Trends in Food Science & Technology 45 (2015) 319–326

Contents lists available at ScienceDirect

Trends in Food Science & Technology

journal homepage: http://www.journals.elsevier.com/trends-in-food-science-and-technology

Commentary

Next-generation sequencing as a tool for the molecular characterisation and risk assessment of genetically modified plants: Added value or not?

Katia Pauwels [a, *, 1], Sigrid C.J. De Keersmaecker [b, **, 1], Adinda De Schrijver [a], Patrick du Jardin [c], Nancy H.C. Roosens [b, 2], Philippe Herman [a, 2]

[a] Scientific Institute of Public Health, Biosafety and Biotechnology Unit, J. Wytsmanstraat 14, B-1050 Brussels, Belgium
[b] Scientific Institute of Public Health, Platform of Biotechnology and Molecular Biology, J. Wytsmanstraat 14, B-1050 Brussels, Belgium
[c] University of Liège, Gembloux Agro-Bio Tech, Unit of Plant Biology, Passage des Déportés, B-5030 Gembloux, Belgium

ARTICLE INFO

Article history:
Received 10 January 2015
Received in revised form
19 June 2015
Accepted 14 July 2015
Available online 17 July 2015

Keywords:
Genetically modified organism (GMO)
Genetically modified plants (GMP)
Risk assessment
Molecular characterisation
Next-generation sequencing (NGS)

ABSTRACT

Background: Legislations and international organizations provide a framework to ensure proper risk assessment of **Genetically Modified Organisms** (GMO). With regard to the deliberate release of GMO as food or feed, applications for Genetically Modified Plants (GMP) typically contain data for the molecular characterisation at the nucleic acid level based on Southern blot and polymerase chain reaction analysis in combination with Sanger sequencing. Along with the diverse range of applications of **next-generation sequencing** (NGS) in genomic research, some recent research projects and product developers explored the use of NGS as an alternative tool for meeting the data requirements for the molecular characterisation of GMPs in view of their risk assessment.

Scope and approach: By means of a literature survey and information collected through the organisation of an international workshop, we investigated whether NGS can replace and/or complement the currently used techniques for **molecular characterisation** of GMP taking into account the possibilities and current bottlenecks of NGS technologies and recent developments in molecular breeding.

Key findings and conclusions: We conclude that although NGS might present clear advantages for product developers, NGS currently does not always offer a significant added value with respect to the **risk assessment** of GMPs. However, the approaches used so far may soon be further challenged by the fast evolution in NGS technologies and also by the recent developments in molecular breeding of plants. We postulate that setting up a common workflow for the generation of relevant and interpretable data by NGS would facilitate a scientifically sound assessment of GMPs.

1. Introduction

Several legislations and international organizations provide a framework to ensure proper risk assessment (RA) of Genetically Modified Organisms (GMO) with respect to their possible impact on human/animal health and environment. With regard to the deliberate use of GMO as food or feed, for which new applications in Europe and other parts of the world mainly concerns GM plants (GMP) so far, the aim is to investigate whether the genetic modification could inadvertently increase the potential toxicity or allergenicity of the recipient plant, and alter its nutritional quality. Molecular characterisation of the GMP is required to get insight into the genomic locus/loci of modification, the potential impact on the function of the interrupted endogenous genes (e.g. by gene disruption) or the generation of new open reading frames (ORF) at the site(s) of insertion. In the European Union (EU), molecular characterisation data not only serve RA, they are also a prerequisite for the development of detection and identification techniques in view of the traceability and labelling requirements of GMP prior to their regulatory approval for release (European Parliament, 2003b).

* Corresponding author.
** Corresponding author.
E-mail addresses: Katia.Pauwels@wiv-isp.be (K. Pauwels), Sigrid.DeKeersmaecker@wiv-isp.be (S.C.J. De Keersmaecker), Adinda.DeSchrijver@wiv-isp.be (A. De Schrijver), Patrick.DuJardin@ulg.ac.be (P. du Jardin), Nancy.roosens@wiv-isp.be (N.H.C. Roosens), Philippe.Herman@wiv-isp.be (P. Herman).
[1] Equally shared first author.
[2] Equally shared last author.

http://dx.doi.org/10.1016/j.tifs.2015.07.009

The latter drives enforcement laboratories to continuously update tools and technical capacities that could facilitate standardized, reliable and accurate molecular characterisation on a routine basis (Broeders, De Keersmaecker, & Roosens, 2012; Holst-Jensen, 2009). Molecular characterisation data at the nucleic level of GMP serving the RA are currently obtained by Southern blot (SB) and polymerase chain reaction (PCR) analysis in combination with Sanger sequencing approaches to determine the precise location of the junction between the transgenic insert and the host genome, and to detect the possible presence of backbone sequences of the transformation vector. Indeed, GMP of the first-generation technologies, are obtained by the introduction of recombinant DNA from organisms including bacteria and viruses and non-crossable plants, resulting in non-targeted insertion of a large cassette of foreign DNA and the creation of junction sequences encompassing one of the borders of the foreign DNA-insert and the adjacent native plant DNA-sequence. Recent developments in plant breeding include the increasing number in stacked events (Kok et al., 2014), the expanding variety of transformed species and new traits and novel techniques for genetic modification of plants. The latter concern techniques that enable the introduction of novel traits in a plant species without the introduction of large sequence of DNA from non-crossable species, i.e. techniques such as site-directed nuclease-mediated mutagenesis or oligonucleotide-directed mutagenesis (SDN-1, SDN-2 or ODM), inducing targeted specific changes in the plant genome that only encompass a limited number of base pairs, thereby showing many similarities with plants obtained by classical mutagenesis. Whether plants obtained by such site-directed mutagenesis techniques should fall within the scope of the European GMO regulatory framework is still matter of debate (Breyer et al., 2009; Hartung & Schiemann, 2014; Heap, 2013; Pauwels, Podevin, Breyer, Carroll, & Herman, 2014; Podevin, Devos, Davies, & Nielsen, 2012). However, if regulated, the molecular characterisation requirements by means of currently used approaches (including SB analysis) could clearly pose challenges. Some recent publications (Kovalic et al., 2012; Yang et al., 2013) propose the ongoing advances in DNA-sequencing, the so-called 'next generation sequencing' (NGS) or high-throughput sequencing technologies as an additional, or even replacing, tool for molecular characterisation of GMP.

