

## ORIGINAL ARTICLE

# Variant filters using segregation information improve mapping of nectar-production genes in sunflower (*Helianthus annuus* L.)

Ashley C. Barstow<sup>1</sup>  | James P. McNellie<sup>2</sup>  | Brian C. Smart<sup>1</sup>  | Kyle G. Keepers<sup>3</sup>  |  
Jarrad R. Prasifka<sup>2</sup>  | Nolan C. Kane<sup>3</sup>  | Brent S. Hulke<sup>2</sup> 

<sup>1</sup>Department of Plant Sciences, North Dakota State University, Fargo, North Dakota, USA

<sup>2</sup>USDA-ARS Sunflower Improvement Research Unit, Edward T. Schafer Agricultural Research Center, Fargo, North Dakota, USA

<sup>3</sup>Ecology and Evolutionary Biology Department, University of Colorado, Boulder, Colorado, USA

## Correspondence

Brent S. Hulke, USDA-ARS Sunflower Improvement Research Unit, Edward T. Schafer Agricultural Research Center, Fargo, ND, USA.

Email: [brent.hulke@usda.gov](mailto:brent.hulke@usda.gov)

Assigned to Associate Editor Agnieszka Golicz.

## Funding information

National Sunflower Association, Grant/Award Number: 22-P01; Agricultural Research Service, Grant/Award Number: 3060-21000-047

## Abstract

Accurate variant calling is critical for identifying the genetic basis of complex traits, yet filters used in variant detection may inadvertently exclude valuable genetic information. In this study, we compare common sequencing depth filters, used to eliminate error-prone variants associated with repetitive regions and technical issues, with a biologically relevant filtering approach that targets expected Mendelian segregation. The resulting variant sets were evaluated in the context of nectar volume quantitative trait loci (QTL) mapping in sunflower (*Helianthus annuus* L.). Our previous research failed to detect an interval containing a strong candidate gene for nectar production (*HaCWINV2*). We removed hard filters and implemented a chi-square goodness-of-fit test to retain variants that segregate according to expected genetic ratios. We demonstrate that biologically relevant filtering retains more significant QTL and candidate genes, including *HaCWINV2*, while removing variants due to technical errors more effectively, and accounted for 48.55% of nectar production phenotypic variation. In finding nine putative homologs of *Arabidopsis* genes with nectary function within QTL regions, we demonstrate that this filtering strategy has a higher power of true variant detection in QTL mapping than the commonly used variant depth filtering strategy. Future research will adapt the technique to multiple population contexts, such as genomic selection.

## Plain Language Summary

In genomic research, identifying genetic markers is key to understanding complex traits, but traditional methods for filtering genetic data can sometimes miss important information. In this study, we explored a new data filtering approach and mapped genes related to nectar production in sunflower. We applied a more flexible filtering

**Abbreviations:** ANOVA, analysis of variance; BLAST, Basic Local Alignment Search Tool; CWINV, cell wall invertase; GATK, Genome Analysis Toolkit; HO, high oleic; LOD, logarithm of the odds; MAF, minor allele frequency; minQ, minimum quality; MQM, multiple QTL model; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; VCF, variant call file format.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *The Plant Genome* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

method that considers how markers are expected to segregate in breeding populations. Our previous work failed to identify an important gene previously hypothesized to be involved in nectar production, likely due to overly strict filtering. Our improved approach identified nine sunflower genes related to nectar production genes in the model species *Arabidopsis thaliana*, compared to zero genes identified from the previous filtering strategy. This study highlights the value of using flexible, biologically relevant filtering methods, which can lead to better results in plant genomic studies.

## 1 | INTRODUCTION

Accurate and reliable variant calling is essential for understanding the genetic underpinnings of complex traits. Genomic data possess errors that present challenges to discovering “true” variants, which often necessitate multiple data filtering strategies to remove. For example, the Genome Analysis Toolkit (GATK) includes a range of tunable parameters to remove low-quality variants and artifacts, which are typically referred to as hard filters (De Summa et al., 2017). These hard filters involve setting specific quality parameter thresholds that variants must meet to be included in downstream analysis. Quality filters exclude low-confidence variants based on metrics such as base quality, read depth, and mapping quality. Minor allele frequency (MAF) filters remove variants with low allele frequencies, which may be sequencing errors; and the maximum missingness filter excludes variants or individuals with excessive missing data. Variants with high missingness may skew results or reduce the statistical power of the analysis. Additionally, single copy filters remove variants from repetitive regions to avoid read mapping errors, which were implemented due to the highly repetitive genome of sunflower (Badouin et al., 2017). While hard filters may help reduce noise and false positives, they can also inadvertently exclude valuable genetic information, potentially affecting the power and resolution of genetic studies. This became apparent in our initial study of quantitative trait loci (QTL) linked to nectar volume in sunflower (Barstow et al., 2022). Although marginally significant QTL were identified, no clear candidate genes linked to nectar production underlie those QTL, not even candidates like *CWINV*, which has been linked to nectar production in both *Arabidopsis thaliana* and sunflower (Minami et al., 2021; Prasifka et al., 2018; Ruhlmann et al., 2010). The authors hypothesized that *HaCWINV2* would be detected due to the evidence of its involvement in nectar composition and the expression differences between sunflower lines described in Prasifka et al. (2018). We hypothesized that the limited number of variants, due to overly stringent filtering, might have prevented us from capturing key genetic regions. Removing hard filters, while retaining variants that segregate in a Mendelian

fashion, could lead to improved mapping resolution. Conversely, using erroneous variants in mapping can obscure true genetic relationships and linkage patterns, leading to genetic map inflation, inaccurate mapping of QTL, and misinterpretation of gene interactions (Buetow, 1991; Hackett & Broadfoot, 2003; Lorieux et al., 1995; Lu et al., 2002; Shields et al., 1991).

There is no “one size fits all” approach to selecting parameters and algorithms for modern genomics protocols, as organisms’ molecular genetic characteristics—such as genome size, proportion of repetitive sequences, and presence of structural variants—differ greatly. As our understanding of genomes and technology improves, variant detection methods and quality filters must adapt to the specific attributes of the sequencing data to ensure errors are removed with minimal loss of biological information. Unlike arbitrary thresholds based on depth of coverage, biologically relevant filters that focus on the Mendelian segregation of variants could offer a more meaningful approach to variant quality filtering. We hypothesize variants that segregate according to expected genetic ratios are more likely to be biologically informative. Furthermore, erroneous variants that deviate from full-sibling segregation ratios will naturally be excluded through this filtering. The key advantage is that these decisions are based on biological relevance rather than rigid technical thresholds.

