interpretation of chromatograms. The final component of the sequencing process is data analysis. This requires that the sequence data be verified for quality, corrected when possible, and then compared to a reference sequence in order to provide a final sequencing result, such as the identity of an agent, or agent-specific support for a diagnosis or an epidemiologic investigation.

Quality assessment. It is crucial that raw data, filtered data outputs, and sequencing results are assessed for quality. Specific sequence-analysis software is needed to open and evaluate data files obtained from automated sequence analysis equipment. In addition to the software analysis performed internally by the sequencing instrumentation, a quality analysis review of the data should be performed manually by visual inspection of each strand. The same or similar software used by commercial sequencing facilities as part of their quality assessment before releasing the sequence information to VDLs is available to VDLs for their initial quality assessment prior to data analysis and interpretation of sequence results. Comprehensive DNA-analysis software packages such as the DNASTAR genomic

suite (https://www.dnastar.com/t-nextgen-seqman-ngen.aspx); Vector NTI (https://www.thermofisher.com), Geneious (https://geneious.com), and CLC Main Workbench (https://www.qiagenbioinformatics.com/products/clc-main-workbench/) incorporate sequence quality programs within their software. Examples of straightforward programs to identify and filter unreliable sequences that can be downloaded from the internet at no charge include Chromas 2.1.1 (http://www.softpedia.com/get/Science-CAD/Chromas-Lite.shtml) and Finch TV (https://digitalworldbiology.com/FinchTV). Both programs are commonly used to visualize chromatograms and provide a graphic quality score within the chromatogram (raw data) to give a visual guideline for the viewer. Multiple online and published reviews of the use of various software programs, including potential strengths and weaknesses, are available.

Dependent on the specific sequencing software used, the chromatograms may be received in either ".ab1" and/or ".fa" formats. Sequencing data might be received in files labeled as forward and reverse reads depending on the primer used, or labeled with laboratory-specific unique identifiers. However, it is not unusual to have one of the sequences, either forward or reverse, provided as the reverse complement file. For example, a sequence CAGTA would be displayed as TACTG in reverse complement mode. Some software programs automatically adjust the reading directions. In others, the user must specify and initiate the change.

Most sequences will have unreadable areas, typically located adjacent to the primer-binding sites. Once the chromatograms are accessed in the software, the primer sequences are removed by the user, also known as "trimmed," from both the forward and complementary strand, leaving just the high-quality target NA sequence. In cases in which a plasmid preparation was used to sequence an insert, all vector sequences must be trimmed. It is advisable to use software external to the sequencing instrumentation software to detect, filter, and eliminate unacceptable sequence data in order to provide base-specific quality scores.[14] Most common external software programs used for assessment of unacceptable sequence data are based on a prototype software called *Phred*. Although various commercial and non-commercial packages are available, most use *Phred*-based algorithms to provide the quality estimates that guide the removal or correction of low-quality regions in sequences (http://www.phrap.com/phred/; https://www.thermofisher.com/us/en/home.html).[9] The sequencing software–specific process of assigning bases to chromatogram peaks is referred to as "base calling." Regardless of the manufacturer, commercial software assigns a quality score to each base, which is then used to assess the overall quality of the sequence, areas of low quality, such as the ends, and estimations of consensus sequence accuracy. A score of 20 or higher indicates reliable identification of a nucleotide.

If the quality scores are very low, the software will automatically label the ambiguous base pairs with the letter "N," rather than assigning A, G, C, or T to the indefinite chromatogram peak. It is possible that some of the bases labeled as "N" can be clarified by manual inspection of the chromatogram data. As a rule of thumb, reliable sequence data should have < 5% of the bases identified as ambiguous after trimming primer data from the ends of the sequence; however, different applications such as pathogen strain identification, presence of virulence marker, etc. might require more stringent criteria.

Sequence chromatograms can be clarified by manual review and should show no overlap of peaks (Fig. 1B), which may be indicative of 2 closely related genotypes or indicative of nonspecific binding of primers to genomic material in the matrix. Sequences with overlapping peaks should be given an interpretation of "mixed sequences detected" or, in situations in which evidence of nonspecific binding is detected, the amplification and sequencing steps should be repeated.

