Database, 2015, 1–11
doi: 10.1093/database/bav101
Original Article

Original Article

JRC GMO-Amplicons: a collection of nucleic acid sequences related to genetically modified organisms

Mauro Petrillo[*,†], **Alexandre Angers-Loustau**[†], **Peter Henriksson**, **Laura Bonfini**, **Alex Patak** and **Joachim Kreysa**

Molecular Biology and Genomics Unit, Joint Research Centre, European Commission, Ispra, Italy

[*] **Corresponding author**: Telephone: +39 0332 786232, Fax: +39 0332785483; Email: mauro.petrillo@ec.europa.eu

[†] These authors contributed equally to this work.

Citation details: Petrillo,M., Angers-Loustau,A., Henriksson,P. *et al.* JRC GMO-Amplicons: a collection of nucleic acid sequences related to genetically modified organisms. *Database* (2015) Vol. 2015: article ID bav101; doi:10.1093/database/bav101

This work is dedicated to the memory of our colleague Cristian Geron, who suddenly left us on February 2015. *Ciao Cristian*.

Received 27 May 2015; Revised 9 July 2015; Accepted 10 September 2015

Abstract

The DNA target sequence is the key element in designing detection methods for genetically modified organisms (GMOs). Unfortunately this information is frequently lacking, especially for unauthorized GMOs. In addition, patent sequences are generally poorly annotated, buried in complex and extensive documentation and hard to link to the corresponding GM event. Here, we present the JRC GMO-Amplicons, a database of amplicons collected by screening public nucleotide sequence databanks by *in silico* determination of PCR amplification with reference methods for GMO analysis. The European Union Reference Laboratory for Genetically Modified Food and Feed (EU-RL GMFF) provides these methods in the GMOMETHODS database to support enforcement of EU legislation and GM food/feed control. The JRC GMO-Amplicons database is composed of more than 240 000 amplicons, which can be easily accessed and screened through a web interface. To our knowledge, this is the first attempt at pooling and collecting publicly available sequences related to GMOs in food and feed. The JRC GMO-Amplicons supports control laboratories in the design and assessment of GMO methods, providing inter-alia *in silico* prediction of primers specificity and GM targets coverage. The new tool can assist the laboratories in the analysis of complex issues, such as the detection and identification of unauthorized GMOs. Notably, the JRC GMO-Amplicons database allows the retrieval and characterization of GMO-related sequences included in patents documentation. Finally, it can help annotating poorly described GM sequences and identifying new relevant GMO-related sequences in public databases. The JRC GMO-Amplicons is freely accessible through a web-based portal that is hosted on the EU-RL GMFF website.

Database URL: http://gmo-crl.jrc.ec.europa.eu/jrcgmoamplicons/

Introduction

The European Union (EU) has defined a legal framework that strictly regulates cultivation of genetically modified organisms (GMOs) and their marketing as food and feed (1–3). Implementation of EU regulations on GMOs and their traceability requirements strictly depends on the availability of reliable and reproducible analytical methods. According to Article 32 of Regulation (EC) N. 882/2004 (4) the European Union Reference Laboratory for Genetically Modified Food and Feed (EU-RL GMFF) must provide National Reference Laboratories (NRLs) with reference methods for GMO analysis. To this end the EU-RL GMFF, hosted by the Joint Research Centre (JRC) of the European Commission, has compiled a state-of-the-art catalogue of DNA-based methods and made them freely accessible in its Database of Reference Methods called GMOMETHODS (5). The GMOMETHODS database only includes methods that have been validated through a collaborative trial according to the international standard ISO 5725 (6) and/or the IUPAC protocol (Union of Pure and Applied Chemistry, 7). These methods have demonstrated performance characteristics in line with the criteria established by the European Network of GMO laboratories (ENGL) and can therefore assure reliable, reproducible, sensitive and accurate determination of GMO content in food and feed. Other applications, such as the GMO Detection Method Database (GMDD, 8), do not perform such a stringent selection and cannot therefore assure the same performance reliability for the methods that they provide.

The GMOMETHODS detection methods are currently based on the polymerase chain reaction (PCR) technology and are grouped according to the specificity of the target sequence:

- *Event-specific* methods have the highest levels of specificity, as they target regions that are unique to each DNA integration event.
- *Construct-specific* methods are specific for the transgenic constructs, as they target a region that spans artificially joined sequences, such as a coding sequence and the neighbouring control regions.
- *Element-specific* methods target DNA sequences that are solely confined to one particular sequence element of a GMO, such as the terminator of the *nopaline synthase* gene of *Agrobacterium tumefaciens* (T-nos) or a gene coding region.
- *Taxon-specific* (or *species-specific*) methods target DNA sequences that are specific to a particular taxon or species. They are used for defining the composition and relative GM content of a food/feed product at ingredient level.

For the analysis of GMOs in the food and feed chains, the GMO-control laboratories generally perform an initial 'screening' step. In order to reduce the list of possible GMOs present, a set of element- or construct-specific methods (also called *screening methods*) is used for detecting target sequences commonly present in known GM events. In case of positive signals, event-specific methods are employed to identify the GMOs containing the previously detected GM elements. Moreover, to assist the GMO-control laboratories in the whole screening phase, information matrices of the expected PCR amplification patterns of known GMOs for a defined range of element- and construct-specific detection methods have been developed. Some of the matrices available online, such as GMOseek (9) and GMOfinder (10), integrate information from laboratory testing and publicly available sources, such as public versions of dossiers submitted by the applicants to regulatory authorities. However, their predictions are not always reliable as elements with the same name can vary at the level of their DNA sequences in different GMOs, and a set of primers and probes designed for detecting a specific element may or may not efficiently amplify a slightly different version of the same element. Moreover, the DNA sequences themselves may not always be readily available for allowing such predictions. To avoid these shortcomings we have recently made available the JRC GMO-Matrix, a matrix-based decision support system that is unique for its reliance on DNA sequence data (11). This new interactive tool relies on data stored in the Central Core DNA Sequences Information System (CCSIS) (12), a database of annotated GM event sequences retrieved from public databases or provided by the applicants for the EU authorization procedure and therefore covering, for each event, the entire insert with their respective flanking regions. A direct comparison of the GM and methods sequences allows a more reliable *in silico* prediction of the PCR results for every combination of reference method (primers/probes) and GM event.

Given the confidential nature of some of these sequences information, the CCSIS database is not directly publicly available but processed, in the JRC GMO-Matrix, in the form of precomputed values of PCR predictions. The dependency of these simulations on the availability of complete and accurate GM sequences can limit the coverage of GM events found in the JRC GMO-Matrix, in particular for unknown or poorly characterized GMOs. For these, information can be complemented by other sources of information, e.g. by compiling experimental test results or by mining public sequence databases for transgenic targets. GMO-related sequence data may be stored in public sequence repositories as a contribution from published literature, grouped sometimes in broad sections, such as the transgenic (*tgn*) data class division of the European Nucleotide Archive (ENA) (13). Patented sequence datasets are another putative source, as GMOs are often patented. Sequence data present in patents are usually retrieved and grouped in the public databases, such as the NCBI *patnt* collection (14) or the *pat* data class division of ENA (13), but are characterized by a high level of redundancy. Few attempts have been made to address this redundancy, such as the development of the *nrnl1* database (15). Moreover, patented sequences are poorly annotated and can be linked to the corresponding GMO only by extensive review of the original documentation.

To our knowledge, there are currently no databases aimed specifically at storing GMO-related sequences. Here, we present such a resource, called JRC GMO-Amplicons. JRC GMO-Amplicons contains and makes available results of a bioinformatic pipeline that regularly screens public nucleotide sequence databanks, including patents and available whole plant genomes, through *in silico* determination of PCR amplification with the reference methods of the GMOMETHODS database.

