AGRONOMY SCIENCE
wcześniej – formerly
Annales UMCS sectio E Agricultura

VOL. LXXIII (4) 2018

http://dx.doi.org/10.24326/asx.2018.4.9

[1] National Centre for Plant Genetic Resources, Plant Breeding and Acclimatization Institute – National Research Institute, Radzików
[2] Department of Plant Breeding and Genetics, Plant Breeding and Acclimatization Institute – National Research Institute, Radzików
[*]e-mail: m.puchta@ihar.edu.pl

MARTA PUCHTA[1*], PAULINA BOLC[1], URSZULA PIECHOTA[2]

Review of genome sampling methods in sequencing libraries preparation protocols

Metody redukcji złożoności genomu w protokołach tworzenia bibliotek do sekwencjonowania

Summary. Since the publication of the full sequence of the *Arabidopsis thaliana* genome in 2000 [AGI Initiative 2000], a period of dynamic genome exploration began. In the last decade, with the revolution in the next generation sequencing, the number of scientific reports based on sequence analysis has increased exponentially. New, fast, high throughput and relatively inexpensive nucleic acid sequencing technologies have become available and widespread, opening up the possibility of making extensive use of molecular tools in science and breeding practice. These new methods include whole genome sequencing and various methods of reduced representation sequencing. The multitude of available methods, varying in terms of availability and costs, generating different types of result data, dedicated to different research and application purposes, may make it difficult to choose the best variant [Poland et al. 2012]. The aim of this paper is to familiarize the reader with the possibilities and application of selected genotyping techniques with the use of sequencing.

Key words: next generation sequencing, reduced representation sequencing, DNA library, RADSeq, ddRADSeq, GBS, DArTSeq

INTRODUCTION

Initial DNA sequencing attempts date back to the first half of the 20th century. In 1977, the first method of efficient DNA sequencing called the Sanger method, was developed [Sanger et al. 1977]. In the second half of the 20th century, the intensive development of innovative sequencing methods, which continues to this day, has begun. Continuous automation and miniaturization of devices enable easier and faster understanding of nucleotide sequences. Dynamic development of sequencing technology has led to the division into three generations of sequencing: the first is the Sanger chain termination method, the second is based on parallel mass sequencing from thousands to millions of different matrices (libraries) and the third enabling a direct reading of sequences from a single DNA molecule (single molecule sequencing) without the need for prior amplification [Sanger et al. 1977, Kotowska and Zakrzewska-Czerwińska 2010]. Since 2001, a great progress has been observed in the development of these methods, and nowadays genetic sequencing is increasingly used in the next generation sequencing (NGS). This technology has a variety of advantages. First of all, it is less labor-consuming, more reliable and providing a larger amount of result data. Systems for identifying nucleotide DNA variations based on NGS sequencing are full-genome sequencing [Hillier et al. 2008], sequencing with methylome analysis [Brunner et al. 2009] and sequencing of the reduced genome fraction, e.g., targeted sequencing of coding regions (*exome capture*) [Ng et al. 2009] or products obtained after digestion by restriction enzymes [Davey et al. 2011]. There are many SNP (*single nucleotide polymorphism*) platforms and systems available [Fan et al. 2006]. However, unlike microarray genotyping which requires prior panel development for a given species, genotyping using NGS can be used *de novo* for organisms with an unknown genome. Microarrays panels are often useful only within the population, for which they were developed.

With the advent of compact sequencers such as MiSeq Illumina [https://www.illumina.com/systems/sequencing-platforms/miseq.html], sequencing has become a technique available for every molecular laboratory. The throughput of this device allows simultaneous analysis of up to 1,500 amplicons from 96 samples (96 indices) simultaneously and allows for cumulative coverage of up to 650 kbp within one work cycle. It allows to identify even tens of thousands of variations of a single nucleotide or short insertions and deletions (In/Del) on one reaction plate. Short readings generated with the use of NGS sequencers (2x300 bp for Illumina MiSeq) are, however, insufficient for sequencing the full-genome composite plant genomes and *de novo* sequence assembly. Full-genome sequencing of each individual object from the population is not necessary for many analyzes; it is also time-consuming and expensive. Therefore, along with the increase in the availability of NGS, protocols have been developed to enable reliable, fast and economical genotyping of organisms with large and complex genomes based on sequencing of the reduced genome fraction [Altshuler et al. 2000]. The described methods make it possible to obtain the repeatability of the sequenced pool of DNA fragments in subsequent, analyzed samples. Repeatability is achieved in the sequencing protocols of targeted PCR products, library enrichment through pre-hybridization with probes or fractionation using restriction digestion [Bybee et al. 2011, Elshire et al. 2011, Myllykangas et al. 2011]. Targeted sequencing includes genes or regions that are selected from genomic DNA prior to sequencing. The reference sequence is known in advance. The procedure for preparing such libraries may be based on a probe hybridization or multiplex PCR [Cheng et al. 2010, Hedges et al. 2011, Kiialainen et al. 2011]. In targeted sequencing, complementary primers to the sequence of the analyzed regions are designed. Amplicons are multiplexed and simultaneously sequenced [Gasc et al. 2016]. Due to this approach, only areas that are interesting from the point of view of the experiment can be analyzed, reducing the labor intensity and cost of the experiment [Gasc et al. 2016]. One of the methods of targeted sequencing is the sequencing of encoding regions (*exome capture*). It is assumed that only 1–2% of the genome of many crop plants contains encoding regions [Mascher et al. 2013].

The sequencing techniques of the reduced genome fraction allow obtaining information on SNP and In/Del variations (*insertion and deletion polymorphisms*). Variant analysis can then be used in genetic diversity [Song et al. 2013], evolutionary studies, and associative analysis (GWAS, *genome-wide association study*) [Guo-Qian et al. 2016], searching for candidate genes and markers for genomic selection [Collard et al. 2008], for the creation of genetic maps, quantitative character *loci* mapping (QTL) and single genes based on the genotyping of population segregating [Andolfatto et al. 2011, Huang et al. 2010, Elshire et al. 2011]. Protocols for constructing RRS (*reduced representation sequencing)*, based libraries to NGS allow different degree of the genome saturation. For the study of wild populations and in the absence of the reference genome, a high density of markers is required, like for example in the RADSeq method (r*estriction-site associated DNA sequencing*). However, when we deal with marker-assisted selection, or QTL mapping in genotypes with a high degree of homozygosity and known parental plant genome, low-coverage genotyping is used, which is obtained in the GBS (g*enotyping by sequencing*) method [Gore et al. 2009, Huang et al. 2009, Xie et al. 2010].

