Food Anal. Methods (2013) 6:1718–1727
DOI 10.1007/s12161-013-9673-x

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

Daniela Wahler · Leif Schauser · Joachim Bendiek · Lutz Grohmann

Received: 1 February 2013 / Accepted: 1 July 2013 / Published online: 28 July 2013

Abstract Precise molecular characterisation of genetic modifications integrated into the genomes of genetically modified organisms (GMOs) and of their flanking genomic regions forms a key component for the development of event-specific detection methods. In the EU, this information is of particular importance for risk management in cases where genetic modifications of unauthorised GM food, feed or seeds are detected. PCR-based chromosome walking approaches are commonly used for DNA sequence determination of the genetic modifications and of the flanking genomic regions in yet undescribed GM plants. If the plant contains complex and re-arranged modifications, sequencing and molecular characterisation are often difficult and laborious. Next-generation sequencing (NGS) of DNA is a powerful alternative tool to rapidly generate primary sequence data on the genome of so far uncharacterised sample material if pure GMO material is available. Recently, robust NGS platforms and affordable sequencing services are accessible for food and feed control laboratories. We here present a NGS-based study for whole-genome sequencing of the GM rice event LLRice62 as a proof-of-principle experiment to develop bioinformatics easy-to-use data analysis tools for rapid molecular characterisation. A total of 171,657,155 read mate pairs of approximately 75 bp each were obtained. Sequence reads belonging to the genetic modifications and their flanking genomic regions in LLRice62 were identified by bioinformatic comparison to the corresponding *Oryza sativa* ssp. *japonica* reference genome sequence using the Illumina InDel caller software and subsequent iterative mapping of retrieved NGS reads. An entire genetic modification of 1,493 bp in the genome of the LLRice62 sample material was determined and correctly mapped on chromosome 6. The determined nucleotide sequence coincides to the genetic modification described by the developer of this rice event. This study demonstrates for the first time the applicability of NGS for molecular characterisation of uncharacterised GMOs.

Keywords GMO · Molecular characterisation · Next-generation sequencing · Re-sequencing · Genetically modified · Detection · Rice

Introduction

Agricultural products derived from genetically modified organisms (GMOs) are increasingly entering the food and feed supply chain. Moreover, a growing number of diverse agricultural relevant traits, genes and genetic regulatory elements are inserted into crop genomes by genetic engineering. Contradictorily, for less than 30 % of all worldwide known genetically modified (GM) plant events, validated methods for event-specific identification are described (Gürtler & Meissner 2011).

To ensure food and feed safety and to take into account the consumers' demand for informed choice on food-related aspects, in many countries regulations have been established. However, GMOs cultivated in different countries may not be authorised for commercial use in other countries, a fact known as asynchronous authorisation. Within the European Union (EU), GMOs and GMO-derived food and feed products

D. Wahler · J. Bendiek · L. Grohmann (✉)
Federal Office of Consumer Protection and Food Safety,
Mauerstr. 39-42, 10117 Berlin, Germany
e-mail: lutz.grohmann@bvl.bund.de

L. Schauser
Interdisciplinary Nanoscience Center (iNANO), Aarhus University,
Gustav Wieds Vej 14, 8000 Aarhus C, Denmark

need to undergo an authorisation process and have to be approved for placing on the market according to Regulation (EC) No. 1829/2003 and considering Directive (EC) No. 2001/18. Unauthorised GMOs and food and feed products derived thereof are not allowed on the EU market. Under these circumstances, the food and feed industry and breeders as well as inspection and control services are challenged and need tools to successfully comply with their task in terms of GMO detection.

Generally, PCR-based GMO detection methods are used, which rely on the prior knowledge of the nucleotide sequence of the genetic elements integrated into the GMO genome (Waiblinger et al. 2010; Spalinskas et al. 2013). Therefore, GMOs with genetic modifications insufficiently characterised at the nucleotide level pose a particular challenge to detection. Here, strategies for the molecular characterisation and identification need to be established. In the past, direct sequencing as well as sequencing by chromosome walking strategies has been described for the molecular characterisation of GMOs (Spalinskas et al. 2013; Babekova et al. 2009; Milcamps et al. 2009). However, these strategies are very laborious until the complete sequence of the genetic modification of an unassigned GMO is determined and at least short stretches of the integrated nucleotide sequence need to be known. Occasionally, the precise sequence of the inserted genetic elements cannot be determined. In the maize event Bt10, for example, a very complex structure with rearrangements, inversions and multiple copies of inserted elements was observed (Milcamps et al. 2009).

Next-generation sequencing (NGS) is a high-throughput method for sequencing of whole genomes. Although currently still too expensive for routine analyses the fast development of NGS promises massive cost reduction. As a consequence, NGS may be expected to become more and more affordable and thus attractive also to food and feed analysts. During NGS, a genomic DNA sample is sheared into a library of small fragments that are sequenced in millions of parallel reactions. The identified strings of bases, called reads, are then either aligned and compared to a reference genome (re-sequencing) or, in the absence of a reference genome, are de novo assembled. The full set of aligned or assembled reads reveals the entire sequence of the genome of the DNA sample. NGS has already been successfully applied to analyse the genomes of plants, e.g. the genomes of rice, *Brachypodium distachyon* and various other crops (The International Brachypodium Initiative 2010; Arai-Kichise et al. 2011). Moreover, the genetic diversity in different ecotypes of *Arabidopsis thaliana* was studied by NGS to identify common polymorphisms as well as small and large insertions and deletions (InDels) across the entire native range of this species by re-sequencing (Cao et al. 2011). Recently, the successful application of NGS for analysis of a genetically modified soybean event with known nucleotide sequence of the insert has been reported by a biotech company to achieve complete molecular characterisation of this GM crop plant (Kovalic et al. 2012).

