BMC Genomics

RESEARCH

Open Access

Distribution of runs of homozygosity in Lactuca species and its implications for plant breeding and evolutionary conservation

Qianqian Zhang[1,2*†], Wenjun Lv[1†], Defu Liu[3†], Xueying Xie[1], Ke Yang[1], You Tang[3] and Johann Solkner[2]

Abstract

Runs of homozygosity (ROH) have been extensively investigated to uncover the genomic inbred regions that reflect past population and breeding histories. In this study, we have explored the distribution and number of ROH in different *Lactuca* species including the cultivated lettuce varieties and their wild relatives. Next generation sequencing (NGS) technology provides the unique opportunity to study the genomes with resolution up to per-base-pair and we could compute ROH in the highest accuracy using NGS data. Our study reveals that *Lactuca sativa* has the longest average ROH length and fewest number of ROHs, while wild species show shorter, more numerous ROHs as expected. We found that these cultivated varieties exhibit relatively stable number of ROH and ROH lengths, with the largest median ROH count observed in Oilseed and the largest average ROH length in Crisphead. There is a significant proportion of medium-length ROHs (100 kb-1 Mb) enriched in *L. sativa* and *L. serriola*, with the highest number observed in *L. serriola*, while *L. saligna* has more short ROHs (< 10 KB), and the highest number of ROHs in the 10 KB-100 KB range were observed in Butterhead, with Stalk and Oilseed showing fewer and shorter ROHs overall. It suggests that Stalk and Oilseed were still in a process of breeding. The comparison between PLINK computation and our developed in-house algorithm shows that PLINK tends to detect longer ROH, whereas our algorithm adopts a more conservative approach, resulting in fewer and shorter ROH segments detected with higher precision more suitable for NGS data. We further analyze the distribution of ROH hotspots with a higher frequency occurred across cultivated species genomes, which has identified key genes such as *DREB2B, NHL12, RPV1,* and *EIX2*, which play crucial roles in plant stress tolerance and immune responses, enhancing adaptability to extreme environments and providing resistance to various diseases. These findings provide fresh scientific insights into lettuce breeding, germplasm conservation, and sustainable production, highlighting the importance of understanding and managing genetic diversity in global agricultural practices.

Keywords Lactuca species, Lettuce, Plant breeding, Inbreeding, Runs of homozygosity

Introduction

Runs of Homozygosity (ROH) refer to extended segments of the genome where both copies of the chromosomes are identical, showing no heterozygosity [1, 2]. These regions typically arise due to inbreeding or a higher inbreeding coefficient within a population. The formation of ROH is often associated with phenomena due to such as inbreeding, random genetic drift, population bottlenecks, and both natural and artificial selection [2]. These continuous

†Qianqian Zhang, Wenjun Lv and Defu Liu contributed equally to this work.

*Correspondence:
Qianqian Zhang
zhangqianqian186@hotmail.com
Full list of author information is available at the end of the article

extended regions of homozygosity can be detected at higher resolution up til a per-base-pair resolution without ascertainment bias using modern genomic technologies i.e. whole-genome sequencing [3–5].

The distribution of the ROH length can indirectly indicate the effective population size, while larger number of long ROH regions are often resulting from recent inbreeding, where alleles appear more frequently within individuals. Usually, ROH hotspots are signal regions of the genome that have undergone strong selection pressure, leading to a rapid increase in the frequency of alleles hitchhiked in the genomes across population. Moreover, long ROH regions are associated with genetic disease susceptibility, as they may harbor genes more likely linked to these conditions. By analyzing ROH, scientists can infer historical events such as migration and population bottlenecks, providing deeper insights into a population's evolutionary history [6]. At the same time, the accumulation of larger number of longer ROH regions reduce genetic diversity, weakening a population's ability to adapt in the environments. Prolonged inbreeding regions also result in higher genetic load, diminishing the population's resilience to environmental changes. In summary, study of ROH on genome not only sheds light on population structure and evolution, but also helps explore genetic diversity in terms of inbreeding affecting adaptability and disease susceptibility in the examined population.

ROH have been widely used to study population history and trait's genetic architecture in humans, livestock, and self-incompatible plants. However, their application in self-pollinating crops has yet to be reported [4]. In livestock, analysis of ROH has been a valuable tool for assessing and monitoring the levels of inbreeding and population genetic diversity. For example, in Holstein cattle, researchers utilize ROH to estimate inbreeding coefficients, which helps instruct livestock genetic improvement and breeding programs [7]. ROH analysis in self-incompatible pears revealed that European pear varieties have significantly higher numbers and total lengths of ROH compared to Asian pear varieties. The average length and number of ROH in European pears were also greater than those in Asian pears. F_{ROH} (i.e. the genomic inbreeding coefficient) showed significant correlations with various fruit traits, such as fruit weight, firmness, Brix value, and acidity. This suggests that systematic breeding of European pears may have started earlier than that of Asian pears, and the ROH patterns in European pears could be a result of population bottlenecks caused by glacial events in Europe, unlike in Asia [8]. Studying ROH patterns provides deeper insights into their distinct histories. The number, length, distribution, and frequency of ROH in plant genomes offer valuable information of genetic background. Additionally, pears exhibit gametophytic self-incompatibility, where the interaction between the pollen S-gene and the pistil S-gene prevents self-pollination during the fertilization process [9–12]. This mechanism helps maintain high genetic diversity. In contrast, self-pollinating plants are expected to more likely experience higher levels of inbreeding, resulting in lower genetic diversity and an increased risk of genetic load. Studying and reporting the patterns and distribution of ROH in self-pollinating plants genomes will also shed light on the genetic mechanism of self-pollinating of inbred lines and why it has difficulties in crossing with other inbred lines.

The genus Lactuca belongs to the subtribe *Lactucinae* within the tribe *Cichorieae* of the family *Asteraceae*. *Lactuca sativa* is one of the oldest domesticated plants and vegetable crops [13], widely cultivated around the world, while the preferences for different types of lettuce are varied across different countries and regions. Through long-term selective breeding, various cultivated types and inbred lines have been developed to meet diverse needs in breeding objectives. Officially, there are seven recognized types of lettuce inbred lines: Cos, Butterhead, Leaf, Stalk, Crisphead, Latin, and Oilseed [14]. Inbreeding [15] could increase the level of homozygosity [16] and reduce recombination frequency in the genome. In addition to inbreeding, long continuous stretches of homozygosity in the genome can also arise through mechanisms such as natural and artificial selection, genetic drift, and population bottlenecks [2]. Increased homozygosity can negatively correlated with the production, reproduction, and survival traits [17, 18], often resulting in reduced vigor and reproductive fitness in the offspring of inbred species [19, 20] and significant difficulty in crossing with other inbred lines.

The genetic variation and resistance to various biotic stresses in wild Lactuca germplasm have made significant contributions to the improvement of cultivated lettuce, particularly in enhancing disease and pest resistance [21]. Throughout the domestication and breeding process of cultivated lettuce, exogenous segments from wild germplasm containing resistance genes have been introgressed into cultivated varieties through interspecific hybridization [22]. For example, resistance to cucumber mosaic virus (CMV) comes from *Lactuca saligna* [23], while resistance to lettuce big-vein virus (LBV) is derived from *Lactuca virosa* [24].

Using Restriction Fragment Length Polymorphism (RFLP) molecular markers, researchers have identified *Lactuca serriola* as the closest relative of cultivated lettuce. The study also suggests that lettuce may form a polyphyletic group, suggesting that different cultivated types have its own distinct origins [25]. Currently, most

scholars agree that *L. serriola* is the wild progenitor of lettuce [26]. *L. sativa* and its wild relative, *L. serriola*, readily hybridize, producing fertile and vigorous hybrids, which are characteristic of a crop-wild complex [27]. Although the outcrossing rate between the two species is limited [28, 29], population genetics analysis methods could identify crop-wild hybrids within natural populations of *L. serriola*, likely due to spontaneous gene flow between the species [30]. Analyzing ROH can assess the genetic diversity and population structure of *Lactuca* species, shedding light on their evolutionary history and adaptive mechanisms. Studying ROH regions also helps identify genes associated with traits such as environmental adaptation and pest resistance, aiding in the development of stress-tolerant crops such as lettuce and providing a theoretical basis for breeding. ROH analysis serves as a tool for selecting superior genotypes and enhancing crop genetic diversity, facilitating precision breeding of lettuce and other *Lactuca* species to improve yield and quality. Understanding the genetic structure of Lactuca species also enables the formulation of more effective conservation strategies, ensuring their survival and reproduction in changing environmrents, which is crucial for preseving and utilizing their genetic resources.

