

Physical mapping and InDel marker development for the restorer gene *Rf2* in cytoplasmic male sterile CMS-D8 cotton

Juanjuan Feng

Institute of Cotton Research of Chinese Academy of Agriculture Science <https://orcid.org/0000-0001-9329-8990>

Xuexian Zhang

Institute of Cotton Research of Chinese Academy of Agricultural Sciences

Meng Zhang

Institute of Cotton Research of Chinese Academy of Agricultural Sciences

Liping Guo

Institution of Cotton Research of Chinese Academy of Agricultural Sciences

Tingxiang Qi

Institute of Cotton Research of Chinese Academy of Agricultural Sciences

Huini Tang

institute of Cotton Research of Chinese Academy of Agricultural Sciences

Haiyong Zhu

Western Agriculture Research Center of Chinese Academy of agricultural Sciences

Hailin Wang

Institute of Cotton Research of Chinese Academy of Agricultural Sciences

Xiuqin Qiao

Institute of Cotton Research of Chinese Academy of Agricultural Sciences

Chaozhu Xing

Institute of Cotton Research of Chinese Academy of Agricultural Sciences

Jianyong Wu (✉ dr.wujianyong@live.cn)

Institute of Cotton Research of Chinese Academy of Agricultural Sciences <https://orcid.org/0000-0003-1392-4790>

Research article

Keywords: Cotton, CMS, *Rf2*, BSA, SNP, InDel

Posted Date: August 21st, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-57917/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at BMC Genomics on January 6th, 2021. See the published version at <https://doi.org/10.1186/s12864-020-07342-y>.

Abstract

Background

Cytoplasmic male sterile (CMS) with cytoplasm from *Gossypium Trilobum* (D8) fails to produce functional pollen. It is useful for commercial hybrid cotton seed production. The restore line of CMS-D8 containing *Rf₂* gene can restore the fertility of the corresponding sterile line. This study combined the whole genome resequencing bulked segregant analysis (BSA) with high-throughput SNP genotyping to accelerate the physical mapping of *Rf₂* locus in CMS-D8 cotton.

Methods

The fertility of backcross population ((sterile line × restorer line) × maintainer line) comprising of 1623 individuals was investigated in the field. The fertile pool (100 plants with fertile phenotypes, F-pool) and the sterile pool (100 plants with sterile phenotypes, S-pool) were constructed for BSA resequencing. Selection of 24 SNPs through high-throughput genotyping and development InDel markers narrow down the candidate interval. The PPR genes and restore line up-regulated genes in the candidate interval were analyzed by RT-PCR.

Results

The fertility investigation results showed that fertile and sterile separation ratio was consistent with 1:1. BSA resequencing technology, high-throughput SNP genotyping, and InDel markers were used to identify *Rf₂* locus on candidate interval of 1.48 Mb on Chromosome D05. Furthermore, it was quantified in this experiment that InDel markers co-segregated with *Rf₂* enhanced the selection of the restorer line. The qRT-PCR analysis revealed PPR family gene *Gh_D05G3391* located in candidate interval had significantly lower expression than sterile and maintainer lines. In addition, utilization of previous anther RNA-Seq data of CMS-D8 identified that the expression level of *Gh_D05G3374* encoding NB-ARC domain-containing disease resistance protein in restorer lines was significantly higher than that in sterile and maintainer lines.

Conclusions

This study not only enabled us to precise locate the restore gene *Rf₂* but also evaluated the utilization of InDel markers for marker assisted selection in CMS-D8 *Rf₂* cotton breeding line. The results of this study provide an important foundation for further studies on the mapping and cloning of restorer genes.

Background

Cytoplasmic male sterility (CMS) system plays an important role in utilization of crop heterosis. CMS is a maternally inherited trait, includes degenerate anthers, aborted pollen with carpelloid and petaloid stamens [1]. Current research determined that CMS phenotype is caused by mutations in the mitochondrial genome linked genes and reserved by fertility restorer genes in the nuclear genome [2–4]. CMS system avoids removal of anthers, thereby through hybrid technology enabling the generation of dramatically superior F_1 progenies. These offspring's display significant advantages over their parents and existing popular cultivars in terms of yield, stress tolerance, adaptability, etc. [5]. The CMS phenomenon exists in more than 150 plants and is also used for hybrid breeding of crops, such as maize [6, 7], rice [8, 9], pepper [10] and sorghum [11].

Cotton (*Gossypium hirsutum* L.) is vital source of fiber, oil, and most important economic crop for textile industry in the world. The challenge of low yield in cotton can be mitigated with CMS hybrid breeding system. In cotton up till now, *Harknessii* (D_2-2) cytoplasmic male sterile (CMS-D2) lines [12, 13], *Trilobum* (D8) cytoplasmic male sterile (CMS-D8) lines [14], and upland cotton cytoplasmic male sterile (104-7A, Xiangyuan A, Jin A) have been established and utilized [15]. Theoretically, sources of cotton CMS lines are different therefore restorer genes are inconsistent, and so are the CMS recovery mechanisms. The restorer gene Rf_1 of CMS-D2 could restore the fertility of CMS-D2 and CMS-D8 sterile lines, while fertility of CMS-D8 sterile lines could only be restore with Rf_2 [16]. Furthermore, the Rf_1 gene functions in sporophytes, whereas the Rf_2 gene has a gametophytic restoration system. Previous studies revealed Rf_1 gene loci and Rf_2 gene loci are not allelic, but these genes are tightly linked at a genetic distance of 0.93 cM on chromosome D05. The mapping and identification of the molecular marker linked with Rf_1 restorer gene in cotton has already been progressed. For example, Liu et al. [17] identified 2 RAPD and 3 SSR markers closed linked to Rf_1 . However, Feng et al. [18] developed 4 STS markers associated to Rf_1 . The study of Yin et al. [19] not only identified 5 new SSRs and 2 new STS markers for Rf_1 , but also made high-resolution genetic and physical maps of 15 markers at a genetic distance of 0.9 cM. More to this, Wu et al. [20, 21] recognized new BNL3535 SSR markers and developed 4 InDel markers by whole-genome resequencing. Zhao et al. [22] used super-BSA and successfully mapped Rf_1 to 1.35 Mb region of chromosome D05.

With the increase in crop functional genome research, Rf genes have been successfully cloned in maize (Rf_2) [23], petunia ($Rf-PPR592$) [24], radish (Rfo) [25, 26], rice ($Rf1a$, $Rf1b$, $Rf2$) [27–30], sorghum ($Rf1$) [31], and sugar beet ($Rf1$) [32]. Most of these genes encode PPR proteins, but Rf_2 in maize CMS-T, $Rf17$ in CMS-CW and $Rf2$ in CMS-LD encode aldehyde dehydrogenase, 178-amino-acid mitochondrial sorting protein and mitochondrial glycine-rich protein, respectively [23, 33, 34]. At present, the major bottleneck of cotton CMS breeding system is narrow source of restorer genes and lack of excellent restorer lines compatible with a given sterile line. Unfortunately, no restorer gene has been cloned in cotton. Therefore, fine mapping and isolation of new restorer genes Rf_2 in upland cotton are highly needed for efficient breeding. Interestingly, bulked segregant analysis (BSA) make possible to quickly locate molecular markers closely linked to the target gene by analysing the differences between SNPs and InDels in

segregating population pools [35]. This method has already been used in gene mapping of *Arabidopsis thaliana* [36], rice [37–39] maize [40] and tomato [41]. Single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) are the most abundant type of DNA sequence polymorphisms, found within the genomic sequence of each species [42, 43], and used in QTL analysis. These markers have widely been used in cultivar identification, construction of genetic maps, genetic diversity, map-based cloning, the detection of genotype/phenotype associations, and marker-assisted breeding [44–46]. In recent years, the release of the upland cotton genomic sequence [47–49] and the rapid development of sequencing technology have enhanced detection and application of SNPs and InDels. Furthermore, the application of high-throughput genotyping methods makes SNPs highly attractive genetic markers [50, 51].

The objectives of this study were to physical map restorer gene *Rf₂* and to develop InDel markers co-separated with *Rf₂*. A 1.88 Mb candidate interval was obtained by combining BSA with high-throughput SNP genotyping using a separate BC₁F₁ population. Based on the InDel variation in the 1.88 Mb interval, the InDel markers were developed and used to narrow down a 1.48 Mb candidate interval. The PPR family genes and the genes selected by transcriptome data in candidate region were analysed by qRT-PCR. The InDel markers co-separated with *Rf₂* will be useful to trace *Rf₂* breeding restorer lines in cotton.

Results

Anther observation and BC₁F₁ fertility analysis

The anthers of fertile plants had a large amount of pollen, while the sterile plants had no pollen, and their anthers did not crack. Overall, a total of 1623 BC₁F₁ plants were classified as 850 fertile and 773 sterile plants, and the ratio of the number of fertile plants (850) to the number of sterile plants (773) fit a 1:1 segregation ($\chi^2 = 3.6531 < \chi^2_{(0.05,1)} = 3.84$), confirming that fertility restoration is conditioned by one dominant restorer gene, *Rf₂*. This result is consistent with the results of Zhang et al. [52].