In the present pilot study, the ability of NGS and of subsequent bioinformatics analysis to identify and characterise genetic modifications in a DNA sample by comparison to a corresponding untransformed reference genome has been tested experimentally on the GM rice event LLRice62. In contrast to the study of Kovalic et al. (2012), here the GMO identity was unknown to the analysing laboratory. To our knowledge, this is the first time that the NGS technique is applied as a GMO analysis strategy to identify and characterise at the molecular level a GM plant event unauthorised in the EU with an a priori unknown nucleotide sequence of the integrated genetic modification to the analysing laboratory.

Material and Methods

Procedure

For sequencing and bioinformatics analysis, a service company selected via public tendering was commissioned. Two micrograms of genetically modified rice genomic DNA was sent to the company in a sealed vial with encrypted labelling. No information was given to the company with respect to the genetic modification or to the GMO event. The order comprised the nucleotide sequencing of the genomic DNA sample using NGS, mapping of the sequence to the *Oryza sativa* reference genome sequence publicly available at the Ensembl server (http://plants.ensembl.org/Oryza_sativa), identification of DNA sequences that could not be mapped, identification of putative transgenic DNA inserts and of the flanking regions in the *O. sativa* genome.

Sample

Genomic DNA of LLRice62, a glufosinate-tolerant rice event developed by Bayer Crop Science (unique identifier: ACS-OS002-5), was used as sample material. Certified reference material of LLRice62 (AOCS No. 0306-I3) was purchased from AOCS, Urbana, USA. The sample contained purified genomic DNA from leaf tissue. Taxonomically, the event LLRice62 belongs to *O. sativa* ssp. *japonica*, cultivar Bengal.

Reference Genome Data

As a reference genome sequence, data of the *O. sativa* ssp. *japonica* MSU6 (cultivar Nipponbare) obtained from the

Ensembl server (http://plants.ensembl.org/Oryza_sativa) were used.

Whole-Genome Sequencing

LLRice62 genomic DNA (1.2 μg) was fragmented into a target size distribution of 300 to 400 bp. This was achieved by ultrasound shearing (Covaris S2 instrument) and gel extraction following Illumina TrueSeq preparation protocols (Illumina Inc., San Diego, USA). The fragments were sequenced on a HiSeq GAII (Illumina Inc., San Diego, USA). One lane of HiSeq GAII (Illumina Inc., San Diego, USA) was loaded with fragmented sample DNA and used for paired-end sequencing (100 bases read length).

Determination of Transgenic Insertions

The Illumina Genome Analysis System (Genome analyzer and CASAVA 1.8.2) software was used for base calling, mapping of the reads to the reference genome and variant calling. The Elandv2e alignment algorithm of CASAVA was configured to report single nucleotide polymorphisms (SNPs), InDels and insertion sequences at breakpoints. A custom Python script was deployed to parse the CASAVA InDel call report in order to extract the nucleotide sequences of the 5′- and 3′-end of insertions at breakpoints, so-called breakpoint border sequences (Fig. 1). These sequences were searched against the DNA sequence of pCAMBIA-1300 (NCBI accession AF234296) using MegaBLAST at default

