DNA Research, 2024, **31**, 1–11
https://doi.org/10.1093/dnares/dsad027
Advance access publication 22 December 2023
Research article

OXFORD

Research article

MCPtaggR: R package for accurate genotype calling in reduced representation sequencing data by eliminating error-prone markers based on genome comparison

Tomoyuki Furuta[1,*] and Toshio Yamamoto[1]

[1]Institute of Plant Science and Resources, Okayama University, Kurashiki, Okayama, Japan

[*]To whom correspondence should be addressed. Tel: +81-86-434-1208. Fax: +81-86-434-1208. Email: f.tomoyuki@okayama-u.ac.jp.

Abstract

Reduced representation sequencing (RRS) offers cost-effective, high-throughput genotyping platforms such as genotyping-by-sequencing (GBS). RRS reads are typically mapped onto a reference genome. However, mapping reads harbouring mismatches against the reference can potentially result in mismapping and biased mapping, leading to the detection of error-prone markers that provide incorrect genotype information. We established a genotype-calling pipeline named mappable collinear polymorphic tag genotyping (MCPtagg) to achieve accurate genotyping by eliminating error-prone markers. MCPtagg was designed for the RRS-based genotyping of a population derived from a biparental cross. The MCPtagg pipeline filters out error-prone markers prior to genotype calling based on marker collinearity information obtained by comparing the genome sequences of the parents of a population to be genotyped. A performance evaluation on real GBS data from a rice F_2 population confirmed its effectiveness. Furthermore, our performance test using a genome assembly that was obtained by genome sequence polishing on an available genome assembly suggests that our pipeline performs well with converted genomes, rather than necessitating *de novo* assembly. This demonstrates its flexibility and scalability. The R package, MCPtaggR, was developed to provide functions for the pipeline and is available at https://github.com/tomoyukif/MCPtaggR.

Key words: genotyping, genome comparison, next-generation sequencing, R package

1. Introduction

Next-generation sequencing (NGS) has enabled us to simultaneously obtain high-resolution genotypic information from a large number of samples.[1] NGS provides genotyping platforms with remarkably improved flexibility and throughput and is now routinely used in diverse fields of biology.[2–4] Nevertheless, obtaining genome-wide genotype information via whole-genome resequencing (WGR) is an expensive option for researchers.[5] Numerous reduced representation sequencing (RRS) methods have been introduced to meet the demand for cost-effective genotyping systems using dense markers.[6] Since restriction site-associated DNA sequencing (RAD-seq) was first published in 2007,[7] several derivative methods have been introduced, including ddRADseq[8] and genotyping-by-sequencing (GBS).[9,10]

To identify genotypes at single-nucleotide polymorphism (SNP) markers, NGS reads are typically mapped onto one reference sequence. However, several studies have reported that mapping reads on one reference causes the detection of error-prone SNP markers owing to mismapping and biased mapping of NGS reads derived from non-reference genomes.[11–13] In addition to genotyping a population derived from a distant cross, such as interspecific crosses conducted in plants,[14] these error-prone markers have also been observed in genetic studies of bacteria and humans.[11–13] Read mismapping may have been caused by structural differences between the reference genome and genomes that were genotyped. Translocations of genome segments potentially result in reads that can be mapped to the same location in the reference genome but actually originate from another location in the genomes of genotyped samples.[15] Biased mapping, also known as reference bias, occurs because of the methodological nature of read alignment.[16,17] Sequencing reads derived from a non-reference genome have mismatches with reference sequences. These mismatches cause sequencing reads to be simultaneously mapped to multiple locations with the same probability or even unmapped.[16] Because read alignment is achieved by evaluating the sequence matches between a query sequence and a subject sequence, the reference allele reads can be preferentially mapped to the reference genome. In such situations, a genome graph representing a pan-genome sequence is a solution to avoid these errors.[18] Variations, including SNPs, insertions and deletions (indels), and translocations, represented in the genome graph prevent mismapping and biased mapping. Although this approach effectively improves the genotyping accuracy, genome graph construction and read alignment remain computationally challenging. Other types of graph-based approaches have also been proposed.[19–21] These graph-based approaches require a list of SNPs validated as reliable and call genotypes based on the SNP list. However, these lists of validated SNPs

Received 18 October 2023; Revised 11 December 2023; Accepted 18 December 2023

are typically not available for genetic mapping and plant breeding populations because scientists and breeders tend to use a unique combination of parents to create a population. Therefore, a simple and flexible SNP-genotyping pipeline is required to avoid mismapping and biased mapping.

This study introduces the mappable collinear polymorphic tag genotyping (MCPtagg) pipeline and the R package MCPtaggR, which provides functions for MCPtagg. MCPtagg is a pipeline for processing RRS data to precisely detect SNPs and count reads on reliable tags (short genome segments) that were validated based on genome comparison information. The current version of the pipeline supports a population derived from a biparental cross. First, a simulation study illustrates how structural differences in genomes can lead to the emergence of error-prone markers in GBS data. We then assess the performance of MCPtaggR using real data, comparing it with the well-established pipeline TASSEL-GBS.[22] Finally, the simplicity, efficiency, and flexibility of the MCPtagg pipeline and the MCPtaggR implementation are discussed.

2. Materials and methods

2.1. *In silico* genome digestion and digested-tag alignment

To simulate the digestion of genomic DNA in GBS library preparation, *in silico* digestion was performed on the genome sequences of the cultivated rice *Oryza sativa* cv. Nipponbare (NB), and wild rice *O. longistaminta* (OL). The NB and OL genome sequences were downloaded from RAP-DB[23] and *Oryza longistaminata* Information Resource,[24] respectively. The obtained FASTA files of the genome sequences were loaded into the R environment using the Biostrings package, version 2.68.1.[25] Since the real sequencing reads used in this study were obtained by GBS with KpnI and MspI (see Section 2.5), the recognition sequences of KpnI and MspI were searched in the given genome sequences using the DigestDNA function in the DECIPHER package, version 2.0.[26] The *in silico* digested fragments were filtered to retain only those with a KpnI cut site at one end and an MspI cut site at the other end, which could potentially be sequenced in real GBS. To simulate the sequencing reads generated by NGS, fragment sequences were cropped to retain a maximum of 75 bp from both ends. The 75-bp cropped sequences, referred to as digested-tags, from the KpnI and MspI cut sites, were treated as Read1 and Read2 of typical paired-end sequencing output and stored in FASTQ files, respectively. Phred scores for the reads were set to ';' indicating the quality score is 26 for all nucleotides, which was the default behaviour of the writeXStringSet function in the Biostrings package. Each set of digested-tags generated from the NB and OL genomes was aligned to the NB genome using BWA-MEM, version 0.7.17.[27] The flags '-S -P' were set to run BWA-MEM with skipping read pairing and mate rescue to allow the tags mapped to their best matching locations regardless of mate read mapping locations.

