ISSN 1831-9424
Ref. Ares(2025)4669853 - 12/06/2025

European Commission

Sequencing strategies for the traceability of GMOs - methods and related quality aspects

*De Keersmaecker, S.C., *Jacchia, S., *Kok, E.J., *Roosens, N., *Zaoui, X., Angers-Loustau, A., Burns, M., De Loose, M., Dobnik, D., Keiss, N., Geuthner, A., Help, H., Hochegger, R., Lämke, J., Lee, D., Mazzara, M., Ovesna, J., Pallarz, S., Rolland, M., Savini, C., Schäfers, C., Simoes, F., Sowa, S., Wilkes, T.

*equal first-author contribution

European Network of GMO Laboratories

2025

Joint Research Centre

EUR 31987

Contact information
ENGL Secretariat
Food & Feed Compliance Unit (F.5)
Directorate F - Health and Food
European Commission - Joint Research Centre (JRC)
Via Enrico Fermi 2749. TP 201
I-21027 Ispra (VA), Italy

E-mail: JRC-ENGL-SECRETARIAT@ec.europa.eu

EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC137607

EUR 31987

PDF ISBN 978-92-68-18730-2 ISSN 1831-9424 doi:10.2760/890142 KJ-NA-31-987-EN-N

Luxembourg: Publications Office of the European Union, 2025

How to cite this report: De Keersmaecker, S.C., Jacchia, S., Kok, E.J., Roosens, N., Zaoui, X. et al., *Sequencing strategies for the traceability of GMOs - methods and related quality aspects*, Publications Office of the European Union, Luxembourg, 2025, https://data.europa.eu/doi/10.2760/890142, JRC137607.

Contents

Abstract

GMO analysis may increasingly rely on DNA sequencing, a technology that was originally used as a confirmatory step and is now considered for routine testing. However, in order to fully support the enforcement of European GMO legislation, this evolution requires harmonization, standardisation and validation of selected methods. In this context, this report from the European Network of GMO Laboratories (ENGL) seeks to evaluate the impact of sequencing technologies on GMO analysis, itself impacted by the advent of new genomic techniques. In addition, the report discusses quality criteria and good practice in the use of DNA sequencing data and hardware, with a focus on GMO-related aspects further illustrated by four realistic scenarios. Finally, by identifying areas requiring further development, the authors propose a set of recommendations towards the establishment of minimum performance parameters for GMO analyses including DNA sequencing techniques.

Foreword

Working group establishment

The Working Group (WG) on good practice/quality of DNA sequencing data was established at the 33rd meeting of the European Network of GMO Laboratories (ENGL) Steering Committee on 20-21st June 2017, based on a mandate adopted at the 28th Plenary Meeting of the ENGL on 20-21st September 2017. The WG was co-chaired by Nancy Roosens, Sciensano, Belgium and Sigrid C.J. De Keersmaecker, Sciensano, Belgium (2022-2023) and previously by Esther E.J. Kok, Wageningen Food Safety Research, Netherlands (2017-2021). The other members of the Working Group were Alexandre Angers-Loustau, European Commission, Directorate-General Joint Research Centre (DG JRC), Italy; Malcolm Burns, LGC, United Kingdom; Marc De Loose, Flanders research institute for agriculture, fisheries and food (ILVO), Belgium; David Dobnik, National Institute of Biology (NIB), Slovenia; Nina Keiß, Federal Office of Consumer Protection and Food Safety, Germany; Anne-Catrin Geuthner, State Office for Consumer Protection of Saxony-Anhalt, Germany; Hanna Help, Finnish Food Authority, Finland; Rupert Hochegger, Austrian Agency for Health and Food Safety (AGES), Austria; Sara Jacchia, European Commission, Directorate-General Joint Research Centre (DG JRC), Italy; Jörn Lämke, Federal Office of Consumer Protection and Food Safety, Germany; David Lee, HSE, Ireland; Marco Mazzara, European Commission, Directorate-General Joint Research Centre (DG JRC), Italy; Jaroslava Ovesna, Crop Research Institute, Czech Republic; Steffen Pallarz, Federal Office of Consumer Protection and Food Safety, Germany; Mathieu Rolland, ANSES, Plant Health Laboratory, France; Cristian Savini, European Commission, Directorate-General Joint Research Centre (DG JRC), Italy; Christian Schäfers, Hamburg Public Laboratories for Food Safety, Health Protection and Environmental Assessment, Germany; Fernanda Simoes, National Institute for Agricultural and Veterinary Research (INIAV) Portugal; Sławomir Sowa, Plant Breeding and Acclimatization Institute (IHAR) - National Research Institute, Poland; Timothy Wilkes, LGC, United Kingdom; Xavier Zaoui, European Commission, Directorate-General Joint Research Centre (DG JRC), Italy.

Background

GMO analysis, targeting GM plants, animals and microorganisms, is becoming more complex with questions that may relate to the detection, identification and quantification of EU-authorised GMOs and well-characterised unauthorised GMOs. Furthermore, the identification of unauthorised GMOs, for which only limited or fragmentary information is available so far, is also becoming important. In addition, new genomic techniques are increasingly being applied and the determination of the whole genome sequence (WGS) of these organisms may be required for detailed molecular characterisation.

In this context, the identification of DNA sequences by Sanger sequencing and massive parallel DNA sequencing techniques is gaining more importance. This may relate to the sequencing of:

- Polymerase Chain Reaction (PCR) amplicons;
- captured DNA fragments;
- enriched DNA populations;
- whole genomes, including plasmids.

Some of these sequencing approaches are already routinely applied, while others are only used in exceptional cases based on more experimental protocols. However, on a case-by-case basis, all these methods may contribute to the effective and cost-efficient detection and identification of all GMOs in a given sample. At the same time, quality aspects of DNA sequence data in the field of GMO detection and characterisation have so far received only limited attention.

Mandate and tasks

As part of the mandate from the ENGL Steering Committee, the WG was asked to draft guidance to:

- Assess the quality of sequencing data and of the results of sequencing strategies used for GMO detection and identification and molecular characterisation, as well as of the related data analysis workflow;

As a result, the following document was elaborated, providing an overview of the current application of DNA sequencing methodologies for the screening, detection and identification of GMO-related elements, constructs or GM events in single or complex samples. Additionally, this document provides guidance for the development of quality criteria to be followed and implemented for different scenarios deriving from the application of massive parallel DNA sequencing analysis to the screening for GMOs, authorised as well as unauthorised, with the aim of enabling a harmonised assessment of the resulting data.

- The WG was also asked to define minimum performance parameters (MPPs) and their associated acceptance values (AAVs) for sequencing-based analyses as well to define how sequencing information should be reported.

However, these objectives were deemed by the WG to be too ambitious compared to current scientific developments in the field of GMO analysis using DNA sequencing strategies. The current report can be taken as a basis for future work on more detailed minimum performance parameters and reporting formats for these methods to be applied in routine GMO analysis.

Scope

The scope of this report is two-fold: firstly, to evaluate the impact of ongoing developments in sequencing technologies for the quality and validation of related DNA-based methods for GMO detection, identification and, possibly, quantification developments with respect to the required quality of analytical results for the enforcement of European GMO legislation. Secondly, to concisely describe current developments in the area of GMO detection, identification and molecular characterisation with DNA sequencing technologies, with a focus on related quality aspects. The purpose of this exercise is to promote a harmonised assessment of the resulting data, and to identify the related areas that need to be developed and matured before minimum performance parameters and their associated acceptance values can be defined for methods for GMO analyses that include DNA sequencing steps.

Acknowledgements

The WG would like to thank Margriet Hokken (Wageningen Food Safety Research, Netherlands), Raniero Lorenzetti (Experimental Zooprophylactic Institute of Lazio and Tuscany, Italy), Theo Prins (Wageningen Food Safety Research, Netherlands), Martijn Staats (Wageningen Food Safety Research, Netherlands), Menno Van der Voort (Wageningen Food Safety Research, Netherlands) and Christopher Weidner (Federal Office of Consumer Protection and Food Safety, Germany) for their review and comments.

Executive summary

DNA sequencing in GMO analysis has undergone a rapid repositioning from a technology that may be used as an additional confirmatory step to a technique that deserves consideration as a routine methodology. The methodology may be applied in different ways, from initial screening for authorised as well as unauthorised GMOs in complex samples to, for specific cases, a strategy to establish the identity of possible new GMOs, as well as a range of options in between these two alternative applications. As the overall costs of DNA sequencing are reducing, broader applications are becoming feasible for routine testing in GMO laboratories. The actual application, however, requires further harmonisation and standardisation to maintain current quality standards when applying innovative GMO analytical strategies that include DNA sequencing steps.

In the present report, a summary is provided of the results of the dedicated Working Group of the European Network of GMO Laboratories (ENGL) on good practice/quality of DNA sequencing data. The Working Group has addressed both Sanger sequencing as well as massive parallel DNA sequencing within the framework of GMO analysis, with emphasis on the latter. General information is provided in relation to quality aspects of massive parallel DNA sequencing, from sample DNA extraction and preparation, through template amplification and DNA sequencing, to quality assessment of the sequencing results and quality criteria for the use of sequencing data analysis pipelines. In addition, specific GMO-related aspects of DNA sequencing methodologies are discussed considering current GMO detection and identification strategies in different sample types (i.e. simple versus complex food/feed matrices, known versus unknown GMOs), and the intended purpose (i.e. characterisation of the full genome of a genetically modified microorganism (GMM), the full identification of an inserted genetic element and its flanking regions, or the screening for multiple genetically modified (GM) elements in a single genome or in a mixture).

The Working Group identified four scenarios covering current real-life situations in GMO analysis strategies that include massive parallel DNA sequencing steps: two targeted sequencing approaches focusing on multiple known or partially known sequences, respectively, and two non-targeted sequencing approaches focusing on either whole genome sequencing of a single organism or, alternatively, applying metagenomics in complex samples. Quality considerations and criteria common to all scenarios have been tentatively established, as well as specific ones relevant for individual scenarios.

In this report it has been established that all aspects of DNA sequencing strategies for GMO detection and identification will require further harmonisation and standardisation, as well as the validation of selected methods. It is concluded that the determination of strict minimum performance parameters (MPPs) and their associated acceptance values is not (yet) possible. These parameters will be different from current performance parameters in the field of GMO analysis and will need to explicitly include performance parameters for both the molecular biological experimental part of the methodology, as well as for the related bioinformatics workflows. In addition, the increase in DNA sequencing output and the long-term storage and sharing of relevant DNA sequencing results in curated databases will require further consideration, and pragmatic solutions, as access to sequence information is a key factor in successful GMO analyses.

For the complete sequencing-based workflows and the bioinformatics workflows, it is necessary to further establish appropriate validation schemes to ensure accurate and reproducible analysis, either on the basis of *in-silico*, or real-life data. For GMO analysis laboratories, it may be beneficial

to establish shared bioinformatics workflows and a harmonised data management approach at the EU level, including criteria already established by ISO EN ISO 23418:2022 for methods for detection and identification of specific organisms that include DNA sequencing steps [1]. The increased availability of reference genomes will be of benefit in this respect. In the near future, Machine Learning approaches, such as support vector machines (SVM), artificial neural networks (ANN), k-means and others, may increasingly be applied to identify relevant (unauthorised) GMOs on the world market. The use of these advanced bioinformatics strategies for the purpose of GMO analysis is currently, however, still in its early developmental stages.

An important aspect of the use of DNA sequencing tools and strategies for GMO analysis within the frame of enforcing EU GMO regulations is the availability of DNA sequencing hardware and related bioinformatics infrastructure (hardware and expertise) for all official European GMO analysis laboratories. This will also involve the availability of adequate training facilities and training opportunities for all personnel involved. Training should include the discussion and implementation of all quality aspects that have already been established, as presented and discussed in the present report, and regular updates thereof.

On a global level, one harmonised definition of a GMO is no longer applicable, especially for genome-edited organisms that contain minor modifications, such as single nucleotide mutations. For these organisms, there is no international consensus on whether or not they should fall within the scope of the GMO regional legislations. In this context, global discussions on the safety aspects and the traceability of these organisms are affected and the exchange of information on the (potential) presence of GMO in food/feed samples and related raw materials has become hampered. The direct consequence of this divergence in legislation is that if no application in the EU is made, generally no sequence information will be available for those (genome-edited) organisms that are considered GMOs in Europe, but are not considered GMOs in other countries. This will generally affect the likelihood of detecting and identifying EU-unauthorised GMOs on the market. Finally, reference materials might not be available to establish validated methods for identification, which will affect GMO analysis strategies in general, including those that include DNA sequencing steps.

1. Introduction

In an enforcement context, DNA sequencing strategies have undergone a rapid repositioning in recent years, from a technology that may be used as a final confirmation step with relation to detailed issues at hand, to a technique that deserves consideration at each step of the analytical chain, from initial screening through to the final confirmatory analyses. At what point DNA sequencing analysis may be utilised largely depends on the question underlying the analysis. In some cases, for instance when referring to species identification in meat samples, DNA sequencing can often be considered conclusive in the initial screening phase. In other cases, the effectiveness of DNA sequencing in the initial screening step may increase as the quality of the resulting data also improves, thus becoming increasingly informative, in addition to a reduction in the costs of DNA sequencing.

Similarly, the world of GMO detection and identification is changing. Here, the demands are increasing with ongoing developments in plant and animal breeding and the development of new types of microbial strains to produce food and feed additives.

For conventional GMOs, it has been sufficient to be able to detect and identify the newly introduced, relatively large genetic elements that could be distinguished in most cases from the endogenous DNA. With the advent of new breeding techniques and especially more recently of targeted genome-editing techniques, hereafter referred to as new genomic techniques (NGTs), the detection and identification of new genetic elements that may be related to authorised or unauthorised GMOs has become much more challenging. Following the ruling of the Court of Justice of the European Union (CJEU) of 25 July 2018 case C-528/16 [a], organisms obtained by new methods of mutagenesis (e.g. oligo-directed mutagenesis (ODM) or various site directed nucleases like CRISPR (clustered regularly interspaced short palindromic repeats)) are GMOs subject to the GMO regulation. Therefore, the detection and identification of point mutations may be required in order to identify this new category of GMOs. GMO detection, identification and quantification is almost exclusively performed by the use of PCR-based methods in a two- to three- step approach. In the first step, a screening PCR analysis is performed on the basis of a limited number of genetic elements or constructs commonly used in various GMOs. Based on this initial analysis, it is determined which authorised GMOs, or well-characterised unauthorised GMOs, may be present in the sample. In the second and third steps additional quantitative PCR (qPCR) methods may, based on this initial screening, identify and/or quantify the GMOs present in the sample. In recent decades, however, the number of GMOs has gradually increased, leading to a considerable number of PCR analyses that need to be performed for a single sample. The number of GMOs which do not contain any of the common screening elements has increased too, contributing to increasing costs and complexity of the analysis. Another bottleneck of the current approach is that it primarily focuses on authorised GMOs and offers limited chances of identifying unknown unauthorised GMOs, that may be based on undocumented genomic rearrangements and structural variations in addition to point mutations [3].

[a] European Court of Justice, C-528/16 - Judgement of 25 July 2018. See: http://curia.europa.eu/juris/document/document.jsf?docid=204387&mode=req&pageIndex=1&dir=&occ=first&part=1&text=&doclang=EN&cid=515140.

With the increasing number of GMOs globally present in food, feed and seeds, and the expanding range of genetic elements that may be inserted or modified in different variants of the same elements, there is a clear tendency to look for methods that allow for a first screening step that targets a much broader range of known or suspected GMO elements. Ideally, this step would also provide information on the DNA sequence adjacent to the targeted elements, which may often be informative in relation to the presence of authorised versus unauthorised GMOs. With the advent of genome-edited organisms, there is now also an increased need to identify (a series of) shorter DNA mutations.

These developments have led scientists to increasingly look at DNA sequencing methodologies as additional or replacement tools to detect and identify GMOs. It is clear that, with current methodologies, it will not be feasible to use DNA sequencing data for quantification. This may, however, change in the future. Initially, the focus was on the use of traditional Sanger methods, which are based on the sequencing of one specific DNA fragment. High accuracy for read lengths of up to 1 kb can be achieved, but low throughput and relatively high costs per base makes Sanger sequencing mostly suitable for small scale projects. Over the last few decades, Sanger sequencing has been effectively applied for amplicon characterisation, and also in relation to GMO identification. More recently, its application has somewhat shifted, and Sanger sequencing is nowadays increasingly applied (also) for quality assurance in the validation of massive parallel DNA sequencing data. It is used for the analysis of the targeted PCR products to confirm sequences initially identified on the basis of massive parallel DNA sequencing bioinformatics.

Within the last few decades, massive parallel DNA sequencing strategies have been developed and are now already mainstream in some areas. Massive parallel DNA sequencing encompasses technologies providing both short and long sequencing reads. Short-read sequencing is highly accurate and produces read lengths of 50-300 bp. It uses an *in vitro* clonal amplification step to amplify individual DNA molecules, as their molecular detection methods are generally not sensitive enough for single-molecule sequencing [4,5]. Short-read sequencing has been widely used for various applications such as targeted amplicon sequencing, metagenomics, transcriptomics, and WGS. However, when complete genomes are required, and for determining complex genomic regions, longer reads are advantageous. Long-read sequencing systems are capable of generating reads from 10 to > 300 kb in length, but this is currently at the cost of high(er) sequencing error rates. Which technology may be applied primarily depends on what the sequencing data is to be used for. However, other considerations such as the required throughput of sequencing, in addition to more practical issues such as budget and time available, may influence the decision. Developments in the field of DNA sequencing are moving fast, and newer instruments may soon be available that overcome some of the current bottlenecks with a focus on throughput capacity, sequencing accuracy and costs. In Annex 1 to this report, an informative overview of currently available DNA sequencing techniques, with details on massive parallel DNA sequencing systems available at the time of writing of the present report, is provided. Similar to the sequencing technology itself, the software tools for the analyses of the resulting DNA sequences have evolved. Data analysis for GMO detection and identification may be based on the alignment of the obtained data to reference sequences, or on *de novo* assembly. Here, the computational demand (both storage and computing power) for analysing the obtained data varies greatly and might present an obstacle for those wishing to utilize massive parallel DNA sequencing to its full capacity. Current software tools are increasingly able to identify variants in matrices comprising single organisms, where they already, in specific cases, allow the distinction between authorised and unauthorised GMOs. In complex

matrices that may entail the presence of multiple varieties or cultivars within an ingredient of a certain crop species, for instance soy or maize, this is still largely a promise for the future, but the tools are gradually improving here as well to help scientists to optimally mine the information in the massive parallel DNA sequencing data.

Given the wide range of sequencing strategies currently available, the use of different technologies (instrumentation/platforms) and methodologies (e.g. targeted resequencing or *de novo* assembly, whole genome (re)sequencing)) may influence the comparability of results. Similarly, the lack of harmonised terminology and approaches used for data analysis may also have an impact on the use of these new and powerful technologies in official controls. In light of these considerations, stakeholders from relevant fields of biology have instigated a number of initiatives to address the problem of harmonisation and data comparability of DNA-based technologies. In addition, and within the context of accreditation, analysis and reporting will need to be performed in accordance with prevailing quality management systems. However, so far, initiatives of harmonisation/standardisation and quality assurance in the field of GMO detection and identification have primarily covered guidance on the establishment of MPPs and AAVs for method development of other DNA-based technologies, with a main focus on qPCR. Comparatively, limited guidance has been formulated that specifically relates to massive parallel DNA sequencing-based methodologies for GMO analysis. In other application fields, such as in clinical settings, or for the characterization of foodborne bacteria or microbial pathogens, there are already, or soon to be published, well-established guidelines for harmonisation of massive parallel DNA sequencing quality standards [1,6,7,8].