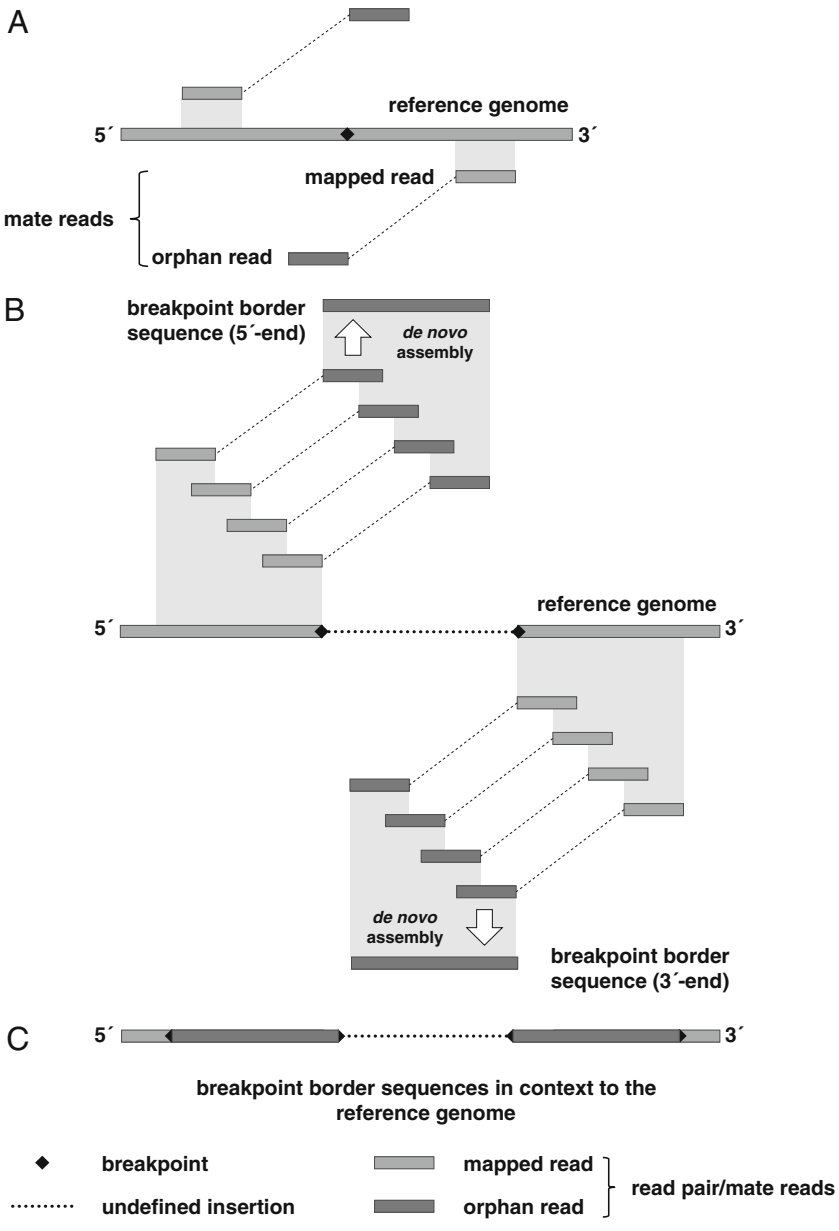

Fig. 1 Schematic illustration of breakpoint border sequence determination. **a** Breakpoints are localized on the reference genome where reads cannot be assembled and the bioinformatic analysis identifies a gap at a certain locus. Mate sequence reads corresponding to the mapped reads remain unmapped to the reference genome and are assigned as orphan reads. **b** Read pairs that have a mate read mapping to the reference genome in close proximity to the breakpoint and an unmapped read (here: orphan read) are determined. All orphan reads are de novo assembled to a breakpoint border sequences, i.e. the 5′- and 3′-sequence of the insertion at the breakpoint. **c** Breakpoint border sequences represent the 5′- and 3′-end of the insertion at a specific breakpoint

parameter settings with pCAMBIA-1300 as the query and the breakpoint border sequences as the database (Hajdukiewicz et al. 1994; Zhang et al. 2000). Breakpoint border sequences exhibiting homology to pCAMBIA-1300 were merged head-to-head to the corresponding breakpoint within the reference genome to form a fusion sequence. All reads were then mapped to this fusion sequence to identify novel breakpoint border sequences. This bioinformatics "read walking" procedure was iterated until no new breakpoint border sequences were identified, thereby identifying the complete nucleotide sequence of the insertion at the specific breakpoint.

Sequence Comparison of NGS Data to Amplicons for Rice Taxon-Specific Detection

In order to demonstrate that the rice taxon-specific detection using established PCR systems (Bonfini et al. 2012) is also possible when working with Bengal rice sample, we used the genes gos9 (NCBI accession X51909), phosholipase D (PLD, NCBI accession AB00191) and sucrose phosphate synthase (SPS, NCBI accession U33175) as reference sequences for mapping the Illumina reads of the whole genome sequencing effort using the Burrows-Wheeler Aligner program (Li & Durbin 2009; Wang et al. 2010). Consensus sequences for the specified sequence stretches of mapped Illumina reads were created by using SAMtools (Li et al. 2009), and pairwise nucleotide sequence alignments were performed using the EMBOSS needle tool (Rice et al. 2000).

Results

A whole-genome sequencing approach using pure genomic DNA of the rice event LLRice62 was performed. The ability of our NGS concept to identify and determine the nucleotide sequence of the genetic modification and the 5′- and 3′-flanking genomic regions of insertion sites in the genome of *O. sativa* ssp. *japonica* (cultivar Bengal) without prior knowledge of the nucleotide sequence of the transgene was challenged. As a general sequencing strategy, the approach of genome re-sequencing was employed.

Quality Control of Sequencing Depth by Mapping to a Reference Genome

Genomic DNA of LLRice62 (certified reference material AOCS 0306-I3) was sheared into a library of fragments of 300 to 400 bp. The nucleotide sequences of the 5′- and 3′-ends were determined by paired-end sequencing. Overall, as much as 343,314,310 single reads of an average read length of 75 bp comprising 171,657,155 mate pairs were obtained. To identify those reads that belong to a genetic modification and to the flanking genomic regions of insertion sites the obtained reads were mapped against the genome of the rice *O. sativa* ssp. *japonica* (cultivar Nipponbare MSU6). This plant genome is completely sequenced and thus could be used as a reference genome. Comprehensive sequence data to the 12 rice chromosomes, the mitochondrial chromosome, the plastome and a small collection of sequences of unknown origin (i.e. which are not yet physically mapped to the genome) are publicly available as so-called pseudochromosomes (IRGSP 2005). Mapping of reads resulted in an average of 65-fold coverage representing 93 % of the rice reference genome (Table 1). In a next step, it was aimed to identify the reads belonging to transgenic insertions in LLRice62.

Calling Insertions of Genetic Modifications Using the Illumina InDel Caller Software

The Illumina 'Consensus Assessment of Sequence and Variation' (CASAVA) software (Elandv2e) reported all DNA polymorphisms (SNPs, insertions, deletions and breakpoints) between the rice reference genome and the reads obtained for the LLRice62 DNA sample. Insertions are defined as sequences where reads exhibit one or more additional nucleotides compared to the reference genome, deletions are sequences that lack one or more nucleotides compared to the reference genome, and breakpoints occur when insertions are larger than the length of a single read and hence cannot be covered by mapping of reads to the reference genome. In total, 176,405 DNA polymorphisms, comprising 72,512 insertions, 81,323 deletions and 22,570 breakpoints (indicating large insertions), were found and attributed to the different pseudochromosomes (Table 2). To develop a screening strategy for genomic polymorphisms caused by genetic engineering we assumed that (1) such phenotypic active polymorphisms are insertions generally larger than 100 bp, i.e. larger than a single read and (2) that they exhibit homology to genetic elements typically used in plant transformation. Only breakpoints indicating insertions larger than 100 bp were further analysed. The 5′- and 3′-borders of inserted sequences at breakpoints were determined (breakpoint border sequences). CASAVA's Elandv2e does this by collecting orphan reads. Orphan reads are reads that do not map to the reference genome sequence but have a mate read that maps to the reference genome at regions that flank a breakpoint. For each specific breakpoint, orphan reads were collected and de novo assembled to breakpoint border sequences (Fig. 1). To distinguish between natural insertions and insertions caused by transformation of the *O. sativa* ssp. *japonica* (cultivar Bengal) genome, all breakpoint border sequences larger than 30 bp were compared against the nucleotide sequence of a plant transformation vector (pCAMBIA-1300) in blast searches using the MegaBLAST

