Adopted: 20 March 2024

DOI: 10.2903/j.efsa.2024.8744

efsa JOURNAL

STATEMENT

Technical Note on the quality of DNA sequencing for the molecular characterisation of genetically modified plants

European Food Safety Authority (EFSA) | Adrián César-Razquin | Josep Casacuberta | Tamas Dalmay | Silvia Federici | Sara Jacchia | Dafni Maria Kagkli | Simon Moxon | Nikoletta Papadopoulou

Correspondence: nif@efsa.europa.eu

Abstract

As part of the risk assessment (RA) requirements for genetically modified (GM) plants, according to Regulation (EU) No 503/2013 and the EFSA guidance on the RA of food and feed from GM plants (EFSA GMO Panel 2011), applicants need to perform a molecular characterisation of the DNA sequences inserted in the GM plant genome. This Technical Note to the applicants puts together requirements and recommendations for the quality assessment of the methodology, analysis and reporting when DNA sequencing is used for the molecular characterisation of GM plants. In particular, it applies to the use of Sanger sequencing and next-generation sequencing for the characterisation of the inserted genetic material and its flanking regions at each insertion site, the determination of the copy number of all detectable inserts and the analysis of the genetic stability of the inserts. This updated document replaces the EFSA 2018 Technical Note and reflects the current knowledge in scientific-technical methods for generating and verifying, in a standardised manner, DNA sequencing data in the context of RA of GM plants. It does not take into consideration the verification and validation of the detection method which remains under the remit of the Joint Research Centre (JRC).

KEYWORDS

DNA sequencing, genetic stability, genetically modified organisms, molecular characterisation, next-generation sequencing, NGS, risk assessment

This is an open access article under the terms of the Creative Commons Attribution-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited and no modifications or adaptations are made.

© 2024 European Food Safety Authority. *EFSA Journal* published by Wiley-VCH GmbH on behalf of European Food Safety Authority.

CONTENTS

Summary

Genetically modified organisms (GMOs) are subject to a risk assessment (RA) and regulatory approval before entering the European market. In this process, the role of the European Food Safety Authority (EFSA) is to independently assess and provide scientific advice to risk managers on any possible risk that the use of GMOs may pose to human and animal health and the environment. As part of the RA requirements for genetically modified (GM) plants, according to Regulation (EU) No 503/2013 and the EFSA guidance on the RA of food and feed from GM plants (EFSA GMO Panel, 2011), applicants need to perform a molecular characterisation of the DNA sequences inserted in the GM plant genome.

In 2018, following a mandate from the European Commission, EFSA published the first version of the Technical Note on the quality of DNA sequencing for the molecular characterisation of genetically modified plants.

This updated document replaces the EFSA, 2018 Technical Note and reflects the current knowledge in scientific-technical methods for generating and verifying, in a standardised manner, DNA sequencing data in the context of RA of GM plants. It puts together requirements and recommendations for the quality assessment of the methodology, analysis and reporting when DNA sequencing is used for the molecular characterisation of GM plants. In particular, it applies to the use of Sanger sequencing and next-generation sequencing (NGS) for the characterisation of the inserted genetic material and its flanking regions at each insertion site, the determination of the copy number of all detectable inserts and the analysis of the genetic stability of the inserts. This note does not take into consideration the verification and validation of the detection method which remains under the remit of the Joint Research Centre (JRC).

A list of information that should be included in GMO applications submitted to EFSA in conjunction with the DNA sequences can be found in Annex 1. In order to assist in the submission of sequencing information in accordance with this document, and to enhance the efficiency of the sequencing quality check, applicants are requested to implement a harmonised structure of such data, follow specific naming conventions for data files and use the appropriate file extensions as described in Annex 2.

1 | INTRODUCTION

Genetically modified organisms (GMOs) are subject to a risk assessment (RA) and regulatory approval before entering the European market. In this process, the role of the European Food Safety Authority (EFSA) is to independently assess and provide scientific advice to risk managers on any possible risk that the use of GMOs may pose to human and animal health and the environment. As part of the RA requirements for genetically modified (GM) plants, according to Regulation (EU) No 503/2013[1] and the EFSA guidance on the RA of food and feed from GM plants (EFSA GMO Panel, 2011), applicants need to perform a molecular characterisation of the DNA sequences inserted in the GM plant genome.