Whole genome resequencing data analysis and evaluation

The two parent lines, F-pool, and S-pool of BC₁F₁ segregation population were sequenced. The Illumina platform was selected to construct the paired-end (PE) library, and the PE fragment was between 300 and 500 bp; 1,251,289,091 reads were obtained (Table 1). The reads from samples were aligned to the reference genome using BWA software, with >82.57% normal efficiency, For the sequencing results, the average Q30 was 94.95%, and the average GC content was 37.22%. A total of 177,874,504 reads were obtained for the R restorer line, with a Q30 value of 94.69%, and average GC content of 37.77%. On the other hand, 174,907,610 reads were obtained for the B maintainer line, with a Q30 value of 94.29%, and average GC content of 36.69%. Finally, 465,660,282 and 432,846,695 reads were obtained for the filial BC₁F₁ generation (fertile and sterile) with a Q30 of 94.83% and 95.97%, and average GC content of 36.93% and 37.22%, respectively (Table 1).

Table 1
Results of high-throughput resequencing data mining

Sample	Reads	Bases	GC(%)	Q20(%)	Q30(%)
B	174,907,610	52,032,498,409	36.69	98.43	94.29
R	177,874,504	52,866,884,655	37.77	98.57	94.69
fertility-bulk	465,660,282	138,509,668,729	37.50	98.53	94.83
sterility-bulk	432,846,695	128,892,417,044	36.93	98.90	95.97
Mean	312,822,273	93,075,367,209	37.22	98.61	94.95
Sum	1,251,289,091	372,301,468,837	-	-	-

BSA combining SNP-index and G' values

The average sequencing depth of the parent lines and the offspring pools was 30.92 \times . Of these, the R restorer line has a sequencing depth of 16.62 \times . The B maintainer line sequencing depth was 16.03 \times , whereas the sequencing depth of the filial BC₁F₁ generation was 47.77 \times + 43.26 \times (Table 2).

Table 2
Sequencing coverage and depth data

Sample	Coverage(%)	Mean Depth
B	82.94	16.03
R	79.95	16.62
fertility-bulk	83.44	47.77
sterility-bulk	83.93	43.26
Mean	82.57	30.92

These reads were mapped onto the reference genome of *Gossypium hirsutum* (Tm-1, <http://mascotton.njau.edu.cn/info/1054/1118.htm>). The average mapping rate was 82.57% for these four samples. A total of 798,286 SNPs was obtained from the two mixed pools, and 72,108 small InDels were obtained from the mixed pools. We used two different methods to map the *Rf*₂ locus responsible for restoring fertility. As shown in Fig. 1 and Fig. 2, only one locus was identified, and both the SNP-index and G' value association algorithms mapped this locus to chromosome D05. More specifically, this locus was located in the region of 25.61 Mb-59.94 Mb (34.33 Mb) using the SNP-index and G' value method.

Fine mapping of the *Rf*₂ gene

It was difficult to determine the candidate gene of *Rf*₂, since the candidate range of 34.33 Mb contains a large amount of genetic information. Thus, it was necessary to fine map *Rf*₂. We developed 24 SNP

markers, and 23 valid SNP markers in this region were used for genotyping an additional 1423 individuals by high-throughput SNP genotyping. We found 6 recombinant plants in the BC₁F₁ population. The position of the *Rf*₂ locus was narrowed down and was located between SNP563981 and SNP597385, a 1.88 Mb region (Supplementary Table S2, the information of the SNP site). Next, we developed InDel markers on the correlated region, and InDel marker analyses revealed that 16 InDel markers were polymorphic. These InDel markers narrowed down the candidate interval to 1.48 Mb for existing recombinants at InDel marker sites. Finally, 10 InDel markers were co-segregated with *Rf*₂ (Fig. 3, Fig. 4, Supplementary Table S3, InDel primers).

Marker-assisted breeding of restorer lines and CMS-D8 hybrid identification

Subsequently, 500 plants were randomly selected from the BC₄F₁ population of CMS-D8, and the InDel 1327 marker was used for genotype analysis. The BC₄F₁ population was typed by visual fertility investigations. The PCR products were analysed by agarose gel electrophoresis, and the results of agarose gel electrophoresis showed two different banding patterns. A single small PCR product was considered homozygous and lacked the restorer gene allele [*S* (*rf*₂*rf*₂)], indicating sterile plants, whereas two fragments were considered heterozygous at the restorer gene locus [*S*(*Rf*₂*rf*₂)], indicating fertile plants. Furthermore, the segregation ratio followed a 1 (*Rf*₂*rf*₂):1 (*rf*₂*rf*₂) (254 *Rf*₂*rf*₂: 246 *rf*₂*rf*₂, $\chi^2_{0.05} = 0.1667 < 3.841$), and the results were consistent with those of the fertility survey.

The plants were scanned with an *atpA* SCAR marker [2] and InDel 1327 markers, as the hybrids of the CMS-D8 system have sterile cytoplasm and *Rf*₂ heterozygous sites. The InDel 1327 primer amplification product produced two bands, and the *atpA* SCAR marker amplification product produced a band with a size of 611 bp (Fig. 5b). Therefore, the CMS-D8 hybrids with heterozygous restorer gene sites and sterile cytoplasm were differentiated by the genotyping of restorer genes and the identification of cytoplasm type.

Candidate gene selection and expression pattern analysis

To determine candidate genes, we adopted a method that combined the 79 genes in the interval with the functional annotation of *Arabidopsis* orthologs and transcriptome data [53]. The 79 genes were subjected to Gene Ontology (GO) analysis. The GO analysis indicated that most of the genes are involved in binding (Fig. 6). According to the successfully isolated restorer genes of other crops belonging to the PPR family, we used qRT-PCR to analyse PPR family genes in candidate interval. Interestingly, the candidate region of the *Rf*₂ locus was found to contain 8 PPR genes (*Gh_D05G3356*, *Gh_D05G3357*, *Gh_D05G3359*, *Gh_D05G3378*, *Gh_D05G3389*, *Gh_D05G3391*, *Gh_D05G3392*, *Gh_D05G3380*) in the region of 1.48 Mb. The relative expression levels of the eight PPR family genes in the restorer line were not significantly higher than those of the maintainer and the sterile lines. However, the relative expression of the

Gh_D05G3391 gene in the restorer line was significantly lower than that in the sterile and the maintainer lines (Fig. 7).

Furthermore, the *Gh_D05G3374*, *Gh_D05G3407* and *Gh_D05G3417* genes were chosen based on the RNA sequence data (Supplementary Table S4), Since, the expression level in the restorer line was significantly higher than that in the sterile and maintainer lines. The qRT-PCR results showed that the *Gh_D05G3417* gene in the restorer line was significantly higher than sterile and maintainer lines (Fig. 7). Finally, two genes (*Gh_D05G3391* and *Gh_D05G3374*) were selected as possible candidate genes.

Discussion

CMS is a common phenomenon that occurs in flowering plants due to the result of interactions between the mitochondrial genome and the nuclear genome [54]. CMS systems have been proven to be a proficient tool in hybrid seed production. Considering the importance of the CMS and restoration systems, numerous molecular mapping studies have been performed on restorer genes in crops, and *Rf* genes have already been isolated in other crops [23–32]. With CMS systems in cotton, fertility can be restored by restoring the genes *Rf*₁ or *Rf*₂. However, these two genes have not yet been identified and cloned. With the availability of upland cotton whole genome sequencing [47–49] and cotton mitochondrial genome sequencing [55], breakthroughs in the study of cotton CMS and restoration of fertility mechanisms can be realized in recent years.

Molecular marker discovery and fine mapping of the fertility restoring gene of CMS cotton

Some researchers have recently been studied cotton CMS systems for molecular marker development and fine mapping of fertility restoration genes. For instance, Yin et al. [19] constructed a BAC library of CMS-D2 restorer lines and reported that *Rf*₁ was located 100 kb between two BAC clone overlapping regions. Yang et al. [56] identified 6 EST-SSR markers (NAU2650, NAU2924, NAU3205, NAU3652, NAU3938, and NAU4040) with a genetic distance of 0.327 cM linked to *Rf*₁ of CMS-D2. Wu et al. [20] screened 13 molecular markers closely linked to *Rf*₁ and located *Rf*₁ between the SSR markers BNL3535 and NAU3652, with a genetic distance of 0.049 cM and 0.078, respectively. Recently, they have reported co-segregated InDel markers such as InDel-1891, InDel-3434, InDel-7525, InDel-9356 and InDel-R [21, 57]. Previous studies have shown that *Rf*₁ and *Rf*₂ are tightly linked at a genetic distance of 0.93 cM on chromosome D05 [16]. The findings of Wang et al. [58] revealed CIR179-250 was strictly linked with both *Rf*₁ and *Rf*₂, which was located on chromosome D05. The present study on the *Rf*₂ gene identified the location of the chromosome D05 base sequence as 54.3-55.78 Mb. Furthermore, the present study developed 10 InDel markers in the correlated region. These markers laid the foundation for locating and fine mapping of *Rf*₂ in CMS-D8 cotton.

Mapping *Rf*₂ using an efficient strategy

Traditional map-based cloning is an efficient approach to isolate genes/QTLs responsible for desired agronomic traits [59–61]. Usually, a genetic map of F_2 , double haploid (DH) or recombinant inbred line (RIL) population based on hundreds of SSR or InDel markers is used to make a primary map. Then a near-isogenic line (NIL) is developed based on marker-assisted selection (MAS; newly discovered in the primary mapping region) to narrow down the region of interest to a sufficient size to screen for a few candidate genes. Unfortunately, this workflow requires relatively more labour and long-time [62, 63]. Compared with genetic mapping, The NGS is a faster and reliable method for mapping [64]. Nevertheless, one mixed pool typically contains approximately 20–100 individuals generally maps the target region at a Mb-level interval [65–69]. Because of insufficient meiotic recombination events, we still have to perform fine mapping or use omics methods such as RNA-seq to further screen the candidate genes [70, 71].