GENOME COMPLEXITY REDUCTION METHODS BASED ON RESTRICTION DIGESTION

The RRS methods (Fig. 1) are based on reducing the number of discriminated SNPs with a fixed position in the genome. The RRS protocols differ in their suitability for various research purposes, but they generate similar costs. They are more applicable to *de novo* sequencing and are more cost-effective when organisms with large genomes or populations with high heterozygosity are analyzed [Poland et al. 2012].

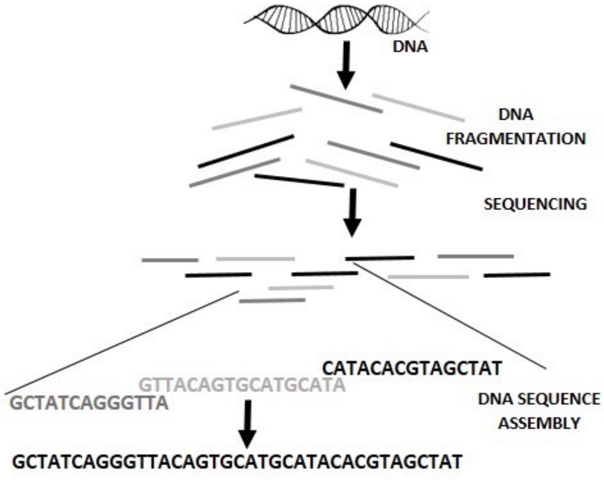

Fig. 1. The stages of the reduced representation sequencing analysis

A large group of RRS methods uses restriction enzymes to digest genomic DNA at the initial stage of library preparation, however protocols may differ in the number and type of enzymes used [Poland et al. 2012]. Baird et al. (2008) described the technique of controlled reduction of genome complexity for the purpose of NGS sequencing using restriction digestion. In the RRS, data on nucleotide sequences adjacent to restriction enzyme cleavage sites are obtained. Restriction digestion allows a large fraction of the same fragments to be obtained in all samples [Davey et al. 2011, Elshire et al. 2011, Sonah et al. 2013]. The flexibility of these methods allows the protocol to be adapted to the material and research objective as well as the budget. Enzymes that recognize longer sequences digest the genome at fewer sites and generate fewer *loci*. Fragment length in methods using two restriction enzymes is determined by the distance between the cutting sites [Andrews et al. 2016]. The number of obtained *loci* can be estimated *in silico* by selecting the enzyme pair and the range of fragments length. Precise information can be generated *in silico* based on the reference genome of the tested object [DaCosta and Sorenson 2014]. Selection and optimization of the restriction used enables the exclusion of low-informative, often methylated repetitive regions from the analysis, through the use of methylation-sensitive enzymes [Baird et al. 2008].

1. RADSeq METHOD

Development of restriction-related sequencing – RADSeq has been recognized as one of the most important breakthroughs in molecular research over the last decade [Poland et al. 2012]. RADSeq sequencing consists in reducing the representation of the genome, by sampling it, while ensuring greater coverage depth [Andrews et al. 2016]. RADSeq techniques are based on high molecular weight genomic DNA, which makes methods unsuitable for working with highly degraded DNA.

RADSeq methods are used as a research tool in the areas of ecological and evolutionary genomics as well as phylogenetics, using the huge efficiency of NGS. The technique allows the discovery of hundreds of thousands of polymorphic genetic markers in the entire genome in one simple and economical experiment [Andrews et al. 2016]. High density of this type of markers in genomes makes them perfect for research on the heredity of genomic regions. Unlike many other sequencing methods, RADSeq does not require any prior information about the genomes of the organisms studied. Therefore, this technology has become the most commonly used genomic approach applied during high-throughput detection and discrimination of SNPs in studies of non-model organisms [Andrews et al. 2016].

The *Neurospora crassa* fungus and the *Gasterosteus aculeatus* three-spined stickleback were the first model organisms, for which the RADSeq technique was used. In a situation where there is no reference to the reference genome, RAD markers can be analyzed based on bio-informatic methods, building the genetic sequence from scratch [Bergey et al. 2013].

Creating a library using the RADSeq method starts with the isolation of genomic DNA (gDNA) with a relatively high molecular weight (Fig. 2a, 2b). In the subsequent step, gDNA is digested with one or more restriction enzymes. Then specific adapters are attached to DNA fragments. Adapters added during RADSeq protocols may contain built-in barcodes, short unique sequences of 6–12 bp, allowing to distinguish between multiplexed samples. Depending on the restrictases used, the following step is to select the size of DNA fragments optimal for sequencing [Andrews et al. 2016]. Selection of defined length fragments limited by two different restriction sites results in high repeatability of sequenced fragments in subsequent analyzed samples [Andrews et al. 2016]. The nucleotide sequence is read during sequencing from a single end of the DNA strand, generating one reading forward or from both ends, generating two readings for each fragment, one reading forward and one backwards.

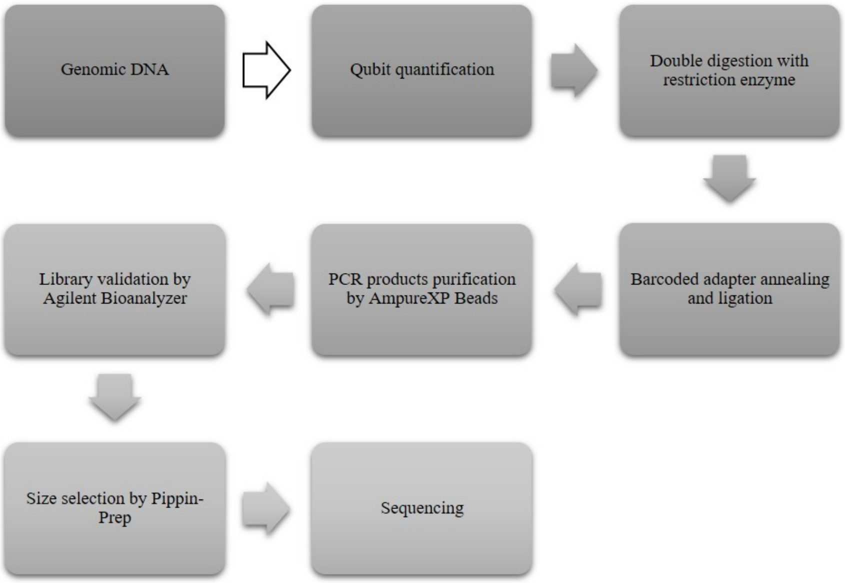
Genomic DNA
Qubit quantification
Double digestion with restriction enzyme
Barcoded adapter annealing and ligation
PCR products purification by AmpureXP Beads
Library validation by Agilent Bioanalyzer
Size selection by Pippin-Prep
Sequencing

Fig. 2a. The stages RADseq (restriction-site associated DNA sequencing) library preparation

The RADSeq method uses a combination of single enzyme digestion and mechanical fragmentation. For sequencing, fragments delimited by a restriction site at one end and a randomly truncated other end, are selected [Andrews et al. 2016].

