

RESEARCH

Open Access



Lactuca super-pangenome reduces bias towards reference genes in lettuce research

Dirk-Jan M. van Workum¹, Sarah L. Mehrem², Basten L. Snoek², Marrit C. Alderkamp³, Dmitry Lapin³, Flip F. M. Mulder⁴, Guido Van den Ackerveken³, Dick de Ridder¹, M. Eric Schranz⁵ and Sandra Smit^{1*}

Abstract

Background Breeding of lettuce (*Lactuca sativa* L.), the most important leafy vegetable worldwide, for enhanced disease resistance and resilience relies on multiple wild relatives to provide the necessary genetic diversity. In this study, we constructed a super-pangenome based on four *Lactuca* species (representing the primary, secondary and tertiary gene pools) and comprising 474 accessions. We include 68 newly sequenced accessions to improve cultivar coverage and add important foundational breeding lines.

Results With the super-pangenome we find substantial presence/absence variation (PAV) and copy-number variation (CNV). Functional enrichment analyses of core and variable genes show that transcriptional regulators are conserved whereas disease resistance genes are variable. PAV-genome-wide association studies (GWAS) and CNV-GWAS are largely congruent with single-nucleotide polymorphism (SNP)-GWAS. Importantly, they also identify several major novel quantitative trait loci (QTL) for resistance against *Bremia lactucae* in variable regions not present in the reference lettuce genome. The usability of the super-pangenome is demonstrated by identifying the likely origin of non-reference resistance loci from the wild relatives *Lactuca serriola*, *Lactuca saligna* and *Lactuca virosa*.

Conclusions The super-pangenome offers a broader view on the gene repertoire of lettuce, revealing relevant loci that are not in the reference genome(s). The provided methodology and data provide a strong basis for research into PAVs, CNVs and other variation underlying important biological traits of lettuce and other crops.

Keywords Lettuce, *Lactuca sativa*, Pangenomics, Super-pangenome, PAV-GWAS

Background

Garden lettuce (*Lactuca sativa*) plays an important role in agriculture worldwide. In 2020, worldwide production of lettuce (including chicory) was about 27 billion tonnes, with a total value of about 20 billion US dollars [1]. Because of its size and easy cultivation, lettuce is the perfect candidate for new farming techniques such as vertical farming, hydroponics and light-emitting diode (LED) illumination. In addition, lettuce is suitable for modern breeding techniques such as genome editing by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas [2] and has many economically important close relatives within the Asteraceae such as

*Correspondence:

Sandra Smit
sandra.smit@wur.nl

¹Bioinformatics Group, Wageningen University & Research, Droeendaalsesteeg 1, Wageningen 6708 PB, the Netherlands

²Theoretical Biology & Bioinformatics, Utrecht University, Padualaan 8, Utrecht 3584 CH, the Netherlands

³Translational Plant Biology, Utrecht University, Padualaan 8, Utrecht 3584 CH, the Netherlands

⁴Department of Genetics, University Medical Center Utrecht, Heidelberglaan 100, Utrecht 3584 CX, the Netherlands

⁵Biosystematics Group, Wageningen University & Research, Droeendaalsesteeg 1, Wageningen 6708 PB, the Netherlands



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

endive, chicory, sunflower and chrysanthemum. Thus, lettuce has the potential to become a model organism for all Asteraceae.

For accurate, contemporary and inter-specific research into lettuce, a complete understanding of the lettuce genome is needed. Currently, a chromosome-level genome assembly is available for *L. sativa* [3] as well as whole-genome sequencing (WGS) data for hundreds of *Lactuca* accessions [4]. Lettuce genomic research has been centred around a single reference genome assembly for a long time, thereby likely introducing so-called “reference bias”. Reference bias is caused by ignoring differences in genomic composition and rearrangements between analysed accessions and the reference accession. Of particularly large impact are gene presence/absence variation (PAV), gene copy-number variation (CNV) and other large structural variations. Absence of regions in a reference assembly due to such variations can result in missing relevant information or incorrect genotyping. Moreover, it is known that variable genes (*i.e.* genes present in only a subset of accessions of a species) are enriched for functions important to agronomic traits such as disease resistance [5–10]. This results in a strong need for expanding our vision beyond the single reference genome to get an integrated overview of genetic variation in both cultivated and wild *Lactuca* accessions. Pangenomics – the genetic study of closely related individuals first introduced by [11] – does this by taking all available genomic variation among related individuals into account.

Similarly to other crops, the availability of large-scale sequencing efforts enables an integrated analysis of genetic variation among both cultivated and wild relatives of lettuce. This allows for the exploitation of genetic variation across species boundaries for the improvement of traits of interest. So far, only single-nucleotide polymorphisms (SNPs) and InDels with respect to *L. sativa* were analysed [4, 12], whereas the wealth of WGS data allows for a more thorough investigation of PAV and CNV across species. Traditionally, wild relatives of a cultivated crop species have been classified into three gene pools based on their ability to cross with it [13]. *Lactuca serriola*, being the evolutionary predecessor of *L. sativa*, belongs to the primary gene pool for lettuce breeding since there are no limitations for crossing with *L. sativa* [14, 15]. The secondary gene pool contains *Lactuca saligna*, which produces partly fertile offspring with *L. sativa*. Finally, the tertiary gene pool contains *Lactuca virosa*, which can only be crossed with *L. sativa* under specific circumstances [14, 16]. For both *L. saligna* and *L. virosa*, near chromosome-level genome assemblies became available recently [17, 18]. All of this data allows for the analysis of a lettuce super-pangenome, a pangenome of both cultivated and wild accessions from

different species [19]. Since its introduction, this concept has already been successfully applied to rice, lentil, maize, soy and tomato [20–24].

Here, we describe the analysis of a 474-accession *Lactuca* super-pangenome comprising variation within the three gene pools available for lettuce breeding. As genes are the building blocks of the genome, we focus on the accurate identification of all genic PAVs and CNVs in accessions that have publicly available WGS data as well as newly generated WGS data for *L. sativa*. We built linear pangenomes per species for obtaining a full inventory of genes before identifying PAV and CNV. PAVs specifically allow for the biological interpretation of the core and variable genome of *Lactuca* species. Next, we show how PAV and CNV information can be used in genome-wide association studies (GWAS) to study genotype-phenotype associations, extending the traditionally used SNP-based analysis [10, 25]. Illustrating the vast amount of information in this super-pangenome resource, we describe a few key observations such as the make-up of the core and variable *Lactuca* genome, the use of PAVs to study the evolution of *Lactuca*, and non-reference resistance loci to *Bremia lactucae*.

Methods

Data selection

We selected a total of 474 lettuce accessions (belonging to *L. sativa*, *L. serriola*, *L. saligna* and *L. virosa*) for analysis (Supplementary Methods – Data selection). Both publicly available and in-house WGS data were used as well as nuclear, mitochondrial and chloroplast genomes for each species (Supplementary Table S1). The accession numbers of the 406 WGS data sets used from [4] can be found in Supplementary Table S2 (column ‘BGI sequencing’). Importantly, the *L. sativa* reference genome was used as reference genome for *L. serriola*. Only for *L. virosa*, a different accession was used for its mitochondrial genome, since none was available for the reference accession and only little sequence variation exists within lettuce mitochondrial genomes [26].

