BMC Plant Biology

RESEARCH

Open Access

# The Dark Side of the pollen: BSA-seq identified genomic regions linked to male sterility in globe artichoke

Matteo Martina[1†], Aldana Zayas[2†], Ezio Portis[1], Giovanna Di Nardo[3], Maria Francesca Polli[1], Cinzia Comino[1], Gianfranco Gilardi[3], Eugenia Martin[2*] and Alberto Acquadro[1*]

## Abstract

Globe artichoke (*Cynara cardunculus* var. *scolymus*; 2n = 2x = 34) is a food crop consumed for its immature flower heads. Traditionally, globe artichoke varietal types are vegetatively propagated. However, seed propagation makes it possible to treat the crop as annual, increasing field uniformity and reducing farmers costs, as well as pathogens diffusion. Despite globe artichoke's significant agricultural value and the critical role of heterosis in the development of superior varieties, the production of hybrids remains challenging without a reliable system for large-scale industrial seed production. Male sterility (MS) presents a promising avenue for overcoming these challenges by simplifying the hybridization process and enabling cost-effective seed production. However, within the *Cynara* genus, genic male sterility has been linked to three recessive loci in globe artichoke, with no definitive genetic mechanism elucidated to date. A 250 offsprings $F_2$ population, derived from a cross between a MS globe artichoke and a male fertile (MF) cultivated cardoon (*C. cardunculus* var. *altilis*) and fitting a monogenic segregation model (3:1), was analyzed through BSA-seq, aiming at the identification of genomic regions/genes affecting male sterility. Four QTL regions were identified on chromosomes 4, 12, and 14. By analyzing the sequence around the highest pick on chromosome 14, a cytochrome P450 (*CYP703A2*) was identified, carrying a deleterious substitution (R/Q) fixed in the male sterile parent. A single dCAPS marker was developed around this SNP, allowing the discrimination between MS and MF genotypes within the population, suitable for applications in plant breeding programs. A 3D model of the protein was generated by homology modeling, revealing that the mutated amino acid is part of a highly conserved motif crucial for protein folding.

## Key message

Globe artichoke (*Cynara cardunculus* var. *scolymus* L.) is a mediterranean allogamous, mainly cultivated for human consumption. The establishment of superior varieties through heterosis has been applied in this species, but their production is complex, and the development of male sterile varieties has been of interest among plant

†Matteo Martina and Aldana Zayas contributed equally to this work.

*Correspondence:
Eugenia Martin
martin@iicar-conicet.gob.ar
Alberto Acquadro
alberto.acquadro@unito.it

Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

scientists and breeders as a cost-effective solution for hybrid seeds production. Here, we report the use of the BSA-seq approach for the identification of a SNP strictly associated with male sterility in an $F_2$ mapping population, we developed a single dCAPS marker allowing the discrimination between MS and MF genotypes within the population, and we discussed a potential role for a candidate gene (CYP702A3), involved in pollen grain vitality, present in the candidate region.

**Keywords** Globe artichoke, BSA-seq, Male sterility, CYP702A3, Pollen vitality

## Introduction

Globe artichoke (*Cynara cardunculus* var. *scolymus* L.) is a highly heterozygous, allogamous species native to the Mediterranean region, traditionally cultivated for human consumption of its fresh, corned, or frozen immature heads, characterized by high nutraceutical value [1, 2]. *C. cardunculus* includes two further taxa: (i) var. *sylvestris*, the progenitor of both cultivated forms, namely wild cardoon [3, 4], and (ii) the var. *altilis*, the cultivated cardoon, grown for the production of fleshy stems [5, 6]. Worldwide, with almost 115,897 ha and 1,516,955 t per year, Italy is the main producer country, accounting for about 24% of the world production [7]. Traditionally, this crop is cultivated in the Mediterranean basin (almost 85% of the world production), but in the last decades the production has developed also in other regions, such as Argentina, Peru, China, and USA [7, 8]. Artichoke production is traditionally carried out through vegetative propagation [9], although sexual propagation is possible [10]. However, in recent years a considerable number of new globe artichoke seed-propagated cultivars, open pollinated or hybrids, have been developed and successfully introduced in the market [11].

Historically, the use of heterosis for the establishment of superior varieties has been applied through the development of hybrids, but their production is complex in the absence of a system that allows large-scale crosses for industrial seeds production. For this reason, the mechanism of male sterility has long been of interest among plant scientists and breeders as a cost-effective solution for hybrid seeds production [12]. Among *Cynara* genus, genic male sterility has been associated with three recessive loci in globe artichoke only, but no clear genetic mechanism has been proposed [13–15]. Recently, Zayas et al. [16] developed an $F_2$ population fitting a monogenic segregation model for MS, which was analyzed by combining the sequence related amplified polymorphism (SRAP) technology and the bulk segregant analysis (BSA) approach.

BSA allows the identification of genetic regions associated with traits of interest by selecting contrasting individuals within a segregating population, bulking the identified genotypes, and analyzing their genetic differences [17, 18]. This technique has been extensively applied with traditional molecular markers, both in qualitative traits investigation and quantitative trait loci (QTLs) mapping. In the NGS-era, its power has been dramatically improved by the application of whole genome sequencing (WGS −[19]). This approach is known as bulked segregant analysis by deep sequencing (BSA-seq) and has been shown to be highly reliable for QTL mapping in many species [19]. Its main advantage is that it can quickly associate a specific locus with candidate genomic regions, which considerably reduces workload and time. Thanks to the availability of a reference genome for the globe artichoke, released and implemented starting from an inbred variety harboring only 10% of heterozygosity [20–22], parental genotypes can then be investigated, and a large number of single nucleotide polymorphism (SNPs) and some insertion/deletion (InDels) can be identified, potentially associated with genomic regions of interest for the focused trait. While the potential of male sterility (MS) in globe artichoke offers a promising pathway for the efficient production of $F_1$ hybrids, our understanding of its genetic underpinnings remains rudimentary, and only a few associated loci have been identified in the *Cynara* genus, without a clear delineation of the underlying genetic pathways. This study aims to address this gap by leveraging an $F_2$ population developed from a cross between a MS globe artichoke and a male fertile cultivated cardoon, analyzed through bulk segregant analysis sequencing (BSA-seq). Here we identified specific genomic regions and candidate genes responsible for male sterility, with a focus on a cytochrome P450 gene variant, CYP703A2, suspected to impair male fertility. Our results provide a better delineation of the genetic architecture of MS in globe artichoke, which will not only fill a significant knowledge gap but also facilitate the development of molecular markers for breeding MS varieties, ultimately advancing globe artichoke breeding programs.

## Materials and methods

Starting from the material developed by Zayas et al., 2019, BSAseq was applied for the identification of QTLs and potential candidate genes for male sterility in globe artichoke. A schematic representation of the experimental workflow is presented in Fig. 1.

### Plant material and bulks construction

An $F_2$ population of 250 individuals was established by selfing an $F_1$ individual obtained by crossing a male

