Genotype calling and haplotype inference from low coverage sequence data in heterozygous plant genome using HetMap

Hao Gong1 · Bin Han2

Received: 4 January 2022 / Accepted: 12 April 2022 / Published online: 3 May 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Key message This study developed a new genotyping method that can accurately infer heterozygous genotype information from the complex plant genome sequence data, which helped discover new alleles in the association studies.

Abstract Many software packages and pipelines had been developed to handle the sequence data of the model species. However, Genotyping from complex heterozygous plant genome needs further improvement on the previous methods. Here we present a new pipeline available at https://github.com/Ncgrhg/HetMapv1) for variant calling and missing genotype imputation from low coverage sequence data for heterozygous plant genomes. To check the performance of the HetMap on the real sequence data, HetMap was applied to both the F1 hybrid rice population, which consists of 1495 samples and the wild rice population with 446 samples. The high coverage sequence data of four hybrid rice accessions and two wild rice accessions, which were also included in low coverage sequence data, were used to validate the accuracy of genotype inference. The validation results showed that HetMap achieved significant improvement in heterozygous genotype inference accuracy (13.65% for hybrid rice, 26.05% for wild rice) and total accuracy compared with similar software packages. The application of the new genotype with the genome-wide association study also showed improvement of association power in wild rice awn length phenotype. It could archive high genotype inference accuracy in low sequence coverage in a small population with both the natural and constructed recombination population. HetMap provided a powerful tool for the heterozygous plant genome sequence data analysis, which may help to discover new phenotype regions for the plant species with the complex heterozygous genome.

Abbreviations

BWA Burrows-Wheeler aligner
MAF Minor allele frequency
SNP Single nucleotide polymorphism
VCF Variant call format
NGS Next-generation sequence
API Application programming interface

GWAS Genome-wide association study
IRGSP International rice genome sequence project

Introduction

Next-generation sequencing (NGS) technology has been widely applied in the genome sequencing area to explore the mysteries in the genome. The short reads of NGS data can be used for assembly of the reference genome, exploration of population structure (Huang et al. 2012), conduction of GWAS (Huang et al. 2011; Flint et al. 2012), and so on. Besides the important model species like human and mouse, genome research on some other plant species, which are of great economic value, such as maize (Lai et al. 2010; Tian et al. 2011) and rice, is also important. One primary step of NGS data analysis is to do genotyping from the short reads data to generate the genome variation map. There are already some excellent pipelines like SNPtools (Wang et al. 2013), GotCloud (Jun et al. 2015), and OutcrossSeq (Chen et al. 2021) that provide the whole workflow from raw sequence
data to imputed genotype matrix. However, most popular software and pipelines are devote to tackling the model species genome sequence data, such as human and mouse, with a high recombination rate and medium repeat sequences. While some heterozygous plant genomes have rich repeat sequences and low recombination rates (For example, wild rice and hybrid maize). There is a great difference in the heterozygous rate of genome variants between the plants and animals (Jaramillo-Correa JP et al. 2010). For example, the heterozygous genotype rate in plants can be reduced dramatically due to the existence of self-pollination. Many software packages designed for model species cannot directly be used to tackle the plant genome sequence data. Here we have developed a pipeline HetMap to solve this problem. Normal cultivated rice often has a low heterozygous rate. A high heterozygous genotype rate in rice can be found in the F1 hybrid rice population, which is caused by the direct cross of two homozygous parent lines with different alleles in the genome. It can also be found in a natural heterozygous population, which is a combination of genetic variants from different wild rice accessions. The sequence data of hybrid rice (Huang et al. 2015) and wild rice (natural population, Huang et al. 2012) are used to test our software package. We archive great improvement in genotype inference accuracy compared with other popular software packages.

Materials and methods

Plant material, DNA isolation and RNA isolation

The detailed information for the four high coverage sequence hybrid rice accessions is described in the previous study (Huang et al. 2012). The DNA extraction, library preparation, and genome sequence with Illumina sequencing machine are also described in the previous publication.