2.2. Evaluation of digested-tag alignments

The alignment results of the digested-tags were evaluated using the following steps. The output files generated by BWA-MEM, which store the alignments of the NB- and OL-digested-tags on the NB genome, were imported into the R environment using the Rsamtools package, version 2.16.0.[28] First, unmapped reads were filtered from the imported alignment data. In general, genotype-calling pipelines filter multiple mapping reads mapped to multiple locations with the same mapping quality. Therefore, multiple mapping reads were filtered out based on the information stored in the XA tag of the SAM file, which shows the additional mapping locations of the reads. This filtering retained only reads uniquely mapped on single locations that were the origins of the digested-tags in the case of the NB-digested-tags. Subsequently, tag alignment locations were searched to identify genome locations where both NB- and OL-digested-tags were mapped (co-mapped sites). The co-mapped sites were classified into two types based on the number of mapped OL-digested-tags: single-tag sites and multi-tag sites, where only one and multiple tags were mapped, respectively (Fig. 1A). Finally, the collinearity of the origins of tags between genomes was evaluated. Collinearity between genomes refers to the conservation of sequence order and orientation across different genomes. Genome regions that have collinearity are defined as collinear blocks. Collinear blocks between the NB and OL genomes (Fig. 1B) were searched using the functions of MUMmer4[29] as follows. Repeat sequences in the genomes were masked using RepeatMasker[30] with the arguments '-excln -s -no_is -species rice' and the RepBase RepeatMasker Edition library downloaded from www.girinst.org prior to processing using MUMmer to reduce the unnecessary computation time to find sequence matches between repeat sequences. The masked genomes were processed using the nucmer function with the parameters '--maxmatch -g 1000' followed by filtering on the outputs using the delta-filter function with the setting '-1'. Collinear block coordinates were summarized using the show-coords function with the setting '-CTlr'. Intervals between collinear blocks can also be assumed to conserve relative positions in the genomes. Therefore, these intervals were considered collinear (Fig. 1B). If the origins of the NB and OL tags mapped to single-tag sites were located in the corresponding collinear blocks or intervals in the NB and OL genomes, respectively, the single-tag sites were recorded as collinear tag sites (Fig. 1C). For each non-collinear tag site, the physical distance was measured between the start position of the block/interval in which the OL tag was mapped and the start position of the collinear block/interval in the NB genome, corresponding to the origin of the OL tag in the OL genome (Fig. 1D). If the OL tags of the non-collinear tag sites originated from chromosomes that were different from the chromosomes where the OL tags were mapped in the NB genome, these sites were counted as interchromosomal alignment sites (Fig. 1E).

2.3. Algorithm of MCPtagg

MCPtagg was designed for the RRS-based genotyping of a population derived from a biparental cross. The pipeline requires the genome sequences of parents and NGS reads obtained using RRS from a population derived from a biparental cross. Before mapping reads, the pipeline prepares a list of MCP tags. First, a whole-genome sequence comparison detects collinear blocks, each of which is a pair of genomic regions sharing identical or similar sequences that can be found at one locus in each parental genome and are arranged in the same order in both parental genomes (Fig. 2A). The SNPs in the collinear blocks are used as SNP markers for subsequent genotype calling. Thereafter, *in silico* digestion of the genomes simulates the restriction fragments generated during library preparation for the RRS (Fig. 2B). To simulate the NGS reads obtained by sequencing

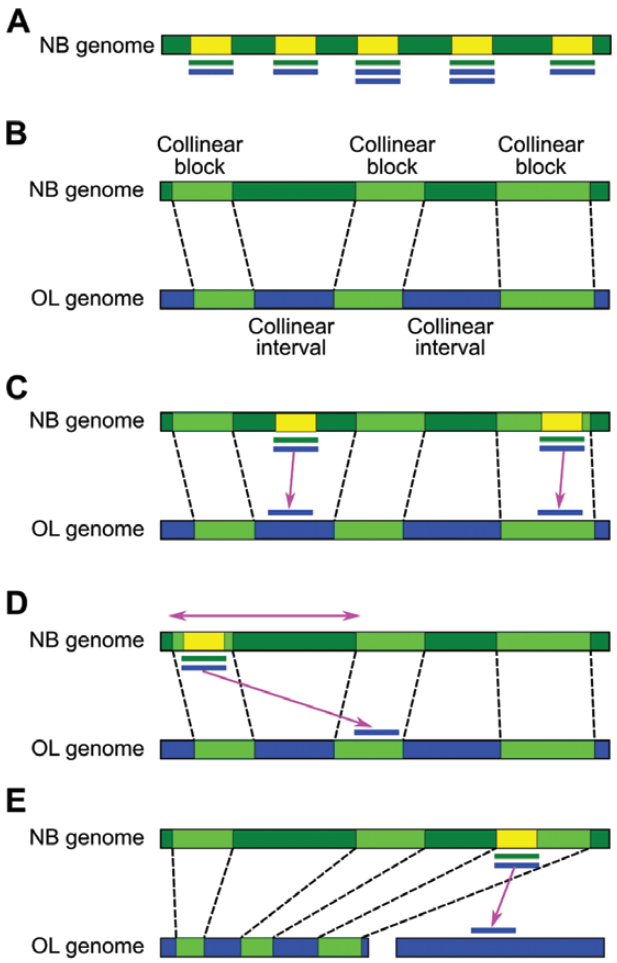

Figure 1. Collinearity assessment of the digested-tags. (A) The locations, at which the NB- and OL-digested-tags were mapped, were classified into single-tag sites and multiple-tag sites. Single-tag sites are the genome locations where single NB and OL tags were mapped, whereas multi-tag sites have multiple OL tags. NB and OL tags are represented by green and blue lines in the schematic image. (B) The NB and OL genomes were segmented in collinear blocks and intervals. (C) A single-tag site was classified as a collinear-tag site if the OL tag that was mapped at the single-tag site originated from the collinear block/interval corresponding to the block/interval of the single-tag site. Magenta arrows point to the origins of the OL tags in the OL genome. (D) If a single-tag site was a non-collinear tag site, the physical distance was measured between the start position of the block/interval in which the OL tag was mapped and the start position of the collinear block/interval in the NB genome corresponding to the origin of the OL tag in the OL genome. The magenta double-headed arrow indicates the distance to be measured. (E) If an OL tag originated from a chromosome that does not correspond to the chromosome where the OL tag was mapped in the NB genome, the site was assigned as an interchromosomal alignment site.

the restriction fragments, the simulated sequences are clipped to ensure that they are less than or equal to the maximum read length. In this step, the algorithm identifies SNPs that are potentially undetectable in genotype calling because there were no restriction sites close enough to be sequenced (Fig. 2B). All the simulated reads from both genomes are simultaneously mapped to the genomes (Fig. 2C). Based on the mapping results, the filtering step retains only SNPs at which the simulated reads can be uniquely mapped onto each genome at the correct positions (Fig. 2D). If a simulated read is mapped to multiple locations in both or either of the genomes, the read in real data can potentially cause mismapping if it is mapped to one reference genome. The algorithm then lists MCP tags that are the genomic sequences/regions where the simulated reads are uniquely and completely mappable, located in the collinear blocks, and harbour SNPs. Genotype calling is accomplished by mapping real NGS reads to MCP tags. Reads mapped to the corresponding SNP positions on the MCP tags in each genome without any mismatches and indels are counted to determine the genotypes at the SNP markers.

2.4. Implementation of the MCPtagg pipeline in MCPtaggR

The R package, MCPtaggR, provides functions for the MCPtagg pipeline (Fig. 2E). Collinear block detection and read mapping require the external tools, MUMmer[29] and Subread,[31] respectively. The run_mummer() function is a wrapper function that internally executes MUMmer functions to identify collinear blocks and SNPs. FASTA files of the reference and alternative genomes set to run_mummer() are used to count the reference and alternative allele reads in the read-counting step. The mummer2SNPs() function converts and organizes the MUMmer outputs into a list of candidate MCP tags. The digestGenome() and alignTAG() functions perform *in silico* genome digestion and simulated read mapping. The outputs from both functions are used as inputs for the findMCPtag() function to list the MCP tags. The function mcptagg() performs read mapping to the MCP tags and outputs read counts and genotype calls in a variant call format (VCF) file[32] and a genomic data structure (GDS) file.[33] The output GDS file can be directly manipulated using the GBScleanR package, which provides functions for filtering, data visualization, and genotyping error correction.[14] The alignTAG() and mcptagg() functions internally execute the align() function in the Rsubread package for read mapping.[34]

2.5. Genotype calling using TASSEL-GBS and MCPtaggR

To assess the advantages of MCPtaggR for genotype calling, its performance was compared with that of the well-established genotype-calling pipeline TASSEL-GBS.[22] Both tools were subjected to genotype calling using previously published GBS reads obtained from 813 F_2 samples derived from a cross between distant rice relatives NB and OL.[35] Reference genome sequences for NB and OL were obtained as described above. The NB genome was used as a reference in the TASSEL-GBS pipeline, and the NB and OL genomes were used as a reference and an alternative to MCPtaggR, respectively. Read data for the F_2 samples were obtained from the sequence read archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) by querying project ID PRJDB5346. These reads were originally obtained in GBS with a KpnI–MspI RE pair using MiSeq with 75-bp paired-end sequencing. Because the reads deposited in the SRA were demultiplexed and had no barcode sequences required for processing using TASSEL-GBS, dummy barcode sequences were concatenated to all reads using the bardcode_faker function available at https://github.com/labroo2/rtassel_supp. The scripts used for genotype calling are summarized in Supplementary Methods. The genotype call data obtained were subjected to SNP filtering using the following criteria:

(1) SNPs that were biallelic between parents and monomorphic in each parent were retained.