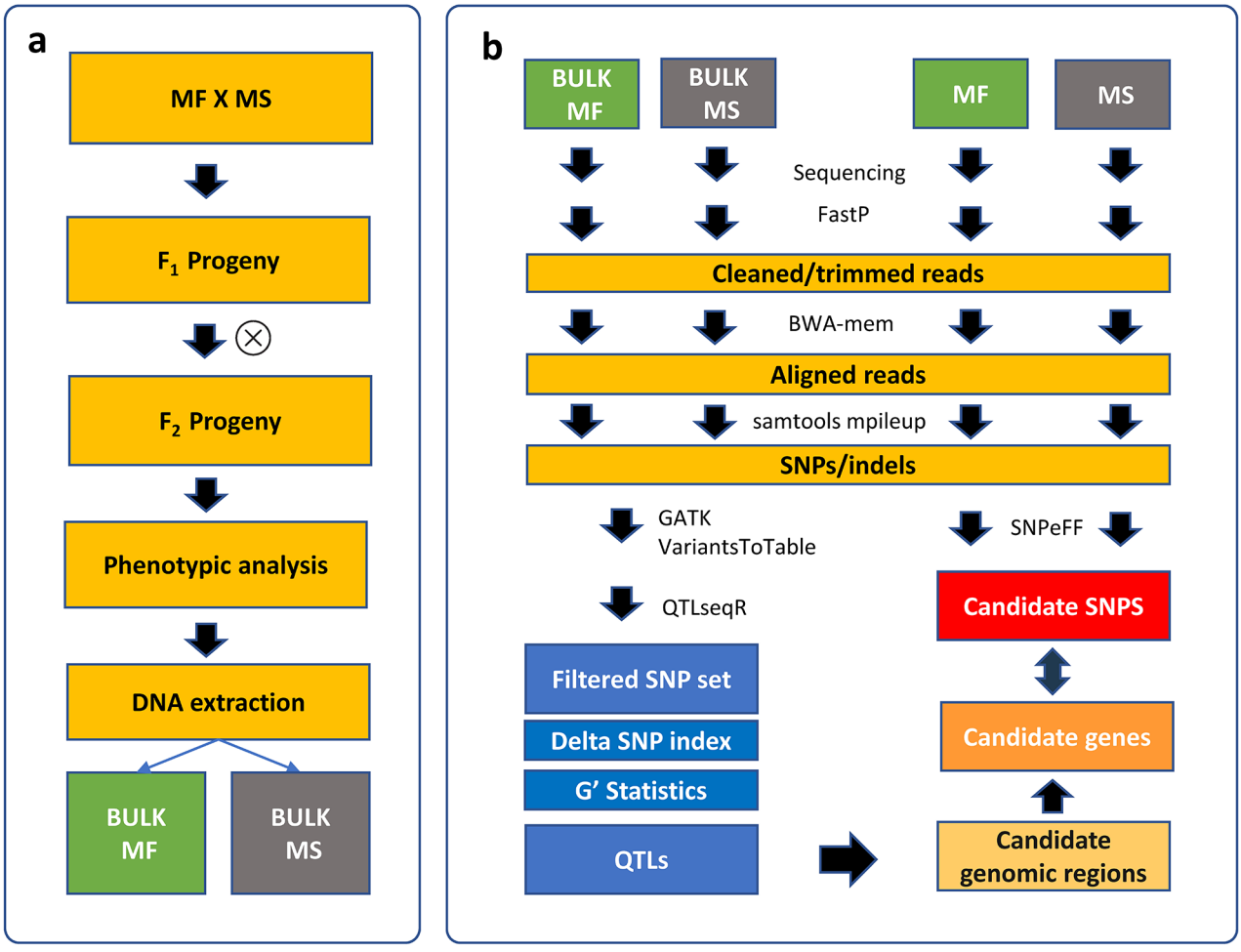


**Fig. 1** Overall presentation of the experimental design

sterile globe artichoke genotype and a cultivated cardoon genotype and was grown as reported by Zayas et al. [16]. Pollen production was screened during two seasons, and two phenotypic groups were identified: male fertile (MF), plants with normal pollen production, and male sterile (MS), plants not producing pollen. Overall, 195 MF plants and 55 MS plants were phenotyped, fitting a 3:1 monogenic mendelian segregation model. Two bulks were constructed, assuming that the MS bulk (Bulk 1) included only homozygous recessive plants, and that the MF bulk (Bulk 2) included heterozygous and homozygous dominant individuals for the locus of interest.

#### DNA extraction and bulk sequencing

Genomic DNA was extracted from fresh leaves of each parental genotype, as well as from 15 individuals of the MS group (Bulk 1) and 15 plants of the MF ones (Bulk 2), using DNeasy Plant mini Kit (QIAGEN). Nucleic acid quantification was performed with the Qubit 2.0 fluorometer (Qubit™ dsDNA BR Assay Kits, Thermo Fisher). Within every bulk, the fifteen genotypes were equimolarly pooled and used for library preparation. The pooled DNA samples were sequenced, together with the parental lines, at the Novogene UK sequencing facility using Illumina's NovaSeq 6000, preparing standard Illumina sequencing libraries with 350 bp insert size. Whole-genome sequencing (PE150) was performed at 45x coverage on both the bulks and at 65x (MS) and 50x (MF) in the parental lines, obtaining an average of 33G raw data per sample.

#### NGS-based BSA analysis

The raw reads obtained from the two bulks were cleaned and mapped on the globe artichoke reference genome (V2 - Acquadro et al. [22], http://www.artichokegenome.unito.it) using *FastP* [23] with standard filtering parameters, and Burrows-Wheeler Aligner program (BWA, v0.7.17, https://sourceforge.net/projects/bio-bwa/files) for the alignment. Variant calling was performed on the aligned sequences, identifying SNPs and INDELs between the sequenced bulks using Samtools mpileup,

with minimum mapping quality equal to 25. SNPs having mapping quality lower than 20 were removed.

NGS-based BSA analysis was performed using the R package QTLseqr (https://github.com/bmansfeld/QTLseqr) developed by Mansfeld and Grumet [24] by calculating SNP index, deltaSNP index, and tricube-smoothed G value as described by Takagi et al. [25], Magwene et al. [18], and Yang et al. [26]. SNPs were filtered with standard parameters suggested by the pipeline, namely: refAlleleFreq=0.20, minTotalDepth=40, maxTotalDepth=400, depthDifference=100, minSampleDepth=20, minGQ=99. Based on the BSA-seq analysis, candidate regions surrounding QTL peaks were recorded. Those genomic coordinates were intersected with annotated SNP datasets.

#### Progeny's parents' resequencing

Parental lines reads were cleaned using *FastP* [23] with standard filtering parameters and mapped onto globe artichoke genome reference (v2, Acquadro et al. [22]; http://www.artichokegenome.unito.it/.) using Burrows-Wheeler Aligner program (BWA, v0.7.17, https://sourceforge.net/projects/bio-bwa/files). Samtools mpileup was used for SNP calling, with minimum mapping quality equal to 25, and filtering SNPs call quality and depth. SNPs having mapping quality lower than 20 were removed. Common variants between parental lines were filtered out and SNPs were then analyzed using the SNPeff (http://pcingola.github.io/SnpEff/) suite to predict their effect on the set of gene models of globe artichoke. The effect of each SNP/indel was classified according to SNPeff software into four classes: (1) "modifier"; (2) "low" impact; (3) "moderate" impact and (4) "high" impact.

#### SNP evaluation of the impact on the biological function

Genomic SNPs in parental lines datasets were analyzed, defining peak regions according to the slope rate around the main SNP in the region. In brief, considering that every significant SNP has to be evaluated as associated with the investigated trait in BSAseq analysis, candidate genes were hypothesized to be more densely located around sharper peaks than flat ones. Given the rapid increase and decrease of G' around the top SNP in sharp peaks, a 500Kb genomic region was selected as candidate for them, while 800Kb was evaluated for flat peaks. Such intervals where further reduced to ~100-200Kb, investigating the closest genes to the peak. In the peak regions, moderate/high impact mutations were considered and the ones in homozygous state were selected. Some candidate moderate impact SNPs were also submitted to Provean analysis (Protein Variation Effect Analyzer algorithm, https://www.jcvi.org/research/provean), to check if the mutation had an impact on the biological protein functions. The score threshold used was set to the standard −2.5 value.

#### Homology modeling of CYP703A2

Homology models for both the WT and Arg424Gln CYP703A2 were built using the software Modeler 9.25 and the online tool SWISS-MODEL [27], using CYP76AH1 crystal structure from *Salvia miltiorrhiza* Bunge (PDB ID 5YLW) as a template. Energy minimization was performed using YASARA Amberff14SB force field and subjected to validation using Molprobity [28] and QMEAN [29]. The best models were selected according to their QMEAN4 score and to the percentage of residues in favored regions. The WT protein showed a QMEAN4 value of −1.34, and the Ramachandran plot showed that 95.79% were in favored regions. The QMEAN4 value for Arg424Gln protein was −1.69 and 96.41% of the residues were in the favored regions of the Ramachandran plot. The difference between the two models in the R424 region were analyzed in the UCSF Chimera software [30], which also allowed the calculation of the hydrogen bonds of the amino acid of interest.