Competing sequences, also referred to as "background noise," are often displayed at the bottom of the chromatogram in the form of secondary lower peaks when compared to the main target sequence (Fig. 1C). The competing peak is acceptable when the peak height is < 20% of the main sequence peak, otherwise the entire sequence should be discarded because interference generates a low-quality, thus unreliable, sequence result.

After a quality assessment of each strand individually, the forward and reverse strands are assembled and evaluated as one sequence. The term "sequence assembling" refers to the arrangement of 2 or more NA sequences for the purpose of generating a complete contig using the overlapping regions. When assembling the forward and reverse sequences, a successful assembly will show near-perfect agreement of the 2 strands. As noted previously, it is best practice to sequence the amplicon target in both the forward and reverse directions; however, if problems with one of the primers are detected, multiple reads of sequences generated from the same orientation showing complete agreement provide the same confidence as an agreement between the forward and reverse read alignment. Successful assemblies without mismatching results for the forward and reverse sequences indicate good quality for the sequencing work. Additionally, the assembly process compares the forward and reverse sequences generated, allowing the correction of sequencing errors. Because the sequencing process includes the use of enzymes, which have innate and documented error rates,[12] these unavoidable errors need to be detected and corrected prior to final analysis of the sequence data. When assembly of the forward and reverse sequences results in insertions or deletions (indels) in only one strand, the re-sequencing of that strand is recommended. Non-trimmed ends often contain indels, which is why the ends should have been discarded prior to assembling. The guidance provided herein is that insertions in sets of 3 nucleotides are more reliable than insertions of 1 or 2 nucleotides, which may introduce a stop codon or frameshift, and are often indicative of a sequence quality problem. Conceptual translation from DNA to 6 protein frames can be used to

find the correct reading frame of the sequence based on the absence of stop codons or a frameshift at the questionable base, which will assist with the assessment of indels occurring in an exon region.

Assembling of the forward and reverse sequences is used to generate a single consensus sequence, defined as the calculated order of most frequent nucleotides (or amino acids) found at each position in a sequence assembly. Bioinformatics software tools are used for calculating and visualizing consensus sequences; the evaluation of individual software products is beyond the scope of these guidelines. Insertions and deletions as noted above and occurring in the consensus sequence should be evaluated with caution. Well-trained technicians experienced in sequencing might also manually read a single strand for part of the analysis provided the sequencing data are of high quality, and the resulting "consensus sequence" will then include data originated from the alignment regions and high-quality single strand data. The consensus sequence is the final product after the quality check of the sequencing material received from outside sources. A consensus sequence can be produced by simply assembling the forward and reverse sequence of one amplicon or by aligning multiple unidirectional sequences (e.g., forward primer only) of one amplicon.

When the quality assessment of the sequence(s) generated is not satisfactory, the laboratory cannot proceed with the next analysis step. Experienced laboratories report that up to 10% of sequencing submissions require repetition; therefore, laboratories should take this into consideration and be cautious when projecting turnaround times for their clients. Sequencing failure rates substantially > 10% should be investigated as amplicon-preparation problems or quality-control problems occurring during the automated sequencing process. Sequencing failures can have multifold origins, and resolution requires a step-by-step troubleshooting approach, as applied in any quality problem investigation. Common examples of sequencing-failure investigations include assessing amplicon preparation, adjusting the ratio of primer to template, and modifying NA-purification protocols, among others. To specifically assess quality problems at the sequencing step, a suggested approach would be to perform a repeatability assessment by submitting sequencing amplicons for a previously sequenced (known) target and compare the sequence results using the previously described alignment process. Some laboratories periodically submit "no amplicon" buffer for sequencing as a "negative control," in order to monitor potential contamination issues that may occur in the amplification steps during the sequencing process.

