Food Chemistry 192 (2016) 788–798

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Statistical framework for detection of genetically modified organisms based on Next Generation Sequencing

Sander Willems[a,b,c,1], Marie-Alice Fraiture[a,b,c,d,1], Dieter Deforce[c,1], Sigrid C.J. De Keersmaecker[a], Marc De Loose[d], Tom Ruttink[e], Philippe Herman[b], Filip Van Nieuwerburgh[c,2], Nancy Roosens[a,*,2]

[a] Scientific Institute of Public Health (WIV-ISP), Platform of Biotechnology and Molecular Biology (PBB), J. Wytsmanstraat 14, 1050 Brussels, Belgium
[b] Scientific Institute of Public Health (WIV-ISP), Biosafety and Biotechnology Unit (SBB), J. Wytsmanstraat 14, 1050 Brussels, Belgium
[c] University of Gent (UGent), Faculty of Pharmaceutical Sciences, Laboratory of Pharmaceutical Biotechnology, Harelbekestraat 72, 9000 Ghent, Belgium
[d] Institute for Agricultural and Fisheries Research (ILVO), Technology and Food Sciences Unit, Burg. Van Gansberghelaan 115, bus 1, 9820 Merelbeke, Belgium
[e] Institute for Agricultural and Fisheries Research (ILVO), Plant Sciences Unit, Caritasstraat 21, 9090 Melle, Belgium

ARTICLE INFO

Article history:
Received 10 December 2014
Received in revised form 26 June 2015
Accepted 18 July 2015
Available online 21 July 2015

Keywords:
NGS
GMO detection
Bioinformatics
GM rice
Statistical framework
Processed food

ABSTRACT

Because the number and diversity of genetically modified (GM) crops has significantly increased, their analysis based on real-time PCR (qPCR) methods is becoming increasingly complex and laborious. While several pioneers already investigated Next Generation Sequencing (NGS) as an alternative to qPCR, its practical use has not been assessed for routine analysis. In this study a statistical framework was developed to predict the number of NGS reads needed to detect transgene sequences, to prove their integration into the host genome and to identify the specific transgene event in a sample with known composition. This framework was validated by applying it to experimental data from food matrices composed of pure GM rice, processed GM rice (noodles) or a 10% GM/non-GM rice mixture, revealing some influential factors. Finally, feasibility of NGS for routine analysis of GM crops was investigated by applying the framework to samples commonly encountered in routine analysis of GM crops.

1. Introduction

In recent years, the number and diversity of genetically modified (GM) crops on the market have drastically increased (James, 2013). Legislations related to GMO (genetically modified organism) commercialisation differ from country to country, but it is internationally agreed that GMOs can only be commercialised after thorough safety assessments. To this end, GMO developers have to perform molecular characterisation of each novel GMO subjected to authorisation. This molecular characterisation includes the determination of the inserted DNA sequence via the evaluation of the number of inserts using Southern blot analysis and Polymerase Chain Reaction (PCR). Furthermore, Sanger sequencing of the junction of the transgene insert and the host genome is used to determine its precise location as well as the detection of possible presence of the backbone sequence of the transformation vector. This approach is relatively time-consuming and requires customised experiments, carefully designed for each event (Kovalic, Garnaat, & Guo, 2012).

The DNA sequence data of the insert junctions is also used for the development and validation of the event-specific detection method, required for subsequent GMO monitoring in food and feed products by EU enforcement laboratories (Commission Regulations EC/1829/2003 (2003) and EC/1830/2003 (2003)). These laboratories use quantitative real-time PCR (qPCR) to screen for the presence of commonly used DNA elements in GMOs and then, using event-specific methods provided by the GMO developers, to identify a GMO (Broeders, De Keersmaecker, & Roosens, 2012). To increase the efficiency of GMO detection, qPCR methods are being used that run on a 96-well plate with multiplex qPCR for simultaneous detection. Moreover, Decision Support Systems have been developed to deal with the complexity of multiple PCR signals (Bahrdt, Krech, Wurz, & Wulff, 2010; Brodmann, Ilg, Berthoud, & Hermann, 2002; Dörries, Remus, Grönewald, Grönewald, & Berghof-Jäger, 2010; Foti, Onori, Donnarumma, De Santis, & Miraglia, 2006; Huber et al., 2013; Köppel, Sendic, & Waiblinger, 2014; Morisset et al., 2014; Van den Bulcke et al., 2010; Waiblinger, Ernst, Anderson, & Pietsch, 2008). If the presence of unauthorised GMOs (UGMs) is suspected, additional analyses, like DNA walking, are performed to identify the junction between the

[*] Corresponding author.
E-mail address: nancy.roosens@wiv-isp.be (N. Roosens).
[1] Equal contribution: first name.
[2] Equal contribution: last name.

http://dx.doi.org/10.1016/j.foodchem.2015.07.074
0308-8146/© 2015 The Authors. Published by Elsevier Ltd.

host genome and the transgene sequence to identify or better characterise the UGM (Fraiture et al., 2014; Ruttink et al., 2010). Although this methodology has been optimised for use by enforcement laboratories, the DNA walking method can be laborious in the case of a complex mixture.

While GMO analysis has benefitted from multiplexing PCR methods, limitations like a maximum of six targets per qPCR experiment (Bahrdt et al., 2010) and unbiased primer design with equal analytical performance for a multiplex assay compared to simplex assays remain. Furthermore, the qPCR strategy *per se* implies the prior knowledge of at least part of the sequence of the transgene integrated in the host genome as well as the subsequent development of an efficient assay targeting this sequence. Collecting these sequences and designing the corresponding method for each new sequence target case by case remains challenging today, especially for unknown GMOs. This poses a major problem as GMOs remain undetectable when no method targeting the transgene element has been used. Recently, Next Generation Sequencing (NGS) has been proposed to tackle these challenges.

NGS, allowing massive parallel DNA fragment sequencing, was of great importance to sequence several complete plant genomes and is being used in the sequencing of many more plant genomes (Michael & Jackson, 2013). As a consequence, the use of NGS has been proposed to provide an informative and cost-effective alternative to the current Southern blot-based method for molecular characterisation of plant GMOs. One of these alternatives assumes the availability of a reference genome of the GM crop and the sequence of the inserted transgene cassette. Based on this information, Kovalic et al. (2012) used NGS to characterise the junctions on both sides of a specific transgene cassette. Other approaches have been developed to exploit the potential of NGS for GMO detection and analysis when a reference genome of the GM crop is available, but only partial or no prior knowledge of the sequence of the transgene insert is available (Wahler, Schauser, Bendiek, & Grohmann, 2013; Yang et al., 2013). Liang et al. (2014) have dealt with GMOs by developing a targeted strategy combining a chromosome walking method, based on SiteFinding-PCR, and NGS technology. In this study, a part of the cassette is known and targeted (partial *a priori* knowledge). The NGS technology is not used for full characterisation of the GM crop but rather as a high-throughput sequencing technology that is more time-efficient than Sanger sequencing to individually sequence DNA fragments.

These pioneer studies in the context of NGS-based GMO detection showed the applicability of NGS to circumvent the limitations posed by the qPCR strategy and Sanger sequencing. The major benefit of NGS is its independence of *a priori* knowledge of the transgene sequence. Because NGS is a relatively new technique applied to GMO detection, the infrastructure and expertise amongst scientists of enforcements laboratories, mainly molecular biologists, is often not present. A key component for short term implementation of NGS is therefore the development of bioinformatics capacity by enforcement laboratories. This includes the availability of computing infrastructure, the development or implementation of adequate software and the development of expertise in order to manage, analyse and gain new information from NGS data. A second challenge is related to the nature of the DNA that needs to be analysed by NGS during GMO analysis in routine; including the large size of plant genomes, lack of good reference genomes for specific varieties or organisms due to large intraspecific genome variability in plants, DNA samples with traces of GMO material and degraded DNA due to food processing. While some of these issues have already been tackled, i.e. large intraspecific variability can be circumvented by an initial alignment against the transgenic cassette (Yang et al., 2013), the applicability of NGS for routine analysis has not been previously investigated.

