Article:
de Kinderen, Maud A J, Ma, Xiao, Vaistij, Fabián E et al. (2026) An inter-specific Amaranthus pangenome captures genetic variation potentially underlying key leafy vegetable traits in this underutilised crop. New Phytologist. pp. 3540-3558. ISSN: 1469-8137

https://doi.org/10.1111/nph.71183

New Phytologist

An inter-specific Amaranthus pangenome captures genetic variation potentially underlying key leafy vegetable traits in this underutilised crop.

Journal:	*New Phytologist*
Manuscript ID	NPH-CR-2025-54284.R1
Manuscript Type:	Full Paper
Date Submitted by the Author:	n/a

Complete List of Authors:	de Kinderen, Maud; University of York, Centre for Novel Agricultural Products, Department of Biology Ma, Xiao; Ghent University, Department of Plant Biotechnology and Bioinformatics Vaistij, Fabian; University of York, CNAP Chang, Jiyang; Ghent University, Department of Plant Biotechnology and Bioinformatics Larson, Tony; University of York, Department of Biology Li, Yi; University of York, Biology, CNAP Dinssa, Fekadu; World Vegetable Center, Eastern and Southern Africa Lin, Jean; World Vegetable Center, Eastern and Southern Africa Suchak, Tulsi; University of York, CNAP Smit, Samuel; University of York, Biology; University of York Centre for Novel Agricultural Products, Jeffares, Daniel; University of York, Department of Biology Van Deynze, Allen; University of California, Plant Sciences Jansen van Rensburg, Willem; Agricultural Research Council, Vegetable, Industrial and Medicinal Plants Research Campus Gerrano, Abe; Agricultural Research Council, Plant breeding and Genetic Resources Management Bairu, Michael; Agricultural Research Council, Vegetable, Industrial and Medicinal Plants Research Campus Venter, Sonja; Agricultural Research Council, Vegetable, Industrial and Medicinal Plants Research Campus Schafleitner, Roland; AVRDC World Vegetable Center, Headquarter Van de Peer, Yves; Ghent University, Department of Plant Biotechnology and Bioinformatics Denby, Katherine; University of York, Centre for Novel Agricultural Products, Department of Biology

Key Words:	Amaranthus, pangenome, MAGIC population, structural variants, metabolomics

Data Availability Statement

Our Mandates Data Policy requires data to be shared and a Data Availability Statement, so please enter one in the space below. Sample statements can be found here. Please note that this statement will be published alongside your manuscript, if it is accepted for publication.

Data presented in this paper can be found in the National Center for Biotechnology Information under Bioproject PRJNA1193012 (genome sequencing, assembly and RNA sequencing data) and in the MASSIVE database under datasets MSV000096703, MSV000096711, MSV000096712, MSV000096713, MSV000096714 and MSV000096870 (as outlined in the Supplementary Information).

All genome assemblies, annotations, protein-coding genes, sequence and structural variants, and the graphical pangenome are deposited in the ORCAE database at https://bioinformatics.psb.ugent.be/gdb/Amranthus/

An inter-specific *Amaranthus* pangenome captures genetic variation potentially underlying key leafy vegetable traits in this underutilised crop.

Maud A.J. de Kinderen[1*], Xiao Ma[2,3*], Fabián E. Vaistij[1], Jiyang Chang[2,3], Tony Larson[1], Yi Li[1], Fekadu Dinssa[4], Lin Chen-yu[4], Tulsi Suchak[1], Samuel J. Smit[1], Daniel C. Jeffares[5], Allen Van Deynze[6], Willem S. Jansen van Rensburg[7], Abe S. Gerrano[7,8], Michael W. Bairu[7], Sonja L. Venter[7], Roland Schafleitner[9], Yves Van de Peer[2,3,10,11] and Katherine Denby[1]

[1] Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Wentworth Way, York YO10 5DD, UK

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9054 Ghent, Belgium

[3] Center for Plant Systems Biology, VIB, 9054 Ghent, Belgium

[4] World Vegetable Center, Eastern and Southern Africa, Duluti, PO Box 10, Arusha, Tanzania

[5] Department of Biology, University of York, Wentworth Way, York YO10 5DD, UK

[6] Seed Biotechnology Center, Department of Plant Sciences, University of California, Davis, USA, 95616; African Orphan Crops Consortium, CIFOR-ICRAF, 1 UN Lane, Nairobi, Kenya

[7] Agricultural Research Council, Vegetable, Industrial and Medicinal Plants Research Campus, Private Bag X293 Pretoria 0001, South Africa

[8] Food Security and Safety Focus Area, Faculty of Natural and Agricultural Science, North-West University, Mmabatho 2735, South Africa

[9] World Vegetable Center, Shanhua, Tainan, 74155, Taiwan

[10] Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

[11] College of Horticulture, Nanjing Agricultural University, Nanjing, 210095, China

[*] Joint first authors

Author email addresses: maud.de.kinderen@york.ac.uk, xiao.ma@psb.vib-ugent.be, fabian.vaistij@york.ac.uk, jiyang.chang@psb.vib-ugent.be, tony.larson@york.ac.uk, yi.li@york.ac.uk, fekadu.dinssa@worldveg.org, jean.lin@worldveg.org, t.suchak@surrey.ac.uk, cobus.smit@york.ac.uk, daniel.jeffares@york.ac.uk, avandeynze@ucdavis.edu, WjvRensburg@arc.agric.za, agerrano@arc.agric.za, BairuM@arc.agric.za, SVenter@arc.agric.za, roland.schafleitner@worldveg.org, yves.vandepeer@ugent.be, katherine.denby@york.ac.uk.