At the nucleic acid level, the molecular characterisation for the RA of GM plants includes, among other analyses, the following three aspects: (1) the determination of the copy number of all detectable inserts, both complete and partial; (2) the determination of the organisation and sequence of the inserted genetic material at each insertion site as well as that of the 5′ and 3′ flanking regions, hereafter referred to as the characterisation of the insert and flanking regions; and (3) the analysis of the genetic stability of the inserts.

DNA sequencing techniques can be used for all three aspects of the molecular characterisation of GM plants in the frame of RA. In GMO applications, Sanger sequencing and next-generation sequencing (NGS) are used for the characterisation of the sequence of the insert as well as to demonstrate genetic stability. Similarly, techniques relying on NGS, such as Southern-by-Sequencing (SbS) and Junction Sequence Analysis (JSA), are often used as an alternative to Southern blot to determine the copy number of detectable insert(s) and demonstrate the genetic stability of the plant insertion sites (Guo et al., 2016; Guttikonda et al., 2016; Kovalic et al., 2012; Pauwels et al., 2015; Yang et al., 2013).

In 2018, following a mandate from the European Commission,[2] EFSA took over from the Joint Research Centre (JRC) the verification and quality assessment of the sequencing data for all GMO applications, and published the first version of the Technical Note on the quality of DNA sequencing for the molecular characterisation of genetically modified plants.[3] The EFSA 2018 Technical Note built on and replaced the EURL-JRC guideline 2016[4] (updated April 2017) on the quality and reliability of submitted information related to sequencing of the insert(s) and flanking regions, integrating and updating it where scientifically justified.

To ensure that the quality parameters used for the sequencing methodologies are in line with up-to-date scientific knowledge, as the technologies advance, the Technical Note 2018 is replaced by this updated document. This document reflects the current knowledge in scientific-technical methods for generating and verifying, in a standardised manner, DNA sequencing data in the context of RA of GM plants. It puts together requirements and recommendations for the quality assessment of the methodology, analysis and reporting when DNA sequencing is used for the molecular characterisation of GM plants. In particular, it applies to the use of Sanger sequencing and NGS for the characterisation of the inserted genetic material and its flanking regions at each insertion site, the determination of the copy number of all detectable inserts and the analysis of the genetic stability of the inserts.

This document does not take into consideration the verification and validation of the detection method, which remains under the remit of the JRC.

A list of information that should be included in GMO applications submitted to EFSA in conjunction with the DNA sequences can be found in Annex 1 (see Supporting information). In order to assist in the submission of sequencing information in accordance with this Technical Note, and to enhance the efficiency of the sequencing quality check, applicants are requested to implement a harmonised structure of such information, follow specific naming conventions for data files and use the appropriate file extensions as described in Annex 2 (see Supporting Information).

The present version is the first update after the publication of the original version on the 11 July 2018 (EFSA, 2018).

2 | DATA AND METHODOLOGIES

To update the EFSA 2018 Technical Note, EFSA staff and experts from the molecular characterisation working group of the GMO panel, specialised in this field, took into account the current knowledge in scientific-technical methods for generating and verifying, in a standardised manner, DNA sequencing data in the context of RA of GM plants. In order to review and update the recommendations previously provided, data from published scientific literature, the experience from the assessment of GMO applications containing data sets generated by Sanger sequencing or NGS since the implementation of the EFSA 2018 Technical Note, and the preparatory work performed by EFSA contractors (OC/EFSA/GMO/2020/01) were considered. The current update also aims at improving data quality and enabling automated quality assessments and data processing to ensure compliance, by clarifying requirements for data formats (particularly the final event sequence file) and introducing naming conventions for files and sequences.

[1]Commission Regulation (EU) No 503/2013 of 3 April 2013 on applications for authorisation of genetically modified food and feed in accordance with Regulation (EC) No 1829/2003 of the European Parliament and of the Council and amending Commission Regulations (EC) No 641/2004 and (EC) No 1981/2006. OJ L157, 8.6.2013, p. 1–48.

[2]European Commission mandate to develop a Technical Note to the applicants on, and checking of, the quality of the methodology, analysis and reporting covering complete sequencing of the insert and flanking regions, insertion site analysis of the genetically modified (GM) event, and generational stability and integrity.

[3]EFSA-Q-2017-00706.