After restriction digestion, the adapters are ligated to the sticky ends of DNA. The P1 adapter contains a complementary sequence for Illumina forward primer, one of 48 unique barcodes and a sequence complementary to the viscous end left after restriction digestion. The P2 adapter contains a complementary sequence for the Illumina reverse primer, the 6-nucleotide Illumina index sequence and four (AATT) unpaired nucleotides [DaCosta and Sorenson 2014]. The free 3’ end of the P2 adapter denoted as the “Y” adapter, causes the reverse primer to not bind the P2 adapter until the complementary sequence is completed. It is filled during the first round of amplification starting with the P1 adapter. Due to this, all sequencing readings start synchronously from the side of the P1 adapter [Baird et al. 2008].

Preparing libraries with the use of adapters containing barcodes enables multiplexing of samples, which in turn reduces the costs and time of subsequent stages of the study. DNA fragments from each sample are identified by a unique combination of two different identifiers of one barcode and one Illumina index of 6–8 bp located near the center of the adapter. The use of designed barcodes is a cheaper alternative to the use of two original Illumina indices. The use of barcodes reduces the total number of adapters required to distinguish between samples, e.g. due to a set of 24 adapters with barcodes and 16 indices, we can clearly identify 384 samples [Andrews et al. 2016]. Adapters to be ligated with DNA fragments are designed to ensure sequencing of only the target fragments adjacent to the restriction cleavage sites [Andrews et al. 2016].

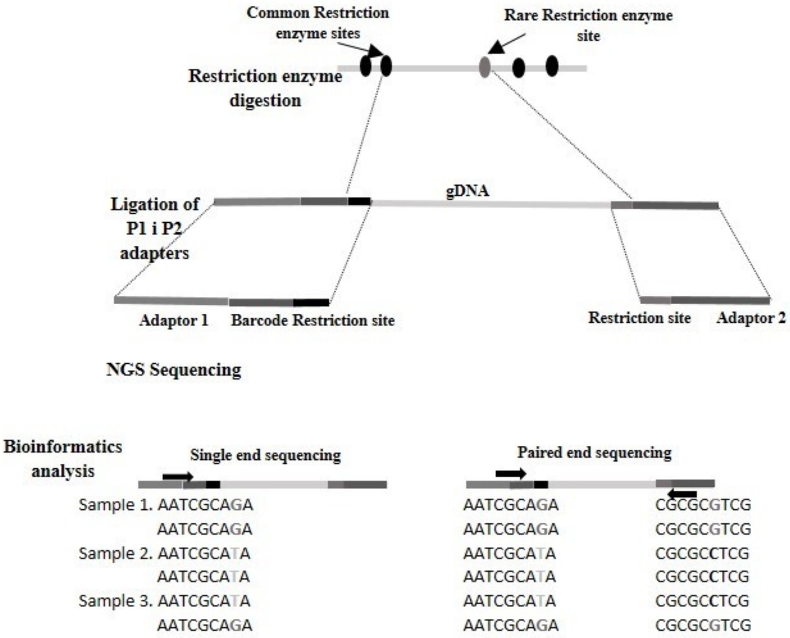

Fig. 2b. Restriction-site associated DNA sequencing (RADSeq) library preparation

The next stage in the preparation of RADSeq libraries is the PCR amplification of DNA fragments containing previously attached adapters. Amplified fragments are then multiplexed and randomly fragmented.

During preparation of NGS libraries, the most popular tool for the selection of DNA fragments and for the purification of residues of components after a PCR reaction is the use of magnetic beads. A purification system based on magnetic bead technology ensures the highest DNA quality without transferring salts, free nucleotides and enzymes [Quail et al. 2012].

One of the final stages of library preparation before sequencing is the precise selection of DNA fragments of a given length by means of an electrophoretic separation [Etter et al. 2011]. The platform for automatic electrophoresis and elution of Pippin Prep (Sage Science, USA), recommended by Illumina, is used for this. Another selection method, still used in the research, is a classic electrophoretic separation in an agarose gel followed by excision of the gel fragment and elution of fragments with selected length range [Mardis and McCombie 2017]. Currently, it is more often recommended to use automatic separation of DNA fragments, significantly facilitating and accelerating construction of the library, and the obtained material is definitely of better quality [http://www.sagescience.com/wpcontent/uploads/2012/11/sage_wp_saygoodbyetomanualgels_0912_6.pdf]. Verification of the length of fragments obtained and the qualitative and quantitative assessment of the library can be carried out by capillary electrophoresis on a micro-plate (Agilent Bioanalyzer).

Table 1. Comparison of the RADSeq method and its modification (developed on the basis of Andrews et al. 2016)

	RADSeq	2bRADSeq	ddRADSeq	ezRADSeq	GBS
Options to adjust the number of *loci*	Enzyme selection	Enzyme selection	Enzyme selection, selection of fragments	Enzyme selection, selection of fragments	Enzyme selection
Number of *loci* per 1 Mb of the genome size	30–500	50–1000	0,3–200	10–800	5–40
The length of *loci* with a single end	≤ 300 pz ≤ 300 bp	33–36 pz 33–36 bp	≤ 300 pz ≤ 300 bp	≤ 300 pz ≤ 300 bp	< 300 pz < 300 bp
Cost of indexed sample	low	low	low	high	low
Using a ready kit	no	no	no	yes	no
Possibility to identify the duplicates	for paired ends sequencing	no	yes	no	yes
Hardware requirements	sonicator	no	Pippin Prep standard gel cutting equipment can be used	Pippin Prep	no
Suitability for large genomes	good	poor	good	good	moderate
Suitability for identifying the *locus de novo*	good	poor	moderate	moderate	moderate

The RADSeq method makes it possible to obtain thousands of single nucleotide polymorphisms. In 2016, Guo-Qian et al. [2016] used the RADSeq method for extensive analysis of many species of eggplant, chickpea, sesame, soy, pumpkin and bamboo. The researchers were developing molecular markers, constructing genetic maps, and mapping the QTL. In addition, Guo-Qian and coworkers performed analyses in the field of population genetics and phylogenetics of the studied species. To assess the efficiency of the RADSeq method, they tested it on the species *Oryza sativa* L. *japonica* and *Zea mays* L. In addition, they verified the reproducibility of the method for the species *Phyllostachys edulis* and *Alloteropsis semialata* [Guo-Qian et al. 2016]. They reconstructed the phylogenetic relationships of two types of woody bamboos *Dendrocalamus* and

Hyllostachys. Readings obtained after sequencing were mapped to the reference genomes of *O. sativa*, *Z. mays*, *A. semialata*, and the bamboo readings to *P. edulis* [Guo-Qian et al. 2016]. During the study, they tested the universality of commonly used restriction enzyme pairs for 23 plant species. On the basis of the result analysis of the conducted research, they chose the combination of *AvaII* and *MspI* as generating the largest number of fragments for the tested angiosperm plant species [Guo-Qian et al. 2016].