With the advancement of high-throughput genomic sequencing technologies, researchers have developed a variety of ROH-based analysis tools and methods. These tools include GERMLINE, which is based on Identity by Descent (IBD) [31], Beagle and BCFtools, both of which utilize Hidden Markov Models (HMM) [32, 33], as well as PLINK, which relies on SNP window scanning [34]. These tools allow scientists to more accurately identify and analyze ROH, providing deeper insights into genetic backgrounds and population structures. Traditional computational methods, such as the PLINK software tool, offer basic ROH analysis functions by identifying homozygous regions in the genome based on specific thresholds, such as the minimum number of SNPs or the minimum ROH length [35]. As research progresses, computational methods continue to evolve to accommodate large-scale datasets and enhance analytical accuracy. For instance, some studies have proposed new algorithms or optimized existing methods to reduce false positive and negative results, while improving the resolution of ROH detection [36]. The adventure of Next-generation Sequencing (NGS) also resulted in an unbiased detection of single nucleotide polymorphism (SNP), which allows more accurate detection of ROH. These advancements include more sophisticated statistical models that better capture an individual's genetic history and population structure [37–39].

This study selected *L. sativa* and five wild *Lactuca* species. Within the gene pool of cultivated lettuce, *L. sativa* and the wild species *L. aculeata* and *L. serriola* form a group capable of complete hybridization. *L. saligna*, part of the secondary gene pool (GP2), is more difficult to cross with cultivated lettuce, and its hybrids have lower fertility [40–42]. *L. virosa* and *L. georgica*, belonging to the tertiary gene pool (GP3), typically produce infertile offspring when crossed with cultivated lettuce. Techniques like bridge crossing are required to obtain some fertile progeny from these hybrids [34]. Studying the pattern and distribution of ROH in these *Lactuca* species would aid in understanding the mechanisms of decreasing fertility or infertility due to inbreeding. We employed a series of computational methods to detect and analyze ROH in the lettuce genome from NGS data, comparing the results with traditional genetic analysis approach. The objectives of this study are 1) analyzing and reporting the distribution of ROH in different *Lactuca* species including the number and the lengths distribution of ROH and reporting the inbreeding coefficients; 2) comparing the algorithms between our developed in-house algorithm and PLINK; 3) develop an algorithm to detect the ROH hotspots in the genomes of *Lactuca* species and examining the genes within these ROH hotspots. We have provided scientific insights and practical guidance for lettuce breeding, germplasm conservation, and sustainable production through precise ROH analysis. This research not only highlights the potential of ROH in modern genetics but also underscores the importance of understanding and managing genetic diversity in global agricultural species.

Materials and methods

Data collection

The sequences of 246 varieties of lettuce genomes were collected and firstly aligned to the reference genome of *Lactuca Sativa* L, includes 78 *L. sativa* varieties (35 Butterhead, 14 Cos, 14 Crisp, 9 Cutting, 2 Latin, 3 Oilseed, and 1 Stalk), 2 *L.* aculeata, 2 *L.* georgica, 33 *L.* saligna, 109 *L.* serriola, and 22 *L.* virosa varieties (Table S1) [40]. The leaf samples were used for genomic DNA extraction using the cetyl trimethylammonium bromide method [40]. the fragments undergo end repair and adapter ligation to create a sequencing library. The sequencing was done using the DNA libraries with an insert size of 250 bp and paired-end reads using BGISEQ-500 platform [40]. The workflow for next-generation sequencing (20 × coverage depth) begins with sample preparation, where high-quality genomic DNA was extracted from *Lactuca* species.

The variants across the varieties were mapped with the reference genome and called using GATK. Post-sequencing, tools FastQC and Trimmomatic were used to assess and trim the raw data. Quality-controlled reads were then

aligned to the reference genome using alignment tools BWA or Bowtie2. Variant calling was performed using the Genome Analysis Toolkit (GATK version 4.0.3.0). A total number of 12,983,735 variants were mapped after sequence alignment, variants calling and quality control from the whole genome sequence data of Lactuca species. In total, 246 samples were used for ROH analysis.

ROH computation

The ROH computation was conducted using in-house developed algorithm adapted to the whole genome sequencing data with consideration of sequencing read depth and the tolerance of genomic heterozygosity in a counting window to define a ROH. The threshold to define a ROH was set to a SNP count maximum of 0.25 ×the genome coverage. We also used PLINK v1.9 software to calculate individual ROH regions [40]. The analysis began with genotype data files, i.e. bfile, as input. ROH calculation for each individual was performed using the `–homozyg` command. To exclude short heterozygous regions and low-density markers, the ROH regions were defined by the following criteria: each region contained a minimum of 100 SNPs (`–homozyg-snp 100`), had a minimum length of 1000 kb (`–homozyg-kb 1000`), and allowed a maximum distance of 100 kb between adjacent SNPs (`–homozyg-gap 100`). Additionally, the sliding window allowed for a maximum of one heterozygous SNP (`–homozyg-window-het 1`) to ensure the continuity of the ROH region. For each sample, PLINK outputed detailed information about the ROH regions, including chromosome number, start and end positions, the length of the ROH, and the marker density within the region. The results generated in the `.hom` file were then used for further analysis and statistical evaluation.

The computed ROH were classified into different lengths classes and the distribution was compared between different Lactuca species. The different length categories of ROH in the comparison are defined as short ROH (less than 10 KB), medium-length ROH (between 100 KB and 1 MB), and long ROH (greater than 1 MB). The inbreeding coefficient was calculated as the followed:

$$f = \frac{L_{ROH}}{L_{Genome}}$$

Table 1 The mean inbreeding coefficient of different species

L. georgica	L. saligna	L. sativa	L. serriola	L. virosa	L. aculeata
0.01	0.017 ±0.017	0.079 ±0.030	0.082 ±0.045	0.0063 ±0.0050	0.03 ±0.028

where L_{ROH} is the total length of ROH across the whole genome and L_{Genome} is the total length of Lettuce genome.

The hotspots of ROH

We firstly calculated linkage disequilibrium (LD) based on the input binary genotype file using PLINK. LD is calculated as follows: 'plink –bfile output_file –r2 –ld-window 10 –ld-window-kb 1000 –ld-window-r2 0.2 –out ld_results'. LD values (r^2) were computed using a window size of 1000 kb and a minimum threshold of 0.2. For each chromosome, the average r^2 value within each 1000 kb window was calculated, and the degree of ROH sharing on each chromosome was quantified by correcting the average LD value in the corresponding window.

The average r^2value was calculated as the squared correlation between two variables coded as 0, 1, or 2. Then we calculated the average LD within a window using 'average_r2_by_window_chr =ld_data.groupby(['CHR_A','window']) ['R2'].mean()'. We calculated the ROH hotspots frequency based on dividing the chromosome into fixed-size bins (10 kb) and counting the number of ROH occurrences within each bin. The Lettuce genomes have long-range of LD. Therefore, we further corrected the ROH sharing distribution data by dividing by the corresponding LD values in the window of the chromosome and a ratio i.e.,

$$H_{\mathrm{ROH}} = \mathrm{C}_{\mathrm{ROH}}/\bar{r}^{2}$$

where C_{ROH} is the ROH count and r^2 is the average LD, which was calculated based on ROH sharing and the average r^2 value for each window. A Manhattan plot was generated using Matplotlib to visualize the distribution of the ROH hotspots by the ROH sharing Count/r^2 Average across all the chromosomes. We further identified the genes within the ROH hotspots by taking an observation of higher number of average ROH counts/r2. By examining the genome annotation data, the genes overlapped within the ROH hotspots were found and annotated.

