

RERSPECTIVE

Design of efficient simplified genomic DNA and bisulfite sequencing in large plant populations

Jinhua Wu, Zewei Luo and Ning Jiang*

Institute of Biostatistics, Fudan University, Shanghai 200438, China

* Correspondence: ningjiang@fudan.edu.cn

Received March 26, 2016; Revised June 10, 2016; Accepted June 13, 2016

The next generation sequencing enables generation of high resolution and high throughput data for structure sequence of any genome at a fast declining cost. This opens opportunity for population based genetic and genomic analyses. In many applications, whole genome sequencing or re-sequencing is unnecessary or prohibited by budget limits. The Reduced Representation Genome Sequencing (RRGS), which sequences only a small proportion of the genome of interest, has been proposed to deal with the situations. Several forms of RRGS are proposed and implemented in the literature. When applied to plant or crop species, the current RRGS protocols shared a key drawback that a significantly high proportion (up to 60%) of sequence reads to be generated may be of non-genomic origin but attributed to chloroplast DNA or rRNA genes, leaving an exceptional low efficiency of the sequencing experiment. We recommended and discussed here the design of optimized simplified genomic DNA and bisulfite sequencing strategies, which may greatly improves efficiency of the sequencing experiments by bringing down the presentation of the undesirable sequencing reads to less than 10% in the whole sequence reads. The optimized RAD-seq and RRBS-seq methods are potentially useful for sequence variant screening and genotyping in large plant/crop populations.

Keywords: plant/crop genomes; next generation sequencing; genotyping; restriction-enzyme sites associated DNA (RAD); DNA methylation; reduced representation bisulfite sequencing

INTRODUCTION

One of the most popular genomics analyses involves identifying genetic and epigenetic variants and genotyping the variants genome-wide in large populations of interest as has been widely implemented to phenotype prediction, gene mapping and isolation, assessment of population diversity in genetics, ecological and evolutionary biology [1–3]. This is made practical in fast developing techniques of sequencing technology during the past decades. Supplementary Table 1 summarizes the major technical features of three generations of genome sequencing variant assay [4].

The Sanger sequencing technique, which was developed in 1977 and based on the principle of selective incorporation of chain-terminating dideoxynucleosides

by DNA polymerase [5], has dominated the DNA sequencing for about 25 years. The major limitation of the Sanger sequencing in throughput and efficiency urges emerging new alternatives. Microarray-based DNA sequencing technique represents the first “poster-Sanger” method for simultaneously profiling the DNA sequencing polymorphisms at genome-wide candidate loci. It has greatly facilitated genome-wide association studies (GWAS) and QTL mapping [6] through accurately genotyping a large population at a high-density set of discontinuous genomic coordinates. The microarray technique requires design of probes as tiling representation of the reference sequence of the genome of interest [7]. There are two key commercial producers providing microarray chips that satisfy different needs of SNP detecting and genotyping in plants or crops species (Supplementary Table 1), for example, Affymetrix has developed chips for crop species including barley, cotton, lettuce, maize, pepper, rice, rose, soybean, strawberry,

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

wheat etc., and Illumina developed chips for maize etc. Although the microarray technology enables the efficient genotyping of large populations in parallel at tens of thousands of SNPs, its relying on the prior knowledge of genomic sequence of the organism under question to design probes has seriously restricted its application to the species whose genomic sequence and polymorphic sequence are yet properly annotated.

Techniques of the next-generation sequencing (NGS) matured in 2005 and have become more and more popular experimental tool for genomic, transcriptomic and epigenetic analyses in plant and crop species [8,9] without need of prior information of the genome sequence. The techniques confer efficient re-sequencing of the genome, transcriptome and epigenome of any diploid species, with high throughput polymorphism detecting and genotyping as a typical example of utilities of the NGS techniques in addition to their many other uses. The whole genome re-sequencing has been implemented in various plant species including *Arabidopsis*, potato (*solanum tuberosum*), rice and maize [10–13] and the cost of the re-sequencing analysis has continuously declined over the past decade. But it is still infeasible and highly costly to implement whole genome DNA sequencing for genotyping large populations of most plant and crop species, which generally have a large genome size, from many millions to even billions of base pairs [14,15]. This report focuses on design and implement of optimized protocols of sequencing based genetic and epigenetic detecting and genotyping in large populations of plant and crop species.

REDUCED-REPRESENTATION SEQUENCING STRATEGIES FOR GENOTYPING LARGE PLANT POPULATIONS

Basic idea behind sequencing based genotyping large populations is reduced representation of genome to be sequenced, addressing the problem of cost limitation of the whole genome sequencing. Supplementary Table 2 summarizes features of 3 different strategies of reduced representation sequencing.

The first “reduced-representation” sequencing approach is to re-sequence transcriptome rather than genome to identify the genetic polymorphisms only in transcribed genomic regions. Compared to the large size of genome in eukaryotic species, the size of transcriptome is significantly reduced. Thus, amount of short-read data required and associated cost can be dramatically reduced. In the past five years, RNA-sequencing has been implemented to identify tens to thousands of genetic variants in human, cottonwood and potato genomes [16–18]. Because expression of genes in different samples can

vary widely and change dynamically, the RNA-seq based approach bears several key inherent weaknesses. For example, the allelic imbalances in RNA-seq data can significantly affect the analysis of the nuclear genotype [19,20]. Furthermore, only the genetic polymorphisms in commonly expressed regions can be identified in most of the sequenced samples. For those locating in tissue-specific expressed regions (or development stage specific expressed regions), they can be detected only in those few specific samples. In 2011, Christodoulou *et al.* introduced a method that tried to enable efficient sequencing of low-abundance RNAs and to decrease the proportion of reads from highly expressed RNAs. However, normalizing range of dynamic expression in NGS is not an easy task [21].

To ease the problem caused by differential expression of genes in transcriptome, the hybrid capture sequencing approach was proposed to extract a small part of whole genomic DNA [22]. If nucleotide sequence information for regions of interest is known, hybrid-capture probes are designed to bind to regions of interest and these regions can be directly isolated for sequencing [23,24]. Recently, Uitdewilligen *et al.* (2013) designed 57,054 oligonucleotide probes to capture DNA fragments for 807 potato genes and inferred sample genotypes from the sequence data [25]. An obvious technical hurdle to this hybrid capture sequencing strategy is requirement of genomic sequence information to enable design of a large number of unique capturing probes for surrogating the predefined genome regions. The capture sequencing strategy could receive highly accurate genotyping result, but may be inappropriate when population under question is considerably divergent from the probe designing reference. The probes thus designed may poorly bind to the regions of interest and result in severe biases in capturing the target regions [2]. Furthermore, the DNA may not be evenly presented by the targeted genomic regions. For instance, the extracted DNA is generally enriched in targeted regions with high GC content, due to the more stable binding [26].