The increasing sequencing throughput possibilities, at continuously decreasing costs, combined with the digital nature and the tuneable resolution of NGS technologies have paved the way for its implementation in several areas of genomic research and in a diverse range of applications. In the area of GMP, NGS can facilitate post-transformation screening during product development by increasing time effectiveness, scalability and automation in the selection of potential valuable events on the basis of their molecular profile (Heck et al., 2005). Additionally, compared to SB analysis, NGS has been reported as an efficient approach to achieve molecular characterisation of GMP (Kovalic et al., 2012). On the other hand, the currently used NGS technologies have some limitations, are more expensive than the SB approach and require a profound expertise in bioinformatics as they are producing a considerable amount of data that needs to be analysed and interpreted.

In this viewpoint we aimed at identifying how NGS data could contribute to the molecular characterisation of GMP in light of their RA and regulatory approval, but also of plants obtained through novel directed molecular techniques, and whether NGS has in this respect an added value as compared to the currently used techniques. To answer this question, several consensus documents issued by FAO, OECD and EFSA were consulted which provide data requirements for the molecular characterisation of GMP (Devos et al., 2014; EFSA Panel on Genetically Modified Organisms (GMO), 2011; Food and Agricultural Organization of the United Nations [FAO], 2003; FAO, 2008; Organisation for Economic Cooperation and Development [OECD], 2010). Furthermore, we used the information gathered during an international workshop on the use of NGS for the molecular characterisation of GMO that we organized with attendees from academia, NGS platforms, companies, GMO laboratories and advisory bodies. The EU regulatory framework was used to investigate whether NGS can replace and/or complement the currently used techniques (European Commission, 2013), considering the possibilities and bottlenecks of the present NGS platforms and the potential challenges for molecular characterisation posed by recent developments in plant breeding. Although some elements discussed in this viewpoint might apply to GMO in general, we focused on GMP because a vast majority of applications submitted under Regulation (EC) No 1829/2003 in the EU, but also submitted in other parts of the world, concerns plants, thereby justifying the more urgent need to address the questions posed in the viewpoint specifically for plants.

2. Data requirements for the molecular characterisation of GMP

In the EU, GMP market registration applications for food and/or feed are submitted under Regulation (EC) No 1829/2003 (European Parliament, 2003a). EFSA's Scientific Panel on GMO (GMO Panel) has developed guidance documents for the preparation and presentation of such applications (e.g. (EFSA Panel on Genetically Modified Organisms (GMO), 2011)). It provides the rationale for data requirements of the different components of RA and describes the information needed for the molecular characterisation of GMP. At the nucleic acid level, the provided information should include the description of nucleic acids intended for transformation and of any (vector) sequence that could be potentially inserted to the recipient plant, and a description of the nucleic acids actually inserted in the plant, including sub-cellular location, copy number and sequence. Further, the genetic stability of the insert over several generations needs to be established as well as the occurrence of novel ORFs in the region spanning the inserted DNA and the native plant DNA (Table 1, first column, point i to vii).

Recently, the RA approach for GM food and/or feed detailed in the EFSA guidance has been incorporated into a legal text, the Implementing Regulation (IR) (EU) No 503/2013, covering only applications concerning GM plants for food and feed uses, and which presents some differences in RA requirements compared with the former. For example for stacked transformation events, comparisons of sequences of the inserts and flanking regions should be carried out between the GMP containing stacked transformation events and their corresponding single events, while the guidance document only demands to control insert integrity in the stacked compared to corresponding single events (Table 1, vii). We used the IR, containing one of the most explicit data requirements for molecular characterisation of GMP intended for food and feed as a starting point to explore potential added values of NGS for the molecular characterisation of GMP compared to currently used techniques (Table 1).

The currently used techniques to fulfil the data requirements for molecular characterisation at the nucleic acid level are SB analysis, a technique enabling a specific restriction fragment to be detected against a background of many other restriction fragments by using a probe (Southern, 1975), and PCR analysis in combination with Sanger sequencing approaches (Table 1, column 2). Up-to-date databases are used for bioinformatic analysis of the sequence at the insertion site. When using appropriate controls and probe/restriction enzyme combinations providing complete coverage of sequences that could be inserted into a given plant genome, SB

Table 1
Possibilities and limitations of NGS, compared to conventional approaches, for molecular characterisation of GMP in view of their risk assessment.