At the time of writing, a thorough evaluation of the impact of ongoing developments in sequencing technologies on the quality and validation of related DNA-based methods for GMO detection, identification and, possibly, quantification had not been undertaken. The objective of the present report is to evaluate these developments with respect to the required quality of analytical results for the enforcement of European GMO legislation. In this report, a summary is provided of the results of the dedicated Working Group of the ENGL on good practice/quality of DNA sequencing data.

Report Organisation

This report provides an overview of DNA-sequencing-based scenarios for the detection and identification of GMOs (plants, animals and micro-organisms). This includes the DNA sequencing of amplified genomic fragments or the entire genome, as well as related quality aspects.

In section 3, information is provided on quality aspects related to Sanger sequencing. This includes the quality of the Sanger sequence reads generated, as well as the subsequent data analysis workflow resulting in the Sanger consensus sequences.

In section 4, general information is provided in relation to quality aspects of massive parallel DNA sequencing, from sample DNA extraction and preparation, through template amplification, and DNA sequencing, to quality assessment of the sequencing results and quality parameters for the use of sequencing data analysis pipelines. As most of the current scientific literature on sequencing is not tailored to GMO detection specifically, but is rather focused on applications in other fields, including microbiology, taxonomy, or clinical settings, reference will be made to these fields, where appropriate. Relevant quality parameters of massive parallel DNA sequencing that have been

identified in these fields may also be considered for use in the field of GMO traceability, if this is not yet the case.

In section 5, specific GMO-related aspects of DNA sequencing methodologies are discussed in depth. These discussions are based on a number of different scenarios, which reflect current GMO detection and identification strategies in single and complex products. These scenarios are the identification of authorised or well-characterised unauthorised GMOs, the detection and (initial) characterisation of unknown or of partially known unauthorised GMOs in complex products. Appropriate sequencing methodologies may range from targeted procedures to whole genome sequencing, depending on the specific product and related research questions, and may range from general screening procedures to targeted confirmatory methods.

Finally, in section 6, the overview presented in sections 3 to 5 is discussed, and an outlook on further developments in the field of DNA sequencing methodologies and related quality assurance issues will be presented. The main conclusions and recommendations presented in this report will also be summarised.

This report thus provides an overview of the current application of DNA sequencing methodologies for the screening, detection and identification of GMO-related elements in single or complex samples. Massive parallel DNA sequencing analysis may be of increased value in the broader screening for GMOs, authorised as well as unauthorised, in the near future. To allow the application of these innovative technologies, it is necessary that agreed quality criteria enable a harmonised assessment of the resulting data. This report is a first attempt to concisely describe current developments in this area with a focus on related quality aspects. Given the rapid progress in the related massive parallel DNA sequencing technologies, future reports will be necessary to monitor the developments and to adjust the quality criteria accordingly.

2. Glossary

Adapter sequence (EN ISO 23418:2022)

Oligonucleotides of a known sequence that are ligated to each end of a DNA/cDNA fragment to facilitate the sequencing process (e.g., annealing to a flow cell).

Annotation (adapted from EN ISO 23418:2022)

The process of identifying genes and other genomic features, their functions in sequence data and the prediction of their function.

Assembly (also Genome assembly) (adapted from EN ISO 23418:2022)

Output from process of aligning and merging sequencing reads into larger contiguous sequences (contigs/scaffolds). Reads may be assembled either *de novo,* without *a priori* sequence information, or reference-based, where reads are mapped to a reference genome by different means.

Binary alignment format (BAM file) (ISO 20397-2)

Compressed format analogous to the SAM format in binary form.

Barcode (or Index) (EN ISO 23418:2022, ISO 20397-1:2022)

Short sequences of typically 6 or more nucleotides used in the process of sequencing library preparation to tag and identify DNA from specific samples, so that multiple samples may be combined (multiplexed) in a sequencing reaction.

Base calling (EN ISO 23418:2022)

The process of assigning nucleotides and quality scores to positions in sequencing reads.

Depending upon the sequencing technology, the assignment may be based on analysis of signals

that includes changes in fluorescence or electrical current.

Bioinformatics (EN ISO 23418:2022)

In relation to this report, this refers to the collection, storage, and analysis of nucleotide sequence data.

Bioinformatics pipeline (EN ISO 23418:2022)

Individual programs, scripts, or pieces of software linked together, where the output of one program is used as input for the next step in data processing. For example, the output from a read trimming program may be used as input to a *de novo* assembler.

Compressed reference-oriented alignment map (CRAM) (ISO 20397-2)

A sequencing read file format that is space efficient by using reference-based compression of sequence data and offers both lossless and lossy modes of compression.

Contig (EN ISO 23418:2022)

A contiguous stretch of DNA sequence that results from the assembly of shorter, overlapping DNA sequencing reads.

Coverage depth (or estimated coverage or theoretical coverage) (adapted from EN ISO 23418:2022)

In relation to this report, this refers to the predicted amount each base in the genome is sequenced. Coverage depth is calculated by dividing the total number of sequenced bases by the expected size of the genome. The coverage is usually expressed with the term "fold", for example, 60-fold means that on average each base in the genome has been sequenced 60 times.

The intended coverage of a sample depends on the aim of the experiment.

Deletion (adapted from ISO 20397-2)

Change in DNA sequence where a part of a chromosome or a sequence of DNA is lost.

FASTQ (adapted from ISO 20397-2)

Can be used as a *de facto* standard format for downstream analysis of the quality of massive parallel DNA sequencing data sets. FASTQ is widely accepted as a cross platform interchange file format. It includes a name and a nucleotide sequence for each sequence read and the corresponding quality values.

Indel (adapted from ISO 20397-2)

Insertion or /and deletion of nucleotides in genomic DNA. Indels are between one and 1000 bases in length.

Insertion (ISO 20397-2)

Addition of one or more nucleotide base pairs into a DNA sequence.

Library (adapted from ISO 20397-1:2022)

Collection of DNA fragments of a defined size range and quality with sequencing adapters attached during library preparation that can be run on a sequencer.

Mapping (adapted from EN ISO 23418:2022)

The process of aligning (or mapping) reads to a reference sequence.

Massive parallel DNA sequencing (adapted from ISO 20397-2)

Defined as a non-Sanger-based high-throughput DNA sequencing technology, where millions or billions of DNA strands can be sequenced in parallel. Includes next (second, third and future) generation sequencing (NGS).

Metagenomics [2]

Also referred to as environmental or community genomics, is the genomic analysis of a mixture of genomes in a sample, e.g. identification of microorganisms by direct extraction of whole genome DNA from a sample of microorganisms.

N50 (also N50 length) (adapted from EN ISO 23418:2022)

The largest contig length L such that 50% of all nucleotides are contained in contigs or scaffolds of at least length L. The N50 is one parameter for the assembly quality. To compare N50 values from different assemblies, the genomic context should be taken into account.

Quality score (Phred score, Q score, QV-score) (adapted from EN ISO 23418:2022)

A measure of the probability that a base is incorrectly assigned at a given position in the sequence. Phred scores are logarithmically related to the base-calling error probabilities. The Phred score can be expressed with the following formula, where P is the probability that a base was wrongly called (the higher P, the lower the Q score): $Q = -10 \log_{10} P$. A Phred score of 30 corresponds to a chance that this base is called incorrectly of 1 in 1000, or 99.9 % chance of a base being correct.

Raw data (adapted from ISO 20397-2)

Primary sequencing data produced by a sequencer usually containing the read sequences and their associated quality scores.

Read (EN ISO 23418:2022)

DNA sequence inferred from a fragment of genomic DNA or cDNA.

Reference sequence (adapted from ISO 20397-2)

A high-quality nucleotide sequence which defines the default state of a sequence. It is used to describe variations or to measure the quality of highly similar sequences.

Run (ISO 20397-2)

A single process cycle of the sequencer from initiation/library loading until the complete raw data is obtained.

Sequence alignment format (SAM file) (ISO 20397-2)

A TAB-delimited text format consisting of a header section, which is optional, and an alignment section. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position and variable number of optional fields for flexible or aligner specific information.

Single nucleotide variant (SNV) (adapted from ISO 20397-2)

A variation of a single nucleotide that occurs at a specific position in the genome relative to the reference sequence.

Unmapped BAM format (uBAM) (ISO 20397-2)

Variant form of the BAM file format in which the read data does not contain mapping information.

Variant calling format (VCF) (adapted from ISO 20397-2)

Data file for sequence variants with called variants annotated using an appropriate specification containing meta-information, a header line, and data lines that each contain information about a position in the genome and genotype information on samples for each position. BCF, or the binary variant call format, is the binary version of VCF.

Whole genome sequencing (WGS) (adapted from EN ISO 23418:2022)

The process of determining the DNA sequence of an organism's genome using total genomic DNA as input for sequencing.

3. Quality aspects related to methods based on Sanger DNA sequencing

3.1. Sanger sequencing in the GMO field

In the field of GMO testing, Sanger sequencing [9] is applied for the molecular characterisation of the DNA inserts both i) in case the GMO is known but its sequence should be determined with high accuracy or ii) when the available sequence information of the insert is only partially known or totally unknown but nearby DNA regions are disclosed. Other uses are also possible, such as for the detection of specific single nucleotide variants (SNVs) in the context of NGTs.

Sanger sequencing is based on the primer extension of a complementary strand from a fixed primer point in the sequence. The sequence determination depends on the use of complementary terminator bases, labelled with a fluorophore, that stop extension and produce extended fragments that are resolved in a capillary electrophoresis (CE). A coupled fluorescence detector allows for the identification of the terminal base of each generated extended fragment.

Determination of known sequences from GM events is easily achieved by direct sequencing of the specific DNA fragment conventionally amplified either by the same PCR used for GMO detection or by cloning the amplicon of interest in a plasmid followed by Sanger sequencing.

When suspecting the occurrence of DNA inserts in an organism with unknown sequences, the determination of unknown sequences relies on devised amplification strategies mainly based on DNA digestion, adapter ligation and PCR. Amplification performed using the GM event-specific primer and adapter complementary primers will generate amplicons covering the flanking sequences of the GM event. Further Sanger sequencing of these amplicons can then reveal unknown events, surrounding a known GM insert. These “anchored” amplification strategies were applied successfully on the determination of insertion junction sequences as shown for a transgenic Roundup Ready soybean line (event GTS 40-3-2) [10] or genetically modified maize, MON863, where disclosed sequences lead to the design of new detection methods for the detection of GM events [11,12]. In case of templates from DNA walking amplicons, Sanger sequencing allowed the confirmation of the presence of GMOs and to discriminate EU-authorized and unauthorised GMOs when few GMOs were present in a sample [13,14]. However, for more complex matrices with a mix of different GMOs, Sanger sequencing seemed to be inefficient, and other methods for multiplex sequencing were more successful [15].

Finally, Sanger sequencing can also be used as a confirmatory tool when DNA variation is deduced by massive parallel DNA sequencing methods [16].

3.2. Quality aspects of cycle sequencing reactions

Sanger sequencing is based on individual cycle sequencing reactions extending from each sequencing strand primer. For Sanger sequencing, the presence of unique templates is a major factor for base calling quality and sequence noise background. However, other critical issues may impact the quality of the DNA sequence to be determined and should be addressed:

Type of templates:

Plasmid templates - DNA purity of the plasmid to be sequenced is crucial and should be determined by checking the OD 260/230 and 260/280 ratios. Ideally the OD 260/280 should be comprised between 1.75 and 2.05 [1]. More general information on DNA extraction procedures can be found in the ENGL document 'DNA extraction from food/feed samples for GMO analysis' [b].

Amplicon quantity and quality - Amplicons should be first checked on a gel or capillary electrophoresis system. If only one amplicon band is detected, it can be directly purified. If more than one band is detected, the band of interest must be excised from the gel and further purified.

Amplicon length - For amplicons exceeding 900-1000 bp, additional internal primers for fragment sequencing should be considered. Shorter amplicons, or amplicons in which more accuracy close to the primer site is needed, are usually sequenced using adapted sequencing mixes.

Difficult templates - Hairpin structures, G:C rich regions or nucleotide repeats are often difficult regions to sequence efficiently. For hairpin structures and G:C rich regions, the quality of the sequencing read can be poor or no reads are available. Plasmid cloning can be an alternative strategy to reach the desired level of depth of coverage (the number of times a single nucleotide is read) to assign the correct sequence. For nucleotide repeats there could be variability in the length/number of repeats present in the returned sequences.

Sequencing primers:

Primers should preferentially be non-degenerated, 100% homologous to the target region and attention should be paid to primer length and melting temperature (Tm). For example, Tm could be comprised between 52 °C and 60 °C and the length between 18 and 25 bp.

Sequencing Reaction Purification*:*

Purification of generated single chain fragments prior to capillary electrophoresis could interfere with the resolution on capillary electrophoresis. Samples containing salts from insufficient purification of templates, PCR products, or sequencing reactions interfere with proper electro kinetic injection. The decision on the purification method used depends on the time available and/or the amount of original template DNA.

3.3. Quality aspects of the generated reads

The following considerations assume that the sequencing runs were completed, and data files were correctly managed by the software instrument as well as being correctly assessed and displayed.

Factors with an impact on raw sequence data such as artefacts, peak heights and resolution, length of reads and background noise should be considered when assigning the correct nucleotide per position. For instance, excessive amplification of non-target DNA (e.g. viral DNA from CaMV) can interfere with the detection of targeted DNA. The degree of interference from non-targeted DNA can be deduced from the electropherogram. Electropherograms produced by capillary electrophoresis are generated from the automatic base calling, according to device´s defined

[b] In preparation

quality criteria. Quality score (Q score) or Phred score [17] is a per-base estimate of the base caller accuracy. It is represented in the electropherogram in colour code (Q score 0 to 9 - unreliable data; Q score 10 to 20 - probably reliable data; Q score > 20 - reliable data). The signal to noise ratio is also an important parameter to be considered in case of nucleotide bases overlapping. High quality data normally yields a signal to noise ratio >100, although accurate base calling can be achieved with values as low as 25 [18]. Electropherogram analysis should evaluate the quality and accuracy of base calling and determine whether a manual correction is necessary.

Sequence and peak analysis of known control sequences (plasmid or synthetic cloned sequences) following the sequencing routine protocol should be periodically performed to monitor data analysis parameters and to assure the optimal quality of sequencing or even to distinguish between chemistry and instrument problems.

3.4. Quality aspects related to the data analysis

The JRC "*Guideline for the submission of DNA sequences derived from genetically modified organisms and associated annotations within the framework of Directive 2001/18/EC and Regulation (EC) No 1829/2003*" [19] and the EFSA "*Technical Note on the quality of DNA sequencing for the molecular characterisation of genetically modified plants*" [20] represent two examples of regulatory documents. Therein, requirements and recommendations on the information to be submitted when sequencing is used for the characterisation of the GMO insert sequence are specified. Basic requirements for Sanger-derived sequences are i) obtained sequence of the GMO insert results from at least two independent PCR amplification reactions; ii) each nucleotide base should be sequenced on the two DNA strands resulting in a final coverage of at least 4 times.

After producing bi-directional fragment sequences (forward and reverse strand), a consensus sequence should be generated by overlapping the common sections (contig) using basic bioinformatics software. A consensus sequence may also contain flanking non-overlapping sequences when two independent sequencing runs were obtained. To produce a consensus sequence, the trimming of all sequencing primer sequences is advisable. Nucleotide bases in the consensus sequence with QV <20 should be noted as N (undefined nucleotide). Finally, the consensus sequence should be prepared to be shown in 5'-3' direction. When overlapping peaks are observed with a good QV, they must be given the ambiguity code according to the International Union of Pure and Applied Chemistry (IUPAC). Any manual editing performed on the sequence (base calling and trimming) should be reported and justified.

For the submission of Sanger sequencing data, guidance on the reporting format is largely available in JRC and EFSA guidances mentioned above. This entails a detailed description of the samples and materials used as the basis of the Sanger sequencing analysis, a confirmation that the samples will be stored for future analysis, a detailed description of the experimental procedure and the sequencing method(s) applied, as well as a full report on the conducted bioinformatics analysis of the sequencing data. The latter will include a detailed description of the used software and tools, including names, versions, selected options and parameters used.

4. Quality aspects related to massive parallel DNA sequencing: general considerations

4.1. Introduction

Although there are many different massive parallel DNA sequencing technologies, these may all be categorised, amongst others, on the basis of the sequence length they typically generate. As a consequence of read length, all current platforms may be classified as belonging to one of two broad groups. These can either be short or long read sequencers, both of which are subject to different advantages and limitations for the purposes of data generation and analysis, see also Annex 1.

Regardless of read length, all current sequencing platforms are subject to a certain level of sequencing error, which is commonly expressed in terms of Phred scores, or QV scores. This metric provides an indication of the probability of a sequenced base being called correctly. In general, Phred scores above 20 (corresponding to a 99 % chance of a base being correct) are normally considered acceptable for short-read sequencing applications [17]. However, this threshold may be adjusted to the specific sequencing approach (e.g. for Illumina sequencing, a Phred score of 30 is widely used [c]). For long read sequencing, the average Phred score is usually lower, but has recently increased (e.g. 7-21 for Oxford Nanopore Technologies [21,22]). However, depending on the run mode being used, the Phred score 30 value may also be reached by long read sequencing platforms, *i.e.* running a high-output mode with a high coverage on the platform in use. More detailed information on the individual characteristics of the different platforms can be found in Annex 1.

The generation of sequencing data may be viewed as a consecutive workflow that consists of the steps outlined in Figure 1. A number of general considerations for each of these steps are outlined in the following sections. Aspects specific to GMO analysis have been addressed in the next part, section 5.

Figure 1. Components of a massive parallel DNA sequencing workflow

Source: ENGL

4.2. Extraction and quality of DNA to be used for sequencing

The aim of this step is to provide DNA of suitable quality and quantity for subsequent library preparation as massive parallel DNA sequencing is dependent on the quality of the DNA used, i.e. on the length, structural integrity and physical-chemical purity of the extracted DNA. Depending on the type of platform used (short read versus long read) and the application (direct DNA sequencing, PCR amplicon sequencing), particular attention needs to be paid to the samples' concentration as well as to DNA integrity.

[c] Illumina, I. (2011). Quality scores for next-generation sequencing. Technical Note: Informatics, 31.

In particular, large quantities of high molecular weight genomic DNA samples must be obtained and further handled with great care when long read sequencing is envisaged.

DNA purity may be assessed by a number of different means including the spectrophotometrically measured OD 260/280 ratio (which is recommended to be between 1.75 and 2.05) [1] and the 260 nm / 230 nm ratio (which is recommended to be between 2.0 and 2.2). In order to accurately determine the DNA quantity, fluorometric methods using DNA intercalating dyes should be used in order to quantitate the amount of intact, double stranded DNA. The DNA integrity (i.e. size distribution of extracted DNA) may be assessed by analysing the average fragment size distribution using capillary electrophoresis-based instruments capable of measuring high-molecular weight DNA, or alternatively, through the use of pulse field electrophoresis. In addition, if the workflow includes a step to enrich the sequence in particular targets (sequence capture approach, targeted PCR, DNA walking, etc.), it is critical that experimental procedures and primers and/or probe design enable the full representation of the targets aimed for.

Similar to other GMO analysis applications, it may sometimes be challenging to extract qualitatively and quantitatively sufficient DNA for massive parallel DNA sequencing analysis. When implementing massive parallel DNA sequencing or when testing new methodologies, it is therefore advisable to use samples of DNA that have been extracted from well-defined materials for which details regarding composition are available. This may either be well defined real-life samples or, if available, specific reference standards [23].