To accommodate NGS within routine GMO detection, a first priority is capturing transgene information with NGS. The focus on a specific sequence (transgene insert) within a given genome, as opposed to reconstructing the entire genome sequence, means that statistical methods for the estimation of sequencing depth versus coverage of whole genomes, like the Lander–Waterman theory (Sims, Sudbery, Ilott, Heger, & Ponting, 2014), are not applicable. Therefore, a novel conceptual statistical framework is developed in this article to draw a better picture of the present feasibility of NGS technology for routine GMO analysis. This statistical framework was validated by NGS data from a GM rice (Bt rice), with known transgene insert and flanking regions, and is based on three approaches: (1) detecting potential transgene inserts, (2) proving their integration in the host genome, (3) identifying the specific junctions. All these approaches start with an alignment against an *a priori* known insert and only the aligned reads are subsequently investigated to avoid large intraspecific variability in plants. To assess the potential applicability of NGS on different types of food matrices, 100% Bt rice grains, 10% Bt rice grains mixed with 90% non-GM rice grains and 100% Bt rice noodles were analysed. To evaluate the robustness of these three approaches, they were implemented on two different data analysis platforms: an easy-to-use commercial software platform, the “CLC Genomics Workbench”, allowing potential use of NGS by “bioinformatics novices”, and a “Command-Line” platform allowing greater control of the workflow and parameters, but demanding a higher level of expertise in bioinformatics. This newly developed statistical framework allows to determine the probability that a given GMO can be detected when its presence in a sample is known. Based on this probability, an estimate of the number of reads necessary to be able to detect a transgene cassette, to prove integration in the host genome and to identify several common GMO events and mixtures can be calculated.

2. Materials and methods

2.1. Statistical framework

Three approaches, addressing different levels of complexity in the analysis of GMOs, are used to analyse shotgun sequencing libraries, sequenced as paired-end reads from a sample that consists of a single GMO. The “detection approach” was used to detect the presence of a transgene cassette, referred to as the insert. The “proof approach” allows to provide the evidence that the insert is effectively integrated in the non-GM genome, referred to as the host genome, and gives a crude localisation of the insert in the host genome. The “identification approach” delivers the precise identification and localisation of the junctions between the host genome and the insert (Fig. 1).

2.1.1. Calculation of probabilities to successfully detect a sequence aligned to a transgene

For each approach, the probability to successfully detect a theoretical read in an NGS sample of a known GMO, $P(+|\mathrm{GMO})$, was calculated. False positives were not considered and as a result the probability of an unknown sample containing a GMO when testing positive $P(\mathrm{GMO}|+)$ was not determined.

For a GMO, the length of the GM genome is the sum of the length of the non-GM genome (H) and the length of the insert (I). A partial insertion is defined as an insert with a large part of the insert deleted. In this case the length of the partial insertion is considered as the length of the insert (I). After sequencing of the GMO, this gives a total of different mates (T_s), with an average read length for each mate (R), equal to $H + I - R + 1$ or a total of

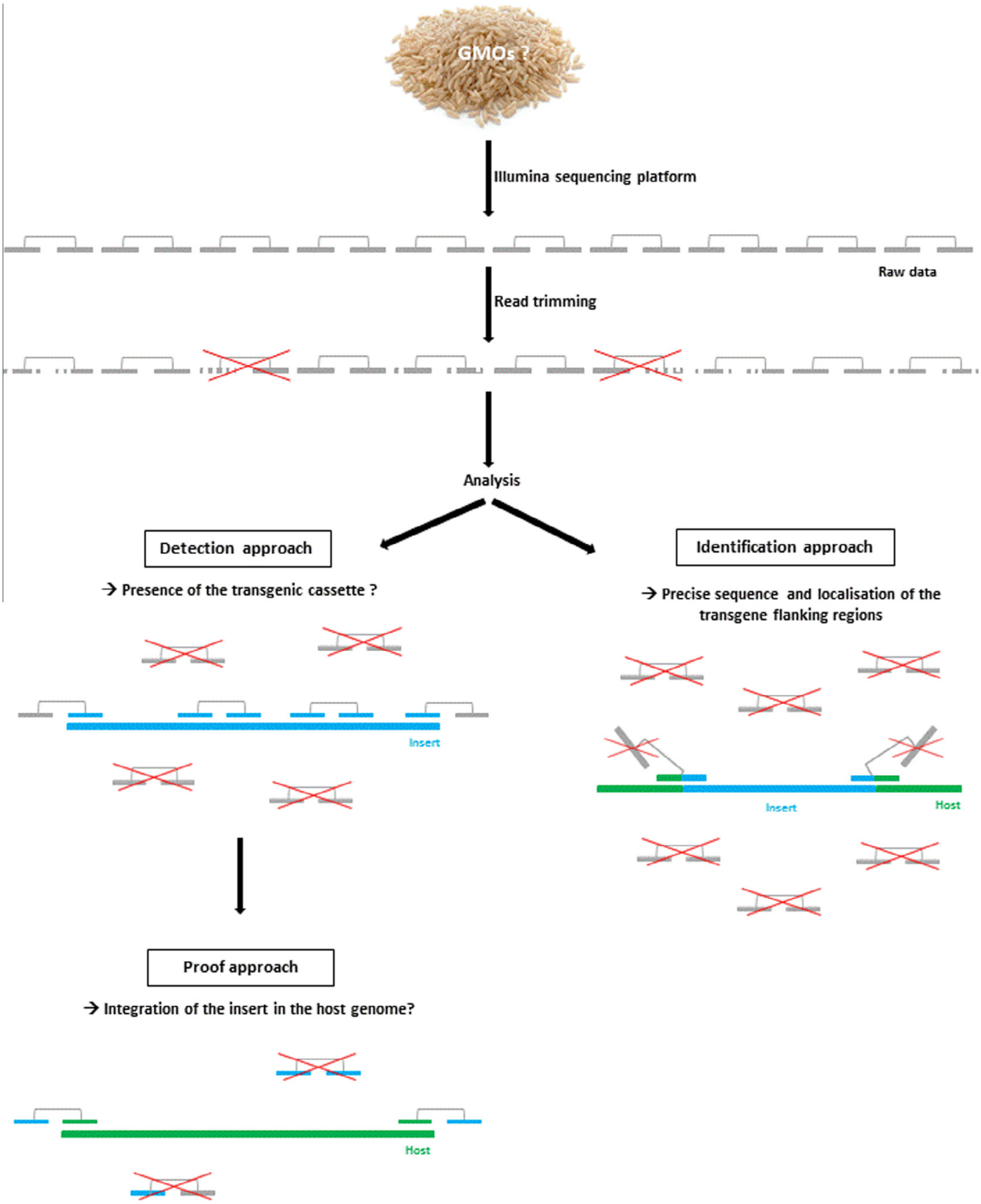
GMOs ?
Illumina sequencing platform
Raw data
Read trimming
Analysis
Detection approach
→ Presence of the transgenic cassette ?
Insert
Identification approach
→ Precise sequence and localisation of the transgene flanking regions
Insert
Host
Proof approach
→ Integration of the insert in the host genome?
Host

Fig. 1. GMO analysis workflow based on NGS. From a given matrix, extracted DNA is used for shotgun library construction and sequenced on an Illumina platform to obtain millions of raw paired-end reads. These are first trimmed based on sequencing quality scores and then filtered so only paired-end reads remain with each mate having a length of 30 bp or larger. To determine the presence of GMOs, the filtered reads are then analysed using three different approaches. On the one hand, the detection approach selects all paired-end reads with one mate globally aligned to the reference sequence of the insert, revealing its presence in the tested sample. The corresponding mates of the detected reads are subsequently analysed in the proof approach to confirm the integration of the transgenic insert in the host genome by globally aligning these mates to the reference sequence of the host genome. This approach also allows a rough localisation of the transgene flanking regions. On the other hand, all filtered reads are analysed with the identification approach to determine the exact localisation and sequence of the flanking regions by locally aligning them to the host genome and transgenic insert simultaneously.

different theoretical paired-end reads (T_p), with an average paired-end distance (D), mates included, equal to $H + I - D + 1$.

To be able to detect the presence of a known insert, only sequences that fall completely in the inserted region can be detected using a global alignment. As a consequence, partial insertions that are smaller than the read length ($I < R$) are impossible to detect with this method. If $I \geqslant R$, there are $I - R + 1$ different theoretical mates that possibly align.