High-throughput SNP genotyping is one of the dimorphic methods in which genotypes are confirmed by direct sequencing [72]. It has been successfully used to genotype the interesting traits in plants [73, 74]. In this regard, Yang et al. [75] developed 1,536 SNP markers to measure genetic diversity by a high-throughput SNP genotyping method.

In this study, SNP-index analysis was used to first position the Rf_2 gene within a 34.33 Mb region. Later on, twenty-three SNP sites selected in this region helped to narrow the Rf_2 gene to a 1.88 Mb region. We developed InDel markers based on InDel variations and used these markers to locate the Rf_2 gene in a 1.48 Mb region. We thus put forward an approach that could rapidly fine map gene loci using only a large BC_1F_1 segregation population, especially for those traits governed by single nuclear-encoded genes. This can be achieved by developing a large segregation population, mapping by sequencing analysis, and high-throughput SNP genotyping in a short time. More to this, rapid and accurate identification of phenotype can be performed with progeny test for desired objectives. Our study results suggested that BSA-seq combined with SNP genotyping can accelerate the mapping of loci controlling quality traits.

Utilization of InDel markers for MAS

Various types of molecular markers closely linked to cotton restoration genes have been developed, but these markers are difficult to use for molecular marker-assisted breeding because of complex experimental processes or because the markers exhibit low sensitivity [21]. There is clear evidence that more genotyping is needed for more accurate breeding. In this study, InDel markers co-segregated with restorer genes tracked Rf_2 for molecular marker-assisted breeding. InDel markers developed on the region showed a higher identification rate of the Rf_2 phenotype than previously developed markers, when applied to breeding of new restorer lines.

Characteristics of the potential candidate gene Rf_2

Currently, Rf genes have been successfully isolated from different crop species [76]. Most of these restorer genes belong to the pentapeptide repeat (PPR) gene family. PPR-type fertility restorer genes have been cloned for petunia [24], Ogura and Kosena cytoplasm in *Raphanus sativus* [25, 26, 77], Boroll CMS in *Oryza sativa* [78], A1 cytoplasm in *Sorghum bicolor* [31], Honglian CMS in rice [27, 28], and for nap

CMS in *Brassica napus* [79]. In this study, we explored the expression patterns of 8 PPR genes of the CMS-D8 system in the candidate interval, and the expression level of most of genes in restorer line was not significantly different from that of the sterile and the maintainer lines. Interestingly, *D05G3391*, a PPR family gene had significant lower expression in restorer line than that of the sterile and maintainer lines. However, non-PPR restorer genes also exist in other crops such as *Rf2* in the maize CMS-T system, *Rf17* in the rice CMS-CW system, and *Rf2* in the rice CMS-LD system encode aldehyde dehydrogenase, 178-amino-acid mitochondrial sorting protein and mitochondrial glycine-rich protein, respectively [23, 33, 34, 80]. Likewise, previous transcriptome data were used to select upregulated genes of restorer line and then analysed by qRT-PCR. The relative expression of *D05G3374*, a NB-ARC disease-resistant protein gene was significantly higher in restorer than sterile and maintainers, which. Based on the results of qRT-PCR, the candidate genes were not determined with desired results in this study. The reason may be that *Rf2* is derived from the nuclear gene of *G. trilobum* [14] and that was might be not available in the reference genome of *G. hirsutum*. Therefore, cloning of fertility restorer genes in cotton CMS systems still need further investigation.

Conclusions

In our study, the BC₁F₁ population was chosen as a genetic population to map *Rf₂* of CMS-D8. Integration of BSA, high-throughput SNP genotyping, and InDel markers identified 1.48 Mb candidate interval on chromosome D05. The InDel markers co-segregated with *Rf₂* can be used to trace *Rf₂* for molecular marker-assisted breeding of restorer lines or hybrids. The qRT-PCR analysis identified *Gh_D05G3391* and *Gh_D05G3374* genes as a putative candidate in this interval. The InDel markers co-segregated with *Rf₂* can not only be used to trace *Rf₂* for molecular marker-assisted CMS breeding but also cornerstone in fine mapping and cloning of restorer genes in cotton.

Methods

Materials and sample collection

The CMS-D8 system, a sterile line (A), maintainer line (B) and restorer line (R) were provided by the Institute of Cotton Research (ICR), Chinese Academy of Agricultural Science, Anyang, Henan, China. A BC₁F₁ ((A line × R line) × B line) segregation population was constructed, and all materials were grown at the Cotton Research Farm at the ICR. Fresh leaves were obtained from the parent lines and BC₁F₁ population. Anthers from buds with 1–2 mm, 3 mm, and 4 mm in length were collected and combined from 100 plants. All harvested samples were snap-frozen in liquid nitrogen and stored at –80 °C before use.

Fertility segregation analysis

During anthesis, visual fertility surveys were conducted for each individual of the BC₁F₁ population of CMS-D8 under field trial conditions, three times per plant. The presence of pollen in a plant indicated

fertility and was determined by squeezing the anthers between the fingers because the male sterility of CMS-D8 occurs during meiosis, the *S(rf)* gametes are sterile, and *S(Rf₂Rf₂)* produces no pollen. The numbers of fertile and sterile BC₁F₁ plants were recorded. The segregation ratio in BC₁F₁ populations was analysed, and the genetic model was inferred and tested using a Chi-square test with SAS software.

Bulked segregant analysis

DNA extraction, library construction and Illumina sequencing

DNA was extracted by the CTAB method [81], and the quality of DNA was assessed by 1.2% agarose gel electrophoresis. The purity of DNA was examined using an Agilent Technologies 2100 Bioanalyzer. The DNA concentration was estimated using a Qubit® DNA Assay Kit in a Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). Equal amounts of DNA (1.5 µg/sample) from 100 BC₁F₁ plants with fertile phenotypes were mixed to form the fertile sample (F-pool), and those from another 100 plants with sterile phenotypes were mixed to form the sterile sample (S-pool). Sequencing libraries were generated using the VAHTS™ Universal DNA Library Prep Kit for Illumina® V3 (Vazyme Biotech) according to the manufacturer's recommendations. Briefly, the DNA samples were fragmented by sonication to a size of 300–500 bp. Then, the DNA fragments were end-polished, A-tailed, and ligated with the full-length adapter for Illumina sequencing by PCR amplification. Consequently, the PCR products were purified (VAHTS™ DNA Clean Beads (Vazyme #N411)), and libraries were analysed for size distribution by an Agilent 2100 Bioanalyzer and quantified by real-time PCR. The libraries constructed above were sequenced by the Illumina HiSeq platform, and 150 bp paired-end reads were generated with an insert size of approximately 350 bp.

Data analysis, data filtering, and alignment

The Fast x-toolkit (v 0.0.14–1) was used to filter out the low-quality reads such as reads with ≥ 10% unidentified nucleotides (N), reads with > 50% bases having phred quality < 5, reads with > 10 nt aligned to the adapter, allowing ≤ 10% mismatches, and putative PCR duplicates generated by PCR amplification in the library construction process (read 1 and read 2 of two paired-end reads were completely identical). The released genome of *Gossypium hirsutum* was downloaded from the Cotton Research Institute (CRI) of Nanjing Agricultural University of China (<http://mascotton.njau.edu.cn/Data.htm>, v1.1) and used as a reference genome [48]. For mapping to the reference genome, BWA (Burrows-Wheeler Aligner) [82] was used to align the clean reads of each sample against the reference genome (settings: mem -t 4 -k 32 -M -R). Alignment files were converted to BAM files using SAMtools software [83] (settings: -bS -t). In addition, potential PCR duplications were removed using the SAMtools command “rmdup”. If multiple read pairs had identical external coordinates, only the pair with the highest mapping quality was retained.

SNP/InDel detection and annotation

Variant calling was performed for all samples by using the unified genotyper function in GATK [84] software. SNP was used as the variant filtration parameter in GATK (settings: `-filterExpression "QD < 4.0 || FS > 60.0 || MQ < 40.0"`, `-G_filter "GQ < 20"`, `-cluster WindowSize 4`). InDel was filtered by Variant Filtration parameter (settings: `-filter Expression "QD < 4.0 || FS > 200.0 || Read PosRankSum < -20.0 || Inbreeding Coeff < -0.8"`). ANNOVAR [85], an efficient software tool, was used to annotate the SNPs and InDels based on the GFF3 files for the reference genome.

Determination of a candidate interval using SNP index and G' values

To obtain a highly accurate SNP set, a range of filters were also employed [86]. The homozygous SNPs between two parents were extracted from the VCF files for SNPs. The read depth information for homozygous SNPs in the above offspring pools was obtained to calculate the SNP index [87]. The SNP index method was used for the analysis, and the SNP-index dot matched curve was obtained by regression fitting as described by Abe et al. [88]. The G' values are used for noise reduction while also addressing linkage disequilibrium (LD) between SNPs [89]. To rule out the effects of unreliable markers, we screened markers based on the SNP index using the following conditions: 1) the sequencing depth of both parents was greater than 8; 2) both pools had a sequencing depth greater than 10; 3) the SNP index values of the two pools was not greater than 0.8 or less than 0.2 at the same time; and 4) the SNP index value difference was greater than 0.8. Sliding window methods were used to present the SNP index of the whole genome. Usually, we used a window size of 2 Mb as the default settings, and used QTL-seq software to calculate Δ (SNPs-index) and G', based on the markers that were scanned [90].