From 2008, a new “reduced-representation” sequencing strategy referred as “RAD-seq” has been widely proposed for genotyping large population in several animal and plant species [27–31]. This RAD-seq strategy mainly combines genotyping-by-sequencing with restriction-enzyme sites associated DNA fragments (RAD) to make this goal considerably more time and cost-effective in comparison to conventional platforms such as microarray based genotyping technologies, RNA-seq and capture sequencing. It can identify and score hundreds to thousands of genetic markers randomly distributed across the interested genome, from a group of sequenced samples. Although there are several modified forms of the original RAD-seq method (Table 1), they still share some

common and significant drawbacks. For example, DNA samples for genetic or genomic analyses are usually collected from leaves to ensure gain of both the integrity and quantity of DNA extracted. However, the current RAD-seq methods ignored a fundamental fact that there are a large copy number (1,000~10,000) of chloroplast sequence and rRNA genes in plant leaf cells [31–34]. Without properly dealing with the issue, the chloroplast DNA and/or rRNA genes may occupy a substantially large proportion (up to 60%) of sequence reads obtained from the standard RAD-seq protocol and their modified versions. This will, in turn, significantly dilute coverage of the sequence reads across the genome under question. To address this issue, we described here an optimized RAD approach whose comparison to other versions of RAD-seq was summarized in Table 1.

Table 1 summarized the type of enzyme, the number of restriction enzymes, need of segment size selection, removal of DNA of chloroplast and RNA genes, multiplex level of sample pooling for four main representatives of the RAD-seq methods and the optimized RAD-seq protocol described here. The optimized RAD-seq method integrated bioinformatic analysis and a computer guided sequencing library construction design in order to maximize uniformity and coverage of sequence reads across the genome under question and to minimize presentation of chloroplast and rRNA DNA in the sequence reads.

AN OPTIMIZED RAD-SEQ STRATEGY FOR GENOTYPING LARGE PLANT POPULATIONS

Based on re-analysis of the current RAD-seq experiments summarized, we observed that a large proportion (30%~60%) of obtained short sequence reads were mapped to chloroplast sequence and rRNA genes [32–34]. And, this represents a serious waste of high-throughput sequencing capacity and resources since these reads are typically directly discarded for downstream analysis. Furthermore, it is crucial to achieve high recovery of DNA fragments within a specified size range during the size-selection

step. However, existing RAD-seq approaches have generally implemented traditional manual gel purification to extract the targeted DNA fragment. This method is obviously superficial, has low reproducibility, and can introduce low molecular weight contamination [35]. In 2012, Peterson et al. introduced the use of Pippin-Prep automated size-selection technology to precisely extract the targeted DNA fragments [36]. Furthermore, Quail *et al.* introduced the double size selection strategy for NGS library construction which could give tighter and more uniform size distribution for selected DNA fragments among the desired size ranges [37]. We described here how this size selection may be effectively implemented in the optimized RAD-seq approach.

The optimization was made to maximize the number of genomic DNA sequence reads that uniformly cover a plant genome under study and to minimize the number of sequence reads representing chloroplast DNA and rRNA genes. The optimized approach provides both a generic bioinformatics tool and a library construction protocol. The bioinformatics pipeline was developed in a user friendly manner to work out the cutting sites from different combinations of restriction enzymes (REs) from 269 possible REs across a given plant genome of interest, and then to determine the optimal combination of REs. The simulation based bioinformatic analysis requires sequence information of the genome and information of the cutting sites of all 269 possible REs. The 4 popular REs (*EcoR* I, *Hind* III, *Msp* I and *Mse* I) could be selected for the first round of *in silico* digestion to make the optimal combination of REs that satisfies the following criteria: (i) recovery of the largest possible number of genomic regions within the target size range required by Illumina sequencing (224 – 424 bp). This number should be at least as large as the predefined number based on consideration of the experimental objectives, including the coverage required per sample, the number of samples to be sequenced, and the density of DNA polymorphisms to be targeted; and (ii) minimization of the number of DNA fragments recovered from chloroplast sequence and rRNA genes. For the second round of digestion, we employed the 269 commercially available REs (<http://insilico.ehu.es/restriction/main/index2.php>) to achieve

Table 1. Basic features of different RAD-seq methods when applied to genotype plant populations.

Methods	Enzyme used	Number of REs*	Size-selection	Removal of chloroplast and/or RNA DNAs	Multiplex level	Ref.
Original RAD-seq	Type II	1	Yes	No	96	[27]
ddRAD-seq	Type II	2	Yes	No	96	[28]
2b-RAD-seq	Type IIB	1	Needless	No	1	[32]
l2b-RAD-seq	Type IIB	1	Needless	No	12	[31]
Optimized RAD-seq	Type II	≤3	Yes	Yes	12	The present

* The number of restriction enzymes

the optimal combination of REs for completely removing the DNA fragments from chloroplast sequence and rRNA genes while keeping the maximum number of selected genomic DNA fragments almost intact. Thus, these two rounds of RE digestion are designed to not only cut genomic DNA sequence into pre-designed fragments but also to remove the large proportion of DNA fragments from chloroplast sequence and rRNA genes during RAD-seq library construction. Here, bioinformatics tools and the corresponding computer programs were generic in design and compilation, enabling their use in the design of RAD-seq experiments without needing specialized bioinformatics or computational skills. Experimentally, we proposed here a double size-selection strategy using the Pippin Prep automated size selection technology, in order to reproducibly extract DNA fragments with a pre-defined size range from the bioinformatic prediction, and at the same time, to enable removal of dimers from the RAD-seq libraries [35,37].

The workflow of library construction for the optimized RAD-seq protocol was shown in Figure 1. In brief, it consists of the following 6 steps: (i) first round of digestion to cut the DNA sequence of each sample; (ii) ligating the barcoded adapters to RE cut sites; (iii) pooling equimolar amounts of DNA fragments for each sample and carrying out precise size-selection using the Pippin Prep automated size selection platform; (iv) second round of digestion to remove the selected DNA fragments from chloroplast sequence and rRNA genes; (v) PCR amplification using Illumina primers for pooled samples; (vi) second round of size-selection using Pippin Prep to ensure high recovery of DNA fragments in the required size range and remove adapter dimers.

To experimentally validate the optimized RAD-seq approach, we have constructed several pooled sequencing libraries from leave tissues from parental lines and their segregating offspring of both diploid and tetraploid *Arabidopsis* and potato. The sequence data analysis shows that the optimized RAD-seq approach designed for the *Arabidopsis* and potato genomes can effectively remove DNA fragments derived from chloroplast sequence and rRNA genes, and the short reads collected are mostly concentrated onto the targeted genomic regions. A balanced representation of sequence reads was obtained from across all pooled samples. Generally, after the second round of digestion, proportion of reads mapped to chloroplast and rRNA genes could significantly reduce from 60% to less than 10%. Meanwhile, these short-reads could precisely and consistently map to the selected genomic regions among different samples. These features demonstrate the robustness and efficiency of the optimized RAD-seq approach developed here and further indicate that one can feasibly design and effectively implement the protocol to achieve expected

coverage of polymorphic sequence markers for large plant populations given information of the genome sequence and ideally, though not necessarily, information of the sequence polymorphism distribution in the genome.