In 2016, Andrews et al. used the RADSeq method to map the utility traits in the genomes of agricultural plant species. The SNP analysis remains a challenge for polyploid species due to numerous paralogy, homologs and repetitive sequences. Wu et al. in 2016 constructed a ddRADSeq library (*double digest restriction associated DNA*) for inbred *B. napus* lines. This plant is an allotetraploid derived from *B. napa* and *B. oleracea* hybridization. Sequencing was performed to look for the SNP type variability and 189 inbred genotypes. Researchers obtained 506.81 million readings with a length of 90 bp, and the average number of readings per line was 2.68 million.

In 2015, using the ddRADSeq method, the exact genetic map of the cultivated strawberry (*Fragaria* × *ananassa* Duch.) was constructed. The constructed map allowed the analysis of conjugations and segregations of interested features at the offspring. The first strawberry genetic map was developed. Constructed maps contained genes of both parental varieties. The ddRADSeq method has become a useful tool for creating genetic maps with good resolution for studying the segregating features [Davik et al. 2015].

Konar et al. [2017] used the ddRADSeq method to construct the genetic map of Northern red oak (*Quercus rubra*), a highly heterozygous, angiospermous tree that does not have a reference genome. The application of ddRADSeq method in combination with the SSR (Simple Sequence Repeats)-based marker analysis allowed researchers to construct a genetic map of the oak. Developed map was used to detect and analyze alleles that contribute to the stress tolerance and evolutionary research [Konar et al. 2017].

Tomato wild relatives carrying resistance genes to tomato disease caused by the fungus *Phytophthora infestans*. Genes of pathogen resistance can be used in breeding to protect cultivars. In the genome of wild tomato *Solanum habrochaites*, a new *locus* of quantitative resistance to *P. infestans* was identified. The ddRADSeq method was applied to the genotyping of the F_2 mapping population derived from interspecies crossbreeding. The association analysis revealed the region of 6.8 Mbp genome on chromosome 6 as a potential disease resistance *locus*. In ddRADSeq analysis for parental lines and F_2 population, 616 763 readings were obtained for each sample. The GWAS analysis was carried out simultaneously in many mapping populations. The use of ddRADSeq technology allowed to identify a new *locus* of resistance to disease caused by *P. infestans* [Arafa et al. 2017].

2. MODIFICATIONS OF RADSeq METHOD

Widespread use of RADSeq resulted in a rapid development of derived methods. The resulting variants have increased the flexibility of the RADSeq method, as well as the reduction of costs and labor-intensive preparation of DNA library [Andrews et al. 2016]. The methods described differ in the number of *loci* used for further studies. The basic tool for optimizing the number of *loci* is the choice of restriction enzymes. The optimal level of genome reduction depends on the research goals and assumptions of each experiment [DaCosta and Sorenson 2014]. The RADSeq method allows the selection of enzymes used and the size of DNA fragments obtained after restriction digestion. Modifications of the RADSeq technique differ in: the order of the library preparation steps, restriction digestion conditions, adapter ligation, barcoding and selection of the size of DNA fragments. In addition, the protocols differ in the number of restriction enzymes used and the frequency, at which they cut the genome [Andrews et al. 2016]. Fig. 3 presents four main modifications of basic RADSeq protocol. The RADSeq and 2bRadSeq protocols aim to obtain the sequence data at all restriction enzyme cleavage sites. In contrast to this approach, other methods are based on the selection of the size of DNA fragments obtained after digestion with two restrictases. Usually, fragments between 300 and 600 bp in length limited by two restriction sites, are selected. Table 1 summarizes the comparison of the RADSeq modification.

