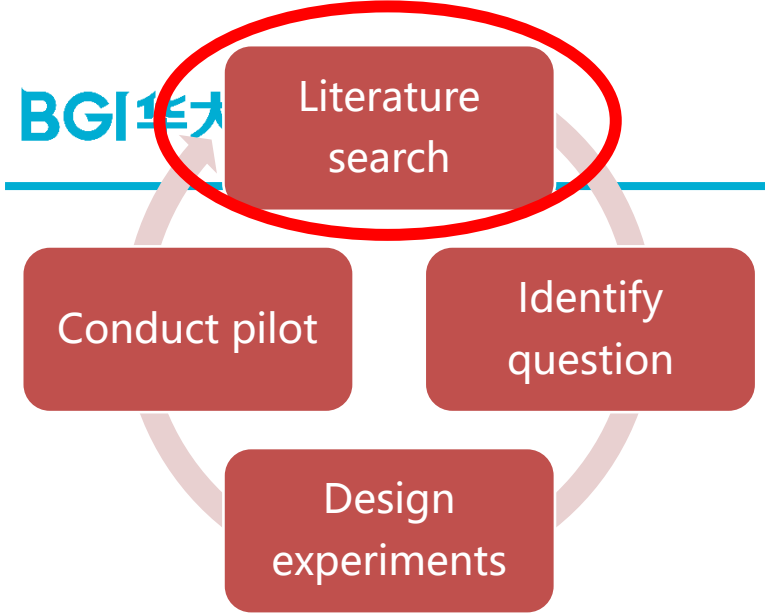
Part II
Scientific Reading

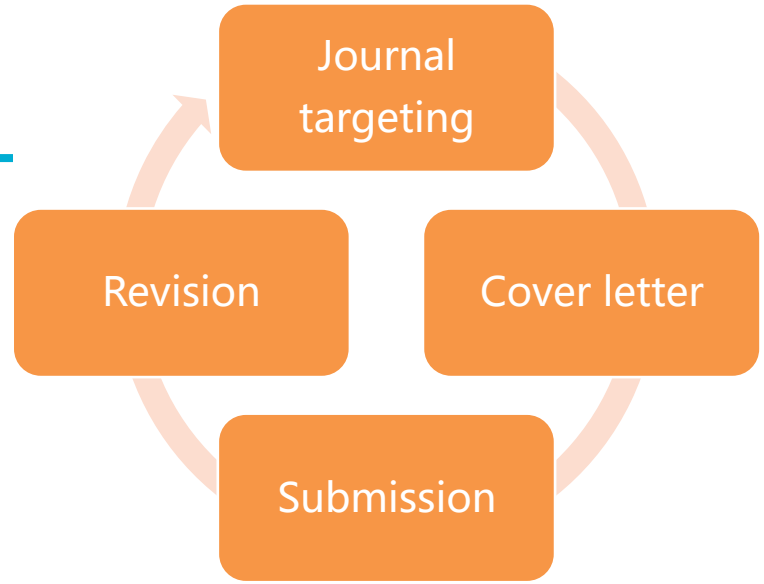
魏桐

2/22/2023

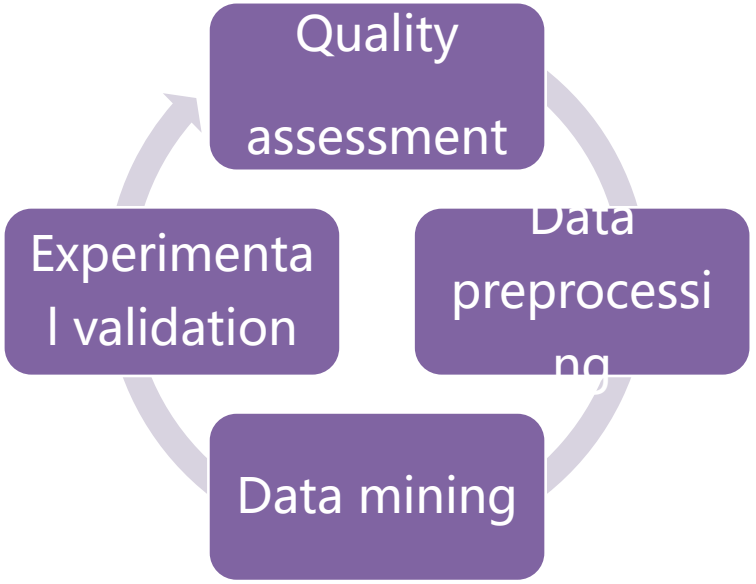
华大生命科学研究院
BGI · research

Research cycle





Project design → Data analysis → Manuscript preparation → Submission & revision



What is scientific reading?

- Scientific reading is
  - to **read and use** literature with **a full and critical understanding**,
  - while **addressing such questions** as content, context, objectives, and its interpretation and application.

BGI华大

How to read a paper

- Know **the structure**
  - Title page, Abstract
  - **Main text in IMRaD** : Introduction, Methods & Materials, Results, and Discussion
  - Other materials: supplementary/supporting materials, data and code, peer review information, etc.

- Understand **the logic** in,
  - Introduction
  - Results + Figures/Tables
  - Discussion

BGI华大

Methodology in reading a paper

- **Why** did they do it?
- **How** did they do it?
- **What** did they get?
- **So what** did it mean?

BGI华大

It applies to Abstract/Conclusion

- Why did they do it?
  - … is important in organ development/crop science; however, the mechanism remains unclear.
- How did they do it?
  - We carried out a multi-omics approach …
- What did they get?
  - The results showed that genes were associated with …
- So what did it mean?
  - Our work discovered key players and shed light on …

BGI华大

It also applies to Results/Discussion

- Why did they do it?
  - To investigate/reveal/study the mechanism of [your research], ..
  - To further explore the population structure, …
- How did they do it?
  - We conducted a transcriptomic approach …
  - We analyzed the SNPs from two populations …
- What did they get?
  - The results showed that genes were differentially expressed …
  - The population analysis revealed additional structure ..
- So what did it mean?
  - Our data indicate the transcriptomic reprogramming during …
  - The results suggest the domestication history of …

BGI华大

It connects sections

Introduction	→	WHY
Methods	→	HOW
Results	→	WHAT
Discussion	→	SO WHAT

Abstract

- **Background + question**
- **A sentence or phrases about methods**
- **Major discoveries**
- **Conclusion + significance**

Title page

- A **concise and meaningful** title with the emphasis on the main discoveries

- Author list
- Affiliations
- Corresponding author(s)

BGI华大

Title example

DONE →

FOUND →

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00831-0

Check for updates

Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce

Tong Wei[1,11], Rob van Treuren[2,11]✉, Xinjiang Liu[1,11], Zhaowu Zhang[1,3], Jiongjiong Chen[4], Yang Liu[1], Shanshan Dong[5], Peinan Sun[4], Ting Yang[1], Tianming Lan[1,6], Xiaogang Wang[7], Zhouquan Xiong[7], Yaqiong Liu[8], Jinpu Wei[8], Haorong Lu[8], Shengping Han[8], Jason C. Chen[8], Xuemei Ni[1], Jian Wang[1,9], Huanming Yang[1,9], Xun Xu[1,10], Hanhui Kuang[4], Theo van Hintum[2], Xin Liu[1]✉ and Huan Liu[1]✉

[1]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, China. [2]Centre for Genetic Resources, the Netherlands, Wageningen, the Netherlands. [3]BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China. [4]Huazhong Agricultural University, Wuhan, China. [5]Fairy Lake Botanical Garden, Shenzhen, China. [6]University of Copenhagen, Copenhagen, Denmark. [7]BGI-Laos, Vientiane, Laos. [8]China National GeneBank, Shenzhen, China. [9]James D. Watson Institute of Genome Sciences, Hangzhou, China. [10]Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen, China. [11]These authors contributed equally: Tong Wei, Rob van Treuren, Xinjiang Liu. ✉e-mail: robbert.vantreuren@wur.nl; liuxin@genomics.cn; liuhuan@genomics.cn

Abstract

- Structure
  - One sentence for **background**
  - One sentence for **scientific question**
  - Two or three sentences for **major discoveries**
  - One sentence for **conclusion and significance**

- Read Abstract as **a miniature article**
- Use as a **guideline**

BGI华大

Abstract example

Lettuce (*Lactuca sativa*) is an important vegetable crop worldwide. Cultivated lettuce is believed to be domesticated from *L. serriola*; however, its origins and domestication history remain to be elucidated. Here, we sequenced a total of 445 *Lactuca* accessions, including major lettuce crop types and wild relative species, and generated a comprehensive map of lettuce genome variations. In-depth analyses of population structure and demography revealed that lettuce was first domesticated near the Caucasus, which was marked by loss of seed shattering. We also identified the genetic architecture of other domestication traits and wild introgressions in major resistance clusters in the lettuce genome. This study provides valuable genomic resources for crop breeding and sheds light on the domestication history of cultivated lettuce.