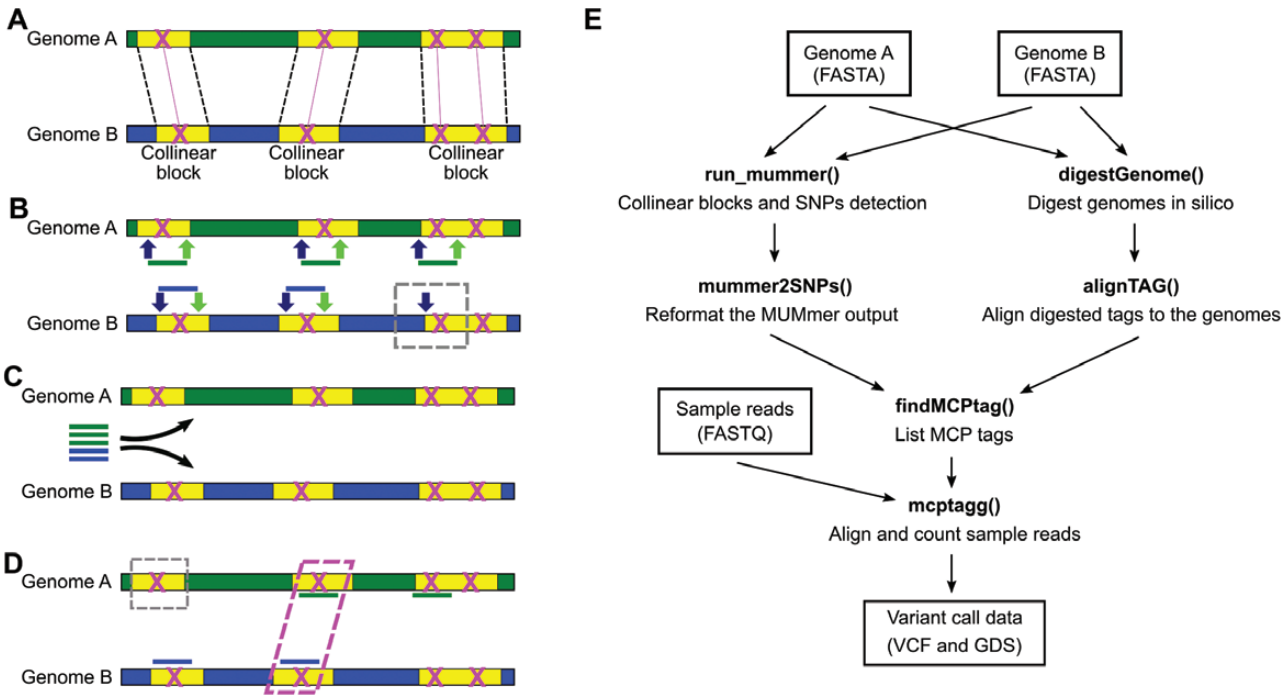

Figure 2. Schematic of the MCPtagg algorithm and the MCPtaggR workflow. (A) The MCPtagg pipeline requires the SNPs (indicated by magenta X symbols) and collinear blocks (indicated by yellow boxes) information that can be obtained by comparing two genomes using the genome sequence comparison tool MUMmer. (B) Subsequently, the pipeline conducts *in silico* digestion of the genomes to simulate sequencing reads generated in an NGS run. Restriction sites are indicated by blue and green arrows. The grey dashed rectangle indicates an example that a restriction site of a restriction enzyme is missing in genome B. (C) The simulated reads are mapped to both genomes simultaneously. (D) Based on the mapping results, the algorithm detects MCP tags (indicated by the magenta dashed rectangle). The grey dashed rectangle indicates a region where the simulated read failed to map uniquely and completely. (E) The schematic image represents the MCPtaggR workflow for genotype calling. The texts in rectangles display input and output files, whereas the bold texts indicate the function names. The arrows represent data flow through the pipeline.

(2) SNPs that were completely missing in the F_2 samples were removed.
(3) The first SNP was retained when multiple SNPs were found within a 75-bp stretch.

Because the genotype data generated by MCPtaggR were supposed to contain only SNPs that met criterion (1), filtering based only on criteria (2) and (3) was applied to the MCPtaggR output. RRS-based genotype data usually contain numerous errors, including heterozygote undercalling, which is a typical error that incorrectly identifies a homozygote at a heterozygous site. Therefore, the obtained genotype data were processed using GBScleanR to correct errors and estimate allele dosage.[14] The scripts used to execute the filtering and dosage estimation are available in Supplementary Methods.

2.6. Evaluation of genotype call data

Genotype call data obtained using MCPtaggR and TASSEL-GBS were evaluated based on the estimated dosage generated by GBScleanR as raw genotype data containing a large number of missing values and erroneous calls, including heterozygote undercallings, which could prevent the systematic evaluation of software performance. The number of recombinations in which the genotype calls changed from one to another was counted in the estimated dosage data. The middle points of the physical positions of the markers, indicating recombination, were considered recombination breakpoints. The number of double crossovers was also evaluated by counting the number of genome segments in which the estimated dosage was heterozygous and flanked by homozygous segments of the same allele; for example, a heterozygous segment flanked by reference homozygous segments on both sides.

It was expected that SNP markers supported by non-collinear tags would show more frequent mismapped reads than those supported by collinear tags. Therefore, the mismapping rate per marker was estimated by calculating the proportion of genotype calls that were heterozygous in the raw genotype data, but homozygous in the estimated dosage data. The SNP markers were classified into three groups: (1) those commonly found by MCPtaggR and TASSEL-GBS, (2) TASSEL-GBS only, and (3) MCPtaggR only. The statistical significance of the difference in mismapping rates between the three marker groups was tested using the Wilcoxon rank-sum test. The obtained *P*-values were adjusted using Bonferroni's method.

To further assess the properties of the SNP markers obtained, the origins of the reads used by TASSEL-GBS to call genotypes were assessed by mapping the tags listed by TASSEL-GBS. The tags of TASSEL-GBS, referred to as tassel tags, are unique sequences found in the input read data and stored in an SQLite database output by the pipeline. The SNP position information and corresponding tag sequences used for genotype calling at the SNPs were retrieved from the SQLite database using SQLite, version 3.41.2, which is available at https://www.sqlite.org. The tag sequences were then mapped to the NB and OL genomes using the aln function of BWA, as the TASSEL-GBS pipeline uses this function instead of BWA-MEM. Tag alignment results were analysed in a manner similar to that described in Section 2.2, except for the method used to determine tag origins in the OL genome.

Because the true origins of the tags were unknown, NB and OL tassel tags were defined as tags that showed no mismatch or mismatch(es) with the NB genome sequence, respectively. The best matching locations of the OL tassel tags were assigned as the origins of the tags. If the origin of an OL tassel tag was located on a chromosome different from the chromosome on which the tag was mapped in the NB genome, this OL tag was recorded as an interchromosomal alignment. The collinearity of tag origins between the genomes was evaluated as described in Section 2.2. The SNP markers identified by TASSEL-GBS were grouped into four classes based on the properties of the tags supporting each SNP marker: (1) assigned as a multi-tag site (Fig. 1A), (2) supported by collinear tags (Fig. 1C), (3) assigned as a single-tag site but not supported by collinear tags (Fig. 1D), and (4) assigned as an interchromosomal alignment site (Fig. 1E). The statistical significance of the difference in mismapping rates between the four marker classes was tested using the Wilcoxon rank-sum test. The obtained *P*-values were adjusted using Bonferroni's method.