EU data requirements for molecular characterisation (according to IR)[a]	Covered by current techniques (SB and Sanger)	Covered by NGS today or in the near future	Added value of NGS (now or in the near future) compared to current techniques[b]
i) Insert(s) 1) Size and copy number of all detectable inserts (complete or partial) 2) Absence of vector backbone	Yes, depending on the overlap between the probe and the hybridised restriction fragment, the hybridization/washing stringency conditions and size limitation (both maximal and minimal) of the detected fragments	1) Yes: number of junction sequences ('chimeric' sequences) found 2) Yes, if no unintended junction sequences are detected	Whole genome resequencing: small inserts will be spotted, standardized procedure (independent of probes contrary to target enrichment or SB). Less starting material (DNA) is required
ii) Sub-cellular location(s) of insert(s) e.g. nucleus, chloroplast, mitochondria or maintained in non-integrated form	Not possible with SB only, need for additional Sanger sequencing of the flanking regions	Yes, via sequence analysis of the flanking region	Today: no added value Near future: increased cost-effectiveness and streamlined procedure
iii) Organisation and sequence of the inserted genetic material at each insertion site (Also required for stacked events)	Not possible with SB only, need for additional Sanger sequencing	Today: assembling small read lengths is challenging. Paired-end sequencing, longer reads or 'read walking' (Wahler et al., 2013) increases feasibility	Today: Sequencing of insert is possible but still needs to be confirmed by classic PCR and Sanger sequencing. Near future: sequencing of inserts by deep sequencing and longer sequencing reads
iv) Sequence information for both 5′ and 3′ flanking regions at each insertion site	Not possible with SB only, need for additional Sanger sequencing	Yes	Today: no added value Near future: may be possible by the use of longer reads
v) Information on creation of ORFs present within insert and region spanning the junction site and verification of potential similarities of these ORF with known toxins or allergens (bioinformatics analysis)	Not possible with SB only, need for additional Sanger sequencing	Yes. Might be more efficient via capture. ORF analysis requests accurate sequence information	Today: no added value. Near future: may be possible by the use of longer reads
vi) Genetic stability of the events	Yes, stability is shown when SB pattern is consistent among all generations studied	Yes, stability is shown when patterns of detected junction sequence is consistent among all generations studied	Today: No added value Near future: increased cost-effectiveness and streamlined procedure
vii) Stacked events: control of insert integrity in the stack compared to corresponding single events and comparison of sequences of the inserts and the flanking regions obtained from GM plants containing single events and plants containing stacked transformation events.	Yes, but becomes more cumbersome with increasing number of inserts, combination with Sanger sequencing needed	Today sequencing of insert is possible but still needs to be confirmed by classic PCR and Sanger sequencing. Near future: sequencing of inserts by deep sequencing.	Today: Possible advantage in case of multiple insertions Near future: increased cost-effectiveness for stacks with high number of inserts owing to universal, high throughput method combined with standardized procedures
viii) Information on the expression of the intended and unintended inserted/ modified sequences (e.g. data at the protein, metabolite level and/or RNA level)	Covered by approaches such as Western bot, ELISA, Northern blot and methods to determine metabolites such as liquid and, gas chromatography, and capillary electrophoresis. Sanger sequencing of EST populations is a way to perform transcriptomic analysis.	Today: For the RNA level*: Yes, RNA-Seq * NGS is a method involving nucleic acids, so information at the protein or metabolite level cannot be directly measured with this technology	In the future: potential transcriptome profiling, sRNA profiling in case of RNAi-based GMP

[a] Several consensus documents issued by FAO, OECD and EFSA, were consulted which provide data requirements for the molecular characterisation of GMO (Devos et al., 2014; EFSA Panel on Genetically Modified Organisms (GMO), 2011; FAO, 2003; FAO, 2008; OECD, 2010). However, the Implementing Regulation (IR) (EU) No 503/2013 was taken as source of data requirements to elaborate on the possibilities and limitations of NGS for molecular characterisation of GMP in view of their risk assessment.
[b] Some of these points have also been discussed during the international workshop on the use of NGS for the molecular characterisation of GMP that we organised.

analysis is used to reveal the number of insertion sites, the copy number at each insertion site, the genetic elements (e.g. promoters, enhancers or backbone sequences) inserted and genetic stability (Table 1, column 2, i and vi). SB analysis is sometimes complemented with real-time PCR (possibly also with digital PCR) to determine the copy-number of the insert. PCR analysis of the entire insert as well as of the flanking sequences, combined with Sanger sequencing, is used to determine the exact DNA-sequence of the transgenic locus (Table 1, column 2, iii, iv and v). Limitations of SB include the high amount of input DNA required, multiple manual work interventions, the use of agarose gels, case specific use of restriction enzymes and design of radioactive probes (thereby already anticipating the results), and the impossibility to clearly identify a high number of copies, all aspects that may become cumbersome when dealing with GMP with stacked transformation events. Also the genetic changes obtained by new molecular breeding techniques are difficult (impossible) to be characterised by SB. Recently, NGS has been proposed as an alternative to these current approaches for molecular characterisation of GMP (Kovalic et al., 2012; Yang et al., 2013). Another requirement for molecular characterisation of GMP is to demonstrate whether the inserted/ modified sequence results in intended changes at the protein, RNA or metabolite level (Table 1, column 1, viii). The current approaches to verify this include Western blot, ELISA, Northern blot and compositional analysis (e.g. gas liquid chromatography). However, except for data at the RNA level, NGS does not offer a direct alternative to these approaches.

3. Possibilities offered by next-generation sequencing for RA of GMP

Theoretically, starting from the (complete) genome sequence of a GMP, the data requirements at the nucleic acid level as set out in IR (see Table 1) could be collected. However, it is not known to which extent NGS can accurately deliver this (complete) genome sequence.

NGS DNA sample preparation involves the shearing of the genomic DNA, using starting material in lesser quantities than needed for SB (up to more than 10-fold less), the selection of DNA-fragments of the appropriate size and the subsequent library

construction of DNA-fragments that are sequenced in parallel reactions. The obtained strings of bases, called reads, are then aligned or assembled to theoretically reveal the entire sequence of the DNA sample. Most interesting is the preparation of a paired-end library, allowing to sequence both ends of a DNA fragment where the ends are separated by a known distance. This will result in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome, which is an important issue for plant genomes (Imelfort & Edwards, 2009).