ORCID IDs:

Maud A.J. de Kinderen 0000-0002-9439-7314
Xiao Ma 0000-0003-4787-9318
Fabián E. Vaistij 0000-0002-1469-0589
Jiyang Chang 0000-0002-3646-2878
Tony Larson 0000-0003-1337-3482
Yi Li 0000-0002-4526-7185

Fekadu Dinssa 0000-0002-9277-1413
Lin Chen-yu
Samuel Smit 0000-0002-7382-5113
Dan Jeffares 0000-0001-7320-0706
Allen van Deynze 0000-0002-2093-0577
Willem S. Jansen van Rensburg 0000-0002-5684-6174
Abe S. Gerrano 0000-0001-7472-8246
Michael W. Bairu 0000-0002-8904-1663
Sonja L. Venter 0000-0003-0513-1029
Roland Schafleitner 0000-0003-3637-7603
Yves van de Peer 0000-0003-4327-3730
Katherine Denby 0000-0002-7857-6814

Running Head: Amaranthus inter-specific pangenome

Key Words: Amaranthus, pangenome, MAGIC population, structural variants, field trial, metabolomics

Summary

Amaranthus species (particularly *Amaranthus cruentus*, *Amaranthus hypochondriacus*, *Amaranthus tricolor* and *Amaranthus caudatus*) are traditional underutilised crops with the potential to contribute to sustainable, healthy food systems. We focus on amaranth as a leafy vegetable aiming to develop improved lines for cultivation by smallholder farmers in Sub-Saharan Africa. We demonstrate differences in leaf yield and metabolites relevant to human nutrition across eight amaranth accessions: four *A. cruentus* and four *A. hypochondriacus*. These accessions are founders of an inter-specific Multi-parent Advanced Generation InterCross (MAGIC) population. We generated high-quality genome assemblies and annotations for these founder lines and identified sequence and structural variants compared to a reference *A. cruentus* genome. Pangenome analysis (also including *A. cruentus*, *A. hypochondriacus* and *A. tricolor* reference genomes) identified core, dispensable and private gene families. 50% of gene families were core, highlighting the value, in terms of gene discovery, of sequencing additional accessions and the inclusion of three Amaranthus species. A graphical pangenome was constructed using structural variants and demonstrated variation in copy number of genes with a likely role in disease resistance. This inter-specific pangenome will be highly valuable for future research on amaranth and facilitate usage of structural variants in trait mapping and causal gene discovery.

Introduction

The majority of Amaranthus species originate in the Americas, but there are over 70 species now distributed around the world(Riggins and Mumm, 2021). Three species – *Amaranthus cruentus*, *Amaranthus hypochondriacus* and *Amaranthus caudatus* – are commonly grown for their grain and considered pseudo-cereals. Species such as *Amaranthus tricolor* are grown as a leafy vegetable, however, leaf and grain amaranths cannot be distinguished purely by species,

since there are accessions of grain amaranth species used for vegetable production (Lin *et al.*, 2022).

In Sub-Saharan African, amaranths are viewed as a traditional crop and are one of the most commonly consumed African leafy vegetables. Their consumption is on the rise in East Africa but has been declining in South Africa, at least in part due to urban migration, shifts away from traditional diets and a lack of these crops in formal markets (Nyonje *et al.*, 2022). However, leafy amaranth is recognized, along with other underutilised or orphan crops, for its potential in contributing to sustainable and healthy food systems (Mabhaudhi *et al.*, 2016) and several amaranth leafy vegetable species were identified as priority crops for enhancing nutrition and climate-resilient agriculture in Sub-Saharan Africa (Zonneveld *et al.*, 2023). Amaranths exhibit C_4 photosynthesis (Kadereit *et al.*, 2003), meaning photosynthesis is more efficient under high temperatures, high light intensities and reduced water availability (Edwards *et al.*, 2004). Amaranths grow well in low input agricultural systems typical of smallholder farmers and leaves contain high levels of vitamins, and antioxidants, as well as essential amino acids (Achigan-Dako *et al.*, 2014; Venskutonis and Kraujalis, 2013). Despite these beneficial traits, there are still challenges to overcome to realise the potential of this traditional crop. Amaranth has not been fully domesticated and the traits that have been targeted are those relevant to grain crops (e.g. seed colour (Anuradha *et al.*, 2023)). Limited research has been carried out on leafy amaranth (Emmanuel and Babalola, 2021), with limited improvement of lines and molecular genetic characterisation of traits (Riggins and Mumm, 2021). Capitalising the potential of amaranths to enhance nutrition, dietary diversity, smallholder livelihoods and climate-resilient agriculture, will require the development of improved lines suitable for different agro-ecologies.