Sequencing

We grew and sequenced 68 accessions that are not part of the WGS data published by [4] to increase the resolution of our dataset (Supplementary Table S2; column ‘LK sequencing’). Seeds were obtained from the Centre for Genetic Resources, the Netherlands (CGN) and multiplied by single-seed descent, except for *L. saligna* 275-5 (PV15242) which was multiplied by regular propagation. Plants were grown on rockwool cubes under long day conditions (21°C/19°C, 16 h light, ~100 $\mu\text{mol}/\text{m}^2/\text{sec}$). The plants received 1x Hyponex on the day of sowing, 0.5x Hyponex on day 7, and tap water on the other days. Samples of the first and second true leaves were collected

at 17 days in the form of two 10 mm leaf discs for two plants per accession. These were snap-frozen in liquid nitrogen and stored at -80°C . DNA extraction was performed with MagMAX™ plant DNA isolation kit using a Kingfisher robot. Library preparation was performed using the Truseq DNA nano protocol and paired-end sequencing (2×150 bp) was performed on the Illumina NovaSeq6000 machine with an S4 flowcell at Utrecht Sequencing Facility (USEQ).

Linear pangenome construction

A custom pipeline was built based on the “iterative map and build” approach [5], partly inspired by the methods described by [27] and [28] (Fig. 1, Supplementary Methods – Pipeline construction strategy). After trimming the input paired-end Illumina reads with Trimmomatic v0.39 [29], each dataset was mapped to its closest reference genome with bwa-mem2 v2.1 [30]. Using samtools v1.11 [31], the alignments were sorted, merged and indexed for extracting all unmapped reads per accession. These unmapped reads were assembled per accession with

MEGAHIT v1.2.9 [32]; in the resulting assemblies, contigs with a length below 1kbp, a GC percentage higher than 50% or a kraken2 v2.1.2 [33] classification of Bacteria, Arthropoda and Amoebozoa were removed. This mapping/assembly stage was performed in parallel to speed up processing of hundreds of samples.

Next, a linear pangenome was built by iteratively adding the newly assembled sequences to the reference genome. Whether a new sequence was already part of the pangenome was determined by performing a mini-map2 v2.22 (parameters ‘-cxasm5 --secondary=no’) [34] alignment of the new genomic content per accession to the previous pangenome build. Whenever an alignment was found, only the longer sequence was retained. Then, all new sequences in the pangenome were clustered with CD-HIT-EST v4.8.1 [35] at 90% similarity to further remove possible duplicates. Finally, blobtools v1.1.1 was run per set of novel sequences in a linear pangenome according to standard instructions and based on its output we only kept contigs with at least 10X coverage and either “no-hit” or “Steptophyta” taxonomy.

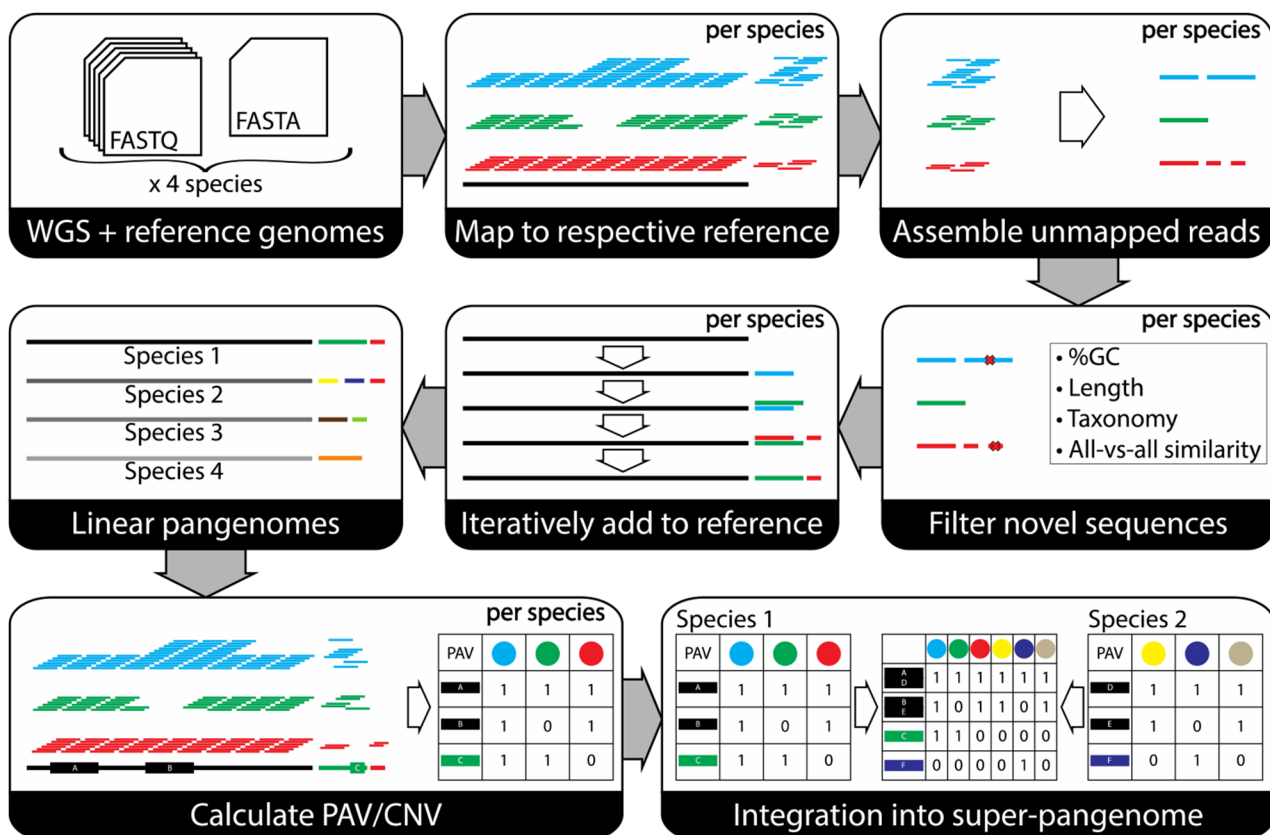


Fig. 1 Overview of the process of going from WGS data for multiple species with their respective reference genomes to a super-pangenome in which all data is integrated across accessions and species. For each species separately, we mapped all available WGS sets per accession and extracted the unmapped reads. These unmapped reads were assembled and filtered for contamination, length and duplications. Then, we iteratively extended the reference genome to a linear pangenome with these novel sequences as to prevent any duplications. Mapping the WGS data per species against their respective linear pangenomes revealed PAV and CNV values per transcript per accession. Finally, the linear pangenomes were integrated into a super-pangenome for all four species using PanTools

Before starting annotation, we created a repeat database per linear pangenome with RepeatModeler and used it for masking the novel sequences with RepeatMasker (both tools run in the TETools v1.5 container) [36]. With PanTools v3.3.4, we retrieved the core set of proteins shared between the *L. sativa*, *L. saligna* and *L. virosa* reference genomes [37, 38]. With only these core genes, we trained a new AUGUSTUS v3.4.0 model for *Lactuca*, which we applied to all novel sequences in the pangenome for gene prediction [39]. Finally, we searched all newly predicted proteins in the nr database with mmseqs2 v13.45111 and only kept contigs that had at least one hit to a species with taxonomy “unknown” or “Eukaryota” [40].

The full pipeline was implemented in Snakemake [41] and is available via:

https://github.com/LettuceKnow/linear_pangenome_building.

Presence/absence and copy-number variation analysis

For calling PAV and CNV of genes, we mapped all WGS data to the linear pangenome of the same species (*i.e.* closest genome) with bwa-mem2. Horizontal coverage (breadth; the fraction of the length of a transcript covered by at least one read) was calculated as a proxy for PAVs. Vertical coverage (depth; the average sequencing depth over a transcript divided by the median value of all average sequencing depths) was calculated as a proxy for CNVs. We calculated these metrics with bedtools ‘coverage’ (v2.30.0) [42], only considering exon locations. Metrics on exons were combined into metrics for transcripts based on their “Parent” attribute according to GFF3 information, thus ignoring coverage information for intronic regions. Organellar genes were excluded from these analyses.