Result

The basic statistics of ROH distributions and inbreeding coefficients

We calculated the inbreeding coefficients of the *Lactuca* species including *L. Sativa*, *L.* aculeata, *L.* georgica, *L.* saligna, *L.* serriola, and *L.* virosa. Figure 1 showed the average inbreeding coefficients of different species and cultivated species ranging from 0.01 for *L. Georgica* to 0.82 for *L. serriola*. Specifically, *L. georgica* had an inbreeding coefficient of 0.01, *L. saligna* 0.017 ±0.017, *L. sativa* 0.079 ±0.030, *L. serriola* 0.082 ±0.045, *L. virosa* 0.0063 ±0.0050, and *L. aculeata* 0.03 ±0.028 (Table 1). Among these, *L. serriola* and *L. sativa* exhibited the

Table 2 The mean inbreeding coefficient of different cultivars

Butterhead	Cos	Crisp	Cutting	Latin	Oilseed	Stalk
0.070 ± 0.012	0.075 ± 0.017	0.070 ± 0.0047	0.06 ± 0.042	0.070	0.070	0.070

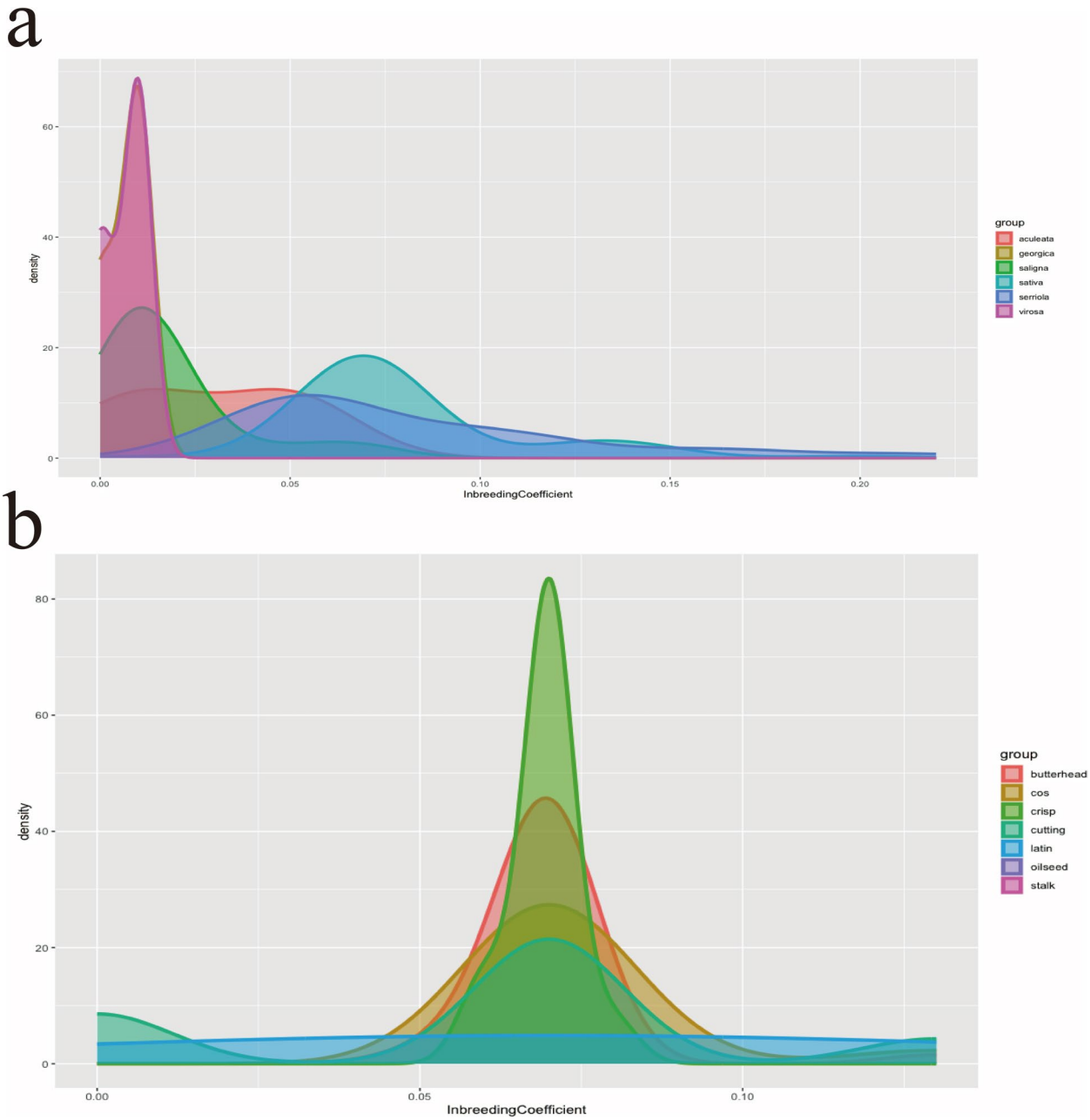

Fig. 1 Distribution of Genomic Inbreeding Coefficient Measurements. **a** Distribution of genomic inbreeding coefficient measurements based on homozygosity across different Lactuca species. The table shows the average inbreeding coefficients for different species. **b** Distribution of genomic inbreeding coefficient measurements based on homozygosity across different cultivars. The table shows the average inbreeding coefficient of different cultivars

highest inbreeding coefficients (0.082 and 0.079, respectively), while *L. virosa* showed the lowest inbreeding coefficient (0.0063). The F_{ROH} for *L. Sativa* vary between different varieties ranging from 0.06 for Cutting to 0.075 for Cos, while Cutting had a slightly lower inbreeding coefficient (0.06 ± 0.042), reflecting relatively higher heterozygosity in its genome (Table 2). For the Crisp and Cos, a higher proportion of individual genomes exhibit moderate levels of inbreeding compared with other lettuce varieties. It is as expected that the inbreeding coefficient of *Lactuca Sativa* was relatively higher among these species as it has been domesticated and cultivated by breeders for years. The inbreeding coefficient of *L. virosa* is primarily concentrated around 0.01. In contrast, other species, such as *L. sativa* and *L. serriola*, showed a broader distribution of inbreeding coefficients (Fig. 1a). Most cultivars i.e. Crisphead, Butterhead, Cos, Cutting, Latin, Oilseed and Stalk show similar inbreeding coefficients, which were main in-bred types of leaf lettuce in the market.

The distribution of the length and number of ROH

The mean ROH length of *L. sativa* is the longest among the six species analyzed, at 65 ± 105.00 KB. The median length of ROH length of *L. sativa* is 25 KB, which is higher than any of the other species, and the highest ROH length in the non-outlier dataset reaches 146 KB (Fig. 2a). *L. sativa* also has the fewest number of ROHs, with a mean of 287 ROH regions per chromosome (Fig. 2b). In contrast, the median ROH length of *L. georgica, L. saligna, and L. virosa* is only 2 KB. *L. saligna* exhibits the highest number of ROHs, with an average of 948 per chromosome.

The general ROH length and number of ROH are distributed similar among the different cultivated varieties. The median ROH length of cultivated varieties is relatively concentrated, with the longest being observed in Latin with median length of 32 KB. The shortest median length is observed in Oilseed, with a median length of 21 KB (Fig. 2c). Oilseed also has the largest median number of ROH regions, with 367 ROH regions, while Cutting