During extraction, the use of a no template extraction control (NTC) is suggested to check for contamination. When working under quality management standards, documentation defining the specific requirements should be in place, examples of which have been described by Hendriksen and colleagues [24]. The DNA extraction procedure, the possible enrichment steps, the methods to assess the purity, concentration and integrity of the DNA should be accurately described, together with their expected performance criteria.

4.3. General considerations for library preparation and sequencing

Next-generation sequencing involves three fundamental steps, starting from the extracted DNA. These are: library preparation, library sequencing, and data analysis (Figure 1). Library preparation is the critical first step in the massive parallel DNA sequencing workflow. This step prepares the DNA sample to be compatible with the sequencing platform employed. Depending on the platform used and the goal envisaged, different requirements need to be met and different protocols may be used to create a sequencing library. General criteria for library quality assessment include: DNA concentration, insert size distribution and assessment of possible contamination. The concentration of the library (determined fluorometrically) and its average fragment size distribution should be determined, taking into account sample characteristics (e.g. genome size, DNA amplicon abundancy, etc.) and the massive parallel DNA sequencing platform that will be used. Further details for sequencing library quality control criteria are listed in Annex A of EN ISO 23418:2022 [1]. Additional key points to consider include the importance of not over- or underloading an instrument when seeding flow cells/sequencing chips. Hence, library quantification needs to be carefully and precisely performed. Particularly when performing paired end sequencing, a narrow insert size distribution of the initial library can give important additional information when assembling obtained data, and is hence advisable to control for this aspect. Considering that shorter DNA fragments bind more easily to the flow-cell of the sequencer than long fragments (which might lead to low output of sequencing data), in case of partly degraded samples with large DNA size distribution, one could consider specific measures to select for longer DNA strands prior to library construction.

When using a targeted approach (to only target specific genomic regions), multiplex sequencing is useful to reduce the cost and the time of the massive parallel DNA sequencing analysis. Individual "barcode" (i.e. indexed) sequences are added to each DNA fragment during library preparation so that each read can be identified and assigned to a specific sample during the primary data analysis before starting the secondary data analysis. Depending on the genome size of the species present in the sample and on the required coverage, multiplexing can also be used for whole genome sequencing. Normalisation of barcoded library DNA quantities is essential to ensure that all individual samples achieve adequate coverage even in a multiplex approach.

The use of appropriate positive and negative controls is also an important aspect of library preparation. Guidance on recommended use of controls during library preparation has been given in Annex A of the EN ISO 23418:2022 document [1]. Controls for the sequencing process itself commonly take the form of—spike-in controls, such as PhiX-DNA [d] that is used for Illumina instruments. Hendriksen and colleagues [24] provide practical recommendations for library preparation when using Illumina technology, including normalization, size control, and controls to ensure adequate sequence coverage.

As the different steps of the library preparation workflow may provide clues to help solve any problems encountered with a specific library, information on the library preparation method used, along with any other informative metrics (e.g. % of DNA with the expected size, % of adapter dimers), should be carefully documented. Producing a DNA sequencing library that satisfies the quality control criteria presented here would be technically challenging if the input DNA were not of a high quality.

The type of sequencing platform used and the number of samples to be pooled in one sequencing run should be carefully evaluated, depending on the intended purpose (e.g. complexity of mixture, size of genome, expected output of instrument, required coverage or read depth, required accuracy). Following suitable library generation and use in a sequencing run on an instrument, the run data should be carefully evaluated and recorded. Platform-specific sequencing metrics (e.g. cluster density, number of reads, average base quality, etc.) should be evaluated for each sequencing run to guarantee its quality. EN ISO 23418:2022 [1] provides run acceptance parameters for Illumina instruments (QV 30 coverage, PhiX error rate, reads passing filter and negative control results). Hendriksen and colleagues (Appendix B) [24] provide some practical recommendations for Illumina and IonTorrent sequencing platforms as well as on-machine quality metrics calculation. In addition, a broad list of quality control metrics for massive parallel DNA sequencing analyses (as developed for pathogen detection) have been compiled and discussed by Schlaberg and colleagues [25], including a checklist for the quality control of associated wet bench processes (supplementary table 2).

4.4. Primary data analysis: base calling

The raw signals generated by the sequencing instrument are translated into base calls and ultimately into nucleotide sequences or “reads”. Primary data analysis may be defined as the machine-specific steps required to call bases and compute quality scores for those calls. This typically results in a FASTQ or uBAM file with reads of a few hundred bases for short read platforms (such as Illumina) and several thousand bases for long read platforms (e.g. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)). The FASTQ/uBAM file also contains both the actual sequence data: A, C, G and T bases, or N for undefined positions, and an associated Phred score for each of those bases.

[d] The PhiX Control is derived from the small, well-characterized bacteriophage genome, PhiX. It is a concentrated Illumina library (10 nM in 10 µl) that has an average size of 500 bp and consists of balanced base composition at ~45 % GC and ~55 % AT. https://support.illumina.com/bulletins/2017/02/what-is-the-phix-control-v3-library-and-what-is-its-function-in-.html

It is imperative that appropriate quality checks are performed for the raw sequencing reads. Example quality checks include using FASTQC, assessing insert size, sequence length distribution, number of reads, average Phred score, average read length, coverage and assessment of base composition. For the latter, the AT/GC balance or TAGC (taxon annotated GC-coverage should be considered, etc.).

Primary analysis frequently includes the demultiplexing of multiple samples which have been indexed and pooled into a single sequencing run (in order to allow the pooling of multiple samples prior to sequencing). Each of the reads is assigned to a specific sample, based on the incorporated barcode that is also sequenced. To improve secondary analysis (see 6.5), read quality is often increased by removing (or trimming) low-quality bases from the beginning and ends of reads, and filtering out reads of insufficient quality.

A number of different approaches may be employed to qualify the performance of a sequencing run such as described in Annex A in EN ISO 23418:2022 [1]. This annex contains a summary of general information for sequencing analyses, covering multiplexed sample normalization, identification and estimation of inter-run carryover, and instrument performance as well as critical criteria for sequence data quality, read length and coverage.

Information on quality control and pre-processing for raw data is also provided in ISO 20397-2:2021 [8]. Importantly, raw data files should be generated using instrument-specific software, and should report individual sequencing reads as well as the quality score for each nucleotide. Several quality metrics are also proposed, such as: length distribution, quality score per base and average quality score per sequence, sequence duplication level, adapter sequence contamination. In general, however, relevant quality metrics may differ depending on the specific sequencing platform.

In a further publication, Hendriksen and colleagues [24](Appendix B) provide practical recommendations for Illumina and IonTorrent sequencing platforms for post-sequencing data processing, including demultiplexing, quality control, and data analysis.

In addition to run acceptance parameters (see above), Annex A of EN ISO 23418:2022 [1] also provides criteria for governing sample acceptance (i.e. estimated coverage, Phred score, contamination check).

As a concluding remark, metrics to evaluate the quality of raw sequencing reads should always be documented. Phred scores and raw read numbers for each sequencing run should be described. The removal of poor-quality reads and trimming of sequencing adapters or low-quality ends of reads is recommended to obtain a dataset of acceptable quality. The quality control implemented strategy should be clearly described (e.g., which tools are used to check and trim reads, and which threshold values are used for quality control checks such as Phred score etc. that are considered acceptable).

The output from the primary data analysis (base calling) should then be ready for processing in a secondary data analysis pipeline.

4.5. Secondary and tertiary data analyses

Once the raw sequence data are obtained from the massive parallel DNA sequencing instrument, secondary data analysis may commence. This is normally a computationally intensive step that is accomplished through the execution of a series of publicly and/or commercially available or in-house developed pipelines that enable the goal of the particular analysis (i.e. its intended purpose). This process consists of the secondary data analysis - typically considered as the steps that process data and can be semi-automated in order to transform read data into a format that can be used for further analysis and interpretation, including, for instance, additional read trimming, read mapping and/or assembly. This is then followed by a tertiary analysis of the data (i.e. the 'sense-making' that is required in order to address the specific question being asked).

The read depth, particularly when short read technologies are used, is often a key factor in the evaluation of the quality of the data, and to guarantee that the data processed through the bioinformatics pipelines will meet the acceptance criteria's thresholds. As such, it should be defined early on to be sufficient for the intended purpose (see section 4.3). Therefore, for each pipeline, if feasible, an acceptance threshold for read depth needs to be specified. Additionally, it should be checked that the obtained read depth is in accordance with the criteria defined before the analysis. In addition, the detailed information on the read depth over each specific position, the average read depth, and the coverage uniformity along the genome in case of WGS needs to be included in the results generated by any particular pipeline.

Parameters for the quality assessment of secondary as well as tertiary massive parallel DNA sequencing analyses are also listed in Annex G of EN ISO 23418:2022 [1], and include the N50, sequencing depth, breadth of coverage, mean contig length, number of contigs, and the size of the assembled genome.

ISO 20397-2:2021 [8] provides information on sequence alignment (file format, quality control metrics, post-processing) and variant calling (data file, quality control metrics, processing of false positive variants, annotation). Sequence alignment requires a proper reference genome that should be chosen according to the experiment (e.g. masked/unmasked genomes, i.e. reference genomes with nucleotide stretches being replaced by 'n' in order to suppress the alignment to repetitive or unwanted regions, of parental line or strain genotype). The aligned reads should be stored in BAM, SAM or CRAM format and relevant quality metrics should be provided (e.g. mapping rate, reads with multiple hits, properly mapped mate pairs for paired-end sequencing, and coverage statistics). Regarding the variant calling, results should be provided in VCF/BCF format, which reports for each variant relevant annotations such as variant quality, allelic frequency, and strand bias. All variant calling parameters and threshold should be clearly indicated.

The bioinformatics pipeline that is used in the end-to-end process should be validated. In other scientific fields, primarily microbial genomics, information on the validation of bioinformatics pipelines is available, such as that described in ISO 20397-2:2021 [8] and EN ISO 23418:2022 [1]. A wealth of useful considerations for the validation of the bioinformatics of massive parallel DNA sequencing regarding pathogen detection are referenced in Schlaberg *et al.* [25] (2017), including a quality control checklist for the bioinformatics process (supplementary table 1). EN ISO 23418:2022 [1] (field: foodborne bacteria) provides information on test data to verify that bioinformatics pipelines are installed correctly and function as expected; are included standard (or benchmark) data sets (sequence data that has been made publicly available) as well as sample data and simulated data-applications (synthetic sequence read data from real genome sequence data). Lambert and colleagues [26] have proposed a list of guidelines for the generation of reliable genomic data, with a focus on WGS applications for microbiological food safety testing. They emphasize the importance of a systematic application of quality criteria for both the utilization of bioinformatics analyses and reporting of WGS data, as well as addressing the generation and standardization of metadata, good laboratory practice, data processing, management, interpretation and reporting of data.

If any component of the underlying bioinformatics workflow is updated (e.g. a software update or bug fix), revalidation will normally be necessary, but will be dependent on which components were affected (a change to the underlying algorithms for data processing warrants a revalidation whereas, for instance, a simple change in the layout of the program or output report does not necessarily warrant a revalidation). It is important that a log be kept of all steps in the data analysis pipeline, including the version of the bioinformatics workflow applied.

Designing benchmark strategies for the bioinformatics pipelines of massive parallel DNA sequencing technologies is challenging, but this is a mandatory requirement for the correct evaluation, validation and quality control of the bioinformatics component of the process [27].

To date, a limited number of benchmarking studies on massive parallel DNA sequencing have been undertaken. Angers-Loustau and colleagues [27] have developed a benchmark strategy for the evaluation of bioinformatics pipelines that transforms a set of massive parallel DNA sequencing reads to a characterised antimicrobial resistance (AMR) profile. Bogaerts and colleagues have outlined a validation strategy of the bioinformatics analysis of a bacterial whole genome sequencing workflow, in the context of pathogen characterization [28–30]. Hendriksen and colleagues [24] have also provided bioinformatics benchmarking exercises of different *de novo* assembly tools for microbial genomics. This publication also lists and discusses the merits of relevant and commonly used bioinformatics software and tools, which can also be used for capacity building. These include tools for quality assessment, trimming, assembly, annotation, alignment or sequence searching, mapping, assembly refinement, assembly statistics and quality assessment, variant calling, or comparative genomics, all with a focus on microbial genomics.

To define suitable thresholds for these quality control parameters, several components of the entire sequencing process need to be considered. These include: the characteristics of the genome (GC content, etc.), the sequencing approach used (targeted versus WGS), and the sequencing technology used.

4.6. Best practices to ensure overall quality of the bioinformatics pipelines

For the development of bioinformatics pipelines, it is advised that certain 'best practices' are followed to ensure the overall quality of the pipeline. Such 'best practices' do not represent a strict set of rigid rules that need to be followed at all times, but rather represent a set of recommendations that facilitate the life cycle of bioinformatics pipelines and, therefore, indirectly also positively affect the quality of the end product.

Software program code versioning is the process by which a unique id is assigned to a currently existing version of the program code. This has several advantages. Firstly, it allows a complete and long-term change history of every file that is part of the code base, allowing the tracing of all of the changes that have been applied to a bioinformatics pipeline. Secondly, it enables 'branching', which is the process by which different versions of the same code base are created. This allows for the creation of changes to the bioinformatics workflow on different branches, which can be intensively tested before being released without affecting the functionality of the main workflow. Thirdly, it facilitates collaboration between team members employed on the same code base because each can be working independently and merge the different changes afterwards.

It is recommended that some form of disaster recovery policy is established such that, if one individual computer fails, the central copy can still be restored.

In addition, code conventions should be adhered to, since they promote the standardization of the structure and coding style and therefore have an indirect positive effect on the quality of the bioinformatics workflow. By adhering to conventions, the code may also be more easily read and understood by other programmers because more intuitive, precise and unambiguous source code is easier to maintain and also debug. One example for the different existing code conventions (per programming language) is the 'PEP 8' convention for Python[e] .

During code review, code produced by one developer undergoes peer-review by another developer to ensure that errors and bugs are corrected, and that all code conventions have been adhered to. However, proper code review is difficult to implement in practice due to the time and effort it requires. Making bioinformatics workflows publicly available is a suitable and simple alternative.

[e] https://www.python.org/dev/peps/pep-0008/

Publishing bioinformatics workflows ensures that the source code used in validated bioinformatics workflows is available to the scientific community, and can thus be employed and tested by external scientists. It therefore also indirectly allows code review. Although making code open source cannot be enforced, it is strongly recommended and a variety of different licenses are available for this very purpose (e.g. General Public License, Berkeley Software Distribution) and should be considered.

Documented procedures for testing and updating bioinformatics workflows is also of primary importance. It is recommended that the 'DTAP' (development-test-acceptance-production) life cycle of software products be followed, wherein bioinformatics workflows exist in different phases, ideally also on separate computational environments (e.g. a different server). The development stage encompasses the active development of the workflow. The test and acceptance stages encompass verification that the prototype works as expected by the developer by taking the prototype through a series of steps that assess stability etc., and verification that the prototype meets the needs and requirements as expected by the end user of the program, respectively. Lastly, the production stage encompasses the process of taking the bioinformatics workflow into production for routine analysis, complying with the performance criteria evaluated during the validation.

Proper documentation of the bioinformatics workflow at all of the different levels is highly recommended and can exist at different descriptive levels. Firstly, proper documentation of the source-code inline to allow other programmers to read and understand the code. Secondly, proper documentation of the technical properties of the bioinformatics workflow to provide an overall overview to other programmers so that they can suggest the making of functional changes to the overall workflow. And lastly, proper documentation on how to use the bioinformatics workflow by the end user in order to ensure that the bioinformatics workflow is properly employed for routine analysis.

4.7. Reference databases and data storage

As discussed in section 4.5, fundamental parts of many bioinformatics pipelines are based on the comparison of the gained sequences to reference sequences (e.g. alignment). Those sequences are predominantly stored in publicly or privately accessible databases. The databases are kept up to date either by a private entity, a community or a combination of the two, where in community members suggest updates which are then evaluated and incorporated into the database by a governing body of experts. Since there is a growing number of reference databases available, it is important to choose databases which are maintained and updated regularly and to always note the version of the databases used during analysis. Examples of such reference databases in the microbial world are the Resfinder [31] (for antimicrobial resistance gene detection) and VirulenceFinder [32] (for virulence gene detection) databases from the Center for Genomic Epidemiology, National Food Institute, Technical University of Denmark. For GMO detection, the NCBI database [33] can be used for preliminary analyses. Other databases containing sequence information specific for GMOs and CRISPR-mediated genome-edited plants exist, such as the EUginius, the JRC GMO amplicon, the CrisprGE and the Nexplorer databases, which include amplicon sequences. Some of these databases can be directly queried using BLAST or similar tools. A non-exhaustive list including information on the databases containing GMO sequence data available is reported in Table 1 below. A list of the most important databases containing information on insert structure and GMO approval status can be found in the table present in Annex 2 to this report.

Table 1. Databases containing sequence information on GM and genome-edited plants. It should be noted that generally databases cannot be used as such for bioinformatics analyses, this will require additional steps restructuring the available sequencing information.

GMO sequence databases	
The European GMO database (EUginius.eu) https://euginius.eu/	The database provides and collects detailed information on issues regarding the presence, detection and identification of GMOs with a focus on the situation in the European Union as well as worldwide coverage. Where available, the information on GMOs includes the description of the genetic elements including related DNA sequences. It is an initiative of BVL - the Federal Office of Consumer Protection and Food Safety (Berlin, DE) and WFSR - Wageningen Food Safety Research of Wageningen UR (Wageningen, NL), presently supported by further member states of the EU. EUginius´ intention is to support competent authorities and private users who seek information on GMOs. Direct query of the database is not possible, data present on the database needs to be first downloaded by the user and queried locally.
JRC GMO-Amplicons database https://gmo-crl.jrc.ec.europa.eu/jrcgmoamplicons	The database collects putative GMO-related nucleotide sequences, obtained by PCR simulation screening of public nucleotide sequence databanks, including patents and available whole plant genomes. It was developed by the European Commission Joint Research Centre. The JRC GMO-Amplicons database is meant to assist the validation of new GMO detection methods, and to verify the quality and validity of already validated methods. It is possible to perform a direct query of the database.
JRC Central Core DNA Sequence Information System (CCSIS) https://gmo-crl.jrc.ec.europa.eu/jrcgmomatrix/	The database collects GMO sequence information received from companies as part of their legal obligations or extracted from nucleotide / patent sequences databases. It is maintained by the European Commission Joint Research Centre. The database sequence information can only be exploited through the use of the JRC GMO-Matrix, Event finder and Prespotted plates applications. Direct query of the database is not possible.
Plant Genome Editing Database (PGED) http://plantcrispr.org/cgi-bin/crispr/index.cgi	The database currently provides information about plants that have been generated using the CRISPR/Cas9 technology in order to study economically important traits. Information includes the transformation experiment, the name of the transformed plant variety, the DNA construct used, including the guide RNA sequence and primers used to characterize resulting mutations, and details about the mutant plant line including the altered sequence, whether it is heterozygous or homozygous, and any phenotypes that have been observed. This database is supported by National Science Foundation and hosted by Boyce Thompson Institute. Direct query of the database is not possible.
CrisprGE http://crdd.osdd.net/servers/crisprge/index.php	The database presently provides information about CRISPR/Cas-based genome-edited organisms, genes, target gene sequences, genetic modifications, modifications length, genome-editing efficiency, cell lines, assays, etc. It is developed by the Council of Scientific and Industrial Research and the Institute of Microbial Technology in India, as an assistance to accelerate research in the field of genome engineering. It is possible to perform a direct query of the database.