For some downstream analyses, we could not make use of continuous PAV values. In these cases, we “binarised” the PAV values by calling genes present when $PAV > 0.8$ and absent otherwise (see Supplementary Methods – PAV threshold for presence/absence). This binary PAV matrix we call binary PAV (bPAV).

Functional analysis

For each linear pangenome, only the longest isoform per gene was functionally annotated. Extracting the longest isoforms and translating these to protein sequences was done with AGAT v0.5.0 [43]. InterProScan v5.53-87.0 [44] was used for functional annotation. Next to searching for associated InterPro domains, GO terms and pathways, we searched for TIGRFAM, SUPERFAMILY, PANTHER, Gene3D, Coils, Pfam and MobiDBLite annotations.

Super-pangenome construction

The four linear species panproteomes (using protein sequences only) were integrated into one genic super-pangenome using PanTools [38]. Homology grouping was found to be optimal (highest F1-score) at the ‘D3’ setting based on the distribution of ‘eudicots_odb10’ (Benchmarking Universal Single-Copy Orthologs (BUSCO)) orthologues across homology groups. As no method existed to integrate called PAVs across genomes (*i.e.* proteomes), we developed novel functionality to add bPAV information to the graph database (‘add_pavs’; included in PanTools v4.2.0 and onwards). The bPAV information has to be in “binarised” format, *i.e.* each transcript has to be either present (‘1’) or absent (‘0’). This allows us to calculate the number of core and accessory transcripts in the super-pangenome and provide an overview of the gene space for all 474 accessions based on homology. Finally, we also extended the PanTools pangenome growth and gene classification functions to use bPAV information, from PanTools v4.2.0 onwards. Based on gene distances between the accessions, a neighbour-joining tree was generated to gain insight in the phylogeny of all four species.

GWAS

To associate transcript presence/absence and copy numbers with phenotypes, we used the PAV and CNV matrices for *L. sativa*. Since many transcripts show no variation (Fig. 2), we first reduced the size of both PAV and CNV matrices by removing those transcripts that show little to no variation in the following way: We created bins per one decimal for PAV values and we rounded CNV values to the closest integer value (these bins and rounded values are only used for the purpose of filtering out uninformative transcripts). Next, depending on availability of phenotypes for a given accession, we removed those accessions for which no phenotype value was recorded and subsequently applied two filters to the bins: (1) We removed all transcripts that showed no variation across accessions. (2) We removed all transcripts where more than 95% of all accessions had the same (rounded) value (corresponding to a minor allele frequency of 5%). Phenotypes were retrieved from the CGN database and from [45].

GWAS was then performed with the two (filtered) matrices: PAVs and CNVs. The kinship (based on PAV) was calculated as the covariance of the genotypes, using the cov function in R v4.2. GWAS was conducted through a linear mixed model with the kinship included as random effect, using the lme4QTL package in R [46]. Bonferroni multiple testing correction was used, calling associations significant at $p < (0.05/\text{number of variants})$.

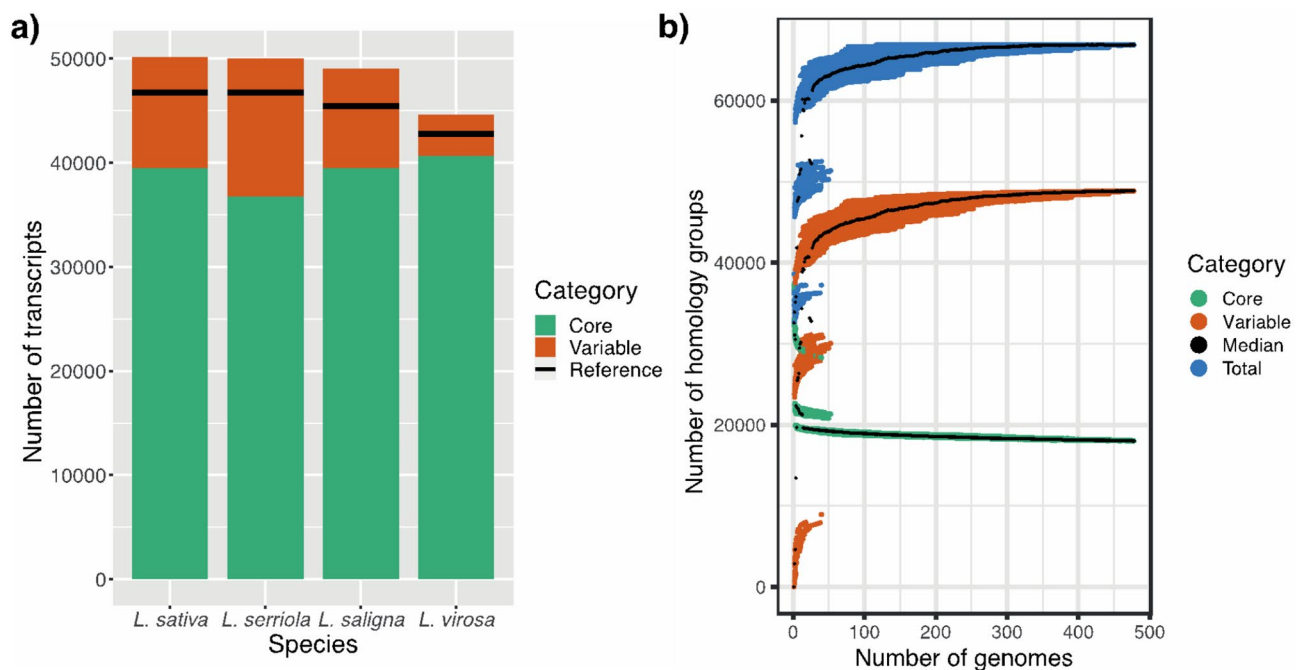


Fig. 2 **a)** Pangenome size for each of the *L. sativa*, *L. serriola*, *L. saligna* and *L. virosa* linear pangenomes. **b)** Growth curve of the super-pangenome constructed from combining the four linear pangenomes (474 genomes total). Curves for core genes (green), variable genes (orange) and total genes (blue) are shown. These growth curves are created by calculating core, variable and total homology group (gene cluster) size for a random selection of genomes in the super-pangenome. This process was repeated 100 times for each number of genomes in the same fashion as introduced by [11]. The “jumps” in each colour are striking and represent sampling genomes from one, two or three gene pools

Results

Constructing a *Lactuca* super-pangenome

Following [19], a *Lactuca* super-pangenome across four species was constructed, combining variation found in 474 cultivated and wild accessions (Supplementary Table S1, S2). This super-pangenome was built from four linear pangenomes, one for each of the four species relevant to lettuce breeding (*L. sativa*, *L. serriola*, *L. saligna*, *L. virosa*) (Fig. 1). Per species, we combined the reference genome with non-reference sequence identified and assembled from low-coverage re-sequenced accessions of that species, and annotated the non-reference genes therein. No duplicated or contaminated sequences were added in the process (Supplementary Methods – Contamination filtering). For each species, we discovered between 10 and 18 Mbp non-reference sequence containing about 2,000 to 3,000 novel genes (Table 1). Finally, we quantified PAVs and CNVs from the alignment of WGS data per linear pangenome. A PAV is calculated as the fraction over the length of a predicted transcript that is covered by WGS reads. A CNV is calculated as the average depth over a transcript by WGS reads. This means that neither PAV nor CNV is a discrete value. However, some applications need discrete values for which we created a bPAV matrix by applying a threshold.

These linear pangenomes were then combined in a super-pangenome graph database using PanTools [38].