BGI华大

Abstract example

WHY

HOW

WHAT

SO WHAT

1. Lettuce (*Lactuca sativa*) is an important vegetable crop worldwide.
2. Cultivated lettuce is believed to be domesticated from *L. serriola*; however, its origins and domestication history remain to be elucidated.
3. Here, we sequenced a total of 445 *Lactuca* accessions, including major lettuce crop types and wild relative species, and generated a comprehensive map of lettuce genome variations.
4. In-depth analyses of population structure and demography revealed that lettuce was first domesticated near the Caucasus, which was marked by loss of seed shattering.
5. We also identified the genetic architecture of other domestication traits and wild introgressions in major resistance clusters in the lettuce genome.
6. This study provides valuable genomic resources for crop breeding and sheds light on the domestication history of cultivated lettuce.

Introduction

- Structure
  - The 1st, 2nd, … paragraphs are background, which describes **the importance** of your area of study, and review **the major findings related** to this work
  - In the end of the 2nd last paragraph, raise **the scientific question(s)** in a logical way
  - The last paragraph state the **major discoveries**

- Read and understand the general background
- Pay attention on **how the questions are raised**
- Take a glance at the major findings

BGI华大

Introduction example: background

Lettuce (*Lactuca sativa* L.) is an important vegetable crop in the Compositae (also known as Asteraceae) family and is widely consumed as salad greens in many countries. Lettuce was first depicted on wall paintings of Egyptian tombs around 2,500 BC[1,2], making it one of the oldest known vegetable crops. It is believed that cultivated lettuce was domesticated from its progenitor *L. serriola*, and several hypotheses were proposed regarding the domestication center of lettuce, including Egypt, the Mediterranean area, the Middle East and Southwest Asia[1–3]. Modern lettuce varieties are classified based on morphological characteristics into leaf lettuce types (namely cos, butterhead, crisp, Latin and cutting) and non-leaf types (stalk and oilseed)[4]. Oilseed, mostly grown in Egypt for seed oil, is considered the most primitive type, while cos lettuce represents the predecessor of leaf types[2,5]. Despite the morphological variations, different lettuce types share common agronomic traits, such as entire leaf morphology, loss of seed shattering and an absence of spines along the leaf midvein, which are recognized as the domestication syndrome in cultivated lettuce[2].

BGI华大

Introduction example: question(s)

Advances in DNA sequencing technology make it feasible to study the genetic architecture in such germplasm collections. A previous RNA sequencing (RNA-seq) study of 240 lettuce accessions demonstrated that different crop types of cultivated lettuce were derived from a single domestication event[9]. However, the domestication history of cultivated lettuce and the genetic basis of human selection remain largely unknown.

BGI华大

Introduction example: findings

In this study, we sequenced 445 *Lactuca* accessions from 47 countries, comprising the major lettuce crop types and wild relative species. A comprehensive variation map, including 179 million single-nucleotide polymorphisms (SNPs), 30 million insertions/deletions (indels) and 244,866 structural variants, was constructed, from which we analyzed the phylogenetic relationship within the gene pool species and the domestication history of cultivated lettuce. The genetic architecture of domestication traits and introgression regions in resistance clusters were also identified. These sequencing results provide a valuable resource for lettuce research and breeding in the future.

BGI华大

Methods

- Organized by pipelines/experiments
- Read useful sections
- Pay attention **in details**, software, parameters, filtering criteria, etc

ARTICLES NATURE GENETICS

Methods

Plant materials and sequencing. The collection of *Lactuca* SSD lines (http://www.wur.eu/cgnsc002) used in this study comprises a core subset of the regular collection of the Centre for Genetic Resources, the Netherlands (CGN) and includes all crop types of cultivated lettuce and main wild relatives used in plant breeding[35]. The total study set of 445 SSD lines included 131 cultivated lettuce (*L. sativa*) accessions collected worldwide, 201 *L. serriola* accessions, 57 *L. saligna* accessions, 37 *L. virosa* accessions and 19 lines from another eight *Lactuca* species (Supplementary Table 1). Ten seeds were sown for each accession in December 2017 and transplanted in January 2018 in a greenhouse at BGI-Laos. Leaf samples were harvested from representative plants in March for genomic DNA extraction using the cetyl trimethylammonium bromide method[33]. Libraries with an insert size of 250 bp were constructed and paired-end reads (2 × 100 bp) were produced on a BGISEQ-500 platform at BGI-Shenzhen following the manufacturer's procedures[36]. All of the samples were sequenced with 20× depth except 12 wild species sequenced with higher depth for de novo assembly (Supplementary Table 2).

Genome assembly of wild *Lactuca* species. To construct genome assemblies for 12 wild *Lactuca* species, raw reads were filtered using Trimmomatic (version 0.27)[37] with the parameters ILLUMINACLIP:adapter.fa:2:35:4:12:true LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:50. The genome size for each species was estimated using KmerFreq (version 5.0)[38] with $K = 17$, and genome heterozygosity and repeat ratios were assessed using GCE (version 1.0.0)[39]. The DNA *C* values were retrieved from the Plant DNA C-values Database of Kew Gardens (https://cvalues.science.kew.org)[37]. For each species, genome assembly was run using SOAPdenovo (version 2.04) with ten different *k*-mer values, and the completeness was assessed by BUSCO (Supplementary Table 3). The assemblies with the highest BUSCO scores were selected for genome annotation with multiple pipelines, as previously reported[38]. Briefly, transposable elements were identified using a combination of homology-based and de novo approaches. RepeatMasker (version open-4.0.6)[39] and RepeatProteinMask (version 1.36)[39] were used to identify transposable elements with the known repeats from Repbase (release 25.03)[40] and custom repeat libraries annotated by RepeatModeler (version open-1.0.8)[41]. Tandem Repeats Finder (version 4.07b)[42] was used to find tandem repeats. All of the repeats identified by different approaches were masked before gene prediction. Ab initio prediction was carried out using AUGUSTUS (version 3.2.3)[43] and GeneMark (version 1.0)[44]. RNA-seq data obtained from the National Center for Biotechnology Information (NCBI) Sequence Read Archive database (details in Supplementary Table 3) were assembled into transcripts using Bridger (version r2014-12-01)[45]. The resulting transcripts and expressed sequence tag assemblies downloaded from PlantGDB[46] were aligned against the genome assemblies using BLASTN[47]. Homology-based prediction was performed by TBLASTN (evalue < 1×10^{-5}) using five published plant genomes downloaded from the NCBI (https://www.ncbi.nlm.nih.gov/), including *L. sativa* (version 8.0)[12], *H. annuus* (HanXRQr1.0)[48], *Cynara cardunculus* (CcrdV1)[48], *Arabidopsis thaliana* (TAIR10)[20] and *Solanum lycopersicum* (SL3.0)[21]. All evidences were combined into a final consensus gene set using MAKER (version 2.31.8)[52].