Short reads alignment and SNP calling

BWA (version: 0.7.17-r1188) is used to align all the pair-end sequence data to rice reference genome (IRGSP version4) with the default parameters to generate the SAM (Sequence Alignment/Map) format file. Samtools (version 1.7) is used to convert the SAM format file to the BAM format. They are sorted by the genome coordination to generate the sorted BAM files, which are subject to local realignment and base recalibration module of GATK package (version: 4.2.5.0) with default parameters. All the BAM files of wild rice and hybrid rice are also submitted to other five different genotype callers to make a detailed comparison. The five genotype callers are SNPtools (version: 1.0), freebayes (version: 1.3.6, Garrison et al. 2012), deepvariant (version: v0.9.21, Poplin et al. 2018), bcftools (Samtools mpileup version: 1.7, Li 2011), and GATK (DePristo et al. 2011). They are processed as described in the user manual to generate input file for the downstream analysis (Yao et al. 2020, Sandmann et al. 2017). In this study, HetMap first parses the BAM file using Samtools API. We search for the bases in the sequencing reads that are different from the reference genome. The reads with mapping quality lower than 60 and reads mapping ratio smaller than 0.82 are discarded. HetMap only outputs the polymorphic sites that at least appear twice in the population. All the bases that cover the polymorphic sites detected in the previous step and their related sequencing quality score are extracted from the BAM file. HetMap summarizes all the extracted bases for each polymorphic site to generate the missing rate and minor allele frequency. Only the polymorphic sites with minor allele frequency higher than 0.05, missing rate lower than 0.4, and physical distance with neighboring SNPs bigger than 10 bp are kept. Using the chosen polymorphic sites, we calculate genotype likelihood probability using the reads depth and sequence quality score information. Beagle is used to infer the genotype from the raw genotype matrix files generated by all six variant callers.

SNP validation

HetMap uses the high coverage sequence data of four hybrid rice accession and two wild rice accessions for the SNP inference accuracy validation. They are also included in the low coverage sequence data. We compare the genotype called in the low coverage sequence data with the genotype called in the high coverage sequence data to check the inference accuracy.

The penalty of reads count based on Phred score information

The Phred quality score (q) for the base in the fastq format was calculated using the log-transformed error probability (p) by the following equation.

\[ q = -10 \times \log 10(p) \]

we can calculate the error probability of the base using the equation \( p = 10^{q/(-10)} \). For example, if we get three reads supporting the reference allele (Phred score: a1, a2, and a3) for a specific position. If the Phred score for a base is higher than 19, there is no penalty, and the base count is set as 1. If the Phred score of a base is smaller than 20, we integrate the penalty by reducing the base count based on the error probability (refined count = 1*(1 - 10^{q/(-10)})). If the Phred score for all the three bases (a1, a2, and a3) were higher than 19, the base count is 3 for our example datasets. If all the three bases were below 20, the base
count is calculated as follows. refined count = 1*(1− 10^{a1/(-10)}) + 1*(1− 10^{a2/(-10)}) + 1*(1− 10^{a3/(-10)})

Genotype likelihood calculation

For all the three possible kinds of genotype likelihood, we set them as the homozygous reference allele, homozygous alternative allele, and heterozygous allele. The base coverage information for the different alleles is integrated with an improved Binomial model. As there is a high sequencing error for the second-generation sequence data, we give a penalty score based on the base Phred score information in the integration of reads coverage for the genotype likelihood calculation. After we calculate the genotype likelihood for all three kinds of genotypes, we normalize them to get the final probability.

\[ Pr_{ref} = \prod_{i=1}^{m} Pr_{ref} \times \prod_{i=1}^{n} Er_{alt} \]  

(1)

\[ Pr_{alt} = \prod_{i=1}^{m} Pr_{alt} \times \prod_{i=1}^{n} Er_{ref} \]  

(2)

\[ Pr_{Het} = \prod_{i=1}^{m} Pr_{het} \times \prod_{i=1}^{n} Pr_{het} \]  

(3)

\[ ProbAA = Pr_{ref} / (Pr_{ref} + Pr_{alt} + Pr_{Het}) \]  

(4)

\[ ProbAB = Pr_{Het} / (Pr_{ref} + Pr_{alt} + Pr_{Het}) \]  

(5)

\[ ProbBB = Pr_{alt} / (Pr_{ref} + Pr_{alt} + Pr_{Het}) \]  

(6)

ProbAA, ProbAB, and ProbAB represent the final normalized value for the three kinds of genotype likelihood.