Transcripts from all linear pangenomes (based on the proteins they encode) were grouped in homology groups, with settings optimised based on the distribution of BUSCO genes. Then, a super-pangenome bPAV matrix was created by combining the four linear pangenome bPAV tables based on homology (as illustrated in bottom right panel of Fig. 1). The super-pangenome bPAV matrix thus indicates the number of genes present per homology group per accession. This resource enables gene set analyses, the comparison of PAVs and translation of biological information across species.

bPAV analysis shows that the variable part of the pangenome differs per species (Fig. 2a). Based on the thresholds we set (see Supplementary Methods – PAV threshold for presence/absence), per species we found between 5,000 and 15,000 transcripts that show PAV (the so-called ‘variable genome’), which includes about 2,000 to 3,000 non-reference PAVs. Each core genome (per species) contains between 37,000 and 41,000 transcripts. Although the total pangenome size for *L. sativa* and *L. serriola* is comparable, *L. serriola* has a smaller core pangenome size and thus exhibits more variation between accessions. We hypothesise this is because *L. sativa* has potentially gone through a genetic bottleneck, which has removed (presence/absence) variation [4]. Specifically compared to *L. saligna*, which has about the same sample size as *L. virosa*, *L. virosa* exhibits very

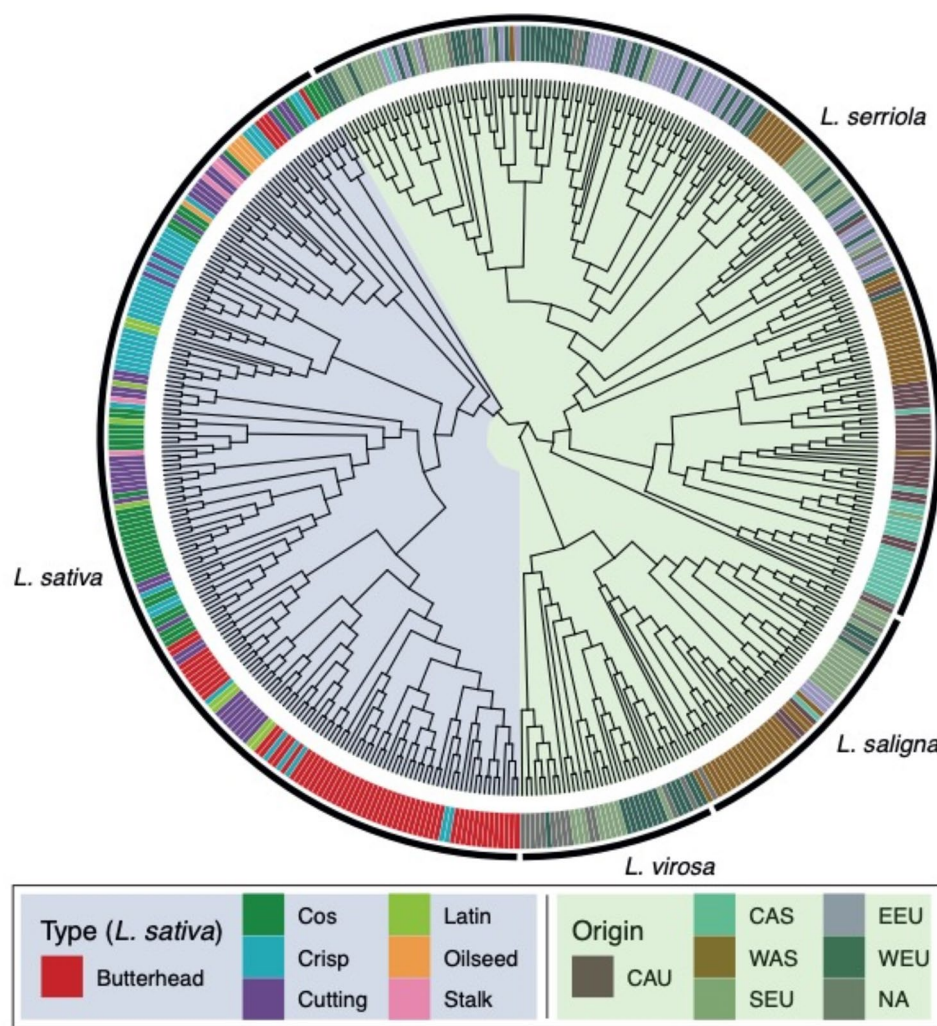


Fig. 3 The super-pangenome bPAV matrix was used to construct a neighbour-joining tree of 474 *Lactuca* accessions. To indicate relatedness of accessions, type (also known as ‘subgroup’) names are used for *L. sativa* and geographical origin for the wild lettuce relatives. The classification of origin as CAU (Caucasus), CAS (Central Asia), WAS (Western Asia), SEU (Southern Europe), EEU (Eastern Europe) and WEU (Western Europe) was taken from [4]

little PAV. This corresponds to the geographical sampling spread of these species: the *L. virosa* accessions sampled are all located in Western Europe, whereas *L. serriola* and *L. saligna* accessions were sampled in both Europe and Asia [4].

Across the four species in the *Lactuca* super-pangenome, bPAV analysis reveals a clear distinction between the three gene pools and shows that about 50% of genes are shared across these pools (Fig. 2b, Supplementary Data 7). In the super-pangenome growth curve, two clear jumps can be seen (in core, variable and total curves) which occur when sampling genomes across phylogenetic boundaries. This corresponds to the three distinct gene pools present: (1) *L. sativa* and *L. serriola*, (2) *L. saligna* and (3) *L. virosa*. Also, the number of core homology groups quickly converges to about 18,000, indicating that the current set of accessions provides a good view on

the core genes that make up *Lactuca*. The jumps in the variable genome size, on the other hand, are larger than for the core genome. Nevertheless, the Heaps’ Law alpha value (1.09) is higher than 1, indicating a closed pangenome [47].

PAVs enable transfer of biological knowledge across species

The super-pangenome facilitates integration of genomic information across species. This allows for high-level analyses such as those of gene sets, but also provides the possibility to approximate the phylogeny of all accessions without the need for variant calling and to perform detailed analyses such as the translation of gene knowledge between species.

The 18,009 core homology groups contain about half of the genes in each accession. Interestingly, this is not

