

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

Ashley C. Barstow¹, James P. McNellie², Brian C. Smart¹, Kyle G. Keepers³, Jarrad R. Prasifka², Nolan C. Kane³, and Brent S. Hulke²

¹Department of Plant Sciences, North Dakota State University, Fargo, 58102, ND, United States

²Sunflower Improvement Research Unit, Edward T. Schafer Agricultural Research Center, United States Department of Agriculture (USDA)-Agricultural Research Service, Fargo, 58102, ND, United States

³Ecology and Evolutionary Biology Department, University of Colorado, Boulder, 80309, CO United States

Correspondence

Brent Hulke, Sunflower Improvement Research Unit, Edward T. Schafer Agricultural Research Center, United States Department of Agriculture (USDA)-Agricultural Research Service, Fargo, ND, United States

Email: brent.hulke@usda.gov

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.**

ABSTRACT

Accurate variant calling is critical for identifying the genetic basis of complex traits, yet filters used in variant detection and validation 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 population-level Mendelian segregation. The resulting variant sets were evaluated in the context of nectar volume QTL mapping in sunflower (*Helianthus annuus* L.). Our previous research failed to detect a significant interval containing a strong candidate gene for nectar production (*HaCWINV2*). We removed certain hard filters and implemented a Chi-square goodness-of-fit test to retain variants that segregate according to expected genetic ratios. We hypothesized that this will enhance mapping resolution and capture key genetic regions previously missed. 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 phenotypic variation. In finding nine putative homologs of *Arabidopsis* genes with nectary function within QTL regions, we demonstrate that this filtering strategy, which considers biological contexts, has a higher power of true variant detection than the commonly used variant depth filtering strategy.

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 for mapping genes related to nectar production in sunflower. We applied a more flexible filtering 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 fil-

tering. Our improved approach identified nine sunflower genes related to nectar production genes in the model species *Arabidopsis thaliana*, as 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.

ABBREVIATIONS

ANOVA	Analysis of variance
BLAST	Basic Local Alignment Search Tool
cM	centimorgans
CWINV	cell wall invertase
GATK	Genome Analysis Toolkit
HA	<i>Helianthus annuus</i>
HO	high oleic
IBD	identity-by-descent
LOD	logarithm of the odds
minQ	minimum quality
maf	minor allele frequency
MQM	multiple QTL model
QTL	quantitative trait loci
SNP	single nucleotide polymorphism
vcf	variant call file format

1 INTRODUCTION

Accurate and reliable variant calling is essential for understanding the genetic underpinnings of complex traits. Genomic data possesses 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. While hard filters 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 significant QTL were identified, a sunflower homolog (cell wall invertase, *HaCWINV2*) of a well-studied *Arabidopsis thaliana* gene, *CWINV4*, both with evidence of involvement in nectar production, was not detected (Ruhlmann et al., 2010; Prasifka et al., 2018; Minami et al., 2021). We hypothesized that the limited number of variants, due to overly stringent filtering, might have prevented us from capturing key genetic regions. Relaxing 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 (Shields et al, 1991; Buetow, 1991; Lorieux et al., 1995; Lu et al., 2002; Hackett & Broadfoot, 2003).

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 under-

standing 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 that variants that segregate according to expected genetic ratios are more likely to be biologically informative. Furthermore, erroneous variants that deviate from population-level 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 complex trait of interest as hybrid sunflower seed production is fully dependent on pollinators (Greenleaf and 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 (Simpson and Neff, 1983; Greenleaf and Kremen, 2006; Prasifka et al., 2018). 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 (Reeves et al., 2012; Lee et al., 2005; Seo et al., 2001; Bender et al., 2013; Schmitt et al., 2018). Discovering loci associated with nectar production will allow breeders to more effectively make selections via variant assisted selection and improve yield by increasing pollination rates.