RAD BASED BISULFITE SEQUENCING FOR PLANT SPECIES

DNA methylation plays a fundamental role in the regulation of gene expression and is widespread in the genome of eukaryotic species. Currently, bisulfite sequencing (BS) approach is the sole and standard method that can accurately measure the DNA methylation at both single-base resolution at a genome scale [38]. BS method uses bisulfite treatment in combination with high-throughput sequencing to draw the most complete picture of a DNA methylome. It has been successfully applied to elucidate the methylomes of human cells as well as of other species such as mouse, *Arabidopsis*, maize [39–42]. But, as the same as whole genome sequencing, the genome wide BS still bears the high cost and need for exceptional depth of sequencing, and this hinders BS for profiling large populations. In 2005, Meissner *et al.* firstly proposed “reduced representation bisulfite sequencing” (RRBS) strategy for covering only 1% of the mouse genome [43]. This approach basically combines “*Bgl* II” restriction-enzyme sites associated DNA fragments (RAD) with the standard BS platform to enrich for the regions of genome that have a high CpG content. And, this RRBS approach has been widely applied to the DNA methylation studies for several mammal species such as human and mouse genomes [43–46].

The idea of the optimized RAD-seq method aforementioned can be extended to develop a RRBS method specific to plant and crop species. Basically, the plant/crop specific RRBS method is proposed to minimize presentation of chloroplast and rRNA DNA in the sequencing library and, in turn, in sequence reads finally generated. Figure 2 illustrates diagrammatic procedure of a plant/crop specific RRBS experiment with *Arabidopsis* as an experimental model. Compared to the optimized RAD-seq experiments, the optimized RRBS experiment differs in the following aspects.

Firstly, in the first round of digestion of the RAD-seq, REs were selected from *EcoR* I, *Hind* III, *Msp* I and *Mse* I. In the RRSB, we used 9 methylation insensitive restriction enzymes in the simulation prediction of the cutting sites across the genome under study so as to determine the optimal combination of REs in the RRBS-seq experiment (Table 2). The first 5 are 6 bp or 5 bp REs and can identify “cytosine” sites in any plant genome. This ensures at least one “CpG”, “CHH” or “CHG” site in every selected DNA fragment. The remaining 4 except for “*Mse* I” are 4 bp REs and can identify “CpG” sites. Thus,

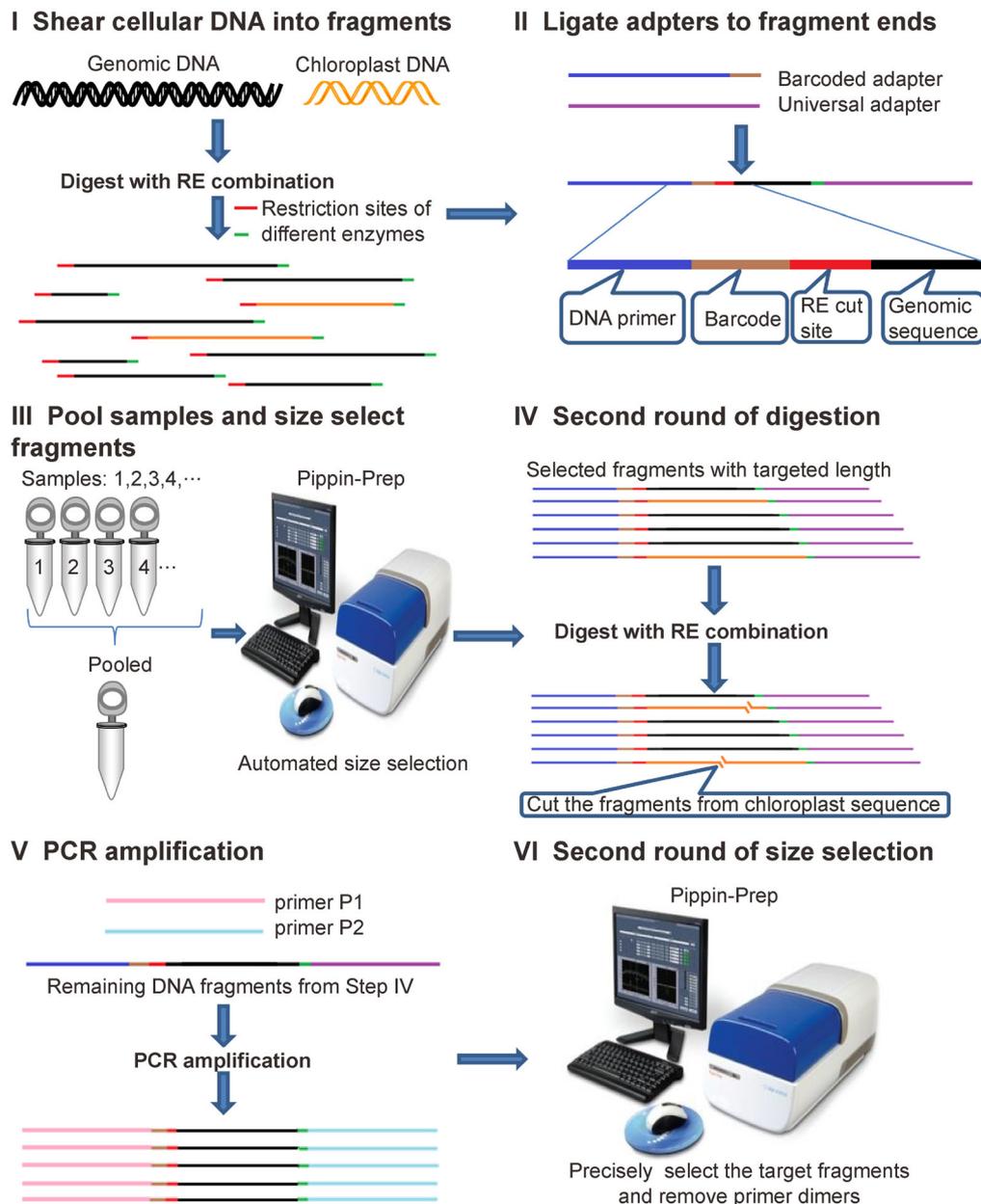


Figure 1. Workflow of the optimized RAD-seq library construction.

use of these REs ensures the selected DNA fragments could detect at least one “CpG” site. Thus, use of the above the 9 REs guarantees 2 candidate methylation sites (“CpG”, “CHH” or “CHG”) in every DNA fragment selected for sequencing library construction. The numbers of produced DNA fragments varied markedly over different REs’ combinations when applied to shear the *Arabidopsis* genome (Table 3). The current Illumina HiSeq sequencer generates short reads with length varying

among 2×100 bp, 2×125 bp and 2×150 bp. Thus, we determined the size range of selected DNA fragments from 224 bp and 424 bp. The numbers of selected DNA fragments for different REs’ combinations were shown in Table 3. For the RRBS-seq experimental design, we will be able to detect the methylation events from about 20,000 small DNA regions evenly distributed across the whole *Arabidopsis* genome. To reach this objective, we selected 4 candidate combinations of Res for the first