The JRC GMO-Amplicons can be queried by control laboratories to evaluate results of their screening/identification analysis or for developing new detection methods and assessing *in silico* their primers specificity and GMO coverage. Moreover, with the advent of next-generation sequencing (NGS), the availability of an up-to-date source of sequences related to GMOs is crucial for the molecular characterization of GM events. Finally, the JRC GMO-Amplicons database can aid in annotating poorly described GM events or in recognizing new GM-related sequences in public repositories.

Methods

Platform deployment

The platform supporting the JRC GMO-Amplicons functions comprises three servers: a server dedicated to the database (PostgreSQL, http://www.postgresql.org/) storing all DNA sequences and related information, a server hosting a Ruby on Rails web application (http://rubyonrails.org/) that acts as a user interface, and a high-performance computer that regularly runs a compute-intensive pipeline.

Data and resources

The GMOMETHODS database includes, as of January 2015, 140 different PCR methods, specifically 60 event-specific, 20 construct-specific, 26 element-specific and 37 taxon-specific methods.

In the current implementation, the pipeline runs on a regular basis (i.e. every 3 months) on the following public data sets:

- EBI non-redundant patent nucleotide database *nrnl1*, available as a single file at ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/patent/nrnl1.gz (release of 2 January 2013)
- the NCBI *nt* collection, available as a single file at ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nt.gz (release of 6 December 2014)
- the NCBI patnt collection, available as a single file at ftp://ftp.ncbi.nih.gov/blast/db/FASTA/patnt.gz (release of 6 December 2014)
- the *pat* (patent) data class division of the ENA, available as multiple em_rel_pat_*.gz files at ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/emblrelease/ (release 121)
- the plant *conexp* (contig expanded) class division of the ENA, available as multiple rel_exp_con_pln_*.dat.gz files at ftp://ftp.ebi.ac.uk/pub/databases/ena/sequence/release/expanded_con (release 121)
- the *tgn* (transgenic) data class division of the ENA, available as multiple em_rel_*_tgn.gz files at ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/emblrelease/ (release 121)

These data sets are regularly downloaded and installed as local copies for the analysis and updated when new release versions become available.

Pipeline

The pipeline was written in the PHP5 (http://php.net) object-oriented programming language. Given a set of nucleic acid sequences and the array of reference methods in the GMOMETHODS database, the pipeline sequentially performs the tasks listed below:

1. *In silico* PCR simulation on the whole sequence collection with the primers pairs of the reference methods, by running *e-PCR* (16), a tool developed at NCBI and installed locally. *e-PCR* is performed with the following options: -d50-500 (-*d*='default size range'), -n2 (-*n*='Max mismatches allowed per primer'), -g2 (-*g*='Max number of

indels allowed per primer'), -t3 (*-t*='output in tabular format'). Amplicons are selected and further processed in the following steps only if the total number of mismatches and gaps is less or equal to 2.
2. Retrieval of the matching sequences, annotation of their record ID and extraction of the relative sub-regions (amplicons) obtained by *e-PCR*;
3. If a probe is present, alignment of the probe and amplicon sequences is performed by *blast2seq*, a sequence comparison tool of the NCBI-BLAST package (17);
4. Species assignment of the amplicon sequences by querying the Taxonomy database (18, 19) with the target record ID;
5. Generation of the output as TAB delimited file with all metadata to populate/update the database.

Web interface construction and features

The JRC GMO-Amplicons interface is a Ruby on Rails application (http://rubyonrails.org/) integrated within the EU-RL GMFF public website (http://gmo-crl.jrc.ec.europa.eu/). It is backed by a local PostgreSQL database which is automatically updated by local scripts that parse the TAB delimited files generated in the annotation pipeline above.

The application also integrates, as a frame, a local copy of the SequenceServer software (Priyam A, Woodcroft BJ, Rai V & Wurm Y SequenceServer: BLAST searching made easy, in preparation, http://sequenceserver.com) to add the BLAST-based similarity search functionality. Modifications to the original source code (version 0.8.7) were limited to visual elements, color scheme and security improvements. BLAST datasets are generated using the *makeblastdb* command of the NCBI-BLAST package (17).

Clustering

Clustering of amplicon sequences in order to generate the non-redundant amplicon set used for BLAST is accomplished by running *cd-hit*, version V4.6.1 (20) with the following options: -aL = 0.9 (alignment coverage for the longer sequence), -aS = 0.9 (alignment coverage for the shorter sequence).

Results and discussion

Analysis of pipeline output

Using the reference methods included in the GMOMETHODS database and the sequences from the selected public resources, the pipeline produced a set of more than 240 000 amplicons, which ultimately populate the JRC GMO-Amplicons database. Out of the available 140 methods in the GMOMETHODS database, 132 (94%) were able to pick up at least one sequence in the sequence datasets and 128 (91%) had a perfect match in the annealing region of the primers with their corresponding DNA target. The set of detected amplicons sequences can be clustered in 722 distinct groups each composed of 100% identical sequences (184 groups by setting both 90% identity and 90% sequence length coverage for each sequence pair comparison). This set of 722 sequences defines a 'non-redundant' set of amplicons that can be queried by similarity search tools.

Element-specific methods are expected to target many sequences, including those of the elements' donor organisms, in all the input datasets. In contrast, construct- and (to a larger extent) event-specific methods are expected to target fewer sequences, which should be found mostly in the patent-related datasets. In order to verify the consistency of the matches between the datasets and methods we used, amplicons were analysed with respect to the dataset of origin and the type of methods used to detect them. These analyses revealed a non-random distribution of the amplicons consistent with both the method type and the nature of the datasets, in agreement with our hypothesis, as highlighted in Figure 1.

Figure 1A shows the percentage of methods, classified by specificity, that generate *in silico* at least one amplicon in the different sequence datasets. It can be observed that:

- GM-specific methods (i.e. event-, construct-, element-specific) do not detect any sequence in the ENA plants contigs dataset (*emblconexp* bars in Figure 1A), where GM-sequences are not expected to be present.
- Conversely, taxon-specific methods do not detect any sequence in the ENA transgenic dataset (*emblreltgn* bars in Figure 1A), where plant genomic sequences are presumably not occurring.
- Most of the event-specific methods find putative amplicons in the patent-related sets (i.e. *patnt*, *emblrelpat*, *nrnl1* bars in Figure 1A) and only 20% in the NCBI nt section, thus reflecting the fact that GMO sequences are usually patented and not deposited in the publicly available DNA sequence resources.
- All the element-specific methods, that are not specific for individual GMOs, detect sequences in all the analysed datasets except the ENA plants contigs (*emblconexp* bar in Figure 1A).
- The distribution for each type of GM-specific method (i.e. event-, construct-, element-specific) is practically identical in the three patent-related datasets (i.e. *patnt*, *emblrelpat*, *nrnl1*), with the exception of three recently added event-specific detection methods (QT-EVE-GM-012, QT-EVE-GM-013 and QT-EVE-GM-001), whose