Phylogenetic inference of *Lactuca* species. To construct the phylogenetic relationship of the investigated *Lactuca* species, 4,513 single-copy genes were defined by OrthoMCL (version 5)[53] from the genome of an outgroup *H. annuus*, the lettuce reference genome and the genome assemblies of 11 wild *Lactuca* species, excluding the tetraploid species *L. canadensis*. TranslatorX (http://translatorx.co.uk/)[54] was used to translate DNA sequences into amino acids using the standard genetic code and to create amino acid alignments using MAFFT (version 5.0)[55], which was used as a guide for nucleotide sequence alignment after trimming ambiguous alignment portions using GBLOCKS (version 0.91b)[56]. Individual gene trees were constructed using RAxML (version 7.2.3)[57] with the GTR + GAMMA model, and the species tree was summarized by ASTRAL-III[58]. The Phyparts software (version 0.0.1; https://bitbucket.org/blackrim/phyparts) was used to demonstrate topological concordances and conflicts between individual gene trees and the species tree. Two additional phylogenetic trees were constructed

reads were filtered by Trimmomatic using the same parameter as for the genome assembly, and aligned to the *L. sativa* cv. Salinas reference genome (version 8.0)[12] using BWA-MEM with default parameters (version 0.7.12)[66]. Twelve wild accessions with higher sequencing depth were downsampled to about 20×. Five samples from four distantly related species, *L. canadensis*, *L. homblei*, *L. indica* and *L. palmensis*, were removed from variant calling because of the low mapping rates on the lettuce reference genome. The alignment bam files were then sorted and PCR duplicates were marked by MarkDuplicates, and HaplotypeCaller was run on each bam file in a genomic variant call format (GVCF) mode. The GVCF files from 440 accessions were consolidated into a single GVCF file, from which SNPs and small indels were identified using a joint calling approach. The SNPs and indels were further filtered using the following criteria: (1) SNPs were filtered with QD < 2.0 | FS > 60.0 | MQ < 40.0 | SOR > 3.0 | MQRankSum < −12.5 | ReadPosRankSum < −8.0', and indels with "QD < 2.0 | FS > 200.0 | SOR > 10.0 | MQRankSum < −12.5 | ReadPosRankSum < −8.0; (2) genotype calls with a depth <2 or >50, (3) variants with more than two alleles and (4) variants with a missing rate of >10% or a minor allele frequency (MAF) of <0.05 were removed, resulting in a set of 13 million filtered SNPs used for population genetic analyses; and (5) linkage disequilibrium pruning was performed with PLINK (version 1.9) using a window size of 10 kb with a step size of one SNP and an r^2 threshold of 0.5, resulting a set of 2.77 million pruned SNPs for clustering analysis. The variants from 332 *L. sativa* and *L. serriola* accessions were extracted from the quality-filtered VCF file and filtered with a missing rate of <10% and a MAF of >0.05, resulting in a set of 25.6 million SNPs for demography and GWAS.

Structural variant calling was performed on PCR-duplicate-marked bam files using three programs: Delly (version 0.8.1)[67], Manta (version 1.5.0)[68] and Breakdancer (1.3.6)[69]. Delly and Manta were run with the default parameters and structural variant calls with imprecise breakpoints were removed (flag IMPRECISE). BreakDancer was performed with the parameters -m 10000000 -q 30 -y 30 -r 2. Structural variants identified in each accession were integrated from different programs and then merged among all 440 accessions using SURVIVOR (version 1.0.7)[70]. The structural variants identified by at least two programs were kept for further analysis.

All of the variants were annotated using SnpEff (version 4.3r)[71]. SNPs, indels and structural variants were categorized based on their positions on the chromosome (including intergenic regions, exons, introns, splicing sites, untranslated regions and 1-kb upstream and downstream regions) and on their effects (including missense, start codon gain or loss, stop codon gain or loss and splicing mutations).

Population genetic analysis. PCA was performed on the filtered SNP set using GCTA (version 1.91.4beta3)[72]. A neighbor-joining tree was constructed with 100 bootstraps using PHYLIP (version 3.696)[73] and the tree layout was generated using the online tool iTOL (http://itol.embl.de). The population structure was analyzed with the cluster number *K* ranging from 1–20 by ADMIXTURE (version 1.3.0)[74] using a default fivefold cross-validation (--cv = 5). Each *K* was run with 20 replicates and the outputs were aligned by CLUMPP (version 1.1.2)[75]. Considering a cultivated ancestry of 20% ($K = 10$), we labeled one *L. sagittata*, two *L. dregeana* and seven *L. serriola* accessions as admixed samples (Supplementary Table 5).

Genetic differentiation (F_{ST}) and nucleotide diversity (π) were calculated within a nonoverlapping 100-kb window using VCFtools (version 0.1.13)[76]. For a given species, only biallelic SNPs with a missing rate of <0.1 and a MAF of >0.05 were used. Linkage disequilibrium was calculated on SNP pairs within a 500-kb window using PopLDdecay (version 3.31; https://github.com/BGI-shenzhen/PopLDdecay). Linkage disequilibrium decay measured the distance at which the Pearson's correlation efficient (r^2) dropped to half of the maximum. Singleton SNPs in each accession were calculated from the hard-filtering SNP set using VCFtools (version 0.1.13)[76], and the geographic distribution of singletons was assessed using the kriging regression function Krig implemented in the R package fields[77].

Demographic analysis of lettuce evolutionary history. *D* statistics[78] were calculated within a nonoverlapping 100-kb window to detect asymmetric gene

Method example

Read mapping and variant calling. Variant calling was carried out following the Genome Analysis Toolkit (GATK version 4.0.3.0) Best Practices[64,65]. Raw reads were filtered by Trimmomatic using the same parameter as for the genome assembly, and aligned to the *L. sativa* cv. Salinas reference genome (version 8.0)[12] using BWA-MEM with default parameters (version 0.7.12)[66]. Twelve wild accessions with higher sequencing depth were downsampled to about 20×. Five samples from four distantly related species, *L. canadensis*, *L. homblei*, *L. indica* and *L. palmensis*, were removed from variant calling because of the low mapping rates on the lettuce reference genome. The alignment bam files were then sorted and PCR duplicates were marked by MarkDuplicates, and HaplotypeCaller was run on each bam file in a genomic variant call format (GVCF) mode. The GVCF files from 440 accessions were consolidated into a single GVCF file, from which SNPs and small indels were identified using a joint calling approach.

BGI华大

Method example: details in pipeline

- Variant calling was carried out following the Genome Analysis ToolKit (GATK version 4.0.3.0) Best Practices [41]. Raw reads were filtered by Trimmomatic using the same parameter as in the genome assembly procedure, and aligned to the *L. sativa* cv. Salinas reference genome (v8.0, downloaded from http://genomevolution.org/coge)[14] using BWA mem with default parameters (version 0.7.12)[42].
- Twelve wild accessions with higher sequencing depth were downsampled to obtain a similar sequencing depth to the rest of the samples. The alignment bam files were then sorted and PCR duplicates were marked, and a GATK tool HaplotypeCaller was run on each bam file in a GVCF (genomic variant call format) mode. The gVCF files from 440 accessions were consolidated into a single gVCF file, from which single nucleotide polymorphisms (SNPs) and small indels were identified using a joint calling approach.

BGI华大

Results

- Organized under subtitles
- Read section by section
- Find and follow **the story line**, follow-up analyses, functional validation, etc.

into crisp lettuce detected here agrees with the pedigree record of using an *L. virosa* line to breed crisp varieties in the United States[16].

As little archeological evidence has been discovered for lettuce, we estimated the domestication time by assessing the change of effective population sizes (N_e) of *L. sativa* and *L. serriola*. Our result showed that both *L. sativa* and *L. serriola* experienced a gradual decline of N_e starting from 10,000 years ago, but the N_e contraction in *L. sativa* was stronger than that in *L. serriola* and continued until 2,000 years ago (Extended Data Fig. 5a,b). The N_e of six *L. serriola* groups and four *L. sativa* crop types displayed a similar pattern of population contraction and expansion to wild and cultivated lettuce, respectively (Supplementary Fig. 5). The divergence time between *L. sativa* and *L. serriola* was estimated to be around 6,000 years ago (Extended Data Fig. 5c,d), denoting the domestication of cultivated lettuce. All of these results indicate that lettuce was first domesticated near the Caucasus, and that the European *L. serriola* population played an important role in crop improvement after lettuce was spread to mainland Europe.

Human selection of domestication traits in lettuce. In addition to a reduced genetic diversity caused by population bottlenecks, domestication often results in purifying selection of genomic regions that control agronomic traits favored by humans. In search for the signature of selection, we scanned the lettuce genome using a cross-population composite likelihood ratio test (XP-CLR) and identified 4,089 selective sweeps covering 4.66% (107.7 Mb) of the assembled genome and harboring 2,304 genes in total (Fig. 3a and Supplementary Table 8). Genes involved in fatty acid and carbohydrate metabolism are enriched among those under selection, probably caused by human selection on nutritional values (Supplementary Table 9).

