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

Hao Gong (mygonghao@163.com)
Huizhou University https://orcid.org/0000-0002-0327-6673
Bin Han
National Center for Gene Research Chinese Academy of Sciences

Research Article

Keywords: genotyping, genome wide association study, heterozygous genotype, low coverage sequence

Posted Date: January 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1220819/v1

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 F_1 hybrid rice population which consists of 1495 samples and wild rice population with 446 samples. Four high coverage sequence hybrid rice accessions and two high coverage sequence wild rice accessions, which were also included in low coverage sequence data, are used to validate the genotype inference accuracy. The validation results showed that HetMap archived significant improvement in heterozygous genotype inference accuracy (13.65% for hybrid rice, 26.05% for wild rice) and total accuracy compared with other similar software packages. The application of the new genotype with the genome wide association study also showed improvement of association power in two wild rice phenotypes. It could archive high genotype inference accuracy with low sequence coverage with a small population size with both the natural population and constructed recombination population. HetMap provided a powerful tool for the heterozygous plant genome sequence data analysis, which may help the discover of new phenotype regions for the plant species with complex heterozygous genome.

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.

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 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 populate software and pipelines are devoted to tackle the model species genome sequence data. For example human, which has high recombination rate and medium repeat sequences. While some heterozygous plant genomes have rich repeat sequences and low recombination rate (For example, wild rice and hybrid maize speices). There is much difference in the heterozygous rate of SNP between the

plants and animals (Jaramillo-Correa JP et al. 2010). For example, heterozygous genotype rate in plants can be reduced dramatically due to existence of self-pollination. Many software packages designed for model species can not be used directly to tackle the plant genome sequence data. Here we have developed a pipeline HetMap to fill this gap. Normal cultivated rice often has a low heterozygous rate. High heterozygous genotype rate in rice can be found in the F_1 hybrid rice population, which is caused by the direct cross of two homologue parent lines with different alleles in the genome. It can also be found in 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) is used to test our software package. We archive significant improvement in genotype inference accuracy compared with another popular software package (chi-square test P-value < 0.01).

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 is used to align all the short reads to rice reference genome (IRGSP version4) with the default parameters for the pair-end sequence data to generate the SAM format file. Samtools is used to convert the SAM format file to the bam format. They are sorted by the genome coordination to generate the sorted bam file. All the bam files of wild rice and hybrid rice are also submitted to the SNPtools pipeline. They are processed as described in the manual. In this study we first parse the bam file using Samtools API. We search for the bases in the sequencing reads that are different from the reference genome. The reads that have 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 that have minor allele frequency higher than 0.05, missing rate lower than 0.4 and physical distance with neighboring SNPs bigger than 10bp are kept. Using the chosen the polymorphic sites we calculate genotype likelihood probability using the reads depth and sequence quality score information. Beagle is used to infer the genotype using the output genotype likelihood file.

SNP validation

HetMap uses four high coverage sequence hybrid rice and two high coverage sequence wild rice accessions for the SNP inference accuracy validation of hybrid rice and wild rice. They are also included

in the low coverage sequence data. We compare the genotype called in the low coverage sequence data against the genotype called in the high coverage sequence data to check the inference accuracy.

Genotype likelihood calculation

For all the three possible kinds of genotype likelihood we set them as the homologue reference allele, homologue alternative allele and heterozygous allele. The base coverage information for the different alleles is integrated with an improved Binomial model. As there is 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 final probability.

Prref= $\prod_{i=1}^{m} \mathrm{Pr}_{\mathrm{ref}} \times \prod_{i=1}^{n} \mathrm{Er}_{\mathrm{alt}}$ (1)

Pralt= $\prod_{i=1}^{m} \mathrm{Pr}_{\mathrm{alt}} \times \prod_{i=1}^{n} \mathrm{Er}_{\mathrm{ref}}$ (2)

PrHet= $\prod_{i=1}^{m} \mathrm{Pr}_{\mathrm{het}} \times \prod_{i=1}^{n} \mathrm{Pr}_{\mathrm{het}}$ (3)

ProbAA= Prref /(Prref + Pralt + PrHet) (4)

ProbAB= PrHet /(Prref + Pralt + PrHet) (5)

ProbBB= Pralt /(Prref + Pralt + PrHet) (6)

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

Cultivation of the wild rice and the phenotyping of the wild rice traits

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

Data simulation

We simulate different sequence coverage ranging from the 0.25 fold coverage to 30 fold coverage 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 bam format and generate the sorted bam file. HetMap and SNPtools are used to generate the genotype likelihood files using these sorted bam files..The genotype likelihood files are submited to beagle

imputation. The genotype called in the simulated data is 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 Figure 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 the all the genetic variants of the whole population to generate the candidate polymorphic sites by using some filtering parameters. Thirdly, “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 whole population for the whole polymorphic sites. The output file, which is in VCF format (version 4.0) can be used as input file for many popular genotype imputation software packages.