[4]European Union Reference Laboratory for Genetically Modified Food and Feed; Guideline for the submission of DNA sequences and associated annotations within the framework of Directive 2001/18/EC and Regulation (EC) No 1829/2003.

This document takes into account the requirements of the guidance on the RA of food and feed from GM plants (EFSA GMO Panel, 2011) and of Regulation (EU) No 503/2013 and applies when sequencing is used for the characterisation of the inserted genetic material and its flanking regions at each insertion site, the determination of the copy number of all detectable inserts and the analysis of the genetic stability of the inserts.

3 | REQUIREMENTS FOR THE MATERIAL AND DNA SAMPLE PREPARATION

The material used for DNA sample preparation should be derived from the GM plant to be assessed. In case of stacks, the material should come from the GM plant (containing all events) under assessment. The applicant should provide a report clearly describing the source of the plant material specifically indicating the GM event(s) in the GM plant, with the unique identifier corresponding to the GMO, the plant species and the generation in the breeding tree. The applicant should also include a description on how and on what year the plant material was collected, specify the organ and/or tissues as well as the number of plants from which the DNA sample(s) used for sequencing was prepared. The DNA extraction protocol should be included in the report. If multiple DNA extractions are needed, it is strongly recommended to use one single DNA extraction protocol for sample preparation to minimise any putative effect of the extraction protocol to the downstream analyses.

It is also recommended that an amount of sample sufficient for at least three further sequencing experiments i.e. activities that lead to the generation of a final event sequence, should be stored from the submission of the application in case reanalysis is requested.

4 | REQUIREMENTS FOR THE SEQUENCING QUALITY, SPECIFIC TO THE TECHNOLOGY USED

This section provides requirements and recommendations on the information to be submitted in GMO applications when DNA sequencing approaches have been used for any of the molecular characterisation aspects that could be addressed by DNA sequencing. The two main technologies that are currently used in the context of RA of GM plants are Sanger sequencing and NGS.

4.1 | Sanger sequencing

This section provides the general requirements and recommendations on the information to be submitted when Sanger sequencing is used in GMO applications.

For the characterisation of the event(s), the final sequence submitted for each event (hereafter referred to as final event sequence) should be generated from the sequencing of two independent PCR amplicons, each one sequenced from the forward and the reverse strand, giving rise to a sequence covering each nucleotide at least four times i.e. two forward and two reverse.

The applicant should provide a report describing, as a minimum, the overall strategy to obtain the DNA fragment(s) (e.g. subcloning, long-run PCR) used for sequencing, the sequencing strategy and the details of the methodology and experimental design used to obtain the final event sequence.

The applicant should submit all individual sequences, alignments and final event sequence(s) for the GM plant under assessment as described in Section 6. Any uncertainty observed in the raw data and any manual editing performed on the sequence (base calling and trimming) should be reported and justified.

4.2 | Next-generation sequencing

This section provides requirements when NGS is used in GMO applications and describes the most relevant parameters to be considered when NGS methodology and generated data sets are assessed in applications. A report on the sequencing strategy and the details of the experiment has to be provided for the final event(s) sequence(s) submitted. This report should include, at least, the details of the experimental design, the description of the technology used and the sequencing method. The applicant should submit all sequences, alignments and final event(s) sequence(s) as described in Section 6.

4.2.1 | Library preparation and sequencing strategy

Information on the library construction method has to be provided. A detailed description of how each of the sequencing libraries was prepared along with details of the sequencing chemistry, strategy and platform should be given. In addition, if a sequence capture approach is used, it is critical that applicants thoroughly describe all experimental procedures and probe design, as well as how hybridisation conditions and capture efficiency were assessed.

4.2.2 | Quality of data sets

In order to assess NGS data sets, information on sequencing platforms used to generate the data together with the number and quality statistics of reads generated for each experiment must be described; all described information shall be included in the application. Providing this information is especially important when reads are not aligned to a reference genome, as such alignments would allow for an additional estimation of quality. Common tools providing quality statistics of reads, like average base quality across the reads and flagging potential quality issues, including over-represented reads (or k-mers) and contamination, are available and should be used. For example, FASTQC is a widely used tool for checking the quality of NGS data sets.[5]

The applicant should provide raw read numbers for each sequencing run. As trimming and quality filtering are often used to remove poor quality and low complexity reads and to trim sequencing adaptors or low-quality ends, the strategy for sequence trimming, the number of reads discarded and trimmed, as well as the average read lengths after each step of filtering and trimming, should all be discussed in the application to allow assessment of the methodology.