To better understand the impact of human selection on the lettuce genome, we carried out genome-wide association studies (GWAS) of the domestication traits in cultivated lettuce (Supplementary Table 10). As entire and lobed leaf morphology were both recorded in cultivated and wild lettuce, we analyzed the leaf morphology in *L. sativa* and *L. serriola* accessions separately. The major GWAS signals were detected within a 600-kb region on chromosome 3 in *L. sativa* (leading SNP $P = 2.45 \times 10^{-22}$), which overlapped with the major signals detected in the same region in *L. serriola* (leading SNP $P = 6.62 \times 10^{-30}$; Fig. 3c, Supplementary Fig. 6 and Supplementary Tables 11 and 12). Our results align with two previous studies using quantitative trait locus (QTL) mapping and bulked segregant analysis[17,18], which linked the same region to lobed leaf morphology. The nucleotide diversity of this interval was markedly reduced in cultivated lettuce and in butterhead, crisp and cos crop types (Fig. 3b). However, the nucleotide diversity in the cutting type was comparable to that in *L. serriola*, consistent with four out of 12 cutting accessions developing lobed leaves. We then carried out a phylogenetic analysis using the SNPs within this region to explore the underlying genetic architecture. The entire leaf lettuce accessions, including three primitive oilseed samples, formed a single clade and clustered with entire leaf *L. serriola* accessions from SEU (Supplementary

detected several major signals across the lettuce chromosomes (Fig. 3e and Supplementary Fig. 8a) and the signals on chromosome 6 agree with a previously identified QTL, *qSHT*, which explained 85% of the phenotypic variation of seed shattering[17]. The associated region was overlapped with an extended selective region spanning 9–23 Mb, within which the nucleotide diversity was markedly reduced in *L. sativa* compared with *L. serriola* (Fig. 3d). A transcription factor gene, encoding a homolog to *NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1* (*NST1*), which controls pod dehiscence in *Arabidopsis*[22], was found within this region as an intriguing candidate (Supplementary Fig. 8b,c). The phylogeny within this region showed that all of the cultivated lettuce accessions and nonshattering admixed samples formed a tight clade that is sister to the CAU group of *L. serriola* (Extended Data Fig. 6c). These results suggest that the loss of seed shattering in cultivated lettuce was probably caused by spontaneous mutation(s) derived from a wild ancestor near the Caucasus.

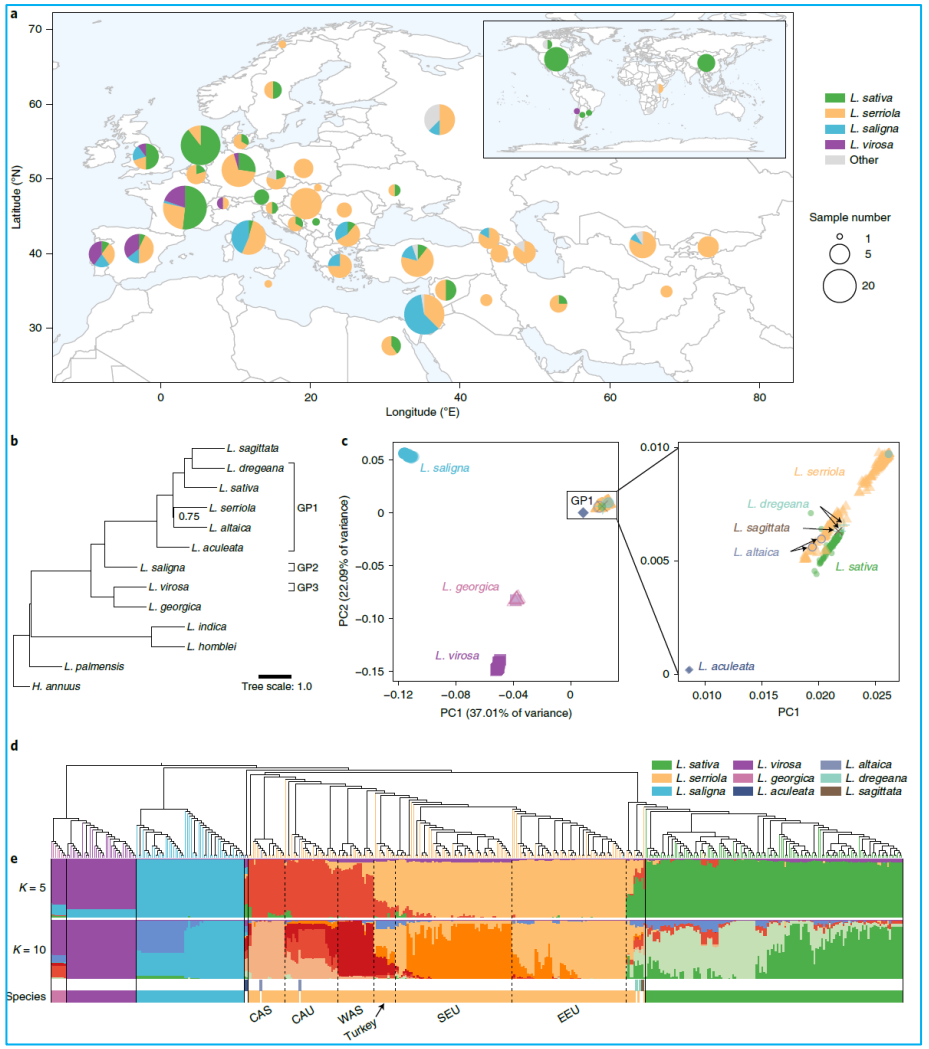
Wild plants develop spinose leaf structures such as spines and thorns as essential physical defense traits against herbivores, but these primitive traits in vegetables, especially leafy ones, are undesirable for human consumption. To identify the associated genomic regions, we carried out an association analysis in the investigated *L. serriola* samples that were recorded with or without spines in leaf midveins. The major GWAS signals were detected within 306.22–310.60 Mb on chromosome 5 (leading SNP with $P = 3.45 \times 10^{-29}$; Fig. 3g). Additionally, we sequenced two bulked F2 populations with or without leaf spines, which derived from a TKI-143 × GLHZ cross (an *L. serriola* accession with spines crossed with a cutting lettuce accession without spines; Fig. 3f). We calculated the differences in SNP indices (ΔSNP index) between the bulked samples and identified a single region overlapped with the GWAS signals on chromosome 5. Our result was also supported by a previously reported QTL, *qSPN*, which explained 82% of the phenotypic variation of spine presence[17]. Among the 70 candidate genes in this region, 21 were differentially regulated between *L. sativa* and *L. serriola*, which can be investigated further by genetic approaches (Supplementary Table 12). The phylogeny in this region suggests an early divergence from the CAU group of *L. serriola* (Extended Data Fig. 6d), although no selection signals were detected within it.

Identification of loci related to agronomic traits. Modern lettuce cultivars display diverse characteristics in many agronomic traits, such as flowering time, anthocyanin biosynthesis and leaf development. Among them, prolonged vegetative growth and delayed flowering have been recognized as important agronomic traits in cultivated lettuce. Our GWAS analysis detected a strong signal around 164.5 Mb on chromosome 7 ($P = 3.45 \times 10^{-14}$), where a *PHYTOCHROME C* (*PHYC*) gene resides (Extended Data Fig. 7). Two major haplotypes of *PHYC* were discovered in cultivated lettuce, and the accessions carrying the reference G allele displayed a significant delay in flowering date accompanied by reduced *PHYC* expression (Extended Data Fig. 7d and Supplementary Tables 12 and 14), resembling a wheat *PHYC* mutant that showed delayed