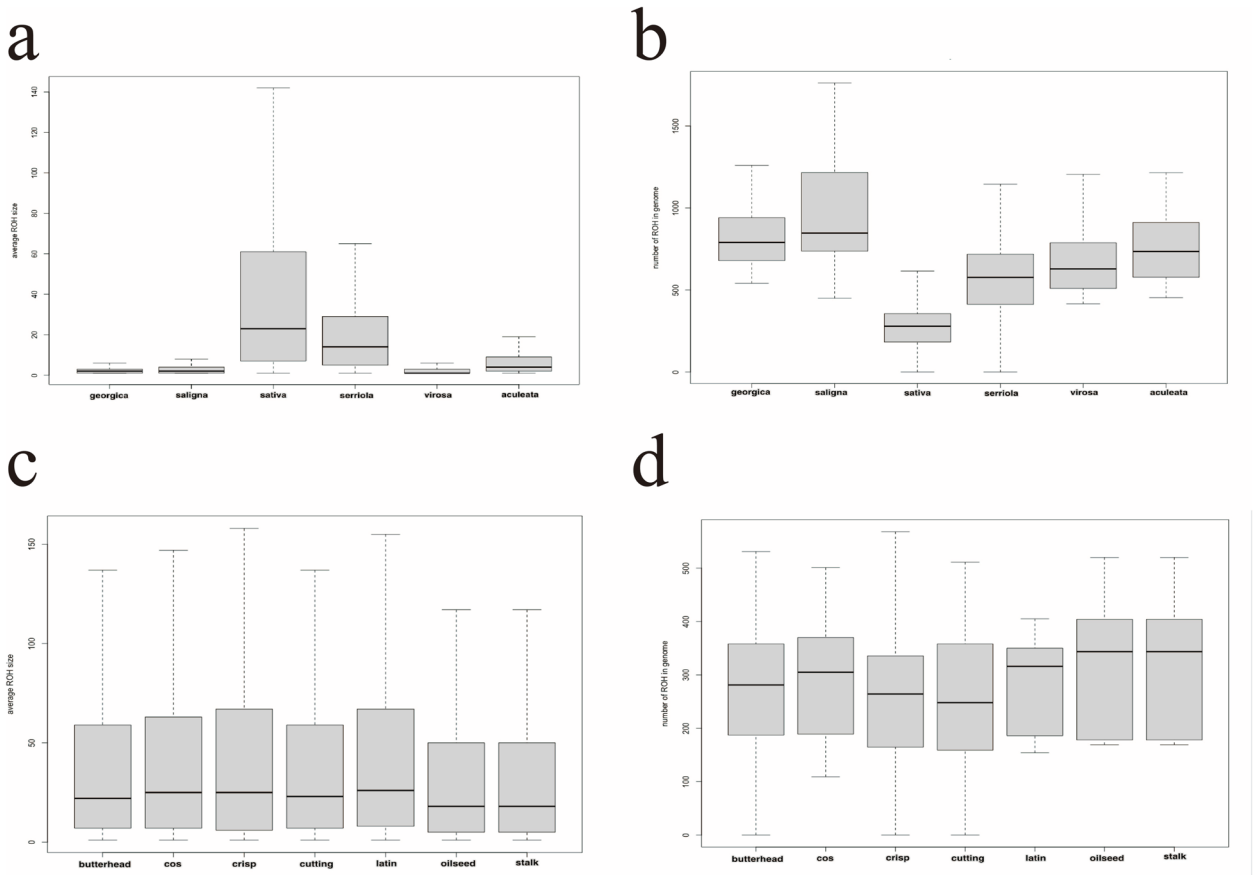

Fig. 2 Comparison of Genomic ROH Characteristics Across Different Species and Cultivars. **a** Comparison of average ROH length between species. **b** Comparison of ROH count between species. **c** Comparison of average ROH length between cultivars. **d** Comparison of ROH count between cultivars

has the smallest median number of ROH, with 227 ROH regions (Fig. 2d). Crisp has the longest ROH length with mean at 74.56 KB, while Oilseed has the shortest ROH length of mean at 49.8 KB. Overall, the number and length of ROH regions in cultivated varieties show relatively low variation and remain relatively stable across species (i.e. the mean number of individual cultivar species are 61.97 ± 103.93 for Butterhead, 67.85 ± 100.79 for Cos, 74.56 ± 119.75 for Crisphead, 66.47 ± 102.92 for Cutting, 68.90 ± 100.15 for Latin, Oilseed is 49.80 ± 80.86 for Oilseed, 61.44 ± 90.18 for Stalk).

Total genome length and number covered by ROH

We firstly calculated ROH using our in-house developed algorithm. In *L. sativa* and *L. serriola*, medium-length ROHs (100 kb-1 Mb) make up a significant proportion (Fig. 3a). In contrast, the total length of ROH in other species is notably lower than in *L. sativa* and *L. serriola* (Fig. 3b). *L. serriola* has the highest number of ROHs, particularly in the range from 100 kb to 1 Mb. In *L. saligna*, there are more number of ROHs with length shorter than 10 KB, but fewer number of ROHs with length longer than 10 KB. Butterhead has the highest number of ROHs, particularly in the length range from 10 to 100 Kb. In contrast, Stalk and Oilseed have fewer number of ROHs (Fig. 3c). The sum length of ROH in Butterhead was ranging from 100 kb to 1 Mb (Fig. 3d). The total ROH length of Stalk and Oilseed is relatively shorter compared with other cultivars (Fig. 3d).

In order to show the robustness of our ROH algorithm, we compare the computational results with ROH computed with PLINK. The general trend of the observations in terms of ROH length, the numbers of ROH are quite similar between two algorithms. The medium and long ROH are predominantly distributed in the *Lactuca* species, while our algorithm designed for NGS data is more able to detect the short or medium ROH especially for the ROH with length at 10kbp. We observe that the majority of ROH counts are distributed in the range between 10 to 100 Kb, particularly in *L. saligna* and *L. virosa* (Supplementary Fig. 1a). *L. sativa* and *L. serriola* exhibit a higher number of long ROH with lengths

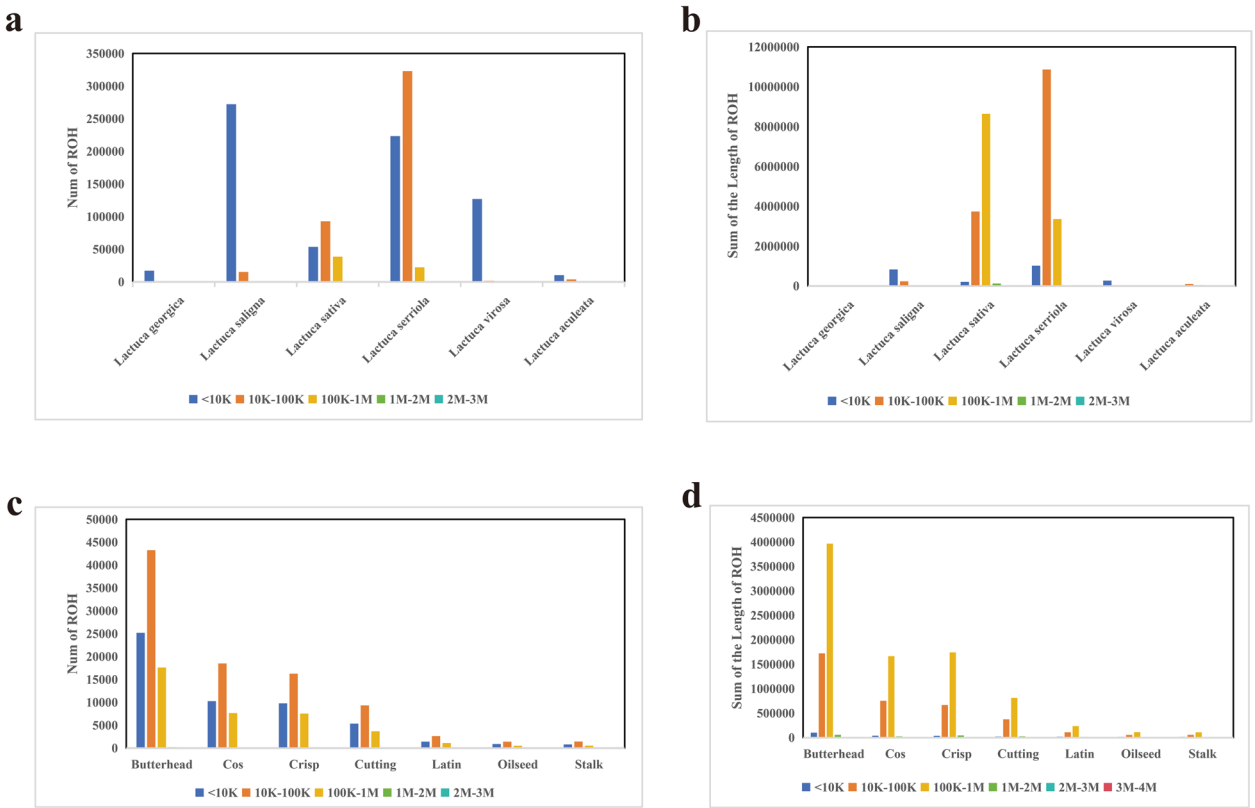

Fig. 3 ROH (Runs of Homozygosity) Statistics Across Different Lettuce Species and Cultivars. The figure presents the statistical results of ROH across various lettuce species and cultivars, categorized by ROH length ranges (< 10 K, 10 K-100 K, 100 K-1 M, 1 M-2 M, and 2 M-3 M). Each subplot uses color to represent different ROH length intervals: blue (0–100 kb), orange (100 kb-1 Mb), and green (> 1 Mb). **a** Bar plot showing the number of ROH segments across different lettuce species. **b** Bar plot showing the total ROH length across different lettuce species. **c** Bar plot showing the number of ROH segments across different lettuce cultivars. **d** Bar plot showing the total ROH length across different lettuce cultivars