2.7. Genotype calling using MCPtaggR and an OL-nized genome sequence

Pilon software is a well-established high-performance genome sequence polisher that corrects genome sequences based on short-read alignment information.[36] Theoretically, using Pilon, one can convert one reference genome sequence to another by aligning the WGR reads derived from a related species. Therefore, instead of the publicly available OL genome assembly, the NB genome sequence was converted into an OL genome sequence (OL-nized) using Pilon. The WGR data of an OL plant using HiSeq 2500, deposited under project ID PRJDB6339,[24] were obtained from SRA. The WGR reads obtained were aligned to the NB genome using BWA-MEM. The resultant BAM file was supplied as input to Pilon. In typical cases of genome polishing for *de novo* genome assembly, multiple rounds are performed by repeating the read alignment on a polished genome, followed by polishing using Pilon to improve sequence accuracy. Thus, OL-nized genome sequences were generated by performing one, two, three, and four rounds of polishing. Genotype calling using MCPtaggR followed by error correction and dosage estimation using GBScleanR was conducted as described in Section 2.5, using each OL-nized genome sequence as an alternative genome. The number of recombinations per chromosome was counted in the resultant estimated dosage data as described in Section 2.6.

2.8. Data availability

MCPtaggR is available on GitHub (https://github.com/tomoyukif/MCPtaggR). More details on package installation and usage can be found in the GitHub repository and the vignette available at https://tomoyukif.github.io/MCPtaggR/. The test data used in this study were described in our previous publication.[35] Because the data size of the full test data exceeds 12 GB, sample FASTQ files, which consist of 10 samples, are installed with MCPtaggR to test the functions.

3. Results

3.1. GBS-read alignment simulation and error-prone detection

In silico digestion at the KpnI and MspI recognition sites generated 155,474 and 125,932 digested-tags from the NB and OL genomes, respectively (Table 1). Tag mapping using BWA-MEM resulted in 20,210 and 31,040 unmapped NB- and OL-digested-tags due to short tag sequences, which were less than 30 bp and were ignored by the aligner, and Ns in the tag sequences, which were added to fill the gaps between contigs of the genome assemblies. Multiple alignments were found in 6,295 and 5,308 of the NB- and OL-digested-tags, respectively. Notably, BWA-MEM failed to align 15,919 NB-digested-tags back to their original locations, without any additional alignment information. Manual searches of the tag sequences in the NB genome revealed that these tags had identical sequences at multiple locations. Therefore, these tags were treated as multiple alignment tags. After filtering out the unmapped and multiple-mapped tags, 113,050 and 89,584 NB- and OL-digested-tags were retained, respectively. In the retained OL-digested-tags, 19,705 tags showed interchromosomal alignments and were filtered out. To further assess the properties of the mapped tags, overlaps of tag alignment locations were searched and found in 43,003 sites. Of the 43,003 co-mapped sites, 38,499, and 4504 sites had single and multiple OL-digested-tags, respectively (Fig. 1A and Table 2). More than one mismatch pattern was found at 2,238 multiple-tag sites. These sites potentially lead to incorrect genotyping calls owing to different mismatch patterns, resulting in different contributions to read counts as reference and alternative alleles at SNPs. Filtering of single-tag sites with no mismatches and interchromosomal alignments removed 15,760 and 1582 sites, respectively (Table 2). As the number of retained single-tag sites was 21,157, approximately 7% of the single-tag sites had SNPs that could be potentially detectable in GBS, but would be derived from interchromosomal alignments, leading to incorrect genotyping calls (Fig. 1E). As the

Table 1. Classification of digested-tags

Classification	NB-digested-tags	OL-digested-tags
Total	155,474	125,932
Unmapped	20,210	31,040
Multiple alignments	6,295	5,308
Silent multiple alignments[a]	15,919	NA
Interchromosomal alignments	NA	19,705
Retained	113,050	69,879

[a]NB-digested-tags that failed to be aligned back to the original locations.

Table 2. Classification of co-mapped sites

Classification	No. of tags
Total sites	43,003
Multiple-tag sites	4,504
Multiple mismatching patterns	2,238
Non- or one mismatching pattern	2,266
Single-tag sites	38,499
Non-mismatching pattern	15,760
Interchromosomal alignments	1,582
Non-collinear sites	4,360
Collinear sites	16,797

final step of the simulation study, the collinearity of the origins of the digested-tags in the NB and OL genomes was confirmed (Fig. 1B-D). Neither the NB nor OL tag at 105 single-tag sites could be assigned to any collinear block/interval because, for these tag sites, one of the tags was assigned to a block/interval in a genome, whereas the other tag was assigned to an interval in which the corresponding interval was missing in the corresponding genome. Imperfect collinear block detection due to local genomic rearrangements may cause these misassignments. Among the remaining single-tag sites, 435 and 172 sites had OL-digested-tags that were derived from the collinear blocks/intervals located more than 100 kb and 1 Mb away from the blocks/intervals where NB-digested-tags were assigned, respectively (Fig. 1D). These sites can also potentially result in incorrect genotype calls by mismapping reads derived from distant genomic locations. In this simulation, error-prone SNP markers that potentially provide incorrect genotype calls in GBS were found at 4,255 locations, which included multiple-tag sites with multiple missing patterns, single-tag sites with interchromosomal alignments, and single-tag sites with tags from non-collinear origins. In contrast, 16,797 sites were identified as reliable SNP markers (Table 2). However, this simulation also demonstrated that more than 20% of the SNP marker sites could be error-prone in GBS.

3.2. Performance of MCPtaggR for accurate genotype calling

To eliminate the detection of error-prone markers in RRS-based genotyping, including GBS, the MCPtagg pipeline was developed and implemented in the R package MCPtaggR, as described in Sections 2.3 and 2.4. MCPtaggR detected 19,714 and 19,695 MCP tags in the NB and OL genomes, respectively (Supplementary Table S1). The GBS reads were mapped on 6418 and 6701 MCP tags in the NB and OL genomes, respectively. The differences in the number of MCP tags were caused by the differences in RE cut sites between the NB and OL genomes. We confirmed the classifications of the read-mapped OL MCP tags in the classes defined by mapping the digested-tags on the NB genome (Tables 1 and 2). Out of 6,701 read-mapped tags, 4,589 tags matched the digested-tags that were mapped on the collinear single-tag sites (Supplementary Table S1). The rest of the read-mapped OL MCP tags were classified as tags that were potentially invalid for genotype calling if only the NB genome was used as a reference for read mapping. In other words, MCPtaggR successfully called genotypes using reads that might be potentially mapped on incorrect locations if the NB genome was solely used as a reference.