Table 1 Summary statistics of the reads mapped to the reference genome

Pseudochromosome	Sites	Known sites	Bases mapped at known sites	Mean depth at known sites/coverage (times)	Fraction of known sites mapped (%)
1.fa	43,268,879	43,262,479	2,921,489,941	67.53	94.49
3.fa	36,406,689	36,398,488	2,622,129,094	72.040	97.23
2.fa	35,930,381	35,926,281	2,486,504,791	69.21	95.47
4.fa	35,278,225	35,267,008	2,068,603,952	58.66	87.82
5.fa	29,894,789	29,886,575	1,983,877,602	66.38	95.48
6.fa	31,246,789	31,226,940	2,018,151,759	64.63	92.36
7.fa	29,696,629	29,694,025	1,933,558,002	65.12	93.45
8.fa	28,439,308	28,437,608	1,836,697,256	64.59	94.16
9.fa	23,011,239	23,005,030	1,548,907,896	67.33	95.68
10.fa	23,134,759	23,121,254	1,361,360,869	58.88	88.50
11.fa	28,512,666	28,506,126	1,648,881,613	57.84	88.29
12.fa	27,497,214	27,494,995	1,771,890,729	64.44	94.64
Sy.fa	526,915	486,515	18,124,942	37.25	66.23
Mt.fa	490,520	490,516	106,997,500	218.13	26.94
Pt.fa	134,525	134,525	3,834,932	28.51	18.09
Total	373,469,527	373,338,365	24,331,010,878	65.17	93.10

Sites length of the pseudochromosomes obtained by mapping of sequence reads, including detected SNPs and InDels, *known sites* length of the reference pseudochromosomes; *bases mapped at known sites* number of read nucleotides mapped to the reference genome; *mean depth at known sites/coverage* calculated average frequency at which the sample genome nucleotides mapped to the reference genome were sequenced; *fraction of known sites mapped* completeness of the sequence for each pseudochromosome (1.fa to 12.fa), the mitochondrial chromosome (Mt.fa), the plastome (Pt.fa) and for not specified sequences (Sy.fa; not physically mapped to the genome)

algorithm (Zhang et al. 2000). Although pCAMBIA-1300 has not been used as the transformation vector to generate GM rice LLRice62 it contains nucleotide sequences of the pUC18 multiple cloning site, the aadA-gene, the cauliflower mosaic virus (CaMV) 3′UTR, the hptII-gene and the 35S promoter from CaMV which are commonly used in plant transformation. The aim was to develop a general approach for a multitude of GM crops to enable identification of likely transgenic breakpoint border sequences. Using this approach, two breakpoint border sequences belonging to the same breakpoint were identified that have sequence identity to the pCAMBIA-1300 sequence. At the left breakpoint border, a sequence of 50 bp and at the right breakpoint border a sequence of 116 bp showed 100 % sequence identity to the pCAMBIA-1300 sequence (Fig. 2a, b). The left breakpoint border sequence is 79 bp, while the right breakpoint border is 275 bp long (Fig. 2c, d). This specific breakpoint is located on chromosome 6, position 25064304 (left breakpoint) and 25064317 (right breakpoint) in the rice reference genome. The proximity of the identified breakpoint positions indicates that a short deletion occurred during plant transformation.

Table 2 InDel category summary for the variant caller of CASAVA software Elandv2e on the LLRice62 sample sequence data

Pseudochromosome	Insertions	Deletions	Breakpoints
1.fa	11,017	12,389	2,944
3.fa	4,619	4,978	1,249
2.fa	6,745	7,458	1,869
4.fa	9,405	10,978	3,053
6.fa	4,930	5,641	1,663
5.fa	3,922	4,565	1,039
7.fa	4,934	5,626	1,674
11.fa	9,003	9,760	3,391
8.fa	4,383	4,898	1,266
12.fa	3,727	4,207	1,134
10.fa	7,877	8,507	2,683
9.fa	1,746	2,113	529
Sy.fa	194	198	73
Mt.fa	7	4	2
Pt.fa	3	1	1
Total	72,512	81,323	22,570

The numbers of identified insertions, deletions and breakpoints relative to the reference genome are shown for each pseudochromosome

Retrieving the Entire Nucleotide Sequence of the Insertion

In order to determine the complete nucleotide sequence of the genetic modification in the rice reference genome at the breakpoint identified on chromosome 6, 200 bp of the reference genome flanking the breakpoint up- and downstream were merged with the identified breakpoint border sequences. This composed sequence was used as a new target

A

```
Score = 93.5 bits (50),  Expect = 4e-23
Identities = 50/50 (100%), Gaps = 0/50 (0%)
Strand=Plus/Minus

AF234296       50 GCATGCCTGCAGGTCGACTCTAGAGGATCCCCGGGTACCGAGCTCGAATT 1
                  ||||||||||||||||||||||||||||||||||||||||||||||||||
lseq-25064304   1 GCATGCCTGCAGGTCGACTCTAGAGGATCCCCGGGTACCGAGCTCGAATT 50
```