Cultivation of the wild rice and the phenotyping of the traits

In this study, we have planted and phenotyped the 446 wild rice accession in Lin shui country Hai Nan province China from 2012 to 2014. Only few cultivated rice accessions have a short awn length, and most cultivated rice accessions have no awn. However, Most wild rice accessions have a long awn length. The awn is stripped from the seed before we measure the awn length and the seed length. The awn is pulled straight and measured with a ruler. We measure the awn length for five seeds, and their average value is used as the final awn length for each wild rice accession.

Data simulation

Reads data of different sequence coverage, which ranges from the 0.25 fold coverage to 30 fold coverage, are simulated by randomly sampling from the high coverage sequence data. The sequence reads are aligned against the reference genome to generate the SAM file. Samtools is used to convert the SAM format to the sorted BAM file. HetMap and SNPTools are used to generate the genotype likelihood files using these sorted BAM files. The genome variants are called for each sample using the default parameters for the rest of four different genotype callers (GATK, freebayes, Samtools, and deepvariant). Self-customized perl scripts are used to merge the sample genotype information to generate the raw genotype matrix for the population with missing data. All the genotype likelihood files and raw genotype matrix are submitted to beagle imputation. The genotypes called in the simulated data are compared against those called in the high coverage sequence data to get the accuracy.

General workflow of the HetMap pipeline

HetMap pipeline is implemented in C++ language. It takes standard BAM files as input. The output file is in VCF format. The workflow from raw sequence reads to the final imputed genotype is outlined in Fig. 1. Many genotype imputation software packages can be used to handle the output file of HetMap. Here Beagle is chosen as an example.

The whole pipeline can be divided into five sections. The HetMap pipeline starts with the sorted BAM file. Generally, multiple sequence alignment software packages can be used to align the short reads to the reference genome to generate the input file. Firstly, the “Call_snp” program is used to detect the genetic variants from the sequence data that are different from the reference genome. Secondly, the “Sum_snp” program can summarize all the genetic variants of the whole population to generate the candidate polymorphic sites by using some filtering parameters. Thirdly, the “Chose_pile” program can extract all the bases that are aligned to the candidate polymorphic sites. Fourthly, the “filter_snp” program extracts the whole base information that’s aligned to the candidate polymorphic sites. Fifthly, the “Call_prob” program can calculate the three kinds of genotype likelihood in the entire population for the whole polymorphic sites. The output file in VCF format (version 4.0) can be used as an input file for many popular genotype imputation software packages.

Detection of polymorphic sites

HetMap pipeline uses the application programming interface (API) provided by the Samtools to parse the compressed BAM file. Before detecting genetic variants from
the next-generation sequence data, HetMap first filters the sequence reads that has low mapping score (default, < 60) and small reads matching ratio (default, < 0.82). All the variant sites in the sequence data, which are different from the reference genome, can be genotyped for each accession. As some variant sites in the sequence reads may be caused by sequencing error or mismatch of sequence reads, only non-reference alleles that appear at least twice in the whole population are kept to generate the primary polymorphic sites. All the genetic variants are considered in the determination of primary polymorphic sites, we need all the four base information to get the final polymorphic sites. All the sequenced bases, which are aligned to the primary polymorphic sites, are pulled down from the sequence alignment files. The base information and Phred score information are all included in the output file.