Nectar production is a genetically complex trait, which is of interest as hybrid sunflower seed production is fully dependent on pollinators (Greenleaf & Kremen, 2006). Nectar is offered as a reward to increase pollinator visitation, leading to enhanced pollination and increased seed yield in stressful seed and commercial production environments (Greenleaf & Kremen, 2006; Prasifka et al., 2018; Simpson & Neff, 1983). The genetic architecture of nectar production in plants is complex, involving multiple genes that operate within different pathways including carbohydrate metabolism, sugar transport, hormonal regulation, and developmental processes (Bender et al., 2013; J.-Y. Lee et al., 2005; Reeves et al., 2012; Schmitt et al., 2018; Seo et al., 2001). Discovering loci associated with nectar production will allow breeders to more effectively make selections via marker assisted selection and improve yield by increasing pollination rates.

In this study, we propose an alternative to stringent hard filters previously used in Barstow et al. (2022) and many other studies. Single nucleotide polymorphism (SNP) marker variants were filtered only using the chi-square goodness of fit test with the expected Mendelian ratio inferred from the filial generation of the population and assuming no known selection. In this study, the terms “variants” and “SNPs” are used interchangeably, referring specifically to SNPs. Previously, a chi-square filter was used in addition to single copy filters, quality filters, minor allele frequency filters, and maximum missingness filters, which resulted in a limited number of variants (Barstow et al., 2022), and in other studies, only hard filters were used (Pogoda et al., 2021; Reinert et al., 2020). We compare the results from different filtering approaches on QTL mapping to illustrate the importance of a more nuanced curation of sequencing data.

## 2 | MATERIALS AND METHODS

### 2.1 | Experimental design

The experimental design, including selection of plant materials, controlled-environment phenotyping, in-field validation of phenotypes, and genotyping methods and materials, is reported in Barstow et al. (2022). Briefly, the parental lines used to create the mapping population had contrasting nectar volume and sucrose content (Mallinger & Prasifka, 2017). HA 434 has high oleic (HO) acid in the seed oil, and a relatively high volume of dilute, hexose-rich (glucose and fructose) nectar, while HA 456 has HO seed oil and a lower volume of nectar with an unusually high concentration of sucrose (Miller et al., 2004, 2006). HA 456 is an inbred line derived from HA 434 crossed with S-16 YU (Miller et al., 2006). The high selection pressure for the HO haplotype in the development of parent line HA 456 resulted in HA 456 inheriting the entire chromosome 14 associated with *FAD2-1* from HA 434 (Barstow et al., 2022; Schuppert et al., 2006). The mapping population was founded from a single  $F_1$  individual and underwent single seed descent until  $F_6$  seeds of 198 recombinant inbred lines were produced.

### 2.2 | Data curation

Demultiplexed data were downloaded directly from Novogene's servers (Novogene, Sacramento, CA, USA). From the demultiplexed data, three datasets were developed in this experiment. The first dataset was made with only hard filters and no additional chi-square filtering. The second dataset was the same variant set from the initial nectar volume QTL mapping study (Barstow et al., 2022) with both hard filters and chi-square filtering as described below. The third

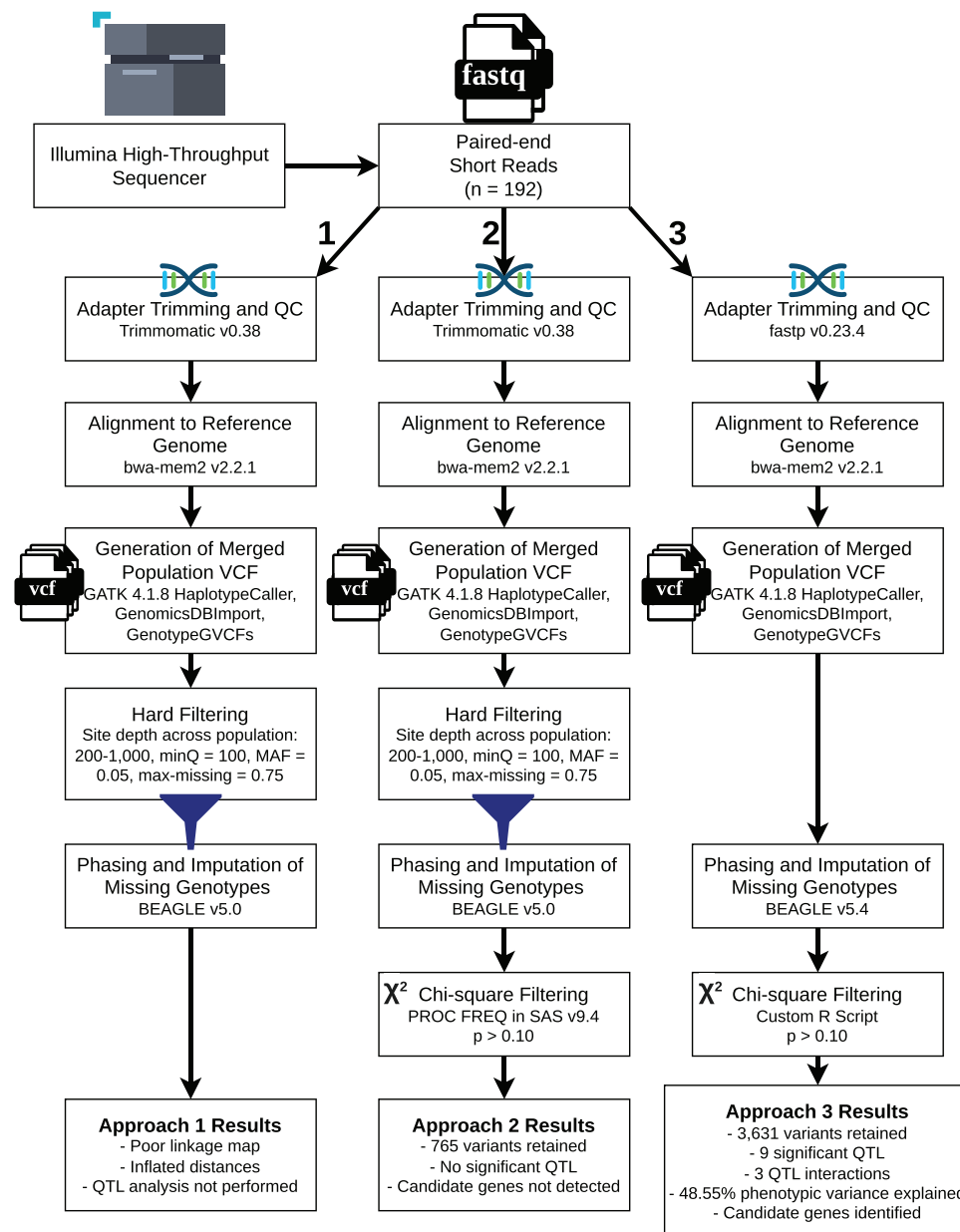
#### Core Ideas

- Discovering biologically meaningful variants from sequence data requires a careful and critical view of bioinformatic workflows.
- The use of arbitrary filters can remove significant genomic variation that contributes to the phenotype of interest.
- Arbitrary filters can also fail to remove variant call errors.
- A chi-square filtering strategy based on segregation ratio retained a larger number of valid variants.
- More candidate regions with putative nectar-related genes and better statistical support were discovered.

was the experimental dataset curated without hard filters and only the population-level, segregation ratio filter, based on a chi-square goodness of fit test.