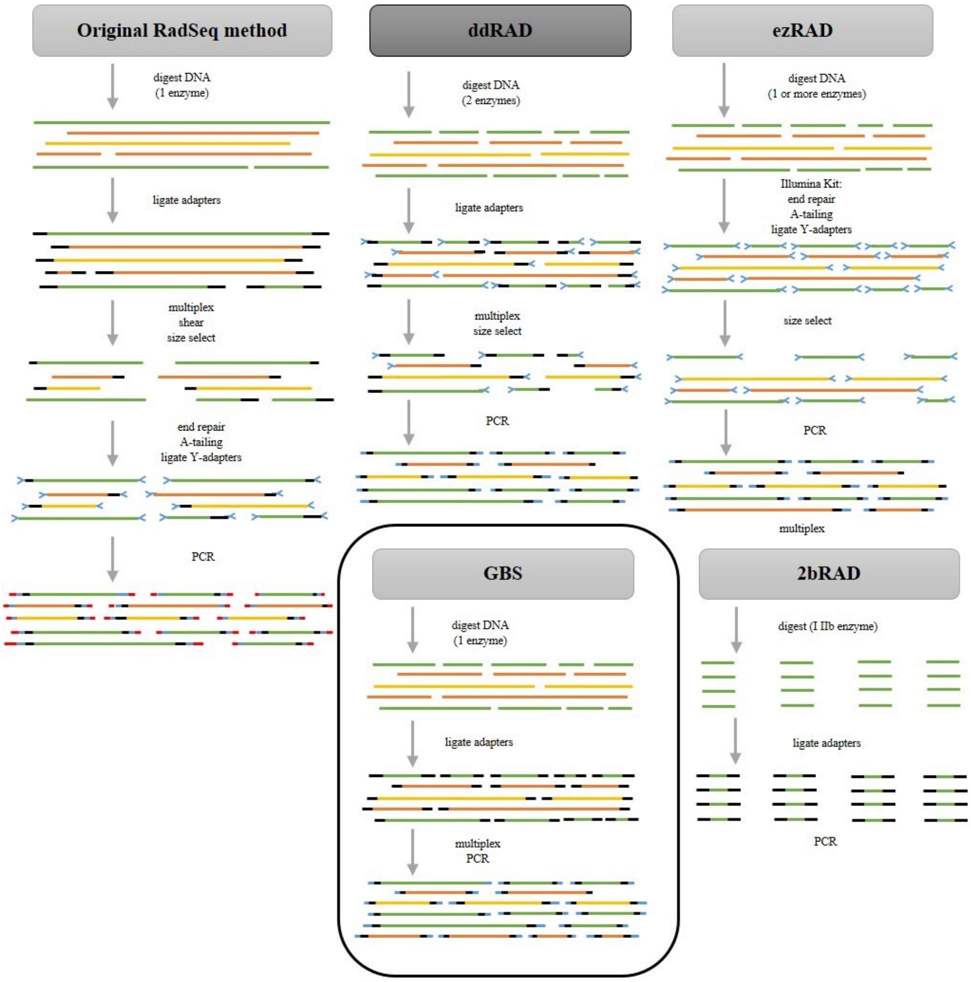

Fig. 3. Comparison of restriction-site associated DNA sequencing methods

2bRADSeq

The 2bRADSeq method (*type IIB endonucleases restriction site associated DNA*) uses II type of restriction endonucleases, which cut DNA at a defined place, in the area of recognized sequence or in its vicinity. Type II endonucleases recognize palindrome sequences and their activity depends on the availability of Mg^{2+} ions in the reaction environment. The 2bRADSeq method is based on the selection of short 33–36 bp depending on the enzyme used (e.g. 33 bp (*Bsa*XI) or 36 bp (*Alf*I) fragments of DNA. Short DNA fragments can limit the potential to detect SNPs. Protocol for the construction of 2bRADSeq library omits several stages, during which there is the possibility of DNA loss. In addition, the method eliminates random cutting and repair of DNA ends. The simplicity of 2bRADSeq protocol makes it particularly suitable for high-throughput genotyping required for mapping the conjugation and profiling of genetic variation in natural populations [Wang et al. 2012].

ddRADSeq

The ddRADSeq method is used in population studies and in phylogenetic analyses. It is based on the reduction of the genome, in which 0.1 to 10% of genomic DNA is used to construct libraries [Davey et al. 2011]. The reduction is performed by digesting DNA with two restriction enzymes. The method adjusts the number of DNA fragments obtained using two different restriction enzymes (frequent-cutters recognizing the sequences with a length of 4 bp and rare-cutters recognizing the sequences of 6-8 bp in length) and selection of the size of obtained fragments. One of the enzymes used should cut the DNA more often than the other one to minimize the number of fragments limited by the same restriction site at both ends [DaCosta and Sorenson 2014]. The ddRADSeq library contains fragments of a certain length, limited by defined restriction sites [Peterson et al. 2012]. The use of two restriction enzymes enables sequencing of paired readings at identical *loci* of many samples. From the point of view of high mapping accuracy, even in complex genomes, ddRADSeq has an advantage over GBS and RADSeq [Kenta et al. 2016].

DdRADSeq technology allows for the creation of a large number of markers for high-quality SNPs, which enables the construction the creation of high quality genetic maps of moderate density. The ddRADSeq method is a good approach to the analysis of organisms for which the reference genome is not available. The method generates thousands of SNPs allowing for the construction of phylogenetic trees, accurate genetic maps and detection of genetic variation of a population [Konar et al. 2017]. The ddRADSeq libraries can be sequenced as single or paired readings using only the barcodes to distinguish the samples. This approach is flexible and cost-effective, but multiplexing a large number of samples requires the appropriate tools for analysis [Peterson et al. 2012].

ezRADSeq

ezRadSeq differs from other RADSeq methods primarily in the application of standard Illumina TruSeq library preparation kits. The method uses two isoschizomers of restriction enzymes specific to the same recognition DNA sequence (GATC). ezRADSeq also allows the flexibility of using any restriction enzyme (or combination of enzymes) that digests frequently enough to generate fragments with the desired size range, without the need to purchase specific adapters for each restriction site. In the ezRADSeq procedure, like in ddRADSeq, there is a fragment size selection step to eliminate DNA fragments of undesirable length from the pool for sequencing. The ezRADSeq method was used in taxonomic studies, in studies of non-model organisms, the search for new, non-described polymorphisms, and targeted sequencing of a defined fragment in natural study upon populations [Guo-Qian et al. 2016]. Commercial Illumina TruSeq kits enable easy library preparation by means of ezRADSeq.

3. GBS METHOD

GBS is a relatively simple technique for sequencing large genomes in diverse organisms. It allows multiplexing a large number of samples in one reaction. GBS generates large number of SNPs used for genetic analyses and genotyping [He et al. 2014]. The number of SNP markers generated by the GBS method depends on the size of genome and its complexity [Elshire et al. 2011]. The GBS procedure consists of several stages including preparation of DNA samples, digestion using restriction enzymes, attachment of adapters and library multiplexing. Product amplification precedes sequencing and analysis of results. The advantages of this technique are: short time of sample preparation, lack of fractionation, no selection stage of fragment sizes, fewer PCR and purification reactions, and application of indices [Pachota et al. 2016]. The GBS technique is used, among others, in QTL mapping, identification of candidate genes, construction of haplotype maps, analysis of genetic diversity and molecular phylogenetics [He et al. 2014]. Original GBS method may not be appropriate for highly methylated genomes, large MAS projects or in phylo-geographical studies of wild populations [Davey et al. 2011]. Selection of enzymes depends on a desired marker density and should be experimentally developed for a species [Van Tassell et al. 2008, Elshire et al. 2011].

4. DArTseq METHOD

DArTseq is the genotyping method offered by Diversity Arrays Technology Pty. Ltd Canberra Australia [https://www.diversityarrays.com/] since 2011. The platform offers genotyping using the next-generation sequencing in Illumina technology [Illumina Inc., San Diego, CA]. The DArTseq method allows to achieve up to three times more dominant markers compared to classical DArT genotyping method using DNA microarray hybridization [Jaccoud et al. 2001, Sansaloni et al. 2011]. Obtained PIC parameters (*polymorphic information content*) [Roldan-Ruiz et al. 2000] are on average higher for classic DArT markers, however, higher number of markers obtained by DArTseq method allows to obtain better parameters in analyses [von Cruz et al. 2013].

DArTseq genotyping is a versatile method. It is applicable both to species with developed reference sequence and in *de novo* sequencing. If there is reference data, the physical location of the DArTseq markers in the genome can be determined. Reduction of the genome complexity in the DArTseq genotyping method is obtained by digestion with two restriction enzymes. Briefly, the DArTseq procedure involves DNA isolation, evaluation of its integrity and purity, digestion of optimized-restriction DNA, adapter ligation, short-reading sequencing, and analysis of results. Finally, only fragments limited by two different restriction sites enter the sequencing reaction. Optimization of the combination of enzymes used in the DArTseq method allows selective sequencing of highly informative fragments, and the elimination of a repetitive sequences from the procedure. In plant genomes, repetitive sequences are largely methylated [Bennetzen et al. 1994], therefore, the DArTseq method uses methylation sensitive restriction enzymes. Optimization is also carried out for a particular material being analyzed, for example, for genotyping the objects of *Secale* genus, *Pst*I – *Hpa*II enzymes were used [Al-Beyroutiová et al. 2016], and for *Triticum* and *Pisum Pst*I – *Mse*I [Dracatos et al. 2016, Barilli et al. 2018]. As a result, up to 90% of the markers obtained are complementary to the unique sequences of the genome [Courtois et al. 2013, von Cruz et al. 2013]. Researchers also showed a positive correlation of DArTseq marker density with the encoding regions in the *Brassicaceae* genome [von Cruz et al. 2013]. The DArTseq method produces short sequence reads of 69 nucleotides. Sequence complementary to the sequencing primer contains only one of the adapters, thus generated unpaired readings always start with the restriction sequence for one of the two enzymes used.