The performance was evaluated by comparing the genotype calls generated by MCPtaggR and TASSEL-GBS. After filtering SNP markers as described in Section 2.5, 5,486 and 2,973 SNP markers were retained for the data generated by MCPtaggR and TASSEL-GBS, respectively. Because the raw genotype data output from the tools contained a large number of missing values and heterozygote undercallings, error correction and allele dosage estimation were performed using GBScleanR. The performance of the tools for genotype calling was indirectly evaluated based on the estimated allele dosage instead of raw genotype data. Figure 3 depicts a representative case in which MCPtaggR and TASSEL-GBS yielded genotypic calls that resulted in different dosage estimations. TASSEL-GBS detected alternative allele reads at a larger number of markers where the estimated dosage

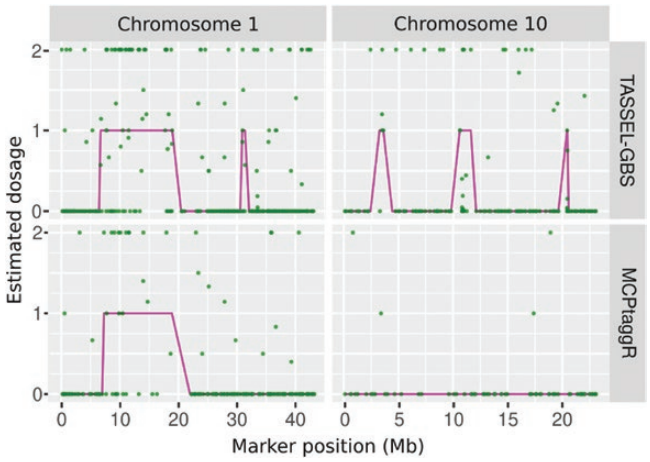

Figure 3. Comparison of obtained read counts and estimate dosages. Estimated dosage (magenta lines) and allele read ratios per marker (green dots) on chromosomes 1 and 10 in a representative sample. The green dots at 0 and 2 on the *y*-axis represent the markers at which only the reference and alternative allele reads were mapped, respectively, whereas the dots between them represent markers that had both allele reads.

indicated reference homozygotes at the surrounding markers. As a result, GBScleanR estimated the dosage to change from 0 to 1 and then return to 0 within short intervals, indicating a double crossover (upper panels in Fig. 3). Theoretically, double crossovers at short intervals are rarely observed in F_2 populations. These unexpected alternative allele reads were reduced, and double crossovers were eliminated from the MCPtaggR results (Fig. 3, lower panels).

To statistically assess the differences in genotype calling using these tools, we first calculated the concordance rates of the raw genotype calls and estimated dosages between MCPtaggR and TASSEL-GBS. Only 507 SNPs were commonly detected by both tools, whereas 4,979 and 2,466 SNPs were uniquely detected by MCPtaggR and TASSEL-GBS, respectively (Supplementary Table S2). As the number and positions of obtained SNPs were different between the tools, we compared raw genotype calls and estimated dosages between the nearest SNPs detected by MCPtaggR and TASSEL-GBS. The raw genotype calls showed a concordance rate of 67.34%, whereas the estimated dosages exhibited a concordance rate of 98.76% (Supplementary Table S2). Thus, although the raw genotype calls were largely different between MCPtaggR and TASSEL-GBS, highly similar dosages were estimated by GBScleanR from both genotype data. Nevertheless, as shown in Fig. 3, there were unexpected double crossovers at short intervals. To further assess the occurrence of unexpected crossovers due to error-prone markers, the average recombination frequency per chromosome was calculated (Fig. 4A and Supplementary Table S3). The expected recombination frequencies in the chromosomes were estimated from their physical lengths, assuming an average recombination rate of 0.04 per Mb. Based on a previous study, the expected recombination frequency across the entire genome of a rice F_2 plant is approximately 15.2.[37] As this value matched the sum of the expected recombination frequencies in the chromosomes (14.94), the estimation of recombination frequencies was reliable and acceptable. Although the data generated by both TASSEL-GBS and MCPtaggR showed more frequent recombination than expected, MCPtaggR yielded values

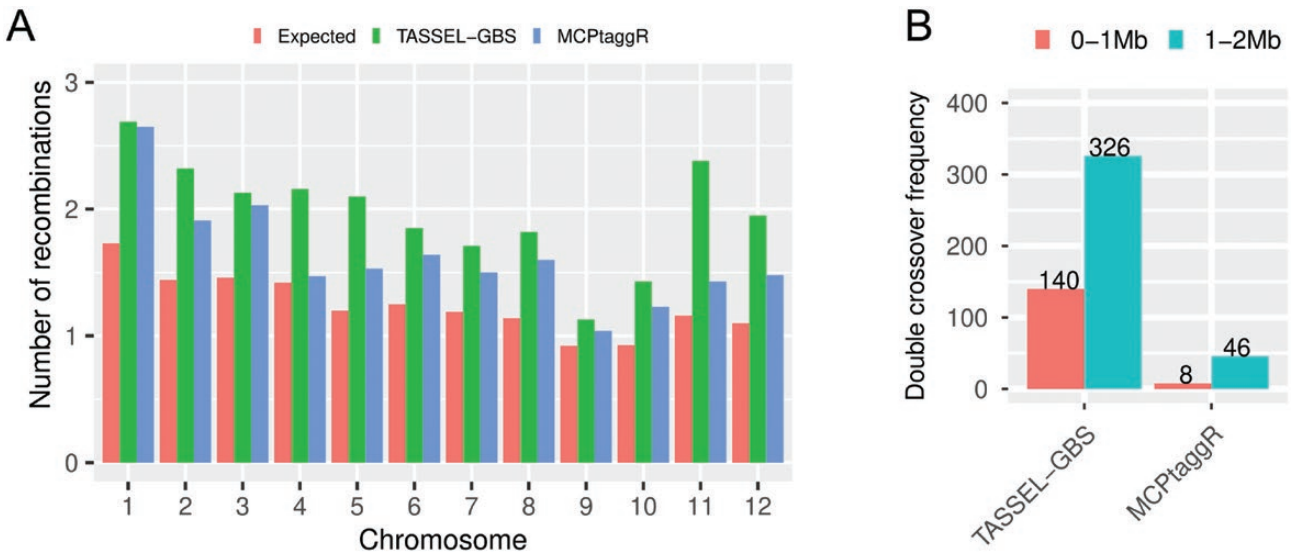

Figure 4. Genotyping quality assessment. (A) Expected and observed average numbers of recombinations per chromosome in the estimated dosage data. (B) Total number of double crossovers within stretches of 0–1 Mb and 1–2 Mb in the estimated dosage data.

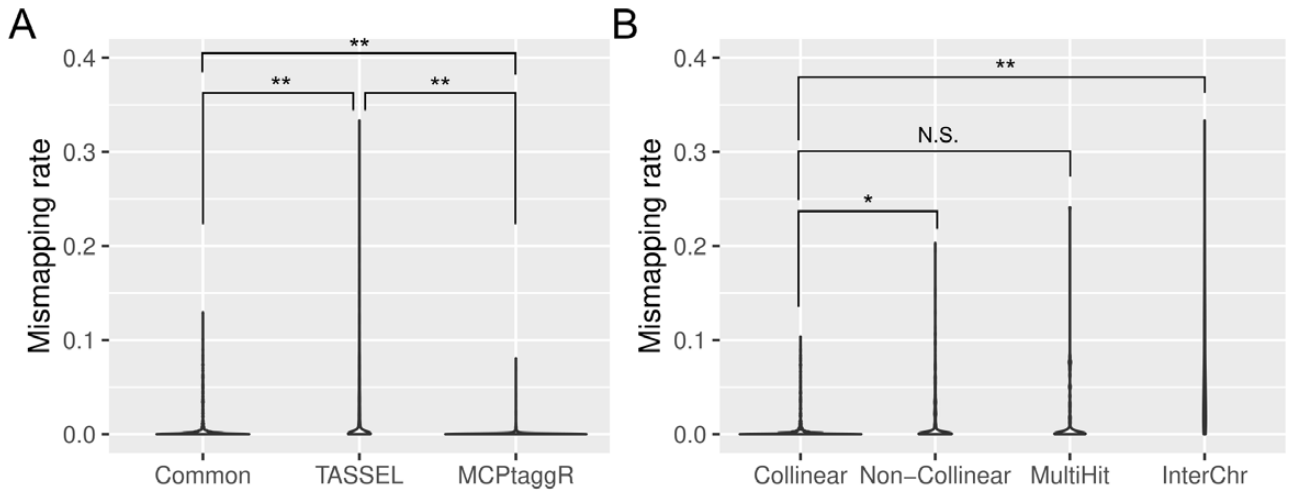

Figure 5. Mismapping rate per marker. The mismapping rate per marker is shown in violin plots for each class of markers, classified based on the commonness between the tools (A) and the collinearity of mapped tags (B). 'Common', 'TASSEL', and 'MCPtaggR' in panel A indicate the markers commonly detected by both tools, only detected by TASSEL-GBS, and MCPtaggR, respectively. The mismapping rate of the 'Common' markers was calculated from the data obtained using TASSEL-GBS. 'Collinear', 'Non-Collinear', 'MultiHit', and 'InterChr' indicate the markers supported by collinear tags, non-collinear tags, multiple alignment tags, and interchromosomal alignment tags, respectively. The significance of the differences in mismapping rates between classes is indicated by $^{N.S.}P > 0.05$, $^{*}P < 0.05$, and $^{**}P < 0.01$.

that were closer to the expected frequencies compared with those of TASSEL-GBS (Fig. 4A). The number of double crossovers within stretches of 0–1 and 1–2 Mb in the estimated dosage data was also counted (Fig. 4B and Supplementary Table S4). A large reduction in unexpected double crossovers within short intervals was observed in the genotype data of MCPtaggR compared with that of TASSEL-GBS.