Fine mapping of Rf_2 using high-throughput SNP genotyping and InDel makers

For the analysis, DNA was isolated from the two parental lines and the BC_1F_1 population. The informative molecular markers were used for genotyping each plant of the BC_1F_1 population, various recombinants in the target region were identified, and the linkage relationship between markers and the Rf_2 locus was analysed for gene mapping. According to the results of sequencing mutation detection and BSA of sequencing candidate intervals, one SNP was selected for approximately every 1.5 Mb of physical distance. The selected SNPs met the requirement of having a variation index close to 0.5 in the F-pool and 0 in the S-pool. A subset of 24 selected SNPs was used for genotyping by the Illumina HiSeq PE150 sequence. Individual BC_1F_1 population (except 200 plants that formed pools) plants were genotyped using high-throughput SNP genotyping. Based on the SNP locus exchange of individual plants, we narrowed down the range in which the target gene was located.

The InDel markers were developed based on the insertion deletion mutation within the selection interval. Primers were designed using Oligo7 software [91] and synthesized commercially (TSINGKE Biological Technology, Zhengzhou, China). The PCR system consisted of 20 μ L of PCR mixture that contained 1 \times reaction buffer, 2.0 mM $MgCl_2$, 0.2 mM dNTPs, 0.5 mM each primer, 1 U Taq DNA polymerase (Takara, Japan), and 50 ng of DNA template. The PCR amplification conditions were as follows: 35 cycles of

denaturation at 94 °C for 30 s, annealing at 56 °C-58 °C for 30 s, and extension at 72 °C for 30 s. Then, the reaction was held at 4 °C. The PCR products were visualized by agarose gel electrophoresis. Based on the difference between the genotypes as assessed using polymorphic markers, recombinants were identified in the BC₁F₁ population (plants) and used to fine map *Rf*₂.

Marker-assisted breeding of restorer lines

The utility of the InDel markers for marker-assisted selection was determined in a segregating population. First, the restorer line [*N(Rf*₂*Rf*₂*)*] of CMS-D8 was crossed with the recurrent parent [*N(rf*₂*rf*₂*)*], which has excellent agronomic characteristics. Beginning in the BC₁F₁ generation, the co-dominant InDel marker 1327 was used to track the restorer gene in each generation, and the other markers were used for further verification. Only those individuals verified by the markers were chosen as the female parent for successive backcrosses. In the BC₄F₁ population, 120 individuals were randomly selected, and then InDel 1327 was used to perform segregation analysis. The individuals verified by the markers as homozygous at the restorer gene locus were test-crossed with the sterile line [*S(rf*₂*rf*₂*)*] to determine the segregation of the fertility phenotype in the offspring under field conditions. The *atpA* SCAR marker distinguishes the CMS cytoplasm from other types of cytoplasm[2]. Here, the InDel 1327 marker was combined with the *atpA* SCAR marker to identify hybrids in the CMS-D8 system.

Real-time quantitative PCR (qRT-PCR) analysis

The annotated genes in the interval were analysed, and real-time quantitative PCR was performed to identify PPR family genes and differentially expressed genes that were selected based on previous anther transcriptome data of the CMS-D8 system. Total RNA was isolated using the Sigma Spectrum Plant Total RNA Kit (Sigma-Aldrich, USA) according to the manufacturer's protocol. Reverse transcription was conducted using the PrimeScript™ RT Reagent Kit (TaKaRa, Beijing, China). Trans Start Top Green qPCR Super Mix (Trans gen, Beijing, China) was used according to the manufacturer's instructions to conduct qRT-PCR of the genes. The internal control gene used for qRT-PCR was the cotton *His3* gene (i.e., *histone 3*), and relative gene expression levels were calculated using the 2^{-ΔΔCT} method [92]. All primers are listed in Table S1.

Abbreviations

BSA: Bulk segregant analysis; CMS: Cytoplasmic male sterility; MAS: Marker assisted selection; PPR: Pentapeptide repeat; *Rf*: Restorer of fertility; SNP: Single nucleotide polymorphisms;

Declarations

Ethics approval and consent to participate

All the cotton lines used and analyzed were public and available for non-commercial purpose. This article did not contain any studies with human participants or animals performed by any of the authors.

Availability of data and materials

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Tianshan Youth Program (2018Q010). The funding body had no role in the design, analysis, interpretation of data, and in writing the manuscript.

Authors' contributions

JYW and CZX designed the experiments. LPG, TXQ, HNT, HYZ, XQQ, and HLW constructed BC₁F₁ population and did field management. JJF, XXZ, and MZ performed fertility survey, data analysis and qRT-PCR. JJF and JYW contributed to the preparation of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors appreciate the field staff for their anonymous work.

References

1. Chase CD. Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. *Trends Genet.* 2007;23(2):81–90.
2. Wu J, Gong Y, Cui M, Qi T, Guo L, Zhang J, Xing CJE. **Molecular characterization of cytoplasmic male sterility conditioned by *Gossypium harknessii* cytoplasm (CMS-D2) in upland cotton.** 2011, 181(1):17–29.
3. Linke B, Börner TJM. **Mitochondrial effects on flower and pollen development.** 2005, 5(6):389–402.
4. Hanson MR. Plant mitochondrial mutations and male sterility. *Annual review of genetics.* 1991;25:461–86.

5. Bohra A, Jha UC, Adhimoolam P, Bisht D, Singh NP. Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. *Plant Cell Rep.* 2016;35(5):967–93.
6. Laughnan JR, Gabay-Laughnan S. Cytoplasmic male sterility in maize. *Annual review of genetics.* 1983;17:27–48.
7. Jaqueth JS, Hou Z, Zheng P, Ren R, Nagel BA, Cutter G, Niu X, Vollbrecht E, Greene TW, Kumpatla SP. **Fertility Restoration of Maize CMS-C Altered by a Single Amino Acid Substitution within the *Rf4* bHLH Transcription Factor.** *The Plant journal: for cell and molecular biology* 2019.
8. Akagi H, Sakamoto M, Shinjyo C, Shimada H, Fujimura TJCG. **A unique sequence located downstream from the rice mitochondrial *atp6* may cause male sterility.** 1994, 25(1):52–58.
9. Chang Z, Chen Z, Wang N, Xie G, Lu J, Yan W, Zhou J, Tang X, Deng XW. Construction of a male sterility system for hybrid rice breeding and seed production using a nuclear male sterility gene. *Proc Natl Acad Sci USA.* 2016;113(49):14145–50.
10. Ma Y, Huang W, Ji JJ, Gong ZH, Yin CC, Ahmed SS, Zhao ZL. **Maintaining and restoring cytoplasmic male sterility systems in pepper (*Capsicum annuum* L.).** *Genet Mol Res.* 2013;12(3):2320–31.
11. Tang HV, Pring DRShaw LC, Salazar RA, Muza FR, Yan B, Schertz KFJPJ. **TRANSCRIPT PROCESSING INTERNAL TO A MITOCHONDRIAL OPEN READING FRAME IS CORRELATED WITH FERTILITY RESTORATION IN MALE-STERILE SORGHUM.** 2010, 10(1):123–133.
12. Meyer VG. **Male Sterility From *Gossypium harknessii*.** *Journal of Heredity* 1975, 62(1).
13. Weaver DB, Weaver JB. Inheritance of Pollen Fertility Restoration in Cytoplasmic Male-Sterile Upland Cotton1. *Crop Sci.* 1977;17(4):497–9.
14. Stewart J, Dugger C, Richter D: **A new cytoplasmic male sterility and restorer for cotton.** In: *Proceedings of Beltwide Cotton Conferences, Nashville, TN: 1992*, 1992: 610.
15. Zhang C: **Preliminary studies and breeding on cytoplasmic male sterility three-line of *Gossypium hirsutum* L.:** HuaZhong Agricultural University; 2005.
16. Zhang JF, Stewart JM. **Inheritance and Genetic Relationships of the D8 and D2-2 Restorer Genes for Cotton Cytoplasmic Male Sterility.** *Crop Science* 2001, 41(2).
17. Liu L, Guo W, Zhu X, Zhang TJT, Genetics A. **Inheritance and fine mapping of fertility restoration for cytoplasmic male sterility in *Gossypium hirsutum* L.** 2003, 106(3):461–469.
18. Feng CD, Stewart JMD, Zhang JFJT, Genetics A. **STS markers linked to the *Rf1* fertility restorer gene of cotton.** 2005, 110(2):237–243.
19. Yin J, Guo W, Yang L, Liu L, Zhang T. **Physical mapping of the *Rf₁* fertility-restoring gene to a 100 kb region in cotton.** *Theor Appl Genet.* 2006;112(7):1318–25.
20. Wu J, Cao X, Guo L, Qi T, Wang H, Tang H, Zhang J, Xing C. **Development of a candidate gene marker for *Rf₁* based on a PPR gene in cytoplasmic male sterile CMS-D2 Upland cotton.** *Mol Breeding.* 2014;34(1):231–40.
21. Wu J, Meng Z, Zhang X, Guo L, Qi T, Wang H, Tang H, Zhang J, Xing C. **Development of InDel markers for the restorer gene *Rf₁* and assessment of their utility for marker-assisted selection in cotton.**