A hash function is often used to get selected polymorphic sites from the sequence reads in many software packages. However, it needs a large amount of calculation to get all the bases aligned to the candidate polymorphic sites by scanning whole sequence data. As the sample size used for the population genetics analysis is often very huge, counting the frequency of all the four bases for the whole genome becomes a challenging problem. Sort and compare is another method to pull down chosen polymorphic sites from the massive sequence data. By comparing the sorted reads mapping position with the polymorphic sites, the bases aligned to the polymorphic sites can also be extracted. It only needs a simple comparison instead of the complex hash function for the input BAM file. The simple sort and compare method is impractical for overlapping the mapping position with the sequence reads. In the HetMap pipeline, an improved sort
and compare method is used. We use a dual sorting method to compare the mapping position of the reads with the polymorphic site. The genome position of the polymorphic sites can increase or decrease dynamically with the change of reads mapping range. After we get all the four base information for each candidate polymorphic site, we can filter the polymorphic sites with multiple criteria to get the final polymorphic sites.

**Genotype likelihood calculation using improved binomial model**

There are three kinds of genotype for each polymorphic site that consists of two alleles. They are homozygous reference allele (RR), homozygous alternative allele (AA), and heterozygous allele (RA). HetMap doesn’t consider the tri-allele polymorphic sites here, as they only occupy a small part of the genome. However, HetMap outputs all four base information for the polymorphic site. Tri-allele sites can still be detected from the output file with some easy customization. HetMap calculates genotype likelihood for all possible three kinds of genotypes (Eqs. 1–6, methods). As there is high sequencing error for the short reads sequence data, some bases with low Phred quality score (< 20) are often discarded in the genotype likelihood calculation process. However, the number of reads covering a specific polymorphic site is small in the low coverage sequence data. Hence HetMap integrates the penalty method based on their Phred score information to take the bases that have low Phred scores into consideration. If both an alternative allele and a reference allele are detected in the polymorphic site. The probability for this allele to be a homozygous reference allele is a combination of the probability of this reference allele plus the error probability of the alternative allele. HetMap outputs the genotype likelihood file in VCF format. Multiple software packages can be used for downstream analysis.

**Results**

**Application of HetMap to the F1 hybrid rice population**

The published sequence data for the F1 population (Huang et al. 2015) are used to test the application of HetMap to the F1 population. The data contain 1495 pair-end sequenced accessions with the reads length being 96 bp and the average sequence coverage being 2.2 fold. The short sequence reads are aligned to the rice reference genome to generate standard BAM file (Li et al. 2009). After detecting the polymorphic sites with HetMap, it results in 21,016,942 non-singleton SNPs. Of them 17,670,585 SNPs have a minor allele frequency bigger than 0.05, with the average missing rate being 0.29 across the whole population (Supplementary Fig. 1). For all the polymorphic sites, only those SNPs with the missing rate smaller than 0.35 and the minor allele frequency bigger than 0.05 are kept for the downstream analysis. After filtering all these positions, we got a total of 1,432,221 SNPs and of them, 125,768 are coding SNPs (Supplementary Fig. 2, Supplementary Table 1). The genotype likelihood for the polymorphic sites is submitted to Beagle (version 4.2) to infer the missing genotype (Brown et al. 2007).

To validate the accuracy of genotype inference, four high coverage sequence hybrid rice accessions are used as the gold standard, which is sequenced with an average of 47.5 fold coverage. The four accessions are also included in the low coverage sequence data. Genotypes called by the HetMap pipeline from the low coverage sequence data are compared with those inferred from high coverage sequence data to check genotype inference accuracy. The average total accuracy is found to be 98.20%. The heterozygous genotype accuracy is 98.31%, and the homozygous genotype accuracy is 98.15% (Table 1).

To compare with other software packages in the genotype inference accuracy, another 5 different genotype callers are selected (SNPtools, GATK, deepvariant, and Samtools mpileup) to analyze the whole F1 samples again. We run the software as they are suggested in its reference manual with the default parameters. Except for SNPtools, which can directly generate genotype likelihood file for beagle imputation, the left four software packages call genotype for each sample separately. Self-customized perl scripts are used to merge the sample data to generate the raw genotype matrix with missing data. All the matrix datasets are also submitted to beagle for genotype imputation with the same parameter. SNPtools was chosen as an example to make a