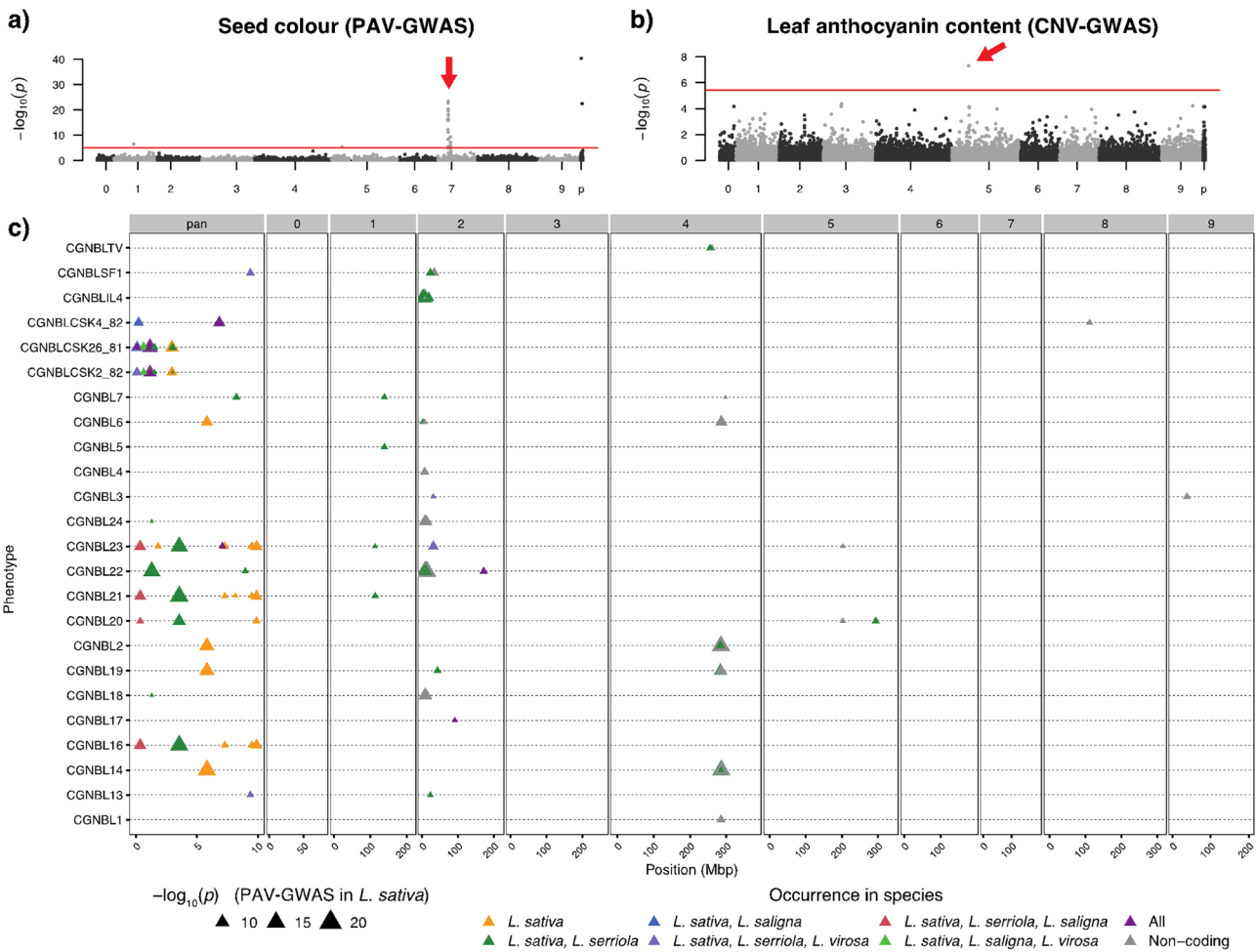


Fig. 4 Manhattan plots for PAV-GWAS and CNV-GWAS show they are complementary to SNP-GWAS. For each PAV, only the longest isoform per gene is visualised. In **a**), we show a clear association between seed colour of *L. sativa* and PAV genotypes that was found previously using SNP-GWAS [48]. **b**) shows that CNV-GWAS can be used for the discovery of the locus responsible for the leaf anthocyanin phenotype [49]. In both **a**) and **b**), the red line indicates the significance threshold according to Bonferroni multiple testing correction ($-\log_{10}(p) > \text{threshold}$). Red arrows indicate the peaks mentioned in the main text. **c**) PAVs in non-reference genes are highly associated with resistance to specific races of lettuce downy mildew *B. lactucae*. The phenotype is the severity score as observed by CGN. Only significant PAVs (after Bonferroni multiple testing correction: $-\log_{10}(p) > 5.3$) are shown, with marker colour indicating occurrence in other species of the super-pangenome and size corresponding to p -value. As non-coding genes cannot be used for homology grouping, they are coloured grey

Table 1 General statistics for all used reference genomes and the linear pangenomes created for these species. Note that the same reference genome was used for both *L. sativa* and *L. serriola*: as *L. serriola* lacks a high-quality reference genome and is closely related to *L. sativa*, we used the *L. sativa* reference genome. The total length is the total length of the assembly in basepairs. For all linear pangenomes, both total and the difference (Δ) compared to the genome of the same species are calculated

Assembly	Origin data	Total length (Δ)	#sequences (Δ)	#transcripts (Δ)
<i>L. sativa</i> genome	GCF_002870075.2	2,391,578,241	8,325	45,246
<i>L. sativa</i> pangenome	REF + 198 accessions	2,401,543,966 (+ 9,965,725)	14,458 (+ 6,133)	48,689 (+ 3,443)
<i>L. serriola</i> genome	GCF_002870075.2	2,391,578,241	8,325	45,246
<i>L. serriola</i> pangenome	REF + 191 accessions	2,408,948,407 (+ 17,370,166)	18,928 (+ 10,603)	48,505 (+ 3,259)
<i>L. saligna</i> genome	PRJEB56287	2,166,282,974	12	42,908
<i>L. saligna</i> pangenome	REF + 49 accessions	2,184,875,879 (+ 18,592,905)	10,509 (+ 10,497)	46,484 (+ 3,576)
<i>L. virosa</i> genome	GCA_928212115.1	3,447,207,058	5,857	39,887
<i>L. virosa</i> pangenome	REF + 36 accessions	3,459,562,374 (+ 12,355,316)	13,510 (+ 7,653)	41,743 (+ 1,856)

a large deviation from what we previously reported on a three genome assembly comparison between *L. sativa*, *L. saligna* and *L. virosa* [18]. A functional enrichment of the core genes mainly identified transcription factor (TF)-related domains (Supplementary Figure S1). This shows that TFs, often master regulators of essential processes, play a role in defining the *Lactuca* genus.

The variable part of genomes of other crops are often enriched for disease resistance genes. These genes are known to be highly variable, both in terms of SNPs and InDels, and in terms of larger variation such as PAVs [50]. This is relevant to plant breeding, since crops need to be protected against ever-evolving pathogens. The available variation among resistance genes provides breeders material to combat the evolution of pathogens. We found the variable genome of *Lactuca* to be not only significantly enriched for resistance genes but also for transposon-related genes (Supplementary Figure S2, Supplementary Data 9). Transposons are known to be both highly variable and a source of phenotypic variation [5, 51]. Some major traits in plants are shown to be caused by transposons, e.g. the insertion of a transposon in *LsKN1* (*L. sativa knotted 1*) is responsible for head formation in lettuce [52].

As SNPs are genomic markers of evolutionary history, they can be used for a good approximation of the phylogeny of a species. Therefore, we reason that PAV should contain similar information. PAVs – as opposed to SNPs – do not involve variant calling and take non-reference genomic content into account. We built a neighbourjoining tree for all accessions in the super-pangenome based on bPAV information (Fig. 3). The *L. serriola* clade in the tree is very similar to the SNP-based tree by [4], which was shown to reflect geographic origin. Both *L. saligna* and *L. virosa* clades also closely follow geographic distribution. We were able to distinguish most types (also called: subgroups) of *L. sativa* accessions. Especially the main breeding types Butterhead, Oilseed and Crisp lettuce clearly cluster together based on PAVs.

PAV-GWAS and CNV-GWAS are complementary to SNP-GWAS

To investigate where PAVs and CNVs are associated to phenotypes, we used the PAV and CNV calls as genotypes for GWAS. First, we performed a validation with the locus that is responsible for variation in seed colour. This locus has previously been identified with SNP-GWAS and a SNP in the *LsTT2* gene has subsequently been confirmed to be the cause of the black/white phenotype [45, 48, 53, 54]. With PAV-GWAS we indeed find the same locus for seed colour as traditional SNP-GWAS. As expected, *LsTT2* itself shows no PAV (the cause is a SNP) but neighbouring genes that do enable us to find the locus (Fig. 4a). Second [49], found a CNV of the *RLL2B*

gene to cause a difference in anthocyanin content in *L. sativa* plants, with a tandem duplication in red accessions but only one copy in green accessions. This duplication locus was previously found with traditional SNP-GWAS and is confirmed in our CNV data (Supplementary Figure S3) [4]. Specifically, when running CNV-GWAS with the anthocyanin phenotypes available through the database from CGN, we found two neighbouring CNVs of *RLL2B* on chromosome 5 to be significant (both correlate with the phenotype and *RLL2B*; thus identifying the correct locus (Fig. 4b)).