Our first dataset was developed using the following methods. Raw genomic libraries were trimmed using Trimmomatic Version 0.38 (Bolger et al., 2014) with the code: NexteraPE-fa:2:30:10 LEADING:3 TRAILING:3 MINLEN:100, and NexteraPE-PE-fa containing the standard set of Nextera adapters to be trimmed from reads. The resulting trimmed reads were aligned to a *Helianthus annuus* chromosome-scale reference genome, HA 412 HOv2.0 (Badouin et al., 2017). Variant calling was performed using GATK best practices (Van der Auwera & O'Connor, 2020; Van der Auwera et al., 2013), resulting in a single variant call file (VCF). This VCF table was then filtered for single copy sites based on depth (sites with depth across all sunflower lines [ $n = 192$ ] between 200 and 1000 were retained). This range of depths was selected by creating a histogram of the variant depths in the VCF table, which produces a prominent, normally distributed peak at the mode single-copy depth sampling of the summed aligned libraries. Variants at depths that are too low were discarded for possibly representing sequencing errors, whereas variants at excessively high depths were discarded for potentially belonging to repetitive content, which is a major contributor of noise to a dataset. Additional filtering selected variants above a minimum quality score of 100 ( $\text{minQ} = 100$ ), minor allele frequency of 5% or greater ( $\text{MAF} \geq 0.05$ ), and no more than 75% of samples having missing data for a given variant ( $\text{max-missing} = 0.75$ ) (Figure 1). All remaining missing data were imputed using BEAGLE version 5.0 with default settings retained (Browning et al., 2018).

Our second approach was the original variant set from a previous nectar volume QTL mapping study (Barstow et al., 2022). All previous steps from the first set were followed.



**FIGURE 1** Procedural comparison of three VCF filtering approaches for sunflower genomic data analysis, with a synopsis of results. MAF, minor allele frequency; QC, quality control; QTL, quantitative trait loci; VCF, variant call format file.

Following imputation, polymorphic SNPs were then filtered using a custom script invoking PROC FREQ of SAS v. 9.4 (SAS Institute, 2016) to exclude variants that did not fit the expected  $F_6$  segregation ratio, from a chi-square goodness of fit test ( $p > 0.10$ ). This dataset was curated using both a chi-square filter and hard filters (Figure 1).

The third approach using chi-square filtering was prepared as follows. First, *fastp* was used to perform quality control and exclude adapters (Chen et al., 2018). Using BWA-mem2, FASTQ files were aligned with the most recent sunflower genome assembly, HA 412 HOv2.0 (Badouin et al., 2017; Vasimuddin et al., 2019). None of the previously described filters (minQ, MAF, max-missing, or variant-depth) were used. Missing data were imputed using BEAGLE version 5.4 with

default settings retained (Browning et al., 2018). Polymorphic SNPs were then filtered using a custom R script to exclude variants that did not fit the expected  $F_6$  segregation ratio, from a chi-square goodness of fit test ( $p > 0.10$ ) (Figure 1). The custom R script produces identical output to the PROC FREQ script used previously but had improved run time with larger variant sets.

## 2.3 | Linkage mapping and QTL analysis

For the hard filtered dataset and the chi-square filtered dataset (first and third approaches), a custom R script was used to prune additional SNPs based on distance, aiming

to reduce variant numbers to meet software requirements for QTL mapping. For each chromosome, the distance between consecutive sets of three variants was calculated and if the distance was less than 125,000 base pairs, one is randomly selected to be kept; otherwise, it keeps all variants in that range (*marker\_filt\_dist*; <https://github.com/BrianSmart/SegregationFilteringSunflower>). Additionally, for all datasets, a custom R script was used to remove co-localized variants by comparing variants at the same genetic map position and retaining the SNP with the higher *p*-value from chi-square test of Mendelian segregation (*thinning\_loop*; <https://github.com/BrianSmart/SegregationFilteringSunflower>).

The construction of linkage maps for all datasets was carried out in R/qtl package version 1.60 (Broman et al., 2003). Genetic distances were calculated using the Kosambi map function (Kosambi, 1943). Erroneous SNPs were identified and manually removed based on recombination maps and SNP frequencies, following R/qtl best practices. For the QTL analysis, the function *scanone* was employed to identify putative QTL with datasets that created a biologically feasible linkage map (Broman & Sen, 2009). Additionally, two-dimensional scans were conducted using *scantwo* with Haley–Knott regression (Broman & Sen, 2009; Haley & Knott, 1992) to explore interactions between QTL. The thresholds were established based on the results of 1000 permutations at a  $p = 0.05$  significance level (Broman et al., 2003).

Building upon the results from single QTL model and the two-dimensional genome scan, multiple QTL models (MQM) were fit to examine the presence of additional QTL and QTL-by-QTL interactions as described by Broman and Sen (2009). QTL above the permutation threshold formed the initial model, and additional model terms were discovered using the *addqtl* and *addint* functions (Broman & Sen, 2009). The resulting final MQM model incorporated all significant QTL and interactions above a logarithm of the odds (LOD) score of 3. The effects of individual QTL were evaluated by comparing the full model to one where the individual QTL was omitted from the full model. LOD scores, estimated additive effects, and the percentage of phenotypic variance explained by each QTL and QTL interaction were obtained from the drop-one analysis of variance (ANOVA) table. Using the resulting model, an additional ANOVA was used to verify results.

## 2.4 | Candidate gene analysis

Using previous knowledge of the cloned genes in *Arabidopsis* with implicated functions in nectaries and nectar (Table 2 from Roy et al., 2017), the corresponding protein sequences were queried against the reference genome HA 412 HOv2.0 (Badouin et al., 2017) using *tblastn* as implemented in

BLAST+ 2.11.0 (Altschul et al., 1990). The protein sequences were obtained from The Arabidopsis Information Resource (TAIR) website (<https://www.arabidopsis.org/>). The candidate genes were considered if found within the 2.0 LOD drop interval of the identified loci. The corresponding candidate gene nucleotide sequences were extracted, and the functional annotation was cross-checked with a TBLASTX 2.16.0+ query against the core\_nt database. Those genes that retained putative functional characterization, as determined by best result, were retained. An additional search was conducted with the most up-to-date annotation of HA 412 HOv2.0 to determine the total number of genes within each QTL region (Supporting Information).