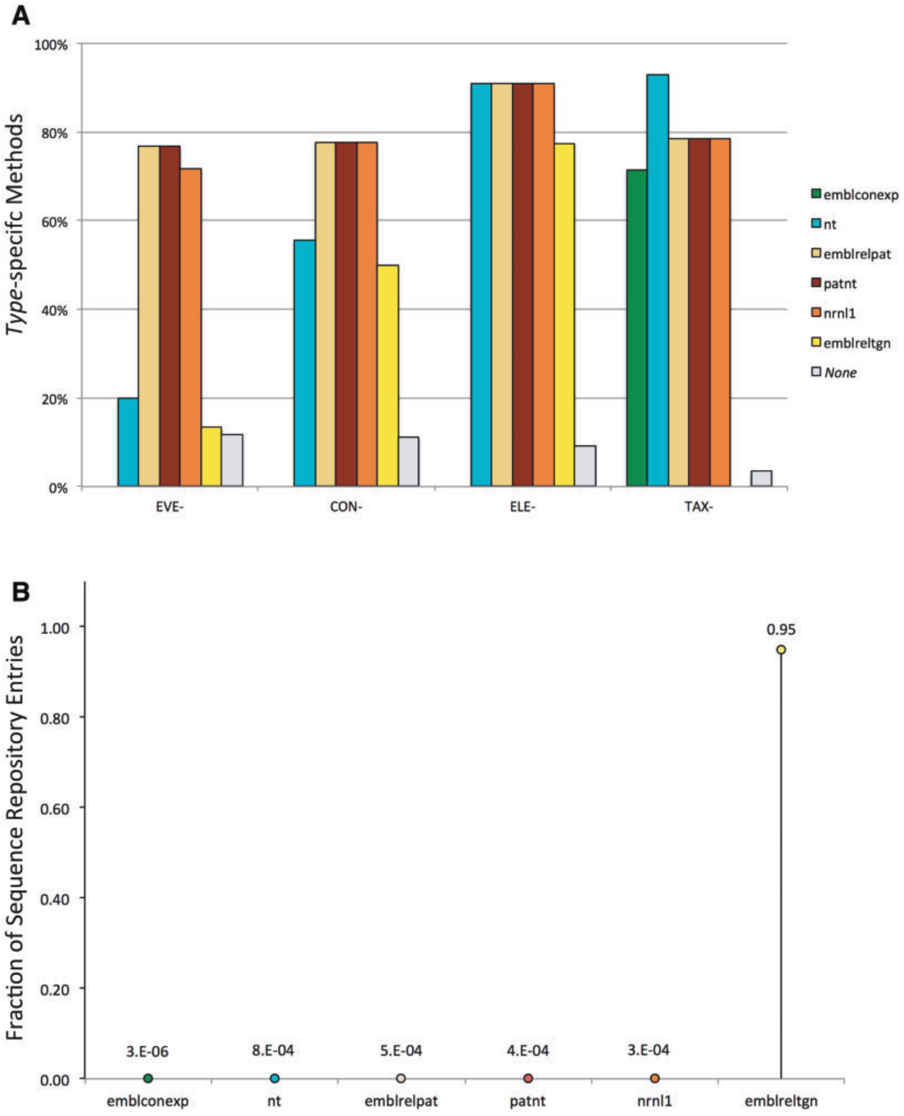


Figure 1. Evaluation of the consistency of the pipeline output. Detected amplicons are evaluated with respect to the dataset of origin and the type of methods used to detect them, revealing a non-random distribution of the amplicons consistent with both the method type and the dataset composition. (**A**) Percentage of methods from the GMOMETHODS database, grouped by specificity, that generate *in silico* at least one amplicon in the different sequence datasets; (**B**) Fraction of records in each of the analysed datasets that can be detected by the methods from the GMOMETHODS database. Only amplicons with 0 gaps and 0 mismatches in primers and probe annealing were taken into account.

targets were not present at the time of the last *nrnl1* release (2013).

Figure 1B, conversely, shows the fraction of records in each of the analysed datasets that can be detected by the methods from the GMOMETHODS database. Consistent with the nature of these datasets, it can be shown that:

- 95% (19 812 out of 20 800) of the *emblreltgn* records are detected by at least one of the GM-specific methods. This suggests that the methods available in the GMOMETHODS database cover 95% of the public nucleic acid sequences classified as transgenic. This fraction is considerably higher than in the other datasets, where only an insignificant portion (on average about 1/10 000) were detected.
- Plant contigs records (*emblconexp* dataset) are not expected to be related to GMOs. In fact, only few sequences (1/1 000 000) are positive to detection methods and all of these are, as expected, taxon-specific methods.

Some records were found to produce an amplicon with more than one method. Once again, this trend differs according to the sequence dataset and the specificity of the method. As shown in Figure 2 event-specific (A), element-specific (B), construct-specific (C) and taxon-specific (D)

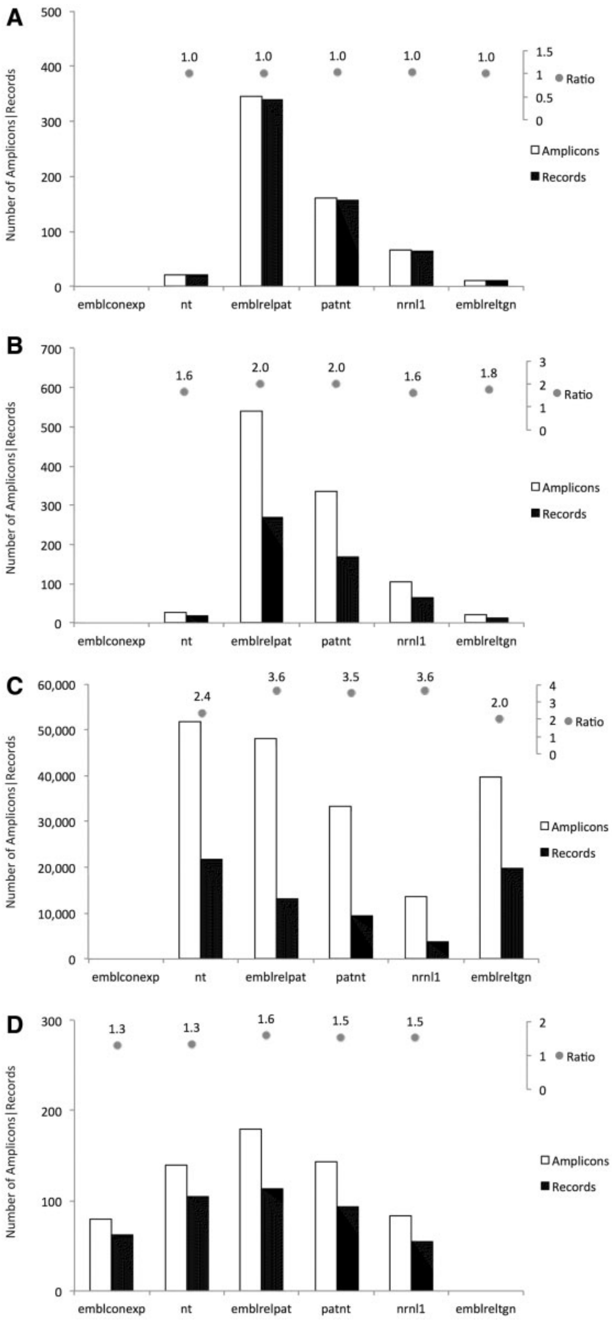


Figure 2. Number of amplicons and number of target records found in the selected datasets. For each type of detection methods, amplicons and target records are plotted as bars: (**A**) Event-specific methods; (**B**) Construct-specific methods; (**C**) Element-specific methods; (**D**) Taxon-specific methods. On top of each panel, the ratio of these numbers is also reported. Only amplicons with 0 gaps and 0 mismatches in primers and probe annealing were taken into account.