To be able to prove that the insert is integrated in the host genome as well as to give a rough location, a theoretical paired-end read needs one mate globally aligned to the host genome and the other mate globally aligned to the insert. Similarly to the detection of an insert, it is only possible to find such kind of sequences when $I \geqslant R$. If the paired-end distance of a theoretical paired-end read (D) is large enough ($D \geqslant I + R$), a mate globally aligned to the insert will always have a mate globally aligned to the host genome. Otherwise, a theoretical paired-end read, with each mate globally aligned to either the insert or the host genome, will span a junction only if the junction is not located on either of the mates. The length of this sequence is equal to $D - 2 \cdot R$, so there are $D - 2 \cdot R + 1$ theoretical paired-end reads for each different junction.

To be able to identify these junctions, a sequence needs to locally align its 5′ and 3′ tail to respectively the host genome and the insert with a minimum overlap of nucleotides (M) for each tail or vice versa. This is impossible in cases with a very small partial insertion ($M > I$). If the read length R is large compared to the insert ($I + M < R$), a theoretical mate that locally aligns to the insert with overlap O is then guaranteed to have at least M basepairs overlap with the host genome reference. In this case, there are exactly $I - M + 1$ different theoretical mates that locally align to the insert with O basepairs overlap. Finally, if $I \geqslant R - M$, sequence covers a junction only if the junction is covered by the $R - 2 \cdot M$ bp in the middle of the sequence. For each junction there are thus $R - 2 \cdot M + 1$ different theoretical mates.

The ratio of targeted theoretical reads over the total of theoretical reads is now given by the following formulae:

(1) p_1: The ratio of theoretical mates globally aligned to the insert, as needed for the detection approach.

(a) $p_1 = \frac{I-R+1}{T_s}$ if $I \geqslant R$
(b) $p_1 = 0$ if $I < R$ (small partial inserts)

(2) p_2: The ratio of theoretical paired-end reads covering a single junction with one mate globally aligned to the host genome and the other to the insert, as needed for the proof approach.

(a) $p_2 = \frac{D-2 \cdot R+1}{T_p}$ if $I > D - R$
(b) $p_2 = \frac{I-R+1}{T_p}$ if $I \leqslant D - R$ and $I \geqslant R$
(c) $p_2 = 0$ if $I < R$ (small partial inserts)

(3) p_3: The ratio of theoretical mates covering a single junction with one mate locally aligned to the host genome and the other locally aligned to the insert, as needed for the identification approach.

(a) $p_3 = \frac{R-2 \cdot M+1}{T_s}$ if $I > R - M$
(b) $p_3 = \frac{I-M+1}{T_s}$ if $I \leqslant R - M$ and $I \geqslant M$(small partial inserts)
(c) $p_3 = 0$ if $I < M$(very small partial inserts)

Given these ratios of theoretical paired-end reads or mates, it is straightforward to calculate the probability $P(-|\text{GMO})$ that no targeted reads are found after sequencing N paired-end reads, all originating from a GMO:

(1) P_1 (no insert detected, while reads originate from a GMO) = $(1 - p_1)^{2N}$, if N paired-end reads are considered as $2N$ independent mates.
(2) P_2 (no proof of integration of insert detected, while reads originate from a GMO) = $(1 - 2p_2)^N$, assuming either one of the two junctions suffices as a proof of this insertion.
(3) P_3 (no identification of junctions possible, while reads originate from a GMO) = $(1 - 2p_3)^{2N}$, if N paired-end reads are considered as $2N$ independent mates and either one of the two junctions suffices for identification.

Conversely, the probability to detect at least one read in a pure sample extracted from a GMO is equal to $P(+|\text{GMO}) = 1 - P(-|\text{GMO})$.

2.1.2. Estimation of the number of paired-end reads needed to have a probability P of finding at least one targeted read

For a GMO, the number N of paired-end reads that are needed to have a probability P of finding at least one targeted read can easily be retrieved by rewriting the formulae in the previous paragraph and is given by the next formulae where p_i is defined as before:

(1) $N_1 = \frac{\ln(1-P)}{2\ln(1-p_1)}$ paired-end reads are needed for the detection approach.
(2) $N_2 = \frac{\ln(1-P)}{\ln(1-2p_2)}$ paired-end reads are needed for the proof approach.
(3) $N_3 = \frac{\ln(1-P)}{2\ln(1-2p_3)}$ paired-end reads are needed for the identification approach.

while the above formula are completely general and can be used for any pure sample of a GMO, they can be greatly simplified for most common cases. In general the host genome length H is large compared to the insert length I, which in turn is large compared to the paired-end distance D and read length R. As a result the probabilities p_1, p_2 and p_3 will be small, so $\ln(1 - p_i) \approx -p_i$. Furthermore, the total number of reads T_s and T_p can be simplified to H, and the constant 1 in the numerator can be omitted for all probabilities p_i. In summary, this yields the following simple approximations for the number of paired-end reads N needed to have a probability P of finding at least one targeted read:

(1) $N_1 \approx \frac{H}{2 \cdot (R-I)} \cdot \ln(1 - P)$
(2) $N_2 \approx \frac{H}{2 \cdot (2R-D)} \cdot \ln(1 - P)$
(3) $N_3 \approx \frac{H}{4 \cdot (2 \cdot M-R)} \cdot \ln(1 - P)$

The following parameters are thus of importance to make an estimate:

- *A priori* known or estimated
  - o I: length of the insert reference.
  - o H: length of the host genome reference.
- Definable by user
  - o R: sequenced read length (average).
  - o D: sequenced paired-end distance (average), including mates, thus larger than twice the read length.
  - o M: minimum overlap length between each tail of a mate and the host genome /insert reference, thus smaller than halve a read length. Software often has default parameters for M, dependent on read length R.
  - o P: probability to find at least one targeted read.
- Calculated result
  - o N: number of quality filtered paired-end reads needed to have a probability P of finding at least one targeted read.

2.1.3. Modifications for more complex cases

It is possible to adjust the presented formulae to different scenarios that better reflect food and feed matrices complexity. These matrices usually contain only traces of a GMO or might contain a mixture of different ingredients. In such cases the ratio of targeted reads should be multiplied by the DNA ratio r of the GMO over the rest of the sample. For instance a mixture of 10% GM rice (genome size of 400 Mbp) and 90% non-GM maize (genome size of 2300 Mbp) has a DNA ratio of $\frac{10 \cdot 400}{90 \cdot 2300 + 10 \cdot 400} \approx 0.019$ for the GM rice. Since in the simplified version the number of needed paired-end reads is linearly dependent on the ratio of targeted reads, it follows that the linear dependence of the DNA ratio is not only valid for ratios of targeted reads, but also for the number of paired-end reads needed to be able to detect at least a single read for each approach. By calculating the probability of detecting exactly $x - 1$, $x - 2, \ldots,$ 0 reads, it is also possible to calculate the probability of detecting at least x reads instead of detecting at least one read.

2.2. Generation of NGS data from different food matrices

Three DNA samples were generated from transgenic Bt rice (see Supplementary text S1 for a description of DNA extraction and the inserted cassette): (1) Bt rice grains (named 100% Bt rice sample), (2) Bt rice grains processed into noodles as described in Fraiture et al. (2015) (named 100% Bt noodles sample), (3) mixture of 10% Bt rice DNA with 90% of the corresponding non-GM rice DNA (named 10% Bt rice sample).

2.2.1. Library preparation and sequencing

Two Illumina shotgun sequencing libraries were generated, one from 5 μg of the 100% Bt rice sample and the other from 5 μg of the 10% Bt rice sample. DNA was fragmented to 300–400 bp using Covaris S2 sonication and an indexed sequencing library was prepared using an Illumina TruSeq DNA Sample Preparation kit. The two resulting libraries were sequenced simultaneously on a single Rapid Run flow cell with 2 lanes, one per library, with an Illumina HiSeq 2000 sequencer, generating 2 × 100 bp paired-end reads for each sequenced fragment. After base calling using the Illumina CASAVA version 1.8 software, raw sequences were obtained.

The 100% Bt noodles sample was sequenced several months later, using updated protocols and techniques. In this case, an Illumina shotgun sequence library was generated from 1 μg of the 100% Bt noodles sample. DNA was fragmented to ±400 bp using Covaris S2 sonication and a sequencing library was made using the NEBNext Ultra DNA Library Prep Kit with 8 enrichment PCR cycles. Size selection was performed on the resulting library using an Invitrogen 2% E-gel, selecting fragments between 400 and 600 bp. The library was sequenced on half of a Rapid Run flow cell lane on an Illumina HiSeq 1500 sequencer, generating 2 × 100 bp paired-end reads for each sequenced fragment. Base calling and primary quality assessments were performed using Illumina's Basespace genomics cloud computing environment.