| Material   | Method  | Accession | Total    | Heterozygous | Homolog  |
|------------|---------|-----------|----------|--------------|----------|
| Hybrid rice | SNPtools | Z175      | 0.9471   | 0.8488       | 0.9855   |
| Hybrid rice | SNPtools | Z341      | 0.9653   | 0.9028       | 0.9861   |
| Hybrid rice | SNPtools | Z446      | 0.9585   | 0.8679       | 0.9881   |
| Hybrid rice | SNPtools | Z789      | 0.9463   | 0.8409       | 0.9865   |
| Hybrid rice | Hetamp  | Z175      | 0.9803   | 0.9826       | 0.9795   |
| Hybrid rice | Hetamp  | Z341      | 0.9820   | 0.9858       | 0.9808   |
| Hybrid rice | Hetamp  | Z446      | 0.9841   | 0.9840       | 0.9842   |
| Hybrid rice | Hetamp  | Z789      | 0.9813   | 0.9802       | 0.9818   |
| Wild rice  | SNPtools | W0600     | 0.8683   | 0.5328       | 0.8852   |
| Wild rice  | SNPtools | W1943     | 0.9077   | 0.6446       | 0.9304   |
| Wild rice  | Hetamp  | W0600     | 0.9125   | 0.7200       | 0.9375   |
| Wild rice  | Hetamp  | W1943     | 0.9595   | 0.7694       | 0.9659   |
For SNPtools, we get 7,603,949 polymorphic sites with an average SNP density of 19.90/kb across the genome. The allele frequency of 58.16% polymorphic sites is below 0.05 (Supplementary Fig. 1, Supplementary Fig. 3). Of the two pipelines, 1,268,169 SNPs are detected by both methods (Fig. 2). A similar overlapping rate was also found for each of the six genotype callers (Supplementary Fig. 4). The genotype inference accuracy is also validated using the high coverage sequence data. The average total accuracy is 95.43% (98.66% for the homolog genotype and 86.50% for the heterozygous genotype, Table 1, Supplementary Table 2). Except for the SNPtools, the accuracy for the rest four genotype callers is much lower than the HetMap method. While the homozygous genotype inference accuracy of our pipeline is similar to that of SNPtools, HetMap gets significant improvement in the heterozygous genotype accuracy (chi-square test, \( P < 0.01 \)) and the total accuracy (chi-square test, \( P < 0.01 \)). About an average 21.75% of polymorphic sites in the genome of hybrid rice are heterozygous genotypes. Hence, a high accuracy rate for the heterozygous genotype is essential for the downstream analysis.

**Application of HetMap to natural wild rice population**

As hybrid rice is the F1 population, we also test our pipeline on the natural heterozygous population. HetMap is applied to the 446 published wild rice genome sequence data (Huang et al. 2012), which is sequenced with average two fold coverage. Different from the normal cultivated rice there is existence of outcross in wild rice. Former studies estimated that 1.56% of the wild rice genome was heterozygous, which was much smaller than that of hybrid rice (Huang et al. 2015). The sequence data of wild rice are processed in the same way as that of hybrid rice. After filtering the SNPs with default parameters, it results in 2,360,987 SNPs. Of them, 126,930 are coding SNPs (Supplementary Fig. 5, Supplementary Table 1). The average SNP density across the genome is 6.17/kb. The minor allele frequency of 86% SNPs is bigger than 0.05 (Supplementary Fig. 6).

To validate the accuracy of the HetMap in the wild rice population, two high sequence coverage wild rice accessions are used as the golden standard. The total genotype inference accuracy for wild rice is average 93.6% (95.17% for the homozygous genotype, 74.47% for the heterozygous genotype).

The other five different genotype callers are also applied to the 446 wild rice samples. The SNPtools is chosen as an example to be studied. It results in a total of 30,179,926 SNPs. The minor allele frequency for 22.30% of the polymorphic sites is bigger than 0.05 (Supplementary Figs. 6 and 7). The total accuracy for the wild rice sample is 88.80% (90.78% for the homozygous genotype, 58.87% for the heterozygous genotype). It’s found that the HetMap pipeline significantly surpasses the SNPtools accuracy in the wild rice population in the genotype inference including both the heterozygous genotypes and homozygous genotypes. Compared with HetMap, the rest of the four genotype callers generally showed a much lower genotype calling accuracy (Supplementary Table 2).