4.2.3 | Read depth

Different sequencing technologies producing reads of variable quality and length are currently available. The number of reads that cover a particular position (read depth) needed to obtain the final event sequence depends, among other factors, on the quality, the length of the individual reads and the purpose of the sequencing experiment. In order to assess the NGS data submitted by the applicant, information on the read depth, and where relevant (e.g. for whole genome sequencing; WGS) the average read depth and its variation, should be provided. The applicant should justify the (average) read depth based on the methodology and technology used.

Average read depth description when using WGS

When WGS is used (e.g. for the identification of the insert(s) and the possible insertion(s) of backbone sequences), there is a need to estimate the average read depth across the whole genome. To estimate the average read depth, different approaches are possible. In cases where a reference genome is available, reads must be aligned to the entire sequence to calculate average read depth. Only in cases where a reference genome is not available, reads could be aligned to a representative number of reference genes/genomic regions from different genomic locations to assess read depth locally. If the available reference genome is not considered to be representative of the plant variety under subject, this must be justified and both the above-mentioned approaches to estimate average read depth must be used. In cases where no genomic resources exist, theoretical average read depth metrics for whole-genome sequencing data, derived from the equation discussed below, should be used. Applicants should have a good estimate of the genome size of the GM plant and therefore should be able to calculate the number of reads required to cover the genome to a specified depth. This can be achieved using the Lander–Waterman formula (Lander & Waterman, 1988), where average read depth is indicated as coverage:

$$\text{Coverage (average read depth)} = \frac{\text{number of reads} \times \text{read length}}{\text{estimated genome size}}$$

The Lander–Waterman equation gives a theoretical average read depth. However, this equation does not consider platform- and sequence-specific biases (Ross et al., 2013) and provides estimate of the average read depth, which is a limitation as the read depth is not necessarily uniform across the genome (Sims et al., 2014). The applicant should also consider evaluating the number of reads corresponding to mitochondrial or plastid DNA and justify the read depth of the nuclear DNA, since the technology used and the genome of the respective GM plant may affect the average read depth calculation (Lutz et al., 2011).

4.2.4 | Description of bioinformatic analysis

The applicant can choose any appropriate bioinformatics pipeline for the various analyses of NGS data sets; however, the methodology and tools used should be thoroughly described by the applicant. In particular for unpublished or in-house tools, a full description along with the scripts, source code and pipelines, inputs and outputs of each of the steps in the analysis, and other parameters used should be provided. Any filtering of results and thresholds should be described and justified. In addition, a flow chart of the analysis showing how the raw data were processed, from start to end, in order to obtain the final results should be formulated and submitted for each final event sequence (Ekblom & Wolf, 2014). When common bioinformatics software such as BLAST and common tools for read filtering and trimming are used, the exact tool version must be provided. Since each tool includes multiple parameters and options, the exact parameters and options applied should be specified and justified in order to flag potential issues and ensure transparency.

[5] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

5 | ADDITIONAL CONSIDERATIONS FOR THE SEQUENCING QUALITY, SPECIFIC TO THE MOLECULAR CHARACTERISATION ASPECTS

This section describes considerations and requirements, in addition to those described in Sections 4 and 6, on the information to be submitted in GMO applications for each of the three specific aspects of the molecular characterisation that could be addressed by one or a combination of the DNA sequencing approaches described in Section 4.

5.1 | Sequencing for the characterisation of the insert(s) and flanking regions

In order to risk assess a GM event, the applicant has to characterise the sequence of the insert(s) and genomic flanking regions (EFSA GMO Panel, 2011, and Regulation (EU) No 503/2013). Regardless of the sequencing method used, all bases in the insert and flanking regions must be covered.

In cases where the applicant has previously submitted the sequence of an event to the European Commission, EFSA or EURL-GMFF, they are required to compare the sequence of the GM event under assessment with all previously submitted sequences of this event. The applicant has to provide an alignment including all those sequences, report the differences identified (if any) and discuss the reasons for these differences. This requirement applies to both the case of renewals and stack applications.

The final event sequence and the alignment(s) have to be submitted as described in Sections 6.1 and 6.4, respectively.

5.1.1 | Considerations when Sanger sequencing is used

When Sanger sequencing is used for the characterisation of the insert and genomic flanking regions, the applicant should comply with the requirements discussed above in Sections 3 and 4.1.