In this paper, we propose an alternative to stringent hard filters previously used in Barstow et al. (2022). Single nucleotide polymorphic (SNP) variants were filtered using the Chi-square goodness of fit test with the expected Mendelian ratio assumed from the filial generation of the population and no known selection. Previously, a Chi-square filter was used in addition to

single copy filters, quality filters, minor allele frequency filters, and maximum missingness filters, which results in a limited number of variants (Reinert et al., 2019; Pogoda et al., 2020; Barstow et al., 2022). We compare the results from different filtering approaches on QTL mapping results 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 are reported in Barstow et al. (2022). Briefly, the parental lines used to create the mapping population had contrasting nectar volume and sucrose content (Mallinger and 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; Miller et al., 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 acid haplotype in the development of parent lines HA 434 and HA 456 resulted in HA 456 inheriting the haplotype associated with *FAD2-1* from HA 434 (Schuppert et al., 2006). The mapping population was founded from a single F₁ individual and underwent single seed descent until F₆ seeds of 198 recombinant inbred lines were produced.

2.2 Data Curation

Demultiplexed data was downloaded directly from Novogene's servers (Novogene, Sacramento, CA). 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 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 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 hard filters. Raw genomic libraries were trimmed using Trimmomatic Version 0.38 (Bolger et al., 2014) using the code: NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 MINLEN:100, with 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 et al., 2013; Van der Auwera & O'Connor, 2020), 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 200-1000 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 was done to select variants above a minimum quality score of 100 (minQ=100), minor allele frequency of 0.05 or greater (maf=0.05) and no more than 25% of samples having missing data for a given variant (max-missing=0.75). 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. 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 analysis goodness of fit test ($p > 0.10$). This dataset was curated using both a Chi-square filter and hard filters.

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 HO v2.0 (Vasimuddin et al., 2019; Badouin et al., 2017). 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 ($p > 0.10$). 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 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*; Supplemental Material). Additionally, for all datasets, a custom R script was used to remove redundant 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*; Supplemental Material).

The construction of linkage maps for all datasets were carried out in R/qlt package version 1.60 (Broman et al., 2003). Genetic distances were calculated using the Kosambi map function (Kosambi, 1944). Erroneous SNPs were identified and manually removed based on recombination maps and SNP frequencies. For the QTL analysis, the function *scanone* was employed to identify putative QTL with datasets that created a biologically feasible linkage map. Additionally, two-dimensional scans were conducted using *scantwo* with Haley-Knott regression (Haley and Knott, 1992) to explore interactions between QTL. The thresholds were established based on the results of 1,000 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. The resulting final MQM model incorporated all significant QTL and interactions above a LOD score of 3. The effects of individual QTL were evaluated by comparing the full model to one where the respective term was omitted. 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 nectars (Table 2 from Roy et al., 2017), the corresponding protein sequences were queried against the reference genome HA 412 HO v2.0 (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 HO v2.0 (Kyle Keepers, unpublished data) to determine the total number of genes within each QTL region.

3 RESULTS

3.1 Construction of Linkage Maps

The hard filtered dataset had 289,678 variants, which is beyond the capacity for most linkage mapping software to efficiently analyze. For mapping efficiency, redundant and erroneous variants 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 centimorgans (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 and no variant segregation, leading to clustering of SNPs across the genome. However, this was not observed as the linkage map had limited clustering; SNPs were dispersed across all 17 chromosomes, 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).

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. 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, redundant 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 cM to 667.1 cM.

3.2 QTL analysis of nectar volume

QTL analysis was not performed on the hard filtered dataset because of the poor quality of the linkage map.

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 iden-

tified as the largest main effect QTL in Barstow et al., 2022. With no significant QTL, MQM was not performed for the dataset using hard 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. 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 1). 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 florets spans 0.03 μ l/floret to 0.91 μ l/floret in the 192 entries sampled, exhibiting similar nectar volumes to the parents, but also with values that appeared intermediate (Figure 1 from Barstow et al., 2022). The estimated additive effects of individual QTL ranged from -0.0277 μ l/floret (chromosome 1 at 88.0 cM) to 0.0220 μ l/floret (chromosome 1 at 191.0 cM). Positive effect estimates indicate that the allele contributed by the high parent (HA 434) increased nectar volume per floret and negative effects meaning the low parent (HA456) contributed the allele increasing nectar volume. The interaction effects ranged from 0.0103 to 0.0208 μ l/floret. QTL identified in this study, though modest in individual contributions, holds biological relevance when considered collectively and within the broader phenotypic range.