Next, we applied PAV-GWAS to a set of publicly available phenotypes of susceptibility scores of *L. sativa* accessions to *B. lactucae* races from CGN. Resistance to pathogens, such as the downy mildew *B. lactucae*, is mostly conveyed by nucleotidebinding leucine-rich repeat, and in more exceptional cases by receptor-like kinase and receptor-like protein families [55]. In lettuce, these genes are commonly found in major resistance clusters that are known to display PAV [56, 57]. Confirming this, we found loci in PAV-GWAS of *B. lactucae* infection phenotypes mostly located in the known major resistance cluster of chromosome 2 (Fig. 4c). This is in agreement with what was found previously with SNP-GWAS [4] (*cfr.* their Supplementary Fig. 11). Furthermore, 36 of all 137 transcripts whose PAV is significantly associated with *B. lactucae* phenotypes do not appear in the reference genome, illustrating the potential impact of reference bias and underlining the importance of a pangenome (Supplementary Table S3). These PAV-containing loci provide starting points for further investigation of these regions.

One of the 36 significant, non-reference transcripts to *B. lactucae* was classified as a disease resistance protein: *g3691.t1*. The encoded protein has a TIR-NBS-LRR domain (PTHR11017:SF386) according to PANTHER. Since the presence of *g3691.t1* is associated with resistance, it could be the causal gene for resistance against *B. lactucae* BL23. From the super-pangenome, we know that this transcript is present in all four species, but in *L. sativa* and *L. serriola* it is absent from their reference genome. The homologues of this gene in *L. saligna* and *L. virosa* are both located at the start of chromosome 2, indicating that this gene is likely part of the major resistance cluster MRC2, which complicates the use of this finding in lettuce breeding.

The non-reference gene whose transcript was significantly associated to most *B. lactucae* phenotypes was *g767*. The PAV pattern of this transcript is significantly associated to 5 of the 24 *B. lactucae* phenotypes collected from the CGN database: CGNBL16, CGNBL18, CGNBL20, CGNBL21, CGNBL23 (Fig. 4c). This transcript is present in 19 of the 198 *L. sativa* accessions and its presence is always negatively correlated to *B. lactucae*

susceptibility, making it a potential resistance locus (Supplementary Table S3). Using correlation with SNPs, we determined this gene likely to be located within the major resistance cluster of chromosome 2 as well. No functional domains could be annotated for the encoded protein by InterProScan. However, in the *L. serriola* pangenome we found a homologue of this gene that is present in 13 of the 191 *L. serriola* accessions, possibly the origin of this locus.

Similarly to SNP-GWAS, GWAS based on CNVs/PAVs is able to identify a locus significantly associated to a phenotype if the causal gene is either a CNV/PAV itself or in linkage disequilibrium (LD) with a CNV/PAV. This makes CNV-GWAS and PAV-GWAS complementary to SNP-GWAS, as CNVs do not necessarily contain SNPs and SNPs are impossible to call for absent regions. Finally, the super-pangenome (including cultivated and wild lettuce) enables us to trace back the potential origin of genes.

Discussion

We described the analysis of a 474-accession *Lactuca* super-pangenome, spanning four species. We demonstrated the added value of a pangenome for GWAS based on gene presence/absence and copy-number variations over reference-based SNP-GWAS. Also, the super-pangenome enabled us to functionally characterise (species-specific) core and variable sets of genes for *Lactuca*. It thereby allowed for reasoning across and comparison between wild and cultivated *Lactuca* species. Below, we discuss a number of general findings.

Quality matters for a reliable pangenome

The quality and completeness of a pangenome critically depend on its underlying data. Thus, from the publicly available genome assemblies we selected those with the highest quality as a reference for mapping and PAV determination, even though this resulted in using the *L. sativa* genome for *L. serriola* accessions (Supplementary Methods – Data selection). At the time of analysis, we used the only available genome assembly, GCF_002870075.2 (*L. sativa* var. Salinas v8). The most recent version (v11: GCF_002870075.4) is a novel genome assembly based on HiFi sequencing instead of Illumina and consequently contains almost no gaps. However, in terms of genic content v11 is almost as complete as v8 (Supplementary Figure S4).

The linear pangenomes are constructed from short-read data, which has drawbacks for the quality of the resulting pangenome. The assembly of all reads that did not map to the high-quality reference genome contains genomic material of other species than only lettuce. Even though this contamination should not be present in a high-quality reference genome, it will be assembled as novel content in the linear pangenome. For separating

Lactuca from non-*Lactuca* content, several metrics were used (see Supplementary Methods – Contamination filtering). Nevertheless, a number of accession-specific PAVs for some accessions in the final PAV analysis are potentially due to unknown contamination. Given short-read data only, we consider it impossible to design a general-purpose pipeline to construct linear pangenomes that are complete, non-redundant and fully free of contamination.

To unequivocally assess homology relationships, ideally all gene sequences in an accession should be available. This requires complete genome assemblies for each accession, which is difficult given only short-read data. However, with the current surge in HiFi sequencing, we expect more of such complete assemblies to become available for *Lactuca* in the near future.

The synergy between PAVs/CNVs and SNPs for GWAS

Even though a genotype may be significantly associated to a specific phenotype by GWAS, this does not imply a causal relationship. In SNP-based GWAS, LD causes SNPs genetically linked to the actual causal genotype to be significantly associated to a phenotype as well. The same is true for PAV- and CNV-GWAS, where LD is clearly visible for e.g. seed colour (Fig. 4a). Other studies where PAVs were used for GWAS showed the presence of LD as well [9, 58]. We found that PAV- and CNV-GWAS peaks overlap loci previously identified by SNP-GWAS in [4, 45]. The causal genotype may be the significant PAV or CNV (as e.g. shown by [9]), but confirmation still requires experimental validation.

An important novelty in this study is the use of continuous genotype values for PAVs and CNVs. In traditional SNP-GWAS, but also in other PAV-GWAS studies [9, 10, 25, 58], presence/absence and the number of copies of a transcript are quantified as integer values. Thresholds are used to distinguish presence from absence, e.g.: The software SGSGeneLoss used a threshold of at least 5% coverage and a minimum depth of 1 for all exons of a transcript [59]; the rice pangenome reported in [60] used 95% coverage for the coding region and 85% coverage for the genic region as threshold; the tomato pangenome by [6] used 20% coverage and a minimum depth of 2 for all exons of a transcript; and the pipeline Panoramic used 75% coverage or 50% coverage and a minimum depth of 3 (depending on quality) for the entire genic region [27]. These choices unnecessarily simplify the underlying data and involve a sometimes rather arbitrary selection of thresholds. Since GWAS is based on a linear model and does not require independent variables to be integer, we made use of the continuous PAV and CNV values as computed from the underlying short-read alignments (see Supplementary Methods – PAV and CNV calculation) to avoid the use of an arbitrary threshold.

Pangenomics is the future for lettuce breeding

A pangenomic approach makes it possible to identify novel genes and find associations between phenotypes and genotypes that are not necessarily present in the reference genome. The power of pangenomics lies in the possibility of combining all types of variation – SNPs, PAV, CNV and larger structural variations –, use it to identify novel breeding targets and interpret them across accessions or even species. The linear pangenomes and the analyses introduced here provide an interesting first look into the *Lactuca* super-pangenome, although ideally the need for the intermediate linear pangenomes will disappear with improved methodology.

To bring pangenomics to full application in plant breeding, however, further investigations are necessary. Firstly, expanding pangenomics to include other wild relatives of lettuce (such as *Lactuca dregeana*, *Lactuca aculeata* and *Lactuca altaica*) will provide a more comprehensive understanding of its genetic diversity. Additionally, generating and including additional genome assemblies will enable researchers to explore intergenic regions, inversions and translocations in detail. These will allow us to help confirm and physically position the here identified non-reference sequences. With the availability of more accurate and complete genome assemblies, we anticipate a clearer picture of (highly) variable regions within the *Lactuca* genome and their biological implications. Such insights can inform targeted breeding strategies and contribute to the development of improved lettuce varieties.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-024-05712-2>.