2.3. Implementation of the framework

Two different platforms were used to analyse the NGS data: (1) freely available programs such as BWA (Li & Durbin, 2009) and Bowtie2 (Langmead & Salzberg, 2012) combined with Python and Perl scripts were used on a computer running Linux Ubuntu 14, referred to in this manuscript as the Command-Line-Tools, (2) the commercial software package CLC Genomics Workbench 7 (http://www.clcbio.com) running on Windows 7 Enterprise, referred to in this manuscript as the CLC Genomics Workbench.

For the host genome reference the sequence of *Oryza sativa* was used, more specifically the MSU6 build of *O. sativa* of length 374,332,026 bp available via Illumina's IGenomes https://support.illumina.com/sequencing/sequencing_software/igenome.ilmn, which includes pseudomolecules representing the mitochondria, plastids and Syngenta sequences. The reference of the inserted pCAMBIA cassette was obtained from Breitler et al. (2004) as a personal communication and consists of 7002 bp.

2.3.1. Command-Line-Tools

Using a custom Perl script, the sequenced paired-end reads were trimmed when the average quality in a sliding window of 10 bp fell below Q20 and were filtered for sequences shorter than 30 bp after trimming (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013). Only paired-end reads were retained.

For the detection approach, the mates (each paired-end read consists of two mates) of all quality filtered paired-end reads were considered as single-ended and were aligned end-to-end (global alignment) to the insert using BWA with default parameters (BWA manual version 0.7.7-r441). Results were converted to SAM format (Li et al., 2009) and only aligned mates were selected.

For the proof approach, corresponding mates of those previously aligned in the detection approach were retrieved with a custom python script. These mates were then aligned to the host reference genome using default BWA parameters, similarly to module 1 described by Yang et al. (2013). Results were then converted to SAM format and unaligned mates were discarded.

For the identification approach, the mates of all quality filtered paired-end reads were considered as single-end and were partially aligned (local alignment) to the insert using Bowtie2. A length of 20 bp for the part initially aligned before elongation starts (seed), located at the beginning or end of a sequence with a maximum of one mismatch, was used instead of default Bowtie2 parameters, as found in the Bowtie2 manual version 2.2.1. Only mates that aligned were selected from the resulting SAM file. Mates with a CIGAR string (Li et al., 2009) matching a global alignment were discarded and the remaining mates were aligned against the host genome reference with the same parameters. Only aligned mates were selected from the resulting SAM file, again discarding mates with a global alignment. The resulting mates were divided in groups corresponding to different junctions, similar to the study published by Kovalic et al. (2012).

2.3.2. CLC Genomics Workbench

Similar to the Command-Line-Tools, a separate stand-alone analysis was done with the CLC Genomics Workbench.

All sequenced paired-end reads were trimmed with the NGS Core Tool "Trim Sequences" with an ambiguous trim length of 2, quality limit of 0.05 and minimum length of 30. Only paired-end reads were retained.

The quality filtered paired-end reads were globally aligned to both the insert and the host genome simultaneously using the NGS Core Tool "Map Reads to Reference" with similarity fraction 0.8, length fraction 1.0 and default parameters, as found in the CLC Genomics Workbench 7 Manual. Only paired-end reads with at least one mate aligned to the insert were selected for downstream analysis.

To verify that this insert was effectively integrated in the host genome, the option "Find Broken Pair Mates" was used on the selected paired-end reads to retrieve mates that did not align to the insert but to the host genome instead.

All quality filtered paired-end reads that were not globally aligned to the insert nor the host genome, were selected to identify the junction sequences. These paired-end reads were locally aligned against the insert with the NGS Core Tool "Map Reads to Reference" with similarity fraction 0.8 and length fraction 0.3. Aligned paired-end reads were selected and realigned against the host genome with the same parameters.

3. Results and Discussion

3.1. Statistical framework

In Section 2.1, a statistical framework was developed to investigate the use of NGS in routine analysis of GMOs. The formulae of this framework predict the number of NGS reads needed to have a probability P to detect transgene sequences, to prove their integration into the host genome or to identify the specific transgene event in a sample with known composition based on a number of parameters. To verify if the developed statistical formulae are good predictors, they were implemented using Command-Line-Tools and compared with the experimental results from the 100% Bt rice, 10% Bt rice and 100% Bt noodles samples (Section 3.2). We identify and discuss several influential factors that have an impact on the formulae of the statistical framework.

3.1.1. Validation based on experimental results

The statistical framework takes several parameters as input. To estimate some of these values, a global alignment against the host genome reference was carried out with all N quality filtered paired-end reads, using BWA with default parameters.

The *a priori* parameters used for all samples were $I = 7002$ bp and $H = 374,332,026$ bp. Experimentally, two different insertions were previously identified in Bt rice, one of length $I_2 = 6868$ on chromosome II and one of length $I_3 = 6936$ on chromosome III. Furthermore, literature suggests that the length H of the host genome of *O. sativa japonica* is actually 385 Mbp (Kawahara et al., 2013) instead of 374 Mbp, the length of the used reference. The ratio r of the genome reference length over the actual genome length of 0.97 was used to correct the ratio of targeted reads.

For the 100% Bt rice sample, the user definable parameters were $R = 100$, $N = 28$ and $D = 350$ (average). However, the experimental values R and D were respectively approximated at 86.4 (average) and 208.35 (average). Although the average read length R is usually a good approximation, it should be noted that the average paired-end distance has a large spread and is skewed (data not shown). To calculate the probability P of detecting no single reads, the number of quality filtered reads $91,371,164$ is used as N, instead of the number of raw reads. Similar parameters were applicable for the 10% Bt rice and 100% Bt noodles samples (Supplemental Table S1).

In the 100% Bt rice sample, 26.6% of the reads originate from the mitochondria and plastids according to the global alignment against the host genome, while the statistical framework was developed for pure genomic DNA. The ratio p_i for each approach was thus corrected by a ratio r of 0.734 to correct for the abundance of mitochondrial DNA. For the 10% Bt rice and 100% Bt noodles samples, the percentage of reads that aligned to the mitochondria/plastids was respectively 18.2% and 11.3%.

The ratios of targeted reads p_i were used to estimate the number of paired-end reads that are to be expected ($E = p_i \cdot N$) after sequencing N quality filtered paired-end reads. Since two identical insertions were present, the ratio of targeted reads for the detection approach was equal to the sum of the ratio for each individual approach. The probabilities for the proof and identification approach were calculated separately for each insertion, since the insertions are independent. The number of expected reads (Table 1) was compared to the number of experimental reads (Table 2) and were found to be in agreement. The largest deviations concern the 10% Bt rice sample where absolute values are low (< 5) and these results were disregarded due to the low statistical significance of few reads.

3.1.2. Identification of influential factors

It should be noted that several assumptions and simplifications have been made to develop the formulae. First, all theoretical reads are assumed to be perfect and to not contain any errors. Although this assumption is not true in reality, it affects both targeted and untargeted reads. It can thus be assumed that the ratio of targeted reads over total reads (p) is mostly unaffected by this property, even though the number of quality filtered reads will be reduced when errors are present. Similarities between the insert and the host genome add an extra level of complexity to the analysis. Currently, the host genome and insert of a GMO are often of a different species or even of a different kingdom, i.e. Plantae and Bacteria, and are genetically different. However, in a near future, many new GMOs are expected to be developed with *cis*-genic inserts and there might be cases where this property has a major influence on the analysis (Espinoza et al., 2013; Holme, Wendt, & Holm, 2013). Another important assumption is that all reads are equally likely to be sequenced. However, it has been shown that regions with a high GC content are underrepresented in i.e. Illumina sequencing and it was found that some regions cannot be sequenced at all with Illumina (Rieber et al., 2013). With enough prior knowledge about these issues though, the calculated probabilities can be adjusted accordingly.