#### dCAPS primer design and experimental validation of DNA polymorphisms

To verify the MS mutation highlighted by Illumina alignment, Sanger sequencing was performed on a Applied Biosystems 3500 Series Genetic Analyzer using the BrilliantDie™ Terminator Cycle Sequencing kit by NimaGen according to standard protocols, followed by dCAPS marker development (http://helix.wustl.edu/dcaps/) using *HpaII* restriction enzyme. 1 µl of the genomic DNA was used as template in a 20 µl PCR containing 10 pmol of forward (GGACACTTTCTCTTTCCTGCA) and reverse primer (TGATATGGGATGATATCAACGTG), 1.5 mM $MgCl_2$, 1 mM dNTP and 1 U Taq polymerase (GoTaq® DNA Polymerase) in the manufacturer's buffer. The amplification program was 94°C/120", 25 cycles of 94 °C/30", 55 °C/30", 72 °C/60", and 10' incubation at 72 °C. PCR products were digested by HpaII restriction endonuclease according to the producer's instructions and run in a 1.5% agarose gel-based electrophoresis to verify polymorphisms and segregation pattern of the dCAPs marker.

## Results

To identify and locate the genomic regions and genes responsible for male sterility in globe artichoke, as well as to develop molecular markers to be applied in breeding programs, BSA-seq approach was applied in an $F_2$ segregating population, previously developed and phenotyped for MS by Zayas et al. [16].

### Whole-genome sequencing of male sterile (MS) and male fertile (MF) bulks

We performed Illumina sequencing (45X coverage; 2×150 bp) of two bulks (MS bulk, and MF bulk), each one containing 15 genotypes concordant for the investigated trait, as well as the two parental lines (65X and 50X coverage for MS and MF, respectively). Genome sequencing of the two bulks yielded 447 million raw pair-end reads (in 224 million clusters), while the two parents yielded ~569 million raw pair-end reads (in 284 million clusters, Table 1).

The sequence data have been deposited into NCBI as Short Read Archive (SRA) files under the Bioproject PRJNA892759.

The sequenced reads of the two bulks were aligned to the reference genome (v2, Acquadro et al. [16]), detecting a total of 7,023,150 SNP/Indel variants. The two parents showed 8,203,723 and 4,748,963 SNPs/indels for MS and MF, respectively. A SnpEff analysis focused on coding regions (CDS) was also conducted on the MS parent. A total of 202,645 SNPs/indel in CDS were found (~2.5% of the total genomic variants). The majority (~51.2%) were non-synonymous (missense), followed by synonymous (silent; ~47.6%), and nonsense (~1.2%) mutations.

### BSA-seq analysis

To analyze the differences between contrasting phenotypes, two statistics are calculated (ΔSNP and G'), based on allele counts. The first statistic, ΔSNP, is calculated using the SNP index of the two bulks, which is determined by dividing the alternate allele depth by the total read depth. ΔSNP is the difference between the SNP index of the bulk defined as high and the bulk defined as low. The second statistic, G', is a tricube-smoothed G statistic. Based on the alleles count, BSAseq can quickly associate a specific locus with candidate genomic regions.

BSA-seq analysis revealed four chromosomal regions (one on chromosome 4 and 12; two on chromosome 14) putatively involved in male sterility, showing G' values above the threshold (G' > 6.5; Fig. 2a, Suppl. Table 1).

All the regions surrounding the G' peaks were investigated in the parental genomes showing many genes and SNPs. Sub-regions at the top of the peaks were selected, narrowing the max G' SNPs. The QTL with the highest G' value (10.33) was present in the first part of chromosome 14 (namely chr. 14a - Table 2); a second region (14b) was spotted in the last part of this chromosome with a max G' of 8.31. In chromosomes 4 and 12, peaks were detected at 8.18 Mb (4) and 2.11 Mb (12), with a max G' value of 8.34 and 6.34, respectively (Table 2).

### Candidate genes over QTL regions

*Chromosome 4* - A 18.2 Mb QTL region was identified (5.53–23.74 Mb, Fig. 2b). The G' peak (8.34, Fig. 2f) was located at 8.17 Mb, and the narrowed region around it was ~800Kb long (7.54-8.43 Mb). Sixty genes were present in the peak region, and 18,417 SNPs were identified between the parents (Suppl. Table S2). Among them, 263 SNPs showed a moderate impact, whilst 14 were categorized as high impact SNPs. A 200Kb interval (8.07 Mb-8.27 Mb) around the peak was further scanned and nine genes were present in this interval (Suppl. Table 2): (i) a Leucine-Rich Repeats (LRR) receptor-like serine/threonine-protein kinase, (ii) the Protein IMPAIRED IN BABA-INDUCED STERILITY 1 (IBS1), involved in female fertility [31], (iii) two isoforms of Formin-like protein 20 (FH20), (iv) the E3 ubiquitin-protein ligase PRT1, and (v) four unknown proteins. No high impact mutations were present in these genes, and the ones with moderate impact and in homozygous state were checked with Provean without predicting any deleterious effects (data not shown).

*Chromosome 12* - A 2.07 Mb QTL region was highlighted (1.70-3.77 Mb, Fig. 2c). The G' peak (6.34) was located at 2.11 Mb (Fig. 2g), and the narrowed region around it was ~800Kb long (2.04–2.90 Mb). Seventy-nine genes were annotated in this region, as well as 10,740 SNPs (Suppl. Table S2). Among them, 68 SNPs showed a moderate impact, whilst three were categorized as high impact SNPs. A 100 kb interval (2.05–2.14 Mb) around the peak was scanned and six genes were identified: (i) an aldehyde oxidase (GLOX1), (ii) a mitochondrial outer membrane protein porin of 36 kDa, (iii) a B3 domain-containing protein, (iv) a pentatricopeptide repeat-containing protein (PCMP-E22), belonging to a class of genes reported as involved in cytoplasmatic male sterility restoration in rice and *Arabidopsis* [32, 33], (v) a LRR receptor-like serine/threonine-protein kinase, and (vi) a MYB124 transcription factor. No high impact mutations were present in these genes, and the ones with

**Table 1** Sequencing statistics

| DNA | clusters | N° raw reads | bp | clusters | N° clean reads | bp | % cleaned | Final Coverage (X) | phenotype |
|---|---|---|---|---|---|---|---|---|---|
| MS parent | 122,236,815 | 244,473,630 | 36,671,044,500 | 122,035,424 | 244,070,848 | 36,430,808,266 | 99.84% | 50.2 | sterile |
| MF parent | 162,159,276 | 324,318,552 | 48,647,782,800 | 161,470,716 | 322,941,432 | 48,117,090,189 | 99.58% | 66.3 | fertile |
| Bulk male sterile | 113,565,092 | 227,130,184 | 34,069,527,600 | 113,097,153 | 226,194,306 | 33,677,359,194 | 99.59% | 46.4 | sterile |
| Bulk male fertile | 110,141,514 | 220,283,028 | 33,042,454,200 | 109,811,254 | 219,622,508 | 32,701,145,992 | 99.70% | 45.0 | fertile |