Besides research on the genome sequence data, detecting variants from the RNA-sequence data is also of great importance. HetMap has an extended function to detect variants from the RNA sequence data. As the HetMap pipeline is compatible with many popular software packages, its individual section can also be used separately in the analysis process.

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 that are different from the reference genome can be detected for each accession. Because 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. As only the genetic variants are taken into consideration 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.

Hash function is often used to get selected polymorphic sties from the sequence reads in many software packages. However, it needs large amount of calculation to get all the bases that are aligned to the candidate polymorphic sites by scanning whole sequence data using hash function. As the sample number which is used for the population genetics analysis is often huge, it becomes a more serious 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 that are 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. For the overlap of mapping position with the sequence reads, simple sort and compare method is impractical. In HetMap pipeline an improved sort and compare method is used. We use a dual sorting method in the comparison of the mapping position of the reads with the polymorphic site. The genome position of the polymorphic sites can increase and drecrease 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

For each polymorphic site that consists of two alleles, there are three kinds of genotype. They are homologue reference allele (RR), homologue alternative allele (AA) and heterozygous allele (RA). HetMap doesn't take the tri-allele polymorphic sites into consideration here, as they only occupy a small part of the genome. However, HetMap outputs all the 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 (Equation 1-6, methods). As there is high sequencing error for the short reads sequence data, some bases which have 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 penalty method based on their phred score information to take the bases that have low phred score 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 homologue 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 F_1 hybrid rice population

The published sequence data of 1495 samples (Huang et al. 2015) for the F_1 population are used to test the application of HetMap to F_1 population. The data contains 1495 pair end sequenced accessions with the reads length being 96bp and the average sequence coverage being 2.2 fold. The short sequence reads are aligned to rice reference genome to generate standard bam file for HetMap (Li et al. 2009). After detecting the polymorphic sites with HetMap it results in a total of 21,016,942 no-singleton SNPs. Of

them 17,670,585 SNPs are found to have a minor allele frequency bigger than 0.05 with the average missing rate being 0.29 across the population (Supplementary Figure 1). For all the polymorphic sites only those SNPs with the missing rate being smaller than 0.35 and the minor allele frequency being 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 Figure 2). The genotype likelihood for the polymorphic sites are submitted to Beagle (version 4.2) to infer the missing genotype (Browning 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 average 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%. Of them the heterozygous genotype accuracy is 98.31% and the homologue genotype accuracy is 98.15% (Table 1).

Table 1
Genotype inference accuracy of the two methods in different materials. We list all the total genotype inference accuracy for the two methods and the two populations.

Material	Method	Accession	Total	Heterozygous	Homologue
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

To compare with other software packages in the genotype inference accuracy a state of art pipeline SNPtools is selected (Wang et al. 2013) to analyze the whole F_1 samples again. We run the pipeline as it's suggested in its reference manual with the default parameters. As the software can also generate genotype likelihood file for beagle imputation, the final genotype likelihood files are also submitted to beagle for genotype imputation. We get a total of 7,603,949 polymorphic sites with average SNP density

being 19.90/kb across the genome. The allele frequency of 58.16% polymorphic sites is below 0.05 (Supplementary Figure 1, Supplementary Figure 3). Of the two pipelines 1,268,169 SNPs are detected by both methods (Figure 2). The genotype inference accuracy is also validated by using the high coverage sequence data. The total average accuracy is found to be 95.43% (98.66% for the homologue genotype and 86.50% for the heterozygous genotype, Table 1). While the homologue 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 21.75% polymorphic sites in the genome of hybrid rice are heterozygous genotypes. Hence, a high accuracy rate for the heterozygous genotype is important for the downstream analysis.

Application of HetMap to natural wild rice population

As hybrid rice is the F_1 population we also test our pipeline on natural heterozygous population. HetMap is applied to the 446 published wild rice genome sequence data (Huang et al. 2012) which is sequenced with average 2 fold coverage. Different from the normal cultivated rice there is existence of outcross in wild rice. Former studies estimated that 1.56% of wild rice genome was heterozygous, which was much smaller than that of the hybrid rice (Huang et al. 2015). The sequence data of wild rice is processed in the same way as that of hybrid rice. After filtering the SNPs with default parameters, it results in a total of 2,360,987 SNPs. Of them 126,930 are coding SNPs (Supplementary Figure 4).The average SNP density across the genome is 6.17/kb. The minor allele frequency of 86% SNPs is bigger than 0.05 (Supplementary Figure 5).