Aside from these assumptions and simplifications, there are some limitations in defining all parameters, although they have a major influence on the analysis. First, a proper reference sequence of both the insert and the host genome is required. In reality this is not always the case, as shown in this study where the reference genome was only 374 Mbp as opposed to the literature suggesting it should be 385 Mbp (Kawahara et al., 2013), implying a different sequence. Since only reads with at least one mate aligned against the insert are used for downstream analysis, differences in the reference genome sequence will only be a limitation if they are near an insert site. Therefore, older draft genomes or other cultivars, as used in this study, are expected to have little influence. However, the plant genome size may vary greatly within a species or between cultivars (Greilhuber, 2005; Ohri, 1998), having a large effect on the statistical formulae. Furthermore, the formulae were developed for a pure sample. This study showed that even for a

Table 1
Theoretical formulae of the statistical framework applied to the 100% Bt rice, 10% Bt rice and 100% Bt noodles samples for all three approaches. For the detection approach, reads from both inserts were analysed simultaneously, as they cannot be identified separately. For the proof and identification approach the reads of the inserts were investigated independently. All used parameters are shown in Supplemental Table S1. Experimentally detected true positive reads from the Command-Line-Tools are shown in brackets.

		Detection approach	Proof approach		Identification approach	
			Chromosome II	Chromosome III	Chromosome II	Chromosome III
100% Bt rice	Ratio of theoretical targeted reads over theoretical possible reads (millionfold)	25.891	0.070	0.070	0.060	0.060
	Probability P to detect at least one read	1.00	1.00	1.00	1.00	1.00
	Expected reads (truly detected reads)	4,731 (3,186)	13 (23)	13 (17)	22 (12)	22 (22)
10% Bt rice	Ratio of theoretical targeted reads over theoretical possible reads (millionfold)	2.921	0.008	0.008	0.007	0.007
	Probability P to detect at least one read	1.00	0.68	0.68	0.85	0.85
	Expected reads (truly detected reads)	406 (284)	1 (2)	1 (0)	2 (0)	2 (2)
Bt Noodles	Ratio of theoretical targeted reads over theoretical possible reads (millionfold)	31.573	0.082	0.082	0.100	0.100
	Probability P to detect at least one read	1.00	1.00	1.00	1.00	1.00
	Expected reads (truly detected reads)	4,265 (5,485)	11 (30)	11 (11)	27 (22)	27 (25)

Table 2
Overview of the number of detected reads per approach for both the Command-Line-Tools and the CLC Genomics Workbench applied to the three samples; 100% Bt rice, 10% Bt rice and 100% Bt noodles. The detection approach was designed to detect an insert by finding reads that align to the used insert reference. The proof approach was designed to prove integration of the insert in the host genome by finding mates of detected reads in the detection approach that align to the host genome reference. The identification approach identified junctions between the host genome reference and the insert by locally aligning reads to both the insert and the host genome. In brackets true/false positives are shown.

Sample name		100% Bt rice	10% Bt rice	Bt noodles
Total paired-end reads		123,574,914	93,206,312	69,931,700
Command-Line-Tools	Quality filtered paired-end reads	91,371,164	69,464,211	67,539,855
	Detection approach	3,186	284	5,485
	Proof approach	51 (40/11)	2 (2/0)	98 (41/57)
	Identification approach	49 (34/15)	9 (2/7)	77 (47/30)
CLC Genomics Workbench	Quality filtered paired-end reads	107,455,990	81,491,366	68,981,939
	Detection approach	3,876	339	5,691
	Proof approach	88 (74/14)	6 (4/2)	134 (55/79)
	Identification approach	952 (20/932)	538 (1/537)	514 (24/490)

pure sample a significant part of the sequence data is not derived from the GMOs chromosomal DNA, but from the mitochondrial genome instead. This is not surprising for rice with a single diploid nuclear genome of almost 400 Mbp and a mitochondrial genome of almost 500 Kbp with a copy number of potentially over a 100 per cell (Bendich & Gauriloff, 1984). The main difficulty is that it is not easy to estimate the relative amount of mitochondrial DNA *a priori*. Not only do different species have a different mitochondrial DNA size, but even within a single organism the number of mitochondria per cell is variable between tissues or organs (Mackenzie & McIntosh, 1999; Tian, Zheng, Hu, & Yu, 2006). In addition, it can be difficult to determine/control experimental parameters properly. For instance, the number of high quality paired-end reads N is difficult to know beforehand (Kircher, Heyn, & Kelso, 2011). It is highly dependent on the quality of the raw unfiltered reads. These reads are produced by a whole sequencing process where different batches of reagents are used, errors in detection of fluorescence are possible and cluster density is variable. Most experimental procedures for library preparation generate a range of DNA fragment sizes. This uncertainty can greatly be reduced by size selecting fragments of the library on a gel. In addition, despite variation in library insert sizes, the paired-end distance D after sequencing on an Illumina HiSeq instrument is typically in the range of 100–300 bp, due to the competitive efficiency of small fragments during the cluster formation by bridge-PCR in Illumina instruments.

3.2. Experimental results

3.2.1. Results of the 100% Bt rice sample

Two different platforms, the CLC Genomics Workbench and a combination of Command-Line-Tools, were used to implement the statistical framework. After quality filtering, the detection, proof and identification approach (Fig. 1) were applied to the

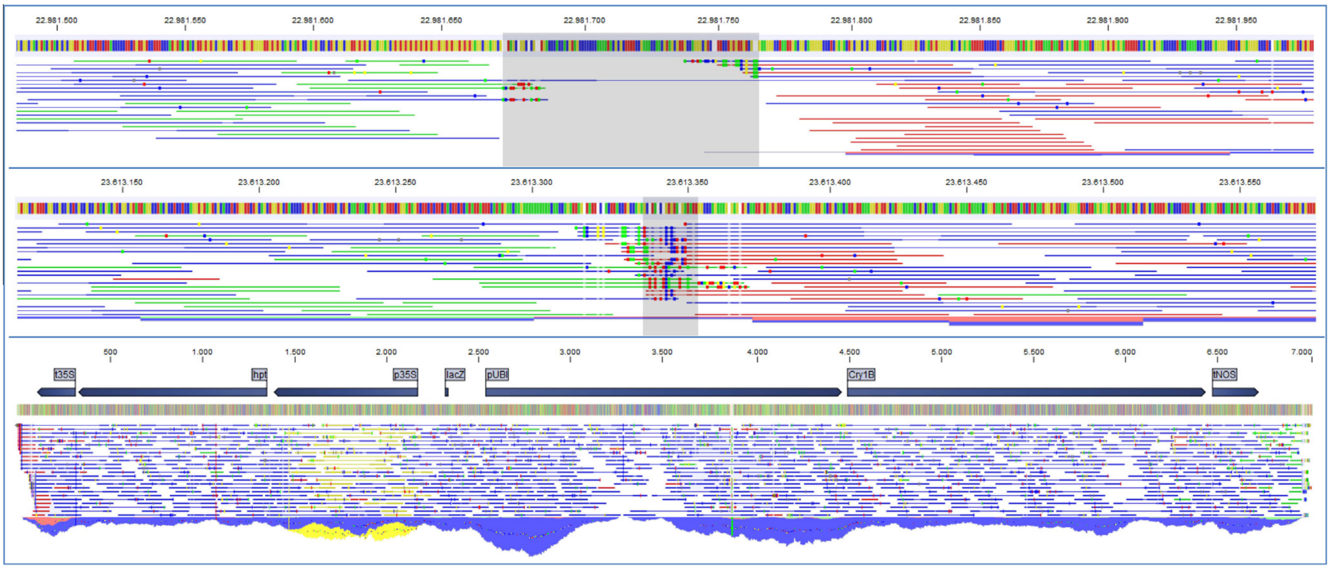

Fig. 2. Global alignment of 100% Bt rice reads using CLC Genomics Workbench. Region 22,981,500–22,981,950 of chromosome II of the host genome (top), region 23,613,150–23,613,550 of chromosome III of the host genome (centre) and the complete insert of length 7002 (bottom), including globally aligned reads using the CLC Genomics Workbench. Reference nucleotides are shown as vertical bars with the four different bases in different colours on top of each image. Below this reference all reads from the 100% Bt rice sample that are globally aligned to this region are shown. Reads with corresponding mates are indicated in blue with a thin line connecting them. Green and red coloured reads do not have their respectively reverse and forward mate pairs aligned in this region. Yellow coloured reads indicate ambiguous reads with multiple possible alignments, in this case corresponding to a repeated region in the promoter p35S on the insert. Mismatches are shown on each read. A clear deletion is present on chromosome II, while a smaller one is detected on chromosome III, indicated by a grey shaded box, this part is replaced by the insert. When the end of a read originates from the insert, but is aligned to chromosome II or chromosome III, multiple mismatches can be detected, i.e. around position 22,981,750 of chromosome II. A single read seems to span the complete insert on chromosome II, although this is unlikely and it is more plausible that a minor contamination with the non-GM type occurred.