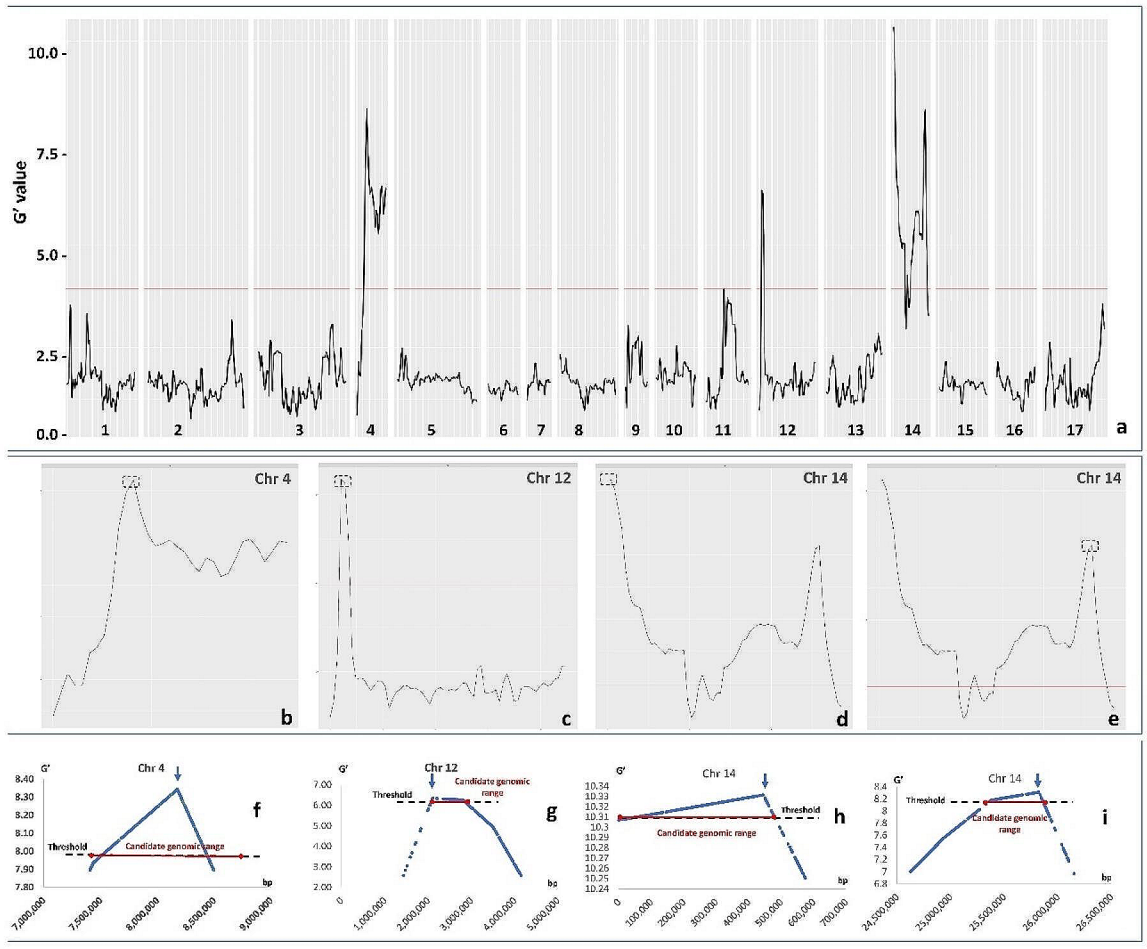


**Fig. 2** Quantitative trait loci (QTL) for male sterility identified by QTLseqr. Plots produced by the plotQTLStats() function with a 1 Mb sliding window: **a**) The tricube-smoothed G' value. **b**-**e**) Detailed plots of the tricube-smoothed G' value produced over chromosomes 4, 12 and 14; **f**-**i**) details of the peak regions and selected thresholds

moderate impact and in homozygous state were checked with Provean without predicting any deleterious effects (data not shown).

*Chromosome 14* - Two G' peaks were spotted (named 14a and 14b, Fig. 2d and e, respectively). The first one included a 9.45 Mb long region (0.2–9.65 Mb, Fig. 3) with a G' peak (10.33) at 0.44 Mb (Fig. 2h). By focusing on a region of ~500 kb (0-0.57 Mb) around the peak, fifty-five genes and 7,818 SNPs were identified between the parents (Suppl. Table S2), including a 3-ketoacyl-CoA synthase-like gene, a class of proteins that have been reported to be involved in the pathways of cuticular wax and cutin biosynthesis [34]. Among them, 108 SNPs showed a moderate impact, whilst 12 were categorized as high impact SNPs. This region was further investigated within a ~100 kb interval (0.43–0.54 Mb), highlighting three high impact mutations, together with several moderate impact variants. Overall, twelve genes were identified, including (i) a pentatricopeptide repeat-containing protein, a class of genes indicated has players in pollen development and cytoplasmatic male sterility restoration in rice and *Arabidopsis* [32, 33], (ii) a calcium-transporting ATPase (LCA1), (iii) a serine/arginine-rich splicing factor 31 (RS31), (iv) a bifunctional fucokinase/fucose pyrophosphorylase (FKGP), (v) an aluminum-activated malate transporter (ALMT10), (vi) a cytochrome P450 (CYP703A2), belonging to the CYP703 family, reported to provide sporopollenin building blocks during pollen development [35], and (vii) a histone-lysine N-methyltransferase (ATX4). Three high impact mutations were detected in three different genes with unknown function, together with some moderate impact SNPs in homozygous state. All the missense variants were checked with Provean and none, except two,

**Table 2** Statistics on selected QTL for male sterility: details on the peak regions, genes and SNP with their potential impact are reported

| | | | | | | | | | | SNP impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Chr** | **sub-region** | **Start** | **end** | **Length** | ***N°* SNPs** | **avg. SNPs/Mb** | **max G'** | **pos. Max G'** | **G' candidate range (bp)** | **genes** | **LOW** | **MODERATE** | **HIGH** |
| 4 | - | 5,527,229 | 23,739,762 | 18,212,533 | 72,638 | 3,988 | 8.34 | 8,176,221 | chr4:7,544,909-8,425,655 | 60 | 257 | 263 | 14 |
| 12 | - | 1,702,450 | 3,773,932 | 2,071,482 | 3,794 | 1,832 | 6.34 | 2,111,758 | chr12:2,040,412-2,945,755 | 82 | 55 | 69 | 3 |
| 14 | a | 202 | 9,655,187 | 9,654,985 | 34,986 | 3,624 | 10.33 | 444,700 | chr14:202–574,964 | 55 | 123 | 108 | 12 |
| 14 | b | 13,943,952 | 27,631,802 | 13,687,850 | 56,370 | 4,118 | 8.31 | 25,822,452 | chr14:25,330,927 – 25,875,057 | 43 | 61 | 41 | 2 |

were predicted to have deleterious effects. The first was in the ATX4 gene (104 kb far from the G' peak), while the second was highly closed to the G' peak (10 kb apart). The latter (455,565 bp) was a missense variant (G1271A) present in the CYP703A2 gene (V2_14g000490.1), a cytochrome P450 involved in the exin synthesis (Suppl. Table 2). The mutation was predicted to produce an amino-acid substitution (Arg424Gln), which Provean analysis reported as highly deleterious (score −6.392).

A second peak (14b) surrounded a 13.69 Mb long QTL region (13.94-27.63 Mb, Fig. 2e) with a G' peak (8.31) at 25.82 Mb (Fig. 2i). By including a region of ~500 Kb (25.3–25.9 Mb) around the peak, 43 genes and 5,650 SNPs were highlighted between parents (Suppl. Table S2). Among them, 41 SNPs showed a moderate impact, whilst 2 were categorized as high impact SNPs. By focusing on a 100 kb region around the peak (14b; 25.72–25.92 Mb), two genes were identified. Both genes coded for unknown proteins, showing one SNPs in heterozygous state, with a low/moderate impact.

### Derived cleaved amplified polymorphic sequences (dCAPs) marker design

To allow the discrimination between MS and MF genotypes within the $F_2$ population, a single dCAPs marker was developed around the SNP located at 455,565 bp in chromosome 14. As this SNP was not associated with any restriction enzyme, a neighbor one (455,570 bp), belonging to the same male sterile haplotype, was selected for the development of a marker for the locus, as part of the HpaII (CCGG) restriction site (Fig. 4). The developed marker (named *14-455570-ms*) was validated on the $F_2$ population, demonstrating to be useful in the assessment of the MS/MF phenotype (Supp. Figure 1 and Supp. Table 3). As depicted in Fig. 3, the non-reference allele (AA) identified in the MF parent allowed the restriction cut of the amplified region in two co-migrating sequences of 95 bp and 58 bp. The cut (MF - AA) and uncut (MS - aa) sequences can be easily visualized by gel electrophoresis (Fig. 5), and their segregation in the population allowed the recognition of the Bulk 1 from the Bulk 2. As expected, "aa" genotypes were only present in the MS bulk (Bulk 1), while dominant homozygous (AA) and heterozygous (Aa) states were present in the MF bulk (Bulk 2). The validation of the marker on the whole $F_2$ population confirmed a 3:1 (MF: MS) segregation of the locus, strongly associating the dCAPs marker (*14-455570-ms*) with the observed phenotypic segregation.