BGI华大

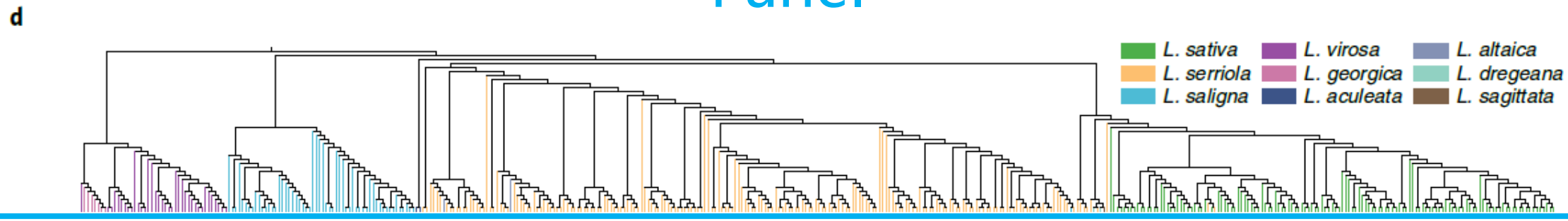
Result example

We explored the phylogenetic relationships among the 440 *Lactuca* accessions using 13 million high-quality SNPs. Our result showed that all of the *L. sativa* accessions formed a monophyletic clade, suggesting a single domestication event for cultivated lettuce (Fig. 1d). The phylogenetic positions of other wild species are consistent with the species tree, with most GP1 species having a close relationship with cultivated lettuce except that *L. georgica* was found close to the GP3 species *L. virosa*. Model-based clustering analysis revealed additional inter- and intraspecies relationships. Asia- and Europe-originated accessions formed two groups in both *L. serriola* and *L. saligna*, reflecting the spatial genetic variations within these species (Fig. 1e and Extended Data Fig. 2). Subgroups were further revealed in *L. serriola* when a higher number of *K* was assumed. These consisted of accessions from Central Asia, the Caucasus, Western Asia and Southern and Eastern Europe. Admixture between cultivated lettuce and wild species was detected in seven *L. serriola* accessions as well as two *L. dregeana* and one *L. sagittata* samples (Supplementary Table 5). We also observed admixture between Western Asian and Southern European compositions in Turkish accessions (Extended Data Fig. 2 and Supplementary Fig. 3). The phylogenetic relationships were revealed by the principal component analysis (PCA) as well, in which the GP1 species except *L. georgica* grouped together and *L. serriola* accessions were split into two distinct groups with Asian and European origins (Fig. 1c and Extended Data Fig. 3a,b).



Result example: figure + legend

Panel



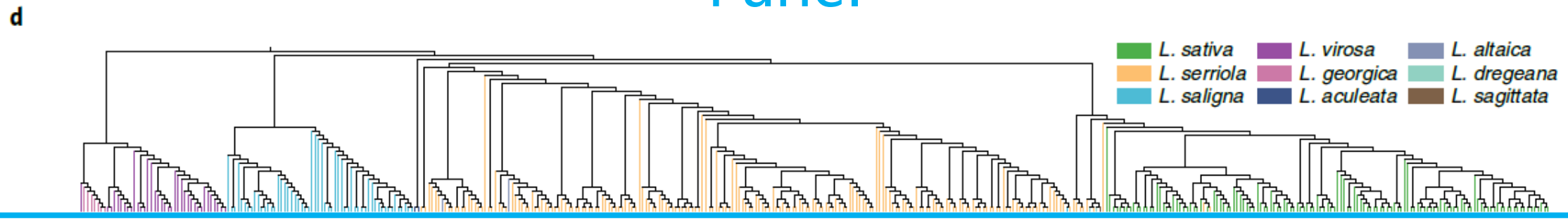
Legend

d, A neighbor-joining tree of 440 *Lactuca* accessions. Branch colors denote species.

BGI华大

Result example: figure + text

Panel



Result

1. We explored the phylogenetic relationships among the 440 *Lactuca* accessions using 13 million high-quality SNPs.
2. Our result showed that all of the *L. sativa* accessions formed a monophyletic clade,
3. suggesting a single domestication event for cultivated lettuce

BGI华大

Result example: logic within paragraph

We explored the phylogenetic relationships among the 440 *Lactuca* accessions using 13 million high-quality SNPs. Our result showed that all of the *L. sativa* accessions formed a monophyletic clade, suggesting a single domestication event for cultivated lettuce (Fig. 1d). The phylogenetic positions of other wild species are consistent with the species tree, with most GP1 species having a close relationship with cultivated lettuce except that *L. georgica* was found close to the GP3 species *L. virosa*. Model-based clustering analysis revealed additional inter- and intraspecies relationships. Asia- and Europe-originated accessions formed two groups in both *L. serriola* and *L. saligna*, reflecting the spatial genetic variations within these species (Fig. 1e and Extended Data Fig. 2). Subgroups were further revealed in *L. serriola* when a higher number of *K* was assumed. These consisted of accessions from Central Asia, the Caucasus, Western Asia and Southern and Eastern Europe. Admixture between cultivated lettuce and wild species was detected in seven *L. serriola* accessions as well as two *L. dregeana* and one *L. sagittata* samples (Supplementary Table 5). We also observed admixture between Western Asian and Southern European compositions in Turkish accessions (Extended Data Fig. 2 and Supplementary Fig. 3). The phylogenetic relationships were revealed by the principal component analysis (PCA) as well, in which the GP1 species except *L. georgica* grouped together and *L. serriola* accessions were split into two distinct groups with Asian and European origins (Fig. 1c and Extended Data Fig. 3a,b).

1. Fig. 1d: Our result showed that all of the *L. sativa* accessions formed a monophyletic clade, … (Fig. 1d).
2. Fig. 1e: Model-based clustering analysis revealed additional inter- and intraspecies relationships. Asia- and Europe-originated accessions formed two groups … (Fig. 1e and Extended Data Fig. 2).
3. Fig. 1c: The phylogenetic relationships were revealed by the principal component analysis (PCA) as well, …(Fig. 1c and Extended Data Fig. 3a,b).

Result example: logic between sections

Fig. 1. Population structure of *Lactuca* accessions

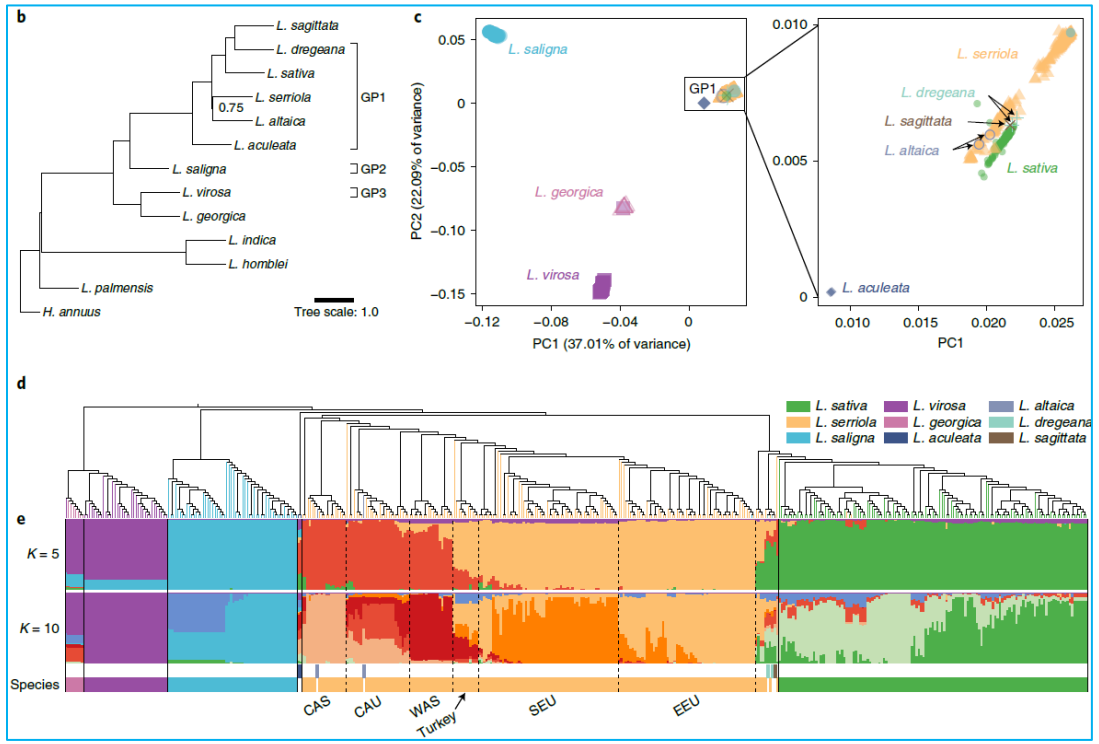
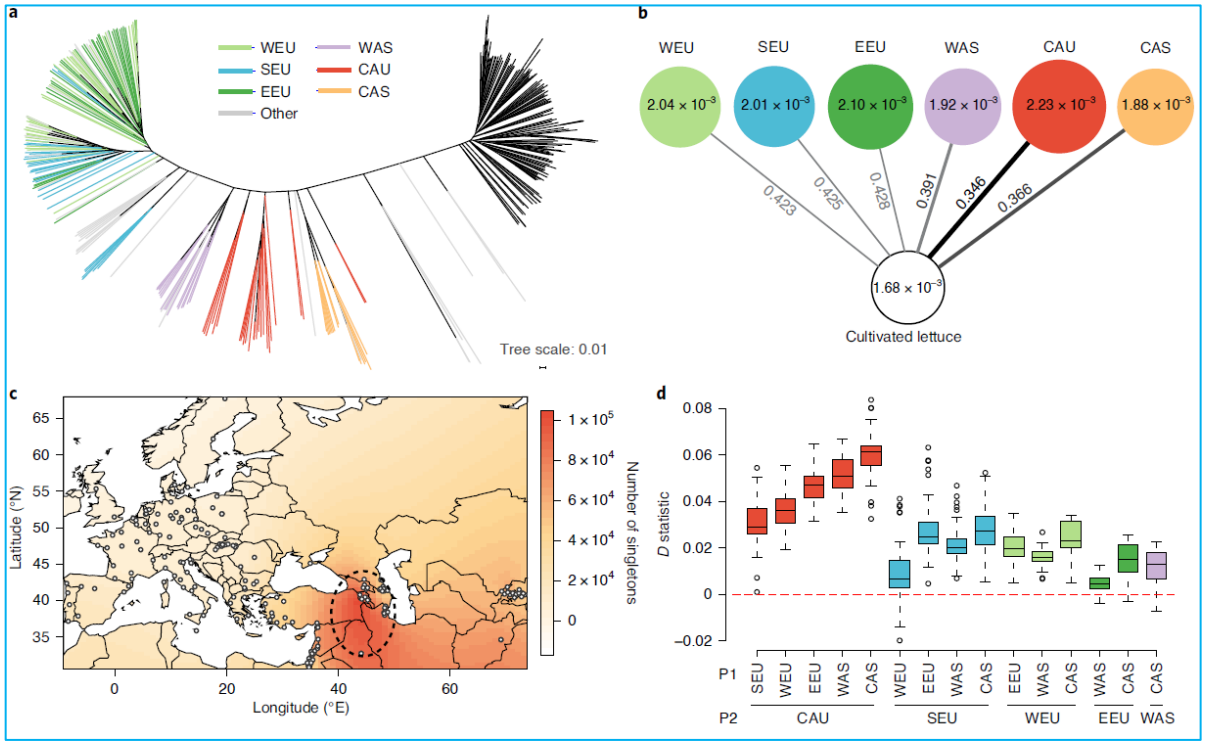