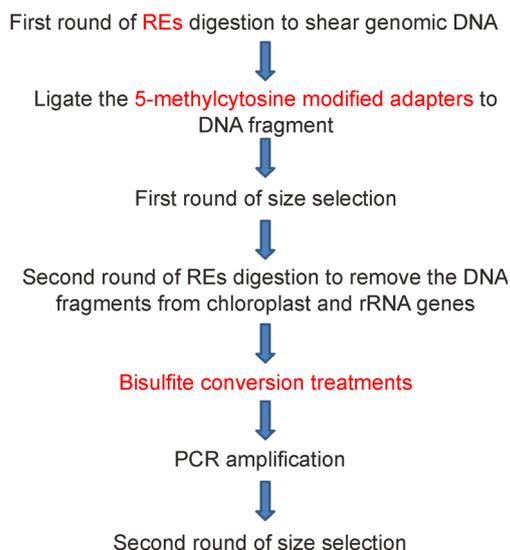


Figure 2. A diagrammatic illustration of an optimized plant/crop specific RRBS experiment.

Table 2. Nine restriction enzymes selected for shearing the *Arabidopsis* genome.

Restriction enzyme	Cutting site
<i>Bsa</i> I	C C ¹ NNGG
<i>Cac</i> 8 I	GCN NGC
<i>Fnu</i> 4H I	GC NGC
<i>Mfe</i> I	C AATTG
<i>Dde</i> I	C TNAG
<i>Hha</i> I	GCG C
<i>Msp</i> I	C CGG
<i>Taq</i> I	T CGA
<i>Mse</i> I	T TAA

¹ indicates the cutting site

N = A, C, G or T

round of digestion to shear the *Arabidopsis* genome: “*Cac8* I/*Taq* I”, “*Fnu4H* I/*Taq* I”, “*Bsa* I/*Cac8* I/*Taq* I” and “*Bsa* I/*Dde* I/*Mse* I”. For these 4 different REs’ combinations, each of REs’ combinations could produce enough genomic DNA fragments (> 30,000) with length from 224 bp to 424 bp while the number of selected DNA fragments from chloroplast sequence and rRNA genes was controlled at a relatively low level (Table 3).

From the selected 4 RE combinations, we found that there were 40–53 DNA fragments from the rRNA genes and chloroplast sequence in the selected genomic DNA fragments. Due to an extremely high copy number of rRNA genes and chloroplast sequence, a very large proportion of reads could be generated for these undesirable segments from the RRBS-seq sequencing (Table 4A). In order to remove these undesirables, we

searched the cutting sites of 269 different restriction enzymes in the selected DNA fragments for the optimal combination to minimize the DNA fragments representing the chloroplast sequence and rRNA genes at the same time to keep the number of the selected genomic DNA fragments as much intact as possible after the second round of digestion (Table 5).

The analysis presented above recommends use of the “*Bsa* I/*Dde* I/*Mse* I” combination for the first round digestion for shearing *Arabidopsis* genome and the “*Xmn* I/*Pct* I/*Tfi* I” combination in the second round digesting to remove the DNA fragments representing the chloroplast sequence and rRNA genes. After the two rounds of digestion, there were about 15,011 genomic DNA fragments remained as sequencing templates but 2 undesirable fragments from the chloroplast sequence. Here, we carried out numerical analysis to work out proportion of short-reads that might be mapped on the chloroplast genome. It is clear from Table 4B that about 93% and 96% of sequence reads generated from the design would be mapped to the genomic DNA in the *Arabidopsis* diploid and tetraploid species respectively, leaving only 7% and 4% of the undesirable sequence reads for the two species.

Furthermore, the 15,011 selected genomic DNA fragments would cover 3.7 Mb genomic regions (based on 2×125 bp paired-ends sequencing strategy). For size of the *Arabidopsis* genome (119 M), the sequence reads account for 3.1% (3.7/119) of the whole genome. Furthermore, we explored distribution of detectable “candidate methylation sites” among 15,011 selected genomic DNA fragments. Table 6 indicates the number of detectable “candidate methylation sites” in each of the *Arabidopsis* chromosome and shows there would be a total of 718,904 candidate methylation sites to be detected through the RRBS-seq experiment designed. Compared to the total 21,468,778 candidate methylation sites in the *Arabidopsis* genome, 3.4% (718,904/21,468,778) of the whole candidate methylation sites would be targeted in the RRBS experiment as summarized in Figure 3, which shows that the detected “candidate methylation sites” could randomly and high-density disperse among the whole *Arabidopsis* genomic sequences.

Secondly, the adapter oligonucleotides in the optimized RRBS have all cytosines (C nucleotides) replaced by 5’ methyl-cytosines, so as to prevent deamination of these cytosines during the bisulfite conversion reaction.

Thirdly, in bisulfite conversion reaction, the unmethylated cytosine is deaminated into a uracil in the selected DNA fragments. Here, the bisulfite conversion efficiency is extremely important for the RRBS experiments. With a low efficiency of the bisulfite conversion, biases may be introduced into the methylation profiles. In the optimized RRBS experiment, we used the commercial “EZ DNA

Table 3. Numbers of total DNA fragments and number of selected DNA fragments for different REs combinations.

REs' combination	In genomic sequences		In chloroplast sequence and rRNA	
	Number of DNA fragment	Number of selected fragments*	Number of DNA fragment	Number of selected fragments*
<i>BsaJ</i> I/ <i>Msp</i> I	115,738	20,731	245	39
<i>BsaJ</i> I/ <i>Taq</i> I	173,542	34,145	343	52
<i>BsaJ</i> I/ <i>Hha</i> I	103,498	18,063	217	41
<i>Cac8</i> I/ <i>Msp</i> I	104,346	18,883	161	37
<i>Cac8</i> I /<i>Taq</i> I	151,872	31,836	164	40
<i>Cac8</i> I / <i>Hha</i> I	93,950	16,452	148	33
<i>Fnu4H</i> I / <i>Msp</i> I	128,425	22,406	183	35
<i>Fnu4H</i> I /<i>Taq</i> I	191,648	37,209	210	51
<i>Fnu4H</i> I / <i>Hha</i> I	104,259	17,757	154	22
<i>BsaJ</i> I/ <i>Cac8</i> I/ <i>Msp</i> I	146,683	26,954	268	35
<i>BsaJ</i> I/ <i>Fnu4H</i> I/ <i>Msp</i> I	158,875	27,135	285	40
<i>Cac8</i> I/ <i>Fnu4H</i> I/ <i>Msp</i> I	153,767	27,272	236	52
<i>BsaJ</i> I/<i>Cac8</i> I/<i>Taq</i> I	254,438	46,922	419	53
<i>BsaJ</i> I/ <i>Fnu4H</i> I/ <i>Taq</i> I	277,133	47,741	432	54
<i>Cac8</i> I/ <i>Fnu4H</i> I/ <i>Taq</i> I	261,726	47,568	306	58
<i>BsaJ</i> I/ <i>Cac8</i> I/ <i>Hha</i> I	126,993	23,179	233	46
<i>BsaJ</i> I/ <i>Fnu4H</i> I/ <i>Hha</i> I	126,468	22,111	236	43
<i>Cac8</i> I/ <i>Fnu4H</i> I/ <i>Hha</i> I	121,329	21,885	190	36
<i>BsaJ</i> I/<i>Dde</i> I/<i>Mse</i> I	513,878	39,830	639	43
<i>BsaJ</i> I/ <i>Dde</i> I/ <i>Taq</i> I	366,534	54,165	609	67
<i>BsaJ</i> I/ <i>Mfe</i> I/ <i>Taq</i> I	204,754	40,832	415	58