## 3 | RESULTS

### 3.1 | Comparison of SNPs between datasets

The first dataset, filtered only with hard filters, produced 289,678 SNPs. The second dataset, being a subset of the first, resulted in 765 SNPs. Therefore, 765 SNPs were shared between the hard filtered dataset and the dataset with both filters (chi-square and hard filters). The third dataset, filtered only with the chi-square filter, consisted of 1,151,856 SNPs, 1569 of which were shared with the hard filtered dataset and 670 shared with the dataset curated with both filters (Table 1). Between all three approaches, 670 SNPs were shared.

### 3.2 | Construction of linkage maps

The hard filtered dataset had more variants than most linkage mapping software can efficiently analyze. For mapping efficiency, co-localized variants and erroneous variants detected during estimation of recombination frequencies were removed, as previously described, resulting in 7425 variants. The resulting linkage map had substantially inflated genetic distances that exceed biologically plausible lengths. The number of variants on each chromosome ranged from 362 to 538 with genetic distances ranging from 43808.3 to 65346.4 cM. Due to the relationship between the parental lines used to create the mapping population, we expected that there would be genomic regions with conserved haplotypes between the parents resulting in no variant segregation, leading to clustering of SNPs across the genome. However, this was not observed as the linkage map had no clustering; SNPs were dispersed somewhat evenly across all 17 chromosomes (Figure 2a), including chromosome 14 despite both parents having the same HO haplotype associated with the *FAD2-1* locus from high selection pressure (Miller et al., 2006; Schuppert et al., 2006). Due to the poor quality of the linkage map, further QTL analysis was not conducted on this dataset.



**TABLE 1** Single nucleotide polymorphism (SNP) number comparisons between datasets and different filtering approaches.

Datasets	Raw SNPs	Shared with first dataset	Shared with second dataset
Approach 1: Hard filters only	289,678		
Approach 2: Hard Filters + chi-square filter	765	765	
Approach 3: Chi-square only	1,151,856	1569	670

For the dataset with both filters (chi-square and hard filters), the variant set started with 765 variants spanning 16 different chromosomes, excluding chromosome 14 due to a lack of polymorphic SNPs (for reasons already described). Chromosome 6 only included one variant, which was excluded from the linkage map. The remaining number of variants ranged from two to 124 per chromosome, forming localized clusters on each due to large segments of the genomes of the two parents being identical by descent (Figure 2b). Additional erroneous variants were identified based on the recombination map and manually dropped, resulting in a linkage map containing 748 variants over 15 chromosomes. Gaps of 27.9, 25.9, 38.7, 34.0, and 26.8 cM were observed on chromosomes 3, 5, 11, 12, and 15, respectively. The average variant density by chromosome ranged from 0.3 to 19.7 variants per cM due to the low-density nature of the map.

The chi-square filtered dataset had 1,151,856 SNPs. Similarly, for mapping efficiency, co-localized and erroneous variants were removed, resulting in 3631 variants. The variants span across 16 chromosomes, excluding chromosome 14. The number of variants per chromosome ranged from 42 to 479, and the average variant density by chromosome ranged from 0.5 to 5.2 variants per cM, with an overall average of 1.4. Chromosome length ranged from 159.5 to 667.1 cM, and the expected clustering of markers was observed (Figure 2c).

### 3.3 | QTL analysis of nectar volume

Using the dataset with both filters (chi-square and hard filters), no significant loci were identified using the single-QTL model. However, the variant with the highest LOD score (2.59), although not statistically significant, mapped to chromosome 16 and was within the region identified as the largest main effect QTL in Barstow et al. (2022). With no significant QTL, MQM was not performed for the dataset using both filters.

For the chi-square filtered dataset, the QTL analysis identified nine significant QTL and three QTL interactions contributing to the trait variation in the dataset (Figure 2c). The full model, developed using Haley–Knott regression and model fitting, explained 48.55% of the phenotypic variance (LOD = 27.42,  $p$ -value < 0.001). When dropping one QTL at a time, the most significant QTL was identified on chromosome 10 at 402.9 cM (LOD = 12.64, 18.44% variance

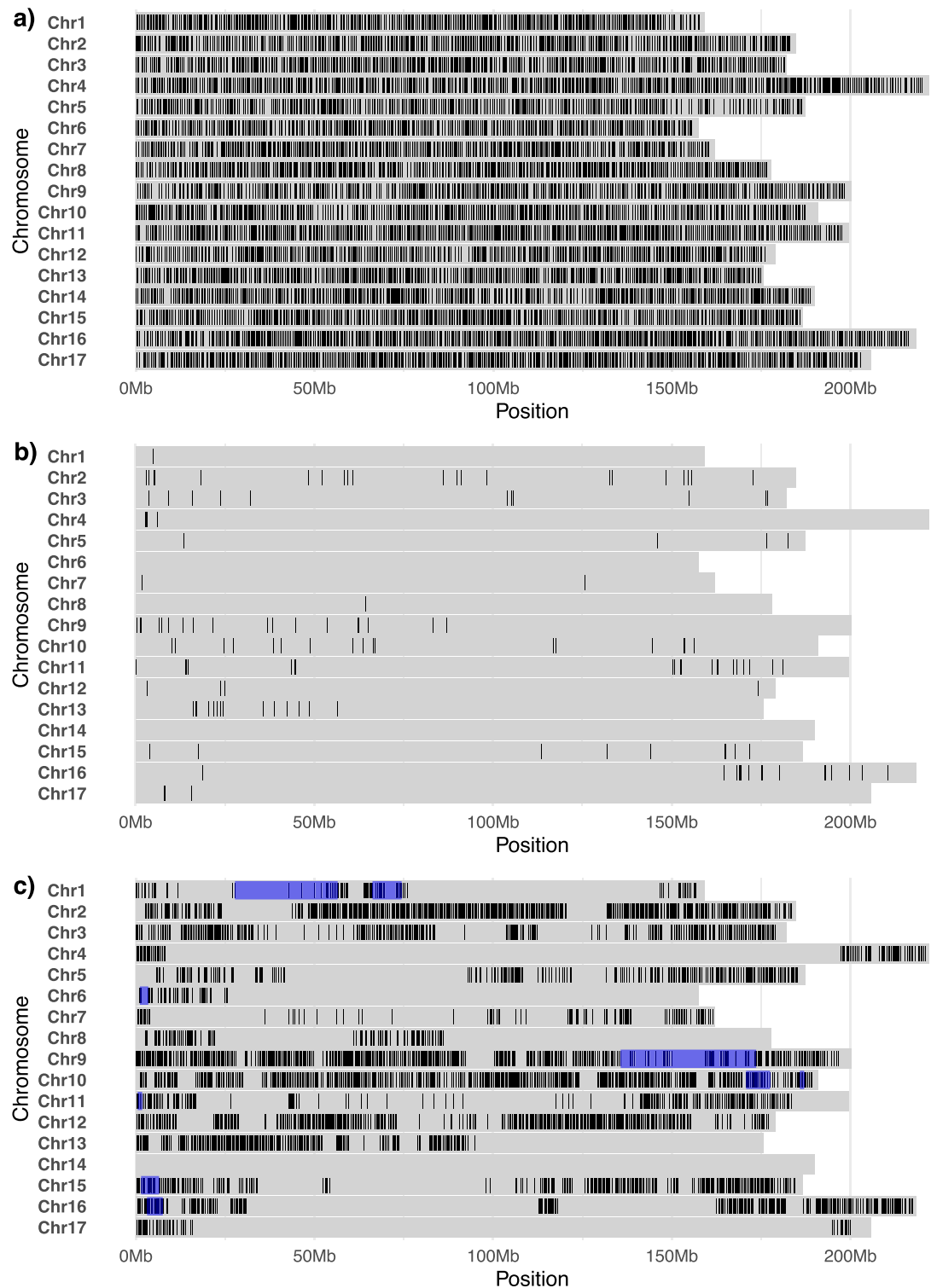
explained; Table 2). Overall, the main effect QTL identified ranged from LOD of 3.17 to 12.64. Three QTL interactions were detected: chromosome 10 at 402.9 cM with chromosome 11 at 3.2 cM (LOD = 10.47, 14.86% variance explained), chromosome 10 at 402.9 cM with chromosome 16 at 32.9 cM (LOD = 3.40, 4.42% variance explained), and chromosome 15 at 16.2 cM with chromosome 16 at 32.9 cM (LOD = 7.07, 9.61% variance explained).