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 homologue genotype, 74.47% for the heterozygous genotype).

SNPtools pipeline is also applied to the 446 wild rice samples. 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 Figure 5, Supplementary Figure 6). The total accuracy for the wild rice sample is 88.80% (90.78% for the homologue 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 homologue genotypes.

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 both the HetMap and SNPtools pipelines. The simulated sequence data ranges from 0.25 fold coverage to 30 fold coverage. (Supplementary Figure 7). The simulation result shows that HetMap can archive the

accuracy of 85% with only 1 fold coverage (Figure 3) in the hybrid rice sample. The accuracy rate doesn't increase much after 4 fold 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 6 fold genome sequence coverage. Compared with SNPtools HetMap can archive better accuracy with low coverage sequence data when the sequence coverage is below 16 (Figure 3). However, when sequence coverage is bigger than 16 there is no significant difference in the genotype inference accuracy.

Besides the difference of genotype inference accuracy between 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 big difference in the genotype inference accuracy between the wild rice population and hybrid rice population when using the two haplotype inference pipelines to call SNP from the sequence data. The difference of genotype inference accuracy between the two populations is smaller when HetMap pipeline is applied to the two populations than that of SNPtools pipeline.

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. As previous study has used the homologue genotype variation for the wild rice and performed genome wide association with one phenotypes. In this study we decide conduct association with the phenotypes using both the homologue genotype variation map generated in the previous study and the heterozygous genotype used in this study to check their difference. To compare their difference of 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 homologue genotype in nearly both cases (Figure 4, Supplementary Table 2). The association result using the heterozygous genotype has better association accuracy and signal detection rate than the homologue genotype.

The heterozygous genotype can not only improve GWAS association power, it can also detect association signal which is missing by using the homologue 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 by using the heterozygous genotype. While only one significant signal was detected in the awn phenotype association by using the homologue 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 peak association signal of awn phenotype association result using the heterozygous genotype. The peak signal of awn phenotype in the chromosome 10 also overlap 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 overlap with the peak signal in chromosome 4 detected using the homologue genotype. Besides more association signals in the awn phenotype result archived with the heterozygous genotype than the homologue genotype, the heterozygous genotype also explains more phenotype variation in the highest association signals (Supplementary Table 2). For awn phenotype association result the highest association signal with the heterozygous genotype can explain 40.04 percent phenotype variation, while the highest association signal with the homologue genotype can only explain 27.77 percent phenotype variation. We find that the heterozygous rate for the highest association signal in awn phenotype is as high as 30.71%. So the heterozygous genotype plays an important role in the association of awn phenotype. For all the 2 traits the association studies using the heterozygous genotype can detect more association signals than the homologue genotype. In the peak signal that is associated with the wild rice phenotype, the heterozygous genotype can explain phenotype variation.

Discussion

In this study we show that HetMap pipeline can be used to tackle heterozygous plant genome with both the F_1 population (hybrid rice) and natural heterozygous population (wild rice). Construction of F_1 population is important for gene mapping and dissection of hybrid vigor, which exists in the species in like rice and maize (Duvick et al. 2001, Zhang Q F. 2007, Huang et al. 2015). For example, many natural plants like *Miscanthus Sinensis* have high heterozygous genotype rate (Ma et al. 2012). It is not suitable to submit them to popular genotype processers (DePristo et al.2011, Li H. 2011). In this study we show that a fine variation map can be obtained by using shallow sequence coverage with HetMap pipeline for the heterozygous plant genome sequence data. By comparing with other software, HetMap pipeline shows higher genotyping accuracy rate in both the F_1 population of hybrid rice and the natural wild rice population (Figure 3). Our software also offers more options for users to customize the input data and filter the variation data by their own need.

Besides the difference of genotyping accuracy between the SNPtools and HetMap software, the genotyping accuracy between the natural wild rice population and F_1 hybrid rice population is also different (Figure 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 and 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 F_1 population is a raw combination of the different alleles between the two parents. It's even 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 (Phuong 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 (Supplementary Figure 7).