* Size range: from 224 bp to 424 bp

Table 4. Prediction of the proportion of sequence reads to be mapped to genomic regions, rRNA genes and chloroplast sequence in the *Arabidopsis* RRBS-seq experiment.

A) Based on original RRBS-seq approach which only used “*BsaJ* I/*Dde* I/*Mse* I” REs’ combination to shear the reference sequences.

	Diploid <i>Arabidopsis</i>			Tetraploid <i>Arabidopsis</i>		
	Genomic	rRNA	Chloroplast	Genomic	rRNA	Chloroplast
Predicted number of selected DNA fragments per haploid genome	39,830	2	41	39,830	2	41
Copy number	2	1,400	1,200	4	2,800	1,200
Predicted number of selected DNA fragments per cell	79,660	2,800	49,200	159,320	5,600	49,200
Total selected cellular DNA fragments per cell		131,660			214,120	
Theoretical proportion of reads mapped to different regions	60%	2%	38%	74%	3%	23%

B) Based on our optimal RAD-seq approach which used the “*Xmn* I/*Pct* I/*Tfi* I” REs’ combination to remove chloroplast and rRNA derived fragments for the second round of digestion.

	Diploid <i>Arabidopsis</i>			Tetraploid <i>Arabidopsis</i>		
	Genomic	rRNA	Chloroplast	Genomic	rRNA	Chloroplast
Predicted number of selected DNA fragments per haploid genome	15,011	0	2	15,011	0	2
Copy number	2	2×700	1,200	4	4×700	1,200

(Continued)

	Diploid <i>Arabidopsis</i>			Tetraploid <i>Arabidopsis</i>		
	Genomic	rRNA	Chloroplast	Genomic	rRNA	Chloroplast
Predicted number of selected DNA fragments per cell	30,022	0	2,400	60,044	0	2,400
Total selected cellular DNA fragments per cell		32,422			62,444	
Theoretical proportion of reads mapped to different regions	93%	0%	7%	96%	0%	4%

Table 5. DNA fragments representing chloroplast sequence and rRNA genes to be generated from using different combinations of restriction enzymes (Res).

A) Based on the “*Cac8 I/Taq I*” combination.

REs' combination	Cut sites within <i>Arabidopsis</i>			Remained genomic DNA fragments	Remained chloroplast and rRNA fragment
	rRNA fragments	Chloroplast fragments	Genomic DNA fragments		
<i>Afi I/BamH I/CviA II</i>	0	40	25,094	6,702	0
<i>Acc II/EcoR II/BseB I</i>	0	39	21,296	10,500	1
<i>Aco I/Bfa I/MspA1 I</i>	0	38	19,448	12,348	2
<i>Aco I/Bfa I/BseY I</i>	0	37	18,622	13,174	3
<i>Afi I/Pvu II/Ssp I</i>	0	36	15,044	16,752	4
<i>Afi I/BstX I/Ssp I</i>	0	35	14,958	16,838	5
<i>Afi I/Eco81 I/Ssp I</i>	0	34	14,597	17,199	6

B) Based on the “*Fnu4H I/Taq I*” combination.

REs' combination	Cut sites within <i>Arabidopsis</i>			Remained genomic DNA fragments	Remained chloroplast and rRNA fragment
	rRNA fragments	Chloroplast fragments	Genomic DNA fragments		
<i>Aci I/Sdu I/Tfi I</i>	4	47	28,292	8,870	0
<i>Afi I/Xmn I/Tfi I</i>	4	46	26,117	11,045	1
<i>Afi I/BamH I/Tfi I</i>	4	45	25,251	11,911	2
<i>Aci I/BsoB I/BsaJ I</i>	4	44	21,250	15,912	3
<i>Aci I/Apa I/BsaJ I</i>	4	43	20,820	16,342	4
<i>Aci I/Afi I/XmaC I</i>	4	42	20,396	16,766	5
<i>Aci I/BsoB I/MspA1 I</i>	4	41	19,581	17,581	6

C) Based on the “*BsaJ I/Cac8 I/Taq I*” combination.

REs' combination	Cut sites within <i>Arabidopsis</i>			Remained genomic DNA fragments	Remained chloroplast and rRNA fragment
	rRNA fragments	Chloroplast fragments	Genomic DNA fragments		
<i>Aci I/Afa I/Hpy188 III</i>	0	53	40,485	6,384	0
<i>Aci I/Ase I/Hpy188 III</i>	0	52	37,991	8,878	1
<i>Acs I/Rsa I/Tse I</i>	0	51	37,708	9,161	2
<i>Acs I/Bis I/BseB I</i>	0	50	34,111	12,758	3
<i>Aci I/Acs I/Eco81 I</i>	0	49	33,084	13,785	4
<i>Acs I/ScrF I/Tau I</i>	0	48	30,966	15,903	5
<i>Acs I/Nci I/Tau I</i>	0	47	27,871	18,998	6

D) Based on the “*Bsa*I/*Dde*I/*Mse*I” combination.

REs' combination	Cut sites within <i>Arabidopsis</i>			Remained genomic DNA fragments	Remained chloroplast and rRNA fragment
	rRNA fragments	Chloroplast fragments	Genomic DNA fragments		
<i>Tfi</i> I/ <i>Pal</i> I/ <i>Sfu</i> I	2	41	27,230	12,557	0
<i>Tfi</i> I/ <i>Pal</i> I/ <i>Xba</i> I	2	41	27,412	12,375	0
<i>Stu</i> I/ <i>Tfi</i> I/ <i>Pal</i> I	2	40	26,352	13,435	1
<i>Vsp</i> I/ <i>Tfi</i> I/ <i>Pal</i> I	2	40	26,352	13,435	1
<i>Xmn</i> I/<i>Pct</i> I/<i>Tfi</i> I	2	39	24,776	15,011	2

Table 6. The number of detectable “candidate methylation sites” in each of the *Arabidopsis* chromosomes in the optimized RRBS design.