Assembly of (a selected set of) the reads can be done by mapping to a reference sequence, i.e. a known (genome) sequence (reference-based mapping or re-sequencing). Subsequently, the sequence of the aligned reads can be subjected to further characterisation. For obtaining molecular characterisation data for RA of GMP, the applicability of NGS by the reference-based approach has been demonstrated for a GM soybean (Kovalic et al., 2012) (Table 1, i column 3). The transformation vector and the reference genome of the native soybean were used as queries for the selection of reads, which were then further subjected to bioinformatic analysis in order to characterise the number and identity of novel ORF obtained upon genetic modification. ORF at the junction region (the so-called junction sequences) were characterised based on the presence of reads covering both the transformation plasmid sequence and a sequence likely derived from either the native plant genome flanking the insertion or either rearranged transformation plasmid DNA. Further, the backbone portion of the transformation vector served to reveal whether backbone sequences were introduced into the plant genome during the transformation process. In the same study, the reference-based approach was also able to examine a GM soybean event with multiple inserts and rearrangements (Kovalic et al., 2012). Another study showed the potential of NGS for the molecular characterisation of GM rice events, using the sequence of the host reference genome and the plasmid to map the reads, thereby revealing an unintended transgene insert (599 bp) which had not been identified using the state-of-the-art SB approach (Yang et al., 2013) (Table 1, i and iii, column 3). It is important to note that whilst NGS results identified the presence and localisation of inserts, and can also sketch a map of the insert sequence (Yang et al., 2013), the insert sequence was still verified and validated by PCR and Sanger sequencing, because with the NGS analysis, no accurate assembly of the total long insert is possible yet due to the short sequence reads length (Kovalic et al., 2012; Yang et al., 2013).

Assembly of reads is also possible in the absence of a reference sequence (*de novo* assembly, *de novo* sequencing). In the field of GMP, only *de novo* assembly restricted to the region of the insert sequence has been reported yet (Liang et al., 2014; Wahler, Schauser, Bendiek, & Grohmann, 2013; Yang et al., 2013). Combined with the appropriate bioinformatic analysis, a *de novo* assembly limited to the reads not mapping to the host reference genome sequence could offer interesting opportunities for the identification and characterisation of (un)authorized GMP for which the insert sequences in the genome are *a priori* unknown and thus cannot be used as a query (Liang et al., 2014; Wahler et al., 2013; Yang et al., 2013). Indeed, research projects related to detection and identification strategies of GMP (both authorized and unauthorized) currently also explore the use of NGS and appropriate bioinformatics for data analysis and hence share similar approaches and questions with regard to the generation of relevant NGS data (Liang et al., 2014; Wahler et al., 2013; Yang et al., 2013).

NGS offers the ability for whole genome sequencing. For an accurate assembly of the complete genome to be technically possible using currently available sequence assembly tools, sufficiently high sequence coverage and long sequence reads are needed (Kovalic et al., 2012; Sims, Sudbery, Ilott, Heger, & Ponting, 2014). The coverage and length of the sequence reads depend on the size of the genome, the number of samples included in one sequencing run, the type of library constructed and the type of NGS platform used. As elaborated above, the currently used approaches for NGS data analysis to characterize GMP do not involve complete genome assembly, given the large genome size of GMP and the concomitant high cost linked to sequencing the full genome, although the complete genome is used as input DNA in the sequence analysis for the identification of the insertion loci. A prior physical enrichment (i.e. capture approach) of the targeted sequence by specific capture through probes designed complementary to the regions of interest has been developed by several companies as an efficient alternative method for the identification of the insertion loci (Bodi et al., 2013; Hunst, 2013). In the subsequent targeted sequencing approach, a library from genomic DNA of the captured regions only is used as input. One such commercially developed method was recently used to identify T-DNA insertion sites in *Arabidopsis* mutants (Lepage, Zampini, Boyle, & Brisson, 2013). By using the transformation vector sequence as template to design the probes, the DNA-library was enriched with fragments containing the insert and junction sequences. Instead of *in silico* selection of sequencing reads mapping to the transformation vector (Kovalic et al., 2012), the NGS data are enriched in reads of the insert and junction sequence. The data will be more straightforward to interpret, without the need for a reference host genome sequence. As less reads are used for regions which are not of interest (native host genome sequence), more samples can be pooled in one analysis, thereby seriously reducing the cost.

It is important to note that studies using NGS for the molecular characterisation of GMP so far have only been done on GMP obtained by first-generation technologies. It is worth considering the possibilities offered by NGS technologies in light of the use of novel plant breeding techniques. NGS has already been applied for the identification of small insertions and deletions (indels) in *Arabidopsis thaliana* (Cao et al., 2011) and rice (Wahler et al., 2013), so NGS could potentially offer an advantage compared to SB in terms of the detection of small sequence modifications.

NGS also enables transcriptome profiling (Chu & Corey, 2012) and can elucidate potential altered expression profiles of genes flanking the transgene insert (Table 1, viii). It could also inform on the relative abundance or composition of the pool of small regulatory RNAs (siRNA pool) in GMP designed to induce silencing of target genes (referred as RNAi-based GMP in Fig. 1) (Guo, Li, Wang, & Liang, 2015). However, some NGS methods determining siRNA pools are confronted with bias and caution is needed in the interpretation of data for risk assessment purposes (Ramon et al., 2014).

4. Technical bottlenecks of NGS in view of molecular characterisation of GMP

Despite the advantages of NGS for some of the aspects of molecular characterisation of GMP, there are also some bottlenecks linked to the limitations of the currently available NGS technologies (McGinn & Gut, 2013; van Dijk, Auger, Jaszczyszyn, & Thermes, 2014), for both data production and data analysis (Table 1, columns 3 and 4).