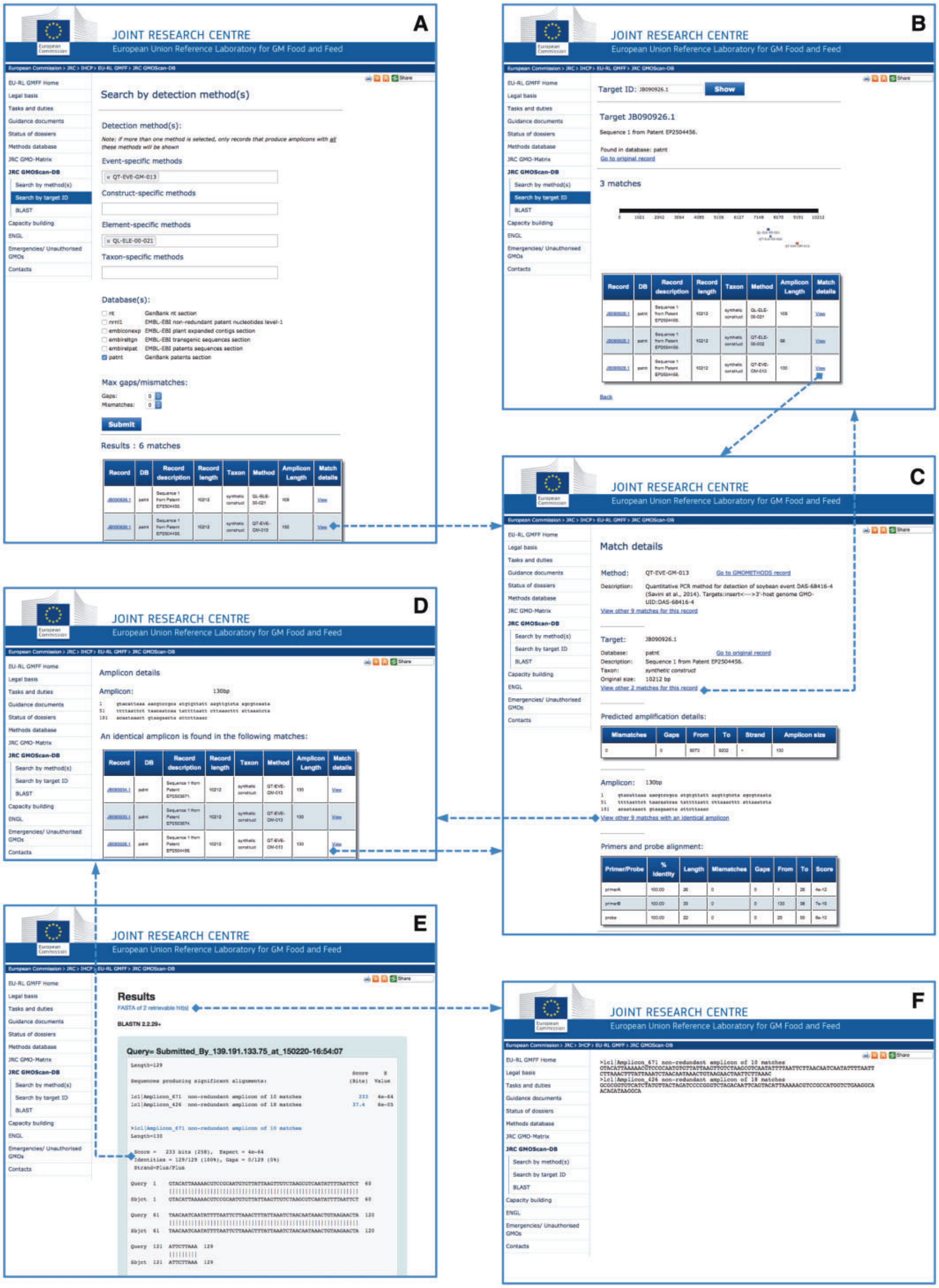
methods display distinctive patterns of PCR amplification predictions in the different datasets:

- Event-specific methods produced a single amplicons for each retrieved record (Figure 2A), thus in a 1:1 relationship, as they target regions that are unique in each record. The only exceptions relate to different methods targeting the same GM event, as for example in record nrnl1:NRN_AX685147 which was retrieved by the QL-EVE-BN-002 and QT-EVE-BN-004 methods, both specific for the detection of oilseed rape event GT73.
- Element-specific methods produced *in silico* hundreds of thousands of amplicons (about 1,000 times more than the other ones, see Figure 2B) as they were designed to detect sequences of elements commonly used in GMOs development. Moreover, a single record can be targeted by different element-specific methods if its recombinant molecule contains more than one element. As expected, element-specific methods target a lot of sequences that correspond to 'natural elements', i.e. elements existing in the sequence of their donor organism outside of their use in GMOs. For example, more than 40 000 amplicons were found for the two detection methods against the neomycin phosphotransferase II (*nptII*) gene, identifying both the natural counterpart and a panel of cloning vectors.
- Construct-specific methods (Figure 2C) share behaviour similar to the element-specific methods, but since they target the junction between elements in synthetic constructs, they generate a smaller number of amplicons. Moreover, most of the records are retrieved from patent-related datasets, as synthetic constructs are often patented.
- Taxon-specific methods (Figure 2D) may, in some instances, target the same record, since different regions of the same gene have sometimes been used for the development of multiple methods. One example is the soybean lectin gene (*Le1*), where distinct regions are targeted by different soybean-specific methods.

All the above observations showing the non-random but logical selection of records indicate a successful generation of a consistent, enriched, GMO-related set of sequences. Moreover, according to these findings, the number of false positives and negatives is expected to be very low.

Database access

The JRC GMO-Amplicons is accessible through a web portal (Figure 3), where queries can be performed according to the methods (Figure 3A) or the identification number of the target record (Figure 3B). The system provides a list of matches, and in the latter case a graphical representation of the amplicons mapped on the target. From the results page it is possible to inspect the match details (Figure 3C) or further examine the 'amplicon detail page' (Figure 3D), i.e. a page reporting a list of identical amplicons retrieved from all the records of the scanned sequence resources.



A

Search by detection method(s)

Detection method(s):

Note: if more than one method is selected, only records that produce amplicons with all these methods will be shown

Event-specific methods: × QT-EVE-GM-013

Construct-specific methods:

Element-specific methods: × QL-ELE-00-021

Taxon-specific methods:

Database(s):

□ nt	GenBank nt section
□ nrnl1	EMBL-EBI non-redundant patent nucleotides level-1
□ emblconexp	EMBL-EBI plant expanded contigs section
□ emblreltgn	EMBL-EBI transgenic sequences section
□ emblrelpat	EMBL-EBI patents sequences section
☑ patnt	GenBank patents section

Max gaps/mismatches:

Gaps: 0

Mismatches: 0

Results : 6 matches

Record	DB	Record description	Record length	Taxon	Method	Amplicon Length	Match details
JB090926.1	patnt	Sequence 1 from Patent EP2504456.	10212	synthetic construct	QL-ELE-00-021	108	View
JB090926.1	patnt	Sequence 1 from Patent EP2504456.	10212	synthetic construct	QT-EVE-GM-013	130	View

B

Target ID: JB090926.1

Target JB090926.1

Sequence 1 from Patent EP2504456.

Found in database: patnt

Go to original record

3 matches

Record	DB	Record description	Record length	Taxon	Method	Amplicon Length	Match details
JB090926.1	patnt	Sequence 1 from Patent EP2504456.	10212	synthetic construct	QL-ELE-00-021	108	View
JB090926.1	patnt	Sequence 1 from Patent EP2504456.	10212	synthetic construct	QT-ELE-00-002	98	View
JB090926.1	patnt	Sequence 1 from Patent EP2504456.	10212	synthetic construct	QT-EVE-GM-013	130	View

C

Match details

Method: QT-EVE-GM-013 Go to GMOMETHODS record

Description: Quantitative PCR method for detection of soybean event DAS-68416-4 (Savini et al., 2014). Targets:insert<--->3'-host genome GMO-UID:DAS-68416-4

View other 9 matches for this record

Target: JB090926.1

Database: patnt Go to original record

Description: Sequence 1 from Patent EP2504456.