Chromosome	Total C sites	“CpG”	“CHH”	“CHG”
Chr1	181,656	23,351	132,254	26,051
Chr2	121,556	15,637	88,826	17,093
Chr3	146,407	19,064	106,523	20,820
Chr4	111,992	14,602	81,636	15,754
Chr5	157,293	20,541	113,962	22,790
Total	718,904	93,195	523,201	102,508

Table 7. Distribution of the number of sequenced short-reads data allocated to each samples in pooled sequencing library.

Sample ID	Tetraploid pooled samples		
	Read pairs*	Proportion	Expected
F2_1	1.74	12.82%	14.29%
F2_2	1.72	12.71%	14.29%
F2_3	1.99	14.69%	14.29%
F2_4	1.95	14.42%	14.29%
P1	2.99	22.15%	14.29%
P2	2.15	15.84%	14.29%
Undetermined	0.99	7.33%	0.00%
Total	13.53	100.00%	100.00%
Coefficient of variation		22.49%	0.0%

* millions of reads

methylation-gold kit” (Zymo Research, Orange County, CA, USA) to improve efficiency of the bisulfite conversion reaction. Although it has been proposed that the conversion efficiency of “EZ DNA methylation-gold kit” can be reached to over 99%, we tested performance of the kit for its bisulfite conversion efficiency. The fact that chloroplast DNA in any plant is free of methylation provided us with a unique opportunity to achieve direct assessment of the bisulfite conversion efficiency by deliberately remaining a few chloroplast DNA fragments in the sequencing library. The conversion efficiency can thus be assessed by checking conversion states of the chloroplast DNA fragments.

A pilot experiment designed by the optimized RRBS protocol was carried out for 2 autotetraploid *Arabidopsis*

parental lines (ABRC: CS3151 and CS3900) and their 4 F2 offspring. The raw sequenced short-reads data showed an even distribution among these pooled samples: the proportion of reads from each sequenced sample ranged from 12.71% to 22.15% (one sample student *t*-test, *P*-value > 0.05; the coefficient of variation = 22.49%), only about 7.33% of short-reads were not assigned to samples (Table 7). Furthermore, the raw short-reads data were aligned to the *Arabidopsis* reference sequence using bismark [47]. The alignment result clearly showed that over 55% of short-reads could be successfully aligned to the genome sequence, while only a small minority of reads (3%) aligned to the chloroplast sequence and rRNA genes (Figure 4). These results were consistent with our simulated prediction based on the high copy number of

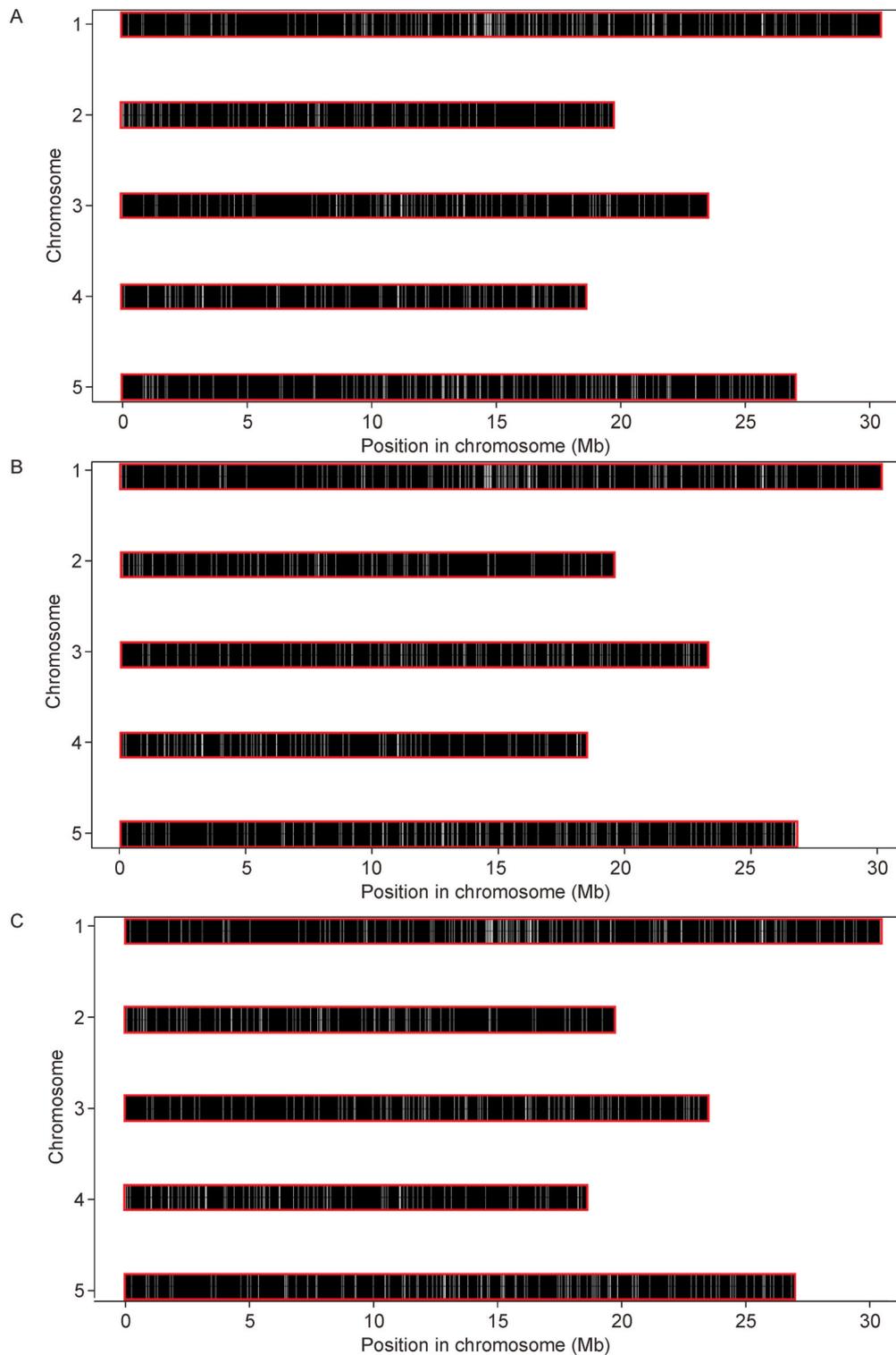


Figure 3. Distribution of "candidate methylation sites" in the *Arabidopsis* genomes. (A) Distribution of detectable "CpG" sites across the *Arabidopsis* genomes. (B) Distribution of detectable "CHH" sites across the *Arabidopsis* genomes. (C) Distribution of detectable "CHG" sites across the *Arabidopsis* genomes.