100% Bt rice sample on both platforms. Multiple mates globally aligned to the insert on both platforms by using the detection approach. By using the proof approach, their corresponding mates aligned in a small range on chromosome II in the region 22,981,000–22,982,000 and on chromosome III in the region 23,613,000–23,614,000, indicating two independent insert sites (Fig. 2). By using the identification approach, both platforms identified four junctions; (1) at position 22,981,764 of chromosome II and at position 94 of the insert; (2) at position 22,981,674 of chromosome II and at position 6962 of the insert; (3) at position 23,613,353 of chromosome III and position 22 of the insert; (4) at position 23,613,341 of chromosome III and position 6958 of the insert (Fig. 3). False positives, due to PCR artefacts, chimeric reads or genomic similarities between the insert sequence and the host genome, were filtered out by inspection of the alignments and their quality and mapping scores. These results are summarised in Table 2.

3.2.2. Effect of different samples

The 10% Bt rice and 100% Bt noodles samples were analysed in a similar way as the 100% Bt rice sample (Table 2).

Degraded DNA in the Bt noodles sample did not impair the construction of the shotgun sequencing library, because the fragment size of the degraded DNA was larger than the selected fragment size of 300–400 bp for sonication (Supplemental Fig. S1).

A factor 10–20 more reads, aligned to the insert, were detected in the 100% Bt rice sample compared to the 10% Bt rice sample in all the described approaches and platforms. Since the number of quality filtered reads is 1.3 times higher for the 100% Bt rice sample than for the 10% Bt rice sample, the results agreed with an expected factor 13.

3.2.3. Effect of different approaches

The detection approach was used to detect the presence of the insert in the sample and provided a minimum of 284 hits for all the samples (Table 2). Analysis of the read mapping (i.e. Fig. 2 for the 100% Bt rice sample) at nucleotide resolution showed few mismatches in their global alignments, suggesting no or few false positive hits. This number of properly aligned mates highlights the power and significance of the detection approach.

The proof approach was used to prove the integration of the insert within the host genome. In the design of this approach only a subset of the quality filtered paired-end reads, those with aligned mates in the detection approach, was used. For all samples, multiple true positive hits were found, although some false positive hits were observed as well (Table 2). For the 100% Bt rice and 100% Bt noodles samples, the proof approach has provided strong evidence of insert integration into the host genome, while the result for the 10% Bt rice sample is of low significance with a minimum of two detected mates.

Mates covering junctions were detected using the identification approach. A sufficient number of true positive hits were detected to make a strong identification for the 100% Bt rice and 100% Bt noodles samples. The identification of each specific junction, compared to the detection of either the left or right junction necessary for identification, proved to be less reliable, as suggested by the presence of only two identifiable mates for the junction on position 22,981,674 of chromosome II and position 6962 of the insert of the 100% Bt rice sample with the Command-Line-Tools. A moderate and high level of false positives was respectively found for the Command-Line-Tools and for the CLC Genomics Workbench. Since the overlap was relatively small (28 bp) and some mismatches were allowed, the presence of false positive hits was

```
                                       A: Insert – Chromosome II
.................TTCCCGCCTTCAGTTTAAACTATCAGTGTTTGAC GCCACCGCCACTCTCTCCTCTCCTCCTCTCT..........................
.................TTCCCGCCTTCAGTTTAAACTATCAGTGTTTGAC GCCACCGCCACTCTCTCCTCTCCTCCTCTCTCTCTCTCTCTCGCACCACCGCTCTCT
CTTGGATCAGATTGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTGAC GCCACCGCCACTCTCTCCTCTCCTCCTCTCT..........................

                                       B: Chromosome II – Insert
.............GGCCCGCGCCGCTGACACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATTTTGGTTTTAGGAATTAGA.
................................TGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATTTTGGTTTTAGGAATTAGAA
...............................GTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATTTTGGTTTTAGGAATTAGA.
..............GCCCGCGCCGCTGACACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATTT..................
............................CACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTT..............................
...........................ACACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATTTTGGTTTTAGGAAT.....
...............CCCGCGCCGCTGACACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATTTT.................
.............GGCCCGCGCCGCTGACACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATT...................
..............................CGTGGGCCCCCAGCCGACTTAATGGCCGAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTCCTGGATTTTGGTTTTA..........
CCCCCGCGAGCGTGGCCCGCGCCGCTGACACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGA..................................
.............................ACGTGGGCCCCCAGCCGACTTAATGGCCCAAC AACGCCGAATTAATTCGGGGGATCTGGATTTTAGTACTGGATTTTGGTTTTAGGAATTA...

                                       C: Insert – Chromosome III
.CTAGAGCAGCTTGAGCTTGGATCAGATTGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTAGCATTACAGGAAGAAAAAAAAAACAAGAACTATAGTTT...
.....................................CCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTAGCATTACAGGAAGAAAAAAAAAACAAGAACTATAGTTTATT
................................GTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTATCATTACAGGAAGAAAAAAAAAACAAGAACTATAGTTT...
............................TGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTATCATTACAGGAAGAAAAAAAAAACAAG..............
............................TGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTATCATTACAGGAAGAAAAAAAAAA..................
...................GGATCAGATTGTCGTTTCCCGCCTTCAGTTTAAACTATCAGAGTTTG CACAGGCTGGATTGAGTTATCATTACAGGAAGAAAAAAAAAA..................
..................TGGATCAGATTGTCGTTTCCCGCCTTCCGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTATCATTACAGGAAGAAAAAAAAA...................
.............GAGCTTGGATCAGATTGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTATCACTACAGGAAGAAAAAA......................
..........................................TTCAGTTTAAACTCTCAGTGTTTG CACAGGCTGGATTGAGTTATCATTACAGGAAGAAAAA.......................
............TGAGCTTGGATCAGATTGCCGTTTCCCGCCTTGAGTTTAAACTATCAGTGTTTG CACAGGCTGCATTGAGTTATCATTACAGGAAGAAAAA.......................
..TAGAGCAGCTTGAGCTTGGATCAGATTGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTA CACAGGCTGGATTGAGTTATCATTACA.................................
.CTAGAGCAGCTTGAGCTTGGATCAGATTGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTATCATTAC..................................
GCTAGAGCAGCTTGAGCTTGGATCAGATTGTCGTTTCCCGCCTTCAGTTTAAACTATCAGTGTTTG CACAGGCTGGATTGAGTTATCATTA...................................

                                       D: Chromosome III – Insert
.TTCTTAGAAGCTGACTTCCAAACAGCGGATTCTCATCTTAAGCTCACCAAACAGGGCC TATTGTGGTGTAAACAAATTGACGCTTAGACAACTTAATAACACA...................
...CTTAGAAGCTGACTTCCAAACAGCGGATTCTCATCTTAAGCTCACCAAACAGGGCC TATTGTGGTGTAAACAAATTGACGCTTAGACAACT............................
ATTCTTAGAAGCTGACTTCCAAACAGCGGATTCTCATCTTAAGCTCACCAAACAGGGCC TATTGTGGTGTAAACAAATTGACGCTTAGACA..............................
.........................GCGGATTCTCATCTTAAGCTCACCAAACAGGGCC TATTGTGGTGTAAACAAATTGACGCTTAGACAACTTAATAACACATTGCGGAC..........
.TTCTTAGAAGCTGACTTCCAAACAGCGGATTCTCATCTTAAGCTCACCAAACAGGGCC TATTGTGGTGTAAACAAATTGACGCTTAGACAA.............................
.............GACTTCCAAACAGCGGATTCTCATCTTAAGCTCACCAAACAGGGCC TATTGTGGTGTAAACAAATTGACGCTTAGACAACTTAATAACACA...................
............TGACTTCCAAACAGCGGATTCTCATCTTAAGCTCACCAAACAGGGCC TATTGTGGTGTAAACAAATTGACGCTTAGACAACTTAATAACAC....................
```

Fig. 3. Single-end reads covering the junctions for the 100% Bt rice sample, detected with the identification approach using the Command-Line-Tools. The consensus sequence is underlined. The transition between host genome and insert is indicated by a gap. The part of each read belonging to the insert is indicated in bold. (A) Junction with transition on insert position 6962 and chromosome II position 22,981,674. (B) Junction with transition on chromosome II position 22,981,764 and insert position 94. (C) Junction with transition on insert position 6958 and chromosome III position 23,613,341. (D) Junction with transition on chromosome III position 23,613,353 and insert position 22.

