Chapter 13
Characterization of Plant Genetic Modifications Using Next-Generation Sequencing

Ana Pérez-González, Álvaro Eseverri, and Elena Caro

Abstract In the last few years, many studies have demonstrated that next-generation sequencing (NGS) technologies can facilitate the detection and molecular characterization of genetically modified organisms (GMOs). T-DNA localization and copy number determination in transgenic plants are very useful in basic research projects because of its implications in transgene expression level and stability, and are absolutely necessary for the commercialization of a GMO. The high throughput of NGS together with its continuously decreasing cost makes it a very rapid, cost-effective, and efficient tool for this task, faster and less laborious than the classical Southern blot and genome walking techniques. Moreover, the recent development of bioinformatics tools designed for users with no specific knowledge of computer science makes this approach affordable to the whole scientific community. Successful wet lab strategies and bioinformatics pipelines reported in the literature will be reviewed and discussed here.

Keywords NGS · GMO · Synthetic biology · WGR · Bioinformatics tools

Plant synthetic biology uses engineering principles to genetically modify plants creating or enhancing beneficial traits or to produce valuable products. With this purpose, genetic modules are combined in different ways to build new modules that exhibit predictable behaviors. These genetic modules are then introduced into a recipient organism's native genome via stable integration of T-DNAs into a plant genome (Fig. 13.1).

Upon *Agrobacterium*-mediated transformation, single or tandem T-DNA copies are usually integrated into one or two loci of the plant genome, sometimes with

Ana Pérez-González and Álvaro Eseverri. These authors contributed equally to the work.

A. Pérez-González · Á. Eseverri · E. Caro (✉)
Centre for Plant Biotechnology and Genomics, Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Campus Montegancedo UPM, Madrid, Spain
e-mail: elena.caro@upm.es

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_13

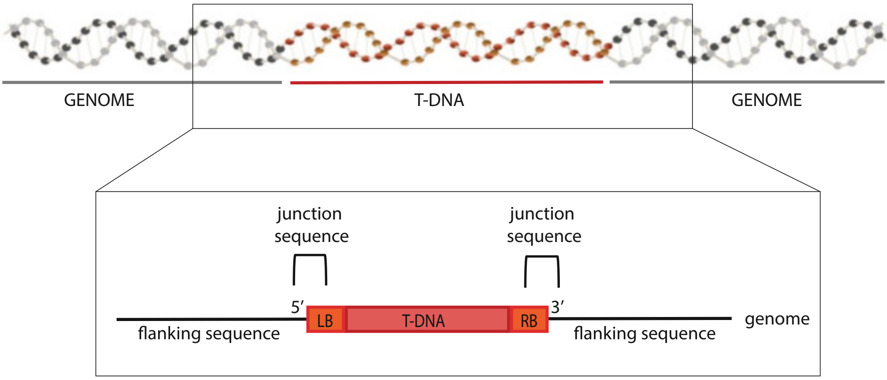

Fig. 13.1 Representation of the stable integration of a T-DNA into a plant genome. Once integrated, junction sequences are created between the T-DNA boundaries and the site of the host genome where it has inserted. Flanking sequences are the original host genome sequences that surround the inserted T-DNA. *LB* left border, *RB* right border

various rearrangements of the target site: T-DNAs are truncated at their left border at low frequency, and in some cases, vector backbone DNA is integrated too. In contrast, direct DNA transfer (like particle bombardment) renders high-copy transgenic loci and extensive rearrangements of the foreign DNA (Pérez-González and Caro 2016).

The existence of repeat-sensitive transcriptional repression mechanisms, described long ago in plants and animals, establishes that single gene copies at a defined locus are expressed much more effectively than reiterated transgenes (Wolffe 1997). Thus, the details about the insertion locus, the number of copies of the inserted T-DNA, and its structure are relevant since they can play an important role in determining if transgenes are correctly expressed or become silenced.

