UPDATE OF THE TECHNICAL NOTE ON SEQUENCING

AD HOC MEETING WITH INDUSTRY REPRESENTATIVES

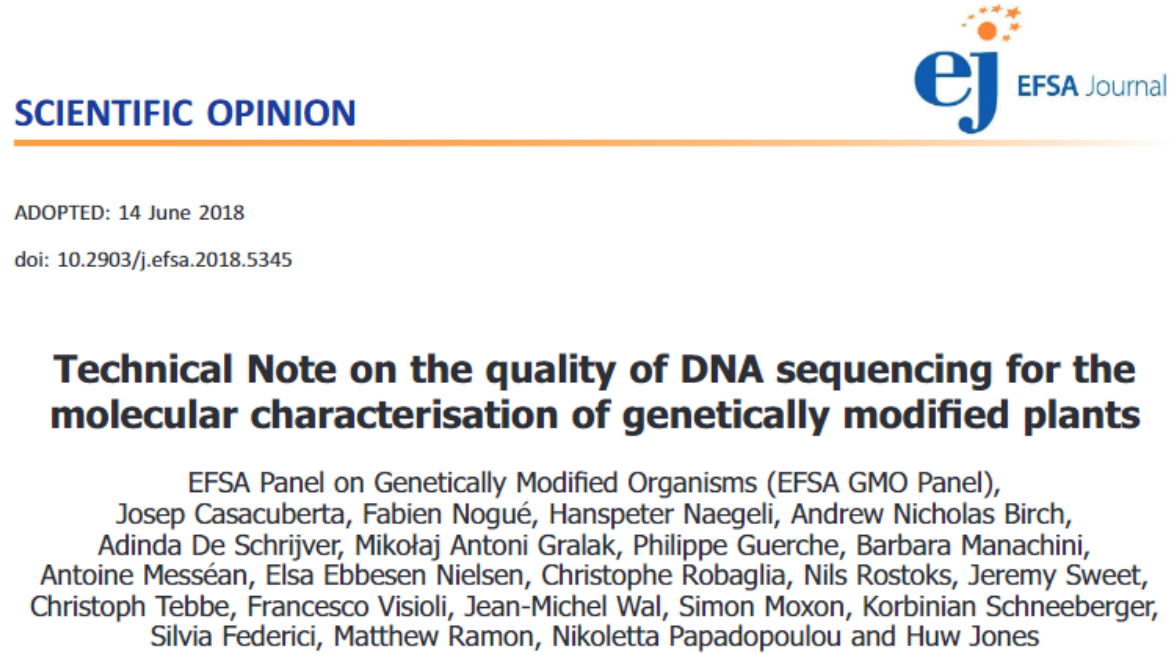
5-6 October 2023

EUROPEAN FOOD SAFETY AUTHORITY

BACKGROUND: EFSA 2018 TECHNICAL NOTE ON SEQUENCING

- **EC Mandate** to develop a *Technical note to applicants on, and checking of, the quality of the methodology, analysis and reporting covering full sequencing and insertion site analysis of GM event, and generational stability and integrity*
- **EFSA Technical Note published** in July 2018
  - presented at meeting with industry Oct 2018
  - **applicable** from 1 October 2018
  - foresees **updates** to keep up to date with scientific knowledge and technological advances
- **Achievements** since the implementation:
  - Increased quality in the data received
  - Reduced significantly clocks for additional clarification

THANKS!!



EFSA Journal

SCIENTIFIC OPINION

ADOPTED: 14 June 2018

doi: 10.2903/j.efsa.2018.5345

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

EFSA Panel on Genetically Modified Organisms (EFSA GMO Panel), Josep Casacuberta, Fabien Nogué, Hanspeter Naegeli, Andrew Nicholas Birch, Adinda De Schrijver, Mikołaj Antoni Gralak, Philippe Guerche, Barbara Manachini, Antoine Messéan, Elsa Ebbesen Nielsen, Christophe Robaglia, Nils Rostoks, Jeremy Sweet, Christoph Tebbe, Francesco Visioli, Jean-Michel Wal, Simon Moxon, Korbinian Schneeberger, Silvia Federici, Matthew Ramon, Nikoletta Papadopoulou and Huw Jones