5.1.2 | Considerations when NGS is used

For the determination of the insert(s) sequence and genomic flanking regions, different NGS approaches may be used, such as WGS or sequence capture approaches to enrich for the target DNA fragments before sequencing (Ekblom & Wolf, 2014; Inagaki et al., 2015). Although in some cases this can be relatively straightforward, some configurations of the inserted sequences can make this more challenging e.g. the presence of sequence rearrangements or duplications within the locus, or the nature of the inserted sequence, including the presence of long and tandem repeats. A combination of approaches, including ultra long reads, sequencing of cloned genomic fragments or PCR amplicons (including by Sanger sequencing) may be needed in such cases. When NGS is performed after PCR amplification, the final event sequence should be generated from the sequencing of at least two independent PCR amplicons. The applicant is required to describe, discuss and justify the rationale of the approach used.

In particular, discussion and justification on read depth has to be provided as indicated in Section 4.2.3. To ensure an unambiguous characterisation of the insert(s) and flanking regions, a minimum read depth of 40× is always recommended.

5.2 | Determining the copy number of all detectable inserts

The determination of the copy number of all detectable inserts is required as part of the molecular characterisation of the GM plant. This can be achieved in a number of ways including junction sequence analysis (e.g. see Kovalic et al., 2012). This approach relies on the computational identification of junction reads that show both sequence identity with the insert or the vector sequence and with the host genome (chimeric reads). Because these reads have a partial match to both the insert/vector and the host genome, reads of sufficient length (approx. 100 bp) are required to accurately identify junctions. Any discarding of possible junction reads should be described and justified as described in Section 4.2.2.

Read depth for junction sequence analysis

As discussed in Section 4.2.3, read depth is a key factor to evaluate the quality of the data. The applicant should include detailed information on (average) read depth, as described in Section 4.2.3. Although this is dependent on the characteristics of the genome and the sequencing technology used, read depth should be sufficiently high to detect junction reads, and its sufficiency shall be justified by the applicant. Willems et al. (2016) have proposed a statistical framework for estimating the number of sequencing reads spanning the junction between the intentionally introduced DNA and the host genome needed to have a certain probability to detect transgene sequences ('identification approach') which may be useful for the applicant to consider when planning such experiments. A combination of approaches could also be used.

5.3 | Genetic stability

In the case of GM plants containing a single event, genetic stability encompassing (a) the Mendelian inheritance of the insert(s) and (b) the molecular stability of the event over several generations, has to be demonstrated. Mendelian inheritance is currently checked by segregation analysis and the Chi-square test. Molecular stability of the event, proving that the insertion site(s) and the structure of the insert(s) are maintained over several generations, can be demonstrated by Southern blot, PCR and/or DNA sequencing techniques. In the case of GM plants containing multiple events, the integrity of each event in the stack should be demonstrated. When sequencing by Sanger or NGS is used to demonstrate either molecular stability or integrity of the event, it should be performed following the requirements and recommendations described in Section 4. Reads must be aligned to the final event sequence provided in accordance with the requirements in Section 6.1 and must cover all bases in the insert and flanking regions of the final event sequence.

6 | DATA FORMAT REQUIREMENTS[6]

6.1 | Data format requirements for the final event sequence

The final event sequence has to be submitted as electronic ASCII text files using either the EMBL/GenBank format[7] or the ASN.1 format used by NCBI,[8] and features shall be annotated according to the INSDC Feature Table Definition (version 11.1 October 2021).[9]

The following **keywords** and features must always be present in the file, and must comply with the requirements described below:

- **LOCUS** (GenBank) | **ID** (EMBL): first line in the file, containing information on e.g. locus name or id, sequence length, molecule type, molecule topology or division code. All fields are mandatory and must follow the file format specifications for the corresponding format.

  For **locus name,** the unique identifier for the GMO as established by Commission Regulation (EC) 65/2004[10] and present in the OECD BioTrack Product Database[11] must be provided. In case of stacks, the following convention must be used: "*[unique_identifier of_single_event]_in_[unique_identifier_of_stack_event]*". In case of multiple inserts, the character "/" followed by the insert number must be added after the unique id of the single event. E.g.:

  - single event: "*ABC-A1236-1*";
  - stack event: "*ABC-A1236-1_in_XYZ-A0102-7xABC-A1236-1xXYZ-A9562-8*";
  - inserts: "*ABC-A1236-1/1*" and "*ABC-A1236-1/2*" (for two inserts of a single event) or "*ABC-A1236-1/1_in_XYZ-A0102-7xABC-A1236-1xXYZ-A9562-8*" (for the first insert of one event in a stack).