Also, detailed descriptions of the genetic modifications are required for risk evaluation prior to commercialization and release of GMOs into the environment. Legislation asks specifically that an applicant must provide information on the size and copy number of inserts and their organization and sequence information for both 5′ and 3′ flanking regions, with the aim of identifying interruptions of known host genes (Guidance for risk assessment of food and feed from genetically modified plants 2011).

In the past, this characterization was carried out using "classical" molecular biology techniques. The traditional way to identify foreign DNA integration was through Southern blot analysis, using a sequence-specific probe homologous to the transgene, which is very time-consuming although efficient in confirming the T-DNA copy number. For the identification of the insertion site, many PCR-based methods have been successfully used, such as thermal asymmetric interlaced PCR (TAIL-PCR) (Liu et al. 1995), adapter-ligated PCR (O'Malley and Ecker 2010), inverse PCR (IPCR) (Ochman et al. 1988), and restriction site extension PCR (RSE-PCR) (Ji and Braam 2010), but these methods present important limitations.

They all rely on having accurate sequence information of the integrated T-DNA, and sometimes huge rearrangements can take place, leading to very complex situations where the results are difficult to interpret. Other limitations of these PCR-based methods are the need of restriction enzymes that cut both the T-DNA and the genome at a convenient distance and the generation of non-specific products through PCR (Ji and Braam 2010). In any case, these approaches are always laborious and expensive and especially difficult to use in high throughput.

Whole genome re-sequencing (WGR) offers a good alternative to these conventional techniques with great effectiveness and a rapidly decreasing cost. Several publications have reported its use to detect the exact copy number, structure, and genomic location of transgenes in different species of plants, apart from detecting the presence of vector backbone and assessing the stability of T-DNAs across generations.

13.1 Workflow

In the last decades, the apparition of next-generation sequencing techniques has had a great impact on research. It has increased the throughput and speed of sequencing and allowed a decrease on the cost of the process, leading to its application in many fields. Commercial second-generation high-throughput sequencing platforms (Roche/454, Illumina HiSeq or MiSeq, ABI SOLiD, etc.), independently of the method they use for sequencing, all follow a somehow similar workflow, as described below (Kulski 2016) (Fig. 13.2).

13.1.1 Library Preparation

A library needs to be prepared by randomly breaking the DNA into small fragments and adding common adapters to their ends. Later, the template DNA will be primed by an oligonucleotide complementary to the adapter sequence.

13.1.2 Clonal Amplification

In order to achieve a detectable signal for sequencing, the DNA fragments from the library need to be clonally amplified. This can be done on a solid surface or on beads while isolated within miniature emulsion droplets or arrays. In any case, after several rounds of amplification, DNA clusters are generated, and sequencing can proceed.