B

```
Score =  215 bits (116),  Expect = 4e-59
Identities = 116/116 (100%), Gaps = 0/116 (0%)
Strand=Plus/Plus

AF234296        59 CACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCC 98
                   ||||||||||||||||||||||||||||||||||||||||
rseq-25064317  160 CACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCC 199

AF234296        99 TGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCT 138
                   ||||||||||||||||||||||||||||||||||||||||
rseq-25064317  200 TGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCT 239

AF234296       139 TTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACC 174
                   ||||||||||||||||||||||||||||||||||
rseq-25064317  240 TTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACC 275
```

C

```
left border sequence of breakpoint at position 25064304 of the
reference genome, Length = 79 bp, [5´→ 3´]

>lseq-25064304
GCATGCCTGCAGGTCGACTCTAGAGGATCCCCGGGTACCGAGCTCGAATTCGAGCTCGCCCTGGA
TTTTGGTTTTAGGA
```

D

```
right border sequence of breakpoint at position 25064317 of the
reference genome, Length = 275 bp, [5´→ 3´]
>rseq-25064317
TTGATATTTTTGGAGTAGACAAGCGTGTCGTGCTCCACCATGTTGACGAAGATTTTCTTCTTGTCA
TTGAGTCGTAAGAGACTCTGTATGAACTGTTCGCCAGTCTTTACGGCGAGTTCTGTTAGGTCCTCT
ATTTGAATCTTTGACTCCATGGGAATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACC
CTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAG
AGGCCCGCACC
```

Fig. 2 Result of the MegaBLAST search with pCAMBIA-1300 (NCBI accession AF234296) as a query sequence and all de novo assembled breakpoint border sequences. Two hits were found which correspond to a breakpoint at position 25064304 (**a**) and position 25064317 (**b**) on chromosome 6 of the reference genome, respectively. Nucleotide sequence of the de novo assembled left (**c**) and right (**d**) breakpoint border sequences

for a further round of mapping and variant calling to identify a new breakpoint and a new cluster of orphan reads to assemble new breakpoint border sequences. This procedure was iterated seven times until no new breakpoint was identified, and a sequence overlap in the determined breakpoint border sequences allowed identification of the full nucleotide sequence of the insertion at this breakpoint (Fig. 3). The assembled nucleotide sequence comprises 1,493 bp for the complete insertion on chromosome 6 (Fig. 4). The nucleotide sequence of the insertion and its flanking genomic regions were uploaded at GenBank and will be accessible under accession number KF036176.

Verification of the Transgenic Origin of the Nucleotide Sequence of Insertion on Chromosome 6

LLRice62 is a rice event with tolerance to the herbicide ingredient glufosinate ammonium. The genetic modification in LLRice62 consists of a gene cassette of the bar gene under the control of the CaMV 35S promoter and the CaMV 35S terminator (EFSA 2007). The alignment of the NGS-determined nucleotide sequence of the LLRice62 genome to GeneBank database entries shows identity to the genetic elements described in the LLRice62 application document according to Regulation (EC) No. 1829/2003 (EFSA 2007). Sequence identity was found to the flanking genomic regions of rice chromosome 6 (NCBI accession NC_008399), to the lacZ alpha gene and the herein embedded multiple cloning site of cloning vector pUC1918 (NCBI accession U09331), the terminator and promoter sequence of the CaMV 35S RNA gene (NCBI accession NC_001497) and to the bar gene of *Streptomyces hygroscopicus* (NCBI accession X05822; Table 3, Fig. 4) (Franck et al. 1980; Yanisch-Perron et al. 1985; Pietrzak et al. 1986; Schweizer 1993). Short segments of missing sequence identity to any GeneBank database entry were identified only at the junctions of the different genetic