UPDATE PURPOSE & METHODOLOGY

- EFSA 2018 Technical Note needed an **update** in order to:
  1. Review and update **recommendations**
  2. Provide **clarifications** and improve text
  3. Improve **data quality** and enable **automated sequencing quality check**, by:
     - clarifying and detailing requirements for **data formats**
     - introducing **name** conventions for **files and sequences**
- Developed by **EFSA Staff** and **experts** from the **MC WG** of the GMO Panel
- The update has taken into account:
  - Published **literature**
  - Experience in the **assessment** since implementation of the 2018 Technical Note
  - Preparatory **work by EFSA contractors**
- The **overall purpose** of the Technical Note remains **unchanged**

TOC & MAIN CHANGES

1. Introduction
2. Data and Methodologies
3. Requirements for the material and DNA sample preparation
4. Requirements for the sequencing quality, specific to the technology used
    1. **Sanger** Sequencing
    2. **Next Generation** Sequencing
5. Additional considerations for the sequencing quality, specific to the molecular characterisation aspects
    1. Sequencing for the **characterisation of the insert(s)** and flanking regions
    2. Determining the **copy number** of all detectable inserts
    3. **Genetic stability**
6. Data format requirements
    1. Data format requirements for the **final event sequence**
    2. Data format requirements for **Sanger experiments**
    3. Data format requirements for **NGS experiments**
    4. Data format requirements for the **alignment to previously submitted sequences**
7. Supporting information
    - Annex 1
    - Annex 2

UPDATES TO SECTION 4 – SEQUENCING QUALITY BY TECHNOLOGY

- **4.1 Sanger sequencing**:
  - Clarification: minimum coverage requirement of 2x forward and 2x reverse

    *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 different 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. 2 forward and 2 reverse**.*

- **4.2 NGS sequencing:**
  - Reordering subsections e.g. moving *Library preparation and sequencing strategy* before *Quality of datasets*
  - **4.2.3 Read depth:**
    - Clarification: for the calculation of **average read depth** when using **WGS**, alignment to several reference genes / regions **only** if a **reference genome is not available**
    - Reordering: **moving** section ***Minimum read depth*** *for characterisation of insert(s) and flanking regions* to **5.1.2**, since section 4 is about general considerations (see next slide)

UPDATES TO SECTION 5 – SEQUENCING QUALITY BY MC ASPECTS

- **5.1 Sequencing for the characterisation of the insert(s) and flanking regions**
  - Clarifications:
    - to submit **alignments to all previously submitted sequences** for **both** the case of **renewals and stack** applications
    - regardless of the sequencing method used, **all bases** in the insert and flanking regions must be **covered** (i.e. breadth of coverage 100%)
  - **5.1.2** Considerations when **NGS** is used: recommendation of a **minimum read depth of 40x** (independently of sequencing technology)
- **5.2 Determining the copy number of all detectable inserts:** no relevant changes
- **5.3 Genetic stability**
  - Clarification: when Sanger or NGS used to demonstrate stability or integrity:
    - reads must be **aligned to the final event sequence** as per section 6.1
    - reads must **cover all bases** in the insert and flanking regions (breadth 100%)

UPDATES TO SECTION 6 – DATA FORMAT REQUIREMENTS

- Two new sections introduced:
  - **6.1** Data format requirements for the **final event sequence**
    - previously part of the introduction of section 6
    - most requirements already existed but are now more clearly detailed, with more examples
    - instructions based on real cases of submitted files
  - **6.4** Data format requirements for the **alignment to previously submitted sequences**
    - new section introducing name convention for sequences in the alignments
- Clarifications and minor changes provided for sections:
  - 6.2 Data format requirements for **Sanger** experiments
  - 6.3 Data format requirements for **NGS** experiments

SECTION 6.1 – FINAL EVENT SEQUENCE FORMAT

1. General: follow **format specifications** of EMBL / GenBank / ASN.1 formats and INSDC Feature Table Definition for annotating features

2. Required **keywords**

   - **LOCUS** (GenBank) | **ID** (EMBL)**:** first line, including id/name, sequence length, molecule type, molecule topology, and division code

Locus name: use of **unique identifiers** of the GMO as per **Commission Regulation (EC) 65/2004:**

→ **single events:** unique identifier *e.g. MON-00810-6*

→ **stack events:** unique identifier of single event "**_in_**" unique identifier of stack event
e.g. MON-00810-6_in_DP-004114-3xMON-00810-6xSYN-IR604-5xMON-00603-6