Taxon: synthetic construct

Original size: 10212 bp

View other 2 matches for this record

Predicted amplification details:

Mismatches	Gaps	From	To	Strand	Amplicon size
0	0	9073	9202	+	130

Amplicon: 130bp

```
1     gtacattaaa aacgtccgca atgtgttatt aagttgtcta agcgtcaata
51    ttttaattct taacaatcaa tattttaatt cttaaacttt attaaatcta
101   acaataaact gtaagaacta attcttaaac
```

View other 9 matches with an identical amplicon

Primers and probe alignment:

Primer/Probe	% Identity	Length	Mismatches	Gaps	From	To	Score
primerA	100.00	26	0	0	1	26	4e-12
primerB	100.00	33	0	0	130	98	7e-15
probe	100.00	22	0	0	29	50	6e-10

D

Amplicon details

Amplicon: 130bp

```
1     gtacattaaa aacgtccgca atgtgttatt aagttgtcta agcgtcaata
51    ttttaattct taacaatcaa tattttaatt cttaaacttt attaaatcta
101   acaataaact gtaagaacta attcttaaac
```

An identical amplicon is found in the following matches:

Record	DB	Record description	Record length	Taxon	Method	Amplicon Length	Match details
JB090934.1	patnt	Sequence 1 from Patent EP2503871.	10212	synthetic construct	QT-EVE-GM-013	130	View
JB090933.1	patnt	Sequence 1 from Patent EP2503874.	10212	synthetic construct	QT-EVE-GM-013	130	View
JB090926.1	patnt	Sequence 1 from Patent EP2504456.	10212	synthetic construct	QT-EVE-GM-013	130	View

E

Results

FASTA of 2 retrievable hit(s)

BLASTN 2.2.29+

Query= Submitted_By_139.191.133.75_at_150220-16:54:07

```
Length=129
                                                              Score     E
Sequences producing significant alignments:                   (Bits)  Value

lcl|Amplicon_671  non-redundant amplicon of 10 matches          233   4e-64
lcl|Amplicon_426  non-redundant amplicon of 18 matches         37.4   6e-05


>lcl|Amplicon_671 non-redundant amplicon of 10 matches
Length=130

 Score =  233 bits (258),  Expect = 4e-64
 Identities = 129/129 (100%), Gaps = 0/129 (0%)
 Strand=Plus/Plus

Query  1    GTACATTAAAAACGTCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATATTTTAATTCT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    GTACATTAAAAACGTCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATATTTTAATTCT  60

Query  61   TAACAATCAATATTTTAATTCTTAAACTTTATTAAATCTAACAATAAACTGTAAGAACTA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TAACAATCAATATTTTAATTCTTAAACTTTATTAAATCTAACAATAAACTGTAAGAACTA  120

Query  121  ATTCTTAAA  129
            |||||||||
Sbjct  121  ATTCTTAAA  129
```

F

```
>lcl|Amplicon_671 non-redundant amplicon of 10 matches
GTACATTAAAAACGTCCGCAATGTGTTATTAAGTTGTCTAAGCGTCAATATTTTAATTCTTAACAATCAATATTTTAATT
CTTAAACTTTATTAAATCTAACAATAAACTGTAAGAACTAATTCTTAAAC
>lcl|Amplicon_426 non-redundant amplicon of 18 matches
GCGCGGTGTCATCTATGTTACTAGATCCCCGGGTCTAGACAATTCAGTACATTAAAAACGTCCGCCATGGTCTGAAGGCA
ACAGATAAGCCA
```

Figure 3. JRC GMO-Amplicons web interface overview. Screenshots of the JRC GMO-Amplicons web interface, showing the different entry points and integrated views to data, which can be easily explored by browsing through web links. See text for details.

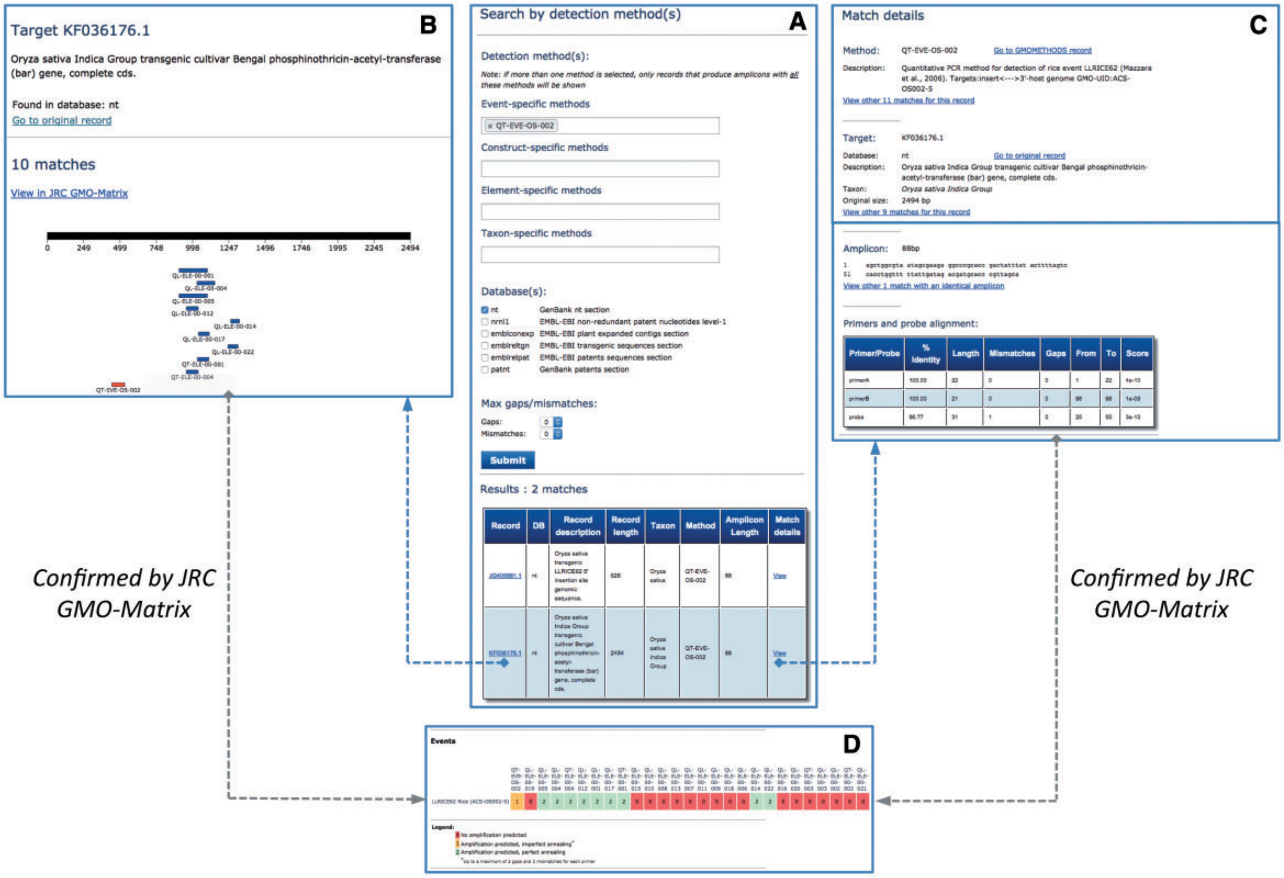
The non-redundant amplicons set and the amplicon sets derived from the different individual datasets can be searched by BLAST (Figure 3E). Matching records can be downloaded for further analyses (Figure 3F).

Examples of JRC GMO-Amplicons use cases

Information mining for poorly characterized GMOs

GMO-related sequences are expected to be present in public databases, but are hard to retrieve, in particular from patent-related datasets. The JRC GMO-Amplicons database facilitates the identification of the full GMO sequences and of the corresponding amplicons in the patent datasets and, therefore, supports control laboratories in the detection and identification of unauthorized GM events. The database can also aid annotating poorly described GM events or recognizing new GM-related sequences in public repositories. Some examples of applications for the JRC GMO-Amplicons database in food/feed control are provided below.