ranging above 4 M. (Supplementary Fig. 1b). Among the different cultivars, Butterhead shows the highest number of ROHs in the range from 10 to 100 Kb (Supplementary Fig. 1c). The total ROH length is longer in cultivars such as Butterhead and Cos compared with Crisp, Cutting, Latin, Oilseed, and Stalk (Supplementary Fig. 1d).

L. serriola has the highest number of ROHs compared to other species, consistently. However, when comparing the results with the algorithm used in this study, the PLINK algorithm detects fewer number of ROHs overall. These ROHs were primarily concentrated within the range from 10 Kb to 100 K, with *L. saligna* and *L. virosa* exhibiting higher counts of ROHs, at 897,043 Mb and 657,100 Mb, respectively (Supplementary Fig. 1). In contrast, *L. serriola* had a relatively lower number of ROHs in this range, at 32,852 Mb. Our own algorithm calculated the number of ROHs predominantly concentrated in the sub-10 Kb range, particularly for *L. saligna* and *L. serriola*, which had 272,183 Mb and 223,331 Mb, respectively (Fig. 3).

ROH hotspots

We have identified the ROH hotspots defined as the regions with highest frequency occurrence across the individual genomes with consideration of the long LD range in *Lactuca* genomes. The prominent peaks in the plot indicate regions in the genome with both high ROH frequency and correcting the strong LD, suggesting potential hotspots of ROH sharing across individual genomes (Figs. 4–5). Interestingly, further analysis of these hotspots in cultivated species revealed several key genes and their functions, which play critical roles in plant stress tolerance and immune responses. We have in total identified 59 genes in the ROH hotspots of the cultivars (Supplementary Table 5), including Gene *Dehydration-responsive element-binding protein 2B (DREB2B),* gene *NDR1/HIN1-like protein 12 (NHL12), Disease resistance protein RPV1 (RPV1), and Receptor-like protein EIX2 (EIX2).*

Discussion

L. sativa, as a cultivated species, exhibits a broad distribution on earth, while wild species such as *L. aculeata* are primarily found in the Middle East. The distribution of *L. georgica* may be limited to specific regions, such as Georgia and neighboring countries (Supplementary Table 3). *L. sativa* is the species that includes both cultivated and wild varieties with a long history of domestication and cultivation. In contrast, other species, such as *L. georgica*, *L. saligna*, and *L. aculeata*, primarily are wild species and have not undergone widespread domestication. *L. sativa* has a larger number of cultivated varieties and fewer wild varieties, reflecting its agricultural significance and degree of domestication. On the other hand, *L. saligna* has a greater number of wild samples, highlighting its broad distribution in natural environments (Supplementary Table 4).

Our in-house developed algorithm could be used to detect ROH from NGS sequencing data with high accuracy especially for short ROHs, while ROH computation algorithm from PLINK seems to have a tendency to detect longer runs of homozygosity (ROH), which

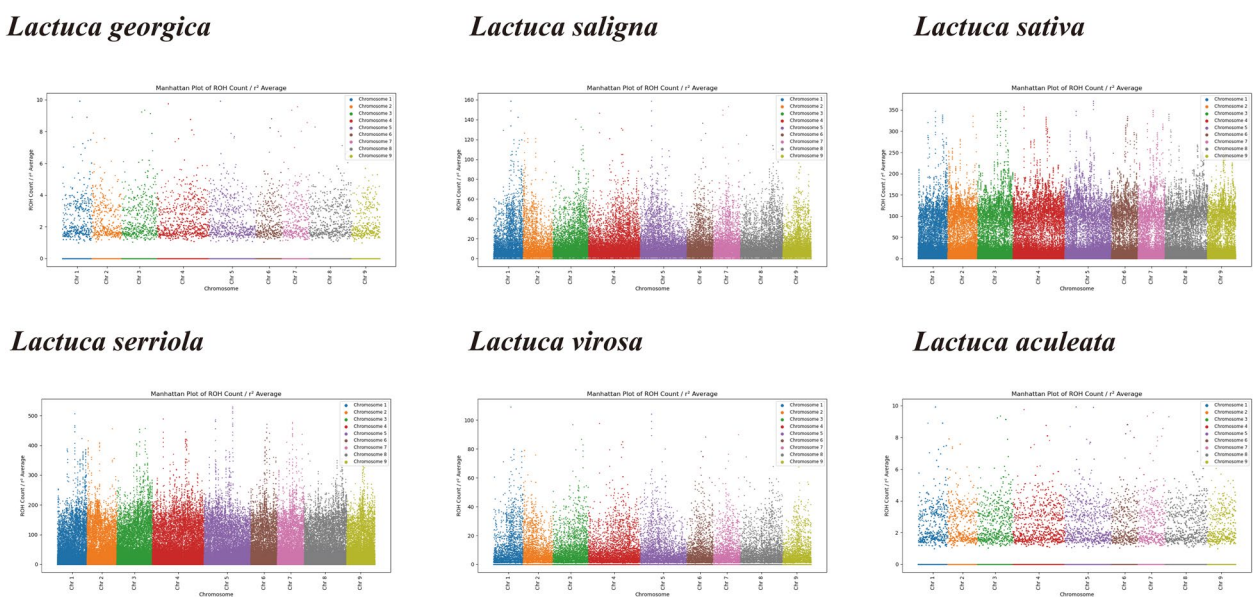

Fig. 4 The ROH hotspots map of different species, with points in different colors representing different chromosomes