Nexplorer https://nexplorer.sciensano.be	Nexplorer is a sequence-based database containing authorised GMO events and links to their respective annotated DNA sequences. These are stored in a structured, searchable and extractable format. The availability of preorganized and annotated sequencing information streamlines bulk analysis of different types of sequencing data, including long read sequencing data obtained from targeted sequencing GM detection methods. It is made available as a web-application offering a user-friendly interface that allows non-IT experts to manage the database and the sequences. It is developed by Sciensano, coordinator of the national reference laboratory for GMO in Belgium. In a proof of concept for efficient analysis of massive parallel DNA sequencing data for the detection and identification of authorized and unauthorized GMOs, the methodology for the analysis of sequencing data of DNA walking libraries of samples containing GMOs using the database was developed [34] . It is possible to perform a direct query of the database.

Source: ENGL

One aspect that is closely related to achieving a high level of capability for GMO analysis across Europe seems to be the creation of a central repository to share data as well as procedures used to create the shared data. Here, a form of controlled vocabulary may be helpful in order to facilitate access and searchability [26]. This repository, if collectively run and maintained, would enable not only a more streamlined bioinformatics analysis (as it would ideally contain assembled genomes, along with tools for doing these analyses), but it would also create an unparalleled data resource for the development of downstream applications for the identification of GMOs entering the European market.

As stated, specific requirements should be considered when creating (and storing) data from massive parallel DNA sequencing. For example, all samples should be processed in a similar manner, according to harmonized protocols and should preferably be treated similarly post sequencing (bioinformatics analyses). To ease data curation, accessibility and usage, the datasets and/or most informative sequences could be catalogued descriptively (based on i.e. species, sample type, sequencing platform used). The datasets produced should be well described according to a sequencing standard (see https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html) so that the harmonized protocol and/or special parameters used to create the data are well-documented (metadata) in a standardized manner.

After bioinformatics analyses and processing, data should be organized and compacted (cleaned from excess and irrelevant material) and accompanying metadata files should be saved together with their datasets. In EN ISO 23418:2022 [1], recommendations for the harmonization of metadata in the context of whole genome sequencing for typing and genomic characterization of foodborne bacteria have been given. In some cases, it might be an interesting option to save only the metadata and key results of a run into a database, while storing the DNA extract, to reduce data storage needs. Since sequencing is becoming increasingly cheaper and storing large sequencing datasets might not always be desirable, this might become a viable option for routine samples. However, in the case of non-routine samples or new accessions (new GMMs or GM varieties) whole bioinformatics datasets should be stored with accompanying metadata.

Massive parallel DNA sequencing data could possibly be uploaded and stored on free public databases (Sequence Read Archive, European Nucleic Archive) [35,36]. However, it should be considered that such databases are public resources. In the case of sensitive and or IPR protected sequences, datasets would have to be protected, with limited access for specified users. This is especially relevant if the samples are non- compliant and/or legal actions have to be taken against certain parties.

In such cases, it should be considered that sensitive/confidential data should not be stored in a public database, and back-ups should also exist somewhere on a stable server/hard drive - up to several years.

Regardless of the database used, storing terabytes of data will be a major effort and will entail major costs. In fact, all matters related to the maintenance, curation and timing of the disposal of unnecessary data should consider national and EU laws. To ease data sharing between member states and competent authorities, the databases and back-up servers should preferably be located at a shared EU database. The computing capacity for massive parallel DNA sequencing runs and bioinformatics analyses is a challenge that needs to be solved, so that this type of analyses can be performed up to harmonised standards in all laboratories of all member states, when needed.

5. Quality aspects related to methods based on Massive Parallel DNA Sequencing (MPS) for GMO analysis - applications to example scenarios and rationale for using the technology

5.1. Introduction

The full workflow used by EU enforcement laboratories for GMO analysis includes an initial screening phase and, if required, the subsequent identification and quantification steps using qPCR. Consequently, there may be several constraints/limitations related to the discovery of all the GMO event(s) possibly present in a particular analytical sample, as described below. Current methods for the screening, identification and quantification of GMO events rely on at least partial knowledge of the sequences that have been modified in the genome of each GMO. Those methods generally relate to highly specific qPCR amplification of one or multiple DNA targets for which primers or probes were designed and are commonly used in reference laboratories. Therefore, during the initial screening phase only the GMO(s) for which a PCR method is available and is applied, can be detected.

In a sample containing both EU-authorised and unauthorised GMOs, this screening system may fail to detect an unauthorised GMO when it shares a common transgenic element or construct targeted by the PCR screening analysis with an authorised GMO, or if none of its transgenic elements are targeted by the screening system. Only in those cases where known unauthorised GMOs harbouring the same common transgenic elements are also included in the confirmatory identification experiments, they may be identified. Unknown GMOs with the same common elements or those which are not targeted by the screening system will not be identified. In addition, the recent introduction of strategies for more targeted modifications at the DNA level, including a range of genome-editing techniques of which CRISPR-Cas has now become most prominent, will require rethinking of the current use of PCR-based methodologies for GMO detection and identification.

Concurrently, a number of massive parallel DNA sequencing techniques have emerged, which continue to rapidly mature in terms of accuracy, capacity and costs. This rapidly evolving field for nucleic acid sequencing has the potential to be used in the new strategies for the detection and identification of conventional as well as genome-edited organisms GMOs. These massive parallel DNA sequencing strategies could overcome the limitations described above, and can be applied to a number of different situations which relate to the detection of GMOs. They could be applied for screening for the presence of known or unknown GMOs in a sample; for the detection and identification of known GMOs in a sample; or for the molecular characterization of a known or unknown GMO sequence. They may probably also be used, in specific cases, for the detection of single and short nucleotide variations that have been obtained with the use of new genomic techniques, such as those referred to in the ENGL report on the 'Detection of food and feed plant products obtained by targeted mutagenesis and cisgenesis' [37].

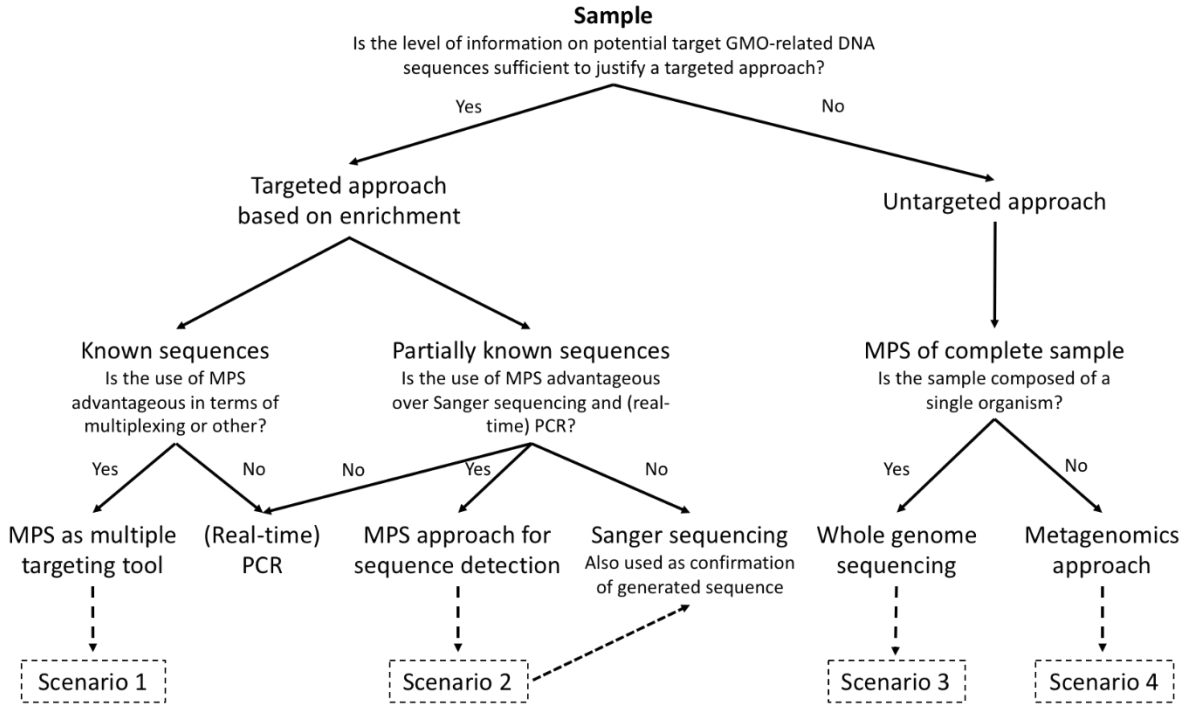
Different scenarios can be envisaged. In rare cases, the sample to be analysed may be simple, containing potentially one or a few known GMOs, and comprised of a simple matrix containing only one organism/crop species. Alternatively, the sample may be complex, with both known and unknown GMOs present, and comprised of a complex mixture of ingredients or matrices. This will obviously have an influence on the strategy that will need to be selected for analysing the sample. The final goal of the massive parallel DNA sequencing approach can be the characterisation of the full genome of a GMM, the full identification of an inserted genetic element and its flanking regions, or the screening for multiple GM elements in a single genome or in a mixture. In a sample consisting of a complex mixture, targeted approaches may be advantageous in order to screen for known and partially known sequences.

The final goal also defines the scale of sequencing required and the approach used. Therefore, each massive parallel DNA sequencing workflow needs to be tailored and performance criteria specific to each step of the analytical procedure need to be adapted depending on the intended purpose. The quality criteria to implement should be flexible in order to address these various applications but should be sufficiently detailed to ensure that procedures and processes used to support regulatory decisions are reliable and reproducible.

The aim of the present section is to describe critical steps associated with quality criteria for each stage of the massive parallel DNA sequencing workflow. It is recommended to take into account the critical steps and associated parameters defined in the current document to form the basis for future iterations of the ENGL guidance "Definition of minimum performance requirements for analytical methods of GMO testing" [38]. However, more data is needed at the experimental level (e.g. appropriate sequence databases and systematic validation data for each stage of the sequencing workflow) before proper performance requirements can be defined.

Depending on the goal of the massive parallel DNA sequencing approach used and on the type of sample to be analysed, the steps described in the previous section are not always of equal importance. Four possible exemplificative scenarios were identified, based on different knowledge levels of the target GMO sequence and on different sample types. Figure 2 outlines a decision tree describing the steps that lead to the different scenarios.

Figure 2. Decision tree for the application of massive parallel DNA sequencing for GMO analysis based on different knowledge levels for the target GMO sequence and on different sample types. MPS = massive parallel DNA sequencing



Source: ENGL

The four scenarios have been defined as follows:

- Scenario 1 describes a targeted sequencing approach using different enrichment strategies, with a focus on multiple known sequences.

- Scenario 2 describes a targeted sequencing approach using different enrichment strategies, with a focus on partially known sequences.

- Scenario 3 describes a non-targeted sequencing approach applying whole genome sequencing for completely unknown GMOs.

- Scenario 4 describes a non-targeted metagenomics approach applying full sequencing of the genetic material present in a sample for screening and identification of completely unknown GMOs.

In real life, a combination of the approaches described above may need to be applied. It is important to consider quality aspects specific for the approach and the workflow applied to each scenario. However, some points of attention common to all scenarios can be identified. Both are described in the following sections. This section will be focused on the application of massive parallel DNA sequencing strategies. References will be made to Sanger sequencing as in some situations this may be employed as a component of possible massive parallel DNA sequencing strategies. Quality aspects related to Sanger sequencing are discussed under section 3 of this report.

5.2. Points of attention common to all scenarios for GMO analysis

The steps that require particular attention in relation to GMO detection and identification with massive parallel DNA sequencing strategies in more or less complex samples will be illustrated in this section and in those describing the four different scenarios mentioned above. The parameters that need to be carefully monitored for each step of the procedure are detailed and summarised in Table 2 below. The aspects that are relevant to achieve effective and reproducible massive parallel DNA sequencing strategies for GMO detection will be highlighted. The following points are of relevance for all scenarios:

DNA preparation: although standard protocols for the extraction of DNA from (highly) processed matrices exist, it may be challenging to extract qualitatively and quantitatively sufficient DNA from a market sample, currently impeding the use of massive parallel DNA sequencing as a tool in GMO analysis for such matrices. The issue of poor-quality DNA is ubiquitous across all molecular biology approaches, but it can have a particular impact on massive parallel DNA sequencing approaches. The use of samples of DNA extracted from well-defined materials with details on the type of material used (well-defined samples or reference material) [19,20] is hence advisable when developing and implementing massive parallel DNA sequencing.

Library synthesis: producing a DNA sequence library that satisfies the quality criteria presented in section 4 would be technically challenging if the input DNA were not of a high quality. As this may be a problem that would be encountered in GMO control, certain criteria may need to be relaxed in order to be able to sequence a specific sample.

Sequencing reaction: the type of sequencing platform used and the number of samples to be pooled in one sequencing run should be carefully evaluated, depending on the scenario (e.g. complexity of mixture, size of genome, expected output of instrument, required coverage, required accuracy).

An important quality parameter of massive parallel DNA sequencing data is the read depth, which should be defined to be sufficient for the intended purpose (e.g. in order to detect junction reads), if realistically feasible. Willems and colleagues [39] have proposed a statistical framework for estimating the probability of sequencing junction reads that span the junction between the known introduced DNA and the host genome DNA. They recommend that, for non-targeted approaches, a typical minimum estimated coverage from 20 to 60-fold for a short-read technology, be implemented. This is in line with JRC (2017) Guideline for submission of DNA sequences and EFSA's (2018) requirements for massive parallel DNA sequencing for the risk assessments of GM plants [19,20]: minimum read depth of 40 for the description of the insert. A potential problem here is that for complex matrices with unknown composition it will not be possible to determine the read depth per species before performing the experiment.

Data analysis: The analysis of massive parallel DNA sequencing data is usually a computationally intensive step that is accomplished through the execution of a series of publicly and/or commercially available, or in-house developed, bioinformatics tools. These tools together constitute a bioinformatics pipeline, which enable the goal of the particular analysis to be achieved. This can be, for example, the characterisation of a sample by screening for known elements-/event-specific sequences in comparison with a database with well-characterised GMO-related sequences (e.g. EUginius and JRC GMO-amplicon databases, see Table 1 in section 4.7), or the detailed characterisation of a GMO, etc. The quality requirements for bioinformatics analyses performed in support of GM plant applications for authorisation under the EU legislation are listed by EFSA (2018) [20]. However, quality requirements for their use in support of GMO testing are not yet available. The data analysis managed by a pipeline can be considered to be comparable to the data analytical methods employed for current GMO testing. Therefore, ENGL documents [38] providing guidance on how these (PCR) methods should be evaluated and validated may be taken into account in order to evaluate and validate a bioinformatics pipeline that is used in the end-to-end analytical process to obtain the final result. In fact, in a similar fashion to current–methods developed for use in GMO testing for enforcement purposes, evidence should be provided that the developed massive parallel DNA sequencing strategy satisfies certain performance criteria, such as specificity, dynamic range, trueness, limit of detection (LOD) and robustness. The same principle applies to bioinformatics pipelines: the performance criteria that have to be applied to the bioinformatics analysis have to be specified *a priori*, and the set of thresholds for acceptance need to be fixed for all developed bioinformatics pipelines (taking into account the specific application). It is therefore necessary to include a relevant and representative set of *in silico* and/or *in vivo* datasets (i.e. data generated by real experiments) which are linked to the expected output, together with considerations for the minimum performance requirements to be met by the pipelines. When used for legal purposes, the results obtained after the massive parallel DNA sequencing strategy should be verified with the use of an alternative method, such as a validated PCR-based method where available, or alternatively PCR followed by Sanger sequencing may be used when other validated methods are not available. *In silico* generated datasets may be used for cases where it is difficult or even impossible to generate experimental data with the corresponding real-time or digital PCR methods on a comparable sample.

For the submission of massive parallel DNA sequencing data, guidance for the reporting is also largely available in JRC (2017) [19] and EFSA (2018) [20] published guidance. This entails, similar to Sanger sequencing reporting, a detailed description of the samples and materials used as the basis of the sequencing analysis, a confirmation that the samples will be stored for future analysis, a detailed description of the experimental procedure and the massive parallel DNA sequencing method(s) applied, including information on sequencing depth, as well as a full report on processing steps of the resulting raw data (format, filtering, trimming etc.) and a detailed description of the conducted bioinformatics analysis of the sequencing data. The latter will include a detailed description of the used software and tools, including names, versions, selected options and parameters used.

The quality aspects presented below have to offer a certain degree of flexibility, given the wide range of platforms currently available and specific combinations of purpose and approach. However, when using a specific massive parallel DNA sequencing strategy, it is important to apply strict performance thresholds that are tailored to the specific aims of the study for which they are intended. Depending on the goal of the study and the type of sample to be analysed, the steps described below may not always be of equal importance.

Table 2. Parameters that need to be carefully monitored for each step of the sequencing process and which are useful as guidance for the development of quality criteria. Where necessary the table details parameters specific to individual scenarios, in light grey in the table. Appropriate controls should be included in all the steps described.

Steps	**Guidance for the development of quality criteria**	
DNA preparation	The selected method for DNA isolation should provide DNA in an appropriate yield and quality for the intended analysis. The yield of the DNA should be sufficient for further analysis (see also requirements for library preparation). Depending on the approach/library kit subsequently used (see manufacturer recommendations), low DNA amounts are recommended for targeted massive parallel DNA sequencing applications with an amplification step (5-15 ng/µL), high amounts for PCR-free approaches (> 50 ng/µL; total DNA up to 5 µg). The quantity of DNA should be determined by fluorometric methods. The purity of the DNA should be checked spectrophotometrically by readings at 230, 260 and 280 nm, although good quality massive parallel DNA sequencing libraries may be obtained from samples with inferior ratios (260/280 ratio should preferably be 1.75 – 2.05 and 260/230 ratio 2.0 - 2.2 for pure DNA). The integrity of the DNA should be checked by standard or capillary gel electrophoresis. See additional scenario-specific criteria below. An enrichment step is foreseen for scenarios 1 and 2, and corresponding guidance is given below. Enrichment is not foreseen for scenarios 3 and 4.	
	Scenarios 1 & 2	DNA integrity: in the case where enriched targets contain adjacent sequences, DNA integrity is more critical for the subsequent massive parallel DNA sequencing analysis. The method of enrichment will be selected on the basis of the characteristics of the sample and of the specific questions to be answered. It is important to monitor the level of enrichment on the basis of selected sequences, if feasible. This will provide an indication of the actual enrichment during the procedure. The method applied will need to be described in detail.
	Scenarios 3 & 4	DNA integrity: for long read sequencing, high concentration of high molecular weight DNA is required.
Library synthesis	Library synthesis includes a number of subsequent steps, including most often DNA fragmentation, adapters ligation, and amplification of the library (if applicable). It is important to monitor the results after each of the (key) steps by assessing the size of the fragments after fragmentation, size and concentration after adapter ligations, and quality /quantity of the resulting library. A crucial part is an accurate quantification of DNA-libraries, fluorometric or qPCR-based methods are recommended. Spectrophotometric methods are not sufficiently accurate.	

