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