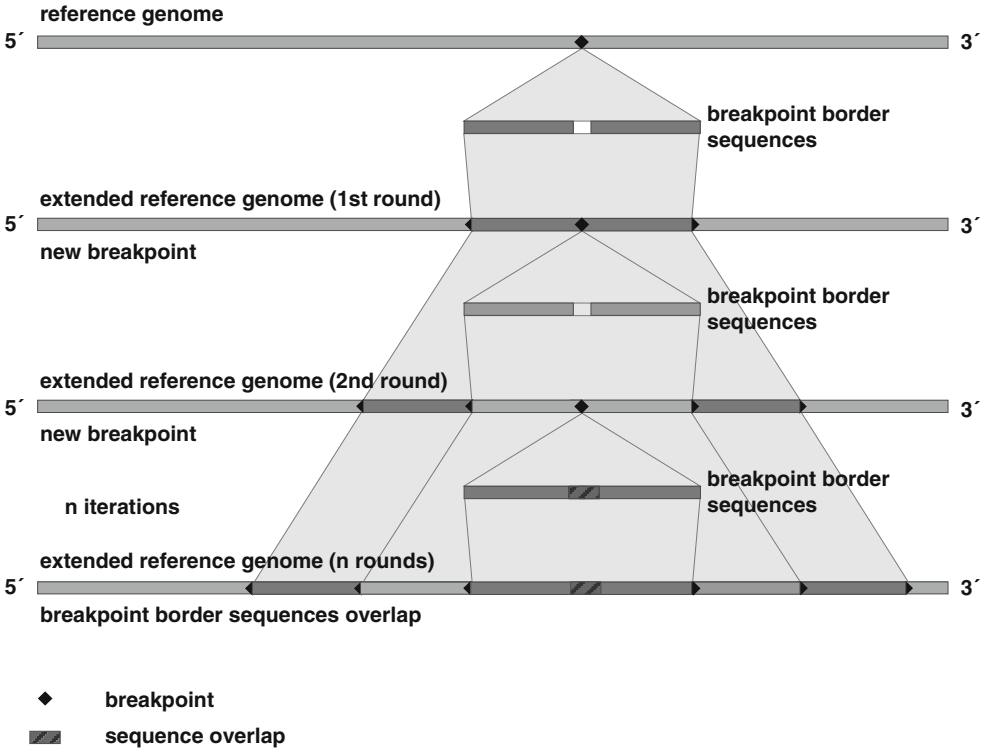

Fig. 3 Schematic illustration of the iterative nucleotide sequence read identification and their assembly at a large insertion at a specific breakpoint. Breakpoint border sequences are assigned to the originally found breakpoint and merged to the reference genome data set to build an extended reference genome. Orphan reads of a mate pair that map to this newly created reference genome at the breakpoint are filtered and are de novo assembled to new breakpoint border sequences. This process is repeated until the filtered breakpoint border sequences overlap the 3′- and 5′-end of breakpoint and map to the newly merged reference genome

elements and represent those defined as polylinker sequences according to the developers' information (EFSA 2007). Furthermore, in alignments, the derived nucleotide sequence is in agreement with nucleotide sequence fragments of LLRice62 recently published (data not shown; (Spalinskas et al. 2013)).

Sequence Comparison of NGS Data to Amplicons for Rice Taxon-Specific Detection

We also used the NGS data to re-sequence the amplicon nucleotide sequences of the gos9 (NCBI accession X51909), the SPS (NCBI accession U33175) and PLD (NCBI accession AB00191) genes which are recommended as European reference methods for PCR-based detection and identification of the rice taxon (Bonfini et al. 2012). These PCR systems are based on consensus sequences derived from the alignment of as much as 58 rice varieties (Wang et al. 2010). The sequences for gos9, SPS and PLD determined by NGS were used to verify the specificity of the respective taxon-specific PCR methods also when working with Bengal rice. SNPs at the primer and probe binding sites would cause reduced sensitivity and later Ct values in quantitative real-time PCR. However, the alignments of the NGS-derived sequence and the predicted amplicon sequence showed no nucleotide polymorphisms in the sequence targeted by the

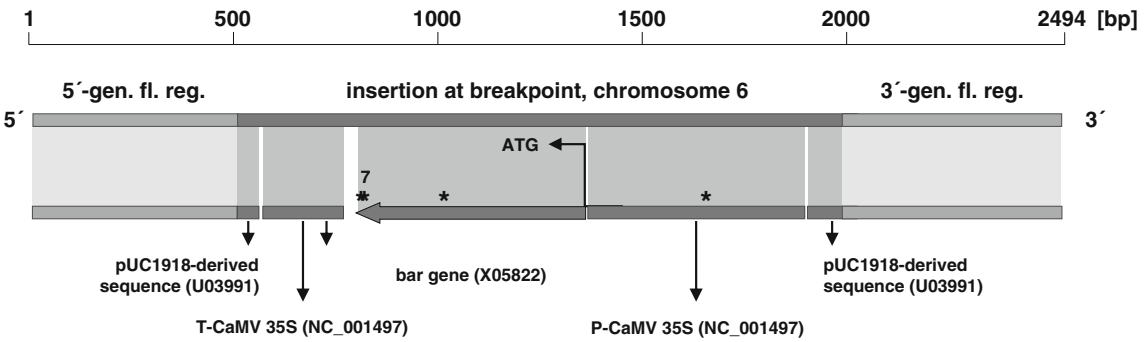

Fig. 4 Schematic alignment of publicly available sequence data and the NGS-based determination of the genetic modification on chromosome 6 in the genome of the LLRice62 sample. Nucleotides 1 to 500 and 1,994 to 2,494 represent the wild-type flanking genomic regions (5′- and 3′-gen. fl. reg.) of rice flanking the transgenic insertion (*light grey*). In *dark grey*, the genetic elements derived from cloning vector pUC1918 (NCBI accession U03991), Cauliflower mosaic virus (NCBI accession NC_001497) and from the bar gene from *S. hygroscopicus* (NCBI accession X05822) are indicated, *asterisks* mark mismatches in the sequence alignment, *digits* indicate number of mismatches, the ATG start codon and direction of the bar gene reading frame are indicated

Table 3 Overview of sequence alignments of NGS-derived sequence data and publicly available sequence data performed with EMBOSS needle algorithm regarding the genetic elements integrated in LLRice62

Annotation	Position on NGS-derived sequence	NCBI accession	Length (bp)	Sequence identity (%)
Rice genome	1–500	NC_008399	500	100
pUC1918	501–552	U03991	53	100
T-CaMV 35S	561–754	NC_001497	194	100
Bar gene	777–1,327	X05822	551	97.7
P-CaMV 35S	1,341–1,874	NC_001497	529	99.8
pUC1918	1,881–1,993	U03991	119	100
Rice genome	1,994–2,494	NC_008399	501	100

For each genetic element (NCBI accession given), the results of the alignment are given in terms of position of the matching sequence on the NGS-derived sequence, the length of the matching sequence and the sequence identity in percentage. Numbering of positions corresponds to those given in Fig. 4

PCR primers and probes confirming the suitability of these taxon-specific PCR methods also for Bengal rice (data not shown).