To date, three chromosomal-level *Amaranthus* reference genomes have been published: *A. hypochondriacus* (Winkler *et al.*, 2024; Lightfoot *et al.*, 2017) , *A. cruentus* (Ma *et al.*, 2021) and *A. tricolor* (Wang *et al.*, 2022). All are diploid with *A. hypochondriacus* and *A. cruentus* having current genome assembly sizes of 404 and 371 Mb and n=16 and n=17 chromosomes, respectively. *A. tricolor* has a larger genome size (current assembly of 520 Mb) distributed over 17 chromosomes. Near chromosomal-level assemblies have been published for *A. tuberculatus*, *A. hybridus* and *A. palmeri* – species renowned as invasive weeds (Montgomery *et al.*, 2020). In addition to these genome assemblies, short read sequencing of 121 accessions of *A. caudatus*, *A. cruentus*, *A. hypochondriacus* and *A. quitensis* was carried out to investigate (incomplete) domestication (Stetter *et al.*, 2020) and a genotype-by-sequencing approach was used to clarify taxonomy of more than 700 accessions including tetraploid *A. dubius* (Lin *et al.*, 2022).

Reference genomes are valuable genetic resources, but they do not capture the full genetic diversity of these species. Comparison and integration of high-quality assemblies of multiple accessions of a species (pangenomic studies) enables better understanding of genome evolution and phenotypic diversity and linking of genetic variation to agronomic traits. Pangenomes are particularly valuable for identification of structural variants (insertions, deletions, copy number variants and translocations), that are known to play a key role in phenotypic variation (Alonge *et al.*, 2020). Multiple model plant and crop pangenomes have been assembled (Jayakodi *et al.*, 2021) significantly expanding the scope of genetic diversity and improving trait mapping and prediction (Coletta *et al.*, 2021). Pangenomic studies have also

captured inter-specific variation between domesticated cultivars and wild relatives (Zhang *et al.*, 2021).

We generated an inter-specific pangenome from high-quality genome assemblies of eight amaranth accessions, four *A. cruentus* and four *A. hypochondriacus*, the parents of a Multi-parent Advanced Generation Inter-Cross (MAGIC) population. We identified sequence and structural variants between the lines compared to the reference *A. cruentus* genome (Ma *et al.*, 2021) and assembled both a graphical and gene-based pangenome. The accessions vary for leafy vegetable traits, including yield and metabolite profiles, and the pangenome highlights copy number variation in disease resistance genes between the lines. Creation of this inter-specific amaranth pangenome will underpin use of the MAGIC population in developing improved lines of this traditional crop.

Materials and Methods

Field trial

The eight MAGIC population parent accessions, VI044371**,** VI044437A (*A. cruentus*), VI061487, VI061494A (*A. cruentus* admixture), VI062433, VI063769, VI050446 (*A. hypochondriacus*) and VI060472 (*A. hypochondriacus* admixture) were grown at WorldVeg for Eastern and Southern Africa in Arusha, Tanzania (coordinates: 3.37° south, 36.8° east). The trial took place in the warm season over 83 days, with the average time between the five harvests 13 days (Supp Dataset 2a). Three replicate plots of each accession were initially planted with 20 plants. At each harvest, the number of surviving plants was verified. Only one replicate of one line had less than 17 surviving plants at harvest 4, and four individual plots had less than 15 plants at the last harvest (Supp Dataset 2c). Leaf yield was measured in grams per plot and subsequently divided by the number of plants to estimate leaf yield per plant. The number of leaves per plant was counted. For leaf length and leaf width, 10 random leaves were picked from 10 random plants per plot at each harvest and measured. After checking for normality, trait data were analysed using linear mixed-effects models producing restricted maximum likelihood (REML) estimates determined by the lmer function in the lme4 package (Bates *et al.*, 2015). The most complete model used to test significance of different variables on phenotypic traits was

$$y_{ijk} = \mu + Accession_i + Harvest_j + Acc.*Harvest_{ij} + Rep_k + e_{ijk}$$

y_{ijk} is the trait prediction, μ is the mean, $Accession_i$ is the fixed effect of line (n=8), $Harvest_j$ is the fixed effect of Harvest (n=5), $Acc.*Harvest_{ij}$ is the fixed effect of accession by harvest interaction, Rep_k is the random effect of replicate plot (n=3) and e_{ijk} is the random residual assumed to be $\sim N(0, I\sigma^2_e)$, with I being an identity matrix and σ^2_e the residual variance. Significance of the fixed variables was tested using Analysis of Variance with the model adjusted in a backward stepwise manner and the model with the lowest Akaike Information Criterion considered best fitting. Rep_k was considered a random model effect. Tukey's HSD test was used for post-hoc comparisons.

Metabolite profiling

The eight MAGIC parent accessions were grown in soil (Levington Advance, F2+S) in a glasshouse with supplemental lighting (240 μm $m^{-2}s^{-1}$ PPFD) for 12 hour days, at ambient humidity and air temperatures of 27°C (+/- 3°C) day; 25°C (+/- 3°C) night. The youngest fully expanded leaves were harvested from at least 3 independent 6-week old plants and methanolic

extracts analysed in triplicate using liquid chromatography with data-dependent mass spectrometry (LC-MS/MS). Detailed methods are provided in Supplementary Information.