### Sequence validation and Arg424Gln investigation through CYP703A2 homology modeling

The genomic region containing the Arg424Gln missense variant in CYP703A2 was inspected in the MS and MF parental lines and it was found fixed in homozygous

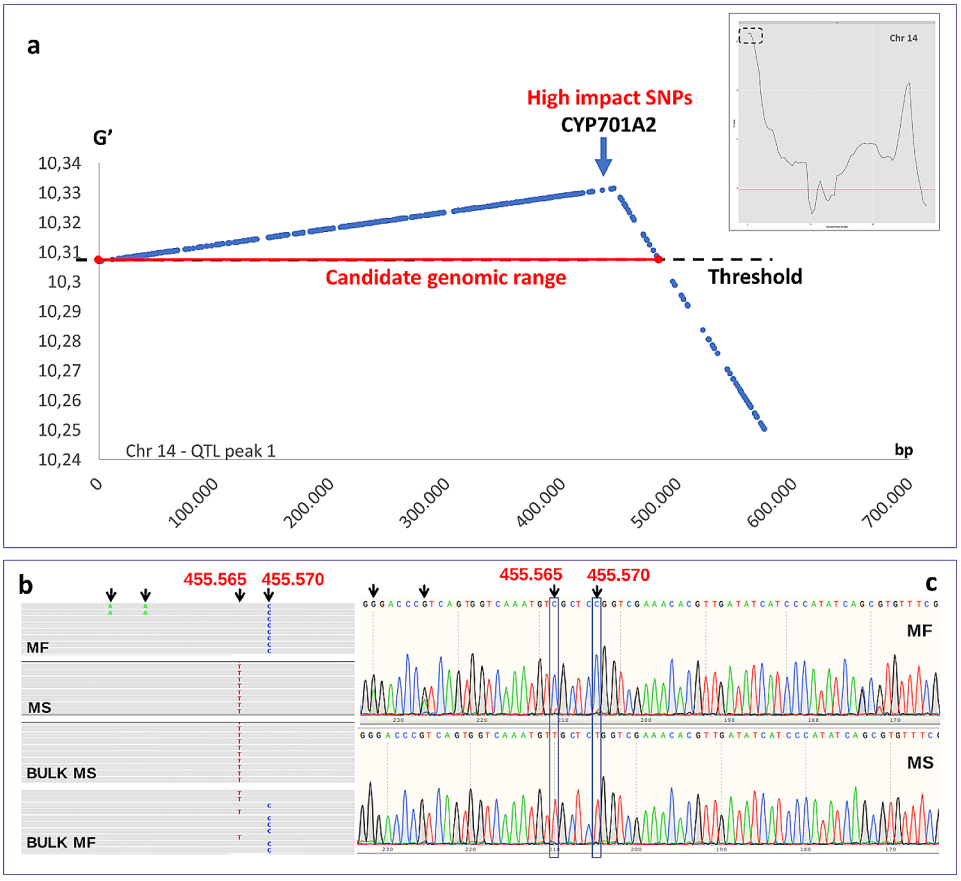


**Fig. 3** **a**) Detail of the 14a, one of the two G' peaks in chromosome 14; **b**) Alignment profile with Illumina based reads of the two parents (MF and MF) and the two bulks (MS and MF) in the peak region. **c**) Sanger validation of the SNPs observed with Illumina sequencing

state in the MS parent (Fig. 3b and c). Such mutations were further validated through Sanger sequencing of the region (Fig. 3c). According to both Sanger and Illumina-based resequencing alignment (Fig. 3b), a cytosine in position 455,565 (reference-like) in the MF parent, was mutated in a thymine in the MS parent (Suppl. Table 2), leading to the substitution of Arg424 with a Gln (Suppl. Table 2). The mutation on CYP703A2 was investigated through protein modeling, to assess the effective impact of the predicted amino acid substitution (Arg424Gln) on protein structure. A homology model was built (see Experimental Procedures) using the best matching crystal structure, CYP76AH1 from *Salvia miltiorrhiza*, sharing 31% of primary sequence identity and 50% of homology, as template (Fig. 3a and b). The analysis highlighted the role of Arg424 as the amino acid involved in a salt bridge with a glutamate residue that is part of the fingerprint motif EX1×2R (Fig. 4a), located on helix K. Together with the E and the R from this motif, Arg424 is part of the so-called ERR triad, a conserved region in the superfamily involved in salt bridges and protein folding [36].

## Discussion

### Male sterility in globe artichoke

Even if globe artichoke is traditionally vegetatively propagated, this form of propagation has a great impact on production costs, which tend to be higher for the intensive labor required for transplanting and plant losses in field preparation. On the other hand, the use of the achene as a reproductive unit makes it possible to treat the crop as annual, increasing field uniformity and reducing planting costs, as well as pathogens diffusion [37]. In the last decades, this has pushed the popularity of seed-propagated cultivars and the production of $F_1$ hybrids which, in some cases through the exploitation of male sterile (MS) genotypes, have been successfully introduced in cultivation [11]. Indeed, to take advantage of the low seed costs and the heterosis phenomenon through hybrid production, it is essential to have an effective globe artichoke male-sterility system available [16], which allows to avoid selfing and easies the crossing process. Several genotypes are today available carrying MS phenotype [11], but the genetic bases of this trait have been, to date, poorly explored.

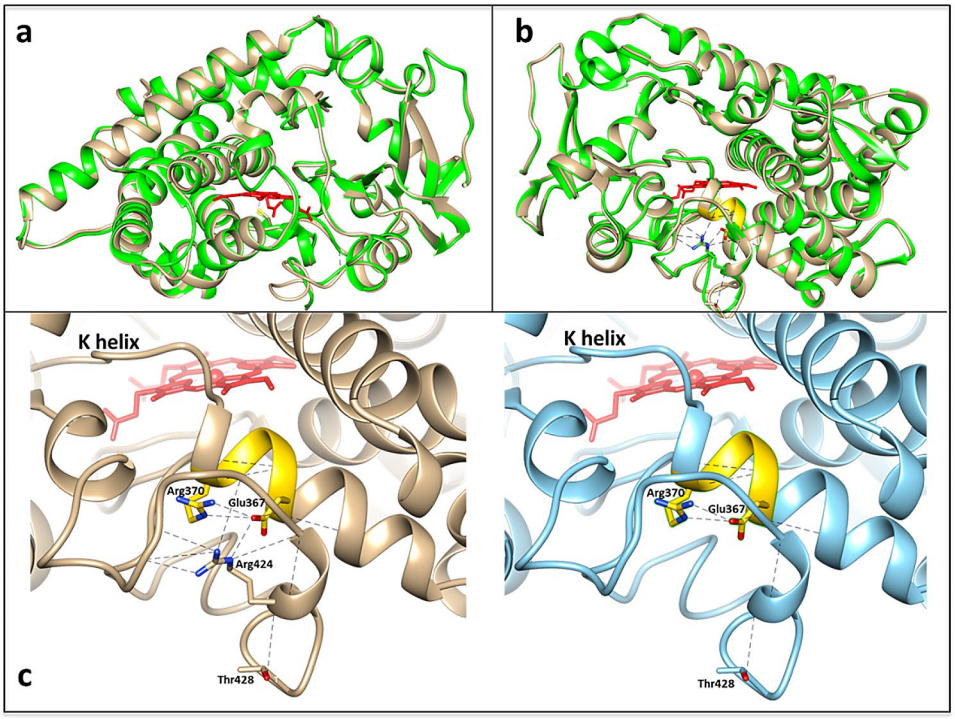


**Fig. 4** **a**) CYP76AH1 crystal structure from *Salvia miltiorrhiza* used for homology modeling of CYP703A2; **b**) Modeled globe artichoke WT CYP703A2 protein structure; **c**) Comparison between the models of CYP703A2 protein in globe artichoke MF (left) and MS (right) in the region of interest. ERR triad is highlighted in yellow on K helix, the amino acids interacting with the mutated site are shown, as well as the heme (in red)

### Investigating globe artichoke genetics with BSA-seq, allowing marker development for breeding

We analyzed an $F_2$ population of 250 offsprings derived from a cross of a MS globe artichoke with a male fertile (MF) cultivated cardoon *(C. cardunculus* var. *altilis*), segregating to produce vital/not vital pollen, and fitting a monogenic Mendelian 3:1 model. This population was here analyzed with the BSA-seq approach [38], previously proven to be efficient for monogenic/oligogenic character spotting and QTL mapping in many different species [39–47]. Such technology was applied to fully address the genetic determinant of MS and overcome limitations due to the lower resolution power of SRAP markers, previously highlighted by Zayas et al [16]. BSA-seq analysis revealed four chromosomal regions (4, 12, 14a and 14b) putatively involved in male sterility (Fig. 2). A QTL with the noticeably highest G' value (10.33) was detected in the first part of chromosome 14. In the paper by Zayas et al [16] five candidate regions were already spotted as a candidate for the MS trait, and one revealed to be present on chromosome 14, in a proximal chromosomal region. By focusing on the 14a region (Fig. 3a), in the 20 kb around the peak position (444,700 bp), four genes were spotted, as well as 5 SNPs, of which only one was predicted to produce a homozygous missense variant. The *14-455570-ms* marker validation, conducted on the whole $F_2$ population, confirmed its strong association with the observed male sterility trait (Supp. Figure 1). The male sterile phenotype was always observed in the population in association with a single recessive mutation in the second exon of gene, while wild type and heterozygous individuals in the progeny had a fertile phenotype. This is consistent with the results in rice [48] and Arabidopsis [35], where phenotypes are alternatively caused by a single recessive mutation in the CYP703A2 gene, causing a coding frame shifting in the second exon, and a homozygous T-DNA insertion in the gene, respectively.