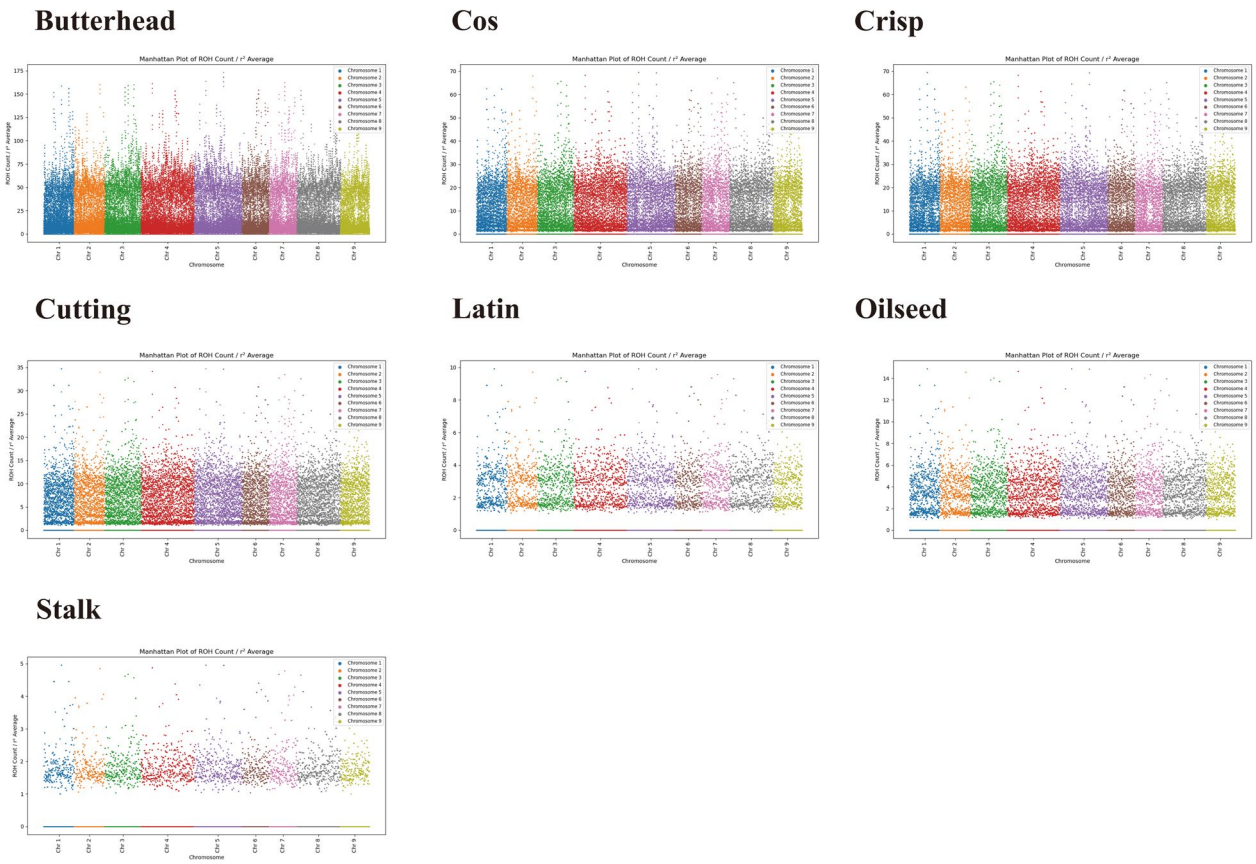

Fig. 5 The ROH hotspots map of different breeds, with points in different colors representing different chromosomes

might be related to its parameter settings or underlying algorithmic logic. In contrast, our own algorithm takes a more conservative approach, resulting in more short ROH detected with higher total lengths of ROH.

The levels of inbreeding measured by the calculated inbreeding coefficients in Lactuca species indicate that *L. virosa* individuals exhibit little to no evidence of inbreeding, while species such as *L. sativa* and *L. serriola* show higher levels of inbreeding. Among these, *L. serriola* and *L. sativa* exhibited the highest inbreeding coefficients (0.082 and 0.079, respectively), indicating a greater accumulation of homozygous segments in their genomes, while *L. virosa* showed the lowest inbreeding coefficient (0.0063), suggesting higher genomic heterozygosity. So it is under expectation that *L. serriola* and *L. sativa* were more inbred than other types, which probably needs further genetic management during the breeding program. The breeders have been emphasizing on crossbreeding between these cultivars, but little success has been achieved [19]. The inbreeding coefficient of all varieties overlapped mostly at about 0.07 (Supplementary Table 2).

In *L. sativa*, the average ROH length is the longest, but the number of ROH segments is the fewest, suggesting that *L. sativa* may have undergone prolonged selective sweeps with artificial long-term selective breeding as expected, resulting in the accumulation of long homozygous segments. This leads to fewer but longer ROHs. As a widely cultivated crop, *L. sativa* has been subject to extensive artificial selection over time. During domestication, the genomic regions and genes associated with specific traits have been artificially selected for, such as yield or disease resistance, leading to increased uniformity in these regions across the population [43]. This uniformity increases the length of extended ROH because heterozygosity in these regions is gradually eliminated, resulted in fixing homozygosity in the population.

At the same time, modern agriculture may introduce new gene introgressed into *L. sativa* through cross breeding [44], which reduces the overall number of ROHs in the genome. However, artificial selection continues to target specific key functional genomic regions, where selective pressure results in longer ROHs in

those genomic areas. The prolonged ROH in *L. serriola* may also reflect a more closed population structure and potential inbreeding, while the shorter ROH in *L. saligna* could indicate a more open breeding environment with stronger gene flow.

In contrast, wild species such as *L. georgica*, *L. virosa*, and *L. aculeata* exhibit more shorter ROH but a greater number of ROH segments. This is likely because these wild species have not been subjected to the same artificial selective pressures as cultivated species and typically possess higher genetic diversity. As natural selection operates in various environments, wild populations tend to accumulate more short ROHs due to natural selection, but without the prolonged artificial selection, the ROH segments remain shorter [45]. If these species have experienced more frequent gene flow (e.g. introgression from other populations in natural environment), this could increase genetic diversity and cause more interruptions in ROH formation. Additionally, these species may have undergone more hybridization or mixed selection with accumulated genetic drift in natural environment, reducing the formation of long homozygous regions and increasing the number of shorter ROH segments [46].

Among the different varieties, Butterhead consistently shows the highest number of ROHs. However, according to the results from our algorithm, the ROHs in Butterhead are primarily concentrated in the range from 10 Kb to100 Kb. Our algorithm still detects more number of ROHs overall in different ranges including short, medium and long. On the other hand, the PLINK algorithm tends to detect longer ROH segments (> 3 Mb). The total ROH length detected by the PLINK algorithm is longer than detected by our algorithm.

The overlap in genomic inbreeding coefficient distributions based on homozygosity across different cultivars suggests that multiple cultivated types share similar levels of inbreeding within certain ranges. This overlap in peaks may be due to the shared genetic resources among these cultivars during breeding. These types likely underwent similar artificial selection pressures, leading to overlapping ROH regions in their genomes [47]. The ROH distribution pattern of Butterhead may suggest that its population has undergone prolonged inbreeding, a population bottleneck, or has relatively low genetic diversity. In contrast, the shorter total ROH length and fewer number of ROHs in Stalk and Oilseed could reflect higher gene flow and better genetic diversity in these species, indicating a more diverse breeding process.

Analysis of the ROH hotspot in cultivars regions revealed several genes closely linked to plant stress tolerance and immune responses. The *DREB2B* gene, a regulator of drought and heat shock stress, likely enhances plant resilience to extreme conditions by modulating physiological processes such as water balance and heat stress response. As climate change intensifies, drought- and heat-resistant genes will provide valuable genetic resources for improving agricultural crops [48–50]. The *NDR1/HIN1-like protein 12* gene may play a role in plant immunity. As a homolog of *NDR1*, it may participate in pathogen-associated molecular pattern recognition and regulate the plant immune response. This discovery offers new insights for further exploring plant immunity mechanisms and may help improve crop resistance to diseases [51, 52]. The *RPV1* gene, a disease resistance (R) protein, activates plant defense responses by promoting cell death, conferring resistance to powdery mildew and downy mildew. It belongs to the disease resistance (R) protein family and contributing to the plant immune response, acting as NADase, which catalyzes the cleavage of NAD into ADP-D-ribose (*ADPR*) and nicotinamide, triggers cell death and initiates plant defense mechanisms [53].The presence of this gene reveals key mechanisms in plant immunity, and future studies could explore its function and potential applications across different plant species [53–55]. The *EIX2* gene, a receptor-like protein, is involved in pathogen recognition and initiation of the plant immune response. It may serve as a critical regulatory factor in the plant immune system, and further research could uncover its specific role in plant immunity [52, 56]. In conclusion, these genes in the ROH hotspot regions play significant roles in plant adaptation and immune responses, providing valuable genetic resources for agricultural improvement and stress resistance research. Future studies will further explore the specific functions and mechanisms of these genes to provide more precise molecular tools for enhancing crop disease resistance and adaptability.

Conclusions

In this study, we investigated the distribution and characteristics of runs of homozygosity (ROH) across various Lactuca species, including both cultivated lettuce varieties and their wild relatives. Our analysis using next-generation sequencing (NGS) technology allowed for highly accurate computation of ROH, providing valuable insights into the genomic history of these species. The results revealed that *L. sativa* (cultivated lettuce) exhibited the longest mean ROH length and the fewest number of ROHs, suggesting a history of inbreeding or limited genetic diversity. In contrast, the wild species showed a greater number of shorter ROHs, indicating higher genetic diversity and a more open breeding system.

Further analysis revealed distinct patterns within specific cultivated varieties, such as Oilseed, which had the

largest median ROH count, and Crisp, which had the longest mean ROH length. Additionally, *L. sativa* and *L. serriola* were found to have a significant proportion of medium-length ROHs, while *L. saligna* had more short ROHs. Notably, Butterhead exhibited the highest number of ROHs in the 10 KB-100 KB range, and the Stalk and Oilseed varieties had fewer and smaller ROHs, reflecting ongoing breeding processes.