Fig. 2. Lettuce domestication center near the Caucasus



Discussion

- Structure
  - The 1st paragraph, restates **the major findings.**
  - The 2nd, 3rd, and …, discuss in details **topic by topic**, and bring out **the significance in the context.**
  - The last one ends with a **conclusion.**

- Read topic by topic
- Pay attention to the interpretation of results **in the context**

BGI华大

Discussion example: major findings

Discussion

In this work, we analyzed the genome sequences of 445 *Lactuca* accessions representing lettuce crop types and its wild gene pool species. More than 208 million sequence variants were identified, from which we revealed the population structure of the genebank collection and the domestication history of cultivated lettuce.

BGI华大

Discussion example: one topic

WHY

As germplasms of major crops are maintained as genebank collections, understanding the population structure and phylogenetic relationships is of great importance for genebank management and utilization. In lettuce breeding, the GP1 species are used widely as

WHAT

there is no reproductive barrier within the group[6,7]. Our phylogenetic analyses clarified several issues regarding the taxonomic status of these GP1 species (see the Supplementary Note for a detailed discussion). First, the presumed GP1 species *L. georgica* should be reassigned as it clustered with the GP3 species *L. virosa*. The *L. dregeana* and *L. sagittata* samples are not to be considered as true wild species. Another GP1 species, *L. altaica*, has been considered as conspecific with *L. serriola*[29,30], but the plastid phylogeny implied an introgression and fixation of a distantly related plastid haplotype in *L. altaica*. Phylogenetic analyses with additional samples will clarify these taxonomic issues in wild species. Our study also pointed out future directions in germplasm collection and utilization. Among the investigated samples, *L. serriola* from the Caucasus represents the most promising resource because the population from this area showed the highest nucleotide diversity. *L. aculeata* represents another potentially important gene pool, as its phylogenetic position distinct from other GP1 species suggests a different genetic repertoire. Thus, our study provided new insights regarding accession identity and genetic resources for crop improvement, demonstrating the value of whole-genome sequencing in the management of crop collections and the utilization thereof.

HOW

SO WHAT

BGI华大

Discussion example: conclusion

Overall, our study constructed phylogenetic relationships within lettuce gene pool species and revealed the genetic basis of human selection during lettuce domestication. The genome sequences and the variation map generated in this study will serve as a valuable resource for lettuce research and breeding in the future.

BGI华大

Other materials

- Extended Data
- Supplementary information

- Source Data
- Data and code availability

- Peer review information
- Accepted time

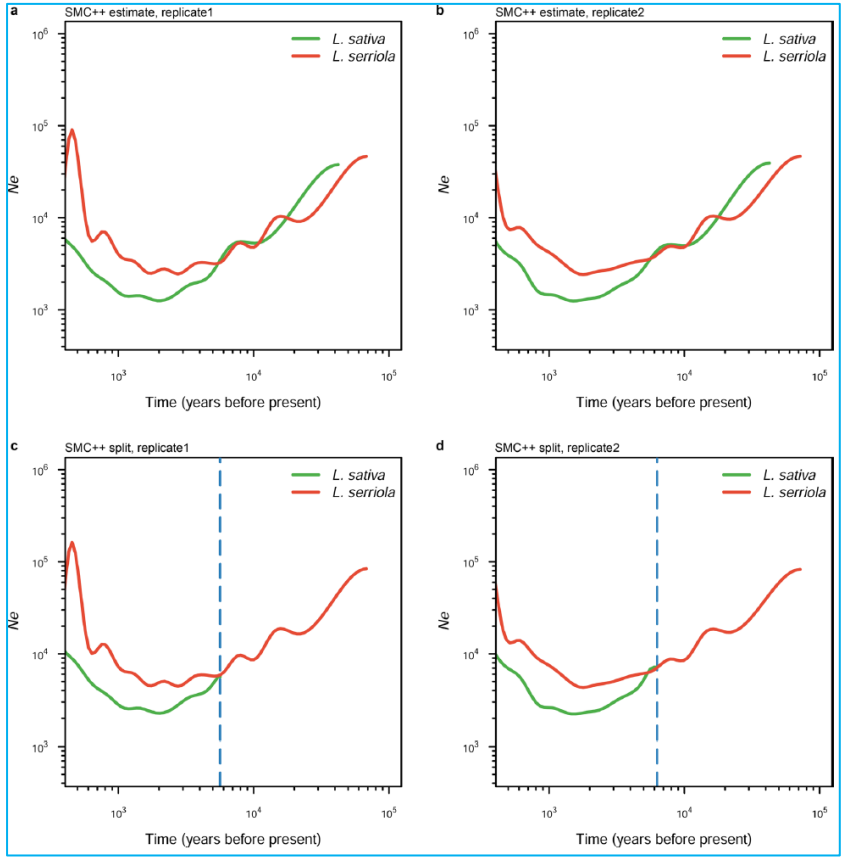
Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00831-0.