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 were often limited to the 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 or the recombination populations. Association studies with two wild rice phenotypes had shown that the improvement of the genotype calling accuracy can help facilitate the discover 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 discover of the associated regions in the complex heterozygous genome species.

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.

Declarations

Acknowledgments

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

Authors’ contributions

B.H. conceived the project. H.G. and B.H. designed and supervised the project and wrote the manuscript. H.G. performed most of the data analysis.

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 was downloaded from the EBI with ERP001143, ERP000729 and ERP000106. The sequence data of the hybrid rice were download from the European Nucleotide Archive with accession PRJEB13735. Two high coverage sequence data for wild rice are also download from public RicePanGenome database (http://www.ncgr.ac.cn/RicePanGenome).

Competing interests

The authors declare that they have no competing interests.

References

1. 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
2. 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
3. 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
4. Duvick DN (2001) Biotechnology in the 1930s: the development of hybrid maize. Nat Rev Genet 2:69–74. https://doi.org/10.1038/35047587
5. Flint J, Eskin E (2012) Genome-wide association studies in mice. Nat Rev Genet 13:807–817. https://doi.org/10.1038/nrg3335
6. 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
7. Hickey LT, Hafeez N, Robinson A et al (2019) Breeding crops to feed 10 billion. Nat Biotechnol 37:744–754. https://doi.org/10.1038/s41587-019-0152-9
8. 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
9. 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
10. 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 44:32–39. https://doi.org/10.1038/ng.1018
11. Hufford MB, Lubinksy 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
12. 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
13. 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
14. 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
15. 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
16. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242
17. 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
18. 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
19. 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
20. 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
21. 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
22. Zhang Q (2007) Strategies for developing Green Super Rice. Proc Natl Acad Sci U S A 104:16402–16409. https://doi.org/10.1073/pnas.0708013104

Figures

Figure 1

General workflow for the HetMap pipeline. Short sequence reads are aligned to the rice reference genome to generate the sorted bam alignment files.

Figure 2

The venn diagram for the four polymorphic datasets.