The phenotypic range of nectar volume in the mapping population spans 0.03  $\mu$ L/floret to 0.91  $\mu$ L/floret in the 192 entries sampled, exhibiting similar nectar volumes to the parents, but also with values that appeared intermediate (fig. 1 from Barstow et al., 2022). The estimated additive effects of individual QTL ranged from  $-0.0277$   $\mu$ L/floret (chromosome 1 at 88.0 cM) to 0.0220  $\mu$ L/floret (chromosome 1 at 191.0 cM) (Table 2). Positive effect estimates indicate that the allele contributed by the high parent (HA 434) increased nectar volume per floret and negative estimates indicate the low parent (HA 456) contributed the allele increasing nectar volume. The interaction effects ranged from 0.0103 to 0.0208  $\mu$ L/floret (Table 2). QTL identified in this study, though modest in individual contributions, hold biological relevance when considered collectively and within the broader phenotypic range.

### 3.4 | Candidate genes

Of the nectar-related genes listed in Roy et al. (2017), nine homologous genes were found within the regions identified from this study with blast bit scores ranging from 46.2 to 418: *CWINV4*, *SWEET9*, *PIN6*, *MYB57*, *MYB21*, *CRABS CLAW*, *ARF6*, *ARF8*, *TIR1* (Table 3). Based on the blast results, the most similar sunflower gene to an *Arabidopsis* homolog was a *TIR1* homolog found within the QTL region on chromosome 15 ( $e$ -value =  $3.35\text{e-}127$ ). Notably, a *CWINV* was found within the QTL region on chromosome 1 ( $e$ -value =  $5.55\text{e-}109$ ) as previously hypothesized from Prasifka et al. (2018). Only candidate genes whose functional annotations matched the corresponding *Arabidopsis* genes were reported (Table 3).

An additional search was conducted with the most up-to-date annotation of HA 412 HOv2.0 to determine the number of genes within each QTL region (within two LOD units). The number of annotated genes within each QTL interval ranged from 29 to 839 genes (Table 4).



**FIGURE 2** Variant overview of the physical map of the mapping population, aligned to the HA 412 HOv2.0 genome. Black lines represent polymorphic markers in the genetic map, while blue regions represent quantitative trait loci (QTL) regions within a two LOD unit support interval of detected peaks. (a) Hard Filtered dataset. (b) Hard filtered and chi-square filtered dataset. (c) Chi-square filtered dataset.

## 4 | DISCUSSION

The challenges associated with variant calling and filtering—such as distinguishing true variants from sequencing artifacts

and ensuring the accuracy and reproducibility of results—highlight the need for more flexible and adaptive approaches in genomic research. We have shown through linkage mapping results that traditional hard filters inadvertently exclude

**TABLE 2** Summary of quantitative trait loci (QTL) effects on the nectar volume trait in sunflower from the dataset with only chi-square filtering.

QTL start	QTL end	QTL start	QTL end	Peak LOD	Effect estimate <sup>a</sup> (μL)
Main effects					
Chr01_27714800	Chr01_56468243			6.26	−0.0277
Chr01_66218687	Chr01_74371288			4.53	0.0220
Chr06_1099127	Chr06_3470114			4.43	−0.0121
Chr09_135674477	Chr09_173620150			3.17	0.0097
Chr10_170786952	Chr10_177468974			3.20	0.0099
Chr10_185841747	Chr10_187032379			12.64	0.0163
Chr11_487649	Chr11_1726832			11.47	0.0066
Chr15_1444681	Chr15_6460927			7.25	0.0023
Chr16_3122022	Chr16_7502195			11.37	−0.0059
Interaction effects					
Chr10_185841747	Chr10_187032379	Chr11_487649	Chr11_1726832	10.47	0.0208
Chr10_185841747	Chr10_187032379	Chr16_3122022	Chr16_7502195	3.40	0.0103
Chr15_1444681	Chr15_6460927	Chr16_3122022	Chr16_7502195	7.10	0.0161

<sup>a</sup>Positive values indicate that the high parent (HA 434) allele increases the nectar volume per floret.

Abbreviation: LOD, logarithm of the odds.

significant genetic variation that contributes to phenotypic traits. By utilizing goodness-of-fit tests based on Mendelian ratios, such as the chi-square test, rather than relying on stringent hard filters, we were able to ensure a more accurate and comprehensive representation of the genetic architecture, preserving valuable variants that may be crucial for QTL mapping and downstream analyses.

Using this approach (chi-square only), we were able to identify nine more QTL and three more interaction effects contributing to the trait of interest, nectar volume in sunflower, than the hard filtering approaches. Our approach generated over 1 million SNPs, from which we selected 3631 variants spanning across 16 chromosomes for mapping. The chi-square filter also removed unreliable variants, addressing another limitation of hard filters. Our hard filter only dataset initially retained 289,678 SNPs, yet only 765 remained after applying the chi-square filter. This suggests that hard filtering discarded informative variants and failed to eliminate erroneous ones (Table 1; Figure 2). This shift in filtering strategy offers the potential to enhance the resolution and power of genetic studies, particularly for polygenic traits such as nectar production in complex genomes like that of sunflower (3.6 gigabases and more than 75% of the genome as long terminal repeat retrotransposons) (Badouin et al., 2017). The amount of phenotypic variation explained by the model using the chi-square filtered dataset was 48.55%. The  $R^2$  value achieved in this study is promising, particularly for a highly quantitative trait, as it suggests that a significant portion of the phenotypic variance can be explained by the genetic variants identified

by our approach. The overall fit of the model used was highly statistically significant ( $p < 0.001$ ), while the hard filtered dataset failed to identify any loci with this mapping algorithm above the permutation thresholds we set. Additional variants could have been included post hoc to increase resolution further in regions of interest.