### Investigating a candidate gene for male sterility

The mutated gene was a cytochrome P450 (CYP703A2, V2_14g000490.1) coding for a protein responsible for sporopollenin synthesis [35]. Many genes are known to be involved in the sporopollenin synthesis/transport [49], and plants with mutations in these genes have severe defects in exin/sexine layers, anther cuticle, and are usually complete male sterile [50]. As example, CYP703A3 belongs to the CYP71 clan, which catalyzes the biochemical pathway of fatty acid hydroxylation [51]. It has been demonstrated that the same cytochrome P450 hydroxylase, in presence of NADPH, is able to catalyze *in-chain* hydroxylation of FAs as sporopollenin precursors, with a preferential hydroxylation of lauric acid at the C-7 position [35]. This gene has also been reported to be essential for the development of anther cuticle and pollen exine in

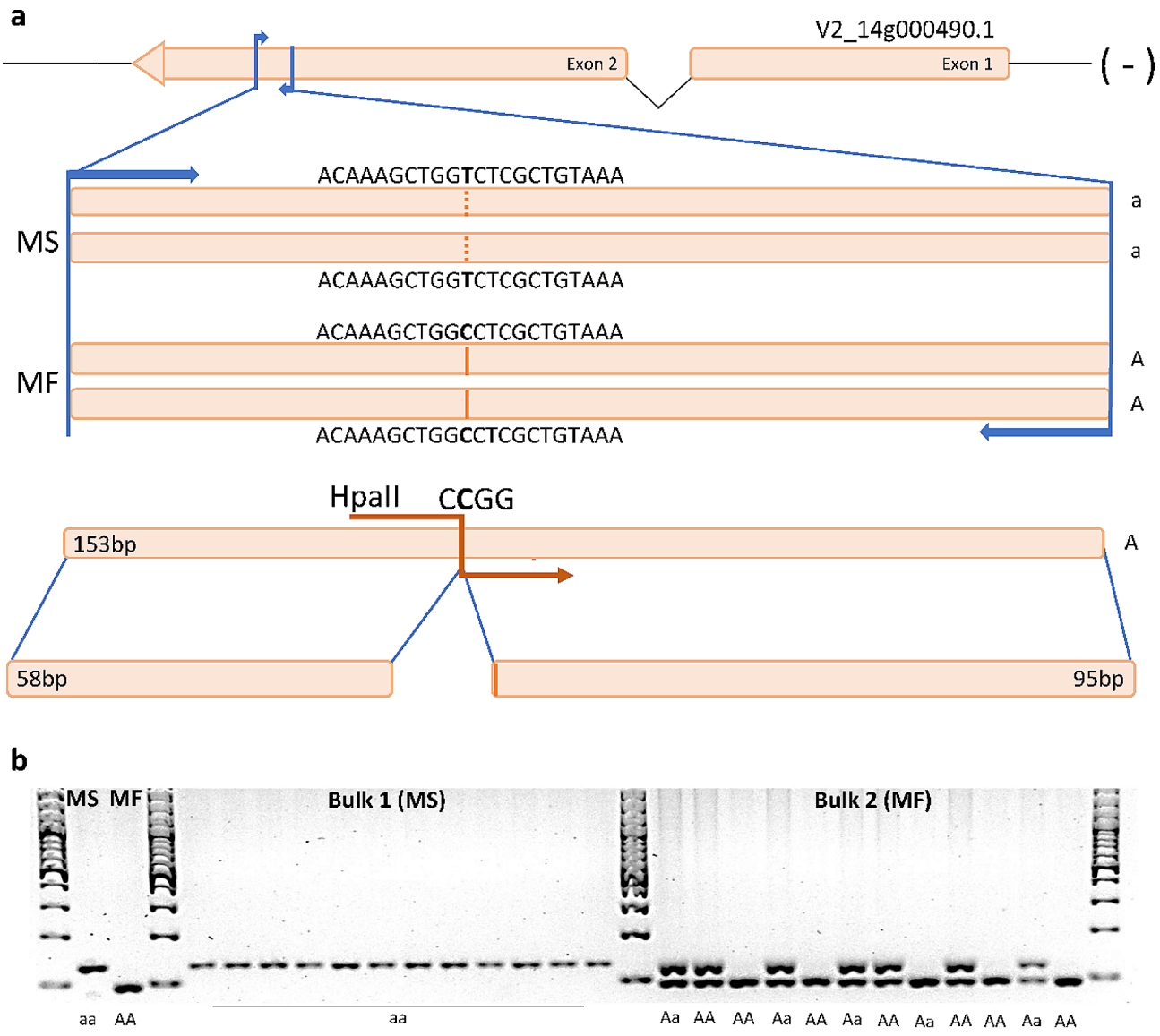


**Fig. 5** **a**) *V2_14g000490.1* (*CYP703A2*) gene structure. The two alleles are reported, as well as the sequence sizing produced by enzymatic cut with HpaII; **b**) Validation of the haplotypic designed CAPs marker (14-455570-ms) in the population. From left to right, parental lines (MS and MF) are shown, followed by Bulk 1 (one uncut fragment) and Bulk 2 (both heterozygous and recessive homozygous states are present)

rice, where mutants showed a fully pollen sterile phenotype [48]. In Arabidopsis, CYP703A2 mutants produce partially sterile pollen grains displaying abnormal exine and sporopollenin deposition [35]. The crucial role of this enzyme in reproductive tissues has been also shown by its overexpression in *Arabidopsis*, where altered expression through transgenesis was able to increase silique size and seed number, altering the contents of fatty acids composition of cutin monomer in the siliques [51]. In our case, the mutation in the globe artichoke CYP703A2 generated an amino-acid substitution (Arg424Gln) with a highly predicted deleterious effect. We investigated this effect through a 3D analysis of the protein (Fig. 4), confirming Arg424 as the amino acid involved in a salt bridge with a glutamate residue within the key motif EX1×2R (ERR) in helix K. The ERR triad is highly conserved across the cytochrome P450 superfamily as it is involved in salt bridges, essentials for a correct folding and heme incorporation [36, 52, 53]. It has been widely reported that mutations of the residues of the ERR triad in different cytochromes P450 resulted in loss of function due to the lack of heme incorporation [54–59]. Here, Gln424 in the mutated protein likely interferes with the salt bridge formation with Glu367 in helix K, potentially leading to the loss of the ERR triad (Fig. 4c). The convergence over this gene in different species (both monocot and dicots) is suffragated by the crucial role of this specific p450 in the sporopollenin synthesis [35, 49], coupled with the

presence of a single copy of the gene in all the investigated plant species [35, 50, 60]. Accordingly, genomic analysis of globe artichoke confirmed the presence of the gene in a single copy. Despite the absence of established protocols for gene knockout in globe artichoke, this enzyme presents itself as a prime target for CRISPR/Cas9 genetic manipulations. However, considering the male sterility observed upon loss of function in CYP703A2 in both monocot and dicot plants [35, 50], together with the extensive list of literature in yeast highlighting the deleterious effects of mutations in the conserved ERR triad [36, 52, 53], it is reasonable to deduce that a similar knockout may yield little additional insight.