We also compared two different computational approaches for detecting ROH, i.e. PLINK and an in-house algorithm we developed. While PLINK tended to identify longer ROHs, our algorithm was more conservative, identifying shorter and fewer ROH segments with greater precision.

Moreover, our research identified ROH hotspots in cultivated species associated with key genes such as *DREB2B, NHL12, RPV1,* and *EIX2*, which are crucial for stress tolerance and immune response in plants. These findings suggest that these genes may play important roles in enhancing the adaptability of lettuce to extreme environmental conditions and disease resistance.

In summary, this study provides valuable insights into the genetic structure of cultivated and wild Lactuca species, with implications for lettuce breeding, germplasm conservation, and sustainable agricultural production. The patterns of ROH distribution reflect the complex interplay of genetic diversity, breeding history, and environmental adaptation in these species.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11674-y.

Supplementary Material 1.

Acknowledgements
The authors acknowledge the fundings of Beijing Novo Program and National Natural Science Foundation of China (there is no order between the fundings).

Authors' contributions
Q.Z. designed the study, coordinated the study, implemented the computational programming part of this study and analyzed the data and edited the manuscript. W.L. analyzed the data and wrote the manuscript. D.L. assisted Q.Z. in this work. X.X. participated in the analysis of the study and edited the manuscript. K.Y. assisted Q.Z. in analysis of the data and drafted the manuscript. Y.T. participated in the analysis of the study and edited the manuscript. J.S. edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
Q. Z. is supported by Beijing Novo Program and Young Scholar Funding (type C) in National Natural Science Foundation of China (there is no order between the fundings).

Data availability
The raw sequencing data are deposited with accession of BioProject PRJNA693894 in the Sequence Read Archive.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Chemical and Biological Engineering, University of Science and Technology, Beijing, China. [2]Department of Agriculture, University of Natural Resources and Life Sciences Vienna BOKU, Vienna, Austria. [3]Electrical and Information Engineering College, Jilin Agricultural Science and Technology University, Jilin, China.

Received: 19 March 2025 Accepted: 6 May 2025
Published online: 14 May 2025

References

1. Liu G, Sun FZ, Zhu FX, Feng HY, Han X. Runs of homozygosity and its application on livestock genome study. Yi Chuan. 2019;41(4):304–17. https://doi.org/10.16288/j.yczz.18-287. PMID: 30992252. Chinese.
2. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nat Rev Genet. 2018;19(4):220–34. https://doi.org/10.1038/nrg.2017.109. Epub 2018 Jan 15 PMID: 29335644.
3. Ceballos FC, Hazelhurst S, Ramsay M. Assessing runs of Homozygosity: a comparison of SNP array and whole genome sequence low coverage data. BMC Genom. 2018;19(1):106. https://doi.org/10.1186/s12864-018-4489-0.
4. Curik I, Ferenčaković M, Sölkner J. Inbreeding and runs of homozygosity: a possible solution to an old problem. Livest Sci. 2014;166:26–34. https://doi.org/10.1016/j.livsci.2014.05.034.
5. Pryce JE, Haile-Mariam M, Goddard ME, Hayes BJ. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet Sel Evol. 2014;46:71. https://doi.org/10.1186/s12711-014-0071-7.
6. Peripolli E, Munari DP, Silva MVGB, Lima ALF, Irgang R, Baldi F. Runs of homozygosity: current knowledge and applications in livestock. Anim Genet. 2017;48(3):255–71. https://doi.org/10.1111/age.12526. Epub 2016 Dec 1 PMID: 27910110.
7. Kumar S, Deng CH, Hunt M, Kirk C, Wiedow C, Rowan D, Wu J, Brewer L. Homozygosity mapping reveals population history and trait architecture in self-incompatible pear (Pyrus spp.). Front Plant Sci. 2021;5(11):590846. https://doi.org/10.3389/fpls.2020.590846. PMID: 33469460; PMCID: PMC7813798.
8. Liu SH, Ma XY, Hassan FU, Gao TY, Deng TX. Genome-wide analysis of runs of homozygosity in Italian Mediterranean buffalo. J Dairy Sci. 2022;105(5):4324–34. https://doi.org/10.3168/jds.2021-21543. Epub 2022 Mar 17 PMID: 35307184.
9. Zuccherelli S, Tassinari P, Broothaerts W, Tar-Tarini S, Dondini L, Sansavinl S. S-allele characterization in self-incompatible pear (Pyrus communis L.). Sex Plant Reprod. 2002;15(3):153–8.
10. Sanzol J, Herrero M. Identification of self-incompatibility alleles in pear cultivars (Pyrus communis L.). Euphytica. 2002;128(3):325–31.
11. Kim HT, Hrata Y, Nou IS. Determination of s-genotypes of pear (Pyrus pyrifolia) cultivars by S-RNase sequencing and PCR-RFLP analyses. Mol Cells. 2002;13(3):444–51.
12. Wang Y, Wang X, Skirpan AL, Kao TH. S-RNase-mediated self-incompatibility. J Exp Bot. 2003;54(380):115–22.
13. Hancock JF. Plant evolution and the origin of crop species[M]. Wallingford: CABI; 2012.
14. Mampholo BM, Maboko MM, Soundy P, Sivakumar D. Phytochemicals and overall quality of leafy lettuce (Lactuca sativa L.) varieties grown in closed