Suggestions for short sequences (< 100 bp). Sequencing procedures are most efficient for amplicons ranging from 100 to 800 bp. Because many rtPCR assays are designed with amplicons of 65–100 bp, obtaining sequence results for rtPCR amplicons presents a challenge. Although sequences from amplicon < 100 bp can technically be generated, quality outcome for these shorter fragments cannot be guaranteed. Whether performed in-house or outsourced, those responsible for the equipment used in generating the sequence data should be alerted to all samples containing short amplicons (e.g., < 100 bp). Dependent on the equipment platform used, it is often possible to apply a sequence-analysis modification within the instrumentation software that will improve the outcome for short targets.

Other possibilities for sequencing short targets include the use of an "outer primer set" that amplifies a larger segment encompassing the specific target segment of interest. Because the primer sites are included, this option is also beneficial for monitoring PCR assays and for performing quality control trend analysis for evaluation of assay efficiency. It is additionally possible to clone the rtPCR amplicon into a vector, followed by sequencing using vector-specific primer sites (M13, SP6, T7, etc.). Some commercial sequencing facilities recommend specific vectors and maintain vector-specific primers in stock for their clients. Finally, M13, which is a common priming site, can provide an option for sequencing short amplicons. The M13 sequences can be tagged onto the original forward and reverse primers followed by a limited amplification (e.g., 15–20 cycles) performed with the tagged primers[27] for sequencing short amplicons resulting from real-time assays. This technique is reported to effectively provide the sequence of small amplicons ~ 50% of the time.

Interpretation of chromatograms—analysis

Once the quality assessment of the received sequences is satisfactory, analysis of the sequencing information can start. Sequence analysis is the interpretation of the NA order in the sequenced area. The process includes the assembling and the alignment (preparation of the consensus sequence) and comparison of the consensus sequence generated to reference sequences using relevant public access databases. A client-specific database can be used to generate phylogenetic trees and identity tables if needed. Various sequence-analysis programs as previously mentioned are available, and choice is user-dependent. It is a good management practice for VDLs to evaluate software before it is used for the analysis of client samples, including software traceability and documentation of its use (e.g., in standard operating procedures) in order to provide an audit trail that allows re-creation of the work performed. Procedures should be established and implemented by the laboratory to ensure that generated data are securely retrievable and approved for use by trained and qualified personnel.

Sequence check against databases (e.g., BLAST)

Generating the consensus sequence, as performed for the quality assessment, is the first step in performing the analysis.

Once the quality of sequences is verified and a consensus sequence established, the consensus sequence is either lined up manually to a reference sequence or compared against multiple reference sequences using online software tools and databases. VDLs often sequence for verification of a PCR amplicon, followed by phylogenetic assessments. For sequence verification, the most commonly used database among VDLs is the Basic Local Alignment Search Tool (BLAST), a bioinformatics tool established by researchers at the National Institutes of Health.[1] The BLAST algorithm allows comparison of newly generated sequences to a library of sequences (https://blast.ncbi.nlm.nih.gov/Blast.cgi). BLAST detects regions of similarity between biological sequences by comparing nucleotide or protein sequences to sequence databases and calculates the statistical significance of such findings. The software is designed to detect functional and evolutionary relationships between sequences as well as to help identify members of gene families.

A second tool called the "Smith–Waterman process" is considered to be much more accurate in finding all possible matches; however, the process is more time consuming and requires more computing power.[20] Since the original development of BLAST in 1990, many updates and improvements have been made. Although various options are available, comparison of nucleotide sequences to the nucleotide databank (blastn) and comparison to the protein data bank (blastx) are most commonly used in VDLs.[11] For use of blastn software, the consensus sequence from the diagnostic amplicon is entered into the web page followed by selection of the appropriate database (e.g., "other" for bacterial and viral sequences). Result optimizations can be selected as highly similar, more similar, or somewhat similar, each setting allowing more or fewer mismatches between the entered sequence and the found database matches. In diagnostic verification of PCR fragments, the agent origin is known and therefore the "highly similar" option is most often chosen. Algorithm parameters are given at the bottom of the page allowing modifications and documentation of the run parameter, which have more implications for research applications and for searching for unknown agents. Once the blastn search is completed, the web page displays the results ranked by similarity. The maximum score is the highest alignment score of a set of aligned segments to the same subject, whereas the bit-score and E-values are numeric values describing the overall quality of the alignment. Because E-value calculations take the length of the fragment into account, shorter sequence fragments yield higher E-score values. For amplicons generated by rtPCR assays (typically 50–150 bp), E-values are not as valuable as they are for amplicons generated for phylogenetic assessments (e.g., gene amplification) or the larger fragments generated from traditional standard PCR assays. If the submitted consensus sequence matches more than one entry in the databank with the same similarity scores, the matches are ordered by GenBank submission date showing the latest submission at the top. The order of the matching sequences can be re-ordered by clicking the headings under "descriptions" (e.g., clicking on "query cover" will render sequences by their percentage of matching regions to the input query sequence; selecting "ident" will order sequence by their percentage of identity to the query sequence). The rule of thumb for blasting shorter amplicons is to match at least 17 base pairs to a reference strain in the public database. Seventeen base pairs translate to 5–6 amino acids; a correct sequence of 5–6 amino acids outside of the primer sequence is considered as verification for the detection of the expected disease agent.[19]