**Assessment of sequencing reads coverage on the genotype inference accuracy through computational simulation**

To check the effects of sequence coverage on the genotype inference accuracy in the wild rice population and hybrid rice population, we simulate sequence reads of different fold coverage to submit them to all the six different genotype callers. The simulated sequence data ranges from 0.25 fold coverage to 30 fold coverage. (Supplementary Figs. 8 and 9). The simulation result shows that HetMap can archive an average accuracy of 85% with only one fold coverage (Fig. 3) in the hybrid rice sample. The accuracy rate doesn’t increase much after fourfold coverage. However, for the wild rice HetMap needs about 1.3 fold coverage to get an average accuracy of 85%. The slow increase of accuracy only occurs after sixfold sequence coverage. Compared with the other five genotype callers, HetMap can archive better accuracy with low coverage sequence data when the sequence coverage is below 16 (Fig. 3, Supplementary Fig. 9). However, when sequence coverage is bigger than 16 there is no significant difference in the genotype inference accuracy.

Besides the difference in genotype inference accuracy within the different software packages, the hybrid rice
population shows consistently higher genotype inference accuracy than the wild rice population. Only when the sequence coverage for the sequenced accession is high (> 18 fold) no significant difference exists between the wild rice population and hybrid rice population. There is a big difference in the genotype inference accuracy between the wild rice population and hybrid rice population for all six variant callers. The difference in genotype inference accuracy between the two populations is smaller when the HetMap pipeline is applied to the two populations than that of the other software packages.

**Improvement of genome-wide association power with the addition of heterozygous information**

In this study, we have characterized heterozygous genotype in the wild rice population with a new method. A previous study, which generated the homozygous genotype variation for the wild rice, had performed genome-wide association with one phenotype. In this study, we decide to conduct association with the phenotypes using both the homozygous genotype variation map generated in the previous research and the heterozygous genotype used in this study to check their difference. To compare their difference in performance, we check the association result for the traits that detect known phenotype-related genes around the associated peaks. The genome-wide association results for the heterozygous genotype surpass the results generated using the homozygous genotype in nearly both cases (Fig. 4, Supplementary Fig. 10). The association result using the heterozygous genotype has better association accuracy and signal detection rate than the homozygous genotype. Genotype matrix generated by the HetMap package is found to archive the highest association signal near the two known awn related genes than the rest of five genotype callers.

Besides the improvement of GWAS association power for the heterozygous genotype, it can also detect the association signal that is missed by using the homozygous genotype. We take the awn phenotype GWAS association result as an example to show this improvement. We can detect five significant association signals for the awn phenotype association using the homozygous genotype. Previous studies have cloned two awn length related genes using one near inbreed line constructed using W1943 (wild rice) and GLA4 (cultivated rice, Luo et al. 2013; Gu et al. 2015). The two known awn related genes are located near the peak association signal of awn phenotype association result using the heterozygous genotype (Supplementary Table 3). The peak signal of awn phenotype in chromosome 10 also overlaps with one awn phenotype QTL locus mapped in the W1943 and GLA4 near inbreed line. However, no QTL mapping locus of awn phenotype in the W1943 and GLA4 population overlaps with the peak signal in chromosome 4 detected using the homozygous genotype. Besides, more association signals in the awn phenotype result are archived with the heterozygous genotype than that of the homozygous genotype, the heterozygous genotype also explains more phenotype variation in the highest association signals (Supplementary Table 3). For awn phenotype association result, the highest association signal with the heterozygous genotype can explain 40.04 percent phenotype variation. In comparison, the highest association signal with the homozygous genotype can only
explain 27.77 percent phenotype variation. The heterozygous rate for the highest association signal in awn phenotype is as high as 30.71%. So, the heterozygous genotype plays a vital role in the association of awn phenotype. Compared with the other five variant callers, the HetMap developed in this study showed the highest association signal. For the awn length phenotype, the association studies using the heterozygous genotype can detect more association signals and archive a better explanation of the phenotype variation than that of the homozygous genotype for the wild rice awn length phenotype.