Although GBScleanR is robust against error-prone markers because it incorporates marker-specific error rates into the hidden Markov model (HMM),[14] it can fail to estimate true genotypes, depending on the distribution and pattern of read counts of error-prone markers. As the estimated dosage of TASSEL-GBS data revealed an increase in recombination frequency and double crossovers, it could be expected that TASSEL-GBS would call genotypes with a larger number of error-prone markers than MCPtaggR. In general, read mismapping caused by the non-collinearity of the origins of the reads results in unexpected heterozygous calls at sites where the true genotypes are homozygotes. Therefore, the mismapping rate per marker was estimated by calculating the proportion of genotype calls that were heterozygous in the raw genotype data, but homozygous in the genotype estimated by GBScleanR. The TASSEL-only markers (2,201 markers) showed a significantly larger mismapping rate than the common markers and MCPtaggR-only markers (772 and 4,150 markers, respectively) (Fig. 5A). In addition, a comparison between common markers and MCPtaggR-only markers indicated a significant reduction in the mismapping

rate of MCPtaggR-only markers. Together with a reduction in the number of recombinations and double crossovers, these results demonstrate that MCPtaggR efficiently eliminates error-prone markers from genotype calling and provides more reliable genotype data in combination with GBScleanR.

3.3. Confirming the cause of mismapped read

To confirm that the mismapped reads originated from non-collinear locations in the NB and OL genomes, the original genomic locations of the tassel tag sequences, which were listed and used by TASSEL-GBS for genotype calling, were determined by aligning the tassel tags to the NB and OL genomes. TASSEL-GBS listed 65,063 unique tags from the read data of the rice F_2 samples, whereas 2,973 of the retained SNP markers were supported by only 6,809 tags (Table 3). Sequence mismatches were found in 3,868 of the 6,809 tags in the tag alignment to the NB genome. Although these mismatched tags may contain tag sequences generated by sequencing error, 3,868 tags were considered to originate from the OL genome. Of the 3,868 OL tags, 577 tags were mapped to the OL chromosomes that were different from the NB chromosomes, where the tags were mapped for genotype calling, indicating interchromosomal alignments (Fig. 1E). Multiple alignments were observed for 94 OL tags. These interchromosomal and multiple alignment tags contributed to genotype calling at the 390 and 77 retained SNP markers, potentially leading to incorrect genotyping. Among the remaining OL tags, 489 tags that were mapped to 402 SNP marker sites failed to show collinearity with the corresponding NB tags that were mapped to the same SNP locations in the NB genome (Fig. 1B–D). Some of the non-collinear tags did not align with either the NB or OL genomes. The incompleteness of genome assemblies may have caused these missing alignments. Finally, 2,104 SNP markers were found to be reliable SNP markers that were supported by 2,708 OL tags, maintaining collinearity with the corresponding NB tags. This collinearity assessment revealed that 29.2% of the SNP markers provided by TASSEL-GBS were genotyped based on reads originating from invalid or unreliable sources.

The mismapping rates estimated from the proportion of mismatches in the raw genotype calls and estimated dosages revealed significant differences in the properties of the SNP markers classified based on collinearity (Fig. 5B). The markers supported by collinear tassel-tags showed significantly smaller mismapping rates than those supported by non-collinear and interchromosomal alignment tags. The highest mismapping rate was observed for markers supported by interchromosomal alignment tags. The distribution of the mismapping rate per marker indicated that markers with interchromosomal alignments had the most deleterious impact on genotype calling (Fig. 5B). Significantly larger proportions of the markers supported by non-collinear and multiple tags were identified as mismapping prone compared with the markers supported by collinear tags, although some of the markers had no mismapping reads. As shown in the previous section, MCPtaggR filtered out these error-prone markers before genotype calling (Fig. 5A). These results suggested that SNP marker filtering based on sequence collinearity is a reasonable and effective method for accurate genotype calling.

Table 3. Classification of tassel-tags

Classification	No. of tags
Total tags	65,063
Valid SNP supporting tags	6,809
NB tags	2,941
OL tags	3,868
Interchromosomal alignments	577
Multiple alignments	94
Non-collinear tags	489
Collinear tags	2,708

3.4. Genotype calling with an OL-nized genome sequence

Considering situations that require RRS-based genotyping, the genome sequences of the parents of a given population are not always available. In contrast, the genome sequence polisher Pilon is a software to correct the sequence of a genome assembly to improve its quality based on the alignment results of WGR data.[36] Therefore, genotype calling using MCPtaggR was performed using the NB genome assembly and an OL genome sequence that was converted from the NB genome assembly, referred to as an OL-nized genome sequence. Pilon is usually applied repeatedly to improve the quality of a *de novo* assembled genome. Thus, OL-nized genome sequences were obtained by polishing one, two, three, and four times. We compared the SNP positions between the genotype data obtained using the OL genome assembly and the OL-nized genome sequences. Approximately 65% of SNPs (3543 out of 5486) detected using the OL genome assembly were also detected using the OL-nized genome sequence after one round of polishing (Table 4). The rate of common SNPs was increased by two rounds of polishing to nearly 70%. Further polishing (three and four rounds) showed negligibly small improvements in the rate of common SNPs. The genotype-calling performance was evaluated based on the number of recombinations per chromosome and double crossovers within short intervals. Only one round of polishing resulted in similar values in the number of recombinations per chromosome and double crossovers within short intervals compared with those observed in the genotype data using the OL genome assembly (Supplementary Table S5). Unexpectedly, an additional round of polishing (two rounds in total) increased the number of recombinations and double crossovers. In addition, more than two rounds of polishing showed negligible improvements. We also compared the concordance rates of raw genotype calls and estimated dosages. Since the number

Table 4. Concordance rates in genotype calls and estimated dosages

Pilon[a]	Number of markers[b]	Concordance rate (%)[c]
1	1,943, 2,170, 3,543	42.53, 39.07
2	1,678, 2,180, 3,808	92.63, 99.73
3	1,671, 2,170, 3,815	92.72, 99.73
4	1,668, 2,168, 3,818	92.70, 99.74

[a]The number of rounds Plilon was applied for polishing.
[b]The number of markers detected solely in the genotype calling using the assembled OL genome, the OL-nized genome, and markers detected in both are presented, separated by commas.
[c]Concordance rates in genotype calls and estimated dosages between those obtained using the OL genome assembly and the OL-nized genome sequence are presented, separated by commas.

and positions of obtained SNPs were different between the genotype data obtained using different genomes (Table 4), we calculated concordance rates of the raw genotype calls and estimated dosages between the nearest SNPs detected using the OL genome assembly and the OL-nized genomes. One round of polishing resulted in poor concordance rates of genotype calls (42.53%) and estimated dosages (39.07%), whereas two rounds of polishing drastically improved concordance rates to 92.63 and 99.74% for genotype calls and estimated dosages, respectively (Table 4). Similar to the other measurements described above, two or more rounds of polishing showed almost identical values of the concordance rates. Together with the number of recombinations and crossovers, only one round of polishing was insufficient to OL-nized the NB genome assembly. The OL-nized genome sequence obtained by one-round polishing exhibited comparable frequencies of recombinations and double crossovers to those observed in the OL-nized genome that underwent multiple rounds of polishing (Supplementary Table S5). However, the genotype calls and estimated dosages were largely different from those obtained using the OL genome assembly (Table 4). We also found that three or more polishing rounds showed limited improvements in the quality of genotype calling (Table 4 and Supplementary Table S5). Two rounds of polishing using Pilon were sufficient to obtain genotype calls at a similar quality compared with genotype calling using the OL genome assembly (Table 4 and Supplementary Table S5). These results demonstrated that the MCPtagg pipeline potentially works sufficiently with an X-nized genome sequence obtained by polishing the genome of a related species, although the positions of markers and raw genotype calls are different from those obtained using a *de novo* assembled genome.