- Euphytica. 2017;213(11):251.
22. Zhao C, Zhao G, Geng Z, Wang Z, Wang K, Liu S, Zhang H, Guo B, Geng J. Physical mapping and candidate gene prediction of fertility restorer gene of cytoplasmic male sterility in cotton. *BMC Genom.* 2018;19(1):6.
 23. Cui X, Wise RP, Schnable PS. **The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize.** *Science.* 1996;272(5266):1334–6.
 24. Bentolila S, Alfonso AA, Hanson MR. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proc Natl Acad Sci USA.* 2002;99(16):10887–92.
 25. Brown GG, Formanova N, Jin H, Wargachuk R, Dendy C, Patil P, Laforest M, Zhang J, Cheung WY, Landry BS. The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. *The Plant journal: for cell molecular biology.* 2003;35(2):262–72.
 26. Desloire S, Gherbi H, Laloui W, Marhadour S, Clouet V, Cattolico L, Falentin C, Giancola S, Renard M, Budar F, et al. **Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family.** *EMBO Rep.* 2003;4(6):588–94.
 27. Hu J, Wang K, Huang W, Liu G, Gao Y, Wang J, Huang Q, Ji Y, Qin X, Wan L, et al. The rice pentatricopeptide repeat protein RF5 restores fertility in Hong-Lian cytoplasmic male-sterile lines via a complex with the glycine-rich protein GRP162. *Plant Cell.* 2012;24(1):109–22.
 28. Huang W, Yu C, Hu J, Wang L, Dan Z, Zhou W, He C, Zeng Y, Yao G, Qi J, et al. Pentatricopeptide-repeat family protein RF6 functions with hexokinase 6 to rescue rice cytoplasmic male sterility. *Proc Natl Acad Sci USA.* 2015;112(48):14984–9.
 29. Kazama T, Toriyama K. **A pentatricopeptide repeat-containing gene that promotes the processing of aberrant *atp6* RNA of cytoplasmic male-sterile rice.** *FEBS Lett.* 2003;544(1–3):99–102.
 30. Komori T, Ohta S, Murai N, Takakura Y, Kuraya Y, Suzuki S, Hiei Y, Imaseki H, Nitta N. **Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa* L.).** *The Plant journal: for cell molecular biology.* 2004;37(3):315–25.
 31. Klein RR, Klein PE, Mullet JE, Minx P, Rooney WL, Schertz KF. **Fertility restorer locus *Rf1* of sorghum (*Sorghum bicolor* L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12.** *Theor Appl Genet.* 2006;112(2):388–8.
 32. Matsuhira H, Kagami H, Kurata M, Kitazaki K, Matsunaga M, Hamaguchi Y, Hagihara E, Ueda M, Harada M, Muramatsu A, et al. **Unusual and typical features of a novel restorer-of-fertility gene of sugar beet (*Beta vulgaris* L.).** *Genetics.* 2012;192(4):1347–58.
 33. Fujii S, Toriyama K. Suppressed expression of Retrograde-Regulated Male Sterility restores pollen fertility in cytoplasmic male sterile rice plants. *Proc Natl Acad Sci USA.* 2009;106(23):9513–8.
 34. Itabashi E, Iwata N, Fujii S, Kazama T, Toriyama K. **The fertility restorer gene, *Rf2*, for Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein.** *The Plant journal: for cell molecular biology.* 2011;65(3):359–67.
 35. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome

- resequencing of DNA from two bulked populations. *The Plant journal: for cell molecular biology*. 2013;74(1):174–83.
36. Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature methods*. 2009;6(8):550–1.
37. Cai Z, Jia P, Zhang J, Gan P, Shao Q, Jin G, Wang L, Jin J, Yang J, Luo J. Genetic analysis and fine mapping of a qualitative trait locus *wpb1* for albino panicle branches in rice. *PLoS One*. 2019;14(9):e0223228.
38. Sun J, Yang L, Wang J, Liu H, Zheng H, Xie D, Zhang M, Feng M, Jia Y, Zhao H *et al.* **Identification of a cold-tolerant locus in rice (*Oryza sativa* L.) using bulked segregant analysis with a next-generation sequencing strategy.** *Rice (New York, NY)* 2018, **11**(1):24.
39. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, Matsumura H, Yoshida K, Mitsuoka C, Tamiru M, et al. Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat Biotechnol*. 2012;30(2):174–8.
40. Klein H, Xiao Y, Conklin PA, Govindarajulu R, Kelly JA, Scanlon MJ, Whipple CJ, Bartlett M: **Bulked-Segregant Analysis Coupled to Whole Genome Sequencing (BSA-Seq) for Rapid Gene Cloning in Maize.** *G3 (Bethesda, Md)* 2018, **8**(11):3583–3592.
41. G L, T Z, X Y, J J, J L, biology XXJBp: **Molecular mapping of the Cf-10 gene by combining SNP/InDel-index and linkage analysis in tomato (*Solanum lycopersicum*).** 2019, **19**(1):15.
42. Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol*. 2002;5(2):94–100.
43. Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet*. 2002;71(4):854–62.
44. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES. Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant journal: for cell molecular biology*. 2005;44(6):1054–64.
45. Simko I, Haynes KG, Ewing EE, Costanzo S, Christ BJ, Jones RW. Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. *Molecular genetics genomics: MGG*. 2004;271(5):522–31.
46. Szalma SJ, Buckler ES, Snook ME, McMullen MD. Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks. *Theor Appl Genet*. 2005;110(7):1324–33.
47. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J. **Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution.** *Nat Biotechnol*. 2015;33(5):524–30.
48. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. **Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement.** *Nat Biotechnol*. 2015;33(5):531–7.

49. Wang M, Tu L, Yuan D, Zhu, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G, et al. **Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*.** Nat Genet. 2019;51(2):224–9.
50. De la Vega FM, Lazaruk KD, Rhodes MD, Wenz MH. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPLex Genotyping System. Mutat Res. 2005;573(1–2):111–35.
51. Shen R, Fan J, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Garcia EW, McBride CJMR. **High-throughput SNP genotyping on universal bead arrays.** 2005, 573(1):70–82.
52. Zhang J, Stewart, JMJs. **CMS-D8 restoration in cotton is conditioned by one dominant gene.** 2001, 41(2):283–288.
53. Ma J, Liu J, Pei W, Ma Q, Science JyJP: **Genome-wide association study of the oil content in upland cotton (*Gossypium hirsutum* L.) and identification of GhPRXR1, a candidate gene for a stable QTLqOC-Dt5-1.** 2019.
54. Touzet P, Meyer EHJM. **Cytoplasmic male sterility and mitochondrial metabolism in plants.** 2014, 19:166–171.
55. Tang M, Chen Z, Grover CE, Wang Y, Li S, Liu G, Ma Z, Wendel JF, Hua JJBG. **Rapid evolutionary divergence of *Gossypium barbadense* and *G. hirsutum* mitochondrial genomes.** 2015, 16(1):770.
56. Yang L. **Map-based cloning of the CMS fertility restorer gene and analysis of the PPR gene family has been conducted in cotton.** PhD dissertation, Nanjing Agricultural University in China. 2009.
57. Bingbing Z, Xuexian Z, Liping G, Tingxiang Q, Hailin W, Huini T, Xiuqin Q, Kashif S, Chaozhu X, Jianyong W. **Genome-wide analysis of Rf-PPR-like(RFL)genes and a new InDel marker development for *Rf₁* gene in cytoplasmic male sterile CMS-D2 Upland cotton..** Journal of Cotton Research. 2018;1(03):12–22.
58. Wang F, Yue B, Hu JG, Stewart JM, Zhang JF. **A target region amplified polymorphism marker for fertility restorer gene *Rf1* and chromosomal localization of *rf1* and *Rf2* in cotton.** Crop Sci. 2009;49(5):1602–8.
59. Zuo W, Chao Q, Zhang N, Ye J, Tan G, Li B, Xing Y, Zhang B, Liu H, Fengler KA, et al. A maize wall-associated kinase confers quantitative resistance to head smut. Nat Genet. 2015;47(2):151–7.
60. Lu X, Xiong Q, Cheng T, Li Q, Liu X, Bi Y, Li W, Zhang W, Ma B, Lai YJMP. **A PP2C-1 Allele Underlying a Quantitative Trait Locus Enhances Soybean 100-Seed Weight.** 2017, 10(5):670–684.
61. Zhang Z, Li J, Pan Y, Li J, Zhou L, Shi H, Zeng Y, Guo H, Yang S, Zheng WJNC. **Natural variation in CTB4a enhances rice adaptation to cold habitats.** 2017, 8(1):14788.
62. Zhu Z, Tan L, Fu Y, Liu F, Cai H, Xie D, Wu F, Wu J, Matsumoto T, Sun CJNC. **Genetic control of inflorescence architecture during rice domestication.** 2013, 4(1):2200–2200.
63. America, QZJPotNAoSotUSo. **PMS1T, producing Phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice.** 2016, 113(52):201619159.