Discussion

In this study, we show that the HetMap pipeline can be used to tackle heterozygous plant genome in both the F1 population (hybrid rice) and natural heterozygous population (wild rice). Construction of the F1 population is important for gene mapping and dissection of hybrid vigor, which exists in the species like rice and maize (Duvick et al. 2001, Zhang 2007, Huang et al. 2015). For example, many natural plants like *Miscanthus Sinensis* have a high heterozygous genotype rate (Ma et al. 2012). It is not suitable to submit them to popular genotype processers (DePristo et al. 2011, Li 2011). This study shows that a fine variation map can be obtained by HetMap pipeline for the heterozygous plant genome sequence data using shallow coverage (Supplementary Figs. 9, 10 and 11). Compared with other software packages, the HetMap pipeline shows a higher genotyping accuracy rate in both the F1 population of hybrid rice and the natural wild rice population (Fig. 3, Supplementary Fig. 9). Our software also offers more options for users to customize the input data and filter the variation data by their own needs.

Besides the difference in genotyping accuracy between the HetMap and other five genotype callers, the genotyping accuracy between the natural wild rice and F1 hybrid rice populations is also different (Fig. 3). They may be caused by several reasons. Firstly, the population size of hybrid rice is larger than that of the wild rice population. There is more reads coverage for a specific polymorphic site in a larger population. The missing genotype has more references in the missing genotype inference process. Secondly, there is significant population structure and high population differentiation in the wild rice population. The genetic distance within the hybrid rice population is small.
and all the accessions generally cluster together in the evolution phylogenetic tree analysis (Huang et al. 2015). The missing genotype will have more reference haplotypes to impute from with smaller genetic distance. Thirdly, the heterozygous genotype of the F₁ population is a raw combination of the different alleles between the two parents. It’s evenly distributed across the genome. However, heterozygous genotype in wild rice is more complex. They can come from genetic drift, genetic introgression, or contamination from other wild rice accessions (Phan et al. 2012; Hufford et al. 2013). Fourthly, as we use the cultivated rice as the reference genome, there exists more difference between the wild rice genome and reference genome than that of the hybrid rice. There is more reads coverage for the hybrid rice than the wild rice when mapping the short reads to the reference genome.

Except for the genotype calling accuracy, the running time for the genotype caller to finish a sample is also important. As the HetMap package developed in this study provides the whole workflow from the input BAM files to the raw genotype matrix, the time to generate the raw genotype matrix from 50 randomly chosen samples is evaluated (Supplementary Fig. 12, Supplementary Table 4). The calling time nearly shows exponential growth with the increase of sequence coverage. Though the calling time of Pileup and HetMap are both fast compared with other software packages, less time is needed for HetMap to finish our test dataset.

Many studies had proposed possible solutions for calling the genotype from the low coverage short sequence data recently (Hickey et al. 2019; Chen et al. 2021). However, they are often limited to model species like human or confined to special recombination populations. Our study provides a powerful tool for calling the genotype from both the nature heterozygous population and the recombination populations. Association studies with wild rice awn phenotypes had shown that the improvement of the genotype calling accuracy could help facilitate the discovery of the associated phenotype-related regions. In total, our study provides a powerful tool to detect the accurate genotype from the low coverage heterozygous genotype, which can help facilitate the discovery of the associated regions in the complex heterozygous genome species.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-022-04105-z.

Acknowledgements This work was partially supported by The Professorial and Doctoral Scientific Research Foundation of Huizhou University (2020JB068).

Authors contribution statement BH conceived the project. HG and BH designed and supervised the project and wrote the manuscript. HG performed most of the data analysis.

Funding The authors have not disclosed any funding.