Comparison to a specific reference strain can be performed using the blastn software by clicking the option of "strain comparison." This feature is helpful when monitoring longer fragments for evolutionary assessment or differentiation of field versus vaccine strains. Phylogenetic assessments are often helpful in visualizing the evolutionary distance between the isolates. The BLAST program has continued to improve over the years, especially with respect to the "naming" isolates. Submission of sequences to GenBank is now a more stringent and consistent process. However, the user is always cautioned that interpretation of the identity results returned from BLAST should be made based using biological knowledge and diagnostic judgment, and not solely on the numeric values returned from the software.

In specific situations of comparing or monitoring virus variants, for example after introducing a new vaccine, blastx searches allow comparison of the translated nucleotide sequence to the protein database, providing valuable epidemiologic or forensic information. This specific tool allows monitoring of virus evolution in highly variable sites coding for the docking sites of the host's neutralizing antibodies and can predict the efficacy of vaccines.

Sequence alignment and phylogenetic analysis

Phylogenetic analyses are most often performed on variable genes to monitor evolution, which are seldom the target of detection assays. Detection assays are instead typically designed to identify conserved regions of the infectious agent to ensure the detection of a wide range of field strains and variants. A few key applications for phylogenetic analyses include: 1) to monitor temporal and spatial changes of the pathogen in order to provide guidance to management; 2) to trace a particular strain during an endemic or an outbreak situation; 3) to compare closeness of the strain to vaccine strains of interest; and 4) to generate sequence information for autogenous subunit vaccine production.

Phylogenetic analysis includes a multiple sequence alignment step requiring that every sequence utilized has gone through a rigorous quality check and is available as a consensus sequence. Multiple alignment options are available in commercial and free sequencing software programs and web pages (e.g., Geneious, Vector NTI, CLC, BLAST, MEGA, ClustalW, Muscle).[8] Free web pages additionally

often provide the link to other helpful web tools such as a link to the BLAST search function; however, the documentation of the functionality of these tools is up to the accredited diagnostic laboratory. Once a multiple alignment is performed, all sequences must be trimmed to the same size. Phylogenetic software uses the length of a sequence in its calculation; making alignments including a variety of sequence lengths is incorrect, and therefore should not be used.

There is a plethora of phylogenetic tree algorithms available. Screening algorithms that are often used for comparing genotypes are the distance-based unweighted pair group method with arithmetic mean (UPGMA) and neighbor-joining algorithms. The most used structure- and character-based methods are parsimony and maximum-likelihood algorithms. An identity table is often generated and used to analyze the relationship of sequences in terms of percentage identity, from which a number can be used for decision making. Interpretation of identity percentages is purpose-dependent and varies among viruses and genes within a virus. As an example, for RNA viruses such as porcine reproductive and respiratory syndrome virus in the open-reading frame 5 region, 2–3% divergence among identity would indicate a different strain.[22]