64. Schneeberger KJNRG. **Using next-generation sequencing to isolate mutant genes from forward genetic screens.** 2014, 15(10):662–676.
65. Uauy CJPBJ. **RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat.** 13(5):613–624.
66. Haase NJ, Beissinger T, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, Buell CR, Kaeppler SM, de Leon N: **Shared Genomic Regions Between Derivatives of a Large Segregating Population of Maize Identified Using Bulk Segregant Analysis Sequencing and Traditional Linkage Analysis.** *G3 (Bethesda, Md)* 2015, 5(8):1593–1602.
67. Zheng W, Wang Y, Wang L, Ma Z, Zhao J, Wang P, Zhang L, Liu Z, Lu X. Genetic mapping and molecular marker development for Pi65(t), a novel broad-spectrum resistance gene to rice blast using next-generation sequencing. *Theor Appl Genet.* 2016;129(5):1035–44.
68. Geng X, Jiang C, Yang J, Wang L, Wu X, Wei WJPO. **Rapid Identification of Candidate Genes for Seed Weight Using the SLAF-Seq Method in *Brassica napus*.** 2016, 11(1).
69. Kayam G, Brand Y, Faigenboim-Doron A, Patil A, Hedvat I, Hovav R. Fine-Mapping the Branching Habit Trait in Cultivated Peanut by Combining Bulk Segregant Analysis and High-Throughput Sequencing. *Front Plant Sci.* 2017;8:467.
70. Song J, Li Z, Liu Z, Guo Y, Qiu LJ. Next-Generation Sequencing from Bulk Segregant Analysis Accelerates the Simultaneous Identification of Two Qualitative Genes in Soybean. *Front Plant Sci.* 2017;8:919.
71. Gu A, Meng C, Chen Y, Wei L, Dong H, Lu Y, Wang Y, Chen X, Zhao J, Shen S. Coupling Seq-BSA and RNA-Seq Analyses Reveal the Molecular Pathway and Genes Associated with Heading Type in Chinese Cabbage. *Frontiers in genetics.* 2017;8:176.
72. Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura YJJoHG. **A high-throughput SNP typing system for genome-wide association studies.** 2001, 46(8):471–477.
73. Grattapaglia D, Silva-Junior OB, Kirst M, de Lima BM, Faria DA, Pappas GJ. High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biol.* 2011;11(1):65.
74. Yan J, Yang X, Shah T, Sánchez-Villeda H, Li J, Warburton M, Zhou Y, Crouch JH, Xu Y. High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol Breeding.* 2010;25(3):441–51.
75. Yang X, Yan J, Shah T, Warburton ML, Li Q, Li L, Gao Y, Chai Y, Fu Z, Zhou Y, et al. Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. *Theor Appl Genet.* 2010;121(3):417–31.
76. Chen L, Liu YG. Male sterility and fertility restoration in crops. *Annu Rev Plant Biol.* 2014;65:579–606.
77. Koizuka N, Imai R, Fujimoto H, Hayakawa T, Kimura Y, Kohno-Murase J, Sakai T, Kawasaki S, Imamura J. **Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosenra radish.** *The Plant journal: for cell molecular biology.* 2003;34(4):407–15.

78. Wang Z, Zou Y, Li X, Zhang Q, Chen L, Wu H, Su D, Chen Y, Guo J, Luo D, et al. Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. *Plant Cell*. 2006;18(3):676–87.
79. Liu Z, Dong F, Wang X, Wang T, Su R, Hong D, Yang G. **A pentatricopeptide repeat protein restores nap cytoplasmic male sterility in *Brassica napus***. *J Exp Bot*. 2017;68(15):4115–23.
80. Liu F, Cui X, Horner HT, Weiner H, Schnable PS. Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize. *Plant Cell*. 2001;13(5):1063–78.
81. Zhang JF, Stewart JM. Economical and rapid method for extracting cotton genomic DNA. *J Cotton Sci*. 2000;4(3):193–201.
82. Li H, Durbin R: **Fast and accurate short read alignment with Burrows Wheeler transform**; 2009.
83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan JJB. **The Sequence Alignment-Map format and SAMtools**. 2009, 25(16):2078–2079.
84. Mckenna A, Hanna ME, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, Depristo MAJGR. **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. 2010, 20(9):1297–1303.
85. Kai W, Mingyao L, Hakon HJNAR. **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. 2010, 38(16):e164.
86. Joke R, Peter DR, Hui Z, Anthony L, Dominiek S, John C, Peter VL, Maarten VDB, Kirsten C, Bernard SJNB. **Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing**. 2012, 30(1):61–68.
87. Hiroki T, Akira A, Kentaro Y, Shunichi K, Satoshi N, Chikako M, Aiko U, Hiroe U, Muluneh T, Shohei TJPJ. **QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations**. 2013, 74(1):174–183.
88. Akira A, Shunichi K, Kentaro Y, Satoshi N, Hiroki T, Hiroyuki K, Hideo M, Kakoto Y, Chikako M, Muluneh TJNB. **Genome sequencing reveals agronomically important loci in rice using MutMap**. 2012, 30(2):174–178.
89. PM M, JH W. biology KJJPC: **The statistics of bulk segregant analysis using next generation sequencing**. 2011, 7(11):e1002255.
90. Mansfeld BN, Grumet R. **QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing**. *The plant genome* 2018, 11(2).
91. Rychlik W: **OLIGO 7 primer analysis software**. In: *PCR primer design*. Springer; 2007: 35–59.
92. Schmittgen TD, Livak KJ, Nature P. **Analyzing real-time PCR data by the comparative C(T) method**. 2008, 3(6):1101–1108.

Figures

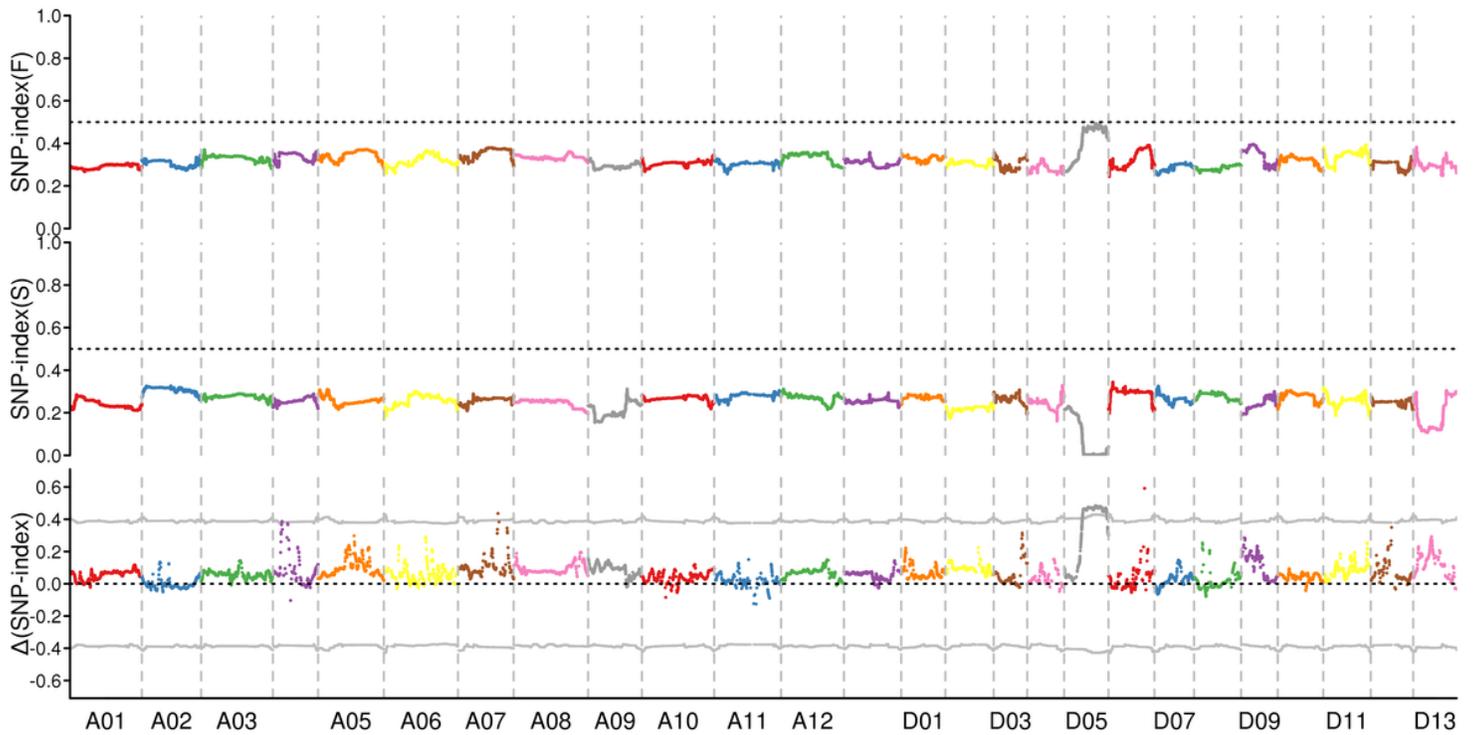


Figure 1

SNP-index algorithm to map the Rf2 gene. The coloured point represents the calculated SNP-index (or Δ SNP-index) value. The top graph illustrates the distribution of the SNP-index values in the F-pool; the middle graph shows the distribution of the SNP-index values in the S-pool; the bottom graph shows the distribution of the Δ SNP-index values, and the grey line represents the theoretical threshold line.