Data availability The software used in this study had been deposited in the open-source platform GitHub. The project home page is https://github.com/Ncgrhg/HetMapv1. We use the published wild rice and hybrid rice sequence data in this study. The data for the wild rice accession were downloaded from the EBI with ERP001143, ERP000729, and ERP000106. The sequence data of the hybrid rice were downloaded from the European Nucleotide Archive with accession PRJEB13735. Two high coverage sequence data for wild rice are also downloaded from public RicePanGenome database (http://www.ncgr.ac.cn/RicePanGenome).

Declarations

Conflict of interests The authors declare that they have no competing interests.

References

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097. https://doi.org/10.1086/521987

Chen M, Fan W, Ji F et al (2021) Genome-wide identification of agronomically important genes in outcrossing crops using Outcrossseq. Mol Plant 14:556–570. https://doi.org/10.1016/j.molp.2021.01.003

DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. https://doi.org/10.1038/ng.806

Duvick DN (2001) Biotechnology in the 1930s: the development of hybrid maize. Nat Rev Genet 2:69–74. https://doi.org/10.1038/35047587

Flint J, Eskin E (2012) Genome-wide association studies in mice. Nat Rev Genet 13:807–817. https://doi.org/10.1038/nrg3335

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN]. https://doi.org/10.48550/arXiv.1207.3907

Gu B, Zhou T, Luo J et al (2015) An-2 encodes a cytokinin synthesis enzyme that regulates awn length and grain production in rice. Mol Plant 8:1635–1650. https://doi.org/10.1016/j.molp.2015.08.001

Hickey LT, Hafeez AN, Robinson H et al (2019) Breeding crops to feed 10 billion. Nat Biotechnol 37(7):744–754. https://doi.org/10.1038/s41587-019-0152-9

Huang X, Zhao Y, Wei X et al (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:39–39. https://doi.org/10.1038/ng.1018

Huang X, Kurata N, Wei X et al (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501. https://doi.org/10.1038/nature11532

Huang X, Yang S, Gong J et al (2015) Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat Commun 6:6258. https://doi.org/10.1038/ncomms7258

Hufford MB, Lubinsky P, Pyhäjärvi T et al (2013) The genomic signature of crop-wild introgression in maize. PLoS Genet 9:e1003477. https://doi.org/10.1371/journal.pgen.1003477

Jaramillo-Correa JP, Verdú M, González-Martínez SC (2010) The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. BMC Evol Biol 10:22. https://doi.org/10.1186/1471-2148-10-22
Jun G, Wing MK, Abecasis GR, Kang HM (2015) An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res 25:918–925. https://doi.org/10.1101/gr.176552.114

Lai J, Li R, Xu X et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1030. https://doi.org/10.1038/ng.684

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genom Hum Genet 10:387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Luo J, Liu H, Zhou T et al (2013) An-1 encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. Plant Cell 25:3360–3376. https://doi.org/10.1105/tpc.113.113589

Ma X-F, Jensen E, Alexandrov N et al (2012) High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid Miscanthus sinensis. PLoS ONE 7:e33821. https://doi.org/10.1371/journal.pone.0033821

Phan PDT, Kageyama H, Ishikawa R, Ishii T (2012) Estimation of the outcrossing rate for annual Asian wild rice under field conditions. Breed Sci 62:256–262. https://doi.org/10.1270/jsbbs.62.256

Poplin R, Chang P-C, Alexander D et al (2018) A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36:983–987. https://doi.org/10.1038/nbt.4235

Sandmann S, de Graaf AO, Karimi M et al (2017) Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep 7:43169. https://doi.org/10.1038/srep43169

Tian F, Bradbury PJ, Brown PJ et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:159–162. https://doi.org/10.1038/ng.746

Wang Y, Lu J, Yu J et al (2013) An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res 23:833–842. https://doi.org/10.1101/gr.146084.112

Yao Z, You FM, N’Diaye A et al (2020) Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinform 21:360. https://doi.org/10.1186/s12859-020-03704-1

Zhang Q (2007) Strategies for developing green super rice. Proc Natl Acad Sci USA 104:16402–16409. https://doi.org/10.1073/pnas.0708013104

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.