This approach not only improved map resolution but also provided a strong foundation for identifying candidate genes associated with nectar production. Building on the previous work done in the model species *Arabidopsis*, candidate gene homologs of *CWINV4*, *SWEET9*, *PIN6*, *MYB21*, *MYB57*, *CRABS CLAW*, *ARF6*, *ARF8*, and *TIR1* were located within QTL regions (Table 3). Previous studies show that the genetic control of nectar secretion in *Arabidopsis* and *Brassica rapa* is complex, but a key player is cell wall invertase, the enzyme known to catalyze the hydrolysis of sucrose into glucose and fructose (Minami et al., 2021; Ruhlmann et al., 2010). This study was able to identify *HaCWINV2* (the closest homolog to *CWINV4* in the sunflower genome) as a contributor to nectar production, on chromosome 1. The proposed model in Roy et al. (2017) of floral nectary regulation suggests that this process relies on a specific cascade of expression events which ultimately lead to nectar production. For all the genes involved in this proposed pathway (fig. 3 in Roy et al., 2017), we were able to find homologs in this study within two LOD units of our identified QTLs, with several regions containing two homologs and two regions containing none from the Roy et al. model. Based on the number of genes within each QTL, strong candidate genes may not be limited



**TABLE 3** Homologs of *Arabidopsis thaliana* nectar production genes within sunflower nectar production quantitative trait locus (QTL) support intervals.

Candidate gene ( <i>Arabidopsis thaliana</i> )	Function and/or mutant phenotype	Sunflower QTL						
		Chromosome	Start location	End Location	Locus tag HanXRQr2	Gene ID	E-value	Bitscore
ARF6	Transcription factor; mutants lack nectaries (Reeves et al., 2012)	Chr09	146,751,503	146,751,357	Chr13g0580481	LOC110900998	2.22E-05	54.7
		Chr11	1,078,087	1,077,932	Chr01g0001331	LOC110942398	1.12E-16	92
ARF8	Transcription factor; mutants lack nectaries (Reeves et al., 2012)	Chr09	149,615,897	149,615,769	Chr13g0580481	LOC110900998	9.00E-08	55.8
		Chr11	1,078,087	1,077,932	Chr01g0001331	LOC110942398	8.00E-17	92.4
CRABS CLAW	Transcription factor needed for nectary development (Bowman & Smyth, 1999)	Chr06	1,332,589	1,332,362	Chr06g0238981	LOC110864276	2.45E-08	58.5
CWINV4	Cell wall invertase responsible for nectar production; mutants do not produce nectar (Ruhlmann et al., 2010)	Chr01	68,660,146	68,659,286	Chr01g0013971	LOC110868021	5.55E-109	365
MYB21	Transcription factor; mutants lack nectar and expression of SWEET9 (Schmitt et al., 2018)	Chr09	153,815,731	153,815,432	Chr09g0397101	LOC110915207	4.70E-22	99.8
		Chr09	146,202,208	146,201,897	Chr09g0394161	LOC110878835	3.27E-21	96.7
MYB57	Transcription factor; mutants have smaller nectaries and produce about half the nectar of wild-type (Bender et al., 2013)	Chr06	1,517,950	1,518,204	Chr06g0239071	LOC110864269	1.48E-25	109
		Chr10	175,680,709	175,680,380	-	LOC113741846	1.98E-21	96.7
PIN6	Transmembrane protein involved in auxin transport; mutants produced 60% less nectar (Bender et al., 2013)	Chr01	54,495,305	54,495,153	Chr01g0010911	LOC110866724	2.66E-17	63.5
		Chr09	136,142,804	136,142,373	Chr09g0390771	LOC110874663	1.07E-32	140

(Continues)

TABLE 3 (Continued)

Candidate gene ( <i>Arabidopsis thaliana</i> )	Function and/or mutant phenotype	Sunflower QTL						
		Chromosome	Start location	End Location	Locus tag HanXRQr2	Gene ID	E-value	Bitscore
<i>SWEET9</i>	Sucrose uniporter; mutants do not produce nectar (Lin et al., 2014)	Chr09	170,112,792	170,113,016	Chr09g0402851	LOC110874867	5.84E-04	46.2
		Chr15	4,825,680	4,825,889	Chr15g0673901	LOC110910200	1.04E-14	79.3
<i>TIR1</i>	Auxin receptor that mediates auxin-regulated transcription (Schmitt et al., 2018)	Chr15	4,842,038	4,841,253	Chr15g0673921	LOC110910197	3.35e-127	418

Abbreviation: QTL, quantitative trait loci.

TABLE 4 Additional candidate genes discovered within a quantitative trait locus (QTL) two logarithm of the odds (LOD) unit interval on the HA 412 HOv2.0 sunflower genome.

Chromosome	Start location	End location	Length (bp)	No. of genes
Chr01	27,714,800	56,468,243	28,753,443	277
Chr01	66,218,687	74,371,288	8,152,601	102
Chr06	1,099,127	3,470,114	2,370,987	59
Chr09	135,674,477	173,620,150	37,945,673	839
Chr10	170,786,952	177,468,974	6,682,022	165
Chr10	185,841,747	187,032,379	1,190,632	29
Chr11	487,649	1,726,832	1,239,183	49
Chr15	1,444,681	6,460,927	5,016,246	202
Chr16	3,122,022	7,502,195	4,380,173	121

to these (Table 4; Supporting Information). Within the loci with no candidate genes from Roy et al., there were candidate genes with support from other studies. Specifically, within the chromosome 10 (185,841,747–187,032,379 bp) confidence interval: xyloglucan endotransglucosylase/hydrolase, heat shock protein 70, leucine-rich repeat (LRR) protein kinase family protein, glycosyl hydrolases family proteins, zinc finger protein, and Golgi nucleotide sugar transporter (Ballerini et al., 2019; Bowman & Smyth, 1999; Silva et al., 2020). Notable candidate genes found within the chromosome 16 (3,122,022–7,502,195 bp) confidence interval included glycosyl hydrolase family proteins, LRR receptor-like serine/threonine-kinase, H[+]-ATPase, zinc finger proteins, transmembrane proteins, and plasma membrane fusion protein (Ballerini et al., 2019; Bowman & Smyth, 1999; Silva et al., 2020).

The utility of this method is currently limited to biparental recombinant inbred line datasets with no selection assumed, in order to properly estimate expected allele frequencies. To

extend the method to plant breeding datasets and diversity panels, we need to adapt this or similar methods to populations that do not fit these basic assumptions. This study is currently underway in the context of breeding populations with a history of artificial selection for two heterotic groups in hybrid oilseed sunflower.