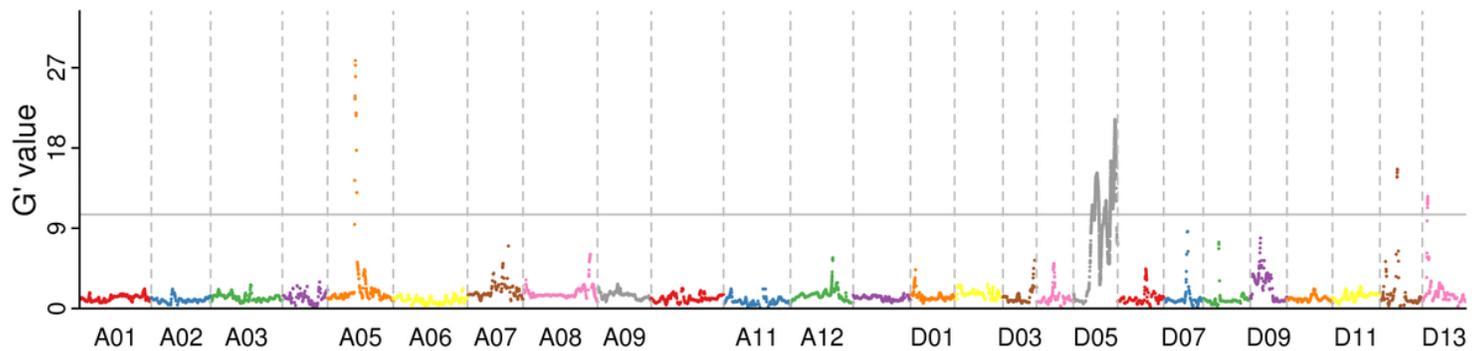


Figure 2

G' algorithm to map the Rf2 gene. The distribution of G' values on the chromosome. The abscissa is the chromosome name. The colour point represents the G' values of each SNP locus. The grey line represents the threshold of significant association. The higher the G' value, better is the correlation effect.

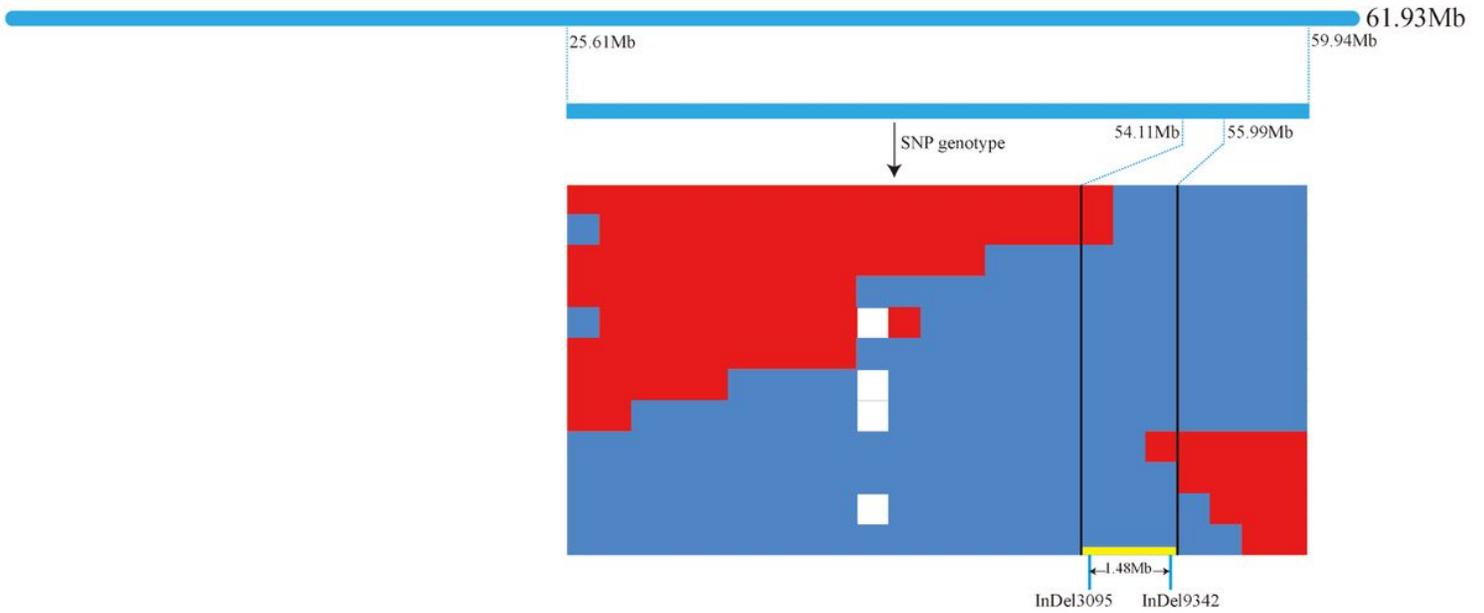


Figure 3

Molecular mapping of the Rf2 gene using the SNP/InDel combinational approach. White indicates a lack of sample, red indicates that the SNP site was exchanged, and blue indicates that the genotype and phenotype were consistent.

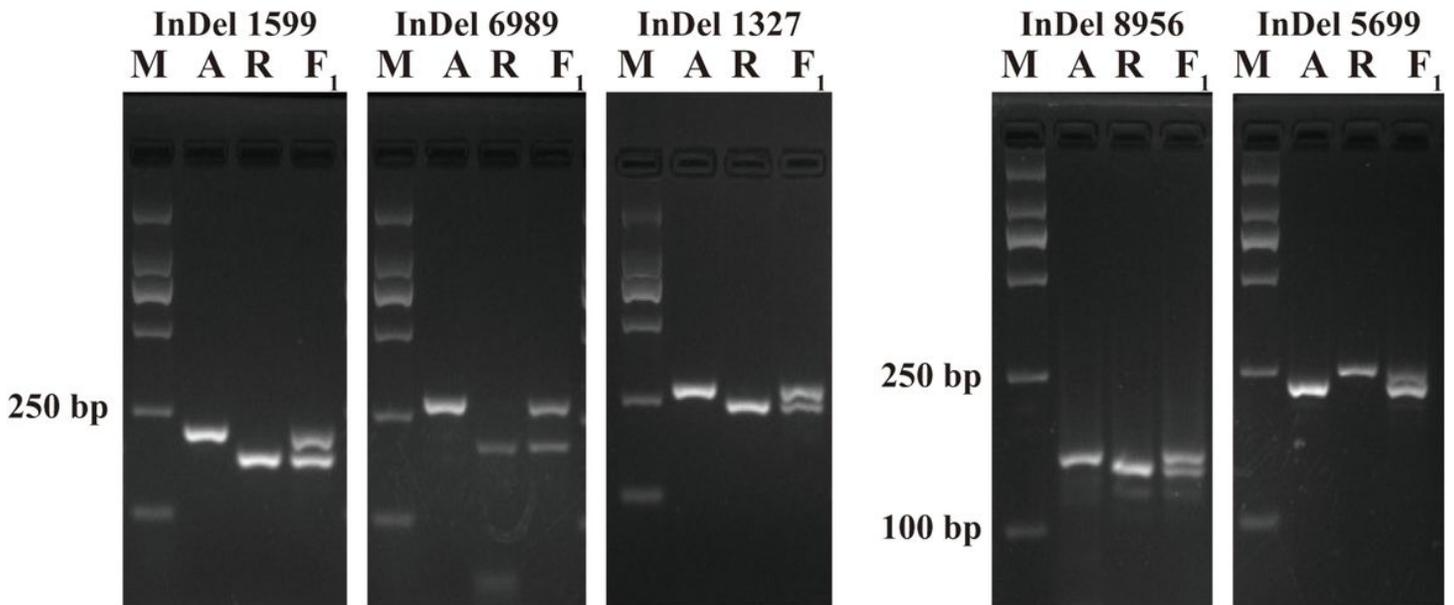


Figure 4

The InDel markers co-separating with Rf2. M: DL2000 DNA marker; A: sterile line; R: restorer line; F1: hybrid of CMS-D8 system.