RNA sequencing and analysis

RNA sequencing was carried out on the same samples as metabolite profiling, with an additional two samples per accession from two independent extractions of a second pool of leaves. Three samples from *A. cruentus* "Arusha": leaf and roots from 6-week old plants and 2-week old seedlings were also included. Illumina short-read transcript sequencing on the NovaSeq 6000 platform generated a minimum of 28 million 150-bp paired-end reads for each sample. Read quality was assessed using FastQC v0.11.7-Java-1.8.0_181 (Wingett and Andrews, 2018) and subsequently, reads were aligned to the relevant transcriptome (i.e. *A. cruentus* reference or a MAGIC population parent line) and transcript abundance measured using Salmon v1.10.0-GCC-11.3.0 (Patro *et al.*, 2017). Tximport R package v1.38.2 was used to generate read counts and length-scaled transcript per million reads (TPMs)(Soneson *et al.*, 2016). Mean expression values were used in Supp Figure 6.

Genome sequencing and assembly

Long-read PacBio HiFi sequencing was carried out for the eight founders, generating between 637,951 and 972,273 raw reads per accession. Genome assembly was performed using the *de novo* assembler Hifiasm v0.8-r279 (Cheng *et al.*, 2022). Primary genome assemblies were purged to a full haploid genome by haplotype separation to obtain contig level assemblies using purge_dups VI.0.1 (Guan *et al.*, 2020). GenomeScope 2.0 (Ranallo-Benavidez *et al.*, 2020) was used for assessing heterozygosity levels in the genomes and genome size of the eight founders with the raw PacBio HiFi sequencing reads. The quality of the genome assemblies was assessed using QUAST v5.2.0 (Gurevich *et al.*, 2013). The quality and completeness of the genome assemblies were evaluated using the LTR Assembly Index (LAI)(Ou *et al.*, 2018). Full-length long terminal repeat retrotransposons (LTR-RTs) were detected using LTRharvest (v1.5.9) and LTR_FINDER_parallel (v1.1), followed by redundancy removal and LAI calculation with LTR_retriever (v2.9.8).

Pseudochromosome construction and genome annotation

The assembled contigs of the four *A. cruentus* and four *A. hypochondriacus* accessions were anchored and oriented to chromosomes using the published reference genomes for each species and the reference-guided software RagTag V2.1.0 (Alonge *et al.*, 2022) with default parameters. A repeat library of the eight newly assembled genomes was *ab initio* constructed using RepeatModeler v.2.0.1 (Flynn *et al.*, 2020). The consensus TE sequences were then aligned to the genomes in RepeatMasker (v4.1.1) with parameters -e rmblast -a -s -norna -xsmall -gff for repetitive element annotation, statistics, and masking.

An integrated gene annotation pipeline was used to annotate the protein-coding genes in the eight amaranth genomes. RNA-seq reads were aligned to their assembled genome using hisat2 v.2.1.0 (Kim *et al.*, 2019) with parameters: --max-intronlen 500000 –dta and transcripts were assembled using StringTie v2.1.5 (Kovaka *et al.*, 2019): -m 150 -t -f 0.3. Transdecoder v.5.0.2 (Haas *et al.*, 2008) was used to predict the ORFs: --retain_blastp_hits. Homology-based prediction was performed using GeMoMa v.1.9 (Keilwagen *et al.*, 2018) based on an improved annotation of the *A. cruentus* reference by manual gene curation. BRAKER2 (Brůna *et al.*, 2021) was used for *de novo* prediction using model training based on the RNA-seq data after

the annotated repeats were soft masked in the assembly. EVidenceModeler v.1.1.1 (Haas *et al.*, 2008) was used to combine all gene model evidence obtained. Subsequently, manual curation of gene annotations in all eight MAGIC founder lines was performed using GenomeView (Abeel *et al.*, 2012). Where there was clear RNAseq evidence, missing genes were included, fragmented genes repaired and adjacent joined genes separated. The quality of the annotation results was evaluated by BUSCO v5.8.3 using the embryophyta Odb12 gene set.

Putative gene function was identified using InterProScan v5.47-82 (Paysan-Lafosse *et al.*, 2023) with different databases: PFAM, Gene3D, PANTHER, CDD, SUPERFAMILY, ProSite and GO. Functional annotation of predicted genes was also obtained by aligning amaranth protein sequences against sequences in public protein databases and the UniProt database using BLASTP (E-value <1 × 10−5).

Gene-based pangenome

A gene-based pangenome was constructed with the 11 amaranth assemblies by estimating core, dispensable and private gene sets based on gene family clustering using OrthoFinder v.2.3.3 (Emms and Kelly, 2019). All protein sequences were subjected to homology searches using BLASTP with an expected value of 1 × 10−5. Protein sequences were clustered into paralogous and orthologous sequences using OrthoFinder with an inflation parameter of 1.5. GO-enrichment was conducted with the topGO package v2.62 (Alexa and Rahnenfuhrer, 2023) using GO term annotations for the MAGIC founder and *A. cruentus* genomes as provided (see Data availability statement). GO term annotations for the *A. hypochondriacus* and *A. tricolor* reference genomes were generated using eggNOG-mapper v2 (Cantalapiedra *et al.*, 2021).