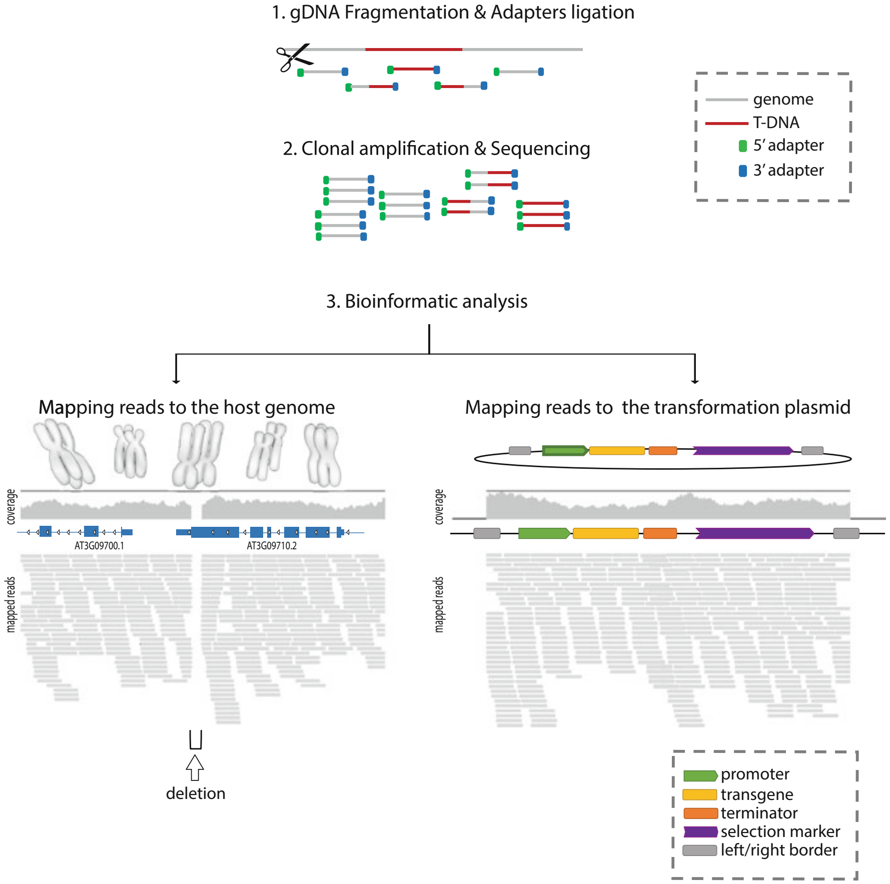
1. gDNA Fragmentation & Adapters ligation
2. Clonal amplification & Sequencing
genome
T-DNA
5′ adapter
3′ adapter
3. Bioinformatic analysis
Mapping reads to the host genome
Mapping reads to the transformation plasmid
coverage
AT3G09700.1
AT3G09710.2
mapped reads
deletion
coverage
mapped reads
promoter
transgene
terminator
selection marker
left/right border

Fig. 13.2 General workflow of a whole genome re-sequencing experiment. (1) Genomic DNA (gDNA) is broken into small fragments, and adapters are added to 5′- and 3′-ends. (2) Small DNA fragments are clonally amplified. (3) Examples of sequencing reads mapping to the host genome and to the transformation plasmid. Deletion refers to a piece of the host genome missing after the insertion of the T-DNA

13.1.3 *Cyclic Array Sequencing*

Sequencing with wash-and-scan techniques is done following different methods. In *sequencing-by-synthesis* approaches, a complementary strand is synthesized in the presence of a polymerase enzyme. In the case of pyrosequencing, the release of pyrophosphate is detected when nucleotides are added to the DNA chain. This is the method applied on the Roche/454 platform, the first to revolutionize sequencing technology.

Another type of sequencing-by-synthesis approach is the case of cyclic reversible termination, in which during each cycle, a fragment of DNA template combined with an adapter incorporates just one nucleotide, since the blocked 3′ group prevents additional incorporations. This is the method applied on the Illumina platform, currently the most cost-effective and most widely used sequencer.

In *sequencing by ligation*, the polymerization reaction is replaced by a ligation reaction. This is the method applied on ABI SOLiD sequencers, which render high accuracy.

A new generation of sequencing methods has been developed around the single-molecule sequencing technology. This so-called third-generation sequencing can be done without creating the time-consuming and costly amplification libraries and, thus, eliminating the errors caused by PCR. The read length of third-generation sequencing methods can reach several kilobases, thereby allowing the resolution of highly complex genomes with many long repetitive elements, copy number, and structural variations. The most popular of the third-generation sequencing platforms is the single-molecule real-time (SMRT) sequencing method used by Pacific Biosciences (PacBio).

13.1.4 Bioinformatics Analysis

Once WGR is performed, bioinformatics tools are used to identify differences between the DNA of the specific individual sequenced (transgenic organism) and that of a reference genome (wild type) (Fig. 13.3).