### Breeding perspectives for male sterility

The discovery and markers design for the locus on top of chromosome 14, and its association with male sterility in globe artichoke undoubtedly opens new breeding possibilities, bringing to light the potential for leveraging male sterility to enhance hybrid seed production. However, the route from genetic discovery to practical application in breeding programs is highly challenging. The complexity of male sterility, influenced by genetic and environmental factors, might benefit from the integration of traditional approached with modern computational strategies, such as machine learning, to successfully integrate this trait in breeding programs. By investigating syntenic loci from model and well-studied species, minor crops can benefit from a mole of research material that could be hardly achievable to produce [61]. Moreover, harnessing ML algorithms to predict male sterility genomic architectures, researchers can gain insights into the underlying genetic interactions and environmental dependencies. This predictive capability can extend beyond traditional breeding approaches, allowing for the anticipation of breeding outcomes in diverse environmental contexts. The use of functional genomics serves as a foundation for these predictive models [62, 63], offering a deeper understanding of gene function and regulation in the context of male sterility. Moreover, the application of predictive breeding promises to enhance the precision and efficiency of breeding programs [64–67]. By leveraging the genomic bases of complex polygenic adaptive trait architectures, predictive molecular breeding can potentially overcome the challenges posed by traits like male sterility, which may exhibit Mendelian inheritance patterns yet are influenced by a multitude of factors. The exploration of these novel perspectives, can unlock new possibilities for the development of superior hybrids, thereby contributing to sustainable agricultural practices and food security.

## Conclusions

The present research on globe artichoke male sterility highlights the potential of using genetic markers, such as the developed dCAPS marker, for breeding programs aimed at enhancing crop yields through hybrid seed production. By elucidating a genetic basis for male sterility, this study provides a template for further research in other plant species, potentially leading to breakthroughs in agricultural productivity and sustainability. Moreover, by enabling the development of more efficient and cost-effective plant breeding, the broader implications of these findings underscore the importance of genetic research in addressing global food security challenges.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-05119-z.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

### Acknowledgements

The authors express their sincere gratitude to Prof. Ben Mansfeld for the support on data analysis interpretation.

### Author contributions

A.A, C.C, E.M, and E.P conceived the study. M.M and A.Z. extracted and bulked the genetic material. M.M, M.P, and A.A performed bioinformatic analysis. A.Z and E.M provided materials. M.M, G.D, and G.G performed the protein modelling. M.M, C.C, A.A and E.P developed and validated the dCAPS marker. M.M, A.Z, E.M and A.A wrote the manuscript. All authors critically revised and approved the manuscript.

### Funding

This work was partially supported by the "GREAT: Genetics and genomics of REproductive systems in plAnTs" project, in the framework of the PRIN 2022 (Research Projects of National Relevance) founding of the Italian Ministry of University and Research (MUR).

### Data availability

Sequencing data used in this study are openly available in the NCBI database (PRJNA892759).

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

### Author details

[1]DISAFA, Plant Genetics and Breeding, University of Turin, Turin, Italy
[2]IICAR (Instituto de Investigaciones en Ciencias Agrarias de Rosario), CONICET, Campo Exp. J.F. Villarino, Zavalla, Santa Fe, Argentina
[3]DBIOS, Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy

Received: 13 December 2023 / Accepted: 8 May 2024
Published online: 17 May 2024

**References**


1. Scaglione D, Lanteri S, Acquadro A, Lai Z, Knapp SJ, Rieseberg L, et al. Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. Plant Biotechnol J. 2012;10:956–69.
2. Moglia A, Acquadro A, Eljounaidi K, Milani AM, Cagliero C, Rubiolo P et al. Genome-wide identification of BAHD acyltransferases and in vivo characterization of HQT-like enzymes involved in Caffeoylquinic Acid Synthesis in Globe Artichoke. Front Plant Sci. 2016;7.
3. Portis E, Mauromicale G, Barchi L, Mauro R, Lanteri S. Population structure and genetic variation in autochthonous globe artichoke germplasm from Sicily Island. Plant Sci. 2005;168:1591–8.
4. Mauro R, Portis E, Acquadro A, Lombardo S, Mauromicale G, Lanteri S. Genetic diversity of globe artichoke landraces from sicilian small-holdings: implications for evolution and domestication of the species. Conserv Genet. 2009;10:431–40.
5. Portis E, Barchi L, Acquadro A, Macua JI, Lanteri S. Genetic diversity assessment in cultivated cardoon by AFLP (amplified fragment length polymorphism) and microsatellite markers. Plant Breeding. 2005;124:299–304.
6. Portis E, Acquadro A, Tirone M, Pesce GR, Mauromicale G, Lanteri S. Mapping the genomic regions encoding biomass-related traits in Cynara cardunculus L. Mol Breed. 2018;38:64.
7. FAOSTAT. Food and Agriculture Organization Corporate Statistical Database. 2021.
8. Rau D, Attene G, Rodriguez M, Baghino L, Pisanu AB, Sanna D et al. The Population structure of a Globe Artichoke Worldwide Collection, as revealed by molecular and phenotypic analyzes. Front Plant Sci. 2022;13.
9. Lanteri S, Acquadro A, Saba E, Portis E. Molecular fingerprinting and evaluation of genetic distances among selected clones of globe artichoke (Cynara cardunculus L. var. scolymus L). J Hortic Sci Biotechnol. 2004;79:863–70.
10. Lanteri S, Portis E, Acquadro A, Mauro RP, Mauromicale G. Morphology and SSR fingerprinting of newly developed Cynara cardunculus genotypes exploitable as ornamentals. Euphytica. 2012;184:311–21.
11. Calabrese N, Cravero V, Pagnotta MA. Cynara cardunculus Propagation. In: Portis E, Acquadro A, Lanteri S, editors. The Globe Artichoke Genome. Cham: Springer International Publishing; 2019. pp. 21–40.
12. Naresh P, Lin S, Lin C, Wang Y, Schafleitner R, Kilian A et al. Molecular markers Associated to two non-allelic genic male sterility genes in Peppers (Capsicum annuum L). Front Plant Sci. 2018;9.
13. Principe JA. Male-sterility in Artichoke. HortScience. 1984;19:864–864.
14. Basnitzki Y, Zohary D. A seed-planted Cultivar of Globe Artichoke. HortScience. 1987;22:678–9.
15. Stamigna C, Micozzi F, Pandozy G, Crinò P, Saccardo F. Produzione Di Ibridi F1 di carciofo mediante impiego di cloni maschiosterili. Italus Hortus. 2004;11:29–33.
16. Zayas A, Martin E, Bianchi M, López Anido F, Cravero V. Elucidating the genetic male sterility in Cynara cardunculusL. Through a BSA approach: identification of associated molecular markers. Euphytica. 2019;216:8.
17. Michelmore RW, Paran I, Kesseli RV. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A. 1991;88:9828–32.
18. Magwene PM, Willis JH, Kelly JK. The statistics of Bulk Segregant Analysis using next generation sequencing. PLoS Comput Biol. 2011;7:e1002255.
19. Li Z, Xu Y. Bulk segregation analysis in the NGS era: a review of its teenage years. Plant J. 2022;109:1355–74.
20. Scaglione D, Reyes-Chin-Wo S, Acquadro A, Froenicke L, Portis E, Beitel C, et al. The genome sequence of the outbreeding globe artichoke constructed de novo incorporating a phase-aware low-pass sequencing strategy of F1 progeny. Sci Rep. 2016;6:19427.
21. Acquadro A, Barchi L, Portis E, Mangino G, Valentino D, Mauromicale G, et al. Genome reconstruction in Cynara cardunculus taxa gains access to chromosome-scale DNA variation. Sci Rep. 2017;7:5617.
22. Acquadro A, Portis E, Valentino D, Barchi L, Lanteri S. Mind the gap: Hi-C Technology boosts contiguity of the Globe Artichoke Genome in Low-Recombination regions. G3 Genes|Genomes|Genetics. 2020;10:3557–64.
23. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
24. Mansfeld BN, Grumet R. QTLseqr: an R Package for Bulk Segregant Analysis with Next-Generation sequencing. Plant Genome. 2018;11:180006.
25. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. 2013;74:174–83.
26. Yang Z, Huang D, Tang W, Zheng Y, Liang K, Cutler AJ, et al. Mapping of quantitative trait loci underlying Cold Tolerance in Rice Seedlings via High-Throughput sequencing of pooled extremes. PLoS ONE. 2013;8:e68433.
27. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46:W296–303.
28. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: more and better reference data for improved all-atom structure validation. Protein Sci. 2018;27:293–315.
29. Benkert P, Tosatto SCE, Schomburg D. QMEAN: a comprehensive scoring function for model quality assessment. Proteins. 2008;71:261–77.
30. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.
31. Jakab G, Cottier V, Toquin V, Rigoli G, Zimmerli L, Métraux J-P, et al. β-Aminobutyric acid-induced resistance in plants. Eur J Plant Pathol. 2001;107:29–37.
32. Zhang Q, Xu Y, Huang J, Zhang K, Xiao H, Qin X, et al. The Rice Pentatricopeptide repeat protein PPR756 is involved in Pollen Development by affecting multiple RNA editing in Mitochondria. Front Plant Sci. 2020;11:749.
33. Durand S, Ricou A, Simon M, Dehaene N, Budar F, Camilleri C. A restorer-of-fertility-like pentatricopeptide repeat protein promotes cytoplasmic male sterility in Arabidopsis thaliana. Plant J. 2021;105:124–35.
34. Zhang M, Liu J, Ma Q, Qin Y, Wang H, Chen P, et al. Deficiencies in the formation and regulation of anther cuticle and tryphine contribute to male sterility in cotton PGMS line. BMC Genomics. 2020;21:825.
35. Morant M, Jørgensen K, Schaller H, Pinot F, Møller BL, Werck-Reichhart D, et al. CYP703 is an ancient cytochrome P450 in land plants catalyzing in-chain hydroxylation of Lauric Acid to provide Building blocks for Sporopollenin Synthesis in Pollen. Plant Cell. 2007;19:1473–87.
36. Bak S, Beisson F, Bishop G, Hamberger B, Höfer R, Paquette S, et al. Cytochromes P450. Arabidopsis Book. 2011;9:e0144.
37. López-Anido F, Martin E. Globe Artichoke (Cynara cardunculus var. scolymus L.) breeding. In: Al-Khayri JM, Jain SM, Johnson DV, editors. Advances in plant breeding strategies: Vegetable crops: volume 10: leaves, flowerheads, Green pods, mushrooms and truffles. Cham: Springer International Publishing; 2021. pp. 303–30.
38. Mardis ER, Next-Generation DNA, Sequencing Methods. Annu Rev Genom Hum Genet. 2008;9:387–402.
39. Meijnen J-P, Randazzo P, Foulquié-Moreno MR, van den Brink J, Vandecruys P, Stojiljkovic M, et al. Polygenic analysis and targeted improvement of the complex trait of high acetic acid tolerance in the yeast Saccharomyces cerevisiae. Biotechnol Biofuels. 2016;9:5.
40. Yaobin Q, Peng C, Yichen C, Yue F, Derun H, Tingxu H, et al. QTL-Seq identified a major QTL for grain length and weight in Rice using Near Isogenic F2 Population. Rice Sci. 2018;25:121–31.
41. Lu H, Lin T, Klein J, Wang S, Qi J, Zhou Q, et al. QTL-seq identifies an early flowering QTL located near flowering locus T in cucumber. Theor Appl Genet. 2014;127:1491–9.
42. Haase NJ, Beissinger T, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, et al. Shared genomic regions between derivatives of a large segregating Population of Maize Identified using bulked segregant analysis sequencing and traditional linkage analysis. G3 (Bethesda). 2015;5:1593–602.
43. Illa-Berenguer E, Van Houten J, Huang Z, van der Knaap E. Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. Theor Appl Genet. 2015;128:1329–42.
44. Kaminski KP, Kørup K, Andersen MN, Sønderkær M, Andersen MS, Kirk HG, et al. Next generation sequencing bulk Segregant Analysis of Potato support that Differential Flux into the cholesterol and stigmasterol metabolite pools is important for Steroidal Glycoalkaloid Content. Potato Res. 2016;59:81–97.
45. Cao Y, Zhang K, Yu H, Chen S, Xu D, Zhao H, et al. Pepper variome reveals the history and key loci associated with fruit domestication and diversification. Mol Plant. 2022;15:1744–58.
46. Imerovski I, Dedić B, Cvejić S, Miladinović D, Jocić S, Owens GL, et al. BSA-seq mapping reveals major QTL for broomrape resistance in four sunflower lines. Mol Breed. 2019;39:41.