Characterization of GM rice event LLRICE62. GM rice event LLRICE62 (Unique identifier ACS-OS002-5) is an *Oryza sativa* transgenic insect-resistant line that is not authorized in the EU market. A quantitative PCR method for LLRICE62 detection has been developed and included in GMOMETHODS (QT-EVE-OS-002, 21). The 5' flanking region of the LLRICE62 event has been published in 2012 (GenBank: JQ406881.1, 22) while the sequence of the full event generated by NGS technology has been made publicly available in 2013 (GenBank: KF036176.1, 23). These sequences can be easily identified by querying the JRC GMO-Amplicons database with the QT-EVE-OS-002 method (Figure 4A). Moreover, by browsing the match detail panel of KF036176.1, it is possible to identify other element-specific methods (Figure 4B) targeting the event, as confirmed by the JRC GMO-Matrix (Figure 4D). The amplicon produced by the QT-EVE-OS-002 method (Figure 4C) is annotated in the database with one mismatch in the probe, as also pointed out by the JRC GMO-Matrix (Figure 4D). Finally, other sequences in the patent databases can be identified in the query as producing the same amplicon: all of them are part of patents claiming the development of glufosinate tolerant rice strains, of which LLRICE62 is a member (see for example AR2338850.1 of



Target KF036176.1

Oryza sativa Indica Group transgenic cultivar Bengal phosphinothricin-acetyl-transferase (bar) gene, complete cds.

Found in database: nt
Go to original record

10 matches

View in JRC GMO-Matrix

Search by detection method(s)

Detection method(s):

Note: If more than one method is selected, only records that produce amplicons with all these methods will be shown

Event-specific methods: × QT-EVE-OS-002

Construct-specific methods:

Element-specific methods:

Taxon-specific methods:

Database(s):

☑	nt	GenBank nt section
☐	nrnl1	EMBL-EBI non-redundant patent nucleotides level-1
☐	emblconexp	EMBL-EBI plant expanded contigs section
☐	emblreltgn	EMBL-EBI transgenic sequences section
☐	emblrelpat	EMBL-EBI patents sequences section
☐	patnt	GenBank patents section

Max gaps/mismatches:

Gaps: 0
Mismatches: 0

Results : 2 matches

Record	DB	Record description	Record length	Taxon	Method	Amplicon Length	Match details
JQ406881.1	nt	Oryza sativa transgenic LLRICE62 5' insertion site genomic sequence.	628	Oryza sativa	QT-EVE-OS-002	88	View
KF036176.1	nt	Oryza sativa Indica Group transgenic cultivar Bengal phosphinothricin-acetyl-transferase (bar) gene, complete cds.	2494	Oryza sativa Indica Group	QT-EVE-OS-002	88	View

Match details

Method: QT-EVE-OS-002 Go to GMOMETHODS record
Description: Quantitative PCR method for detection of rice event LLRICE62 (Mazzara et al., 2006). Targets:insert<--->3'-host genome GMO-UID:ACS-OS002-5
View other 11 matches for this record

Target: KF036176.1
Database: nt Go to original record
Description: Oryza sativa Indica Group transgenic cultivar Bengal phosphinothricin-acetyl-transferase (bar) gene, complete cds.
Taxon: *Oryza sativa Indica Group*
Original size: 2494 bp
View other 9 matches for this record

Amplicon: 88bp

```
1     agctggcgta atagcgaaga ggcccgcacc gactatttat acttttagtc
51    cacctggttt ttattgatag acgatgcacc cgttagca
```

View other 1 match with an identical amplicon

Primers and probe alignment:

Primer/Probe	% Identity	Length	Mismatches	Gaps	From	To	Score
primerA	100.00	22	0	0	1	22	4e-10
primerB	100.00	21	0	0	68	88	1e-09
probe	96.77	31	1	0	26	56	3e-13

Events

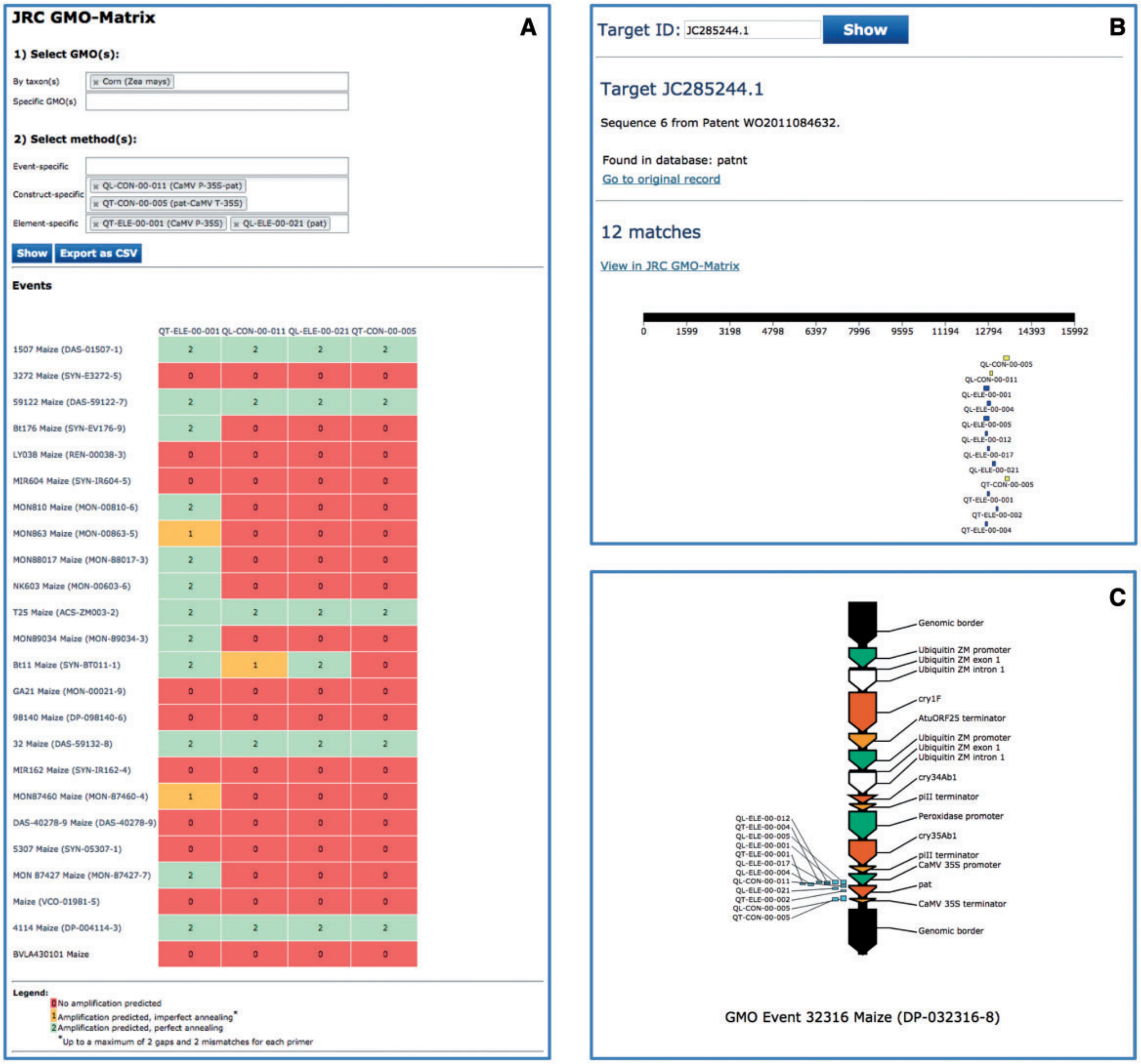
	QT-EVE-OS-002	QL-ELE-00-019	QL-ELE-00-005	QL-ELE-00-004	QT-ELE-00-004	QL-ELE-00-012	QL-ELE-00-001	QL-ELE-00-017	QT-ELE-00-001	QL-ELE-00-015	QL-ELE-00-010	QL-ELE-00-008	QL-ELE-00-013	QL-ELE-00-007	QL-ELE-00-011	QL-ELE-00-009	QL-ELE-00-018	QL-ELE-00-006	QL-ELE-00-014	QL-ELE-00-022	QL-ELE-00-016	QL-ELE-00-020	QT-ELE-00-003	QL-ELE-00-003	QL-ELE-00-002	QT-ELE-00-002	QL-ELE-00-021
LLRICE62 Rice (ACS-OS002-5)	1	0	2	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0