## 5 | CONCLUSION

In this study, three genomic data filtering strategies were compared using nectar-production QTL results in sunflower. Our study revealed that using a chi-square goodness-of-fit test based on the expected population-level segregation ratio outperformed a standard data curation strategy utilizing hard filters in both retaining valid variants and removing errant variants. Using the chi-square filtered data without hard filters, a QTL model was developed that was able to explain over 48% of the phenotypic variation in sunflower nectar

production, while the hard filtered data lacked the ability to identify QTL over the permutation threshold. Nine significant putative QTL and three QTL interactions contributing to the trait variation were found. Additionally, within the identified QTL regions, many genes homologous to previously identified nectar-related genes in other species were found including homologs of *CWINV4*, *SWEET9*, *PIN6*, *MYB21*, *MYB57*, *CRABS CLAW*, *ARF6*, *ARF8*, and *TIR1*. Our results provided us with evidence to reconsider implementation of hard filters in sequencing data curation for genomic studies, and carefully consider biologically relevant alternative methods.

## AUTHOR CONTRIBUTIONS

**Ashley C. Barstow**: Conceptualization; data curation; formal analysis; investigation; methodology; visualization; writing—original draft; writing—review and editing. **James P. McNellie**: Data curation; formal analysis; investigation; methodology; writing—review and editing. **Brian C. Smart**: Data curation; investigation; methodology; visualization; writing—review and editing. **Kyle G. Keepers**: Data curation; investigation; methodology; writing—review and editing. **Jarrad R. Prasifka**: Formal analysis; funding acquisition; investigation; writing—review and editing. **Nolan C. Kane**: Funding acquisition; investigation; writing—review and editing. **Brent S. Hulke**: Conceptualization; funding acquisition; investigation; project administration; writing—review and editing.

## ACKNOWLEDGMENTS

The authors wish to thank Brady Koehler for assisting with plant tissue capture, Dr. Ziv Attia for assisting with sequencing efforts, and numerous undergraduate students who assisted with the development of the mapping population. We appreciate the kind suggestions of Dr. Richard Horsley and Dr. Ana Heilman-Morales, who provided useful feedback at the design stage of this study. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.



## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Data, NCBI BioProject identifiers, and code described in the methods are available in the National Agricultural Library—Ag Data Commons, DOI: 10.15482/USDA.ADC/c.7791377. Code is available from GitHub at <https://github.com/BrianSmart/SegregationFilteringSunflower>.

## ORCID

Ashley C. Barstow  <https://orcid.org/0009-0000-5147-5055>  
James P. McNellie  <https://orcid.org/0000-0001-8067-511X>  
Brian C. Smart  <https://orcid.org/0000-0002-3677-0571>  
Kyle G. Keepers  <https://orcid.org/0000-0002-2288-8018>  
Jarrad R. Prasifka  <https://orcid.org/0000-0002-6165-7319>  
Nolan C. Kane  <https://orcid.org/0000-0001-9133-6543>  
Brent S. Hulke  <https://orcid.org/0000-0001-5380-0827>

## REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Brière, C., Owens, G. L., Carrère, S., Mayjonade, B., Legrand, L., Gill, N., Kane, N. C., Bowers, J. E., Hubner, S., Bellec, A., Bérard, A., Bergès, H., Blanchet, N., ... Langlade, N. B. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, 546(7656), 148–152. <https://doi.org/10.1038/nature22380>
- Ballerini, E. S., Kramer, E. M., & Hodges, S. A. (2019). Comparative transcriptomics of early petal development across four diverse species of Aquilegia reveal few genes consistently associated with nectar spur development. *BMC Genomics*, 20(1), 668. <https://doi.org/10.1186/s12864-019-6002-9>
- Barstow, A. C., Prasifka, J. R., Attia, Z., Kane, N. C., & Hulke, B. S. (2022). Genetic mapping of a pollinator preference trait: Nectar volume in sunflower (*Helianthus annuus* L.). *Frontiers in Plant Science*, 13, 1056278. <https://doi.org/10.3389/fpls.2022.1056278>
- Bender, R. L., Fekete, M. L., Klinkenberg, P. M., Hampton, M., Bauer, B., Malecha, M., Lindgren, K., Maki, J. A., Perera, M. A. D. N., Nikolau, B. J., & Carter, C. J. (2013). *PIN6* is required for nectary auxin response and short stamen development. *The Plant Journal*, 74(6), 893–904. <https://doi.org/10.1111/tpj.12184>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Bowman, J. L., & Smyth, D. R. (1999). *CRABS CLAW*, a gene that regulates carpel and nectary development in *Arabidopsis*, encodes a novel protein with zinc finger and helix-loop-helix domains. *Development*, 126(11), 2387–2396. <https://doi.org/10.1242/dev.126.11.2387>
- Broman, K. W., & Sen, S. (2009). *A guide to QTL mapping with R/qlt*. Springer. <https://doi.org/10.1007/978-0-387-92125-9>
- Broman, K. W., Wu, H., Sen, S., & Churchill, G. A. (2003). R/qlt: QTL mapping in experimental crosses. *Bioinformatics*, 19(7), 889–890. <https://doi.org/10.1093/bioinformatics/btg112>
- Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3), 338–348. <https://doi.org/10.1016/j.ajhg.2018.07.015>
- Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *American Journal of Human Genetics*, 49(5), 985–994.
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>