4. Discussion

The main purpose of our study was to demonstrate the performance of the MCPtagg pipeline. However, we would also like to note the importance of error correction for RRS-based genotype data. In our study, the concordance rate of raw genotype calls was only 67.34% between MCPtaggR and TASSEL-GBS, as described in Section 3.2. However, a high concordance rate of estimated dosages (98.76%) was observed. Even though the raw genotype calls contained largely different genotype information with different error patterns, GBScleanR could estimate highly similar dosages. As GBScleanR was designed to incorporate marker-specific error patterns in the HMM for robust genotype correction against error-prone markers, error correction and dosage estimation could be achieved by picking up reliable genotype information while eliminating negative impact from error-prone markers.

Our study presented the deleterious effects of error-prone markers and the causes of these errors. Although we observed the quite high concordance rate of the estimated dosages between MCPtaggR and TASSEL-GBS, TASSEL-GBS resulted in 20 and 1,650% higher number of recombinations and double crossovers within intervals of 0–1 Mb stretches compared with MCPtaggR (Supplementary Tables S3 and S4). Considering the high concordance rate of the estimated dosages, the increase in the recombinations might be caused by the increase of unexpected double crossovers. Unfortunately, the assessment relying on the overall concordance rate was unable to effectively detect these specific local errors, mainly because such local errors usually have limited effects on the overall measurement. As the present study demonstrated, the MCPtagg pipeline effectively eliminates these local errors that are caused by error-prone markers. RRS-based genotyping in populations derived from crosses between very close relatives might result in a few error-prone markers because their genome sequences have conserved sequence similarity and collinearity. However, the more distant relatives that were crossed to produce populations, the more error-prone markers contaminated the genotype data. Geneticists and breeders use wild relatives of cultivated varieties in distant crosses, particularly in genetic studies and breeding programs. RRS-based genotyping is a popular and standard method used in genetic studies and breeding programs. However, contamination with error-prone markers increases the risk of misleading genotype data.[11–14] Even though some error-prone markers can be filtered out based on marker statistics, such as allele frequency and heterozygosity, there is no gold standard for filtering criteria. In contrast, MCPtaggR provides an easy-to-use reliable genotype calling pipeline that basically never requires filtering based on user-specified arbitrary criteria. In addition, our study also showed efficiency in detecting SNP markers. We performed genotype calling using TASSEL-GBS and MCPtaggR for the same read dataset obtained from a rice F_2 population. Typically, RRS genotype data are subjected to intensive filtering based on the proportion of missing data, heterozygosity, and minor alleles per marker. Genotyping pipelines for bi- and multi-parental populations also filter out SNP markers for which one or more parental samples showed missing genotype calls due to missing read observations. Because sequence reads are acquired stochastically by a sequencer, RRS-based genotype data contain numerous markers that can be filtered out, particularly in low-read-coverage genotype data. Although the parental samples of the rice F_2 population were sequenced at three times higher read coverage, the filtering of the genotype data provided by TASSEL-GBS retained only 2,973 markers while 8,190 markers were obtained just after pruning markers based on their physical distances. In contrast, MCPtaggR provided 5,486 markers, although it filtered out markers by comparing parental genome sequences prior to genotype calling. Missing genotype calls in either NB or OL were found at 3,646 markers in the genotype data generated by MCPtaggR. In a typical filtering procedure, these markers are removed. However, the reliability of the markers was supported by a pre-survey of SNP markers in the MCPtagg pipeline. Thus, MCPtaggR successfully retained more markers than TASSEL-GBS. Although an increase in read coverage might change the result, MCPtaggR enables us to efficiently obtain as numerous markers as possible, even from low-coverage sequence data.

One drawback of the MCPtagg pipeline is that it requires high-quality genome assemblies of both parents. However, even in cases where parental genome assemblies are not available, recent developments in bioinformatics tools have enabled us to instantly obtain genome assemblies without conducting costly *de novo* genome assembly. As demonstrated in this study, available genome assemblies can be converted into others using short-read sequences with the genome polishing tool Pilon[36] and methods introduced in previous studies.[38,39] One round of polishing resulted in poor concordance rates between the estimated dosages obtained using the OL genome assembly and the OL-nized genome sequence. This result indicates that only one round

of polishing is insufficient to convert the NB genome assembly to an OL-nized genome sequence that can provide sufficiently accurate genotype information using MCPtaggR in combination with GBScleanR. Two rounds of polishing by Pilon were sufficient to achieve a genotyping quality similar to that of the OL genome (Table 4 and Supplementary Table S5). As a note, this result never implies that the OL-nized genome sequence is identical to the OL genome assembly, but it suggests that the two-round polishing renders an OL-nized genome sequence with sufficient information for genotyping using MCPtaggR. The number of rounds necessary to X-nize a genome assembly might differ depending on target genomes and there is no prior information to estimate the ideal number of rounds. Thus, multiple rounds of polishing are recommended. Nonetheless, our study suggests that the MCPtagg pipeline sufficiently works with converted genomes. Therefore, if a genome assembly is available for species of interest, WGR data from the parents are sufficient to prepare genome assemblies as input for the pipeline, instead of conducting de novo genome assembly. This fact indicates the flexibility and scalability of the pipeline. The read data used for polishing in this study contained 36,708,088 150-bp pair-end reads that corresponded to approximately 10× coverage of the rice genome. Even though the read coverage was not high but rather low, Pilon successfully optimized the NB genome to sufficient quality for genotype calling by MCPtaggR. Nevertheless, an X-nized genome may not reflect genome rearrangements such as translocations and inversions. This result implies that mismapped reads are more likely due to the presence of similar sequences resulting from mutations in the ancestral genome sequences that were originally located elsewhere, rather than originating from translocated and mutated genomic segments.

Collectively, this study demonstrated the superiority of the MCPtagg pipeline over existing genotype pipelines. Although only TASSEL-GBS was tested with GBS-read data as an example, theoretically similar disadvantages would be found in other RRS-based genotyping methods unless the obtained reads were mapped onto a single reference genome. Another limitation of our study is that we tested the pipeline only using the rice F_2 population derived from a cross between NB and OL. However, this fact never diminishes the value of our study and the pipeline. This point was further discussed and justified in Supplementary Discussion. The current implementation of MCPtaggR treats any sequence as a cut site. It is not necessary to specify RE recognition sequences, but one can specify primer sequences used in genotyping based on amplicon sequences such as MIG-seq[40] and GRAS-Di® (TOYOTA, Aichi, Japan). Genotype data generated by any RRS-based genotyping method can be processed using the MCPtagg pipeline by tweaking the parameters specifying the cut sites and digested-tag length filtering in the MCPtaggR package. In contrast, the current version of MCPtaggR has a limitation in its applicability, only supporting biparental populations derived from a cross between inbred lines or selfing species. Nonetheless, the pipeline can be expanded to handle outbred lines and outcrossing species if haplotype-phased genome assemblies are available.[41] MCPtaggR can also genotype polyploid populations if those were derived from a biparental cross of inbred lines or selfing species. Although the resulting data only includes genotype calls for homozygotes of both alleles and heterozygotes without allele dosage information, polyploid allele dosage estimation tools might help to analyse such genotype call data.[42,43] In addition, the pipeline has the potential to be applied to multiparental populations by conducting comprehensive genome comparisons across all possible combinations of parents. Future updates on the MCPtaggR package will address these expansions of the pipeline. MCPtaggR would increase its applicability as the available genome sequences increase and has the potential to become a standard genotyping method, especially for genotyping populations derived from distant crosses, which may normally cause error-prone SNP detection.