As the NGS technology produces a huge amount of short sequencing reads, the correct and accurate assembly of these reads is a prerequisite to use these kind of data for the molecular characterisation of GMP. However, this assembly is subject to the following technical bottlenecks.

A first limitation is related to the small-size reads produced by the currently available NGS platforms ($\leq$300 nucleotides), their size depending on the library and NGS platform used (McGinn & Gut, 2013; Thudi, Li, Jackson, May, & Varshney, 2012; Wei, Bemmels, &

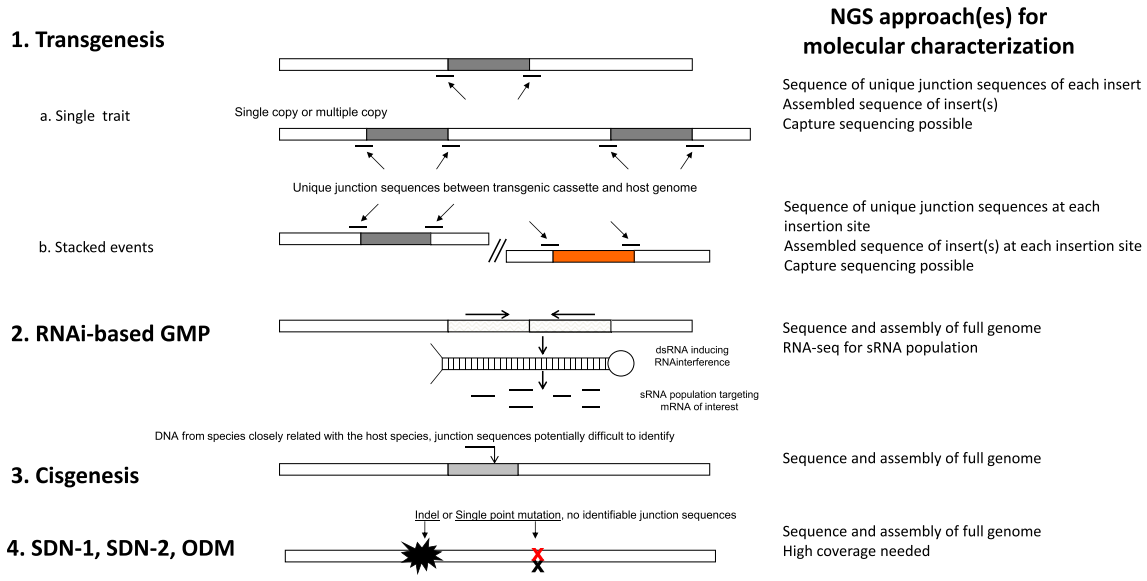

Fig. 1. Schematic illustration of techniques of genetic modification challenging the molecular characterisation and/or identification of the resulting GM plants. 1.a. Transgenic **GMP** obtained by **recombinant DNA techniques** resulting in plant genomes (white bars) with one or several foreign DNA-inserts (black bars) with their respective junctions. A junction is a novel chimeric region encompassing one of the borders of the DNA-insert and the adjacent native genomic DNA sequence. 1.b. **GMP** containing stacked transformation events obtained by conventional crossing of lines obtained by 1.a. and resulting in plant genomes with different foreign DNA-inserts (black bar and red bar). Several junctions need to be characterised.2. **RNAi-based GMP**, designed to induce silencing through RNA interference. This can be obtained by the insertion of a cassette bearing the sense and anti-sense DNA (shaded bars) gene to be silenced.3. **Cisgenic plants**: DNA-inserts originates from the same or crossable species. The inserted DNA (grey bar) may have sequence similarities with the endogenous sequences of the parental organism.4. **GMP obtained by SDN-1, SDN-2** or **ODM**. SDN-1 and SDN-2 are two methods using site-directed nucleases (SDN) generating site-specific double strand breaks. Double strand breaks are repaired either by non-homologous end-joining (SDN-1) or by homologous recombination (SDN-2). Oligonucleotide mediated mutagenesis (ODM) uses chemically synthetized oligonucleotides to induce specific mutations at the target sequence. For SDN-1, SDN-2 or ODM the site specific mutations consists of changes of single or few base pairs, short deletions or insertions (indel) which are not distinguishable from naturally occurring plants. No identification of junction proper to the genetically modified plants are possible. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Dick, 2014). The size will impact the quality of the assembly, i.e. increased read length enables the assembly of larger contigs. This is especially crucial when conducting *de novo* sequencing. High coverage libraries with larger DNA-fragment sizes, combined with paired-end sequencing protocols and the use of NGS platform producing longer read lengths will result in less fragmented assemblies and larger contigs. Although NGS technologies allowing longer read lengths (1–30 kb) exist, these are currently associated with increased error rates which hamper accurate sequencing (Quail et al., 2012). Therefore, with the objective to obtain accurate sequence information of the inserted sequences (often >10 kbp), NGS approaches for molecular characterisation of GMP are still associated with Sanger sequencing for verification and validation, until further progress is made by the NGS platforms in delivering longer and accurate read lengths. Eventually, increased read lengths will also facilitate the capture approach elaborated above, by extending the sequence reads of the flanking regions present in the captured fragments. Additionally, longer read lengths can solve alignment problems associated with repetitive regions which often occur in plant genomes (Liang et al., 2014). There is a need for appropriate assembly tools for GMP since current assembly programs will automatically discard highly repetitive regions. So, unless sequence reads are sufficiently long to extend over the repetitive regions, transforming DNA inserted in a highly repetitive region will not easily be detected.