Supplementary Material 1: Supplementary table S1) Accession numbers used for *L. sativa*, *L. saligna* and *L. virosa* nuclear, mitochondrial and chloroplast assemblies. Supplementary table S2) The accession numbers of all WGS data used, including CGN numbers. Supplementary table S3) Non-reference transcripts significantly associated to *B. lactucae* susceptibility.

Supplementary Material 2: Supplementary figure S1) Functional enrichment of core *Lactuca* genes. Supplementary figure S2) Functional enrichment of variable *Lactuca* genes. Supplementary figure S3) CNV in the *RLL2B* gene in *L. sativa*. Supplementary figure S4) Completeness assessment for both v8 and v11 of *L. sativa* var. Salinas proteome.

Supplementary Material 3: Supplementary methods) Additional information about the key steps we took in the methods and the reasoning behind them.

Acknowledgements

We would like to thank Eef Jonkheer and Robin van Esch for their support in the implementation of novel PanTools functionalities.

Author contributions

DW: Conceptualisation, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Visualisation; SM: Methodology, Software, Formal Analysis, Writing – Review & Editing, Visualisation; BS: Software, Writing – Review & Editing, Supervision; MA: Resources, Data Curation; DL: Resources, Data Curation, Writing – Review & Editing; FM: Resources, Data Curation; GV:

Resources, Data Curation, Writing – Review & Editing, Funding acquisition; DR: Writing – Review & Editing, Supervision; MS: Writing – Review & Editing, Supervision; SS: Conceptualisation, Methodology, Writing – Review & Editing, Supervision.

Funding

This publication is part of the LettuceKnow project (with project number 1.1B of the research Perspective Program P19-17 which is (partly) financed by the Dutch Research Council (NWO) and the breeding companies BASF, Bejo Zaden B.V., Limagrain, Enza Zaden Research & Development B.V., Rijk Zwaan Breeding B.V., Syngenta Seeds B.V., and Takii and Company Ltd.

Data availability

Additional WGS for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB63589 <https://www.ebi.ac.uk/ena/browser/view/PRJEB63589>. The pipeline to create the linear pangenomes is available at https://github.com/LettuceKnow/linear_pangenome_building. PanTools is available at <https://git.wur.nl/bioinformatics/pantools>. We provide a webportal for easy downloading, querying and visualising the data presented here: <https://www.bioinformatics.nl/lettuce/>. Supplementary Data SD1-SD11 are available through 4TU.ResearchData at <https://doi.org/10.4121/c7935d6a-d6ae-42e7-af7e-0ae8cddf70d7>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 1 July 2024 / Accepted: 15 October 2024

Published online: 28 October 2024

References

1. FAOSTAT. FAOSTAT [database on the Internet]. 2023. <https://www.fao.org/faostat/en/#data>.
2. Luo C, Wang S, Ning K, Chen Z, Wang Y, Yang J, et al. The APETALA2 transcription factor LsAP2 regulates seed shape in lettuce. *J Exp Bot*. 2021;03–29(7):2463–76. <https://doi.org/10.1093/jxb/eraa592>.
3. Reyes-Chin-Wo S, Wang Z, Yang X, Kozik A, Arikat S, Song C et al. Genome assembly with *in vitro* proximity ligation data and whole-genome triplication in lettuce. *Nat Commun*. 2017;8(1):14953. <https://doi.org/10.1038/ncomms14953>.
4. Wei T, van Treuren R, Liu X, Zhang Z, Chen J, Liu Y, et al. Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. *Nat Genet*. 2021;05(5):752–60. <https://doi.org/10.1038/s41588-021-00831-0>.
5. Golitz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun*. 2016;7(1):13390. <https://doi.org/10.1038/ncomms13390>.
6. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pangenome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*. 2019;06(6):1044–51. <https://doi.org/10.1038/s41588-019-0410-2>.
7. Li YH, Zhou G, Ma J, Jiang W, Jin Lg, Zhang Z, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol*. 2014;10(10):1045–52. <https://doi.org/10.1038/nbt.2979>.
8. Montenegro JD, Golitz AA, Bayer PE, Hurgobin B, Lee H, Chan CKK, et al. The pangenome of hexaploid bread wheat. *Plant J*. 2017;90(5):1007–13. <https://doi.org/10.1111/tpj.13515>.
9. Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants*. 2020;01(1):34–45. <https://doi.org/10.1038/s41477-019-0577-7>.