Sequencing reaction (platform)	The selection of the sequencing platform is based on the chosen massive parallel DNA sequencing approach for a present question to be answered (see Annex 1). Samples may be pooled for practical reasons.	
	Scenario 1	Short read platforms will generally suffice as the sequence stretch for which validated methods have been developed, is generally short. If samples are mixtures of different ingredients, it may not be possible to determine the (required) read depth for the individual components of the sample.
	Scenario 2	Long read platforms may preferably be used for DNA walking, as assembly may prove difficult for data obtained with short read platforms.
	Scenario 3	Depending on the issue at hand, the use of either short read systems, long read systems or a combination of the two is advisable. Short read systems produce short reads with high quality and great depth, which is advantageous when trying to discover minor modifications and a reference genome is available. Long read systems produce reads of lesser quality and depth, which are advantageous for assembly and to avoid short contigs, and if a reference genome is not available. Longer reads are also useful in genome resequencing, when planning to detect copy number variants or large insertions/deletions. A combination of both systems might increase the analysis performance when, for instance, a reference genome is not present and minor modifications are to be detected. The read depth should be ≥ 30 for GMM isolates [40] and ≥ 50 in other cases [41]. For *de novo* assembly, high coverage and depth is required, and long reads are preferred. The characterization of the chromosomic or plasmid (mobile element) location of GMMs' AMR genes often requires a combination of short and long reads. Additionally, AMR genes are often integrated in very repetitive regions, which makes it difficult to address their location using short reads only.
	Scenario 4	Long read systems are preferred in order to better classify the resulting reads to the corresponding species and to be able to 'reconstruct' the different genome/genome fragments. This is especially true when *de novo* assembly is needed. However, a combination of both long and short read systems might be advantageous in some cases.
Primary data analysis	This is the first step in the data analysis and sequencing instrument-specific. In all cases, raw data and details of the base-calling procedure should be recorded, including any uncertainties observed during the base calling. The software tools used (name and version), as well as the options and settings selected, must be documented. Minimum base quality criteria may be set and need to be registered. Raw massive parallel DNA sequencing-reads in a standard format (e.g. FASTQ, uBAM), already filtered, and eventually trimmed off the used adapters should be kept. The software and parameters used for the filtering and trimming should also be described.	
Secondary data analysis	The secondary analysis covers simple database alignment up to (*de novo*) assembly of the reads obtained depending on the chosen sequencing strategy. The full procedure should be recorded, including any uncertainty observed during the analysis. The software tools used (name and version), as well as the options and settings selected, must be documented.	
	Scenario 1	Secondary analysis focuses on the alignment with well-characterised sequences in dedicated databases (see Table 1) with well-characterised GMO-related sequences, using software for variance, or even SNV assessment.

	Scenario 2	Secondary analysis focuses on the assembly of the reads, *de novo* or based on alignment with well-characterised sequences from curated databases (see Table 1), using software for variance assessment.
	Scenario 3	Secondary analysis focuses on the assembly of the reads, *de novo* or based on alignments with well-characterised sequences (if no reference genome can be applied) followed by variant calling. *De novo* assembly is often required for GMMs.
	Scenario 4	Secondary analysis focuses on the assembly of the reads, *de novo* or based on alignments with well-characterised sequences, using software for variance assessment. In order to resolve mixed samples, a (*de novo*) assembly is needed to produce contigs that can subsequently be blasted. Alternatively, a repeated alignment against expected elements of the sample can be run before assembly, utilizing the non-matching reads of the previous alignment. A further possibility to decrease the sample complexity for assembly is binning.
Tertiary data analysis	This is the interpretation phase. To positively identify known GMO events and elements and as far as possible unknown GMO events present in the sample, it is necessary to develop a standardised approach using curated, up-to-date databases with sequences of authorised and (as far as available) well-characterised unauthorised GMOs (see Table 1 for a list of available databases). It is necessary to register the procedure for interpretation in detail, including the database(s) (name and version) and bioinformatics tools (name and version), options and settings selected, and any uncertainty observed during the analysis. Further analysis of called variants must be undertaken in order to possibly identify and characterize unknown GMOs and/or possible unnatural associations.	
Confirmation of positives	To confirm the massive parallel DNA sequencing results of the tertiary data analysis gold-standard methods (e.g. qPCR, Sanger sequencing) should be applied. On a case-by-case basis, it will need to be assessed which method can confirm the results of a chosen massive parallel DNA sequencing approach. The validation status of the methods used, including the related documentary and the results of the confirmation experiments will need to be recorded. In the case of qPCR methods, reference should be made to the relevant ENGL guidance documents to describe the details of the method(s) used.	
	Scenario 1	Validated qPCR methods will be available.
	Scenarios 2, 3 & 4	Gold-standard methods (e.g. Sanger sequencing) should be applied to confirm the existence of newly discovered GMOs within the original sample in order to avoid false positive declarations based on either sequencing or analytical errors.

Source: ENGL

Long-term storage of sequencing data with all of the related parameters and, optimally, all related metadata of the bioinformatics analyses, should be ensured to help enable data traceability and reproducibility.

In the near future Machine Learning (ML) models, i.e. computer algorithms that improve automatically through experience and by the use of data, may be applied across massive parallel DNA sequencing technologies for the routine analysis of genomic data, as already suggested for food control [42]. Typical applications of ML are the alignment, classification, clustering, and pattern mining in DNA sequence data [43]. The main reasons for using ML are the reduction of human intervention (automation), continuous improvement and simplification of applications, easier handling of multi-dimensional and multi-variety data as well as the recognition of hidden patterns in large data sets [44, 45].

ML tools, such as support vector machines (SVM) [46], artificial neural networks (ANN) [47], k-means and others, could be used, for example, to differentiate between known (GMO and "wild-type") and unknown genotypes in a complex matrix, based on available sequence knowledge of existing GMOs. This would be feasible as a ML model would be trained on many individual genomes of wildtypes and known GMOs. When it discovers a genome (either a complete genome or contigs) in the sequencing result of a complex matrix that does not match anything it was trained with or a generalization thereof, it would flag that genome as 'unknown'. The identity of this 'unknown' (a wildtype variety or truly an unknown GMO) has to be determined in later steps, either through manual analysis or a different ML model. This will provide potential opportunities for increased and efficient use of sequencing technologies in all four scenarios.

5.3. Scenario 1: MPS targeted approach (based on initial enrichment) – focus on known GMO events or GMO elements – complex mixture

Currently, the reference method used by enforcement laboratories for GMO detection is real-time PCR (qPCR). The number and diversity of GMOs have greatly increased in recent years. Consequently, the number of qPCR methods to be applied on complex matrices in order to identify all the potential screening element and event specific GMOs (more than 70) corresponding to these positive elements has drastically increased. Moreover, the high number of sequences and the absence of reference material precludes the development of qPCR methods corresponding to all the sequences present (both for authorized and unauthorised GMOs) in databases such as Eugenius. This means that GMO searches are potentially limited. Massive parallel DNA sequencing approaches have the potential to address the challenge of identifying all GMOs in a sample. At the present time, sequencing the complete sample is not performant enough to detect multiple sequences with sufficient depth inside a complex mixture with several GMOs. Enriching sequences of interest makes the approach more applicable even to mixed products, where insufficient coverage of the different genomes may occur. Therefore, combining massive parallel DNA sequencing with a strategy of enriching the regions of interest is the only available strategy to simultaneously detect a large number of amplicons in a complex sample containing several GMOs (including some at low level). Two types of enrichment strategies followed by massive parallel DNA sequencing analysis are the main approaches used, namely PCR based approaches or DNA hybridisation through probe capture. PCR-based enrichment of multiple known GMO-related sequences in multiplex is generally not of benefit. Indeed, an enrichment strategy based on PCR requires the development of PCR methods and deals with the intrinsic limitation related to PCR target competition when all the PCRs are being run simultaneously on the same DNA template.
Therefore, an enrichment strategy based on sequence hybridisation by probes, facilitating the capture of the DNA fragments of interest, is consequently a preferable alternative, as it does not require initial PCR amplification. An additional advantage of this approach is that some fragments captured by the probes are slightly larger than the probe (theoretically probes of 120 bp may allow capture of fragments up to a maximum of 500 bp), and therefore also cover unknown regions such as the junction between plant DNA and GM constructs, that may potentially facilitate the detection of unknown GMO sequences (related to scenario 2, see below).

In this scenario, enrichment may focus on GMO event-specific or common GMO element-specific sequences from both authorised and unauthorised GMOs. Multiple GM elements might be enriched, but only the GM element for which the sequence is known may be found. Therefore, a strong prerequisite for a successful strategy is the presence of an extensive and reliable sequence database in order to be able to design PCR primers or capture probes.

This methodology relies on the affinity between the probe/primer and the target. Therefore, performance criteria of the method, such as specificity and sensitivity, will depend on this parameter for each target. As no validation strategy exists to assess the “affinity“ of each amplicon

and probe, it is advised to verify the results (identity of the amplicon) by corresponding qPCR or Sanger sequencing of the amplicon.

The parameters that are relevant to the different steps associated with this scenario are listed in Table 2 at the end of section 5.2 as guidance for the development of quality criteria.

Literature on Scenario 1

In 2018, Arulandhu and colleagues [48] developed a massive parallel DNA sequencing-based GMO screening approach based on target PCR amplification and subsequent sequencing with Illumina 150 bp paired-end (PE) technology. The enrichment was based on PCR assays covering 96 GMO targets, and a data analysis pipeline to detect and identify GMOs in complex food or feed samples was developed. When the authors compared this massive parallel DNA sequencing-based GMO screening approach with the qPCR-based GMO screening, it was shown that for targets present in relatively low concentrations (at around or less than 0.1 %), the detection showed discrepancies between the massive parallel DNA sequencing-based screening and the qPCR approach. This study proved the applicability of massive parallel DNA sequencing as a screening method for GMO, but also provided evidence of the non-quantitative nature of sequencing results, and the need to set a threshold for detection accurately, based on a comparison of qPCR and massive parallel DNA sequencing results, in order to avoid false positive results.

As a proof of concept, Debode *et al.* [49] described in 2019 the development of capture probes from GMO sequences publicly available, which contained approximately forty structural elements frequently used for transgenic cassettes. The total size of the enrichment sequences used for the capture probes was approximately 53 kb, but the database is still far from its limit as the methodology can be scaled up to 24 Mb. After enrichment, the DNA libraries were sequenced on an Illumina system (2 x 75 bp), and the results were evaluated for GMO detection capability with a specifically developed bioinformatics pipeline. This bioinformatics pipeline was designed to both estimate the presence of reads above the background level (25 reads per kb per million mapped reads) for the different elements targeted by the enrichment strategy, and to characterise the detected GMOs through the creation of contigs to reconstruct the whole transgene, possibly including the plant-construct junction. Although no systematic analysis was conducted, the results of the study indicated that this strategy could be used to detect a large panel of GMO elements and to partially or completely reconstruct the insert sequence in a single analysis, even at GMO percentages as low as 0.1 % when a single GM is present or in a sample.

5.4. Scenario 2: MPS targeted approach (based on initial enrichment) – GMOs with partially known elements – single or complex mixture

This scenario relates to massive parallel DNA sequencing strategies adopted when a potential unknown GMO that has not yet been described in the scientific literature or in sequence databases is suspected as being present in a sample. Indeed, qPCR methods are designed to target known elements, but have no potential to discover new sequences or therefore to detect and identify unknown GMOs. Massive parallel DNA sequencing approaches have the potential to address the challenge of identifying all GMOs, including unknown ones, in a sample. As previously mentioned, at the present time, sequencing the complete sample is not performant enough to detect multiple sequences with sufficient depth inside a complex mixture consisting of several GMOs, and enriching sequences of interest is the only applicable strategy.

In particular, the detection by qPCR of elements such as P-35S only weakly indicates the presence of EU-unauthorised/unknown GMOs, because these elements are commonly present in a range of GMOs, whether EU authorized or not. With samples composed exclusively of species that do not belong to the list of EU-authorized GMOs, inclusive of rice, wheat, or papaya, the detection of these elements strongly indicates the presence of EU-unauthorised GMOs. Here an approach based on DNA walking can be applied based on GM elements detected during the first screening step.

Enrichment by PCR amplification can be performed and amplicons can be produced using one primer anchored on the identified GM element and several degenerated primers annealing randomly in the genome. The final PCR product could then be sequenced through Sanger sequencing (after purification of a single band from a gel) or massive parallel DNA sequencing technologies. Massive parallel DNA sequencing is preferable to deal with the sequencing of the amplicons produced when a sample contains multiple GMOs. The generated sequences are then analysed through bioinformatics analysis, initially comparing the generated sequences with a database containing at least all of the sequences from EU-authorized GMOs and to analyse further, in a second step, the sequences not matching with this database and therefore potentially related to unknown GMOs. Therefore, a strong prerequisite for a successful strategy is the presence of extensive and reliable sequence databases. Such databases were, however, for a long time not publicly available. The JRC GMO-Amplicon and EUginius databases, containing GMO sequences generated by qPCR screening, represented a first significant step in this direction (see Table 1 for information on available databases). However, in 2022, a proof of concept for the efficient database-guided analysis of massive parallel DNA sequencing data for the detection and identification of authorized and unauthorized GMOs using the Nexplorer database and DNA walking data was presented based on various scenarios that can be encountered in routine GMO analysis [34].

The major advantage of the DNA walking approach is that amplicons of several kb, corresponding to unknown sequences of the GMO, can be produced to help to proof and characterize the unknown GMO by sequencing of unnatural association(s) of element and/or junction between the inserted cassette and the plant genome, and can work on very limited amounts of material. However, firstly the strategy applies only to GMOs containing at least one known sequence corresponding to a common GM element such as P-35S. Secondly, the specificity of the amplification is the result of a successive primer hybridisation covering a specific sequence of 60 to 90 bp in total. Extensive non-specific amplification or chimeric amplification may occur. Therefore, the sequence of each unknown generated fragment must be carefully analysed and its presence in the sample verified by PCR using two specific primers designed on the identified sequence.
When multiple events are in close proximity (less than a read length apart), it may be advisable to enrich the region between the outer most known GMO event-specific or common GMO element-specific sequences, in order to amplify the signal of multiple events between the two flanking sequences simultaneously. This approach will only be feasible when applying long read sequencing. This would reduce the need for qPCR detection of each single element and might lead to the discovery of unknown events in said region. This approach may be feasible when dealing with modern genome-editing methods, resulting in comparatively small changes or when using a long-read system resulting in a large amplified region.

The parameters that are relevant to the different steps associated with this scenario are listed in Table 2 at the end of section 5.2 as guidance for the development of quality criteria.

Literature on Scenario 2

Liang and colleagues [50] provided one of the first examples of the DNA walking strategy. Vip3A (used in conveying insect resistance both to EU approved (*e.g.* MIR162 maize) and EU unapproved (*e.g.* COT102 cotton) crops) was used to anchor the walking. Liang and colleagues used this system as a model to evaluate the screening of *vip3Aa20* genes as a useful approach for the detection of unauthorised GMOs containing a known GMO element (vip3Aa20 CDS) that may be found in the food chain as a result of unwanted contaminations and mixtures [50]. An expanded array of PCR tests for vip3A detection was developed and used for SiteFinding-PCR (using known GMO elements as targets) combined with massive parallel DNA sequencing (Illumina and PacBio platforms) allowing the identification of the new flanking sequences, underpinning the validity of this approach for unauthorized-GMO detection.

Fraiture and colleagues [15,51] developed a DNA walking system anchored on known elements commonly present in GMO (P-35S, T-NOS, and/or T-35S pCAMBIA) for the detection and characterization of a broad spectrum of GMOs in routine analysis of food/feed matrices (mixed samples), that was coupled to a long-read sequencing system. They tested the detection and identification capability on grains containing several levels of GMO but also in processed food and GMO mixtures even at a very low concentration (0.01 % and 0.1 % GMO). In all tested samples, the presence of multiple GMOs was unambiguously proven by the characterization of transgene flanking regions and the combination of elements that are typical for a transgene construct. DNA walking methods, providing high length extension fragments from target sequences (e.g. > 20 kb), may have the advantage of the production of different targeted sequencing fragments extending from a common sequence and revealing if the expected insertions were inserted in different places of the genome (for example, in different chromosomes). The study demonstrated that the DNA walking strategy, fully integrated into routine GMO analysis of GMOs that have incorporated P-35S, T-NOS, and/or T-35S pCAMBIA elements in typical food/feed matrices, , efficiently identifies known and unknown GMOs, detecting the respective GMOs even at trace levels 0.1 % and 0.01 % even in complex mixtures. The same DNA walking data sets were re-analysed by Saltykova and colleagues [34] using a database-guided analysis, which allowed detailed and reliable information to be obtained with limited hands-on time. This study developed a methodology for the analysis of sequencing data of DNA walking libraries of samples containing GMOs using the Nexplorer database, applicable for different scenarios, i.e. sample containing a known GMO at 100%, sample containing an unknown GMO at 100%, samples containing a single GMO at varying concentrations to determine the sensitivity, processed food samples, sample containing a mixture of known GMOs and sample containing a mixture of known and unknown GMOs. Different types of sequencing platform were included (i.e. Oxford Nanopore Technologies and PacBio). This proof-of-concept paves the way for the use of the massive parallel DNA sequencing technology to aid routine detection and identification of GMO.

In 2020 , as an alternative to DNA walking, Boutigny and colleagues [52] developed a protocol to amplify the transgene and its flanking regions based on inverse long-range PCR targeting P-35S on circularized molecules of approximately 6 kb. Sequences of interest were further determined using long read sequencing (on MinION, Oxford Nanopore Technologies). The petunia transgene and its flanking regions were sequenced using this new protocol.

5.5. Scenario 3: Whole Genome Sequencing (WGS) for single organisms

This scenario covers the case in which an isolated/single organism could be obtained and therefore whole genome sequencing (WGS) could be applied. WGS is able to provide information on chromosomes but also on extrachromosomal genetic elements such as plasmids or organelle genomes. In particular, WGS data can provide information for the characterisation of the GMO regarding its potential genetic modification. This is particularly important when an unknown GMO is discovered. Although it is not the scope of the present document, WGS of a single organism is also frequently used for the characterisation of a GMO that will be present as a product in food, and for the characterisation of GMMs that are involved in the production of a variety of food and feed products. The marketing of these products within the EU falls under different legislations, which establish the requirement for a risk assessment for the authorisation of the product including the characterisation of the genetic modification. At the present time, quality criteria necessary for this purpose are described in an EFSA guidance [53].

Theoretically, WGS may also be used to identify single seed materials for which the WGS data has been part of the approval or registration dossier, but in practice, this will generally be too costly in the years to come.

Indeed, WGS has a high cost, compared to targeted sequencing, and the cost is directly related to the size of the target genome and to the coverage required. Therefore, at the present time, it is mostly used by enforcement laboratories for the characterisation of isolated strains of GMMs. Its application to pure organisms is therefore feasible, especially in the case of bacterial GMMs with smaller and less complex genomes than that of plants or animals.

No massive parallel DNA sequencing strategies currently exist which have been validated to characterize GMMs detected in samples in the context of enforcement laboratories. Quality criteria coming from the field of typing and genomic characterization of foodborne bacteria specific to GMM described in EFSA guidance, can be helpful to establish such criteria [1,53]. Description of the sequencing strategy and the quality control based on criteria such as read depth, covering the genetic modification as well as the full genome, need to be established. Regarding assembly and annotation, the use of reference-based read mapping is possible. However, *de novo* assembly is the preferred recommendation in order to more effectively deal with genetic modification. The identification of genetic modifications requires curated public databases for the wild type genome (e.g. REFSEQ). In addition, a BLAST approach against public databases (e.g. NCBI) to identify the gene and the function of the assembled sequence corresponding to the transgene will be necessary. In these instances, criteria regarding the similarity of the obtained sequences and the one in the database will also be necessary [40].

Moreover, it should be noted that GMMs could also carry antimicrobial resistance genes that are a health concern in the context of their potential spread. It is therefore crucial to characterize their chromosomic or plasmid (mobile element) location, which might challenge the *de novo* assembly, mostly due to the presence of repetitive sequences (length of about 6000 - 8000 bp in bacteria). The use of long read sequencing might facilitate the assembly of the chromosome and plasmid (e.g. hybrid analyses of long and short sequencing technologies or non-hybrid analysis via PacBio or Oxford Nanopore Technologies).