Besides repetitive regions, some other parts of the genome seem also more difficult to cover. This can be deduced from the observation that there is always an unpredictable part of the genome missing in the NGS data. This is another technical limitation for the assembly of the NGS reads. Lack of full genome accessibility may be due to GC-rich regions (Sims et al., 2014) or to heterochromatin structures, which could also hamper effective restriction endonuclease digestion of the DNA, and hence also the detection of the corresponding fragments by SB analysis.

With parameters such as the coverage or 'average sequencing depth' (Sims et al., 2014), NGS allows for a quantification of genome accessibility. However, even if this parameter becomes quantifiable, it remains questionable what minimal coverage should be required to comprehensively cover the genomes. For the NGS approach, the higher the coverage, the higher the associated cost, but a sufficiently high coverage is needed to allow for proper quality data analysis (Kovalic et al., 2012; Sims et al., 2014). Therefore a balance or trade-off between coverage and cost of analysis should be made, and this might still be a technical limitation for read assembly. With respect to the IR requirements, the question is raised concerning the minimal required coverage for assessing the presence of all detectable inserts, the absence of backbone sequences or the modification of a limited number of base pairs in a genome. Targeted sequencing, using the appropriate probes, might overcome this challenge, by reserving the generated reads to the fragments that hybridized to the probes, thereby increasing the coverage of the regions of interest.

Theoretically all single nucleotide polymorphisms (SNPs), insertions, deletions that are present in the genome and which are inferior to the length of the reads are retrievable when mapping the reads of a sample with an appropriate reference genome. The use of this reference genome is an additional limitation for the assembly and subsequent data analysis. In practice, accurate and complete data will depend on the quality, availability and the choice of the reference genome, including for those with highly repetitive sequences. Here the plasticity of crop genomes should be taken into account as it leads to genomic diversity (Weber et al., 2012). This will impact the appropriateness of the chosen reference genome and hence the quality of the data analysis. Also in the advent of the

molecular characterisation of GMP obtained by site-directed mutagenesis (Fig. 1) the choice of the genome of reference will also determine the outcome of the data. However, irrespective of the technical feasibility of NGS, deliberately introduced SNPs will hardly be distinguishable from naturally occurring variation that may occur during the breeding process.

For cisgenesis, which is another upcoming plant breeding technique consisting in introducing a gene (cisgene) from a crossable sexually compatible organism (same species or closely related species), the reassembling approach of NGS will be challenging. Indeed, in cisgenic plants the inserted gene includes its native promoter and terminator originating from the same or a closely related species and shows sequence identity or similarity to endogenous sequences of the recipient genome. Therefore, the use of the transforming DNA as a query sequence may be problematic. So, unless a full genome assembly can be obtained, the cisgene could be difficult to detect using NGS. However, while NGS could still offer a possibility for this issue in the future, the molecular characterisation of cisgenic plants by the traditional SB approach will remain fastidious.

Another bottleneck in the use of NGS for molecular characterisation of GMP is that the digital nature of NGS also has a profound impact on the data analysis and interpretation, accessibility, storage, interpretation and visualization (Nekrutenko & Taylor, 2012), both for the companies entering the dossier as for the risk assessors. While the analysis of data obtained by means of SB analysis deals with imaging issues and issues associated to image interpretation, the application of NGS technologies is computationally demanding. The substantial amount of raw data generated implies the need for large data storage capacity, appropriate tools for algorithms, computational infrastructure including high performance computing and bioinformatics skills. Also, to further optimize the data analysis, specific bioinformatics tools need to be developed, such as dedicated assembler tools for GMP NGS data, taking into account the current bottlenecks inherent to the type of material to be sequenced, as described above (Liang et al., 2014). In addition, the use of different analysis programmes or parameters to handle the amount of NGS generated data can lead to different conclusions on the molecular characterisation. This makes the set-up of criteria to retrieve and present relevant and interpretable data of good quality both important and challenging. Clarity on the parameters used for data analysis will enhance transparency and improve reliability of subsequent evaluations. This can be obtained by access to primary data, information on the reference genome in case of read mapping, knowledge of the used software for data analysis and insights in the choice of data visualisation including quality-values (Q-values) for each of the bases determined (Nekrutenko & Taylor, 2012).

5. Will advanced information at the nucleic acid level, offered by NGS, contribute to the RA of GMP?

As elaborated in Table 1 column 3, NGS will cover all data requirements for RA at least to the same extent as the currently used techniques (column 2). Even more, NGS could potentially detect unintended insertions or modifications that due to their small size are not revealed by SB (Yang et al., 2013). Irrespective of the fact that NGS may reveal this information, and clearly extends those in amount currently obtained by SB and Sanger sequencing, it should be questioned whether this extended information at the nucleic acid level would actually contribute to RA. One of the current debates with regard to the molecular characterisation of GMP, irrespective whether this was done by NGS, is to what extent the unintended presence of small inserts in the plant genome as a result of the transformation needs to be identified and characterised taking into account that unintended effects are for GMP also assessed on the basis of agronomic, phenotypic and compositional properties. Depending on the biology of the plant species and on genetic linkage, such small inserts may be separated from the functional insert by genetic segregation in the progeny of the primary transformant. As a consequence, where the transformation event is introduced into different genetic backgrounds by conventional breeding, the small inserts may not be present anymore in the plant material placed on the market. Additionally, unintended effects also arise from conventional breeding through molecular mechanisms that naturally occur in plants (Ossowski et al., 2010; Schnell et al., 2015; Vaughn & Bennetzen, 2014). Undesirable phenotypes are removed during selection and breeding programs and no such risk/safety assessments with molecular characterisation requirements are asked for plants obtained by conventional breeding. It has to be remarked that the assessment of unintended effects might be different for GM animals, as no documents describing which agronomic, phenotypic and compositional data are needed for the risk assessment of GM animals, are yet available.