Identification of variants and generation of graphical pangenome

The HiFi reads of the eight amaranth accessions were aligned to the *A. cruentus* reference using MUMmer v.4.0.0 (Marçais *et al.*, 2018). SNPs were identified using Longshot (Edge and Bansal, 2019) retaining only the variants labeled as "PASS" and filtered using the following criteria: QUAL (the sixth column) greater than 50 and DP greater than 10 (QUAL > 50 and DP > 10). PBSV (*PacificBiosciences/Pbsv*, 2018/2024) and cuteSV (Jiang *et al.*, 2023) were used to identify SVs. For PBSV the following filters were used: --min_insertion_len 50, --max_insertion_len 100000, --min_deletion_len 50, and --max_deletion_len 100000. Insertions and deletions (tags INS and DEL) were categorised as PAVs and duplications (tag DUP) as copy number variations (CNVs). PBSV and cuteSV VCF files were merged to get the non-redundant file using SURVIVOR (Jeffares *et al.*, 2017), only retaining SVs < 100kb. Assemblytics version 1.2.1 (Nattestad and Schatz, 2016) was used to identify PAV > 100kb. Genome alignment was performed using NUCmer (Marçais *et al.*, 2018) with parameters: --maxmatch -l 100 -c 500. Raw alignments were further filtered with delta-filter using: -1 -i 90 -l 100 and then Assemblytics to parse the filtered delta file to obtain SVs. A graphical pangenome was constructed using the 8 MAGIC parent line genome assemblies, as well as the reference *A. cruentus* genome with updated annotation, with the Mini-graph Cactus pipeline (Hickey *et al.*, 2023).

A PCA was made with an annotated biallelic SNP VCF with BCFtools v.1.10.2-foss-2018b (Danecek *et al.*, 2021), filtered for quality > 1st quantile with VCFtools v.0.1.16-foss-2018b-Perl-5.28.0 (Danecek *et al.*, 2011) and for LD with PLINK v.1.9_beta-6.16 (Chang et al., 2015), with pairwise comparisons calculated in 10 kb windows, a step size of 1, and an R Squared (R^2)

threshold of 0.5. The SNP VCF file was split per pangenome component (core and dispensable), and π was calculated in 100 kb windows for each VCF set with VCFtools v.0.1.16-foss-2018b-Perl-5.28.0.

Pairwise synteny analyses were performed using the JCVI v1.2.7 (Tang *et al.*, 2024) implementation of MCScan (Wang *et al.*, 2012). Protein sequences were used for ortholog detection under default settings followed by a maximum of four iterations for MCScan. Microsyntenic blocks were visualised through JCVI and manually supplemented with syntenic links using the output of OrthoFinder outlined above.

Alignment of amaranth whole genome resequencing dataset

Data for the 121 accessions sequenced in Stetter *et al.* (2020) were aligned to the linear *A. cruentus* reference genome (Ma *et al.*, 2021) using Burrows-Wheeler Aligner (BWA) v.0.7.17 (Li and Durbin, 2009) and to the new graphical pangenome using vg giraffe v.1.54.0 (Sirén *et al.*, 2021) and alignment parameters --max-dp-cells 2097152 --max-extensions 200 --rescue-attempts 5 --max-multimaps 1. Prior to alignment, the graph was indexed with vg autoindex (--workflow giraffe) v.1.54.0 (Garrison *et al.*, 2018). Bam files from the linear genome alignment were merged (two per accession) and indexed using samtools v.1.10 (Danecek *et al.*, 2021), while gam files from the graph alignment were merged with cat and indexed with vg gamsort v.1.48.0 (Garrison *et al.*, 2018). Alignment metrics from bam files were generated with samtools flagstat v.1.15, stats v.1.21 and coverage v.1.21 (Danecek *et al.*, 2021) and for gam files with vg stats, pack and depth, all v.1.48.0 (Garrison *et al.*, 2018).

For all bioinformatic analysis in this study, default software parameters were used unless stipulated otherwise.

Results

The eight parent accessions of a Amaranthus MAGIC population show variation in agronomic traits

The World Vegetable Center (WorldVeg) conserves 1,455 Amaranthus accessions, with more than half of them classified as vegetable or leafy amaranth (Lin *et al.*, 2022). From this large diversity panel, eight founder parent accessions were selected to generate a MAGIC population. These parents were classified in the WorldVeg genebank as four *Amaranthus hypochondriacus* and four *Amaranthus cruentus* species, although two of the lines were reclassified following *de novo* SNP calling (Lin *et al.*, 2022)(Supplementary Dataset 1). These two accessions (VI060472 and VI061494A), along with VI061487, did not appear to have a single ancestral genomic component in that study, with maximum Q values (proportion of an individual's genome originating from one population) of 0.8669, 0.5507 and 0.8147. The lines were characterised using morphological descriptors (as part of >700 lines assessed from the WorldVeg collection) and the 8 parents showed variation in a number of traits relevant to improvement for grain or leaf production (Supplementary Dataset 1). Furthermore, the combination of traits across the 8 lines differed, illustrating the diversity captured by these parent lines. *A. hypochondriacus* is typically a grain amaranth whereas *A. cruentus* is grown for both grain and leaf, and considered a traditional leafy vegetable in Sub-Saharan Africa (Zonneveld *et al.*, 2021). Although, in general, *A. cruentus* flower later than grain amaranth, and are taller with larger leaves (Lin *et al.*, 2022), within the 8 lines no traits clearly differ between the two species (Supplementary Dataset

1). However, individual accessions show traits that are of importance to development of improved lines for cultivation. VI063769 has multiple branches emanating from the base of the stem (rather than branches up the stem) which eases manual harvesting of leaves and VI044371 shows reduced seed shattering compared to other accessions. Several accessions show longer times to flowering (lengthening the vegetative phase) and pigmentation in the stem, leaf and petiole, which influences acceptance of fresh produce. The founder accessions also vary for seed colour (a key grain trait) (Supp Dataset 1).