Legend:
0 No amplification predicted
1 Amplification predicted, imperfect annealing*
2 Amplification predicted, perfect annealing
*up to a maximum of 2 gaps and 2 mismatches for each primer

Figure 4. Retrieval of amplicons and target records related to GMO events. Integration of views showing the use of JRC GMO-Amplicons to retrieve GM rice event LLRICE62 targets and corresponding amplicons. Results are in agreement with those obtained by the JRC GMO-Matrix application. See text for details.

the *patnt* database, found in the granted patent US-6468747-B1). Together, these analyses show the use of the JRC GMO-Amplicons database for easily identifying new relevant sequences in public repositories.

Characterization of GM rice event TT51-1. GM rice event TT51-1 (also known as rice Bt63) is an *Oryza sativa* transgenic insect-resistant line that is not authorized in the EU market. Only three qualitative methods of the GMOMETHODS database (QL-CON-00-009, QL-CON-00-007 and QL-ELE-00-016) are described for detecting this event (24–26). In the JRC GMO-Amplicons interface four amplicons from sequence EU880444.1 (*transgenic* section of the ENA database) are identified when using these methods in a simultaneous query. The target record shows two putative amplicons for each of the selected



JRC GMO-Matrix

1) Select GMO(s):

By taxon(s): Corn (Zea mays)

Specific GMO(s):

2) Select method(s):

Event-specific:

Construct-specific: QL-CON-00-011 (CaMV P-35S-pat), QT-CON-00-005 (pat-CaMV T-35S)

Element-specific: QT-ELE-00-001 (CaMV P-35S), QL-ELE-00-021 (pat)

Events

	QT-ELE-00-001	QL-CON-00-011	QL-ELE-00-021	QT-CON-00-005
1507 Maize (DAS-01507-1)	2	2	2	2
3272 Maize (SYN-E3272-5)	0	0	0	0
59122 Maize (DAS-59122-7)	2	2	2	2
Bt176 Maize (SYN-EV176-9)	2	0	0	0
LY038 Maize (REN-00038-3)	0	0	0	0
MIR604 Maize (SYN-IR604-5)	0	0	0	0
MON810 Maize (MON-00810-6)	2	0	0	0
MON863 Maize (MON-00863-5)	1	0	0	0
MON88017 Maize (MON-88017-3)	2	0	0	0
NK603 Maize (MON-00603-6)	2	0	0	0
T25 Maize (ACS-ZM003-2)	2	2	2	2
MON89034 Maize (MON-89034-3)	2	0	0	0
Bt11 Maize (SYN-BT011-1)	2	1	2	0
GA21 Maize (MON-00021-9)	0	0	0	0
98140 Maize (DP-098140-6)	0	0	0	0
32 Maize (DAS-59132-8)	2	2	2	2
MIR162 Maize (SYN-IR162-4)	0	0	0	0
MON87460 Maize (MON-87460-4)	1	0	0	0
DAS-40278-9 Maize (DAS-40278-9)	0	0	0	0
5307 Maize (SYN-05307-1)	0	0	0	0
MON 87427 Maize (MON-87427-7)	2	0	0	0
Maize (VCO-01981-5)	0	0	0	0
4114 Maize (DP-004114-3)	2	2	2	2
BVLA430101 Maize	0	0	0	0

Legend:
0 No amplification predicted
1 Amplification predicted, imperfect annealing*
2 Amplification predicted, perfect annealing
*Up to a maximum of 2 gaps and 2 mismatches for each primer

Target ID: JC285244.1

Target JC285244.1

Sequence 6 from Patent WO2011084632.

Found in database: patnt
Go to original record

12 matches

View in JRC GMO-Matrix

GMO Event 32316 Maize (DP-032316-8)

Figure 5. Use of JRC GMO-Amplicons to assist interpretation of screening results. A case-study of detection and identification of unauthorized GM maize event in food/feed products. A maize sample was found positive to specific methods QL-CON-00-011 (junction between the Cauliflower Mosaic Virus 35S promoter and the *S. viridochromogenes pat* gene), QT-CON-00-005 (junction region between the *pat* gene from and the Cauliflower Mosaic Virus 35S terminator) and to two element detection methods, QT-ELE-00-001 (Cauliflower Mosaic Virus 35S promoter) and QL-ELE-00-021 (*pat* gene). The JRC GMO-Matrix (**A**) predicts 5 GM maize events as potential events positive to all of them, but the sample is negative for the corresponding event-specific methods. By querying the database with these methods and focusing on the unique records of the *patnt*, sequence, JC285244.1 from patent WO2011084632 (33) is found (**B**). According to the annotation, this sequence corresponds to the '*complete sequence of the insert and flanking regions of event DP-032316-8*', a good candidate for which the sample can be easily tested, as the patent describes also a detection method for this event. JC285244.1 was fully characterized (**C**), confirming that it is the full sequence of DP-032316-8 maize event.

methods and, in addition, several other ones from element-specific methods, all of them confirmed by the JRC GMO-Matrix decision support tool. These findings support the hypothesis of the presence of multiple copies of the Bt63 target determined in the development of the selected construct-specific methods (27). Sequence EU880444.1 was published as GM rice event TT51-1 in 2010 (28). Yang *et al.* (29) reported a new version that is slightly different. The pipeline did not find this newer sequence because it was made available only as supplementary material to the article and never submitted to public databases.

Analysis of unexpected detection methods results

Using the JRC GMO-Amplicons, it is possible to identify GM-related sequences that are positive for the elements detected in the analysed samples but negative for the event-specific methods included into the GMOMETHODS database. A case-study example is shown in Figure 5, where a maize sample was found positive for two construct specific methods, QL-CON-00-011 (30) and QT-CON-00-005 (29) and to two element detection methods, QT-ELE-00-001 (31) and QL-ELE-00-021 (32). Through the JRC GMO-Amplicons it was possible to identify a new patent sequence record (33) and to download it for further characterization.

Sequence similarity searches

The JRC GMO-Amplicons gives the possibility to search for similarity by NCBI-BLAST (17). As previously described (see Figure 3E and F), the non-redundant amplicons set and the amplicon sets derived from the different individual datasets can be searched by using a user-defined sequence as query. Through the interface, it is possible to adjust default algorithm parameters, like the e-value (i.e. the expectation value threshold for saving hits) or the penalties for mismatches and introduced gaps. It is also possible to choose to run the application in BLASTN or TBLASTX mode, i.e. by searching nucleotide databases using a nucleotide query or by searching translated nucleotide databases using a translated nucleotide query, respectively. The output is in the format of the classical NCBI BLAST output, with the additional option of downloading all matching amplicons for further analyses. Overall, this functionality is very useful in cases when control laboratories have obtained a sequence from an 'unknown sample' by direct sequencing of a PCR amplicon or by DNA walking anchored on GM elements, the latter already implemented in the context of identification of unauthorized GMOs (34).