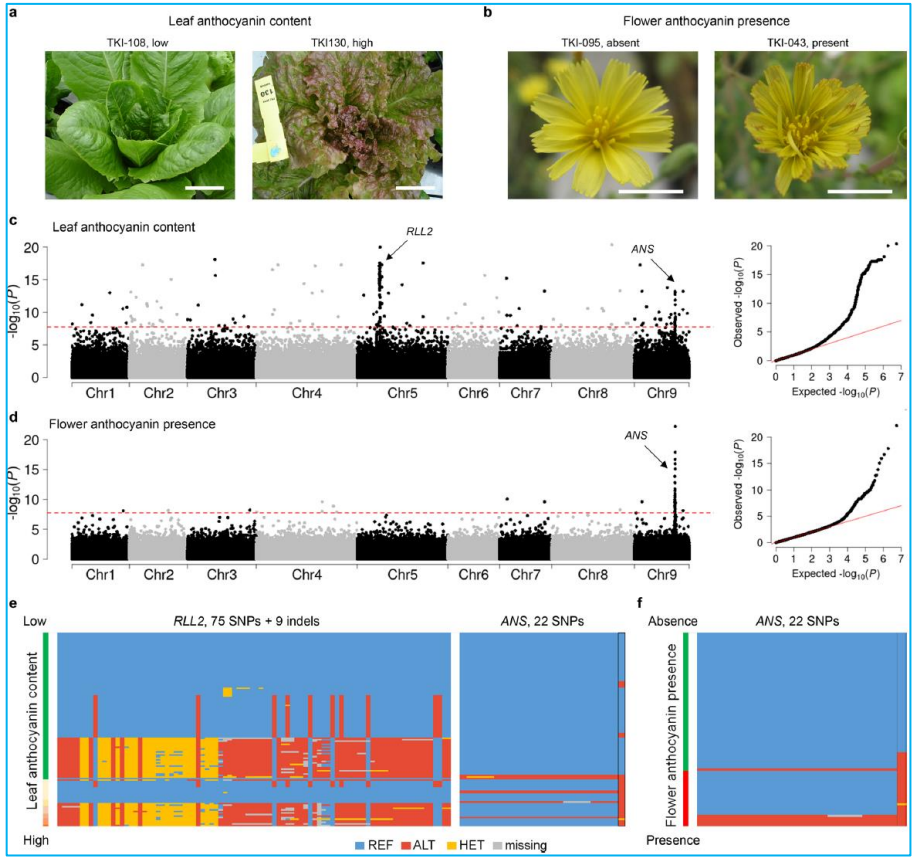
Received: 26 December 2019; Accepted: 1 March 2021;
Published online: 12 April 2021

Extended Data example

Extended Data Fig. 5



Extended Data Fig. 8



Supplementary Information example

Supplementary Note

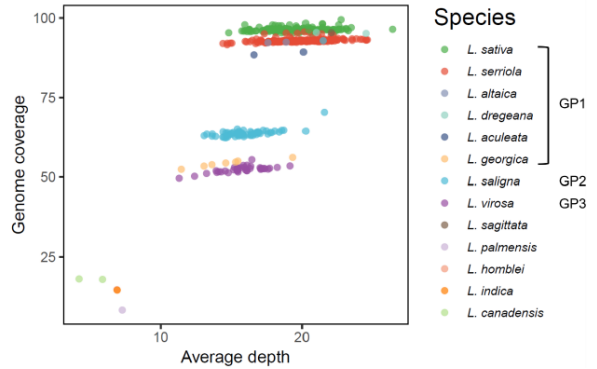
Supplementary Note

1. Lettuce Gene Pool species

Wild relative species are commonly used as a source of novel traits in lettuce breeding, such as in variety improvement for biotic and abiotic stress and for quality characters. The International Plant Name Index (IPNI) includes several hundreds of species of the genus *Lactuca*, but the majority of these taxa refer to synonyms and basionyms or have been reassigned to other genera[1]. Generally, the *Lactuca* genus is considered to consist of about 100 species, of which approximately 20 are part of the lettuce gene pool[2,3]. The primary gene pool (GP1) consists of completely inter-fertile taxa, including the crop species *L. sativa* L. and its wild relatives *L. aculeata* Boiss. & Kotschy, *L. altaica* Fisch. & C.A. Mey., *L. azerbaijanica* Rech. f., *L. dregeana* DC., L. georgica Grossh., *L. scarioloides* Boiss. and *L. serriola* L. The secondary gene pool (GP2) is formed by *L. saligna* L. alone, while the tertiary gene pool (GP3) includes *L. acanthifolia* (Willd.) Boiss., *L. alpestris* (Gand.) Rech. f., *L. aurea* (Vis. & Pančić) Stebbins, *L. longidentata* DC., *L. orientalis* (Boiss.) Boiss., *L. quercina* L., *L. sibirica* (L.) Benth. ex Maxim., *L. tatarica* (L.) C.A. Mey., *L. viminea* (L.) J. Presl & C. Presl, *L. virosa* L. and *L. watsoniana* Trel. *L. serriola*, *L. saligna* and *L. virosa* are the main species that have been extensively used in breeding.

Supplementary Figures

Supplementary Figures



Supplementary Fig. 1. Average sequencing depth and genome coverage of 445 *Lactuca* accessions on the lettuce reference genome. The primary (GP1), secondary (GP2), and tertiary gene pool (GP3) species are indicated in the legend.

Source Data example

Source Data

Source data

Source Data Fig. 1

Statistical source data.

Source Data Fig. 2

Statistical source data.

Source Data Fig. 3

Statistical source data.

Source Data Fig. 4

Statistical source data.

Source Data Extended Data Fig. 1

Statistical source data.

Source Data Extended Data Fig. 2

Statistical source data.

Source Data Extended Data Fig. 3

Statistical source data.

Source Data Fig. 1

	A	B	C	D	E	F	G
	Country	L. sativa	L. serriola	L. saligna	L. virosa	other	Total number
	Afghanistan	0	2	0	0	0	2
	Argentina	1	0	0	0	0	1
	Armenia	0	4	0	0	0	4
	Austria	3	0	0	0	0	3
	Azerbaijan	0	6	0	0	1	7
	Belgium	1	4	0	0	0	5
	Bulgaria	1	5	3	0	0	9
	Canada	1	0	0	0	1	2
	Chile	0	0	0	1	0	1
	China	7	0	0	0	0	7
	Croatia	1	2	0	0	0	3
	Czech Republic	1	3	0	0	1	5
	Denmark	1	2	0	0	0	3
	Egypt	2	3	0	0	0	5
	France	29	15	1	11	0	56
	Georgia	0	5	1	0	0	6
	Germany	6	15	0	1	0	22
	Greece	0	6	2	0	0	8
	Hungary	0	17	0	0	0	17
	Indonesia	0	0	0	0	1	1
	Iran	1	3	0	0	0	4
	Iraq	0	2	0	0	0	2
	Israel	0	15	24	0	1	40
	Italy	1	12	10	0	0	23

Fig1a | Fig1b | Fig1c | Fig1d | Fig1e

BGI华大

Data and Code Availability

Data availability

All raw sequencing data were deposited into the Sequence Read Archive (under BioProject accession PRJNA693894) and CNGB Nucleotide Sequence Archive (CNSA; under the accession number CNP0000335). Variant files, genome assemblies and annotation files are stored in CNSA under the same accession number. Source data are provided with this paper.

Code availability

All of the code used in this study is available at https://github.com/popgenome/lettuce2020.

Tips for reading

- Find and follow **the logic/story line**
- Pay attention in **details** in figures, tables, and methods
- Ask self the four questions in main text

- Make notes

BGI华大

Note example: methods

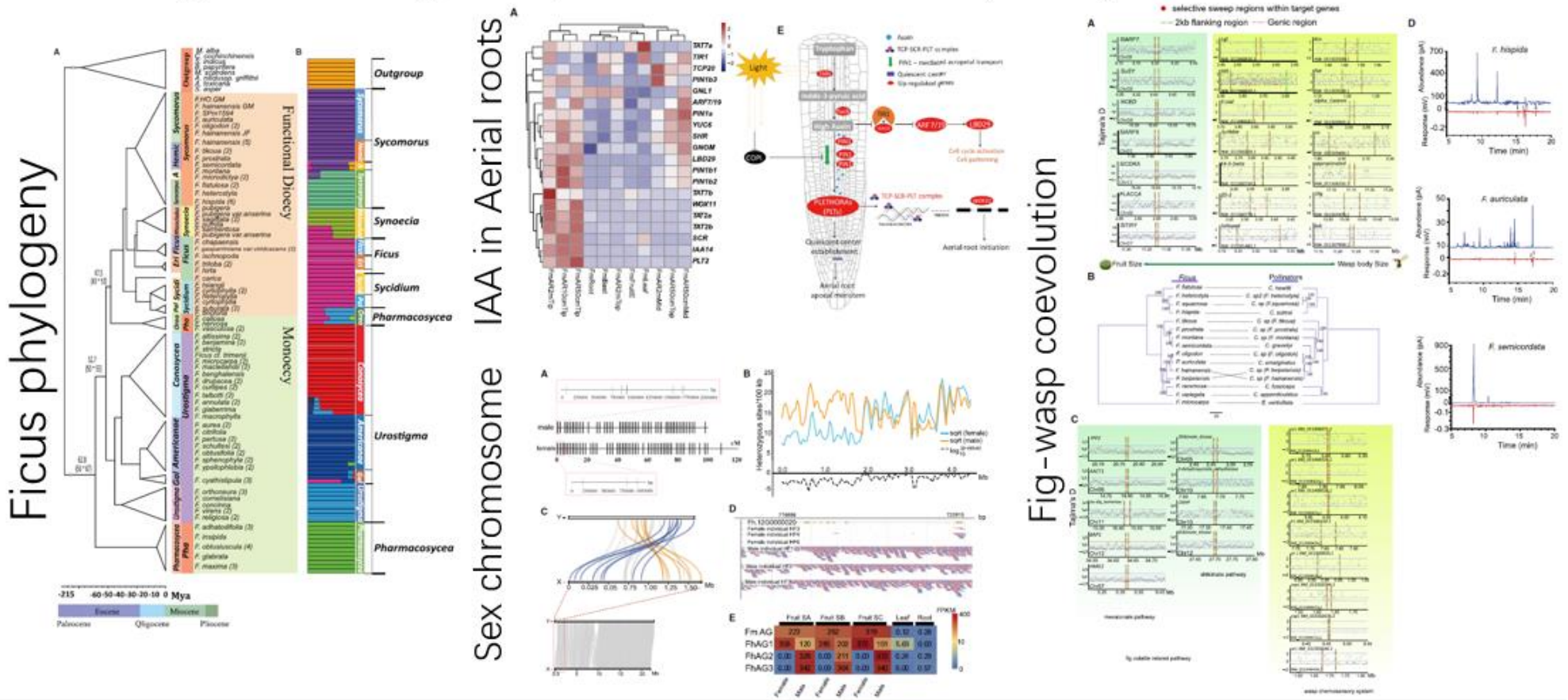
Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. Cell (2020) 183:1