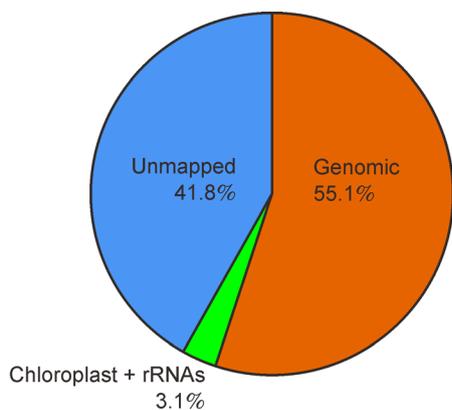


Figure 4. Percentage of RRBS-seq short-reads mapping to different genome regions in *Arabidopsis* reference sequences.

chloroplast and rRNA DNA sequences in *Arabidopsis* leaf cells (Table 4B). Thus, the optimized RRBS-seq strategy developed in the present study have been achieved in effectively removing the reads derived from the chloroplast and rRNA DNA and significantly increasing the proportion of short-reads aligned to the genomic sequence. Additionally, we analyzed the proportion of sequenced fragments that fell within the selected size range (224 bp ~ 424 bp). The length distribution of

the sequenced DNA fragments for the pooled RRBS-seq dataset was shown in Figure 5, which clearly indicates that more than 95% of sequenced fragments are within the target size range.

DISCUSSION

Next generation sequencing (NGS) techniques enhance the power for sequence-based identification and genotyping sequence variants of genetic and epigenetic markers. This opens tremendous research opportunity for understanding genetic and epigenetic basis of complex genetic traits [48,49]. Although the NGS is experiencing a fast decline in costs and provides more robust and informative data, it is still unaffordable for many laboratories to implement the technique for large population genetic and epigenetic analyses. The solution to this problem is the methods so called reduced presentation sequencing such as those described in Gerald *et al.* 2011, Uitdewilligen *et al.* 2013, Baird *et al.* 2008 [16,25,27]. In fact, it is not necessary to have a complete set of genomic sequence variants for many genetic or genomic analyses. For example, mapping resolution in any genetic linkage analysis depends rather on the number of recombinants between genetic markers than on the density of the markers [50]. Thus, use of very dense marker maps will not lead to improved mapping efficiency.

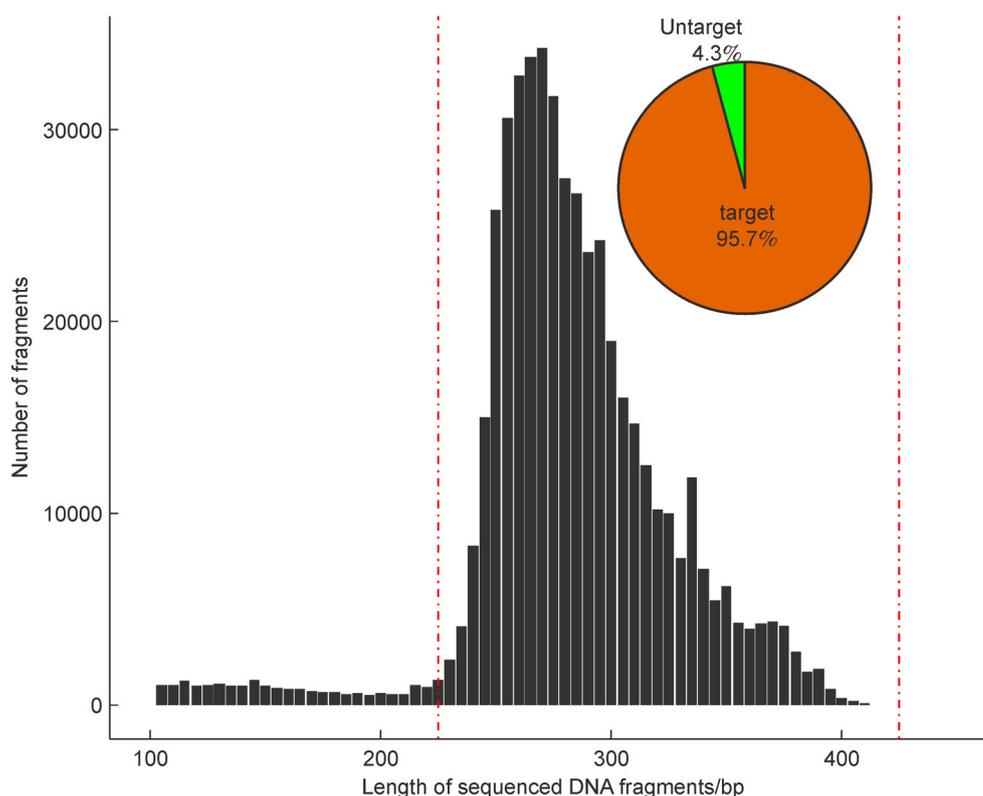


Figure 5. Length distribution of sequenced DNA fragments in pooled RRBS-seq dataset.

Among the reduced presentation sequencing approaches, the restriction enzyme associated DNA sequencing, or RAD-seq in short [27], gains the most popularity for its simple to implement, cost and time effective and flexibility in designing for various purposes. Thus, it represents an idea choice for large population sequence variant scanning and genotyping.

On the other hand, leaf tissues are usually used to extract DNA samples in plant or crop species. A large copy number of the chloroplast genome and rRNA genes in the DNA samples from leaf tissues inevitably lead to significant presentation in the sequence reads to be generated from the DNA samples. Re-analysis of sequence data from several typical plant RAD-seq experiments shows that these undesirable sequence reads may account for up to 60% of all sequence reads generated [32–34]. To address this problem, we describe here the optimized RAD-seq method for large plant population sequence variants screening and genotyping. The method is in fact an *in silico* guided design of restriction enzymes used to shear genomic DNA into segments with prior required lengths and to remove DNA segments representing chloroplast and rRNA genes in sequencing library construction. This in combination with accurate DNA segment selection by use of Pippin-Pre confers the evenness of genome coverage of selected and sequenced DNA segments, the flexibility to design for required sequence coverage over the genome of interest and for targeting genomic regions to be sequenced, effective control of undesirable sequence reads representing nongenomic or high copy number genomic DNA. Moreover, the ideas behind the optimized RAD-seq experiment can be extended to achieve cost and time effective genome-wide profiling of DNA methylation through the reduced representation bisulfite sequencing (RRBS) method. We optimistically anticipate that the desirable features of the optimized RAD-seq and RRBS methods will open a new era for genomic and epigenetic profiling of large plant and crop populations.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-016-0079-9.