  To note, spaces are not allowed in the locus name. In addition, special characters should be avoided e.g. use a 0 (zero) to represent zeroes in the unique identifier instead of the Ø (slashed O) symbol.

- **DEFINITION** (GenBank) | **DE** (EMBL): title describing the sequence record, indicating e.g. the full stack in case of stack applications.
- **SOURCE** (GenBank only): scientific and common name of the organism following the structure "*Scientific_name (common_name)*" e.g. "*Zea mays (maize)*". It must contain the subkeyword ORGANISM (see below).
- **ORGANISM** (GenBank, as subkeyword of SOURCE) | **OS** (EMBL): formal scientific name of the organism (genus and species) e.g. "*Zea mays*". Optionally, additional lines after ORGANISM (GenBank) | starting with OC (EMBL) can be used to list the taxonomic classification levels separated by semicolons.

[6]The data format requirements apply also to the sequence information needed for the verification of the detection method. In addition, for the detection method verification only, if the sequence of a taxon-specific reference gene is included in the submission, the full sequence of the taxon-specific target and its GenBank accession number shall also be submitted (see also EURL GMFF Technical report on the Definition of Minimum Performance Requirements for Analytical Methods of GMO Testing, http://gmo-crl.jrc.ec.europa.eu/guidancedocs.htm).

[7]EMBL format description: https://ena-docs.readthedocs.io/en/latest/submit/fileprep/flat-file-example.html; GenBank format description: https://www.ncbi.nlm.nih.gov/genbank/release/256/.

[8]ASN.1 format description: https://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/ASNLIB.HTML.

[9]INSDC Feature Table Definition: https://www.insdc.org/submitting-standards/feature-table/; ENA's resource for INSDC features and qualifiers: https://www.ebi.ac.uk/ena/WebFeat/.

[10]Commission Regulation (EC) 65/2004: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32004R0065.

[11]https://biotrackproductdatabase.oecd.org.

- **REFERENCE** (GenBank) | **RN** (EMBL): reference including, at least, the subkeywords or additional keywords **TITLE** (GenBank) | **RT** (EMBL), **AUTHORS** (GenBank) | **RA** (EMBL) and **JOURNAL** (GenBank) | **RL** (EMBL). JOURNAL|RL can be used to indicate unpublished reports and studies.

- **FEATURES** (GenBank) | **FT** (EMBL): list of genetic elements and regions of interest annotated in the sequence.

  Only standard feature keys, qualifiers and controlled terminologies (when applicable) as defined by INSDC must be used, and each feature key must include all corresponding **mandatory qualifiers**. In addition, the following qualifiers are also required:

  - *'**/standard_name**'*, for all feature keys: a short text (max. ~30 characters) naming the feature. Additional text can be added in a separate *'/note'* qualifier.
  - *'**/gene**'* and *'**/product**'*, for feature key CDS (in addition to the *'**/translation**'* qualifier, which is mandatory by default)

  Feature locations must indicate the exact start and end positions of the feature in the sequence i.e. location descriptors such as '<', '>', '^' or '.', which express a certain level of uncertainty, are not accepted.

  Feature keys must describe:

  (i) All genetic elements (genes, promoters, terminators, etc.). For genes, use the feature key '**gene**'. For regulatory elements (e.g. promoters, terminators, enhancers), use **'regulatory'** and indicate the type of element under the ***'/regulatory_class'*** qualifier using the standard INSDC vocabulary.[12]

  (ii) All coding sequences, including their translation. Use the feature key **'CDS'** and report the translation in the ***'/translation'*** qualifier. The provided translation must correspond to the result of automatically translating the nucleotide sequence using the information provided in the feature location (start, stop, strand), unless justified otherwise e.g. in a *'/note'* qualifier.

  (iii) Sequence Tagged Site corresponding to the PCR amplicon of the detection method, as well as the names and sequences of forward and reverse primers and probes. For STS, use the **'STS'** feature key. For primers and probes, use the **'primer_bind'** feature key, indicate the type of primers ("forward primer", "reverse primer" or "probe") followed by their name or ID in the ***'/standard_name'*** qualifier (e.g. '/standard_name="forward primer XYZ"'), and provide the sequence using a ***'/note'*** qualifier.