Table 3
Common GMO samples and the number of reads N (in millions) needed to find at least one targeted read with a certainty of $P = 0.95$ for each approach as proposed in the statistical framework. Read length $R = 100$, insert length of GMO cassette $I = 7000$ (unstacked and homozygous), overlap length $M = 30$ and paired-end distance $D = 300$ are assumed for each sample. All samples are assumed to be pure genomic DNA. In case of mixtures the DNA ratio r of the GMO can be calculated as explained in Section 2.1. Experiments requiring more data than currently obtained with a single lane on an Illumina Rapid Run (300 million paired-end reads) are indicated in italics.

Species	Genome size *(H)* in Mbp	Number of reads reads (in millions) needed for the		
		Detection approach	Proof approach	Identification approach
		$N = \frac{(ln1-P)}{2\cdot ln(1-r\cdot\frac{1-R+1}{H-R+1})}$	$N = \frac{(ln1-P)}{ln(1-2\cdot r\cdot\frac{D-2\cdot R+1}{H-D+1})}$	$N = \frac{(ln1-P)}{2\cdot ln(1-2\cdot r\cdot\frac{R-2\cdot M+1}{H-R+1})}$
100% GM Rice (*Oryza sativa*)	385 (diploid)	0.08	5.71	7.03
100% GM Sugar beet (*Beta vulgaris*)	758 (diploid)	0.16	11.24	13.85
100% GM Soybean (*Glycine max*)	1115 (diploid)	0.24	16.54	20.37
100% GM Oilseed rape (*Brassica napus*)	1235 (tetraploid)	0.27	18.32	22.56
100% GM Cotton (*Gossypium hirsutum*)	2250 (tetraploid)	0.49	33.37	41.10
100% GM Maize (*Zea mays*)	2300 (diploid)	0.50	34.11	42.01
100% GM Wheat (*Triticum aestivum*)	17000 (hexaploid)	3.69	252.12	*310.53*
1% GM Rice + 99% WT Rice	385 (diploid)	8.36	*570.97*	*703.28*
0.01% GM Rice + 99.99% WT Rice	385 (diploid)	*835.65*	*57,096.88*	*70,327.91*
1% GM Wheat + 99% WT Wheat	17000 (hexaploid)	*368.99*	*25,211.61*	*31,053.32*
0.01% GM Wheat + 99.99% WT Wheat	17000 (hexaploid)	*36,898.63*	*2,521,083.57*	*3,105,081.42*
50% GM Rice + 50% WT Maize	385 (diploid) + 2300 (diploid)	0.58	39.82	49.05
0.1% GM Soy + 99.9% WT Oilseed	1115 (diploid) + 1235 (tetraploid)	*536.07*	*36,627.44*	*45,114.57*

expected. Inspection of false positive hits relied on the fact that the 5' tail of a mate should align to the insert, while the other tail should align to the host or vice versa in regions that were deemed interesting by the proof approach. The consensus sequence for all junctions was in perfect agreement (100% identity) with the DNA sequences originating from the DNA walking technique of Fraiture et al. (2014; personal communication).

3.2.4. Effect of different platforms

The CLC Genomics Workbench provides intuitive implementation and easily interpretable output formats such as figures and graphs, at the cost of full control of all parameters. Due to this limited control, the results are prone to false positives which are not straightforward to avoid and can be hard to identify graphically. An example is the high number of false positives in the identification approach, due to the lack of a parameter that specifies a seed location for the alignment. The CLC Genomics Workbench thus sacrifices some robustness for user-friendliness.

The Command-Line-Tools rely on textual/tabular information, extendable with other tools for visualisation that were not investigated in this article. Different software tools are available, each with their own benefits and drawbacks (Ruffalo, LaFramboise, & Koyutürk, 2010), but in this article only BWA and Bowtie2 were used in combination with custom Python and Perl scripts. While some false positives are inherent to sequencing technology, textual/tabular representation provides easy identification of false positives since they are often single occurrences with low alignment/mapping qualities, as opposed to true positive hits where multiple hits were found per region. Remaining false positives can often be filtered with the right software tools or custom scripts in subsequent steps. An analysis on the Command-Line-Tools is thus less affected by false positives than the CLC genomics Workbench, but knowledge of several tools and/or programming languages is essential.

3.3. Feasibility of using NGS data for GMO detection

All three approaches used to detect, to prove and to identify GMO events provided the same cassettes, junction sequences and number of insertions as those described in previous studies (Fraiture et al., 2014; personal communication). Command-Line-Tools and commercial software for bioinformatics analysis were able to come to the same results for the used samples. These samples; 100% Bt rice, 10% Bt rice and 100% Bt noodles, were of limited complexity. There are reference sequences available for both the insert and the host genome and the flanking regions of the two insertions are known.

To explore the feasibility of NGS for routine analysis using the statistical framework, it was applied to some theoretical samples containing common GMOs as shown in Table 3. For instance, to identify at least one paired-end read that aligns to a 7 Kbp transgene cassette in the rice genome of 384 Mbp with probability of 0.95, about 7 million paired-end reads need to be generated. Larger genomes, like wheat, will need 300 million paired-end reads to achieve the same result. Based on this information, it can be concluded that pure samples consisting of 100% GMO can, at the time of writing, reasonably be characterised with a single lane on an Illumina Hiseq2500 Rapid Run, yielding roughly 300 million paired-end reads per experiment (http://www.illumina.com/), at a standard price range (https://genohub.com/). The required number of sequencing runs and associated costs increase when samples with only trace amounts of 1% GMO or less are investigated. For instance, for a wheat genome sample with trace amounts of 0.01% GMO, more than 30 billion paired-end reads, equal to a hundred Rapid Run lanes, are necessary to be able to only detect the insert with a probability of 0.95. Even this amount of data does not yield a high probability of detecting reads proving host genome integration or identifying the event.

4. Conclusion

The laborious analysis of an increasing number of GMOs using qPCR technology and the ineffectiveness in detecting "unknown and new GMOs" creates a need for alternatives to the current qPCR technology. In this context, NGS, allowing 'detection-by-sequencing' of GMOs in food and feed matrices, was proposed since it circumvents the need to design specific primers to amplify target sequences for each specific GM event. Although some previous studies have shown the successful use of NGS to detect and identify GMOs, only limited information is available about the feasibility/applicability of NGS for routine GMO analysis, hampering its implementation in enforcement laboratories. In the present study a statistical framework was developed that offers preliminary, yet practical information that needs to be considered before NGS becomes routine use in GMO analysis.

Three approaches are considered in the framework: the "detection approach" to detect transgene sequences, the "proof approach" to prove integration of transgenes into the host genome and the "identification approach" to identify the specific transgene event. For each approach, formulae were developed to calculate the probability P of detecting at least one read in an NGS experiment with N reads or vice versa the number of reads N needed for a probability P to detect at least one targeted read. This framework was validated by using experimental data from a 100% pure Bt rice grains sample, a 10% Bt rice grains mixed with 90% non-GM rice grains sample and a noodles containing 100% pure processed Bt rice sample. Robust experimental results were obtained, regardless of implementation of the framework, on both the CLC Genomics Workbench and by using Command-Line-Tools. The experimental results of all three samples agreed with the theoretical results of the "detection approach", "proof approach" and "identification approach".

There are several assumptions and drawbacks of the approaches in the statistical framework. While whole genome complexities are avoided in the analysis by aligning reads to the transgene reference before aligning them to the host genome reference, the reference sequences of the transgenic cassette and host genome are always required *a priori*. Furthermore, the statistical framework was developed to calculate the probability to detect a GMO in a sample with known composition $P(+|\text{GMO})$, and is not fit to calculate the probability that a GMO is truly present when a sample with unknown composition is analysed $P(\text{GMO}|+)$. To achieve the latter, the statistical framework must be further developed to calculate the probability of false positives when no GMO is present in a sample.