The DArTseq analysis results in two sets of data. The first contains dominant markers – silicoDArT – shown as the matrix 0 and 1 of the presence or absence of a fragment in the sample. The silicoDArT's variants result from differences in susceptibility of a given genome site to restriction enzyme digestion. They may be indicative of mutation or methylation at the restriction site. The second set of results contains DArTSNP codominating markers with specific polymorphisms of single nucleotides. The DArTSNP result matrix contains information which of the SNP variants is present in the sample. In some studies only one set of results is used. An example is the research by Al-Beyroutiová et al. [2016], who used only dominant silicoDArT markers for phylogenetic analysis of objects of the *Secale* genus. The use of silicoDArT in rye research has allowed for the identification of twice as many polymorphisms as compared to DArTSNP [Milczarski et al. 2016].

DArTseq genotyping can be highly economical. After DArTseq analysis, Milczarski et al. [2016] selected markers correlated with the analyzed feature, then converted them to simple PCR markers. Due to this approach, they extended the analysis of relevant markers to 658 rye objects and obtained a high-resolution genetic map containing the *Rfc1* gene for restoring the male fertility. Dracatos et al. [2016] analyzed the doubled haploid lines of wheat. They used polymorphic silicoDArT markers strongly correlated with the resistance trait to create a genetic map containing *locus* of seedling resistance to yellow rust (*P. striiformis* f. sp. *tritici*) [Dracatos et al. 2017], and in the AvocetR line, they identified new resistance genes for this pathogen, *Yr73* and *Yr74*, located on the 3DL and 5BL chromosomes, respectively. In another work [Dracatos et al. 2017], they used the same DArTseq dataset to locate QTL responsible for the racially non-specific AvocetR line resistance at *P. striiformis* f. sp. *pseudo-hordei* on the genetic map [Dracatos et al. 2017].

Deep sequencing used in the DArTseq platform allows the identification of heterozygosity in polyploid organisms such as wheat [Heslot et al. 2013]. Therefore, the genomic selection using DArTseq can be effective regardless of the level of variety ploidy, as demonstrated in potato (*Solanum tuberosum* L.) [Habyarimana et al. 2017] and cultivated strawberry studies (*Fragaria* × *ananassa*) [Sánchez-Sevilla et al. 2015]. Tyrka et al. [2015] used less than 17,000 DArTseq markers to construct a genetic map of triticale

(× *Triticosecale*). The analysis allowed for the assignment of markers to genomes A, B, R and detection of chromosomal rearrangements in the triticale genome.

DArTseq genotyping platform is widely used in research and application. Von Cruz et al. [2013] used the results of DArT and DArTseq analyses for 86 accession of the genera *Physaria* and *Paysonia* (*Brassicaceae*) in studies upon genetic diversity within the taxon. The DArTseq analysis of 84 accessions of genus *Secale* [Al-Beyroutiová et al. 2016] shed new light on the taxonomic division of *Secale* and evolution of one-year rye species, to which cultivated rye belongs. The DArTseq markers were used in the GWAS associative analysis of grain quality traits at 272 rice genotypes (*Oryza sativa* ssp. *indica*) [Qiu et al. 2015], which showed the presence of 33 new unknown QTLs and confirmed the five earlier described ones. Markers correlated with desirable features can be used to introgress traits from wild species related to cultivated varieties. Barilli et al. [2018] analyzed genetic basis of resistance to diseases of fungal origin in wild peas (*Pisum fulvum* Sibthorp & Sm.), an important source of desirable genetic variation for common peas. DArTseq markers can also be used for genomic selection. Analysis of bean plants (*Phaseolus vulgaris* L.) allowed for development of 560 SNP panel adapted for the use in breeding programs [Valdisser et al. 2017].

SUMMARY

NGS sequencing of libraries with a reduced genome fraction is an effective technique for identifying the new and discriminating known SNPs. Flexibility of the presented methods allows to obtain expected coverage and density of markers in the genome. The examples of RRS applications cited in this study indicate the usefulness of such an analytical approach to a wide range of research purposes. Construction of libraries to RRS is relatively simple and less time-consuming. In many protocols, it is also possible to use only the basic equipment of a molecular laboratory. Small and user-friendly compact sequencers available on the market have a throughput adequate to RRS library analysis. The amount of result data generated is sufficient for a wide range of research objectives in the field of plant genetics.

REFERENCES

AGI Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408 (6814), 796–815.

Al-Beyroutiová M., Sabo M., Sleziak P., Dušinský R., Birčák E., Hauptvogel P., Kilian A., Švec M., 2016. Evolutionary relationships in the genus *Secale* revealed by DArTseq DNA polymorphism. Plant Syst. Evol. 302, 1083–1091.

Altshuler D., Pollara V. J., Cowles C. R., Van Etten W. J., Baldwin J., Linton L., Lander E. S., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407 (6803), 513–516.

Andolfatto P., Davison D., Erezyilmaz D., Hu T. T., Mast J., Sunayama-Morita T., Stern D. L., 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 21 (4), 610–617.

Andrews K. R., Good J. M., Miller M. R., Luikart G., Hohenlohe P. A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17(2), 81–92.

Arafa R. A., Rakha M. T., Soliman N. E. K., Moussa O. M., Kamel S. M., Shirasawa K., 2017. Rapid identification of candidate genes for resistance to tomato late blight disease using next-generation sequencing technologies. PLOS One 12(12), e0189951.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLOS One 3(10), e3376.

Barilli E., Cobos M.J., Carrillo E., Kilian A., Carling J., Rubiales D., 2018. A high-density integrated DArTseq SNP-based genetic map of *Pisum fulvum* and identification of QTLs controlling Rust Resistance. Front Plant Sci. 9, 167.

Bennetzen J.L., Schrick K., Springer P.S., Brown W.E., SanMiguel P., 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome 37, 565–576.