Discussion

In this study, we applied NGS for the molecular characterisation of a genetically modified rice variety. We aimed to identify the nucleotide sequence of genetic modification integrated into the plants' genome and the corresponding 5′- and 3′-flanking genomic regions without depending on any sequence information on the transgene. This information was collected since it is required for thorough risk management of GMOs.

By mapping of reads to the rice reference genome breakpoints were detected. Subsequent iterative application of the CASAVA software for in silico analysis of all unmapped reads revealed orphan reads corresponding to breakpoints that were assembled and used for full nucleotide sequence determination of large insertions. Using our approach, it was possible to simultaneously identify the integration site in LLRice62 and to determine the complete nucleotide sequence of this insertion. Theoretically, this approach is applicable to all GMOs with a known reference genome data set.

Our experimental design allowed re-sequencing with a high coverage, indicating that the collected reads virtually cover the complete genome of the LLRice62 sample with as little as 1.2 μg of genomic DNA. In conclusion, also reads corresponding to event-specific insertions in the genome of LLRice62 should be generated. To retrieve these nucleotide sequences, several successive steps of bioinformatics analysis were performed.

As a first step, all DNA polymorphisms between the LLRice62 sample genome and the rice reference genome were collected. As many as 176,405 DNA polymorphisms were reported including those which are attributable to insertions caused by genetic engineering. However, the large majority was expected to be naturally occurring genome polymorphisms since similar numbers of natural polymorphisms among rice varieties have been reported previously. Arai-Kichise et al. (2011) discovered 168,288 DNA polymorphisms in *O. sativa* ssp. *japonica* (land race Omachi), whereas 408,898 DNA polymorphisms have been found in *O. sativa* ssp. *indica* (cultivar 93–11) when compared to *O. sativa* ssp. *japonica* (cultivar Nipponbare)—the reference genome used in this study (Feltus et al. 2004). Therefore, the next analysis step required to separate natural polymorphisms from those caused by genetic engineering.

We assumed that genetic modifications which result in a phenotypic change typically span more than 100 bp, which exceeds the sequence data of a NGS read derived from an Illumina device. This assumption might fail to detect rare events of insertions smaller than 100 bp (Monsanto Company 2000), but this selection criterium allowed for the subsequent analysis that such insertions are indicated as breakpoints on the reference genome. The number of DNA polymorphisms putatively caused by genetic engineering was thereby reduced to a total number of 22,570. However, a recent study of multiple accessions of *A. thaliana* genomes demonstrated that even large insertions simply may occur in plants through natural genome variations (Cao et al. 2011). The bioinformatics strategy should therefore focus on selecting breakpoints potentially caused by plant transformation. Such breakpoints most likely do not exhibit homology towards endogenous nucleotide sequences.

All breakpoint border sequences were thus analysed by running MegaBLAST with a commonly used plant transformation vector as a query sequence. In the present proof-of-concept study, we only used the nucleotide sequence of the vector pCAMBIA-1300, but for more rigorous testing, a systematic review of all plant transformation vectors and genetic elements used in plant transformation should be used to extract an optimised set of query sequences.

If such MegaBLAST searches fail to detect potential transgenic origin of breakpoint border sequences, breakpoint border sequences themselves might be used as a query in blasts searches to check for homology to sequences foreign to the sample specie. However, in the case that the transgenic insertion solely consists of endogenous genetic sequences (cisgene), no detection technique is currently available to distinguish between natural insertions and cisgenesis.

Using our strategy, one breakpoint on chromosome 6 was identified in the LLRice62 sample as a putative insertion site for a genetic modification. This is in complete agreement with the information provided by the LLRice62 breeding company (EFSA 2007). Since the integration site in LLRice62 does not lie within sequence repeats of the wild-type background, the exact breakpoints were not difficult to be determined. Otherwise, subsequent PCR analysis is required to clarify the exact position of the integration.

To completely characterise the genetic modification at the identified breakpoint on chromosome 6, we developed a computationally efficient strategy for nucleotide sequence determination of large insertions. It is based on the ability of CASAVAS variant calling. Identified breakpoint border sequences and iterative use of these sequences as targets allowed filtering for adjacent orphan reads and finally the assembly of the full nucleotide sequence of the insertion. Compared to typically used de novo assembling strategies that use all reads that do not map to the reference genome, our protocol has the advantage that the nucleotide sequence of insertions is computed stepwise. At a time, only a manageable number of reads is de novo assembled to rather small sequence fragments for which the exact position within the insertion can be assigned. Hence, the problem of incorrect sequence assembly, as frequently reported for de novo assembly—especially at sequence repeats—is circumvented (Pop 2009; Miller et al. 2010). Furthermore, even the nucleotide sequence of large and complex insertions can potentially be determined by the stepwise assemblies of breakpoint border sequences.

In the next step, the completed nucleotide sequence of the insertion and its flanking genomic regions was aligned to GeneBank database entries that were specified by the developer as transformed into LLRice62 (Fig. 4; (EFSA 2007; Franck et al. 1980; Yanisch-Perron et al. 1985; Pietrzak et al. 1986). The developer reported that a 1,501-bp-long construct from a pUC19-related transformation vector was used for particle bombardment of which a 1,497-bp-long fragment was integrated into the rice genome. An 18-bp-long deletion within the chromosome 6 was reported at the site of integration (Spalinskas et al. 2013; EFSA 2007). The alignments of the publicly available and our NGS-sequence data could not completely confirm these details. By NGS, we determined an insertion sequence of 1,493 bp. However, these discrepancies can be explained. From the NGS-data analysis, we conclude that at the 3′-flanking genomic region a GACT stretch, perfectly matching the genomic sequence of rice chromosome 6, has not been deleted through plant transformation as previously reported (data not shown; Spalinskas et al. 2013). Apart from these discrepancies, the insertion sequence determined by NGS is in complete agreement to the data published by Spalinskas et al. (2013) (see NCBI accessions JX139719, JQ406880 and JQ406881).