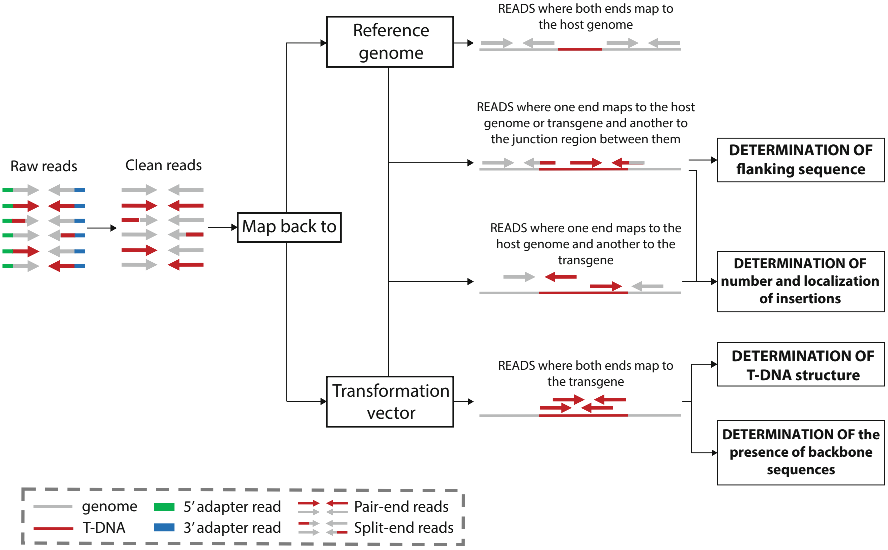
Raw reads
Clean reads
Map back to
Reference genome
Transformation vector
READS where both ends map to the host genome
READS where one end maps to the host genome or transgene and another to the junction region between them
READS where one end maps to the host genome and another to the transgene
READS where both ends map to the transgene
DETERMINATION OF flanking sequence
DETERMINATION OF number and localization of insertions
DETERMINATION OF T-DNA structure
DETERMINATION OF the presence of backbone sequences
genome
T-DNA
5′adapter read
3′adapter read
Pair-end reads
Split-end reads

Fig. 13.3 Detailed bioinformatics pipeline for the analysis of sequencing reads in a whole genome re-sequencing experiment

T-DNA structure and rearrangements can be determined by the calculation of insert size distributions using the reads that map against the transformation plasmid sequence (Park et al. 2017).

If some of these reads map to the vector used for transformation, the presence of *vector backbone-derived sequences* in the genome can also be determined (its absence must be confirmed for safety assessment of GM crops).

And most importantly, the *determination of the T-DNA insertion site and genomic flanking sequences* can be done by analyzing the reads that match the host genome only partially and also map to the T-DNA sequence (Holst-Jensen et al. 2016).

13.2 Reported Strategies

While this is the general workflow of a WGR experiment, several specific approaches have been described to map and characterize T-DNA insertions.

The first report dates to 2012, when Polko et al. (2012) conducted a forward genetic screen to identify molecular components involved in controlling hyponastic growth in *Arabidopsis thaliana*. To identify T-DNA insertion loci in selected candidates, they first tried plasmid rescue and TAIL-PCR, but both methods repeatedly failed. Therefore, they adopted a novel approach using pooled DNA and an Illumina Genome Analyzer to produce over two million of 50-bp paired reads. The reads were mapped to the *Arabidopsis* genome, the T-DNA sequence, or both. The last ones were the ones used to detect the T-DNA insertion loci, showing for the first time an effective method to map T-DNA insertions in *Arabidopsis thaliana* without sequencing entire genomes.

On that same year, Kovalic et al. (2012) published a very thorough characterization of the genetic modification of a soybean line using WGR. They performed paired-end sequencing with an Illumina HiSeq sequencer and generated hundreds of millions of 100-bp reads that covered the genome at 75 times average coverage. Reads mapping partially to both, the transformation vector and the host plant genome, were used to map potential insertion sites that were then confirmed by conventional techniques like PCR and sequencing of the amplification products. The molecular characterization of the genetic modifications present in two herbicide-tolerant transgenic soybeans has been recently carried out following this method for regulatory submissions prior to commercialization and use in breeding programs (Guo et al. 2016).