Agronomic characteristics of the MAGIC population parent accessions were assessed in an irrigated randomised complete block design field trial at WorldVeg in Arusha, Tanzania during the warm season (November 2019 – February 2020, average temperature 21.6°C) (Supp Dataset 2). Leaves were harvested five times during the trial with regrowth of the plants in between; time between harvests varied from 11 to 30 days. The significance of the variation in traits was assessed using linear mixed-effects models including harvest and accession as fixed variables along with their interaction (Supp Table 1). Leaf yield, length, width and number all varied significantly by harvest, with leaf yield peaking at the third harvest and the first and fifth harvests having the lowest yield (Figure 1, Supp Figure 1a). Leaf number followed a similar pattern (Supp Figure 1b) whereas leaf length and width decreased from the first harvest (Supp Figure 1c, d). Crucially there was significant variation between accessions for leaf yield (Supp Table 1), with a significant difference between accessions VI044371 and VI050446 for third harvest yield (Figure 1a) and total yield (i.e. the sum of all five harvests) (Figure 1b). VI044371 and VI050446 also showed a significant difference in leaf length (but not width) across the whole trial (Supp Figure 1e, f), likely underlying the difference in yield as leaf number did not differ significantly between accessions (Supp Table 1). An initial increase and then reduction in yield after repeat harvesting is often seen with repeated harvests of leafy vegetables, for example rocket and spinach (Bantis *et al.*, 2021).