→ **inserts:** add "**/**" followed by a number
e.g. *MON-00810-6/1 or MON-00810-6/2*, for inserts 1 and 2 of a single event
MON-00810-6/1_in_DP-004114-3xMON-00810-6xSYN-IR604-5xMON-00603-6, for insert 1 of an event in the stack

❖ Do not use spaces (not allowed by format specifications)

❖ Do not use special characters: use 0 (zero) to represent zeroes instead of Ø (slashed O)

SECTION 6.1 – FINAL EVENT SEQUENCE FORMAT

1. General: follow **format specifications** of EMBL / GenBank / ASN.1 formats and INSDC Feature Table Definition for annotating features
2. Required **keywords**
   - **LOCUS** (GenBank) | **ID** (EMBL): first line, including id/name, sequence length, molecule type, molecule topology, or division code
   - **DEFINITION** (GenBank) | **DE** (EMBL): description, e.g. the full stack in case of stack applications
   - **SOURCE** (GenBank): scientific and common name of the organism e.g. *Zea mays (maize)*
   - **ORGANISM** (GenBank, as subkeyword of SOURCE) | **OS** (EMBL): scientific name of the organism (genus and species) e.g. *Zea mays*.
   - **REFERENCE** (GenBank) | **RN** (EMBL): reference including, at least, the corresponding additional keywords to report title, authors and journal (possible to indicate 'unpublished')
   - **FEATURES** (GenBank) | **FT** (EMBL): list of genetic elements and regions of interest annotated in the sequence.
   - **ORIGIN** (GenBank) | **SQ** (EMBL): full sequence following the standard format

SECTION 6.1 – FINAL EVENT SEQUENCE FORMAT

1. General: follow **format specifications** of EMBL / GenBank / ASN.1 formats and INSDC Feature Table Definition for annotating features
2. Required **keywords**
3. Required **features**
   - must describe:

Requirement	Element	Feature key	Qualifier	Value or vocabulary
All genetic elements	genes	**gene**		
	regulatory elements	**regulatory**	**/regulatory_class**	INSDC regulatory_class incl. "promoter", "terminator", "enhancer"
All coding sequences	coding sequence	**CDS**	**/translation ***	
Elements for the detection method	sequence tagged site	**STS**		
	forward and reverse primers, and probe	**primer_bind**	**/standard_name**	"forward primer", "reverse primer", OR "probe" (followed by name or ID)
			/note	full sequence of the primer/probe
Flanking regions	flanking regions	**misc_feature**	**/standard_name**	"5' flanking region" OR "3' flanking region"

* must correspond to the automatic translation of the nucleotide sequence in the indicated location (start, stop, strand), unless justified

SECTION 6.1 – FINAL EVENT SEQUENCE FORMAT

1. General: follow **format specifications** of EMBL / GenBank / ASN.1 formats and INSDC Feature Table Definition for annotating features
2. Required **keywords**
3. Required **features**
   - must describe all genetic elements, coding sequences, STS and primers and flanking regions
   - reported using only standard INSDC feature keys, qualifiers and controlled terminologies; feature keys must include all corresponding mandatory qualifiers and, in addition:

     → **"/standard_name",** for all feature keys: a short text (max. ~30 characters) naming the feature

     → **"/gene"** and **"/product":** for CDS (in addition to "/translation", 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 "." are not accepted.

Example GenBank/EMBL files following these criteria will be provided in Annex 2.

SECTION 6.2 – SANGER SEQUENCING FORMATS

- Raw data: ABI or FASTQ formats – no change
- Alignment to final event sequence: CLUSTAL or FASTA format - main changes:
  - Minor additions:
    - **sequence names** must contain the name/id of the primers, indicating their orientation by adding "**_FW**" or "**_RV**" at the end of the name e.g. ">primer_id: 01234_FW" or ">primer_id: 01234_RV"
    - **file names** must follow indications in **Annex 2** (see below)
  - Clarifications:
    - alignment file(s) must always contain the **reference sequence**
    - in case of >1 fragment or amplicon, an **additional file** containing all of the reference sequences aligned to the **consensus sequence** of the final event should be provided
      - sequences and their names should be identical to the reference sequences in the results files
    - the **final event sequence** submitted as described in 6.1 must be **identical** to the consensus/reference sequence(s) in the alignment(s)