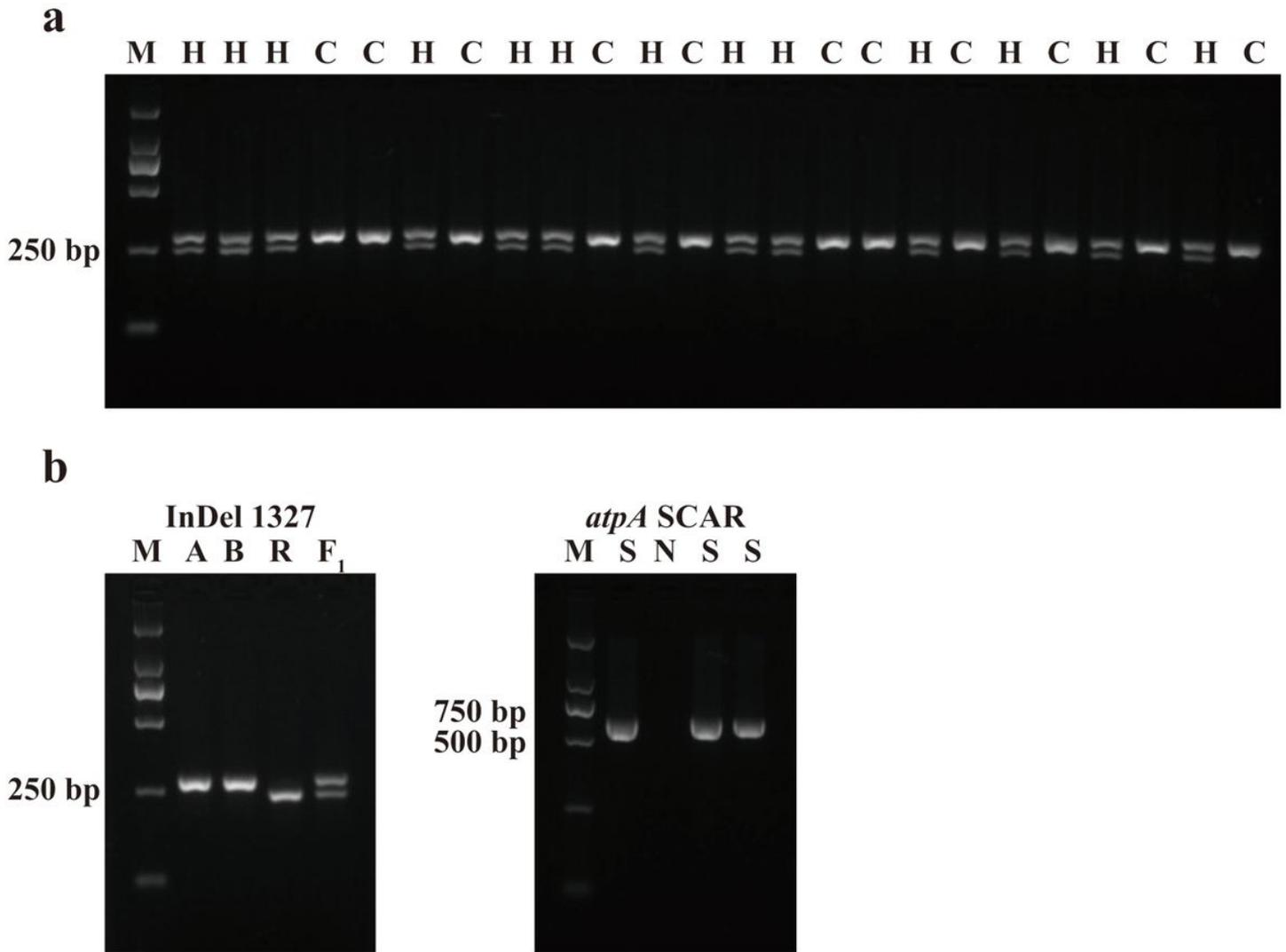


Figure 5

(a) BC4F1 plants were screened with InDel1327, M: DL2000 DNA marker; H: Rf2 heterozygous plants; C: plants lacking the restorer gene Rf2. (b) Molecular identification of the CMS system hybrids and cotton varieties with InDel 1327 and *atpA* SCAR markers. M: DL2000 DNA marker; A: sterile line; B: maintainer line; R: restorer line; F₁: hybrid of CMS-D8 system; S: sterile cytoplasm; N: fertile cytoplasm.

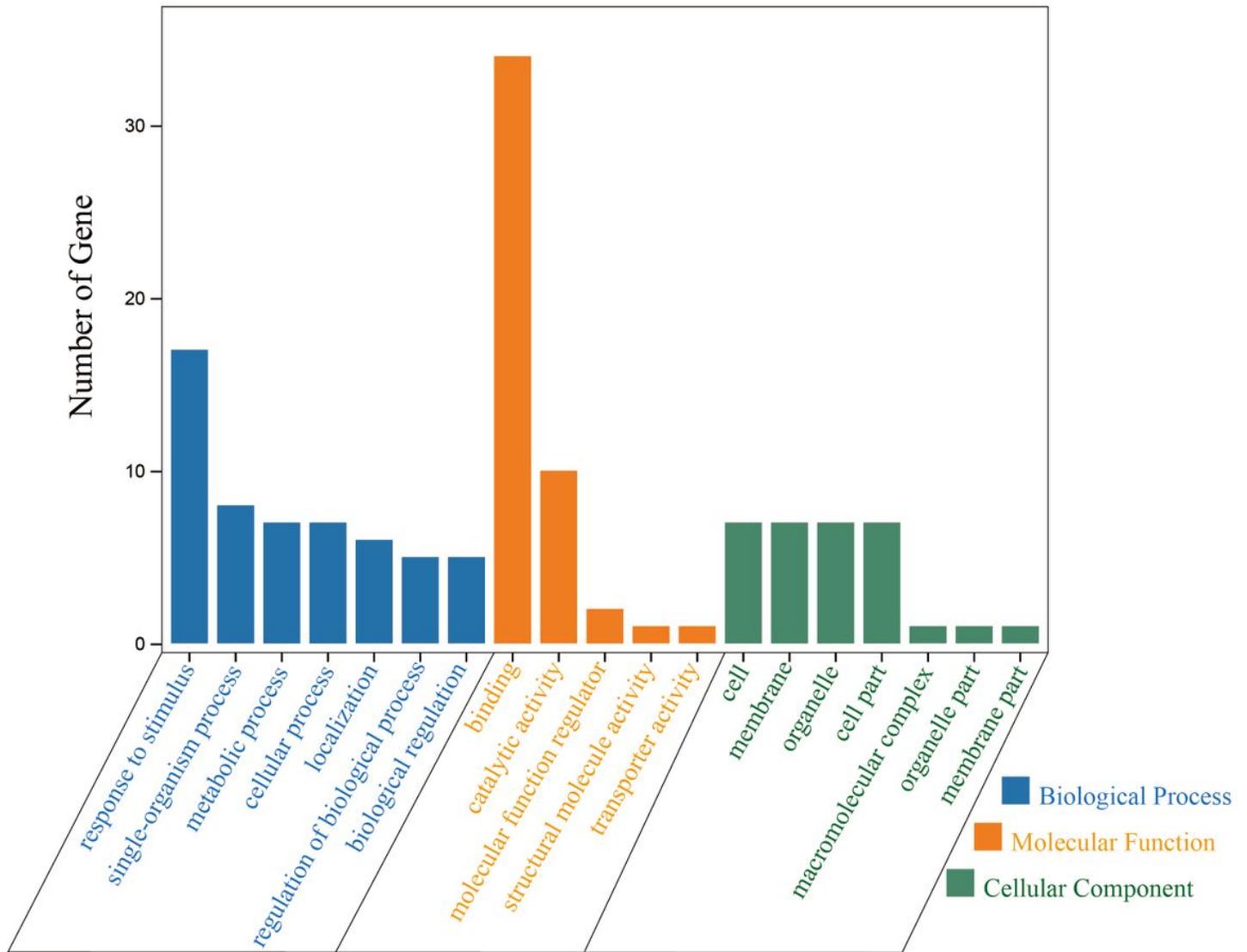


Figure 6

Gene Ontology (GO) analysis of 79 genes in the candidate interval. The ordinate is the number of genes, and the abscissa is the term.

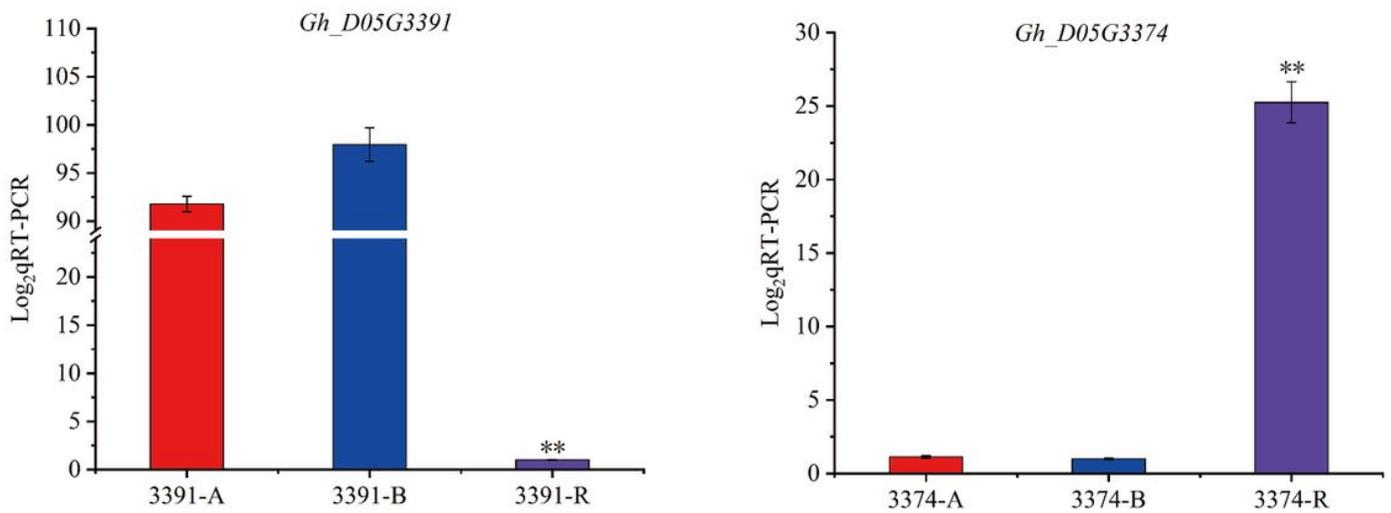


Figure 7

Expression patterns of D05G3391 and D05G3374. (**, $P < 0.01$) The asterisks indicate that the difference in gene expression in the A, B and R lines was highly significant. A: sterile line; B: maintainer line; R: restorer line.

Supplementary Files

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

- [SupplementaryTable.docx](#)