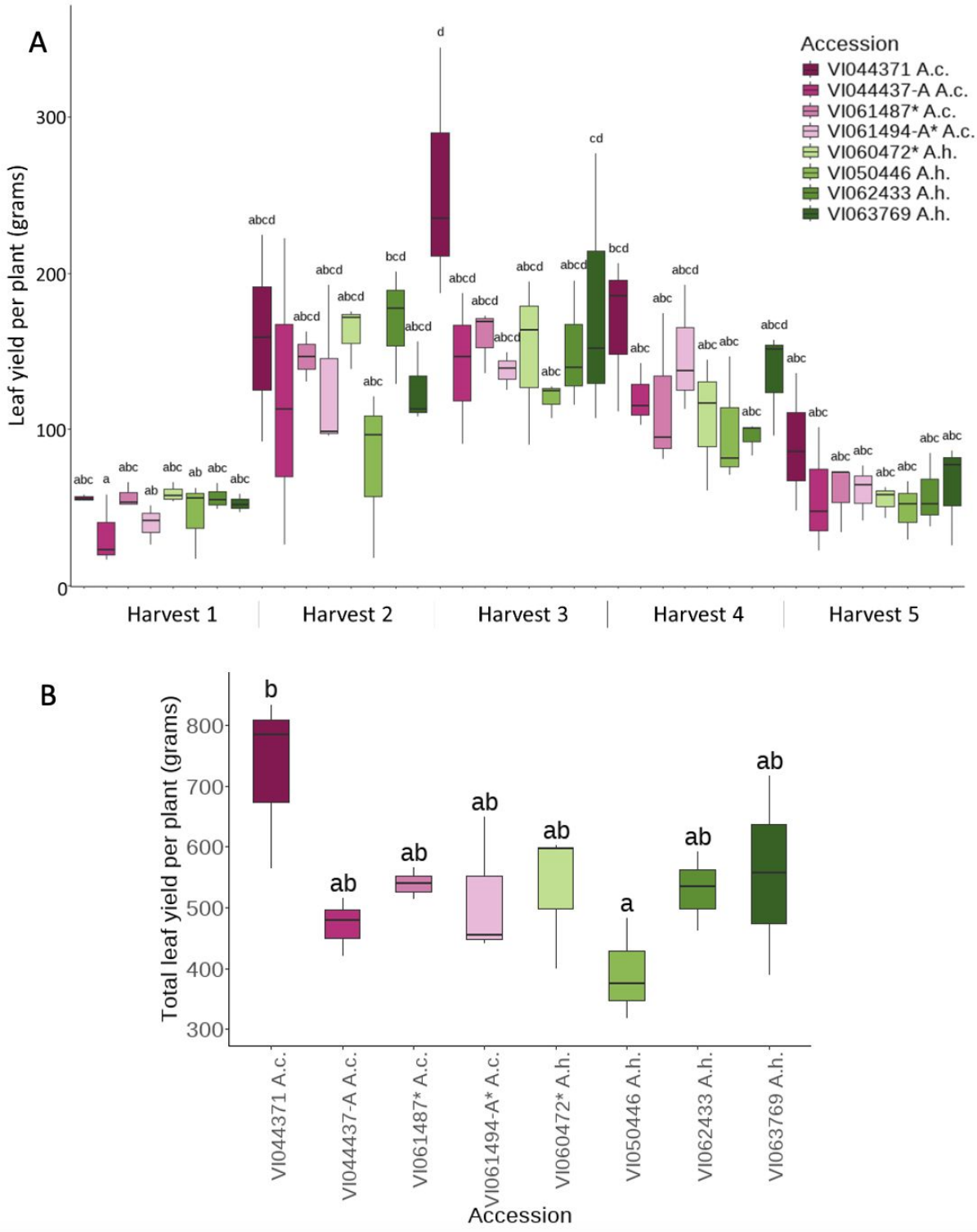


Figure 1. Leaf yield of the eight MAGIC population founder accessions in an irrigated field trial. A) Boxplots show yield measured as fresh weight of harvested leaves per plant. Yield was harvested from three replicate plots of 20 plants with yield per plant calculated to adjust for loss of plants in a plot. Harvest number indicates each subsequent harvest and regrowth of the plants. **B)** Total leaf yield (i.e. the sum of all five harvests) per plant per accession. Letters

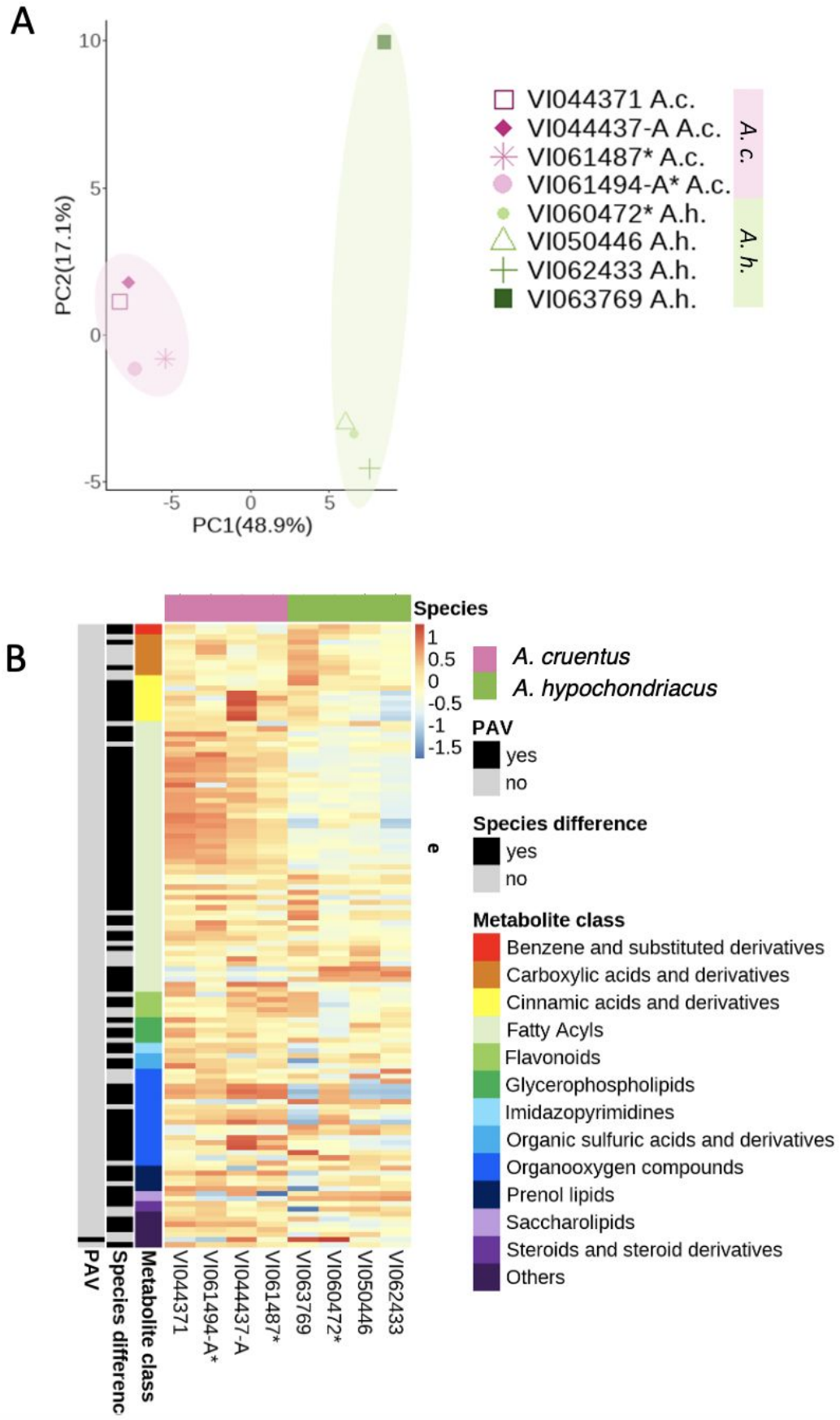
indicate significant differences within each harvest as assessed by Tukey's honest significance test. Initials A.c. and A.h. indicate the species of each line with '*' indicating lines with admixture.