It is also possible that NGS could present an added value with regard to data required for GMP with stacked events since the re-sequencing of the stacked DNA-inserts and their flanking regions is required to assess their integrity (Implementing Regulation and Table 1, vii). However, GMP with stacked events usually have been produced by conventional crossing of parental GMP with one or more single events, for which the insert DNA has been sequenced, and do not involve additional genetic transformation. Moreover, safety assessment evaluations so far, based on SB analysis, did not reveal the concern that insert integrity would be more compromised during conventional crossing of two GM lines compared to crossing of a GM line with a conventional line (Kok et al., 2014; Waigmann, Gomes, Lanzoni, & Perry, 2013).

6. Conclusions and recommendations

Despite practical and economic reasons and the rapid evolution in NGS technologies, it is difficult to predict at which pace NGS-generated data will be presented in regulatory dossiers for RA. Contrary to SB where for each GM crop characterisation, specific probes (sometime radioactively labelled) need to be developed, NGS is a universal, high-throughput platform for generating data. The more GMP are characterised the higher the benefit of its streamlined procedures and its relative cost-effectiveness will become.

It remains an open question to which extent RA offices should repeat the data analysis performed by the applicants to verify the accurateness, reproducibility and quality of the data presented in the regulatory dossier. Should this be necessary, it is also not clear whether risk evaluators should repeat the data analysis only by using algorithms proposed by the applicant or by using algorithms which have been identified as appropriate and of sufficient quality by RA offices themselves. The latter would imply RA offices to have their own trained bioinformaticians and implementing specialized data analysis platforms.

Defining best practices and standardization of bioinformatics tools could facilitate the set-up of a common workflow to be followed by applicants for presenting NGS data for the molecular characterisation of GMP. This would allow RA offices to assess the presentation of data according to a streamlined approach and, if necessary, to focus efforts on developing expertise in certain anticipated data algorithms. The set-up of such a workflow would most likely include issues such as sample/library preparation, NGS platform to be used, required read length to generate, coverage to be used and to be demonstrated to determine the detection limit,

the choice of reference genome, criteria to define artefacts and the use of appropriate data analysis algorithms. Moreover, future possible initiatives for defining such workflow should also take into account the expertise that is currently developed with research projects exploring the feasibility of using NGS for the detection and identification of GMP.

While a common workflow could assist applicants willing to integrate NGS-generated data in their applications, it should be kept in mind that the implementation of NGS technologies are currently not affordable to all potential applicants willing to release a product into the environment. Despite vast decreases in recent years, costs remain substantial. Therefore, risk assessors should pay attention not to request data just because NGS may offer the technical possibilities to do so. Instead, it will be important to focus on (NGS) data that actually fuel the RA.

We identified two developments in plant breeding that could further pave the way for implementing NGS technologies in the preparation and presentation of data for regulatory approval. Firstly, for novel plant breeding products resulting in small modifications, NGS-data generated by the whole genome approach may soon reveal more precise information with respect to the presence of small changes at the nucleic acid level. Secondly, for stacked events, the requirement to re-sequence inserts and flanking regions may also favour high-throughput technologies that can generate data with relative increasing cost-effectiveness and feasibility.

We conclude that although NGS might present clear advantages for product developers as a method at least as well performing as SB, while being faster and more efficient, NGS currently does not always offer a significant added value with respect to the RA of GMPs. Despite some advantages, NGS has currently not the potential to rule out SB and Sanger sequencing for the molecular characterisation of GMP. However, the approaches used so far may soon be further challenged by the fast evolution in NGS technologies and also by the recent developments in molecular breeding of plants, e.g. increasing number of stacked events, expanding variety of transformed species and new traits, and new techniques for genetic modification of crop plants. We therefore plead for a set-up of criteria for the generation of relevant and interpretable NGS data for RA to ensure that data are generated, presented and analysed in a way that facilitates a scientifically sound assessment. This should be done in the first place for GMP, as they constitute the vast majority of GM applications for food and feed uses. However, similar initiatives could be taken for the RA of GM microorganisms and GM animals, as part of the issues discussed in this viewpoint apply to these GMO as well. Recommendations on which and in what way these data should be presented will be beneficial for both applicants in view of the preparation of regulatory dossiers and risk evaluators.

Acknowledgements

The authors thank Fanny Collard (Scientific Institute of Public Health, Brussels, Belgium) for her scientific assistance to the workshop and all the participants of the workshop for their active and valuable contribution to the discussions.

References

Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., et al. (2013). Comparison of commercially available target enrichment methods for next-generation sequencing. *Journal of Biomolecular Techniques, 24*, 73–86.

Breyer, D., Herman, P., Brandenburger, A., Gheysen, G., Remaut, E., Soumillion, P., et al. (2009). Genetic modification through oligonucleotide-mediated mutagenesis. A GMO regulatory challenge? *Environmental Biosafety Research, 8*, 57–64.

Broeders, S., 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, 2012*, 402418.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics, 43*, 956–963.

Chu, Y., & Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics, 22*, 271–274.

Devos, Y., Aguilera, J., Diveki, Z., Gomes, A., Liu, Y., Paoletti, C., et al. (2014). EFSA's scientific activities and achievements on the risk assessment of genetically modified organisms (GMOs) during its first decade of existence: looking back and ahead. *Transgenic Research, 23*, 1–25.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics, 30*, 418–426.

EFSA Panel on Genetically Modified Organisms (GMO). (2011). Scientific opinion: guidance for risk assessment of food and feed from genetically modified plants (Rep. No. 9). *EFSA Journal, 9*, 2150.