Finally, the framework was applied to a range of different samples commonly encountered in routine analysis. It was shown that it is theoretically possible to use NGS to detect and identify samples of 100% GM crops. However, diluted samples and mixtures require large NGS experiments, with billions to trillions of reads and their associated costs, to yield a high probability of finding targeted reads for each approach.

It is concluded that the developed statistical framework can be used to estimate the number of NGS reads needed to detect a GMO in a given sample, and to help decide whether it will be useful to perform a NGS experiment. When the composition of a sample is unknown, the framework can still be used to estimate how many NGS reads are needed to form a hypothesis about the presence of a specific GMO, but no significance testing can be done and any results of an NGS experiment need to be confirmed by targeted molecular analysis in an independent analysis afterwards.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgments

The research that yielded these results, was funded by the Belgian Federal Public Service of Health, Food Chain Safety and Environment through the Project UGMMONITOR (convention RF 11/6242) and an Ylieff Grant. The authors also thank Emmanuel Guiderdoni (CIRAD, UMR AGAP, Biological Systems department, Montpellier, France) for his kindness to provide rice grains.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.foodchem.2015.07.074.

References

Bahrdt, C., Krech, A. B., Wurz, A., & Wulff, D. (2010). Validation of a newly developed hexaplex real-time PCR assay for screening for presence of GMOs in food, feed and seed. *Analytical and Bioanalytical Chemistry, 396*, 2103–2112.

Bendich, A. J., & Gauriloff, L. P. (1984). Morphometric analysis of cucurbit mitochondria: The relationship between chondriome volume and DNA content. *Protoplasma, 119*, 1–7.

Breitler, J. C., Vassal, J. M., del Mar Catala, M., Meynard, D., Marfa, V., Mele, E., et al. (2004). Bt rice harbouring cry genes controlled by a constitutive or wound-inducible promoter: protection and transgene expression under Mediterranean field conditions. *Plant Biotechnology Journal, 2*, 417–430.

Brodmann, P. D., Ilg, E. C., Berthoud, H., & Hermann, A. (2002). Real-time quantitative polymerase chain reaction methods for four genetically modified maize varieties and maize DNA content in food. *Journal of AOAC International, 85*, 646–653.

Broeders, S. R. M., De Keersmaecker, S. C. J., & Roosens, N. H. (2012). How to deal with the upcoming challenges in GMO detection in food and feed. *Journal of Biomedicine and Biotechnology*.

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS ONE, 8*, e85024.

Dörries, H.-H., Remus, I., Grönewald, A., Grönewald, C., & Berghof-Jäger, K. (2010). Development of a qualitative, multiplex real-time PCR kit for screening of genetically modified organisms (GMOs). *Analytical and Bioanalytical Chemistry, 396*, 2043–2054.

Espinoza, C., Schlechter, R., Herrera, D., Torres, E., Serrano, A., Medina, C., et al. (2013). Cisgenesis and intragenesis: New tools for improving crops. *Biological Research, 46*, 323–331.

Foti, N., Onori, R., Donnarumma, E., De Santis, B., & Miraglia, M. (2006). Real-time PCR multiplex method for the quantification of Roundup Ready soybean in raw material and processed food. *European Food Research and Technology, 222*, 209–216.

Fraiture, M. A., Herman, P., Taverniers, I., De Loose, M., Deforce, D., & Roosens, N. H. (2014). An innovative and integrated approach based on DNA walking to identify unauthorised GMOs. *Food Chemistry, 147*, 60–69.

Fraiture, M. A., Herman, P., Taverniers, I., De Loose, M., Nieuwerburgh, F. V., Deforce, D., et al. (2015). Validation of a sensitive DNA walking strategy to characterise unauthorised GMOs using model food matrices mimicking common rice products. *Food Chemistry, 173*, 1259–1265.

Greilhuber, J. (2005). Intraspecific variation in genome size in angiosperms: Identifying its existence. *Annals of Botany, 95*, 91–98.

Holme, I. B., Wendt, T., & Holm, P. B. (2013). Intragenesis and cisgenesis as alternatives to transgenic crop development. *Plant Biotechnology Journal, 11*, 395–407.

Huber, I., Block, A., Sebah, D., Debode, F., Morisset, D., Grohmann, L., et al. (2013). Development and validation of duplex, triplex, and pentaplex real-time pcr screening assays for the detection of genetically modified organisms in food and feed. *Journal of Agricultural and Food Chemistry, 61*, 10293–10301.

James, C. (2013). Global Status of Commercialized Biotech/GM Crops: 2013. ISAAA Brief, 46.

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice, 6*, 4.

Kircher, M., Heyn, P., & Kelso, J. (2011). Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics, 12*, 382.

Köppel, R., Sendic, A., & Waiblinger, H. U. (2014). Two quantitative multiplex real-time PCR systems for the efficient GMO screening of food products. *European Food Research and Technology, 239*, 653–659.

Kovalic, D., Garnaat, C., & Guo, L. (2012). 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*, 149–163.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*, 357–359.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*, 2078–2079.

Liang, C., van Dijk, J. P., Scholtens, I. M., Staats, M., Prins, T. W., Voorhuijzen, M. M., et al. (2014). Detecting authorized and unauthorized genetically modified organisms containing vip3A by real-time PCR and next-generation sequencing. *Analytical and Bioanalytical Chemistry, 406*, 2603–2611.

Mackenzie, S., & McIntosh, L. (1999). Higher plant mitochondria. *The Plant Cell, 11*, 571–585.

Michael, T., & Jackson, S. (2013). The first 50 plant genomes. *The Plant Genome, 6*, 1.

Morisset, D., Novak, P., Zupanic, D., Gruden, K., Lavrac, N., & Zel, J. (2014). GMOseek: A user friendly tool for optimized GMO testing. *BMC Bioinformatics, 15*, 258.

Ohri, D. (1998). Genome Size Variation and Plant Systematics. *Annals of Botany, 82*(Suppl. A), 75–83.

Reg. EC n°1829/2003: REGULATION (EC) No 1829/2003 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 September 2003 on genetically modified food and feed (2003).

Reg. EC n°1830/2003: REGULATION (EC) No 1830/2003 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 September 2003 concerning the traceability and labelling of genetically modified organisms and the traceability of food and feed products produced from genetically modified organisms and amending Directive 2001/18/EC (2003).

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., et al. (2013). Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS ONE, 8*, e66621.

Ruffalo, M., LaFramboise, T., & Koyutürk, M. (2010). Comparative analysis of algorithms for next-generation sequence read alignment. *Bioinformatics, 27*, 2790–2796.

Ruttink, T., Demeyer, R., Van Gulck, E., Van Droogenbroeck, B., Querci, M., Taverniers, I., et al. (2010). Molecular toolbox for the identification of unknown genetically modified organisms. *Analytical and Bioanalytical Chemistry, 396*, 2073–2089.

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

Tian, X., Zheng, J., Hu, S., & Yu, J. (2006). The rice mitochondrial genomes and their variations. *Plant Physiology, 140*, 401–410.

Van den Bulcke, M., Lievens, A., Barbau-Piednoir, E., MbongoloMbella, G., Roosens, N. H., Sneyers, M., et al. (2010). A theoretical introduction to "Combinatory SYBR®Green qPCR Screening", a matrix-based approach for the detection of materials derived from genetically modified plants. *Analytical and Bioanalytical Chemistry, 396*, 2113–2123.

Wahler, D., Schauser, L., Bendiek, J., & Grohmann, L. (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 Analytical Methods, 6*, 1718–1727.

Waiblinger, H. U., Ernst, B., Anderson, A., & Pietsch, K. (2008). Validation and collaborative study of a P35S and T-nos duplex real-time PCR screening method to detect genetically modified organisms in food products. *European Food Research and Technology, 226*, 1221–1228.

Yang, L., Wang, C., Holst-Jensen, A., Morisset, D., Lin, Y., & Zhang, D. (2013). Characterization of GM events by insert knowledge adapted re-sequencing approaches. *Scientific Reports, 3*, 2839.