3.3 Candidate Genes

Of the nectar-related genes listed in Roy et al. (2017), 9 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 2). 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 = 3.35e-127). Notably, a *CWINV* was found within the QTL region on chromosome 1 (e-value = 5.55e-109) as previously hypothesized from Prasifka et al. (2018). Only candidate genes whose functional annotations matched the corresponding *Arabidopsis* genes were reported (Table 2).

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 2 LOD units). The number of annotated genes within each QTL interval ranged from 29 to 839 genes (Table 3).

4 DISCUSSION

The challenges associated with variant calling and filtering highlight the need for more flexible and adaptive approaches in genomic research. We have shown that traditional hard filters inadvertently exclude 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. Conversely, the Chi-square filter helped ensure unreliable variants were filtered out, which did not occur completely in our dataset with standard hard filters. This shift in filtering strategy offers the potential to enhance the resolution and power of

genetic studies, particularly in complex genomes like that of sunflower and polygenic traits such as nectar production.

Using arbitrary thresholds in hard filters removed significant genomic variation that contributed to the trait of interest: nectar volume in sunflower. The use of MAF, quality scores, and a local sequencing depth filter in addition to a Chi-square filter of the segregation ratio of that population, retained a limited number of genetic variants, 746 SNPs over 15 chromosomes. While in Barstow et al., 2022, significant QTLs were identified using that variant set and a different mapping algorithm, using the mapping algorithms and methods described herein, we were unable to identify significant regions. However, after removing hard filters and considering only the Mendelian segregation patterns of that generation to filter the variants, we generated over 1 million SNPs from which we selected 3631 nonredundant, valid variants spanning across 16 chromosomes for mapping. This approach facilitated the retention of a greater number of variants across the genome, effectively increasing representation in regions that were not covered previously (Figure 2). The improved genome coverage increases the likelihood of capturing loci linked to the phenotype. With those improvements in the genetic map, we were able to capture more candidate regions associated with nectar production with better statistical support than our previous study. 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 like nectar production, as it suggests that a significant portion of the phenotypic variance can be explained by the identified genetic variants despite the complex, polygenic nature of the trait. The overall fit of the model used was highly statistically significant ($p < 0.001$) and able to identify nine QTL regions and three epistatic interactions while the hard filtered dataset failed to identify any loci with this mapping algorithm above the permutation

thresholds. This approach not only improved the genetic map's resolution but also provided a strong foundation for identifying candidate genes associated with nectar production. Additional variants could have been included post-hoc to increase resolution further in regions of interest.

Building on the increased genome coverage and the enhanced statistical power of our re-fined genetic map, we were able to move forward with candidate gene discovery, focusing on pinpointing specific genes within the identified QTL regions that may be involved in 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 2). 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 (Ruhlmann et al., 2010; Minami et al., 2021). 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. Our results support that given the additive and interactive QTL, nectar production in *H. annuus* is polygenic, with some strong candidate genes beneath QTL with large effects (Table 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 (Figure 3 in Roy et al., 2017), we were able to find homologs in this study within 2 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 to these (Table 3; Supplemental Material). 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 bp - 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 (Bowman & Smyth, 1999; Ballerini et al., 2019; Silva et al., 2020). Notable candidate genes found within chromosome 16 (3,122,022 bp -7,502,195 bp) confidence interval included glycosyl hydrolases family proteins, LRR receptor-like serine/threonine-kinase, H[+]-ATPase, zinc finger proteins, transmembrane proteins, and plasma membrane fusion protein (Bowman & Smyth, 1999; Ballerini et al., 2019; 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. However, 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 work 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 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 carefully reconsider implementation of hard filters in sequencing data curation for genomic studies.

SUPPLEMENTAL MATERIAL STATEMENT

One table with the putative functions of all genes known in the QTL LOD support intervals, based on the HA 412 HOv.2 reference genome.

DATA AVAILABILITY STATEMENT

Data and code described in the methods will be made available in the National Agricultural Library – Ag Data Commons after acceptance, and the DOI provided here.

AUTHOR CONTRIBUTIONS

Conceptualization, BSH and ACB; Data curation, ACB, BCS, JPM, KGK; Formal analysis, ACB, JPM, JRP; Funding acquisition, BSH, NCK, JRP; Investigation, all authors; Methodology, BCS, ACB, JPM, KGK; Project administration, BSH; Visualization, BCS and ACB; Writing – original draft, ACB; Writing – review & editing, all authors.