  (iv) Flanking regions. The feature key **'misc_feature'** must be used to describe flanking regions, and the qualifier ***'/standard_name'*** must indicate either "5' flanking region" or "3' flanking region". *'/note'* qualifiers can be used to provide further details.

  (v) The feature key **'source'**, including the mandatory qualifiers ***'/organism'*** and ***'/mol_type'*** as well as any other optional qualifiers that help characterising the biological source of the sequence or sequence part.

- **ORIGIN** (GenBank) | **SQ** (EMBL): full sequence following the standard format.

6.2 | Data format requirements for Sanger experiments

The applicant should submit all individual sequences of each event in ABI or FASTQ format.

These sequences should be aligned to generate the final event sequence, and the alignment(s) should be submitted in CLUSTAL or FASTA format following the filename conventions specified in Annex 2.

The alignment(s) must always include the reference sequence. In case of more than one fragment or amplicon, an alignment of all the reference sequences to the consensus sequence of the final event should also be provided. Their names and sequences should be identical to those used in the alignment files with the sequencing results. In addition, sequence names in the alignments with the sequencing results must contain the name/id of the primers used and must clearly indicate their orientation by adding **'_FW'** (for forward) or **'_RV'** (for reverse) at the end e.g. '>primer_id:01234_RV' or '>primer_id:01235_FW'.

The final event sequence should be submitted as described in Section 6.1 and should be identical to the consensus/reference sequence(s) in the alignment(s).

[12] https://www.insdc.org/submitting-standards/controlled-vocabulary-regulatoryclass/.

6.3 | Data format requirements for NGS experiments

The application should include raw NGS reads in compressed (gzip) FASTQ format.

The sequences aligned/mapped to and used to generate the final event sequence should be provided in Sequence Alignment/Map (SAM) format (Li et al., 2009), Binary Alignment/Map (BAM) format (Li et al., 2009) or CRAM[13] format, following the filename conventions specified in Annex 2.

The final event sequence should be submitted as described in Section 6.1 and should be identical to the consensus/reference sequence(s) in the alignment(s).

6.4 | Data format requirements for the alignment to previously submitted sequences

Alignment(s) of the final event sequence to previously submitted sequences in accordance with Section 5.1 should be provided in FASTA or CLUSTAL format, following the filename conventions specified in Annex 2.

In addition, to ensure a correct identification of the sequences aligned, sequence names in the alignment(s) must be composed of the corresponding EFSA question number and the event unique identifier as described in Commission Regulation (EC) 65/2004, separated by an underscore e.g. 'EFSA-Q-2023-00001_ABC-A1236-1'. If an EFSA question number is not available at the time of submission, for the final event sequence of the new application, the EFSA question number shall be replaced exactly by 'EFSA-Q-0000-00000', e.g. 'EFSA-Q-0000-00000_ABC-A1236-1'. As indicated also in Section 6.1, the use of spaces and special characters should be avoided in the sequence names.

7 | SUPPORTING INFORMATION

In addition to providing the data described in Section 6, applicants are required to:

- Provide **Annex 1**, containing the list of information and data that must be included in GMO applications submitted to EFSA, duly filled in and signed;
- Comply with **Annex 2**, which provides the instructions to organise the sequencing information and data to be submitted to EFSA.

Annex 1 and Annex 2 are available under the Supporting Information section of the online version of this Technical Note, together with a supporting folder file named 'Sequencing Info'.

ABBREVIATIONS

BAM	binary alignment/map
EURL GMFF	European Reference Laboratory for Genetically Modified Food and Feed
GM	genetically modified
GMO	genetically modified organism
JRC	Joint Research Centre
JSA	junction sequence analysis
NGS	next-generation sequencing
PCR	polymerase chain reaction
RA	risk assessment
SAM	sequence alignment/Map
WGS	whole-genome sequencing

ACKNOWLEDGEMENTS

EFSA wishes to thank the members of the Working Group on Molecular Characterisation for the preparatory work on this scientific output and EFSA staff member Tommaso Raffaello for the support provided to this scientific output

CONFLICT OF INTEREST

If you wish to access the declaration of interests of any expert contributing to an EFSA scientific assessment, please contact interestmanagement@efsa.europa.eu.