Furthermore, we demonstrated that the sequences targeted by the PCR systems recommended as European reference methods for rice taxon-specific detection are conserved in the Bengal rice genome.

Conclusions

Based on the NGS technique, a proof-of-concept experiment was performed to develop a detailed analysis strategy for identification of genetic modifications without prior knowledge of the nucleotide sequence of the insertion. As a result of this pilot study, some limitations of the strategy were identified. To achieve sufficient depth of coverage, NGS-based GMO analysis requires 100 % GMO material. Contrarily to PCR strategies, population samples cannot be handled yet since the current available software assumes that read mapping is done for homozygous or heterozygous variants. Admittedly, the strategy presented here will not achieve deep sequencing data to characterise a GM event present only in trace amounts what is observed in many cases in food and feed products when unauthorised GMO material is present. However, the approach should be applicable to unprocessed food and feed products such as vegetables and fruits or in cases in which 100 % GMO material is available and the DNA can be extracted in sufficient quantity and quality, e.g. seeds. Recently, unauthorised GM papaya fruits imported from Thailand have been identified and notified in the Rapid Alert System for Food and Feed (notification 2012.0451; http://ec.europa.eu/food/food/rapidalert/index_en.htm). It should be possible to characterise the genetic modifications of this 100 % material by the NGS-based analysis strategy described here. Experimental work to characterise this GM papaya is currently in progress.

The NGS approach described here relies on the availability of a reference genome what might be seen as a limitation. However, the constantly increasing number of reference genome data sets available for many crop genomes and the rapid improvement of the NGS technique will allow to apply the approach of re-sequencing for fast identification and complete molecular characterisation of unauthorised GMO material for which such data are not available to food and feed control services as well as competent authorities in emergency cases. Contrarily to methods used at present,

e.g. PCR-based characterisation of the construct or chromosome walking strategies, the NGS-based GMO analysis should enable complete sequencing of single and of multiple and complex insertions present in a DNA sample in one experiment. The approach described here will complement the pool of methods for GMO analysis since it will become more cost-effective and rapid as the NGS technique will further develop.

Conflict of Interest Daniela Wahler has no conflict of interest. Leif Schauser has no conflict of interest. Joachim Bendiek has no conflict of interest. Lutz Grohmann has no conflict of interest. This article does not contain any studies with human or animals subjects.

References

Arai-Kichise Y, Shiwa Y, Nagasaki H et al (2011) Plant Cell Physiol 52:274–282

Babekova R, Funk T, Pecoraro S, Engel KH, Busch U (2009) Eur Food Res Technol 228:707–716

Bonfini L, van den Bulcke MH, Mazzara M, Ben E, Patak A (2012) J AOAC International 95:1713–1719

Cao J, Schneeberger K, Ossowski S (2011) Nat Genet 43(10):956–963

EFSA (2007) The EFSA Journal 588, 1–25, http://www.efsa.europa.eu/de/efsajournal/pub/588.htm. Accessed 22 July 2013

Feltus A, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) Genome Res 14:1812–1819

Franck G, Guille M, Jonard G, Richards K, Hirth L (1980) Cell 21:285–294

Gürtler P, Meissner E, Busch U (2011) Abschlußbericht zum Projekt Nachweis von nicht zugelassenen gentechnisch veränderten Organismen (GVO): Weltweite Ermittlung, Importanalyse und Entwicklung von Nachweis-Methoden; LGL Oberschleißheim

Hajdukiewicz P, Svab Z, Maliga P (1994) Plant Mol Biol 25:989–994

IRGSP (International Rice Genome Sequencing Project) (2005), Nature 436:793–800

Kovalic D, Garnaat C, Guo L (2012) The Plant Genome 5(3):149–163

Li H, Durbin R (2009) Bioinformatics 25:1754–1760

Li H, Handsaker B, Wysoker A et al (2009) Bioinformatics 25:2078–2079

Milcamps A, Rabe S, Cade R, De Framond AJ, Henriksson P, Kramer V, Lisboa D, Pastor-Benito S, Willits MG, Lawrence D, van den Eede G (2009) J Agric Food Chem 57(8):3156–3163

Miller JR, Koren S, Sutton G (2010) Genomics 95(6):315–327

Monsanto Company (2000) Updated molecular characterization and safety assessment of Roundup Ready soybean event 40-3-2, http://cera-gmc.org/docs/decdocs/gts40-3-2-update.pdf. Accessed 22 July 2013

Pietrzak M, Shillito RD, Hohn T, Potrykus I (1986) Nucleic Acid Res 14(14):5857–5868

Pop M (2009) Brief Bioinform 10(4):354–366

Rice P, Longden I, Bleasby A (2000) Trends Genet 16(6):276–277

Schweizer HP (1993) Gene 134:89–91

Spalinskas R, van den Bulcke M, van den Eede G, Milcamps A (2013) Food Anal Methods 6:705–713

The International Brachypodium Initiative (2010) Nature 463:763–768

Waiblinger H, Grohmann L, Mankertz J, Engelbert D, Pietsch K (2010) Anal Bioanal Chem 396:2065–2072

Wang C, Jiang L, Rao J, Liu Y, Yang L, Zhang D, Agric J (2010) Food Chem 58:11543–11547

Yanisch-Perron C, Vieira J, Messing J (1985) Gene 33:103–119

Zhang Z, Schwartz S, Wagner L, Miller W (2000) J Comput Biol 7(1–2):203–214