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 that assisted with development of the mapping population.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Ashley Barstow: [0009-0000-5147-5055](https://orcid.org/0009-0000-5147-5055)

James McNellie: 0000-0001-8067-511X

Brian Smart: 0000-0002-3677-0571

Kyle Keepers: 0000-0002-2288-8018

Jarrad Prasifka: 0000-0002-6165-7319

Nolan Kane: 0000-0001-9133-6543

Brent Hulke: [0000-0001-5380-0827](https://orcid.org/0000-0001-5380-0827)

REFERENCES

- Auwerda, G. V. der, & O'Connor, B. D. (2020). *Genomics in the cloud: Using Docker, GATK, and WDL in Terra* (First edition). O'Reilly.
- Auwerda, 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). <https://doi.org/10.1002/0471250953.bi1110s43>
- 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., A. Maki, J., Perera, M. A. D. N., Nikolau, B. J., & Carter, C. J. (2013). *PIN6* is re-

- pquired 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/qtl*. Springer New York. <https://doi.org/10.1007/978-0-387-92125-9>
- Broman, K. W., Wu, H., Sen, S., & Churchill, G. A. (2003). R/qtl: 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.
- Burke, J. M., Lai, Z., Salmaso, M., Nakazato, T., Tang, S., Heesacker, A., Knapp, S. J., & Rieseberg, L. H. (2004). Comparative Mapping and Rapid Karyotypic Evolution in the Genus *Helianthus*. *Genetics*, 167(1), 449–457. <https://doi.org/10.1534/genetics.167.1.449>
- 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>

Churchill, G. A., & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3), 963–971. <https://doi.org/10.1093/genetics/138.3.963>

<https://doi.org/10.1101/456103><https://doi.org/10.1101/456103>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>

Huang, K., Jahani, M., Gouzy, J., Legendre, A., Carrere, S., Lázaro-Guevara, J. M., González Segovia, E. G., Todesco, M., Mayjonade, B., Rodde, N., Cauet, S., Dufau, I., Staton, S. E., Pouilly, N., Boniface, M.-C., Tapy, C., Mangin, B., Duhnen, A., Gautier, V., ... Rieseberg, L. H. (2023). The genomics of linkage drag in inbred lines of sunflower. *Proceedings of the National Academy of Sciences*, 120(14), e2205783119.

<https://doi.org/10.1073/pnas.2205783119>

- 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>
- Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., Jiao, Y., Ni, P., Zhang, J., Li, D., Guo, X., Ye, K., Jian, M., Wang, B., Zheng, H., ... Wang, J. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics*, 42(11), 1027–1030. <https://doi.org/10.1038/ng.684>
- Lee, J. A., & Lupski, J. R. (2006). Genomic Rearrangements and Gene Copy-Number Alterations as a Cause of Nervous System Disorders. *Neuron*, 52(1), 103–121. <https://doi.org/10.1016/j.neuron.2006.09.027>
- 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>
- Liu, Q., Xu, J., Zhu, Y., Mo, Y., Yao, X.-F., Wang, R., Ku, W., Huang, Z., Xia, S., Tong, J., Huang, C., Su, Y., Lin, W., Peng, K., Liu, C.-M., & Xiao, L. (2021). The Copy Number Variation of OsMTD1 Regulates Rice Plant Architecture. *Frontiers in Plant Science*, 11, 620282. <https://doi.org/10.3389/fpls.2020.620282>
- 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 popu-

lations. *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>

Mangelsdorf, P. C., & Jones, D. F. (1926). THE EXPRESSION OF MENDELIAN FACTORS IN THE GAMETOPHYTE OF MAIZE. *Genetics*, 11(5), 423–455.

<https://doi.org/10.1093/genetics/11.5.423>

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>

Ohno, S. (2014). *Evolution by Gene Duplication* (Softcover reprint of the original 1st ed. 1970). Springer Berlin.