47. Tassone MR, Bagnaresi P, Desiderio F, Bassolino L, Barchi L, Florio FE, et al. A genomic BSAseq Approach for the characterization of QTLs underlying resistance to Fusarium oxysporum in Eggplant. Cells. 2022;11:2548.
48. Yang X, Wu D, Shi J, He Y, Pinot F, Grausem B, et al. Rice CYP703A3, a cytochrome P450 hydroxylase, is essential for development of anther cuticle and pollen exine. J Integr Plant Biol. 2014;56:979–94.
49. Wang K, Guo Z-L, Zhou W-T, Zhang C, Zhang Z-Y, Lou Y, et al. The regulation of Sporopollenin Biosynthesis genes for Rapid Pollen Wall formation. Plant Physiol. 2018;178:283–94.
50. Han Y, Zhou S-D, Fan J-J, Zhou L, Shi Q-S, Zhang Y-F, et al. OsMS188 is a Key Regulator of Tapetum Development and Sporopollenin Synthesis in Rice. Rice. 2021;14:4.
51. Kim J, Silva J, Park C, Kim Y, Park N, Sukweenadhi J, et al. Overexpression of the Panax ginseng CYP703 alters Cutin Composition of Reproductive Tissues in Arabidopsis. Plants. 2022;11:383.
52. Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J. Structure and function of cytochromes P450: a comparative analysis of three crystal structures. Structure. 1995;3:41–62.
53. Di Nardo G, Zhang C, Marcelli AG, Gilardi G. Molecular and structural evolution of cytochrome P450 aromatase. Int J Mol Sci. 2021;22:631.
54. Furuya H, Shimizu T, Hirano K, Hatano M, Fujii-Kuriyama Y, Raag R, et al. Site-directed mutageneses of rat liver cytochrome P-450d: catalytic activities toward benzphetamine and 7-ethoxycoumarin. Biochemistry. 1989;28:6848–57.
55. Shimizu T, Tateishi T, Hatano M, Fujii-Kuriyama Y. Probing the role of lysines and arginines in the catalytic function of cytochrome P450d by site-directed mutagenesis. Interaction with NADPH-cytochrome P450 reductase. J Biol Chem. 1991;266:3372–5.
56. Yoshikawa K, Noguti T, Tsujimura M, Koga H, Yasukochi T, Horiuchi T, et al. Hydrogen bond network of cytochrome P-450cam: a network connecting the heme group with helix K. Biochim Biophys Acta. 1992;1122:41–4.
57. Kitamura M, Buczko E, Dufau ML. Dissociation of Hydroxylase and lyase activities by Site-Directed mutagenesis of the rat P45017α. Mol Endocrinol. 1991;5:1373–80.
58. Chen S, Zhou D. Functional domains of aromatase cytochrome P450 inferred from comparative analyses of amino acid sequences and substantiated by site-directed mutagenesis experiments. J Biol Chem. 1992;267:22587–94.
59. Zheng Y, i-Min, Henne KR, Charmley P, Kim RB, McCarver DG, Cabacungan ET, et al. Genotyping and site-directed mutagenesis of a cytochrome P450 meander Pro-X-Arg motif critical to CYP4B1 catalysis. Toxicol Appl Pharmcol. 2003;186:119–26.
60. IMAISHI H, MATSUMOTO Y, ISHITOBI U. OHKAWA H. Encoding of a Cytochrome P450-Dependent Lauric Acid Monooxygenase by CYP703A1 Specifically Expressed in the Floral Buds of Petunia hybrida. Bioscience, Biotechnology, and Biochemistry. 1999;63:2082–90.
61. Pancaldi F, Vlegels D, Rijken H, van Loo EN, Trindade LM. Detection and Analysis of Syntenic Quantitative Trait Loci Controlling Cell Wall Quality in Angiosperms. Front Plant Sci. 2022;13.
62. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32.
63. Tong H, Nikoloski Z. Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. J Plant Physiol. 2021;257:153354.
64. Xu Y, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, et al. Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol Plant. 2022;15:1664–95.
65. Jeon D, Kang Y, Lee S, Choi S, Sung Y, Lee T-H et al. Digitalizing breeding in plants: a new trend of next-generation breeding based on genomic prediction. Front Plant Sci. 2023;14.
66. Martina M, De Rosa V, Magon G, Acquadro A, Barchi L, Barcaccia G et al. Revitalizing agriculture: next-generation genotyping and -omics technologies enabling molecular prediction of resilient traits in the Solanaceae family. Front Plant Sci. 2024;15.
67. Magon G, De Rosa V, Martina M, Falchi R, Acquadro A, Barcaccia G et al. Boosting grapevine breeding for climate-smart viticulture: from genetic resources to predictive genomics. Front Plant Sci. 2023;14.


## Publisher's Note

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