REQUESTOR

EFSA

[13] https://www.ebi.ac.uk/ena/software/cram-toolkit.

QUESTION NUMBER

EFSA-Q-2023-00886

COPYRIGHT FOR NON-EFSA CONTENT

EFSA may include images or other content for which it does not hold copyright. In such cases, EFSA indicates the copyright holder and users should seek permission to reproduce the content from the original source.

NOTE

This document replaces the 2018 version https://doi.org/10.2903/j.efsa.2018.5345 published on 11 July 2018.

REFERENCES

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms). (2011). Guidance for risk assessment of food and feed from genetically modified plants. *EFSA Journal*, *9*(5), 2150. https://doi.org/10.2903/j.efsa.2011.2150

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms). (2018). Scientific opinion on the Technical Note on the quality of DNA sequencing for the molecular characterisation of genetically modified plants. *EFSA Journal*, *16*(7), 5345. https://doi.org/10.2903/j.efsa.2018.5345

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*, 1026–1042. https://doi.org/10.1111/eva.12178

Guo, B., Guo, Y., Hong, H., & Qiu, L. J. (2016). Identification of genomic insertion and flanking sequence of G2-EPSPS and GAT transgenes in soybean using whole genome sequencing method. *Frontiers in Plant Science*, *7*, 1009. https://doi.org/10.3389/fpls.2016.01009

Guttikonda, S. K., Marri, P., Mammadov, J., Ye, L., Soe, K., Richey, K., Cruse, J., Zhuang, M., Gao, Z., Evans, C., Rounsley, S., & Kumpatla, S. P. (2016). Molecular characterisation of transgenic events using next generation sequencing approach. *PLoS One*, *11*, e0149515. https://doi.org/10.1371/journal.pone.0149515

Inagaki, S., Henry, I. M., Lieberman, M. C., & Comai, L. (2015). High-throughput analysis of T-DNA location and structure using sequence capture. *PLoS One*, *10*, e0139672. https://doi.org/10.1371/journal.pone.0139672

Kovalic, D., Garnaat, C., Guo, L., Yan, Y., Groat, J., Silvanovich, A., Ralston, L., Huang, M., Tian, Q., Christian, A., Cheikh, N., Hjelle, J., Padgette, S., & Bannon, G. (2012). The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterisation of crops improved through modern biotechnology. *Plant Genome-us*, *5*, 149–163. https://doi.org/10.3835/plantgenome2012.10.0026

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, *2*, 231–239.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lutz, K. A., Wang, W., Zdepski, A., & Michael, T. P. (2011). Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnology*, *11*, 54. https://doi.org/10.1186/1472-6750-11-54

Pauwels, K., de Keersmaecker, S. C. J., de Schrijver, A., du Jardin, P., Roosens, N. H. C., & Herman, P. (2015). Next-generation sequencing as a tool for the molecular characterisation and risk assessment of genetically modified plants: Added value or not? *Trends in Food Science and Technology*, *45*, 319–326. https://doi.org/10.1016/j.tifs.2015.07.009

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*, R51. https://doi.org/10.1186/gb-2013-14-5-r51

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, *15*, 121–132. https://doi.org/10.1038/nrg3642

Willems, S., Fraiture, M. A., Deforce, D., de Keersmaecker, S. C. J., de Loose, M., Ruttink, T., Herman, P., van Nieuwerburgh, F., & Roosens, N. (2016). Statistical framework for detection of genetically modified organisms based on next generation sequencing. *Food Chemistry*, *192*, 788–798. https://doi.org/10.1016/j.foodchem.2015.07.074

Yang, L., Wang, C., Holst-Jensen, A., Morisset, D., Lin, Y., & Zhang, D. (2013). Characterisation of GM events by insert knowledge adapted re-sequencing approaches. *Scientific Reports*, *3*, 2839. https://doi.org/10.1038/srep02839

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: EFSA (European Food Safety Authority), César-Razquin, A., Casacuberta, J., Dalmay, T., Federici, S., Jacchia, S., Kagkli, D. M., Moxon, S., & Papadopoulou, N. (2024). Technical Note on the quality of DNA sequencing for the molecular characterisation of genetically modified plants. *EFSA Journal*, *22*(4), e8744. https://doi.org/10.2903/j.efsa.2024.8744