At the present time, as no WGS methods are validated for GMMs, confirmation of the identified GMO events may require using other methods, such as Sanger sequencing of specific related amplicons.

The parameters that are relevant to the different steps associated with this scenario are listed in Table 2 at the end of section 5.2 as guidance for the development of quality criteria.

Literature on scenario 3:

Regarding the WGS of GM plants, the virus-resistant SunUp papaya was the first transgenic plant genome to be fully sequenced and *de novo* assembled [54]. However, the evidence for the presence of inserts and insert junctions as nuclear copies of papaya chloroplast DNA fragments had to be complemented with Southern blot data due to the low coverage of papaya WGS.

Kovalik and colleagues [55] demonstrated that massive parallel DNA sequencing conjugated with a suitable bioinformatics pipeline for junction sequence analysis provides molecular characterization that is equivalent to the Southern blot and PCR-based methods for GMO detection on typical GM soybean plants. They have demonstrated that massive parallel DNA sequencing using short reads is also capable of characterising complex events including those with multiple T-DNAs and sequence rearrangements. However, this approach was more developed in the context of risk assessment when the genetic modification introduced in the genome is known.

The only real examples of WGS used in the context of enforcement laboratories to characterise an unknown and unauthorised single GMO, are in the context of bacterial GMM. Indeed, GMMs are often exploited for the production of molecules of interest in industry, such as fermentation products including additives, enzymes and flavouring. The underlying sequence information of the modified GMM strains is, however, generally not publicly available.

This seriously hampers the detection and identification of these strains, and thus the enforcement of GMO regulations in this area of application. In 2014, a viable vitamin B2/riboflavin producing *Bacillus subtilis* strain was detected and could be isolated in Germany, in a lot of vitamin B2 feed additive imported from China. The unknown strain was a genetically modified microorganism that was not authorized in the European Union. Whole-genome sequencing on an isolated potential riboflavin secreting bacterial strain revealed the sequence of the non-authorized GM *Bacillus* strain. In this scenario, massive parallel DNA sequencing rapidly provided critical sequence information for GMM identification that was further used to develop a specific qPCR detection method [56,57]. However, the initial results did not clarify whether the sequence targeted by the qPCR method was integrated into the bacterial genome or present on a plasmid. Subsequent massive parallel DNA sequencing of DNA isolated from the above-mentioned GM *B. subtillis* revealed the nucleotide sequence of all identified genetic modifications that had been inserted into the microorganism's genome, and characterized complementing extra-chromosomal recombinant plasmids [58]. As the GMMs also often carry AMR gene(s) that are a health concern in the context of the potential spread of these AMR genes, it is crucial to characterize their chromosomic or plasmid (mobile element) location. As the AMR genes are often integrated in a repetitive region, it is difficult to address their location using short reads only. This challenge was addressed by using long read sequencing (MinION, Oxford Nanopore Technologies) in addition to the short read sequencing for plasmid reconstruction through hybrid assembly [59]. In 2020, the unexpected presence of a viable unauthorised genetically modified bacterium (i.e. *B. velezensis*) in a commercialized food enzyme (protease) product originating from a microbial fermentation process was shown in Belgium (RASFF 2019.3332), based on the use of a massive parallel technology (Illumina MiSeq). WGS was used to characterize the genetic modification comprising a sequence from the pUB110 shuttle vector flanked at each side by the coding sequence of a *Bacillus* protease. This study emphasizes the current key role of WGS in the detection and identification of unknown and unauthorised GMMs [60]. As this protease-producing GM *B. velezensis* was found in and isolated from several other commercial food enzyme products, an in-depth genomic characterization and phylogenomic comparison was subsequently made using both short-read Illumina and long-read Oxford Nanopore Technology sequencing data to employ a *de novo* hybrid assembly strategy. D'aes and colleagues [61] demonstrated that the GMM primarily carry the transgenic construct, with a single copy of the wild-type derived protease encoding gene, on a free high-copy pUB110-derived plasmid. Additionally, transient unstable integration of this transgenic construct into the chromosome can occur. The GMM were genetically almost identical, indicating that they likely originate from the same parental GM strain and presumably manufacturer. This study highlights the added value of a hybrid approach for accurate genomic characterization of GMM (e.g., genomic location of the transgenic construct), and of SNV-based phylogenomic analysis for source-tracking of GMM.

In 2020, Hurel and colleagues [62] developed a bioinformatics pipeline to detect unknown genetic modifications in a bacterial genome without prior assembly of the sequencing data and without taking into account the biological function of the modified sequence. This pipeline is using machine-learning methods to analyse the difference of genomic vocabulary and has been successfully assessed on the data of the above-mentioned GM *B. subtilis.*

5.6. Scenario 4 – non-targeted metagenomics

Often single GM plants are not present or living GMMs cannot be isolated from a food matrix. Scenario 4 covers the current and, more realistically, future potential for the application of non-targeted massive parallel DNA sequencing (i.e. sequencing all DNA present in a sample, referred to as shotgun metagenomics [63]) to the analysis of complex products. These may include complex matrices, for example food and feed market samples containing DNA from different crop species or GMM contaminations in a fermentation product.

In contrast with the previous three scenarios, shotgun metagenomics is a non-targeted approach, sequencing all the DNA present in a sample. This does not require prior isolation or prior knowledge of the sequences, and has the potential to detect in one step, all the species and GM constructs or unnatural associations in a sample. This approach would facilitate the development of a more generalized detection and identification method for both authorized and unauthorised GMOs in any type of sample.

However, the application of metagenomics to complex matrices is at present difficult, even for microbial ones, for two reasons. Firstly, it would require a high coverage to achieve an acceptable limit of detection for the different organisms/species present in the mixture, which translates into a high running cost. Secondly, it would also require complex bioinformatics analyses to resolve the different species/strains and GMOs present, with many practical limitations. If shotgun metagenomics would be applied for the analysis of a mixture with suboptimal coverage, the analysis would potentially miss some targets, and therefore result in false negative results. The Shannon-Wiener-Index is a parameter that has often been applied to check whether the diversity is well represented [64]. However, future developments in massive parallel technologies and the analysis of resulting data, as well as increases in sample complexity, may soon make massive parallel DNA sequencing the method of choice for the analysis of complex samples.

The parameters that are relevant to the different steps associated with this scenario are listed in Table 2 at the end of section 5.2 as guidance for the development of quality criteria.

Literature on scenario 4

Except for the study on the detection and characterization of unauthorized GMMs in microbial fermentation products [65], no shotgun metagenomics approach applied to a real sample for GMO analysis currently exist in the literature and only review papers or reports on theoretical approaches are available.

In 2016, Holst-Jensen and colleagues [66] provided an overview of the scientific literature on the application of massive parallel DNA sequencing for the detection and characterization of genome-edited organisms and derived products. The overview included comparative approaches to identify genetic modifications, as well as *de novo* assembly and characterisation of complete genomes and transcriptome analysis to detect indications for potential genetic modifications. They concluded that sequencing of a complete sample may become the method of choice in the future, but that this will depend on further development of sequencing technologies in terms of cost-efficiency, throughput capacity, availability of high-quality genome data for a broader set of species, and improved and versatile bioinformatics pipelines. Since 2016, there have been major developments in all of these aspects, but not yet to the extent that non-targeted massive parallel DNA sequencing analysis for GMO detection and identification in complex food or feed products is already feasible.

Furthermore, in order to evaluate the feasibility of the sequencing of a complete sample for GMO analysis, a statistical framework was developed to calculate the probability to detect a GMO in a sample with known composition [39]. This approach can be used to estimate the number of (target) reads necessary to detect a GMO in a given sample. This framework considered the search of a target sequence (transgene insert) in a WGS experiment based on theoretical considerations and was validated by massive parallel DNA sequencing data on a GM rice (Bt rice). The massive parallel DNA sequencing experiments used different proportions of GM rice in different matrices. The authors tested the detection of transgene inserts which were potentially present, the integration in the host genome, and the identification of the specific junction sequences. It was shown that it is theoretically possible to use massive parallel DNA sequencing to detect and identify samples of 100 % GM crops. However, diluted samples and mixtures require large massive parallel DNA sequencing experiments, with billions to trillions of reads and their associated costs, to yield a high probability of finding targeted reads for each approach.

As is the case for WGS (scenario 3), the only real example of shotgun metagenomics used in the context of GMO so far, is the study of Buytaers and colleagues [65] on the detection and characterization of an unauthorized bacterial GMM in microbial fermentation products. In this study, a proof-of-concept was presented for a metagenomics-based approach to deliver the proof of presence of a GMM in a microbial fermentation product, with characterization based on the detection of AMR genes and vectors, species and unnatural associations in the GMM genome, without isolating the GMM and without prior knowledge on possible GM elements present. Therefore, this approach mitigates the issues encountered for DNA walking or WGS-based approaches, as elaborated above. This was demonstrated with samples representative of the possible scenarios to occur in a routine setting, i.e. a previously analysed sample containing a GMM *B. subtilis* overproducing vitamin B2 (riboflavin), isolated and fully characterized at that time (RASFF 2014.1249) [57,59,58], a sample positive for some qPCR markers but for which no isolate could be obtained and a sample with no GMM contamination. Both short and long read sequencing were used, including the newly released Flongle as a smaller, more cost-effective alternative to the MinION. The most appropriate data analysis workflow was considered, depending on the sample type (quality of extracted DNA from processed matrix) and applied sequencing technology. The availability of appropriate sequence databases is crucial for this analysis. Theoretically, this method can replace the currently used qPCR first and second line analyses steps for GMM detection and identification in the enforcement labs. However, until the metagenomics approach is appropriately validated, currently it is more likely to be used by the enforcement laboratories as an orientation step, requiring subsequent confirmation of the findings by PCR and/or Sanger sequencing.

6. Conclusions, outlook and recommendations

For routine analysis, enforcement of GMO regulations by European laboratories has been almost exclusively performed by the use of Polymerase Chain Reaction (PCR) methods. This two-step approach includes an initial GMO screening step and a second GMO confirmation and/or quantification step by quantitative PCR (qPCR). For these methods, ENGL has formulated EU-wide recommendations that entail minimum quality performance criteria to ensure reliable methods and hence testing results for routine applications within EU member states [38]. In recent decades, however, the number of GMOs has increased, leading to a considerable number of PCR analyses that need to be performed for a single food or feed sample. These analyses are further complicated as the number of GMOs not containing any of the common screening elements has increased too, contributing to additional costs and complexity of the analysis. Another weakness of the current approach is that it primarily focuses on known GMOs and offers limited capability for identifying unknown ones. The increasing number of authorised and unauthorised GMOs present on the world market, and the advent of various genome-edited organisms has led scientists to look to DNA sequencing methodologies as additional or replacement tools to detect and identify GMOs.

The initial objective of the ENGL WG on DNA Sequencing was to likewise provide minimum quality performance criteria for the methods used for decoding DNA sequences of given samples in relation to GMO detection and identification. The aim of this WG was to draft guidance to ascertain the quality of sequencing data and of the results of sequencing strategies that are used for GMO detection and identification and molecular characterisation, as well as of the related data analysis workflow.

The initial objective of the WG had been to assess the MPPs and their AAVs for sequencing-based analyses alongside the guidelines of the European Union Reference Laboratory for GM Food and Feed (EURL GMFF), amongst others, with the aim of ascertaining the quality of DNA sequence data and of specific applications and strategies. During the initial meetings, however, it became clear that the objectives as formulated in the mandate, were too ambitious in the context of the current state of the art of the use of DNA sequencing strategies in GMO analysis. Whereas the current scientific developments in the field of GMO analysis are rapid and numerous, clearly underlining the added value of these DNA sequencing-based strategies and the related bioinformatics, the applications are still diverse, currently with little focus on aspects of validation and standardisation. Based on this notion, the WG has adjusted its objectives and this resulting draft guidance document by focusing on compiling available insights into quality aspects of methods for GMO detection and identification that include DNA sequencing steps, as a basis for future work on more detailed minimum performance requirements for these methods to be applied in routine GMO analysis.

The present report focuses on the current DNA sequencing methodologies, considering general quality aspects important for DNA isolation from the broad range of target food and feed samples, as well as for current strategies for specific DNA amplification, library synthesis, sequencing, and the primary data analysis. In addition, it assesses the quality aspects related to the current state of bioinformatics workflows for the interpretation of DNA sequencing data and database availability. Theoretically, it may be possible in the future to also quantify the presence of GMOs in a particular sample using sequencing strategies, but as it is not yet clear how this may be achieved in practice, these aspects have not been discussed further, and will require additional consideration in the future.

In addition to massive parallel DNA sequencing, in the present report Sanger sequencing is considered as a means to determine specific GMO-related single sequences with high accuracy. Quality aspects considered for Sanger sequencing focus on the purity and length of the amplicon to be sequenced and quality of the primers, on sequencing data and related data analysis workflows. For Sanger sequencing, it is advised to adhere to available guidelines as have been published by the JRC and EFSA [19,20].

In relation to specific GMO-related applications of massive parallel DNA sequencing, the working group identified four example scenarios covering real-life situations in GMO analysis in which the use of massive parallel DNA sequencing techniques overcomes bottlenecks encountered by the current qPCR-based approaches. These four scenarios comprise two targeted sequencing approaches resulting in the enrichment of the target(s) of interest: one focusing on multiple known sequences, and the other on partially known sequences; and two non-targeted sequencing approaches: one applying whole genome sequencing for completely unknown GMOs if a single GMO is concerned and the other applying metagenomics for the full sequencing of the genetic material present in a sample, with the purpose of screening for and identification of completely unknown GMOs.

The selected studies described in this report, as have been published in the scientific literature, illustrate that the sequencing workflow needs to be designed in relation to the intended purpose of the analysis (e.g. screening for multiple GM elements, identification of unknown GMOs and their corresponding inserted genetic element and its flanking regions, characterisation of the full genome of a GMM) as well the type of sample being analysed (e.g. simple versus complex matrix, known versus unknown GMOs). Therefore, some quality considerations and criteria are common to all scenarios and others are only relevant for specific ones, as described in the present report.

Within the new possibilities offered by massive parallel DNA sequencing, the whole genome sequencing of small genomes for GMO analysis of single organisms is already feasible in a cost-effective way. In this respect, it is important that the costs per sequence run are decreasing over time. In the near future, validated methods could be established for specific cases, for example for the identification of single GMM isolates in a cost-effective way. Massive parallel DNA sequencing may not only replace existing methods that are laborious, but may also provide additional precise information on inserted transgenes, single genetic modifications as well as their localisation in the genome. Additionally, sequencing approaches based on at least one known GM element also offer clear opportunities of detecting/identifying unauthorised GMOs. The use of DNA sequencing for the routine identification of single GM plants and animals is not yet practically feasible mainly due to genomic complexity and related high costs. Similarly, for control purposes for complex samples that may also entail the presence of unauthorised GMOs for which no prior sequence information is available, it is not yet feasible to use metagenomics sequencing approaches for GMO identification, except for GMM. At the moment, this technology will not replace the current GMO analytical strategies, but it should be considered as a valuable tool to be more effective and to gather crucial new information that would be missed by traditional tools. Indeed, it is clear that the targeted sequencing approaches could offer solutions for the screening of the ever-increasing number of authorized GMOs entering the market, some of which do not contain any of the common screening elements. In one single sequencing analysis, all elements could be screened for. If the elements of unauthorized GMOs are known, they can be included in this screening. In fact, with a customised combination of primers and optimal usage of sequencing kits, the costs (material and personnel) for a GMO screening approach using targeted sequencing could be comparable to a qPCR application, especially in cases where many GMOs are present in a single sample.

Besides the cost factor determining the current possibilities of massive parallel DNA sequencing for GMO analysis, another important aspect of the use of DNA sequencing tools and strategies for GMO analysis within the frame of enforcing EU GMO regulations, is the availability of DNA sequencing hardware and related bioinformatics infrastructure (hardware, suitable/appropriate databases and expertise) for all official European GMO analysis laboratories. This will also involve the availability of adequate training facilities and training opportunities for all personnel involved. Training should include the discussion and implementation of all quality aspects that have already been established, as presented and discussed in section 5 of this report, and regular updates thereof, as well as bioinformatics approaches. Training may relate to a specific scenario, e.g. focusing on targeted analysis using available sequence information, or rather to non-targeted analysis based on whole genome sequence analysis or metagenomics (see Figure 2).

In this report, it has been established that all aspects of DNA sequencing strategies for GMO detection and identification will require further harmonisation and standardisation in a timely and effective manner for optimised GMO analysis procedures. It is certain that all methods described in the present report, and their associated quality aspects, will undergo further developments in the (near) future. Developments in DNA sequencing platforms are progressing rapidly with the sequencing of large numbers and long stretches of DNA becoming increasingly feasible, resulting in vast amounts of good quality data generated in a single run. Bioinformatics workflows will need to follow suit and will allow step-by-step improvements in mining the DNA sequence information, directly related to the availability of appropriate databases and/or other references.

When considering quality aspects of GMO analysis strategies that include massive parallel DNA sequencing steps, the working group has formulated a number of observations and recommendations:

- Adequate methods for DNA isolation need to provide both sufficient amounts of DNA and DNA of appropriate integrity to allow subsequent sequencing. Based on this, appropriate enrichment strategies (when applied) may be further standardised to obtain sufficient numbers of the sequences of interest from the sample, thus optimising the chances that the analysis will lead to informative results.

- In order to use the technical capacity of current platforms to sequence long reads, it is essential to be able to preserve DNA integrity during the DNA isolation procedure. Currently, there are adequate DNA isolation procedures and kits to obtain long, intact DNA fragments from samples that have undergone limited processing and thus still contain DNA of high integrity. In highly processed samples, however, DNA will generally be degraded to the extent that the chances of obtaining DNA extracts that still contain sufficient numbers of long DNA stretches for sequencing will generally be limited. It will be important to determine the minimum DNA integrity parameters for the meaningful application of GMO analysis methodologies that include massive parallel DNA sequencing steps.

- The next step will be to select adequate DNA sequencing platforms that provide sufficient sequence data of good quality to enable downstream bioinformatics processing. Bioinformatics data analysis continues to advance and is currently being developed and tentatively validated for specific purposes. Here, however, it is necessary to further establish appropriate validation schemes for bioinformatics workflows to ensure accurate and reproducible analysis, either based on in-silico or real-life data. For GMO analysis laboratories, it may be beneficial to establish a shared bioinformatics workflow and a harmonised data management approach at the EU level, based on the recommendations and 'best practices' as have been summarized in section 4 of the present report, including criteria already established by ISO for methods for detection and identification of specific organisms that include DNA sequencing steps. The increased availability of reference genomes will be of benefit in this respect.

- In the near future, Machine Learning approaches, such as support vector machines (SVM) [46], artificial neural networks (ANN) [47], k-means and others, will be utilised to consolidate the growing number of sequenced genotypes into generalized models. This will allow differentiation between known (GMO + "wildtype") and unknown genotypes, thus significantly simplifying the analysis of complex samples, based on available knowledge of GMOs potentially present on the world market. It is important that these scientific developments are carefully monitored and structurally assessed for their applicability in routine GMO analysis as they may eventually overcome many of the current methodological bottlenecks.

- Nowadays, GMO-related databases have been established that compile all GMO-related data that is (publicly) available (see Table 1 in section 4.7). These databases allow the user to compile, compare and utilise available DNA sequence data for the development of validated methods for specific GMOs, as well as for the development of broader screening strategies. At the same time, available genome information on an increasing range of plants, animals and micro-organisms is rapidly expanding, and will be increasingly helpful for GMO analysis as well. To be able to use methodologies that include DNA sequencing steps for (routine) GMO analysis, the establishment and expansion of both well-curated and annotated GMO sequence databases and standardised bioinformatics workflows will be critical. The development of a first annotated GMO database allowing this analysis is an initial proof of concept in this direction [32]. Furthermore, it is advocated that, in order to achieve a high level of capability for GMO analysis across Europe, it will be necessary to create a central repository to share data as well as procedures used to create the shared data. This repository, for which current European database initiatives can well form the basis, would enable harmonised bioinformatics analyses and at the same time create an unparalleled data resource for the development of downstream applications for the identification of GMOs entering the European market.