- Library: 63 PacBio SMRT cells (30 for Ficus macrocarpa and 33 for F. hispida) sequenced on Pacbio RSII; paired-end libraries sequenced on Illumina X10 with 150-bp cycles and 300-500 bp insert size, Hi-C libraries sequenced on Illumina X10 (80% valid reads evaluated by HiC-Pro); E. verticillata sequenced in 170x on PacBio and 84x on Illumina; RNA libraries sequenced on Illumina HiSeq 2500 or X10.
- Assembly: PacBio reads self-corrected by CANU; assembled by CANU, FALCON, SMARTdenovo and evaluated by N50/size/BUSCO; merge by Quickmerge; assemblies polished by Pilon; uniquely-mapped Hi-C reads corrected by 3D-DNA pipeline and used for scaffolding by ALLHiC; wasp genome assembled by CANU and redundancy removed by Redundans; evaluated by BUSCO, PASA transcripts, Illumina reads, chromatin interactions, and genetic maps based on resequencing of 30 male and female.
- Annotation: a repeat library constructed by RepeatModeler; TEs identified by RepeatMasker, tandem repeats by TRF; TE classified by TEclass; annotation using MAKER pipeline, including training on Trinity/PASA assembled transcripts by SNAP, GENEMARK, AUGUSTUS, a second round of training with AED<0.2 gene models, and combined with HISAT/StringTie assembled transcripts and homologous proteins from 6 species; miRNA predicated by mapping miRBase miRNAs to two genomes by bowtie.
- Comparative genomics: segmental duplication identified by the 1st round of blast of 400-kb unique segments with at least 88% identity and 500 bp and a 2nd round alignment of global alignments at least 90% identity and 1 kb; SVs identified by MUMmer alignment of MaSuRCA assembled contigs and a web-based tool Assemblytics; CNVs significantly different from genome average depth identified by count Illumina reads in 5-kb non-TE sequence.
- Phylogeny: concatenated tree constructed from MUSCLE alignment of single-copy genes 5 species; ML tree constructed on SNPs by IQ-Tree and RAxML/GTRCAT model; divergence time estimated by PAML/MCMCTREE; co-phylogeny analyzed by Jane4, codivergence analyzed by Tajima's D in 14 fig-wasp pair; selective sweep by SweeD.
- Population: ancestry identified by ADMIXTURE; coefficient of genetic relatedness calculated within a 200-kb sliding window using IBD; ABBA analyzed by ANGSD/doAbbababa2.
- Hormone: 50 mg fresh tissues ground in liquid nitrogen, extracted in 0.5 mL methonal/water/formic acid, evaporated in nitrogen gas, reconstituted in 80% methonal, ultraphoniced and filtered for LC-ESI-MS/MS.
- Sex chromosome: phased SNPs called from minimap2 alignment of corrected PacBio alignment by GATK4 with error correction by WhatsHap; SNPs from 40x F. hispida WGS data used to determine sex-phased blocks with at least 70% of sex-specific SNPs; sex chromosome de novo assembly by CANU and scaffolded by ALLHiC; sex-determining regions identified based on heterozygous site density from 15 male and 11 female; expression called from RNAseq from flowers at 3 stages by bowtie-RSEM pipeline.
- Pollinator attraction: odor from fig syconia at receptive phase collected by solid-phase micro-extraction over the headspace for 1h and profiled by GC-MS; stimulus to wasps detected by gas chromatography-electroantennogram detection (GC-EAD) with 9 replicates.

Personal note

Note example: results

Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. Cell (2020) 183:1



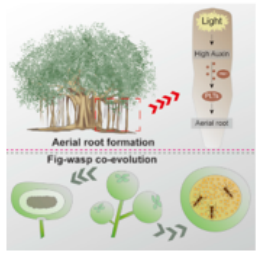
Personal note

BGI华大

Note example: in-depth reading

Background: banyan tree from *Ficus* genus

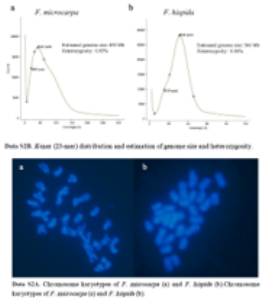
- Ecologically important: provide year-round fruit production in tropical forests
- The largest genus with ~800 species in the Moraceae family (桑科)
- Hemi-epiphytic habit: aerial roots
- A diverse sex determination system
- Fig-wasp obligate mutualism



4

Genome survey

- Genome size, 366 Mb for F. hispida, 430 Mb for F. microcarpa, estimated from Illumina short reads using perl scripts
- Nuclear DNA estimated by flow cytometry analysis



5

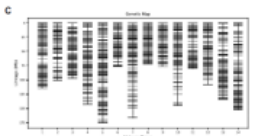
Genome assembly of F. microcarpa, F. hispida and E. verticillata

- 63 PacBio SMRT cells (30 for Ficus macrocarpa and 33 for F. hispida) sequenced on Pacbio RSII; Paired-end libraries sequenced on Illumina X10 with 150-bp cycles and 300-500 bp insert size; Hi-C libraries sequenced on Illumina X10 (80% valid reads evaluated by HiC-Pro)
- RNA libraries sequenced on Illumina HiSeq 2500 or X10.
- PacBio reads self-corrected by CANU; assembled by CANU, FALCON, SMARTdenovo
- Evaluated by N50/size/BUSCO; merge by Quickmerge; assemblies polished by Pilon;
- Uniquely-mapped Hi-C reads corrected by 3D-DNA pipeline and used for scaffolding by ALLHiC.

6

Assembly validation by genetic map

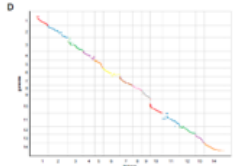
- SNPs called from resequencing of 30 male and 30 female F1 individuals by GATK
- Maternal and paternal genetic maps determined from bin markers



7

Assembly evaluation

- Completeness assessed by BUSCO on 1,440 conserved plant genes
- PASA assembled transcripts mapped to two assemblies
- Illumina reads mapped by bwa
- Genome-wide chromatin interaction



8

Genome annotation

- A repeat library constructed by RepeatModeler; TEs identified by RepeatMasker, tandem repeats by TRF; TE classified by TEclass
- Annotation using MAKER pipeline, including training on Trinity/PASA assembled transcripts by SNAP, GENEMARK, AUGUSTUS, a second round of training with AED<0.2 gene models
- Combined with HISAT/StringTie assembled transcripts and homologous proteins from 6 species.

9

Personal note

BGI华大

Summary for research articles

- Abstract is the essence of the paper

- Introduction sets the stage
- Methods provides details to repeat the work
- Results describe the data and new findings
- Discussion interpretates in the context

BGI华大

Other article types

- Review
- Technical advances
- Data notes
- Editorial, Perspective, Comments, etc.

- Resources: PUBMED, google scholar, sci-hub

Questions?

华大生命科学研究院
BGI · research