Acknowledgements

We would like to thank Editage (www.editage.jp) for the English language editing.

Funding

This work was supported by the Grants-in-Aid for Scientific Research (KAKENHI) from the Japan Society for the Promotion of Science [JP20K15503 and JP23H02185 to T.F.].

Author contributions

T.F. conceptualized the project, developed the software, conducted benchmarking, and wrote the manuscript. T.Y. supervised the project and reviewed the manuscript.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Poland, J.A. and Rife, T.W. 2012, Genotyping-by-sequencing for plant breeding and genetics, *Plant Genome*, **5**. https://doi.org/10.3835/plantgenome2012.05.0005
2. 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.
3. Torkamaneh, D., Boyle, B., and Belzile, F. 2018, Efficient genome-wide genotyping strategies and data integration in crop plants, *Theor. Appl. Genet.*, **131**, 499–511.
4. Unamba, C.I., Nag, A., and Sharma, R.K. 2015, Next generation sequencing technologies: The doorway to the unexplored genomics of non-model plants, *Front. Plant Sci.*, **6**, 1074.
5. Wong, G.K.-S., Soltis, D.E., Leebens-Mack, J., et al. 2020, Sequencing and analyzing the transcriptomes of a thousand species across the tree of life for green plants, *Annu. Rev. Plant Biol.*, **71**, 741–65.
6. Scheben, A., Batley, J., and Edwards, D. 2017, Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application, *Plant Biotechnol. J.*, **15**, 149–61.
7. Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. 2007, Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers, *Genome Res.*, **17**, 240–8.
8. 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.
9. Elshire, R.J., Glaubitz, J.C., Sun, Q., et al. 2011, A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, *PLoS One*, **6**, e19379.

10. Poland, J.A., Brown, P.J., Sorrells, M.E., and 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**, e32253.
11. Günther, T. and Nettelblad, C. 2019, The presence and impact of reference bias on population genomic studies of prehistoric human populations, *PLoS Genet.*, **15**, e1008302.
12. Freeman, T.M., Wang, D., and Harris, J.; Genomics England Research Consortium. 2020, Genomic loci susceptible to systematic sequencing bias in clinical whole genomes, *Genome Res.*, **30**, 415–26.
13. Valiente-Mullor, C., Beamud, B., Ansari, I., et al. 2021, One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads, *PLoS Comput. Biol.*, **17**, e1008678.
14. Furuta, T., Yamamoto, T., and Ashikari, M. 2023, GBScleanR: robust genotyping error correction using a hidden Markov model with error pattern recognition, *Genetics*, **224**, iyad055.
15. Wijnker, E., James, G.V., Ding, J., et al. 2013, The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*, *eLife*, **2**, e01426.
16. Jacob, F.D., John, C.M., Athma, A.P., et al. 2009, Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data, *Bioinformatics*, **25**, 3207–12.
17. Panousis, N.I., Gutierrez-Arcelus, M., Dermitzakis, E.T., and Lappalainen, T. 2014, Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies, *Genome Biol.*, **15**, 467.
18. Ebler, J., Ebert, P., Clarke, W.E., et al. 2022, Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes, *Nat. Genet.*, **54**, 518–25.
19. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. 2012, De novo assembly and genotyping of variants using colored de Bruijn graphs, *Nat. Genet.*, **44**, 226–32.
20. Shajii, A., Yorukoglu, D., Yu, Y.W., and Berger, B. 2016, Fast genotyping of known SNPs through approximate k -mer matching, *Bioinformatics*, **32**, i538–44.
21. Sibbesen, J.A., Maretty, L., and Krogh, A.; Danish Pan-Genome Consortium. 2018, Accurate genotyping across variant classes and lengths using variant graphs, *Nat. Genet.*, **50**, 1054–9.
22. Glaubitz, J.C., Casstevens, T.M., Lu, F., et al. 2104, TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline, *PLoS One*, **9**, e90346.
23. Sakai, H., Lee, S., Tanaka, T., et al. 2013, Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics, *Plant Cell Physiol.*, **54**, e6–e6.
24. Reuscher, S., Furuta, T., Bessho-Uehara, K., et al. 2018, Assembling the genome of the African wild rice *Oryza longistaminata* by exploiting synteny in closely related *Oryza* species, *Commun. Biol.*, **1**, 162.
25. Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S., 2022, Biostrings: efficient manipulation of biological strings. https://doi.org/10.18129/B9.bioc.Biostrings.2022
26. Wright, E.S. 2016, Using DECIPHER v20 to analyze big biological sequence data in R, *R J.*, **8**, 352–9.
27. Li, H. 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997v2, Date of publication 26 May 2013, http://arxiv.org/pdf/1303.3997.pdf, preprint: not peer reviewed.
28. Morgan, M., Pagès, H., Obenchain, V., and Hayden, N. 2023, Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import, Accessed 24 Apr 2023. https://bioconductor.org/packages/Rsamtools
29. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. 2018, MUMmer4: a fast and versatile genome alignment system, *PLoS Comput. Biol.*, **14**, e1005944.
30. Smit, A., Hubley, R., and Green, P. RepeatMasker Open-4.0. 2013-2015, Accessed 4 May 2023. http://www.repeatmasker.org
31. Liao, Y., Smyth, G., and Shi, W. 2013, The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote, *Nucleic Acids Res.*, **41**, e108–e108.
32. Danecek, P., Auton, A., Abecasis, G., et al.; 1000 Genomes Project Analysis Group. 2011, The variant call format and VCFtools, *Bioinformatics*, **27**, 2156–8.
33. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. 2012, A high-performance computing toolset for relatedness and principal component analysis of SNP data, *Bioinformatics*, **28**, 3326–8.
34. Liao, Y., Smyth, G., and Shi, W. 2019, The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads, *Nucleic Acids Res.*, **47**, e47–e47.
35. Furuta, T., Ashikari, M., Jena, K., Doi, K., and Reuscher, S. 2017, Adapting genotyping-by-sequencing for rice F2 populations, *G3–Genes Genom Genet*, **7**, 881–93.
36. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., et al., 2014, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One*, **9**, e112963.
37. Chen, M., Presting, G., Barbazuk, W., et al. 2002, An integrated physical and genetic map of the rice genome, *Plant Cell*, **14**, 537.
38. Schneeberger, K., Ossowski, S., Ott, F., et al. 2011, Reference-guided assembly of four diverse Arabidopsis thaliana genomes, *Proc. Natl. Acad. Sci. USA*, **108**, 10249–54.
39. Lischer, H.E.L. and Shimizu, K.K. 2017, Reference-guided de novo assembly approach improves genome reconstruction for related species, *BMC Bioinf.*, **18**, 474.
40. Suyama, Y. and Matsuki, Y. 2015, MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform, *Sci. Rep.*, **5**, 16963.
41. Guk, J.Y., Jang, M.J., Choi, J.W., Lee, Y.M., and Kim, S. 2022, De novo phasing resolves haplotype sequences in complex plant genomes, *Plant Biotechnol. J.*, **20**, 1031–41.
42. Gerard, D., Ferrão, L.F.V., Garcia, A.A.F., and Stephens, M. 2018, Genotyping polyploids from messy sequencing data, *Genetics*, **210**, 789–807.
43. Clark, L., Lipka, A.E., and Sacks, E.J. 2019, polyRAD: genotype calling with uncertainty from sequencing data in polyploids and diploids, *G3–Genes Genom Genet*, **9**, 663–73.