Conclusion and future direction

To our knowledge, the JRC GMO-Amplicons is the first attempt at making sequences specifically related to GMOs publicly available. The JRC GMO-Amplicons web application is a gateway to mine the results of a pipeline that regularly screens public nucleotide sequence databanks, including patents and available whole plant genomes, through *in silico* determination of PCR amplification with the reference methods of the GMOMETHODS database.

A panel of five different public nucleic acid sequence data sets were selected for screening, even if potentially redundant, to have the most comprehensive set of possible target records and corresponding amplicons. According to our findings, this objective was achieved: 95% of the sequences classified as transgenic in the public databases were picked up by at least one method, and 91% of the selected methods had at least one target record in the selected databases. In addition, target records and corresponding amplicons were distributed in a non-random manner that was consistent with both the method type and the dataset of origin. This strongly suggests that the procedure of pooling a consistent, enriched, GMO-related set of sequences has been successful and specific.

The modular infrastructure behind the implementation of the JRC GMO-Amplicons allows to efficiently maintain up-to-date information when new detection methods and data resources become available. The platform could also allow the straightforward implementation of different tools for scanning patent resources or adding new annotations to the identified sequences. For example, it is scheduled to annotate amplicons extracted from patented sequence records with the original patent publication numbers. Moreover, a similarity clustering procedure will be implemented on the original target records in order to create a non-redundant dataset of GMO-related targets. Finally, work is under way to allow *in silico* determination of PCR amplification with custom primers.

The JRC GMO-Amplicons, together with the JRC GMO-Matrix, will help bridging the gap between the DNA-based detection methods and the important information contained in the GMO DNA sequences themselves, often difficult to obtain.

Conflict of interest. None declared.

Acknowledgements

The authors thank A. Lievens for internal review of the manuscript as well as L. Cengia and E. Ben for their help in setting up and maintaining the application server.

References

1. European Commission. (2001) *Off. J. Eur. Commun.* L, **106**, 1–38.
2. European Commission. (2003) *Off. J. Eur. Union* L, **268**, 1–23.
3. European Commission. (2003) *Off. J. Eur. Union* L, **268**, 24–28.
4. European Commission. (2004) *Off. J. Eur. Union* L, **165**, 61–64.
5. Bonfini,L., Van den Bulcke,M.H., Mazzara,M. *et al.* (2012) GMOMETHODS: the European Union database of reference methods for GMO analysis. *J. AOAC Int.*, **95**, 1713–1719.
6. International Organization for Standardization (ISO) 5725 Accuracy (Trueness and Precision) of Measurement Methods and Results. (1994) Geneva, Switzerland.
7. Horwitz,W. (1995) Protocol for the design, conduct and interpretation of method performance studies. *Pure Appl. Chem.*, **67**, 331–343.
8. Dong,W., Yang,L., Shen,K. *et al.* (2008) GMDD: a database of GMO detection methods. *BMC bioinform.*, **9**, 260.
9. Morisset,D., Novak,P.K., Zupanic,D. *et al.* (2014) GMOseek: a user friendly tool for optimized GMO testing. *BMC Bioinform.*, **15**, 258.
10. Gerdes,L., Busch,U. and Pecoraro,S. (2012) GMOfinder - A GMO Screening Database. *Food Anal. Methods*, **5**, 1368–1376.
11. Angers-Loustau,A., Petrillo,M., Bonfini,L. *et al.* (2014) JRC GMO-Matrix: a web application to support Genetically Modified Organisms detection strategies. *BMC Bioinform.*, **15**, 417.
12. Patak,A. (2011) CCSIS specialist EMBnet node: AGM2011 report. *EMB Net J.*, **17**, 13.
13. Leinonen,R., Akhtar,R., Birney,E. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.*, **39**, D28–31.
14. Coordinators,N.R. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–17.
15. Li,W., Kondratowicz,B., McWilliam,H. *et al.* (2013) The annotation-enriched non-redundant patent sequence databases. *Database*, **2013**, bat005.
16. Rotmistrovsky,K., Jang,W. and Schuler,G.D. (2004) A web server for performing electronic PCR. *Nucleic Acids Res.*, **32**, W108–W112.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Benson,D.A., Clark,K., Karsch-Mizrachi,I. *et al.* (2014) GenBank. *Nucleic Acids Res.*, **42**, D32–D37.
19. Sayers,E.W., Barrett,T., Benson,D.A. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **37**, D5–D15.
20. Fu,L., Niu,B., Zhu,Z. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
21. Mazzara,M., Grazioli,E., Savini,C. *et al.* (2006) Event-specific method for the quantitation of rice line LLRICE62 using real-time PCR. European Commission, Joint Research Centre.
22. Spalinskas,R., Van den Bulcke,M., Van den Eede,G. *et al.* (2013) LT-RADE: an efficient user-friendly genome walking method applied to the molecular characterization of the insertion site of genetically modified maize MON810 and rice LLRICE62. *Food Anal. Methods*, **6**, 705–713.
23. Wahler,D., Schauser,L., Bendiek,J. *et al.* (2013) Next-Generation Sequencing as a tool for detailed molecular characterisation of genomic insertions and flanking regions in genetically modified plants: a pilot study using a rice event unauthorised in the EU. *Food Anal. Methods*, **6**, 1718–1727.
24. Savini,C., Querci,M., Ermolli,M. *et al.* (2008) Report on the verification of the performance of a method for the detection of ‘Bt63’ rice using real-time PCR. European Commission, Joint Research Centre. http://gmo-crl.jrc.ec.europa.eu/gmomethods/docs/Bt63_Rice_verification_report_final.pdf (22 September 2015, data last accessed).
25. Grohmann,L. and Mäde,D. (2009) Detection of genetically modified rice: collaborative validation study of a construct-specific real-time PCR method for detection of transgenic Bt rice. *Eur. Food Res. Technol.*, **228**, 497–500.
26. Official Compendium of Analytical Methods according to § 64 LFGB (Food and Feed law). (2013) Food Analysis, Detection of genetically modified cry1Ab/Ac and P-ubi-cry DNA sequences in rice products with real-time PCR. BVL L 15.06-3. Beuth, Berlin Koln.
27. Mäde,D., Degner,C. and Grohmann,L. (2006) Detection of genetically modified rice: a construct-specific real-time PCR method based on DNA sequences from transgenic Bt rice. *Eur. Food Res. Technol.*, **224**, 271–278.
28. Wu,G., Wu,Y., Nie,S. *et al.* (2010) Real-time PCR method for detection of the transgenic rice event TT51-1. *Food Chem.*, **119**, 417–422.
29. Yang,L., Wang,C., Holst-Jensen,A. *et al.* (2013) Characterization of GM events by insert knowledge adapted resequencing approaches. *Scientific Reports*, **3**, 2839.
30. Official Compendium of Analytical Methods according to § 64 LFGB (Food and Feed law). (2012) Food Analysis, Real-time PCR detection of the P35S-pat genetic construct to screen for genetically modified plants, BVL G 30.40-1:2012-7. Beuth, Berlin Koln.
31. Feinberg,M., Fernandez,S., Cassard,S. *et al.* (2005) Quantitation of 35S promoter in maize DNA extracts from genetically modified organisms using real-time polymerase chain reaction, part 2: interlaboratory study. *J. AOAC Int.*, **88**, 558–573.
32. Barbau-Piednoir,E., Stragier,P., Roosens,N. *et al.* (2014) Interlaboratory Testing of GMO Detection by Combinatory SYBR®Green PCR Screening (CoSYPS). *Food Anal. Methods*, **7**, 1719–1728.
33. Diehn,S., Lu,A.L., Nowatzki,T.M. *et al.* (2011) Patent WO-2011-084632-A1.
34. Fraiture,M.A., Herman,P., Taverniers,M. *et al.* (2014) An innovative and integrated approach based on DNA walking to identify unauthorised GMOs. *Food Chem.*, **147**, 60–69.