SECTION 6.3 – NGS SEQUENCING FORMATS

- Raw data: compressed FASTQ formats, using GZIP
- Alignment to final event sequence: SAM/BAM/CRAM format - main changes:
  - Minor addition: **file names** must follow indications in **Annex 2** (see below)
  - Minor deletion: suggestion to submit **ACE files**
  - Clarification: the **final event sequence** submitted as described in 6.1 must be **identical** to the consensus/reference sequence(s) in the alignment(s)

SECTION 6.4 – ALIGNMENT TO PREVIOUS SEQUENCES FORMAT

- New section
- Format: FASTA or CLUSTAL
- Considerations:
  - **file names** must follow indications in **Annex 2** (see below)
  - **sequence names** must be composed of the corresponding **EFSA question number** and the **unique identifier** of the (single) event as described in Commission Regulation (EC) 65/2004 separated by an underscore "_"
    e.g. EFSA-Q-2023-12345_MON-00810-6

UPDATES TO SECTION 7 – SUPPORTING INFORMATION

- Reduced:

  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*

- Main changes:
  - **Annex 1:** minor update to reflect changes in the Technical Note
  - **Annex 2:** introduced criteria and conventions for file names, file extensions, and compressed folders

SECTION 7 – ANNEX 2

1. **General folder structure**: unchanged
2. **File names, file extensions and compressed folders**: new section, introducing:

Table 1. File extensions

File format / type	Accepted file extension(s)
GenBank	.gb; .gbk; .genbank
EMBL	.embl
ASN.1	.asn
FASTA	.fa; .fas; .fasta (+*.gz)
CLUSTAL	.aln; .clustal
FASTQ	.fastq; .fq*
ABI	.ab1; .abi
BAM/SAM/CRAM	.bam; .sam; .cram
FASTQC	.fastqc; .html

Table 2. File names

File purpose		File format	Name convention	Example(s)
Final event sequence		GenBank; EMBL; ASN.1	**[UNIQUE_ID]_final_event_sequence**	EventXYZ_final_event_sequence.gb
Characterisation of the insert by Sanger sequencing	Alignment of Sanger reads to reference	FASTA; CLUSTAL	**_sanger_sequencing**	EventXYZ_sanger_sequencing.aln
	Alignment of reference sequences to consensus		**_sanger_consensus**	EventXYZ_sanger_consensus.fasta
Characterisation of the insert by NGS -alignment		BAM; SAM; CRAM	**_ngs_characterisation**	EventXYZ_ngs_characterisation.bam
Analysis of genetic stability by NGS		BAM; SAM; CRAM	**_genetic_stability**	EventXYZ_generationX_genetic_stability.bam
Alignment of the event sequence to previous applications		FASTA; CLUSTAL	**_event_alignment**	EventXYZ_event_alignment.fa

Compressed folders

- A single compressed folder
- Use of 7-Zip, Zip

STEPS FORWARD: CONSULTATION AND PUBLICATION

- **Consultation**:
  - Draft update of the Technical Note to be shared **next week**
  - Applicants to provide feedback **by end of October (TBC)**
- **Publication**:
  - Update of Technical Note to be published **by end of November (TBC)**
  - EFSA will make publicly available a **script to check the compliance** of the sequencing package with the requirements in the Technical Note
    - Applicants are encouraged to check for compliance before submitting
    - Your feedback is welcome
  - In addition: planned update of the EFSA Statement on WGS requirements for microorganisms
    - Applicable to GMM applications

STAY CONNECTED

SUBSCRIBE TO
efsa.europa.eu/en/news/newsletters
efsa.europa.eu/en/rss
Careers.efsa.europa.eu – job alerts

LISTEN TO OUR PODCAST
Science on the Menu –Spotify, Apple Podcast and YouTube

FOLLOW US ON TWITTER
@efsa_eu @methods_efsa
@plants_efsa @animals_efsa

FOLLOW US ON LINKEDIN
Linkedin.com/company/efsa

FOLLOW US ON INSTAGRAM
@one_healthenv_eu

CONTACT US
efsa.europa.eu/en/contact/askefsa

efsa
EUROPEAN FOOD SAFETY AUTHORITY