Wahler et al. (2013) applied basically the same approach for the study of event LL62 rice, but in contrast to the situations described before, here the sequence of the transformation vector used for the genetic modification was unknown to the laboratory performing the analysis. Around 172 million of paired-end reads of 75 bp in length were obtained with an Illumina HiSeq sequencer, which represented an average 65 times coverage of the rice genome. The reads were mapped to the rice reference genome, and breakpoints could be detected. The breakpoint border sequences were then mapped against the pCAMBIA-1300 vector sequence, and those that aligned even partially were assembled to form a putative T-DNA sequence. All reads were then mapped to this putative T-DNA sequence in a strategy named "bioinformatics read walking." Although pCAMBIA-1300 was not itself used to generate GM rice LLRice62, it contains nucleotide sequences commonly used in plant genetic engineering, like the pUC18 multiple cloning site, the aadA gene, the hptII gene, and the cauliflower mosaic virus 3′UTR and 35S promoter. The authors succeeded in developing a general approach that allows the identification of likely transgenic breakpoint border sequences, applicable to all GMOs.

Yang et al. (2013) also used WGR and bioinformatics to characterize the insertion of T-DNAs in rice when the a priori knowledge of the insertion is limited. They sequenced two transgenic rice lines with an Illumina HiSeq sequencer and obtained around 100 million of paired-end reads of 90 bp in length, an average of 25 times sequencing coverage of the rice genome. The data was analyzed considering three different scenarios, and for each case, a specific bioinformatics module was designed. Module 1 was designed for situations in which the DNA sequence of the transformation vector is completely known, an approach comparable to that of Kovalic et al. (2012). Module 2 was designed for use when the sequence of the inserted DNA is unknown, but a database of genetic elements and transgene constructs from known GMOs is available and can be used as a reference library. This approach is, thus, comparable to the one taken by Wahler et al. (2013). Module 3 was designed for use when the analyzing laboratory has absolutely no a priori knowledge of the DNA sequence of the T-DNA. In this case, first the reads were mapped to the rice genome, before de novo assembly and BLAST analysis of the retained reads. Although the results obtained were very satisfactory, the bioinformatics workload in case of the use of modules 2 and 3 was still enormous, and an automatic and simplified software needs to be developed for these strategies to be adopted by the scientific community.

A constant in all methods is that only a small fraction of the obtained reads, those mapping to the transformation plasmid, are used for the determination of the insertion site. Most recent methods that use NGS to map T-DNAs do not consider necessary to sequence the whole host genome with great depth but include strategies involving capture enrichment approaches to improve coverage of the inserted and adjacent sequences only.

Lepage et al. (2013) described "targeted genomic sequencing," a new technique that allows the simultaneous identification of multiple insertion sites in a complex DNA sample using biotinylated primers specific for the extremities of the T-DNA. The biotinylated primers were hybridized to a library consisting on a pool of equivalent amounts of genomic DNA from 64 lines, and the recovered DNA was used for Roche/454 sequencing and identification of the region flanking the T-DNA in each line. No analysis of potential rearrangements of the insertions was performed, just identification of the insertion sites.

Another related approach, referred to as "Southern-by-Sequencing," has been reported by Zastrow-Hayes et al. (2015). It also consists on the hybridization of indexed and pooled genomic DNA libraries from transgenic plants to biotinylated probes designed against the sequence of the T-DNA. Sequencing of recovered DNA and analysis of the obtained reads revealed the sequences adjacent to the T-DNA insertions, and thus the number of insertions and their rearrangements could be inferred.

Inagaki et al. (2015) published a work exploring the application of sequence capture-based methods and NGS for high-throughput identification of T-DNA insertion site and structure. They custom-developed methods using a mixture of biotinylated hybridization probes targeted against the various T-DNA ends for target enrichment and bioinformatics tools to determine the location of T-DNA insertions in the *Arabidopsis* genome.