European Commission. (2013). Commission implementing regulation (EU) No 503/2013 on applications for authorisation of genetically modified food and feed in accordance with regulation (EC) No 1829/2003 of the European Parliament and of the council and amending commission regulations (EC) No 641/2004 and (EC) No 1981/2006. *Official Journal of the European Union, L 157*, 1–48.

European Parliament. (2003a). Commission regulation (EC) No 1829/2003 of the European Parliament and of the council of 22 September 2003 on genetically modified food and feed. *Official Journal of the European Union, L 268*, 1–23.

European Parliament. (2003b). 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. *Official Journal of the European Union, L 268*, 24–28.

Food and Agricultural Organization of the United Nations (FAO). (2003). *Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants* (Rep. No. CAC/GL 45–2003).

Food and Agricultural Organization of the United Nations (FAO). (2008). *Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants* (Rep. No. CAC/GL 68–2008).

Guo, C., Li, L., Wang, X., & Liang, C. (2015). Alterations in siRNA and miRNA expression profiles detected by deep dequencing of transgenic rice with siRNA-mediated viral resistance. *PLoS One, 10*, e0116175.

Hartung, F., & Schiemann, J. (2014). Precise plant breeding using new genome editing techniques: opportunities, safety and regulation in the EU. *The Plant Journal, 78*, 742–752.

Heap, B. (2013). Europe should rethink its stance on GM crops. *Nature, 498*, 409.

Heck, G. R., Armstrong, C. L., Astwood, J. D., Behr, C. F., Bookout, J. T., Brown, S. M., et al. (2005). Development and characterization of a CP4 EPSPS-based glyphosate-tolerant corn event. *Crop Science, 44*, 329–339.

Holst-Jensen, A. (2009). Testing for genetically modified organisms (GMOs): past, present and future perspectives. *Biotechnology Advances, 27*, 1071–1082.

Hunst, P. L. (2013). *AEIC Spring meeting 2013 minutes. April 17-18, 2013*. Madison, WI. Available at: www.aeicbiotech.org/meetings/SpringMeetingMinutes2013.pdf.

Imelfort, M., & Edwards, D. (2009). De novo sequencing of plant genomes using second-generation technologies. *Briefings in Bioinformatics, 10*, 609–618.

Kok, E. J., Pedersen, J., Onori, R., Sowa, S., Schauzu, M., De, S. A., et al. (2014). Plants with stacked genetically modified events: to assess or not to assess? *Trends in Biotechnology, 32*, 70–73.

Kovalic, D., Garnaat, C., Guo, L., Yan, Y., Groat, J., Silvanovich, A., et al. (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.

Lepage, E., Zampini, E., Boyle, B., & Brisson, N. (2013). Time- and cost-efficient identification of T-DNA insertion sites through targeted genomic sequencing. *PLoS One, 8*, e70912.

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.

McGinn, S., & Gut, I. G. (2013). DNA sequencing – spanning the generations. *New Biotechnology, 30*, 366–372.

Nekrutenko, A., & Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics, 13*, 667–672.

Organisation for Economic Cooperation and Development (OECD). (2010). *Consensus document on molecular characterisation of plants derived from modern biotechnology Paris, France*.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., et al. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science, 327*, 92–94.

Pauwels, K., Podevin, N., Breyer, D., Carroll, D., & Herman, P. (2014). Engineering nucleases for gene targeting: safety and regulatory considerations. *New Biotechnology, 31*, 18–27.

Podevin, N., Devos, Y., Davies, H. V., & Nielsen, K. M. (2012). Transgenic or not? No simple answer! New biotechnology-based plant breeding techniques and the regulatory landscape. *EMBO Reports, 13*, 1057–1061.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*, 341.

Ramon, M., Devos, Y., Lanzoni, A., Liu, Y., Gomes, A., Gennaro, A., et al. (2014). RNAi-based GM plants: food for thought for risk assessors. *Plant Biotechnology Journal, 12*, 1271–1273.

Schnell, J., Steele, M., Bean, J., Neuspiel, M., Girard, C., Dormann, N., et al. (2015). A comparative analysis of insertional effects in genetically engineered plants: considerations for pre-market assessments. *Transgenic Research, 24*, 1–17.

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.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology, 98*, 503–517.

Thudi, M., Li, Y., Jackson, S. A., May, G. D., & Varshney, R. K. (2012). Current state-of-art of sequencing technologies for plant genomics research. *Briefings in Functional Genomics, 11*, 3–11.

Vaughn, J. N., & Bennetzen, J. L. (2014). Natural insertions in rice commonly form tandem duplications indicative of patch-mediated double-strand break induction and repair. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 6684–6689.

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.

Waigmann, E., Gomes, A., Lanzoni, A., & Perry, J. N. (2013). Editorial: new commission implementing regulation on risk assessment of GM plant applications: novel elements and challenges. *EFSA Journal, 11*, e11121.

Weber, N., Halpin, C., Hannah, L. C., Jez, J. M., Kough, J., & Parrott, W. (2012). Editor's choice: crop genome plasticity and its relevance to food and feed safety of genetically engineered breeding stacks. *Plant Physiology, 160*, 1842–1853.

Wei, N., Bemmels, J. B., & Dick, C. W. (2014). The effects of read length, quality and quantity on microsatellite discovery and primer development: from illumina to PacBio. *Molecular Ecology Resources, 14*, 953–965.

Yang, L., Wang, C., Holst-Jensen, A., Morisset, D., Lin, Y., & Zhan, D. (2013). Characterization of GM events by insert knowledge adapted re-sequencing approaches. *Scientific Reports, 3*, 2839–2847.