10. Zhao J, Bayer PE, Ruperao P, Saxena RK, Khan AW, Goliz AA, et al. Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant Biotechnol J*. 2020;18(9):1946–54. <https://doi.org/10.1111/pbi.13354>.
11. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proc Natl Acad Sci*. 2005;09-27;102(39):13950–5. <https://doi.org/10.1073/pnas.0506758102>.
12. Zhang Z, van Treuren R, Yang T, Hu Y, Zhou W, Liu H, et al. A comprehensive lettuce variation map reveals the impact of structural variations in agronomic traits. *BMC Genomics*. 2023;11–02(1):659. <https://doi.org/10.1186/s12864-023-09739-x>.
13. Harlan JR, de Wet MJ. Toward a rational classification of cultivated plants. *Taxon*. 1971;20(4):509–17. <https://doi.org/10.2307/1218252>.
14. Lindqvist K. On the origin of cultivated lettuce. *Hereditas*. 1960;46(3):319–50. <https://doi.org/10.1111/j.1601-5223.1960.tb03091.x>.
15. Zhang L, Su W, Tao R, Zhang W, Chen J, Wu P et al. RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nat Commun*. 2017-12-22;8(1):2264. <https://doi.org/10.1038/s41467-017-02445-9>.
16. Maisonneuve B, Chupeau MC, Bellec Y, Chupeau Y. Sexual and somatic hybridization in the genus *Lactuca*. *Euphytica*. 1995-02-01;85(1):281–5. <https://doi.org/10.1007/BF00023957>.
17. Xiong W, Berke L, Michelmore R, van Workum DJM, Becker FFM, Schijlen E, et al. The genome of *Lactuca saligna*, a wild relative of lettuce, provides insight into non-host resistance to the downy mildew *Bremia Lactucae*. *Plant J*. 2023. <https://doi.org/10.1111/tpj.16212>.
18. Xiong W, van Workum DJM, Berke L, Bakker LV, Schijlen E, Becker FFM, et al. Genome assembly and analysis of *Lactuca virosa*: implications for lettuce breeding. *G3 Genes|Genomes|Genetics*. 2023-11-01;13(11):jkad204. <https://doi.org/10.1093/g3journal/jkad204>.
19. Khan AW, Garg V, Roorkiwal M, Goliz AA, Edwards D, Varshney RK. SuperPangenome by integrating the wild side of a Species for Accelerated Crop Improvement. *Trends Plant Sci*. 2020-02-01;25(2):148–58. <https://doi.org/10.1016/j.tplants.2019.10.012>.
20. Shang L, Li X, He H, Yuan Q, Song Y, Wei Z et al. A super pan-genomic landscape of rice. *Cell Research*. 2022-10-32(10):878–896. Number: 10 <https://doi.org/10.1038/s41422-022-00685-z>.
21. Gutierrez-Gonzalez JJ, García P, Polanco C, González AI, Vaquero F, Vences FJ, et al. Multi-Species Transcriptome Assemblies of Cultivated and Wild Lentils (*Lens* sp.) Provide a First Glimpse at the Lentil Pangenome. *Agronomy*. 2022;07(7):1619. <https://doi.org/10.3390/agronomy12071619>.
22. Gui S, Wei W, Jiang C, Luo J, Chen L, Wu S et al. A pan-zea genome map for enhancing maize improvement. *Genome Biol*. 2022;23(1):178. <https://doi.org/10.1186/s13059-022-02742-7>.
23. Zhuang Y, Wang X, Li X, Hu J, Fan L, Landis JB, et al. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. *Nat Plants*. 2022;03(3):233–44. <https://doi.org/10.1038/s41477-022-01102-4>.
24. Li N, He Q, Wang J, Wang B, Zhao J, Huang S et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet*. 2023;pp. 1–9. <https://doi.org/10.1038/s41588-023-01340-y>.
25. Sun Y, Wang J, Li Y, Jiang B, Wang X, Xu WH, et al. Pan-genome Analysis reveals the Abundant Gene Presence/Absence variations among different varieties of Melon and their influence on traits. *Front Plant Sci*. 2022;13. <https://doi.org/10.3389/fpls.2022.835496>.
26. Fertet A, Graindorge S, Koehler S, de Boer GJ, Guilloteau-Fonteny E, Gualberto JM. Sequence of the mitochondrial genome of *Lactuca virosa* suggests an unexpected role in *Lactuca sativa*'s evolution. *Front Plant Sci*. 2021;12. <https://doi.org/10.3389/fpls.2021.697136>.
27. Glick L, Mayrose I, Panoramio. A package for constructing eukaryotic pangenomes. *Mol Ecol Resour*. 2021;21(4):1393–403. <https://doi.org/10.1111/1755-0998.13344>.
28. Hu Z, Sun C, Lu K, Chu X, Zhao Y, Lu J et al. EUPAN enables pangenome studies of a large number of eukaryotic genomes. *Bioinformatics*. 2017;33(15):2408–9. <https://doi.org/10.1093/bioinformatics/btx170>.
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.
30. Vasmuddin M, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS); 2019. pp. 314–324. ISSN: 15302075.
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
32. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*. 2015;31(10):1674–6. <https://doi.org/10.1093/bioinformatics/btv033>.
33. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20(1):257. <https://doi.org/10.1186/s13059-019-1891-0>.
34. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–100. <https://doi.org/10.1093/bioinformatics/bty191>.
35. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658–9. <https://doi.org/10.1093/bioinformatics/btl158>.
36. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protocols Bioinf*. 2009;25(1):4.10.1–4.10.14.
37. Sheikhzadeh Anari S, de Ridder D, Schranz ME, Smit S. Efficient inference of homologs in large eukaryotic pan-proteomes. *BMC Bioinformatics*. 2018;19(1):340. <https://doi.org/10.1186/s12859-018-2362-4>.
38. Jonkheer EM, van Workum DJM, Sheikhzadeh Anari S, Brankovics B, de Haan JR, Berke L et al. PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics*. 2022;38(18):4403–5. <https://doi.org/10.1093/bioinformatics/btac506>.
39. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 2003;19(ii215–ii225). <https://doi.org/10.1093/bioinformatics/btg1080>.
40. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35(11):1026–1028. Number: 11. <https://doi.org/10.1038/nbt.3988>.
41. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520–2. <https://doi.org/10.1093/bioinformatics/bts480>.
42. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2. <https://doi.org/10.1093/bioinformatics/btq033>.
43. Dainat J, Hereñú D. NBISweden/AGAT: AGAT-v0.8.0, Zenodo; 2021. <https://zenodo.org/record/5336786>.
44. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C et al. Inter-Scan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236–40. <https://doi.org/10.1093/bioinformatics/btu031>.
45. Mehrem SL, Van den Ackerveken G, Snoek BL. Natural variation in seed coat color in lettuce and wild *Lactuca* species. *bioRxiv*. <https://www.biorxiv.org/content/https://doi.org/10.1101/2024.06.27.600409v1>.
46. Ziyatdinov A, Vázquez-Santiago M, Brunel H, Martínez-Pérez A, Aschard H, Soria JM. lme4qt: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinform*. 2018;19(1):68. <https://doi.org/10.1186/s12859-018-2057-x>.
47. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 2008;11(5):472–7. <https://doi.org/10.1016/j.mib.2008.09.006>.
48. Zhang X, Liang X, He S, Tian H, Liu W, Jia Y, et al. Seed color in lettuce is determined by the LsTT2, LsCHS, and Ls2OGD genes from the flavonoid biosynthesis pathway. *Theor Appl Genet*. 2023;136(12):241. <https://doi.org/10.1007/s00122-023-04491-y>.
49. Su W, Tao R, Liu W, Yu C, Yue Z, He S, et al. Characterization of four polymorphic genes controlling red leaf colour in lettuce that have undergone disruptive selection since domestication. *Plant Biotechnol J*. 2020;18(2):479–90. <https://doi.org/10.1111/pbi.13213>.
50. Tamborski J, Krasileva KV. Evolution of Plant NLRs: From Natural History to Precise Modifications. *Ann Rev Plant Biol*. 2020;71:355–378. <https://doi.org/10.1146/annurev-arplant-081519-035901>.
51. Cai X, Lin R, Liang J, King GJ, Wu J, Wang X. Transposable element insertion: a hidden major source of domesticated phenotypic variation in *Brassica rapa*. *Plant Biotechnol J*. 2022;20(7):1298–310. <https://doi.org/10.1111/pbi.13807>.
52. Yu C, Yan C, Liu Y, Liu Y, Jia Y, Lavelle D et al. Upregulation of a KN1 homolog by transposon insertion promotes leafy head development in lettuce. *Proc Natl Acad Sci*. 2020;117(52):33668–33678. <https://doi.org/10.1073/pnas.2019698117>.
53. Kwon S, Simko I, Hellier B, Mou B, Hu J. Genome-wide association of 10 horticultural traits with expressed sequence tag-derived SNP markers in a collection of lettuce lines. *Crop J*. 2013;10–01(1):25–33. <https://doi.org/10.1016/j.cj.2013.07.014>.

54. Seki K, Komatsu K, Yamaguchi K, Murai Y, Nishida K, Koyama R et al. LsMybW-encoding R2R3-MYB transcription factor is responsible for a shift from black to white in lettuce seed. *Plant Cell Rep.* 2024;43(2):35. <https://doi.org/10.1007/s00299-023-03124-4>.
55. Parra L, Maisonneuve B, Lebeda A, Schut J, Christopoulou M, Jeuken M, et al. Rationalization of genes for resistance to *Bremia lactucae* in lettuce. *Euphytica.* 2016;2103:309–26. <https://doi.org/10.1007/s10681-016-1687-1>.
56. Shen J, Araki H, Chen L, Chen JQ, Tian D. Unique evolutionary mechanism in R-Genes under the Presence/Absence polymorphism in *Arabidopsis thaliana*. *Genetics.* 2006;172(2):1243–50. <https://doi.org/10.1534/genetics.105.047290>.
57. Christopoulou M, Wo SRC, Kozik A, McHale LK, Truco MJ, Wroblewski T et al. Genome-wide Architecture of Disease Resistance genes in lettuce. *G3 Genes[Genomes]Genetics.* 2015;5(12):2655–69. <https://doi.org/10.1534/g3.115.020818>.
58. Liu C, Wang Y, Peng J, Fan B, Xu D, Wu J et al. High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Commun.* 2022;3(6):100352. <https://doi.org/10.1016/j.xplc.2022.100352>.
59. Golicz AA, Martinez PA, Zander M, Patel DA, Van De Wouw AP, Visendi P, et al. Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Funct Integr Genom.* 2015;15(2):189–96. <https://doi.org/10.1007/s10142-014-0412-1>.
60. Sun C, Hu Z, Zheng T, Lu K, Zhao Y, Wang W et al. RPAN: rice pan-genome browser for 3000 rice genomes. *Nucleic Acids Res.* 2017;45(2):597–605. <https://doi.org/10.1093/nar/gkw958>.

Publisher's note

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