ACKNOWLEDGEMENTS

This study is supported by the National Basic Research Program of China (No. 2012CB316505).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Jinhua Wu, Zewei Luo and Ning Jiang declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

- Luikart, G., England, P. R., Tallmon, D., Jordan, S. and Taberlet, P. (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.*, 4, 981–994
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M. and Blaxter, M. L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, 12, 499–510
- Poland, J. A. and Rife, T. W. (2012) Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome*, 5, 92–102
- Hutchison, C. A. III. (2007) DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.*, 35, 6227–6237
- Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 74, 5463–5467
- Bush, W. S. and Moore, J. H. (2012) Chapter 11: Genome-wide association studies. *PLOS Comput. Biol.*, 8, e1002822
- Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, 26, 1135–1145
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bembem, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376–380
- Metzker, M. L. (2010) Sequencing technologies — the next generation. *Nat. Rev. Genet.*, 11, 31–46
- Ashelford, K., Eriksson, M. E., Allen, C. M. D., D'Amore, R., Johansson, M., Gould, P., Kay, S., Millar, A. J., Hall, N. and Hall, A. (2011) Full genome re-sequencing reveals a novel circadian clock mutation in *Arabidopsis*. *Genome Biol.*, 12, R28
- Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, 475, 189–195
- Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., *et al.* (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res.*, 19, 1068–1076
- Chia, J. M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*, 44, 803–807
- Rowe, H. C., Renaut, S. and Guggisberg, A. (2011) RAD in the realm of next-generation sequencing technologies. *Mol. Ecol.*, 20, 3499–3502
- Wetterstrand, K. A. (2012). DNA sequencing costs: Data from the NHGRI large-scale genome sequencing program. National Human Genome Research Institute, Bethesda, MD. <http://www.genome.gov/sequencingcosts>
- Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., Tam, A., Wang, S., Friedmann, M., Birol, I., *et al.* (2011) SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Mol. Ecol. Resour.*, 11, 81–92
- Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., De Jong, W. S., Douches, D. S. and Buell, C. R. (2011) Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genomics*, 12, 302

18. Chepelev, I., Wei, G., Tang, Q. and Zhao, K. (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res.*, 37, e106
19. Nothnagel, M., Wolf, A., Herrmann, A., Szafranski, K., Vater, I., Brosch, M., Huse, K., Siebert, R., Platzer, M., Hampe, J., *et al.* (2011) Statistical inference of allelic imbalance from transcriptome data. *Hum. Mutat.*, 32, 98–106
20. Deelen, P., Zhemakova, D. V., de Haan, M., van der Sijde, M., Bonder, M. J., Karjalainen, J., van der Velde, K. J., Abbott, K. M., Fu, J., Wijmenga, C., *et al.* (2015) Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.*, 7, 30
21. Christodoulou, D. C., Gorham, J. M., Herman, D. S. and Seidman, J. G. (2011) Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Curr. Protoc. Mol. Biol.*, doi: 10.1002/0471142727.mb0412s94
22. Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J. and Turner, D. J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, 7, 111–118
23. Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J. and Snyder, M. (2011) Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, 29, 908–914
24. Kiialainen, A., Karlberg, O., Ahlfors, A., Sigurdsson, S., Lindblad-Toh, K. and 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, e16486
25. Uitdewilligen, J. G., Wolters, A. M., D'hoop, B. B., Borm, T. J., Visser, R. G. and van Eck, H. J. (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One*, 8, e62355
26. Mertes, F., Elsharawy, A., Sauer, S., van Helvoort, J. M., van der Zaag, P. J., Franke, A., Nilsson, M., Lehrach, H. and Brookes, A. J. (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief. Funct. Genomics*, 10, 374–386
27. Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. and Johnson, E. A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, 3, e3376
28. Poland, J. A., Brown, P. J., Sorrells, M. E. and Jannink, J. L. (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, 7, e32253
29. Wang, N., Fang, L., Xin, H., Wang, L. and Li, S. (2012) Construction of a high-density genetic map for grape using next generation restriction-site associated DNA sequencing. *BMC Plant Biol.*, 12, 148
30. Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R. and Bird, C. E. (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, 1, e203
31. Guo, Y., Yuan, H., Fang, D., Song, L., Liu, Y., Liu, Y., Wu, L., Yu, J., Li, Z., Xu, X., *et al.* (2014) An improved 2b-RAD approach (I2b-RAD) offering genotyping tested by a rice (*Oryza sativa* L.) F2 population. *BMC Genomics*, 15, 956
32. Wang, S., Meyer, E., McKay, J. K. and Matz, M. V. (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods*, 9, 808–810
33. Truong, H. T., Ramos, A. M., Yalcin, F., de Ruyter, M., van der Poel, H. J., Huvenaars, K. H., Hogers, R. C., van Enckevort, L. J., Janssen, A., van Orsouw, N. J., *et al.* (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS One*, 7, e37565
34. Chen, X., Li, X., Zhang, B., Xu, J., Wu, Z., Wang, B., Li, H., Younas, M., Huang, L., Luo, Y., *et al.* (2013) Detection and genotyping of restriction fragment associated polymorphisms in polyploid crops with a pseudo-reference sequence: a case study in allotetraploid *Brassica napus*. *BMC Genomics*, 14, 346
35. Life and Technologies. (2015) Pippin Prep™ System (includes instrument and monitor). <http://www.lifetechnologies.com/order/catalog/product/4471271?CID=search-product>
36. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. and 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, e37135
37. Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H. and Turner, D. J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, 5, 1005–1010
38. Hurd, P. J. and Nelson, C. J. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief. Funct. Genomics Proteomics*, 8, 174–183
39. Adey, A. and Shendure, J. (2012) Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res.*, 22, 1139–1143
40. Habibi, E., Brinkman, A. B., Arand, J., Kroeze, L. I., Kerstens, H. H., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N. C., *et al.* (2013) Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell*, 13, 360–369
41. Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M. and Jacobsen, S. E. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*, 452, 215–219
42. Regulski, M., Lu, Z., Kendall, J., Donoghue, M. T., Reinders, J., Llaca, V., Deschamps, S., Smith, A., Levy, D., McCombie, W. R., *et al.* (2013) The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res.*, 23, 1651–1662
43. Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S. and Jaenisch, R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.*, 33, 5868–5877
44. Smith, Z. D., Gu, H., Bock, C., Gnirke, A. and Meissner, A. (2009) High-throughput bisulfite sequencing in mammalian genomes. *Methods*, 48, 226–232
45. Wang, J., Xia, Y., Li, L., Gong, D., Yao, Y., Luo, H., Lu, H., Yi, N., Wu, H., Zhang, X., *et al.* (2013) Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing. *BMC Genomics*, 14, 11
46. Landau, D. A., Clement, K., Ziller, M. J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., *et al.* (2014) Locally disordered methylation forms the basis of intratumor methylome

- variation in chronic lymphocytic leukemia. *Cancer Cell*, 26, 813–825
47. Andolfatto, P., Davison, D., Erezylmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T. and Stern, D. L. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.*, 21, 610–617
48. Yang, X., Kundariya, H., Xu, Y. Z., Sandhu, A., Yu, J., Hutton, S. F., Zhang, M. and Mackenzie, S. A. (2015) MutS HOMOLOG1-derived epigenetic breeding potential in tomato. *Plant Physiol.*, 168, 222–232
49. Tao, S., Chu, J., Liu, X., Zhang, R., Zhang, Z. and Luo, Z. (2002) High-resolution gene mapping using admixture linkage disequilibrium. *Chin. Sci. Bull.*, 47, 1717–1719
50. Krueger, F. and Andrews S. R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27, 1571–1572