Bergey Ch.M., Pozzi L., Disotell T.R., Burrell A.S., 2013. A new method of genome-wide marker development and genotyping holds great promise for molecular primatology. Int. J. Primatol. 34, 303–314.

Brunner A.L., Johnson D.S., Kim S.W., Valouev A., Reddy T.E., Neff N.F., Anton E., Medina C., Nguyen L., Chiao E., Oyolu C.B., Schroth G.P., Absher D.M., Baker J.C., Myers R.M., 2009. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res. 19 (6), 1044–1056.

Bybee S.M., Bracken-Grissom H., Haynes B.D., Hermansen R.A., Byers R.L., Clement M.J., Udall J.A., Wilcox E.R., Crandall K.A., 2011. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biol* Evol. 3, 1312–1323.

Cheng Y., Wang J., Shao J., Chen Q., Mo F., Ma L., Han X., Zhang J., Chen C., Zhang C., Lin S., Yu J., Zheng S., Lin S.C., Lin B., 2010. Identification of novel SNPs by next-generation sequencing of the genomic region containing the APC gene in colorectal cancer patients in China. OMICS 14(3), 315–325.

Collard B. C. Y., Mackill D. J., 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 363 (1491), 557–572.

Courtois B., Audebert A., Dardou A., Roques S., Ghneim-Herrera T., Droc G., Frouin J., Rouan L., Gozé E., Kilian A., Ahmadi N., Dingkuhn M., 2013. Genome-wide association mapping of root traits in a japonica rice panel. PLOS One 8(11), e78037.

von Cruz M., Kilian A., Dierig D.A., 2013. Development of DArT marker platforms and genetic diversity assessment of the U.S. collection of the new oil seed crop Lesquerella and related Species. PLOS One 8(5), e64062.

DaCosta J.M., Sorenson M.D., 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PLOS One 9(9), e106713.

Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M., Blaxter M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. 12(7), 499–510.

Davik J., Sargent D.J., Brurberg M.B., Lien S., Kent M., Alsheikh M., 2015. A ddRAD based linkage map of the cultivated strawberry, *Fragaria* × *ananassa*. PLOS One 10(9), e0137746.

Dracatos P.M., Haghdoust R., Singh D., Park R.F., 2017. Genetic analysis and molecular mapping of resistance to *Puccinia striiformis* f. sp. *pseudo-hordei* in common wheat. Plant Pathol. 66, 285–292.

Dracatos P.M., Zhang P., Park R.F., McIntosh R.A., Wellings C.R., 2016. Complementary resistance genes in wheat selection ‘Avocet R’ confer resistance to stripe rust. Theor. Appl. Genet. 129, 65–76.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. 2011. A robust, Simple Genotyping-by-Sequencing (GBS) approach for high diversity species. PLOS One 6(5), e19379.

Etter P.D., Bassham S., Hohenlohe P.A., Johnson E.A., Cresko W.A., 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol. Biol. 772, 157–178.

Fan J.B., Chee M.S., Gunderson K.L., 2006. Highly parallel genomic assays. Nat. Rev. Genet. 7, 632–644.

Gasc C., Peyretaillade E., Peyret P., 2016. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and non-model organisms. *Nucleic Acids Res.* 44 (10), 4504–4518.

Gore M.A., Chia J.M., Elshire R.J., Sun Q., Ersoz E.S., Hurwitz B.L., Peiffer J.A., McMullen M.D., Grills G.S., Ross-Ibarra J., Ware D.H., Buckler E.S., 2009. A first-generation haplotype map of maize. Science 326 (5956), 1115–1117.

Guo-Qian Y., Chen Y.M., Wang J.P., Guo C., Li L.L., Li D.Z., Guo Z.H., 2016. Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. Plant Methods 12, 39.

Habyarimana E., Parisi B., Mandolino G., 2017. Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (*Solanum tuberosum* L.). Plant Breed. 136, 245–252.

He J., Zhao X., Laroche A., Lu Z.X., Liu H., Li Z., 2014. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front *Plant Sci. 5* (484), 1–6.

Hedges D.J., Guettouche T., Yang S., Bademci G., Diaz A., Andersen A., Hulme W.F., Linker S., Mehta A., Edwards Y.J.K., Beecham G.W., Martin E.R., Pericak-Vance M.A., Zuchner S., Vance J.M., Gilbert J.R., 2011. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. PLOS One 6(4), e18595.

Heslot N., Rutkoski J., Poland J., Jannink J.L., Sorrells M.E., 2013. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLOS One 8(9), e74612.

Hillier L.W., Marth G.T., Quinlan A.R., Dooling D., Fewell G., Barnett D., Fox P., Glasscock J.I., Hickenbotham M., Huang W., Magrini V.J., Richt R.J., Sander S.N., Stewart D.A., Stromberg M., Tsung E.F., Wylie T., Schedl T., Wilson R.K., Mardis E.R., 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat. Methods* 5 (2), 183–188.

Huang X., Wei X., Sang T., Zhao Q., Feng Q., Zhao Y., Jing W., Li W., Lin Z., Buckler E.S., Qian Q., Zhang Q-F., Li J., Han B., 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42, 961–967.

Huang, X., Feng Q., Qian Q., Zhao Q., Wang L., Wang A., Guan J., Fan D., Weng Q., Huang T., Dong G., Sang T., Han B., 2009. High-throughput genotyping by whole-genome resequencing. Genome Res. 19(6), 1068–1076.

Jaccoud D., Peng K., Feinstein D., Kilian A., 2001. Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res. 29(4), 25.

Kenta S., Hideki H., Sachiko I. 2016. Analytical workflow of double-digest Restriction site – associated DNA sequencing based on empirical and in silico optimization in tomato. DNA Res. 23(2), 145–153.

Kiialainen A., Karlberg O., Ahlford A., Sigurdsson S., Lindblad-Toh K., Syvänen A.C., 2011. Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. PLOS One 6(2), e16486.

Konar A., Choudhury O., Bullis R., Fiedler L., Kruser J.M., Stephens M.T., Gailing O., Schlarbaum S., Coggeshall M.V., Staton M.E., Carlson J.E., Emrich S., Severson J.R., 2017. High-quality genetics mapping with ddRADseq in the non-model tree *Quercus rubra*. Genomics 18, 417.

Kotowska M., Zakrzewska-Czerwińska J., 2010. Kurs szybkiego czytania DNA – nowoczesne techniki sekwencjonowania [Fast DNA reading course – modern sequencing techniques]. Biotechnologia 4(91), 24–38.

Mardis E., McCombie W.R., 2017. Agarose gel size selection for DNA sequencing libraries. Cold Spring Harbor Protocols (8).