Discussion

Sequence analysis of PCR amplicons is becoming an increasingly common service offered by VDLs. An abundance of commercial sequence analysis software and public-access, web-based programs is available to VDLs, which makes the processing of sequence data relatively user-friendly for laboratorians. Guidelines for quality assessment, analysis, and interpretation were developed by the Laboratory Technology Committee of the AAVLD and prepared as a consensus document to support efforts to provide high-quality, reliable, and consistent laboratory services. High-quality sequence assessment is a valuable tool for monitoring PCR assays. Our suggestions and guidance emphasize the importance of good quality practices in the handling and submission of diagnostic materials for sequence analysis, and we recommend that only high-quality sequence data are evaluated and interpreted for the laboratory's clients. Our suggested guidance also emphasizes that interpretation of sequence analysis data must be made with knowledge of the agent and strain variation as well as unique features of the genome, such as conserved and variable sites.

Acknowledgments

We thank Monica Reising for contributing to the discussions and editing of the manuscript.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Beate M. Crossley https://orcid.org/0000-0003-2932-7229

References

1. Altschul SF, et al. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
2. Campos P, Gilbert TP. DNA extraction from formalin-fixed material. Methods Mol Biol 2012;840:81–85.
3. Do H, Dobrovic A. Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil-DNA glycosylase. Oncotarget 2012;3: 546–558.
4. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem 2015;61:64–71.
5. Downey N. Extraction of DNA from agarose gels. In: Casali N, Preston A, eds. *E. coli* Plasmid Vectors: Methods and Applications. (Vol. 235, Methods in Molecular Biology series.) Totowa, NJ: Humana Press, 2003:137–139.
6. Gupta AK, Gupta UD. Next generation sequencing and its applications in animal biotechnology. In: Verma AS, Singh A, eds. Animal Biotechnology: Models in Discovery and Translation. Amsterdam, Netherlands: Elsevier/AP, 2014:345–367.
7. Jay E, et al. DNA sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping. Nucleic Acids Res 1974;1:331–353.
8. Kumar S, et al. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 2016;33: 1870–1874.
9. Ledergerber C, Dessimoz C. Base-calling for next-generation sequencing platforms. Brief Bioinform 2011;12:489–497.
10. Liu L, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:1–11.
11. Madden T. The BLAST sequence analysis tool. In: The NCBI Handbook. 2nd ed. Bethesda, MD: National Center for Biotechnology Information, 2013.
12. McInerney P, et al. Error rate comparison during polymerase chain reaction by DNA polymerase. Mol Biol Int 2014;2014:287430.
13. Quach N, et al. In vitro mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. BMC Clin Pathol 2004;4:1.
14. Richterich P. Estimation of errors in "raw" DNA sequences: a validation study. Genome Res 1998;8:251–259.
15. Sambrook J, Russell DW. Molecular Cloning: A Laboratory Manual. 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2001.
16. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94:441–448.
17. Sanger F, et al. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463–5467.
18. Santos F, et al. Purification, concentration and recovery of small fragments of DNA from *Giardia lamblia* and their use for other molecular techniques. MethodsX 2017;4:289–296.

19. Silvanovich A, et al. The value of short amino acid sequence matches for prediction of protein allergenicity. Toxicol Sci 2006;90:252–258.
20. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
21. Stranneheim H, Lundeberg J. Stepping stones in DNA sequencing. Biotechnol J 2012;7:1063–1073.
22. Stricker A, et al. Variation in porcine reproductive and respiratory syndrome virus open reading frame 5 diagnostic sequencing. J Swine Health Prod 2015;23:18–27.
23. Tataurov AV, et al. Predicting ultraviolet spectrum of single stranded and double stranded deoxyribonucleic acids. Biophys Chem 2008;133:66–70.
24. Warburg O, Christian W. Isolierung und kristallisation des gärungsferments enolase [Isolation and crystallization of the fermentation ferment enolase]. Biochem Z 1941;310:384–421. German.
25. Wilfinger WW, et al. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. BioTechniques 1997;22:474–481.
26. Wong SQ, et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. BMC Med Genomics 2014;7:1–10.
27. Zhang X, et al. Quasispecies of bovine enteric and respiratory coronaviruses based on complete genome sequences and genetic changes after tissue culture adaptation. Virology 2007;363:1–10.