- Owens, G. L., Baute, G. J., Hubner, S., & Rieseberg, L. H. (2019). Genomic sequence and copy number evolution during hybrid crop development in sunflowers. *Evolutionary Applications*, 12(1), 54–65. <https://doi.org/10.1111/eva.12603>
- 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, Jarrad, 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>
<https://doi.org/10.1038/s42003-022-03397-7><https://doi.org/10.1038/s42003-022-03397-7>
- Rieseberg, L. H., Linder, C. R., & Seiler, G. J. (1995). Chromosomal and genic barriers to introgression in *Helianthus*. *Genetics*, 141(3), 1163–1171. <https://doi.org/10.1093/genetics/141.3.1163>

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>

Samarakoon, U., Gonzales, J. M., Patel, J. J., Tan, A., Checkley, L., & Ferdig, M. T. (2011). The landscape of inherited and de novo copy number variants in a *Plasmodium falciparum* genetic cross. *BMC Genomics*, 12(1), 457. <https://doi.org/10.1186/1471-2164-12-457>

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, FredyA., 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. In *Handbook of experimental pollination biology* (p. (pg. 142-159)). Van Nostrand Reinhold Company Inc.
- Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddeloh, J. A., Nettleton, D., & Schnable, P. S. (2009). Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genetics*, 5(11), e1000734. <https://doi.org/10.1371/journal.pgen.1000734>
- Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., & Springer, N. M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*, 20(12), 1689–1699. <https://doi.org/10.1101/gr.109165.110>
- Vasimuddin, Md., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. <https://doi.org/10.1109/IPDPS.2019.00041>
<https://doi.org/10.1093/bioinformatics/btz952>
- Żmieńko, A., Samelak, A., Kozłowski, P., & Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theoretical and Applied Genetics*, 127(1), 1–18. <https://doi.org/10.1007/s00122-013-2177-7>

Figure Captions

Figure 1: Procedural comparison of three VCF filtering approaches for sunflower genomic data analysis, with a synopsis of results.

Figure 2: Overview of the physical map of the mapping population, aligned to the HA 412 HOv2 genome. Yellow regions represent polymorphic regions in the genetic map, while blue regions represent QTL regions within 2 LOD units of peak significant variant or loci. (Top) Hard filtered and Chi-square filtered dataset. (Bottom) Chi-square filtered dataset.

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

4

Table 2. Homologous nectar-production related candidate genes of *Arabidopsis thaliana* genes within sunflower genome.

Candidate Gene (<i>Arabidopsis thaliana</i>)	Function and/or mutant phenotype	Sunflower QTL			E-value	Bitscore
		Chromosome	Start Location	End Location		
<i>ARF6</i>	Transcription factor; mutants lack nectaries (Reeves et al., 2012)	Chr09	146,751,503	146,751,357	2.22E-05	54.7
		Chr11	1,078,087	1,077,932	1.12E-16	92
<i>ARF8</i>	Transcription factor; mutants lack nectaries (Reeves et al., 2012)	Chr09	149,615,897	149,615,769	9.00E-08	55.8
		Chr11	1,078,087	1,077,932	8.00E-17	92.4
<i>CRABS CLAW</i>	Transcription factor needed for nectary development (Bowman & Smyth, 1999)	Chr06	1,332,589	1,332,362	2.45E-08	58.5
<i>CWINV4</i>	Cell wall invertase responsible for nectar production; mutants do not produce nectar (Ruhlmann et al., 2010)	Chr01	68,660,146	68,659,286	5.55E-109	365
<i>MYB21</i>	Transcription factor; mutants lack nectar and expression of SWEET9 (Schmitt et al., 2018)	Chr09	153,815,731	153,815,432	4.70E-22	99.8
		Chr09	146,202,208	146,201,897	3.27E-21	96.7
<i>MYB57</i>	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	1.48E-25	109
		Chr10	175,680,709	175,680,380	1.98E-21	96.7
<i>PIN6</i>	Transmembrane protein involved in auxin transport; mutants produced 60% less nectar (Bender et al., 2013)	Chr01	54,495,305	54,495,153	2.66E-17	63.5
		Chr09	136,142,804	136,142,373	1.07E-32	140
<i>SWEET9</i>	Sucrose uniporter; mutants do not produce nectar (Lin et al., 2014)	Chr09	170,112,792	170,113,016	5.84E-04	46.2
		Chr15	4,825,680	4,825,889	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	3.35e-127	418

5

6

7

Table 3. Additional candidate genes discovered within QTL 2 LOD unit interval on the HA412HO v2.0 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

8

9

10