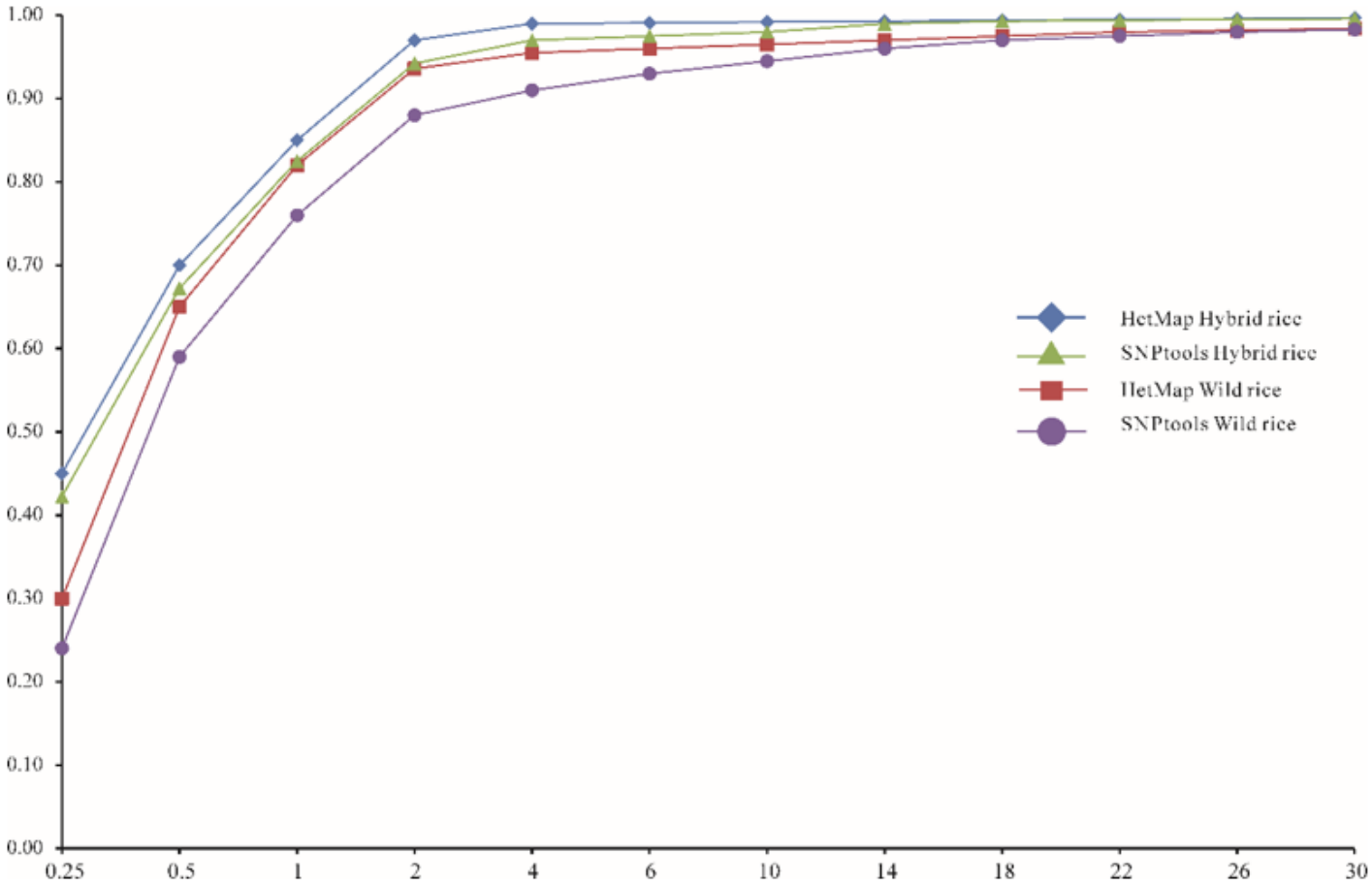

Figure 3

Simulation of different sequence coverage in different populations to check the genotype inference accuracy.

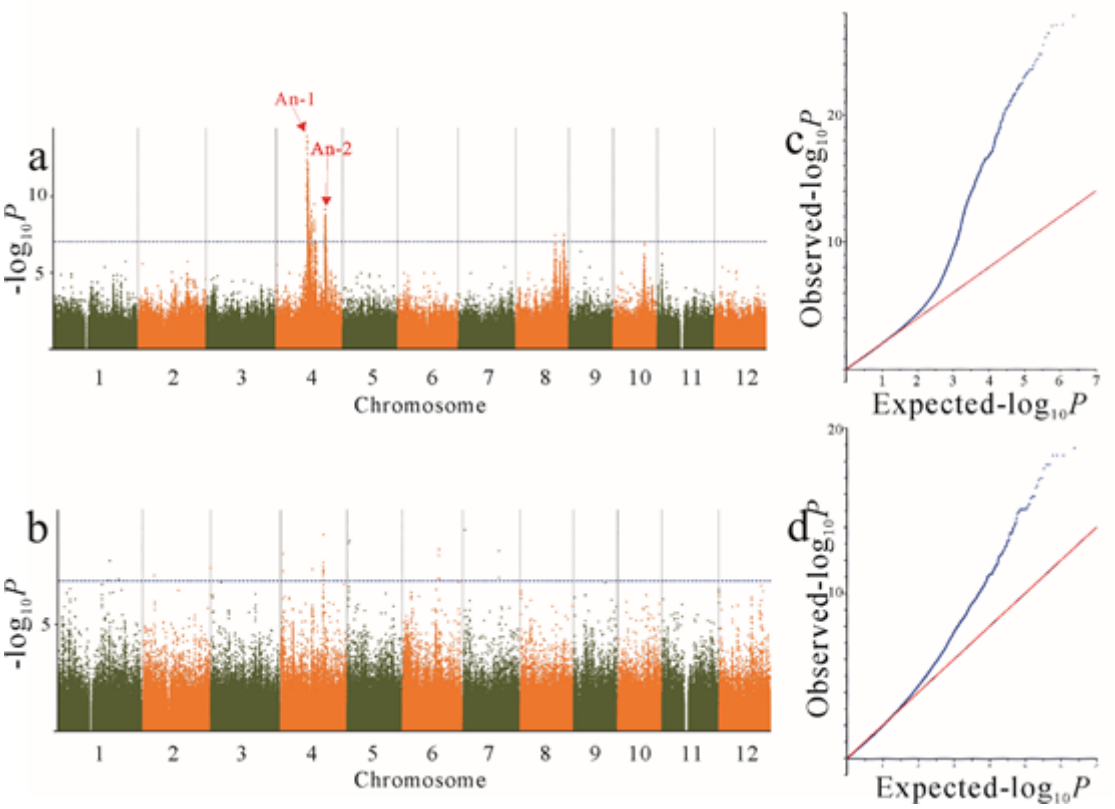

Figure 4

Comparation of the GWAS results for the awn phenotype using different kinds of genotype.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureandtable.pdf