hydroponic system. J Food Qual. 2016;39:805–15. https://doi.org/10.1111/jfq.12234.
15. Hassan MN, Mekkawy SA, Mahdy M, et al. Recent molecular and breeding strategies in lettuce (Lactuca spp.). Genet Resour Crop Evol. 2021;68:3055–79. https://doi.org/10.1007/s10722-021-01246-w.
16. Charlesworth D. Effects of inbreeding on the genetic diversity of populations. Philos Trans R Soc London Ser B Biol Sci. 2003;358:1051–70. https://doi.org/10.1098/rstb.2003.1296.
17. Leroy G. Inbreeding depression in livestock species: review and meta-analysis. Anim Genet. 2014;45:618–28.
18. Baes CF, Makanjuola BO, Miglior F, Marras G, Howard JT, Fleming A, Maltecca C. Symposium review: the genomic architecture of inbreeding: how homozygosity affects health and performance. J Dairy Sci. 2019;102:2807–17.
19. Moore JN, Janick J. Advances in fruit breeding. West Lafayette: Purdue University Press; 1975.
20. Angeloni F, Ouborg NJ, Leimu R. Meta-analysis on the association of population size and life history with inbreeding depression in plants. Biol Conserv. 2011;144:35–43. https://doi.org/10.1016/j.biocon.2010.08.016.
21. Lebeda A, Křístková E, Kitner M, Mieslerová B, Jemelková M, Pink DAC. Wild Lactuca species, their genetic diversity, resistance to diseases and pests, and exploitation in lettuce breeding. Eur J Plant Pathol. 2014;138:597–640.
22. Parra L, Maisonneuve B, Lebeda A, Schut J, Christopoulou M, Jeuken M, McHale L, Truco MJ, Crute I, Michelmore R. Rationalization of genes for resistance to Bremia lactucae in lettuce. Euphytica. 2016;210:309–26.
23. Provvidenti R, Robinson RW, Shail JW. A source of resistance to a strain of cucumber mosaic virus in Lactuca saligna L. HortScience. 1980;15:528–9.
24. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
25. Kesseli R, Ochoa O, Michelmore R. Variation at RFLP loci in Lactuca spp. and origin of cultivated lettuce (L. sativa). Genome. 1991;34:430–6.
26. de Vries IM. Origin and domestication of Lactuca sativa L. Genet Resour Crop Evol. 1997;44:165–74. Mou B. Lettuce. In: Vegetables I. Springer; 2008. p. 75-116.
27. De Vries IM. Crossing experiments of lettuce cultivars and species (Lactuca sect. Lactuca, Compositae). Plant Syst Evol. 1990;171:233–48. https://doi.org/10.1007/BF00940608.
28. D'Andrea L, Felber F, Guadagnuolo R. Hybridization rates between lettuce (Lactuca sativa) and its wild relative (L. serriola) under field conditions. Environ Biosafety Res. 2008;7:61–71. https://doi.org/10.1051/ebr:2008006.
29. Thompson RC, Whitaker TW, Bohn GW, Van Horn CW. Natural cross-pollination in lettuce. Am Soc Hort Sci. 1958;72:403–9.
30. Uwimana B, D'Andrea L, Felber F, Hooftman DAP, den Nijs HCM, Smulders MJM, Visser RGF, van de Wiel CCM. A Bayesian analysis of gene flow from crops to their wild relatives: cultivated (Lactuca sativa L.) and prickly lettuce (L. serriola L.) and the recent expansion of L. serriola in Europe. Mol Ecol. In press. https://doi.org/10.1111/j.1365-294X.2012.05489.x.
31. Gusev A, Lowe J, Stoffel M, Daly M, Altshuler D, Breslow J, Friedman J, Peer I. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 2009;19(2):18–26. https://doi.org/10.1101/gr.081398.108.
32. Browning S, Browning B. High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet. 2010;86(4):526–39. https://doi.org/10.1016/j.ajhg.2010.02.021.
33. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/ROH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32(11):1749–51. https://doi.org/10.1093/bioinformatics/btw044.
34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and populationbased linkage analyses. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10.1086/519795.
35. Smaragdov MG. Identification of homozygosity-rich regions in the Holstein genome. Vavilovskii Zhurnal Genet Selektsii. 2023;27(5):471–9. https://doi.org/10.18699/VJGB-23-57. PMID: 37808215; PMCID: PMC10556852.
36. Ferencakovic M, Hamzic E, Gredler B, Curik I, Johann S. Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. Agric Conspec Sci. 2011;76(4):325–9.
37. Purfield D, Berry D, McParland S, Bradley D. Runs of homozygosity and population history in cattle. BMC Genet. 2012;13:70. https://doi.org/10.1186/1471-2156-13-70.
38. Ferenčaković M, Sölkner J, Curik I. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genet Sel Evol. 2013;45:42.
39. Karimi Z. Runs of homozygosity patterns in taurine and indicine cattle breeds. Doctoral thesis. Vienna: Univ. of Natural Resources and Life Sciences; 2013.
40. Wei T, van Treuren R, Liu X, et al. Whole-genome resequencing of 445 Lactuca accessionsreveals the domestication history of cultivated lettuce. Nat Genet. 2021;53(5):752–60.
41. Treuren RV, Menting FBJ. PGR lettuce: Lettuce crop portal[Z]. 2022. https://www.pgrportal.nl/en/Lettuce-genetic-resources-Portal/Other-info-2/Taxonomy/The-lettuce-gene-pool-1.htm.
42. Lebeda A, Ryder EJ, Grube R, et al. Lettuce (Asteraceae; Lactuca spp.), vol. 3. Boca Raton: CRC Press; 2007. p. 377–472.
43. Simko I, Hayes RJ, Truco MJ. Mapping the genetic basis of agronomic traits in lettuce. Theor Appl Genet. 2014;127(9):1743–61. https://doi.org/10.1007/s00122-014-2335-6.
44. Cao S, Sawettalake N, Li P, Fan S, Shen L. DNA methylation variations underlie lettuce domestication and divergence. Genome Biol. 2024;25(1):158. https://doi.org/10.1186/s13059-024-03310-x. PMID: 38886807; PMCID: PMC11184767.
45. Uwimana B, Smulders MJ, Hooftman DA, Hartman Y, van Tienderen PH, Jansen J, van de Wiel C. Hybridization between crops and wild relatives: the contribution of cultivated lettuce to the vigour of crop-wild hybrids under drought, salinity, and nutrient deficiency conditions. Theor Appl Genet. 2012;125(7):1237–52. https://doi.org/10.1007/s00122-012-1905-7.
46. Dempewolf H, Guarino L. Reaching back through the domestication bottleneck: tapping wild plant biodiversity for crop improvement. Acta Hort. 2015;1101:165–8. https://doi.org/10.17660/ActaHortic.2015.1101.22.
47. Ferencakovic M, Gredler B, Solberg TR, et al. Runs of homozygosity reveal genomewide autozygosity in the Austrian Fleckvieh cattle. Agric Conspec Sci. 2011;76(4):325–8. https://doi.org/10.1111/jbg.12012.
48. Sakuma Y, Liu Q, Dubouzet JG, Abe H, Shinozaki K, Yamaguchi-Shinozaki K. DNA-binding specificity of the ERF/AP2 domain of Arabidopsis DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. Biochem Biophys Res Commun. 2002;290(3):998–1009. https://doi.org/10.1006/bbrc.2001.6299. PMID: 11798174.
49. Zhao JL, Wu Q, Wu HL, Wang AH, Wang XL, Li CL, Zhao HX, Wu Q. FtNAC31, a Tartary buckwheat NAC transcription factor, enhances salt and drought tolerance in transgenic Arabidopsis. Plant Physiol Biochem. 2022;15(191):20–33. https://doi.org/10.1016/j.plaphy.2022.09.016.
50. Sakata T, Oda S, Tsunaga Y, Shomura H, Kawagishi-Kobayashi M, Aya K, Saeki K, Endo T, Nagano K, Kojima M, Sakakibara H, Watanabe M, Matsuoka M, Higashitani A. Reduction of gibberellin by low temperature disrupts pollen development in rice. Plant Physiol. 2014;164(4):2011–9. https://doi.org/10.1104/pp.113.234401.
51. Zheng MS, Takahashi H, Miyazaki A, Hamamoto H, Shah J, Yamaguchi I, Kusano T. Up-regulation of Arabidopsis thaliana NHL10 in the hypersensitive response to cucumber mosaic virus infection and in senescing leaves is controlled by signalling pathways that differ in salicylate involvement. Planta. 2004;218(5):740–50. https://doi.org/10.1007/s00425-003-1169-2. Epub 2003 Dec 10 PMID: 14666423.
52. Wang N, Yin Z, Zhao Y, Li Z, Dou D, Wei L. Two divergent immune receptors of the allopolyploid Nicotiana benthamiana reinforce the recognition of a fungal microbe-associated molecular pattern VdEIX3. Front Plant Sci. 2022;15(13):968562. https://doi.org/10.3389/fpls.2022.968562. PMID: 36046591; PMCID: PMC9421165.
53. Nascimento-Gavioli MC, Agapito-Tenfen SZ, Nodari RO, Welter LJ, Sanchez Mora FD, Saifert L, da Silva AL, Guerra MP. Proteome of Plasmopara viticola-infected Vitis vinifera provides insights into grapevine Rpv1/Rpv3 pyramided resistance to downy mildew. J Proteomics. 2017;16(151):264–74. https://doi.org/10.1016/j.jprot.2016.05.024. Epub 2016 May 26 PMID: 27235723.
54. Decroocq V, Sicard O, Alamillo JM, Lansac M, Eyquard JP, García JA, Candresse T, Le Gall O, Revers F. Multiple resistance traits control Plum pox virus infection in Arabidopsis thaliana. Mol Plant Microbe Interact. 2006;19(5):541–9. https://doi.org/10.1094/MPMI-19-0541.
55. Feechan A, Anderson C, Torregrosa L, Jermakow A, Mestre P, Wiedemann-Merdinoglu S, Merdinoglu D, Walker AR, Cadle-Davidson L, Reisch B,

Aubourg S, Bentahar N, Shrestha B, Bouquet A, Adam-Blondon AF, Thomas MR, Dry IB. Genetic dissection of a TIR-NB-LRR locus from the wild North American grapevine species Muscadinia rotundifolia identifies paralogous genes conferring resistance to major fungal and oomycete pathogens in cultivated grapevine. Plant J. 2013;76(4):661–74. https://doi.org/10.1111/tpj.12327.

56. Ron M, Avni A. The receptor for the fungal elicitor ethylene-inducing xylanase is a member of a resistance-like gene family in tomato. Plant Cell. 2004;16(6):1604–15. https://doi.org/10.1105/tpc.022475.

Publisher's Note