- In close relation to the developments described above, it is crucial to generate additional experimental data, taking each of the critical steps described in sections 4 and 5 into account, with a focus on systematic validation data for each stage of the sequencing workflow. This will help establish relevant minimum performance requirements, comparable to those already established for the qPCR-based analytical methods for GMO testing. Quality control parameters should be agreed to ascertain that the new, sequencing-based methods are repeatable, reproducible and accurate. For this, research projects need to be tailored to generate appropriate data, tools and databases and attention must be paid to the transferability/interoperability of data collection tools and databases.

Other developments may become increasingly relevant for European GMO analysis methodologies. On a global level, one harmonised definition of a GMO is no longer applicable, especially for genome-edited organisms that contain minor modifications, such as single point mutations. For these organisms, there is no international consensus on whether or not they should fall within the scope of the GMO regional legislations. In this context, global discussions on the safety aspects and the traceability of these organisms are affected and the exchange of information on the (potential) presence of GMOs in food/feed samples and related raw materials has been impaired. This will generally affect the likelihood of detecting and identifying such genome-edited organisms on the market, as relevant sequence information will generally not be shared for those organisms that are not considered GMOs in other countries, but are considered GMOs under the EU legislation. Similarly, reference materials might not be available to establish validated methods for identification, and quantification, of these GMOs. This will affect GMO analysis strategies in general, including those that include DNA sequencing steps.

It will be necessary to re-think current procedures in light of these recent developments in plant and animal breeding, as well as in process technologies. Also, on the basis of these developments, it seems likely that the broader application of DNA sequencing strategies will form the core of GMO analysis for known, partly known or unknown unauthorised GMOs, as although these strategies require annotated databases, they do not necessarily require reference materials for identification purposes.

With the advent of new genomic techniques, especially prime editing, and given the various GMO legislations, as mentioned above, it is reasonable to hypothesise that the number of GMOs on the world market with (multiple) minor modifications (SNVs, indels of up to a few hundred base pairs), that have not (yet) been approved for the EU market will increase. This will make effective detection more difficult and massive parallel DNA sequencing could emerge as a useful method to help tackle this problem.

Finally, it is important to consider that DNA sequencing-based GMO detection strategies share many common principles with approaches currently used to identify food related pathogens or allergens, as well as approaches used to test for food authenticity or identify food adulteration. These approaches are dependent upon detection and identification of specific, well-established DNA sequences. In the near future, it may prove feasible to combine different analytical questions in a single DNA sequencing-based strategy. This will, however, require devoting further attention to some quality parameters, such as specificity of the analysis and validation of the results of the bioinformatics workflow. Development of efficient and versatile DNA enrichment strategies would be an important step in this respect, as well as the development of new, dedicated bioinformatics workflows that will be far more complex than the current initial versions.

References

1. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/DIS 23418 MICROBIOLOGY OF THE FOOD CHAIN – WHOLE GENOME SEQUENCING FOR TYPING AND GENOMIC CHARACTERIZATION OF FOODBORNE BACTERIA – GENERAL REQUIREMENTS AND GUIDANCE. HTTPS://WWW.ISO.ORG/CMS/RENDER/LIVE/EN/SITES/ISOORG/CONTENTS/DATA/STANDARD/07/55/75509.HTML.

2. HANDELSMAN, J. METAGENOMICS: APPLICATION OF GENOMICS TO UNCULTURED MICROORGANISMS. *MMBR* 68, 669–685 (2004).

3. HOLST-JENSEN, A. *ET AL.* DETECTING UN-AUTHORIZED GENETICALLY MODIFIED ORGANISMS (GMOS) AND DERIVED MATERIALS. *BIOTECHNOLOGY ADVANCES* 30, 1318–1335 (2012).

4. ARDUI, S., AMEUR, A., VERMEESCH, J. R. & HESTAND, M. S. SINGLE MOLECULE REAL-TIME (SMRT) SEQUENCING COMES OF AGE: APPLICATIONS AND UTILITIES FOR MEDICAL DIAGNOSTICS. *NUCLEIC ACIDS RESEARCH* 46, 2159–2168 (2018).

5. HARRIS, T. D. *ET AL.* SINGLE-MOLECULE DNA SEQUENCING OF A VIRAL GENOME. *SCIENCE* 320, 106–109 (2008).

6. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/DIS 22949-1 MOLECULAR BIOMARKER ANALYSIS – METHODS OF ANALYSIS FOR THE DETECTION AND IDENTIFICATION OF ANIMAL SPECIES IN FOODS AND FOOD PRODUCTS (NUCLEOTIDE SEQUENCING-BASED METHODS) – PART 1: GENERAL REQUIREMENTS. HTTPS://WWW.ISO.ORG/CMS/RENDER/LIVE/EN/SITES/ISOORG/CONTENTS/DATA/STANDARD/07/42/74231.HTML.

7. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/DIS 20397-1 BIOTECHNOLOGY — GENERAL REQUIREMENTS FOR MASSIVELY PARALLEL SEQUENCING — PART 1: NUCLEIC ACID

AND LIBRARY PREPARATION. HTTPS://WWW.ISO.ORG/CMS/RENDER/LIVE/EN/SITES/ISOORG/CONTENTS/DATA/STANDARD/07/40/74054.HTML.

8. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 20397-2:2021 BIOTECHNOLOGY — MASSIVELY PARALLEL SEQUENCING — PART 2: QUALITY EVALUATION OF SEQUENCING DATA. HTTPS://WWW.ISO.ORG/CMS/RENDER/LIVE/EN/SITES/ISOORG/CONTENTS/DATA/STANDARD/06/78/67895.HTML.

9. SANGER, F., NICKLEN, S. & COULSON, A. R. DNA SEQUENCING WITH CHAIN-TERMINATING INHIBITORS. *PROC NATL ACAD SCI U S A* 74, 5463–5467 (1977).

10. WINDELS, P., TAVERNIERS, I., DEPICKER, A., VAN BOCKSTAELE, E. & DE LOOSE, M. CHARACTERISATION OF THE ROUNDUP READY SOYBEAN INSERT. *EUR FOOD RES TECHNOL* 213, 107–112 (2001).

11. PAN, A. *ET AL.* EVENT-SPECIFIC QUALITATIVE AND QUANTITATIVE PCR DETECTION OF MON863 MAIZE BASED UPON THE 3′-TRANSGENE INTEGRATION SEQUENCE. *JOURNAL OF CEREAL SCIENCE* 43, 250–257 (2006).

12. YANG, L. *ET AL.* EVENT SPECIFIC QUALITATIVE AND QUANTITATIVE POLYMERASE CHAIN REACTION DETECTION OF GENETICALLY MODIFIED MON863 MAIZE BASED ON THE 5‘-TRANSGENE INTEGRATION SEQUENCE. *J. AGRIC. FOOD CHEM.* 53, 9312–9318 (2005).

13. FRAITURE, M.-A. *ET AL.* AN INNOVATIVE AND INTEGRATED APPROACH BASED ON DNA WALKING TO IDENTIFY UNAUTHORISED GMOS. *FOOD CHEM* 147, 60–69 (2014).

14. FRAITURE, M.-A. *ET AL.* INTEGRATED DNA WALKING SYSTEM TO CHARACTERIZE A BROAD SPECTRUM OF GMOS IN FOOD/FEED MATRICES. *BMC BIOTECHNOLOGY* 15, 76 (2015).

15. FRAITURE, M.-A. *ET AL.* NANOPORE SEQUENCING TECHNOLOGY: A NEW ROUTE FOR THE FAST DETECTION OF UNAUTHORIZED GMO. *SCIENTIFIC REPORTS* 8, 7903 (2018).

16. FRAITURE, M.-A. *ET AL.* MINION SEQUENCING TECHNOLOGY TO CHARACTERIZE UNAUTHORIZED GM PETUNIA PLANTS CIRCULATING ON THE EUROPEAN UNION MARKET. *SCIENTIFIC REPORTS* 9, 7141 (2019).

17. EWING, B. & GREEN, P. BASE-CALLING OF AUTOMATED SEQUENCER TRACES USING *PHRED.* II. ERROR PROBABILITIES. *GENOME RES.* 8, 186–194 (1998).

18. APPLIED BIOSYSTEMS. DNA SEQUENCING BY CAPILLARY ELECTROPHORESIS. APPLIED BIOSYSTEMS CHEMISTRY GUIDE | SECOND EDITION. PREPRINT AT HTTPS://TOOLS.THERMOFISHER.COM/CONTENT/SFS/MANUALS/CMS_041003.PDF (2009).

19. JOINT RESEARCH CENTER. EU REFERENCE LABORATORY FOR GENETICALLY MODIFIED FOOD AND FEED. GUIDELINE FOR THE SUBMISSION OF DNA SEQUENCES DERIVED FROM GENETICALLY MODIFIED ORGANISMS AND ASSOCIATED ANNOTATIONS WITHIN THE FRAMEWORK OF DIRECTIVE 2001/18/EC AND REGULATION (EC) NO 1829/2003. PREPRINT AT HTTP://GMO-CRL.JRC.EC.EUROPA.EU/GUIDANCEDOCS.HTM (2017).

20. EFSA PANEL ON GENETICALLY MODIFIED ORGANISMS (EFSA GMO PANEL) *ET AL.* TECHNICAL NOTE ON THE QUALITY OF DNA SEQUENCING FOR THE MOLECULAR CHARACTERISATION OF GENETICALLY MODIFIED PLANTS. *EFS2* 16, (2018).

21. DELAHAYE, C. & NICOLAS, J. SEQUENCING DNA WITH NANOPORES: TROUBLES AND BIASES. *PLOS ONE* 16, E0257521 (2021).

22. NI, Y., LIU, X., SIMENEH, Z. M., YANG, M. & LI, R. BENCHMARKING OF NANOPORE R10.4 AND R9.4.1 FLOW CELLS IN SINGLE-CELL WHOLE-GENOME AMPLIFICATION AND WHOLE-GENOME SHOTGUN SEQUENCING. *COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL* 21, 2352–2364 (2023).

23. HARDWICK, S. A., DEVESON, I. W. & MERCER, T. R. REFERENCE STANDARDS FOR NEXT-GENERATION SEQUENCING. *NAT REV GENET* 18, 473–484 (2017).

24. HENDRIKSEN, R. S. *ET AL.* FINAL REPORT OF ENGAGE - ESTABLISHING NEXT GENERATION SEQUENCING ABILITY FOR GENOMIC ANALYSIS IN EUROPE. *EFS3* 15, (2018).

25. SCHLABERG, R. *ET AL.* VALIDATION OF METAGENOMIC NEXT-GENERATION SEQUENCING TESTS FOR UNIVERSAL PATHOGEN DETECTION. *ARCHIVES OF PATHOLOGY & LABORATORY MEDICINE* 141, 776–786 (2017).

26. LAMBERT, D. *ET AL.* BASELINE PRACTICES FOR THE APPLICATION OF GENOMIC DATA SUPPORTING REGULATORY FOOD SAFETY. *JOURNAL OF AOAC INTERNATIONAL* 100, 721–731 (2017).

27. ANGERS-LOUSTAU, A. *ET AL.* THE CHALLENGES OF DESIGNING A BENCHMARK STRATEGY FOR BIOINFORMATICS PIPELINES IN THE IDENTIFICATION OF ANTIMICROBIAL RESISTANCE DETERMINANTS USING NEXT GENERATION SEQUENCING TECHNOLOGIES. *F1000RES* 7, 459 (2018).

28. BOGAERTS, B. *ET AL.* VALIDATION STRATEGY OF A BIOINFORMATICS WHOLE GENOME SEQUENCING WORKFLOW FOR SHIGA TOXIN-PRODUCING *ESCHERICHIA COLI* USING A REFERENCE COLLECTION EXTENSIVELY CHARACTERIZED WITH CONVENTIONAL METHODS. *MICROB GENOM* 7, (2021).

29. BOGAERTS, B. *ET AL.* A BIOINFORMATICS WHOLE-GENOME SEQUENCING WORKFLOW FOR CLINICAL MYCOBACTERIUM TUBERCULOSIS COMPLEX ISOLATE ANALYSIS, VALIDATED USING A REFERENCE COLLECTION EXTENSIVELY CHARACTERIZED WITH CONVENTIONAL METHODS AND IN SILICO APPROACHES. *J CLIN MICROBIOL* 59, E00202-21 (2021).

30. BOGAERTS, B. *ET AL.* VALIDATION OF A BIOINFORMATICS WORKFLOW FOR ROUTINE ANALYSIS OF WHOLE-GENOME SEQUENCING DATA AND RELATED CHALLENGES FOR PATHOGEN TYPING

IN A EUROPEAN NATIONAL REFERENCE CENTER: NEISSERIA MENINGITIDIS AS A PROOF-OF-CONCEPT. *FRONTIERS IN MICROBIOLOGY* 10, 362 (2019).

31. BORTOLAIA, V. *ET AL.* RESFINDER 4.0 FOR PREDICTIONS OF PHENOTYPES FROM GENOTYPES. *J ANTIMICROB CHEMOTHER* 75, 3491–3500 (2020).

32. JOENSEN, K. G. *ET AL.* REAL-TIME WHOLE-GENOME SEQUENCING FOR ROUTINE TYPING, SURVEILLANCE, AND OUTBREAK DETECTION OF VEROTOXIGENIC ESCHERICHIA COLI. *J CLIN MICROBIOL* 52, 1501–1510 (2014).

33. NCBI RESOURCE COORDINATORS. DATABASE RESOURCES OF THE NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. *NUCLEIC ACIDS RES* 46, D8–D13 (2018).

34. SALTYKOVA, A. *ET AL.* DETECTION AND IDENTIFICATION OF AUTHORIZED AND UNAUTHORIZED GMOS USING HIGH-THROUGHPUT SEQUENCING WITH THE SUPPORT OF A SEQUENCE-BASED GMO DATABASE. *FOOD CHEMISTRY: MOLECULAR SCIENCES* 4, 100096 (2022).

35. LEINONEN, R., SUGAWARA, H., SHUMWAY, M., & INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE COLLABORATION. THE SEQUENCE READ ARCHIVE. *NUCLEIC ACIDS RES* 39, D19-21 (2011).

36. LEINONEN, R. *ET AL.* THE EUROPEAN NUCLEOTIDE ARCHIVE. *NUCLEIC ACIDS RESEARCH* 39, D28–D31 (2011).

37. EUROPEAN NETWORK OF GMO LABORATORIES (ENGL). DETECTION OF FOOD AND FEED PLANT PRODUCTS OBTAINED BY NEW MUTAGENESIS TECHNIQUES. PREPRINT AT (2019).

38. EUROPEAN NETWORK OF GMO LABORATORIES (ENGL). DEFINITION OF MINIMUM PERFORMANCE REQUIREMENTS FOR ANALYTICAL METHODS OF GMO TESTING. PREPRINT AT (2015).

39. WILLEMS, S. *ET AL.* STATISTICAL FRAMEWORK FOR DETECTION OF GENETICALLY MODIFIED ORGANISMS BASED ON NEXT GENERATION SEQUENCING. *FOOD CHEM* 192, 788–798 (2016).

40. EFSA STATEMENT ON THE REQUIREMENTS FOR WHOLE GENOME SEQUENCE ANALYSIS OF MICROORGANISMS INTENTIONALLY USED IN THE FOOD CHAIN. *EFSA JOURNAL* 19, E06506 (2021).

41. DESAI, A. *ET AL.* IDENTIFICATION OF OPTIMUM SEQUENCING DEPTH ESPECIALLY FOR DE NOVO GENOME ASSEMBLY OF SMALL GENOMES USING NEXT GENERATION SEQUENCING DATA. *PLOS ONE* 8, E60204 (2013).

42. DENG, X., CAO, S. & HORN, A. L. EMERGING APPLICATIONS OF MACHINE LEARNING IN FOOD SAFETY. *ANNU. REV. FOOD SCI. TECHNOL* 12, 17–43 (2021).

43. CRUZ, J. A. & WISHART, D. S. APPLICATIONS OF MACHINE LEARNING IN CANCER PREDICTION AND PROGNOSIS. *CANCER INFORM* 2, 117693510600200030 (2006).

44. CAMACHO, D. M., COLLINS, K. M., POWERS, R. K., COSTELLO, J. C. & COLLINS, J. J. NEXT-GENERATION MACHINE LEARNING FOR BIOLOGICAL NETWORKS. *CELL* 173, 1581–1592 (2018).

45. NGUYEN, L., VAN HOECK, A. & CUPPEN, E. MACHINE LEARNING-BASED TISSUE OF ORIGIN CLASSIFICATION FOR CANCER OF UNKNOWN PRIMARY DIAGNOSTICS USING GENOME-WIDE MUTATION FEATURES. *NAT COMMUN* 13, 4013 (2022).

46. HUANG, S. *ET AL.* APPLICATIONS OF SUPPORT VECTOR MACHINE (SVM) LEARNING IN CANCER GENOMICS. *CANCER GENOMICS PROTEOMICS* 15, 41–51 (2018).

47. CHANG, P. *ET AL.* DEEP-LEARNING CONVOLUTIONAL NEURAL NETWORKS ACCURATELY CLASSIFY GENETIC MUTATIONS IN GLIOMAS. *AMERICAN JOURNAL OF NEURORADIOLOGY* 39, 1201–1207 (2018).

48. ARULANDHU, A. J. *ET AL.* NGS-BASED AMPLICON SEQUENCING APPROACH; TOWARDS A NEW ERA IN GMO SCREENING AND DETECTION. *FOOD CONTROL* 93, 201–210 (2018).

49. DEBODE, F. *ET AL.* DETECTION AND IDENTIFICATION OF TRANSGENIC EVENTS BY NEXT GENERATION SEQUENCING COMBINED WITH ENRICHMENT TECHNOLOGIES. *SCI REP* 9, 15595 (2019).

50. LIANG, C. *ET AL.* DETECTING AUTHORIZED AND UNAUTHORIZED GENETICALLY MODIFIED ORGANISMS CONTAINING VIP3A BY REAL-TIME PCR AND NEXT-GENERATION SEQUENCING. *ANAL BIOANAL CHEM* 406, 2603–2611 (2014).

51. FRAITURE, M.-A. *ET AL.* AN INTEGRATED STRATEGY COMBINING DNA WALKING AND NGS TO DETECT GMOS. *FOOD CHEMISTRY* 232, 351–358 (2017).

52. BOUTIGNY, A.-L., FIORITI, F. & ROLLAND, M. TARGETED MINION SEQUENCING OF TRANSGENES. *SCIENTIFIC REPORTS* 10, 15144 (2020).

53. GUIDANCE ON THE RISK ASSESSMENT OF GENETICALLY MODIFIED MICROORGANISMS AND THEIR PRODUCTS INTENDED FOR FOOD AND FEED USE. *EFSA JOURNAL* DOI:10.2903/J.EFSA.2011.2193.

54. MING, R. *ET AL.* THE DRAFT GENOME OF THE TRANSGENIC TROPICAL FRUIT TREE PAPAYA (CARICA PAPAYA LINNAEUS). *NATURE* 452, 991–996 (2008).

55. KOVALIC, D. *ET AL.* THE USE OF NEXT GENERATION SEQUENCING AND JUNCTION SEQUENCE ANALYSIS BIOINFORMATICS TO ACHIEVE MOLECULAR CHARACTERIZATION OF CROPS IMPROVED THROUGH MODERN BIOTECHNOLOGY. *THE PLANT GENOME* 5, (2012).