Mascher M., Richmond T.A., Gerhardt D.J., Himmelbach A., Clissold L., Sampath D., Ayling S., Steuernagel B., Pfeifer M., D'Ascenzo M., Akhunov E.D., Hedley P.E., Gonzales A.M., Morrell P.L., Kilian B., Blattner F.R., Scholz U., Mayer K.FX., Flavell A.J., Muehlbauer G.J., Waugh R., Jeddeloh J.A., Stein N., 2013. Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant* J. 76(3), 494–505.

Milczarski P., Hanek M., Tyrka M., Stojałowski S., 2016. The application of GBS markers for extending the dense genetic map of rye (*Secale cereale* L.) and the localization of the *Rfc1* gene restoring male fertility in plants with the C source of sterility-inducing cytoplasm. J. Appl. Genet. 57, 439–451.

Myllykangas S., Natsoulis G., Bell J.M., Ji H.P., 2011. Targeted sequencing library preparation by genomic DNA circularization. *BMC* Biotechnol. 11, 122.

Ng S.B., Turner E.H., Robertson P.D., Flygare S.D., Bigham A.W., Lee C., Shaffer T., Wong M., Bhattacharjee A., Eichler E.E., Bamshad M., Nickerson D.A., Shendure J., 2009. Targeted capture and massively parallel sequencing of twelve human exomes. Nature 461 (7261), 272–276.

Pachota K., Niedziela A., Orłowska R., Bednarek P.T., 2016. Nowoczesne metody genotypowania DArT i GBS w hodowli gatunków roślin użytkowych [Modern methods of DArT and GBS genotyping in the cultivation of utility plant species]. Biul. IHAR 279.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E., 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLOS One 7(5), e37135.

Poland J.A., Brown P.J., Sorreils M.E., Jannink J., 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLOS One 7(2), e32253.

Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P., Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341.

Qiu X., Pang Y., Yuan Z., Xing D., Xu J., Dingkuhn M., Li Z., Ye G., 2015. Genome-wide association study of grain appearance and milling quality in a worldwide collection of indica rice germplasm. PLOS One 10(12), e0145577.

Roldan-Ruiz I., Dendauw J., Van Bockstaele E., Depicker A., De Loose M., 2000. AFLP markers reveal high polymorphic rates in ryegrasses (*Lolium* spp.). Mol. *Breeding* 6, 125–134.

Sánchez-Sevilla J.F., Horvath A., Botella M.A., Gaston A., Folta K., Kilian A., Denoyes B., Amaya I., 2015. Diversity Arrays Technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria* × *ananassa*). PLOS One 10(12), e0144960.

Sansaloni C., Petroli C., Jaccoud D., Carling J., Detering F., Grattapaglia D., Kilian A., 2011. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. BMC Proceedings 5 (Suppl. 7), 54.

Sonah H., Bastien M., Iquira E., Tardivel A., Légaré G., Boyle B., Normandeau E., Laroche J., Larose S., Jean M., Belzile F., 2013. PLOS One 8(1), e54603.

Song K., Ren J., Zhai Z., Liu X., Deng M., Sun F., 2013. Alignment-free sequence comparison based on next-generation sequencing reads. J. Comput. Biol. 20(2), 64–79.

Van Tassell, C.P., Smith T.P., Matukumalli L.K., Taylor J.F., Schnabel R.D., Lawley C.T., Haundenschild C.D., Moore S.S., Warren W.C., Sonstegard T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5(3), 247–252.

Tyrka M., Tyrka D., Wędzony M., 2015. Genetic map of *Triticale* integrating microsatellite, DArT and SNP markers. PLOS One 10(12), e0145714.

Valdisser P.A.M.R., Pereira W.J., Filho J.E.A., Müller B.S.F., Coelho G.R.C, de Menezes I.P.P., Vianna J.P.G., Zucchi M.I., Lanna A.C., Coelho A.S.G., de Oliveira J.P., da Cunha Moraes A., Brondani C., Vianello R.P., 2017. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. BMC Genomics 18, 423.

Wang S., Meyer E., McKay J.K., Matz M.V., 2012. 2b-RAD a simple and flexible method for genome-wide genotyping. Nat. Methods 9(8), 808–810.

Wu Z., Wang B., Chen X., Wu J., King G.J., Xiao Y., Liu K., 2016. Evaluation of linkage disequilibrium pattern and association study on seed oil content in *Brassica napus* using ddRAD sequencing. PLOS One 11(1), 1–15.

Xie, W., Feng Q., Yu H., Huang X., Zhao Q., Xing Y., Yu S., Han B., Zhang Q., 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc. Natl. Acad. Sci. USA. 107(23), 10578–10583.

https://www.diversityarrays.com/ [access 20.09.2018].

https://www.illumina.com/systems/sequencing-platforms/miseq.html [access 20.09.2018].

http://www.sagescience.com/wpcontent/uploads/2012/11/sage_wp_saygoodbyetomanualgels_0912_6.pdf [access 20.09.2018].

The article has been implemented within the framework of Multiannual Programme „Creating the scientific basis of the biological progress and conservation of plant genetic resources as a source of innovation to support sustainable agriculture and food security of the country” coordinated by IHAR and financed by the Polish Ministry of Agriculture and Rural Development.

Streszczenie. Od czasu opublikowania pełnej sekwencji genomu *Arabidopsis thaliana* w 2000 r. [AGI Initiative 2000] rozpoczął się okres dynamicznej eksploracji genomów. W ostatniej dekadzie, wraz z rewolucją w technologii sekwencjonowania nowej generacji, lawinowo wzrosła ilość doniesień naukowych opartych na analizie sekwencji. Nowe, szybkie, wysokoprzepustowe i relatywnie tanie technologie sekwencjonowania kwasów nukleinowych stały się dostępne i powszechne, otwierając możliwość szerokiego wykorzystania narzędzi molekularnych w nauce i praktyce hodowlanej. Te nowe metody obejmują sekwencjonowanie pełnogenomowe oraz wiele metod sekwencjonowania redukowanej frakcji genomu (RRS, ang. r*educed representation sequencing*). Wielość dostępnych metod, zróżnicowanych pod względem przystępności i kosztów, generujących różne rodzaje danych wynikowych, dedykowanych różnym celom badawczym i aplikacyjnym, może sprawić trudność przy wyborze najlepszego wariantu [Poland et al. 2012]. Celem niniejszego opracowania jest przybliżenie czytelnikowi możliwości i zastosowania wybranych protokołów konstrukcji bibliotek do sekwencjonowania zredukowanej frakcji genomu.

Słowa kluczowe: sekwencjonowanie następnej generacji (NGS), metody frakcjonowania genomu, biblioteki DNA, RADSeq, ddRADSeq, GBS, DArTSeq

Received: 16.10.2018
Accepted: 12.12.2018