Guttikonda et al. (2016) showed WGR and target capture can be applied to the molecular characterization for regulatory submissions of single and stacked transgenic events. They characterized two transgenic soybean lines and their hybrid stack using paired-end sequencing and hybridization to a BAC clone including the transformation vector. When they compared their results with those obtained by traditional techniques, NGS showed a significant advantage at detecting small rearrangements of the T-DNAs.

13.3 Discussion

13.3.1 Read Length

The read length obtained after sequencing can be determinant on the feasibility of the generation of genetic maps of transgene inserts to determine insertion sites and T-DNA rearrangements. Roche 454 yields relatively long reads up to 700 bp but at a relatively high price, so most of the published studies from the last years apply Illumina technology, typically yielding short read lengths (50–300 nt). However, Illumina technology makes possible paired-end sequencing, what facilitates detection of the insertion loci of T-DNAs. With recent platforms such as PacBio and MinION, it is possible to obtain very long reads (several thousand base pairs), potentially useful to complete transgenic insert sequences (Holst-Jensen et al. 2016). Combinations of platforms and strategies have proven ideal to obtain detailed and verified information.

For example, Scouten et al. (2017) reported difficulties in assembling short Illumina reads for characterization of inserts, due to multiple inserts per plant, and PacBio sequencing appeared to be helpful to solve this complex assembly.

Another challenge faced when using short reads is the detection of a junction locus within plant-repetitive genome sequences. Illumina short reads may cause an increase in the percentage of ambiguously or incorrectly mapped reads, and this limitation can be overcome also by using sequencing platforms that generate long-read sequencing data (Park et al. 2017).

13.3.2 Coverage

A balance must be achieved between the benefits of a big amount of raw data and its high cost and time consumption. It is obvious that with a deeper sequencing and a higher coverage, the analysis leading to the molecular characterization of a transgenic line becomes easier (Park et al. 2017), but an alternative to increasing sequencing depth can be to perform a fraction enrichment step in the protocol before sequencing to enrich the target sequence (as discussed above Lepage et al. 2013; Zastrow-Hayes et al. 2015; Inagaki et al. 2015; Guttikonda et al. 2016). However, Lambirth et al. (2015) conducted paired-end sequencing of a soy sample on the Illumina HiSeq 2000 system to around five times theoretical genome-wide coverage with 100 base-pair reads, and such low coverage was reported enough to locate and identify a single-copy transgene insertion in a highly complex and repetitive genome like that of soybean. The use of lower coverage can be convenient but is dependent on the existence of a reference genome to facilitate alignments and produces information that needs posterior verification using traditional molecular biology techniques.

13.3.3 Sequenced Reference Genome

Sometimes it is necessary to sequence and de novo assemble a complete genome to study a specific genetic modification. Nowadays it might be feasible on a small computer cluster or even on a single high-quality machine, but the genomes of higher eukaryotes are large and complex and require a certain effort to sequence and assemble (Holst-Jensen et al. 2016).

The "SunUp" papaya was the first transgenic plant where the genome was fully sequenced and assembled in an effort to characterize the genetic modification that conferred it virus resistance (Ming et al. 2008). Sequencing depth was very low though (3 times coverage), and Southern blot experiments were needed to complement the NGS data.

13.3.4 Alignment Algorithm

For high-throughput NGS data processing, a read alignment tool with an algorithm is used, and the choice of that aligner tool can lead to very different efficiencies in genome-T-DNA junction detection and different computational running times. Park et al. (2017) tested different alignment programs and their combinations and compared the results obtained with them in the analysis of a transgenic rice line. In their paper, they report that the BWA-MEM combination showed the best performance and it had longer runtimes than Bowtie or the BWA-aln algorithm, but the results obtained proved more accurate and reliable.

In any case, since many researchers have difficulties in handling large quantities of bioinformatics data and applying algorithms, there are now user-friendly methods that handle NGS data and help in the detection of genome-T-DNA junctions, that do not require almost any computational science background knowledge, and that can be used by most biologists (Park et al. 2017; Lambirth et al. 2015).