- De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V., & Tommasi, S. (2017). GATK hard filtering: Tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics*, 18(Suppl 5), 119. <https://doi.org/10.1186/s12859-017-1537-8>
- Greenleaf, S. S., & Kremen, C. (2006). Wild bees enhance honey bees' pollination of hybrid sunflower. *Proceedings of the National Academy of Sciences*, 103(37), 13890–13895. <https://doi.org/10.1073/pnas.0600929103>
- Hackett, C. A., & Broadfoot, L. B. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity*, 90(1), 33–38. <https://doi.org/10.1038/sj.hdy.6800173>
- Haley, C. S., & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4), 315–324. <https://doi.org/10.1038/hdy.1992.131>
- Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Annals of Eugenics*, 12(1), 172–175. <https://doi.org/10.1111/j.1469-1809.1943.tb02321.x>
- Lee, J.-Y., Baum, S. F., Oh, S.-H., Jiang, C.-Z., Chen, J.-C., & Bowman, J. L. (2005). Recruitment of CRABS CLAW to promote nectary development within the eudicot clade. *Development*, 132(22), 5021–5032. <https://doi.org/10.1242/dev.02067>
- Lin, I. W., Sosso, D., Chen, L.-Q., Gase, K., Kim, S.-G., Kessler, D., Klinkenberg, P. M., Gorder, M. K., Hou, B.-H., Qu, X.-Q., Carter, C. J., Baldwin, I. T., & Frommer, W. B. (2014). Nectar secretion requires sucrose phosphate synthases and the sugar transporter SWEET9. *Nature*, 508(7497), 546–549. <https://doi.org/10.1038/nature13082>
- Lorieux, M., Perrier, X., Goffinet, B., Lanaud, C., & De León, D. G. (1995). Maximum-likelihood models for mapping genetic markers showing segregation distortion. 2. F2 populations. *Theoretical and Applied Genetics*, 90(1), 81–89. <https://doi.org/10.1007/BF00220999>
- Lu, H., Romero-Severson, J., & Bernardo, R. (2002). Chromosomal regions associated with segregation distortion in maize. *Theoretical and Applied Genetics*, 105(4), 622–628. <https://doi.org/10.1007/s00122-002-0970-9>
- Mallinger, R. E., & Prasifka, J. R. (2017). Bee visitation rates to cultivated sunflowers increase with the amount and accessibility of nectar sugars. *Journal of Applied Entomology*, 141(7), 561–573. <https://doi.org/10.1111/jen.12375>
- Miller, J. F., Gulya, T. J., & Vick, B. A. (2004). Registration of two maintainer (HA 434 and HA 435) and three restorer (RHA 436 to RHA 438) high oleic oilseed sunflower germplasms. *Crop Science*, 44(3), 1034–1035. <https://doi.org/10.2135/cropsci2004.1034>
- Miller, J. F., Gulya, T. J., & Vick, B. A. (2006). Registration of three maintainer (HA 456, HA 457, and HA 412 HO) high-oleic oilseed sunflower germplasms. *Crop Science*, 46(6), 2728–2728. <https://doi.org/10.2135/cropsci2006.06.0437>
- Minami, A., Kang, X., & Carter, C. J. (2021). A cell wall invertase controls nectar volume and sugar composition. *The Plant Journal*, 107(4), 1016–1028. <https://doi.org/10.1111/tpj.15357>
- Pogoda, C. S., Reinert, S., Talukder, Z. I., Attia, Z., Collier-zans, E. C. E., Gulya, T. J., Kane, N. C., & Hulke, B. S. (2021). Genetic loci underlying quantitative resistance to necrotrophic pathogens *Sclerotinia* and *Diaporthe* (Phomopsis), and correlated resistance to both pathogens. *Theoretical and Applied Genetics*, 134(1), 249–259. <https://doi.org/10.1007/s00122-020-03694-x>
- Prasifka, J. R., Mallinger, R. E., Portlas, Z. M., Hulke, B. S., Fugate, K. K., Paradis, T., Hampton, M. E., & Carter, C. J. (2018). Using nectar-related traits to enhance crop-pollinator interactions. *Frontiers in Plant Science*, 9, 812. <https://doi.org/10.3389/fpls.2018.00812>
- Reeves, P. H., Ellis, C. M., Ploense, S. E., Wu, M.-F., Yadav, V., Tholl, D., Chételat, A., Haupt, I., Kennerley, B. J., Hodgins, C., Farmer, E. E., Nagpal, P., & Reed, J. W. (2012). A regulatory network for coordinated flower maturation. *PLoS Genetics*, 8(2), e1002506. <https://doi.org/10.1371/journal.pgen.1002506>
- Reinert, S., Gao, Q., Ferguson, B., Portlas, Z. M., Prasifka, J. R., & Hulke, B. S. (2020). Seed and floret size parameters of sunflower are determined by partially overlapping sets of quantitative trait loci with epistatic interactions. *Molecular Genetics and Genomics*, 295(1), 143–154. <https://doi.org/10.1007/s00438-019-01610-7>
- Roy, R., Schmitt, A. J., Thomas, J. B., & Carter, C. J. (2017). Review: Nectar biology: From molecules to ecosystems. *Plant Science*, 262, 148–164. <https://doi.org/10.1016/j.plantsci.2017.04.012>
- Ruhlmann, J. M., Kram, B. W., & Carter, C. J. (2010). CELL WALL INVERTASE 4 is required for nectar production in *Arabidopsis*. *Journal of Experimental Botany*, 61(2), 395–404. <https://doi.org/10.1093/jxb/erp309>
- SAS Institute. (2016). *The SAS system for Windows* (version 9.4). SAS Institute.
- Schmitt, A. J., Roy, R., Klinkenberg, P. M., Jia, M., & Carter, C. J. (2018). The octadecanoid pathway, but not COI1, is required for nectar secretion in *Arabidopsis thaliana*. *Frontiers in Plant Science*, 9, 1060. <https://doi.org/10.3389/fpls.2018.01060>
- Schuppert, G. F., Tang, S., Slabaugh, M. B., & Knapp, S. J. (2006). The sunflower high-oleic mutant Ol carries variable tandem repeats of FAD2-1, a seed-specific oleoyl-phosphatidyl choline desaturase. *Molecular Breeding*, 17(3), 241–256. <https://doi.org/10.1007/s11032-005-5680-y>
- Seo, H. S., Song, J. T., Cheong, J.-J., Lee, Y.-H., Lee, Y.-W., Hwang, I., Lee, J. S., & Choi, Y. D. (2001). Jasmonic acid carboxyl methyltransferase: A key enzyme for jasmonate-regulated plant responses. *Proceedings of the National Academy of Sciences*, 98(8), 4788–4793. <https://doi.org/10.1073/pnas.081557298>
- Shields, D. C., Collins, A., Buetow, K. H., & Morton, N. E. (1991). Error filtration, interference, and the human linkage map. *Proceedings of the National Academy of Sciences*, 88(15), 6501–6505. <https://doi.org/10.1073/pnas.88.15.6501>
- Silva, F. A., Guirgis, A., Von Aderkas, P., Borchers, C. H., & Thornburg, R. (2020). LC-MS/MS based comparative proteomics of floral nectars reveal different mechanisms involved in floral defense of *Nicotiana* spp., *Petunia hybrida* and *Datura stramonium*. *Journal of Proteomics*, 213, 103618. <https://doi.org/10.1016/j.jprot.2019.103618>
- Simpson, B. B., & Neff, J. L. (1983). Evolution and diversity of floral rewards. *Handbook of experimental pollination biology* (pp. 142–159). Van Nostrand Reinhold Company Inc.
- Van der Auwera, G., & O'Connor, B. D. (2020). *Genomics in the cloud: Using Docker, GATK, and WDL in Terra* (1st ed.). O'Reilly.
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best Practices pipeline. *Current Protocols in Bioinformatics*, 43(1), 11.10.1–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (pp. 314–324). IEEE. <https://doi.org/10.1109/IPDPS.2019.00041>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Barstow, A. C., McNellie, J. P., Smart, B. C., Keepers, K. G., Prasifka, J. R., Kane, N. C., & Hulke, B. S. (2025). Variant filters using segregation information improve mapping of nectar-production genes in sunflower (*Helianthus annuus* L.). *The Plant Genome*, 18, e70042. <https://doi.org/10.1002/tpg2.70042>