56. BARBAU-PIEDNOIR, E. *ET AL.* USE OF NEXT GENERATION SEQUENCING DATA TO DEVELOP A QPCR METHOD FOR SPECIFIC DETECTION OF EU-UNAUTHORIZED GENETICALLY MODIFIED BACILLUS SUBTILIS OVERPRODUCING RIBOFLAVIN. *BMC BIOTECHNOL* 15, 103 (2015).

57. BARBAU-PIEDNOIR, E. *ET AL.* GENOME SEQUENCE OF EU-UNAUTHORIZED GENETICALLY MODIFIED *BACILLUS SUBTILIS* STRAIN 2014-3557 OVERPRODUCING RIBOFLAVIN, ISOLATED FROM A VITAMIN B2 80% FEED ADDITIVE. *GENOME ANNOUNC.* 3, E00214-15, /GA/3/2/E00214-15.ATOM (2015).

58. PARACCHINI, V. *ET AL.* MOLECULAR CHARACTERIZATION OF AN UNAUTHORIZED GENETICALLY MODIFIED BACILLUS SUBTILIS PRODUCTION STRAIN IDENTIFIED IN A VITAMIN B 2 FEED ADDITIVE. *FOOD CHEMISTRY* 230, 681–689 (2017).

59. BERBERS, B. *ET AL.* COMBINING SHORT AND LONG READ SEQUENCING TO CHARACTERIZE ANTIMICROBIAL RESISTANCE GENES ON PLASMIDS APPLIED TO AN UNAUTHORIZED GENETICALLY MODIFIED BACILLUS. *SCIENTIFIC REPORTS* 10, 4310 (2020).

60. FRAITURE, M.-A. *ET AL.* IDENTIFICATION OF AN UNAUTHORIZED GENETICALLY MODIFIED BACTERIA IN FOOD ENZYME THROUGH WHOLE-GENOME SEQUENCING. *SCIENTIFIC REPORTS* 10, 7094 (2020).

61. D'AES, J. *ET AL.* CHARACTERIZATION OF GENETICALLY MODIFIED MICROORGANISMS USING SHORT- AND LONG-READ WHOLE-GENOME SEQUENCING REVEALS CONTAMINATIONS OF RELATED ORIGIN IN MULTIPLE COMMERCIAL FOOD ENZYME PRODUCTS. *FOODS* 10, 2637 (2021).

62. HUREL, J. *ET AL.* DUGMO: TOOL FOR THE DETECTION OF UNKNOWN GENETICALLY MODIFIED ORGANISMS WITH HIGH-THROUGHPUT SEQUENCING DATA FOR PURE BACTERIAL SAMPLES. *BMC BIOINFORMATICS* 21, 284 (2020).

63. METAGENOMICS SEQUENCING GUIDE. HTTPS://GENOHUB.COM/SHOTGUN-METAGENOMICS-SEQUENCING/ (2021).

64. SHANNON, C. E. A MATHEMATICAL THEORY OF COMMUNICATION. *BELL SYSTEM TECHNICAL JOURNAL* 27, 379–423 (1948).

65. BUYTAERS, F. E. *ET AL.* A SHOTGUN METAGENOMICS APPROACH TO DETECT AND CHARACTERIZE UNAUTHORIZED GENETICALLY MODIFIED MICROORGANISMS IN MICROBIAL FERMENTATION PRODUCTS. *FOOD CHEMISTRY: MOLECULAR SCIENCES* 2, 100023 (2021).

66. HOLST-JENSEN, A. *ET AL.* APPLICATION OF WHOLE GENOME SHOTGUN SEQUENCING FOR DETECTION AND CHARACTERIZATION OF GENETICALLY MODIFIED ORGANISMS AND DERIVED PRODUCTS. *ANAL BIOANAL CHEM* 408, 4595–4614 (2016).

List of abbreviations

AAV	Associated Acceptance Values
AMR	Antimicrobial Resistance
ANN	Artificial Neural Networks
BAM	Binary Alignment format
BLAST	Basic Local Alignment Search Tool
CJEU	Court of Justice of the European Union
CE	Capillary Electrophoresis
CRAM	Compressed Reference-oriented Alignment Map
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
dPCR	digital Polymerase chain reaction
EFSA	European Food Safety Agency
ENGL	European Network of GMO Laboratories
EU	European Union
EURL GMFF	European Union Reference Laboratory for GM Food and Feed
GM	Genetically Modified
GMM	Genetically Modified Microorganism
GMO	Genetically-Modified Organism
ISO	International Organization for Standardization
JRC	Joint Research Centre
LOD	Limit of Detection
LOQ	Limit of Quantification
ML	Machine Learning
MPP	Minimum Performance Parameters
MPS	Massive Parallel DNA Sequencing
NCBI	National Center for Biotechnology Information
NGT	New Genomic Techniques
NTC	No Template Control
ODM	Oligo-Directed Mutagenesis
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
qPCR	quantitative Polymerase Chain Reaction
Q score	Quality score
RT-PCR	Real-time PCR
SAM	Sequence Alignment format

SNV	Single Nucleotide Variant
SVM	Support Vector Machines
Tm	melting Temperature
UBAM	unmapped BAM format
VCF	Variant Calling Format
WG	Working Group
WGS	Whole Genome Sequence

List of figures and tables

Annexes

Annex 1: Sequencing platforms [f]

1. Traditional (Sanger) sequencing platforms

For standard DNA sequence analysis, the traditional Sanger method is considered the gold standard. Developed by Frederick Sanger and colleagues in 1977 [1], it has been the most widely used sequencing method for approximately 40 years. Sanger sequencing uses dideoxynucleotide base analogues to produce a pool of DNA molecules that are terminated at each residue of the analysed sequence, resulting in the formation of extension products of various lengths. Current DNA sequencing instruments make use of capillary electrophoresis and fluorescently labelled dideoxynucleotides to enable the separation and detection of DNA molecules originated from the analysed sequence. Each of the four dideoxynucleotides (ddA, ddC, ddT, ddG) are labelled with a different fluorescent dye, enabling detection of the extension products by laser-induced fluorescent emission during electrophoretic separation [2]. Different instruments offer different throughput for parallel sequencing of samples, and result in highly accurate and reliable sequencing data with read lengths of up to 1 kb. The low throughput and relatively high cost per base make Sanger sequencing suitable for small scale projects (Table A1).

2. Massive parallel DNA sequencing platforms

Massive parallel DNA sequencing encompasses both massively parallel and single-molecule sequencing, which respectively provide short and long sequencing reads. Short-read sequencing is highly accurate and produces read lengths of 100-300 bp. It uses an *in vitro* cloning step to amplify individual DNA molecules, because the detection systems used are not sensitive enough for single-molecule sequencing [3]. Short-read sequencing has been widely used, and has many applications, including targeted amplicon sequencing, whole genome sequencing, metagenomics, and transcriptomics. However, when it is required to sequence complete genomes or to determine complex genomic regions (e.g. repetitive regions), longer reads might be necessary. Long-read sequencing systems can produce reads up to > 300 kb in length, however at the cost of higher error rates compared to short-read sequencing. Which technology to apply depends on the foreseen use of the generated sequencing data and on the required throughput of the sequencing experiment. As sequencing technology is evolving, so are the instruments.

(a) Short-Read Sequencing Systems

Illumina sequencing systems

Illumina currently provides a range of different instruments[4], which differ in both throughput and read length (Table A1). Sequencing is performed using reversible-terminator sequencing-by-synthesis utilising a four colour or a two-colour reporter technology. Both single- and paired-end sequencing runs can be performed. Increased error frequencies in regions containing inverted repeats and GC-rich sequence motifs (GGC, GGT) have been reported for Illumina platforms [5,6].

[f] Correct as of March 2021.

Ion Torrent sequencing systems

The Ion Torrent massive parallel DNA sequencer developed by Ion Torrent (now ThermoFisher) is based on sequencing by synthesis principle. The detection principle exploits the releases of a hydrogen ion H^+ that occurs when a dNTP is added to a DNA polymer. The release of the hydrogen ions is measured using semiconductors that measures the associated change. This process takes place on a microchip, therefore millions of such changes can be measured simultaneously.

The semiconductor approach not only eliminates the need for light-based detection, but also enables a fast and relatively simple workflow in the laboratory. Instruments with fast runs are more suitable for routine diagnostics where time is of great importance (Table A1). The most common errors in Ion Torrent platforms were observed in homopolymer repeats, resulting in false identification of indels [7,8].

(b) Single-Molecule Long-read Sequencing Systems

Single-molecule sequencing technology allows sequencing of native DNA without cloning or amplification, and permits increased library fragment sizes and read lengths compared to short-read sequencing technology. With reads lengths typically exceeding 5 kb, long-read sequencing enables efficient genome assembly, simplifies phasing of structural genomic variants, and enables sequencing through repeats or complex genomic loci [2]. Currently two long-read sequencing technologies are commercially available, each having a unique sequencing approach: single-molecular real-time (SMRT) sequencing by Pacific Biosciences (PacBio) and nanopore sequencing by Oxford Nanopore Technologies.

PacBio SMRT sequencing systems

PacBio utilizes a sequencing-by-synthesis approach: a DNA polymerase incorporates fluorescently labelled nucleotides to a single DNA molecule, and the fluorescent signal is recorded in real-time by a camera. The sequencing reaction takes place in a SMRT® cell containing millions of wells called zero-mode waveguides (ZMWs). Each ZMW is a structure that contains a single DNA polymerase enzyme: this allows monitoring of the activity of the DNA polymerase at the single molecule level.

A unique feature of PacBio sequencing is that the library preparation creates a circular input molecule. Depending on the sequencing mode, the insert is either sequenced only once to create a very long read. If the insert is shorter than ~10 - 20 kb, each template can be sequenced multiple times. These multiple passes are used to generate a high-quality consensus sequence, known as a circular consensus sequence or HiFi read. Read accuracies of up to ~99.9 % can be achieved for sequences derived from combining several subreads. HiFi read lengths and output, however, may vary based on the sample quality and insert size.

The per-read error rate predominantly manifests itself as indel errors. These errors are randomly distributed within each read and hence sufficiently high coverage can overcome the high error rate.

Nanopore sequencing systems

Nanopore sequencing uses a molecular motor to transport an unknown DNA or RNA molecule through a nanoscale hole (nanopore) embedded in an electro-resistant membrane. An electric current is applied across this membrane. Changes in electrical conductivity arising from passing bases are measured as the molecule traverses the nanopore. The information about the change in

current is then used to identify the nucleobase. The nanopore can be created by proteins puncturing membranes (biological nanopores) or in solid materials (solid-state nanopores). This sequencing technology, in different flow cells and devices (Table A1), has been commercially released by Oxford Nanopore Technologies (ONT).

What is unique is that the flow cells have very low costs and a small size, offering extremely cheap start-up investments and portability. Furthermore, data is generated in real-time without having to wait until the run is complete. Also unique is the option for sequencing of polyadenylated RNA strands (direct RNA-seq) without the cDNA recoding and amplification biases inherent to other sequencing methodologies.

Nanopore flow cells produces a high error rate with many errors being sequence context dependent (homopolymers). However, the technology is constantly being improved and various ONT sequencing systems available to offer nanopore sequencing as a service (Table A1).

Table A1: Key performance indicators for a selection of Sanger, massive parallel DNA short read and long read sequencing platforms. [a]

Manufacturer	Instrument	Characteristics	Samples/ run	Max. read length	Output/ run	Run time	Reads/ run	Data Quality
ThermoFisher Scientific	ABI Genetic analyser 3730XL and 3500 series[b]	capillary Sanger sequencer	96	900 bp	<0,09 Gb	0.5-3 h	96	QV20
Illumina	MiSeq[c] (MiSeqDx)*	Short-read benchtop sequencer	1-96	2 x 300 bp (depending on reagents used)	0.3 - 15 Gb (depending on reagents used)	5.5-56 h (depending on required read length)	1-25 million (single); 2-50 million (paired-end)	>70 % of bases higher than Q30 (depending on reagents used)
Illumina	NextSeq 1000/2000[d]	Short-read compact production-Scale sequencer	1-96	2 x 150 bp (depending on reagents used)	40-360 Gb (depending on reagents used)	11-48 h (depending on required read length and reagents used)	400 million – 1.2 billion (single)	≥85 % of bases higher than Q30
Illumina	NovaSeq 6000[e]	Short-read large production-scale sequencer	1-96	2 x 250 bp	65-3000 Gb	13-44 h (depending on required read length)	0.65 – 10 billion (single)	≥75 % of bases higher than Q30
ThermoFisher Scientific	Ion PGM Dx	Suitable for smaller genome and targeted sequencing	1-16	200 bp	600 Mb-1 Gb	4.4 h	4-5.5 million	Not specified
ThermoFisher Scientific	Ion Proton	Suitable for smaller genome and targeted sequencing	1-96	200 bp	Max. 15 Gb	2-4h	60-80 million	Not specified
ThermoFisher Scientific	Ion GeneStudio S5 systems	Available in 3 versions, each supporting 5 different chips for various throughputs	1-96	200-600 bp**	0.3-25 Gb**	4.5-21.5 h**	2-130 million**	Not specified
Pacific Biosciences	Sequel II/IIe[f]	Long-read single molecule real-time sequencer	1-96	Variable±	up to 2 TB	Up to 30 h	Up to 6 million	>Q30 in Hifi mode

Manufacturer	Instrument	Characteristics	Samples/ run	Max. read length	Output/ run	Run time	Reads/ run	Data Quality
Oxford Nanopore Technologies	MinION Mk1B[g]	Long-read single molecule real-time sequencer	1-96	>4 Mb	Max. 50 Gb	1 min-72 h	Variable	98.3 % modal base accuracy
	Flongle[h] for use with MinION/GridION	Hardware adapter for use with MinION/GridION systems (long-read single molecule real-time sequencing). Designed for smaller/rapid tests.	1-96	>4 Mb	Up to 2.8 Gb	1 min – 16 h	Variable	98.3 % modal base accuracy
	PromethION 48[i]	Long-read single molecule real-time sequencer	1-96	>4 Mb	Max. 14 Tb	1 min-72 h	Variable	98.3 % modal base accuracy

Source: ENGL

a; Correct as of March 2021.

b; https://www.thermofisher.com/order/catalog/product/3730XL#/3730XL, accessed March 2021, as an example for Sanger sequencers; other instruments exists e.g. with lower capacity in number of capillaries.

c; https://emea.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/miseq-system-data-sheet-m-gl-00006/miseq-data-sheet-m-gl-00006.pdf, accessed March 2021.

d;https://emea.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/nextseq-1000-2000-spec-sheet-770-2019-030/nextseq-1000-2000-spec-sheet-770-2019-030.pdf, accessed March 2021.

e; https://emea.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/novaseq-6000-spec-sheet-770-2016-025/novaseq-6000-spec-sheet-770-2016-025.pdf, accessed March 2021.

f; https://www.pacb.com/products-and-services/sequel-system/, accessed March 2021.

g; https://nanoporetech.com/products/comparison?flongle=on&minion1b=on&promethion=on & https://nanoporetech.com/products/minion, accessed June 2021.

h; https://nanoporetech.com/products/comparison?flongle=on&minion1b=on&promethion=on & https://nanoporetech.com/products/flongle, accessed June 2021

i; https://nanoporetech.com/products/comparison?flongle=on&minion1b=on&promethion=on & https://nanoporetech.com/products/promethion, accessed June 2021

*MiSeqDx represents the first Food and Drug Administration (FDA)-regulated and Conformite Europeene in vitro diagnostic (CE-IVD)-marked platform for massive parallel DNA sequencing. It has been designed and optimised for use with in vitro diagnostic (IVD) assays only and is not intended for use with whole genome or de novo sequencing applications.

**Depending on chip type used.

± Read lengths are limited by the molecular fragment lengths in the sample.

References

1. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).

2. McCombie WR, McPherson JD, Mardis ER (2018) Next-generation sequencing technologies. *Cold Spring Harb. Perspect. Med*. doi: 10.1101/cshperspect.a036798

3. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. *Trends in Genetics* **34** (9):666-681

4. Lin J, Ao C, Zhang S, and Rhim J (2018) Illumina Health Care? Life Sciences Tools and Services. *Krause Fund Research* Fall 2018. Accessed on 25 May 2019. (https://tippie.uiowa.edu/sites/tippie.uiowa.edu/files/documents/krause/krause-fund-f18_ilmn.pdf).

5. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**:451

6. Schirmer M, Ljaz UZ, D'Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* **43**(6):e37

7. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* **30**(5):434-439

8. Yeo ZX, Chan M, Yap YS, Ang P, Rozen S, Lee ASG (2012) Improving indel detection specificity of the Ion Torrent PGM benchtop sequencer. *PLoS One* **7**(9):e45798

Annex 2: Databases containing information on insert structure and GMO approval status

GMO detection methods databases	
GMO Detection method Database (GMDD) [7] http://gmdd.sjtu.edu.cn/ *[no longer online]*	The database contained both DNA- and protein-based GMO detection methods. It was maintained by the GMO Detection Laboratory in Shanghai Jiao Tong University (GMODL-SJTU). The database provided detailed information on the methods, including primer sequences, amplicon length, endogenous reference gene primers, PCR programs, validation information and references. The database also contained information of GMO insertion sequences and certified reference materials. Users could upload their own methods and GMO inserted sequences, upon administrator's confirmation.
GMO Methods database - EU Database of Reference Methods for GMO Analysis http://gmo-crl.jrc.ec.europa.eu/gmomethods/	The database contains DNA-based GMO detection methods. It has been developed and is maintained by the Joint Research Centre as the European Union Reference Laboratory for GM Food and Feed (EURL GMFF), in collaboration with the European Network of GMO Laboratories (ENGL). The database provides detailed information on the methods, including primer sequences, amplicon length, endogenous reference gene, reaction setup, PCR programs, validation information and references. The database aims at providing a list of reference methods for GMO analysis that have been validated in a collaborative trial, according to the principles and requirements of ISO 5725 and/or IUPAC protocol or verified by the EURL GMFF in the context of compliance with an EU legislative act.

Source: ENGL

[7] Dong, W., Yang, L., Shen, K. et al. GMDD: a database of GMO detection methods. *BMC Bioinformatics* **9**, 260 (2008).

GMO registry databases	
ISAAA GM Approval Database http://www.isaaa.org/gmapprovaldatabase/default.asp	The database features the Biotech/GM crop events and traits that were approved for commercialization and planting and/or for import for food and feed use with a short description of the crop and the trait. It is maintained by the International Service for the Acquisition of Agri-biotech Applications (ISAAA). Entries in the database are sourced principally from Biotechnology Clearing House of approving countries and country regulatory websites.
GenBit GM crops database https://www.genbitgroup.com/en/gmo/gmodatabase/	The database lists the genetically engineered agricultural crops approved worldwide and includes information on the genetic elements present in the constructs. It is supported by GenBit, a Russian company. The database also includes information on the authorisation status of genetically modified crops in the Russian Federation and in the European Union. The database excludes flowers (carnation, petunia…), as well as lines for which no information is available on the transformation event and/or genetic elements of an insert (e.g., several sugar cane varieties).
Biosafety Clearing House (BCH) https://bch.cbd.int/database/organisms/	The Biosafety Clearing House (BCH) database is used for the international exchange of information based on the Cartagena Protocol treaties on biosafety. The database contains worldwide entries on authorisations of GMOs providing access to a variety of scientific, technical, environmental, legal and capacity building information.

Source: ENGL

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
Joint-research-centre.ec.europa.eu

Publications Office
of the European Union