13.4 Conclusion

NGS can facilitate the molecular characterization of GMOs in light of its high throughput, continuously decreasing costs and the development of a diverse range of bioinformatics tools that allow transgene identification, location, and characterization. Compared with traditional Southern blot and PCR-based methods, whole genome re-sequencing has become a faster, cheaper, simpler, and more effective approach for transgenic modification analysis.

References

Guidance for risk assessment of food and feed from genetically modified plants (2011) EFSA J 9 (5):2150

Guo B, Guo Y, Hong H, Qiu L-J (2016) Identification of genomic insertion and flanking sequence of G2-EPSPS and GAT transgenes in soybean using whole genome sequencing method. Front Plant Sci 7:1009

Guttikonda SK, Marri P, Mammadov J, Ye L, Soe K, Richey K et al (2016) Molecular characterization of transgenic events using next generation sequencing approach ed: Jain M. PLoS One 11(2):e0149515

Holst-Jensen A, Spilsberg B, Arulandhu AJ, Kok E, Shi J, Zel J (2016) Application of whole genome shotgun sequencing for detection and characterization of genetically modified organisms and derived products. Anal Bioanal Chem 408(17):4595–4614

Inagaki S, Henry IM, Lieberman MC, Comai L (2015) High-throughput analysis of T-DNA location and structure using sequence capture ed: Candela H. PLoS One 10(10):e0139672

Ji J, Braam J (2010) Restriction site extension PCR: a novel method for high-throughput characterization of tagged DNA fragments and genome walking ed: Herrera-Estrella A, editor. PLoS One 5(5):e10577

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 5(3)

Kulski JK (2016) Next-generation sequencing — an overview of the history, tools, and "Omic" applications. In: Next generation sequencing – advances, applications and challenges. InTech, Croatia

Lambirth KC, Whaley AM, Schlueter JA, Bost KL, Piller KJ (2015) CONTRAILS: a tool for rapid identification of transgene integration sites in complex, repetitive genomes using low-coverage paired-end sequencing. Genomics Data 6:175–181

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

Liu YG, Mitsukawa N, Oosumi T, Whittier RF (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J 8(3):457–463

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452(7190):991

O'Malley RC, Ecker JR (2010) Linking genotype to phenotype using the Arabidopsis unimutant collection. Plant J 61(6):928–940

Ochman H, Gerber AS, Hartl DL (1988) Genetic applications of an inverse polymerase chain reaction. Genetics 120(3):621–623

Park D, Park S-H, Ban YW, Kim YS, Park K-C, Kim N-S et al (2017) A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data. BMC Biotechnol 17(1):67

Pérez-González A, Caro E (2016) Hindrances to the efficient and stable expression of transgenes in plant synthetic biology approaches. In: Systems biology application in synthetic biology. Springer India, New Delhi, pp 79–89

Polko JK, Temanni M-R, van Zanten M, van Workum W, Iburg S, Pierik R et al (2012) Illumina sequencing technology as a method of identifying T-DNA insertion loci in activation- tagged *Arabidopsis thaliana* plants. Mol Plant 5:948–950

Schouten HJ, van de Geest H, Papadimitriou S, Bemer M, Schaart JG, Smulders MJM et al (2017) Re-sequencing transgenic plants revealed rearrangements at T-DNA inserts, and integration of a short T-DNA fragment, but no increase of small mutations elsewhere. Plant Cell Rep 36 (3):493–504

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 Anal Methods 6 (6):1718–1727

Wolffe AP (1997) Transcription control: repressed repeats express themselves. Curr Biol 7(12): R796–R798

Yang L, Wang C, Holst-Jensen A, Morisset D, Lin Y, Zhang D (2013) Characterization of GM events by insert knowledge adapted re-sequencing approaches. Sci Rep 3(1):2839

Zastrow-Hayes GM, Lin H, Sigmund AL, Hoffman JL, Alarcon CM, Hayes KR et al (2015) Southern-by-sequencing: a robust screening approach for molecular characterization of genetically modified crops. Plant Genome 8(1):0