The eight amaranth accessions vary in metabolite profiles relevant to human nutrition

The nutritional content of traditional crops, such as amaranth, is an important factor in the value of these crops to local food systems. We assessed the eight founder lines for variation in metabolite content of greenhouse-grown leaves using untargeted metabolite profiling of methanolic extracts with three replicates.

Liquid chromatography–mass spectrometry (LC-MS) analysis resulted in 3256 features from the positive and negative ionization modes after correcting for redundancy (Supp Dataset 3). Hierarchical clustering using normalised data for these 3256 non-redundant features demonstrates the similarity of replicates as well as clear separation of the two species of amaranth (Supp Figure 2a). Principal component analysis using the abundance of the 3256 non-redundant features again demonstrated clear variation between the two amaranth species (Supp Fig 2b).

After removing very low abundance features, 1421 non-redundant peaks remained of which 1227 varied significantly in abundance between the 8 parent accessions and 847 showed a significant difference in abundance between the two species (*A. cruentus* and *A. hypochondriacus*) (Supp Dataset 3a). Of the 1227, 122 peaks could be annotated and represented 18 major metabolite classes. After clustering of these 122 features on the basis of their abundance across the samples, features with the same annotation cluster together (Supp Figure 3) providing support for the accuracy of annotations. The abundance of the 122 annotated metabolites was also able to clearly distinguish between the two species groups, consistent with their revised species attribution (Lin et al. 2022) (Figure 2a). Figure 2b highlights the variation in abundance of these 122 metabolites both across and within species. Of interest to our aim of developing improved amaranth lines, many of these annotated metabolites are of relevance to human nutrition.


A
PC2(17.1%)
PC1(48.9%)
VI044371 A.c.
VI044437-A A.c.
VI061487* A.c.
VI061494-A* A.c.
VI060472* A.h.
VI050446 A.h.
VI062433 A.h.
VI063769 A.h.
A. c.
A. h.
B
Species
A. cruentus
A. hypochondriacus
PAV
yes
no
Species difference
yes
no
Metabolite class
Benzene and substituted derivatives
Carboxylic acids and derivatives
Cinnamic acids and derivatives
Fatty Acyls
Flavonoids
Glycerophospholipids
Imidazopyrimidines
Organic sulfuric acids and derivatives
Organooxygen compounds
Prenol lipids
Saccharolipids
Steroids and steroid derivatives
Others
PAV
Species differenc
Metabolite class
VI044371
VI061494-A*
VI044437-A
VI061487*
VI063769
VI060472*
VI050446
VI062433

361

Figure 2. Variation in metabolite content between the eight Amaranthus MAGIC population parent accessions. A) Principal component analysis using the abundance of the 122 annotated metabolites significantly differing between MAGIC parent accessions. **B)** The relative abundance of the 122 metabolites across the 8 amaranth accessions organized by metabolite class. Presence absence values (PAV) based on raw data and metabolites varying significantly between the two species are indicated. '*' indicates accessions with genomic admixture.

The largest group of annotated metabolites varying across the 8 accessions are lipids, including fatty acyls, glycerolipids, glycerophospholipids, prenols, saccharolipids and steroids. Lipids are obviously an important nutritional component, and many of these (particularly fatty acyls) appear to differ between the *A. cruentus* and *A. hypochondriacus* accessions (Figure 2b) with higher levels in the *A. cruentus* accessions. Polyphenolic compounds including flavonoids, benzenoids, sesquiterpenes, tannin and hydroxycinnamic acid esters vary across the 8 MAGIC accessions (Figure 2b). These compounds have been associated with multiple health benefits, most likely through their anti-oxidant properties (Monjotin *et al.*, 2022).

Generation of high-quality annotated genome assemblies for the eight Amaranth accessions

To aid association of phenotypic traits with genome sequence, we wanted to be able to identify both sequence and structural variants in the amaranth MAGIC population founder lines. We generated high-quality *de novo* assemblies of the eight amaranth accessions using PacBio HiFi sequencing data (Supplementary Dataset 4a,b, Supplementary Figure 4a). The resulting genomes range in size from 404 Mb to 420 Mb, with contig N50 values from 1.1 Mb to 1.8 Mb (Table 1), consistent with estimated genome sizes based on k-mer analysis (Table 1). The newly sequenced genomes have LAI scores of at least 12.56 and an average of 14.3, consistent with assemblies from long read sequencing and the reference *A. cruentus* genome (Supp Figure 4b) (Ou *et al.*, 2018) The genomes were annotated using a combination of *ab initio* prediction, homology searches, and RNAseq evidence (Supplementary Dataset 4c), supplemented by rigorous manual curation. The resulting annotations identified between 23,756 and 24,973 gene models in each genome (Table 1). We also carried out manual curation of the *A. cruentus* reference genome (Ma et al. 2021) reducing the number of annotated genes from 25,477 to 23,814, but increasing the average gene length from 4,337 to 4,606 bp (Table 1). This manual curation resulted in the BUSCO score increasing from 90.5% complete genes to 94.8% (an additional 69 genes) and the functional annotation rate improving from 94.3% to 97%. The average length of genes and coding regions (CDS) were very consistent across the eight MAGIC population parent and *A. cruentus* reference genome, ranging from 4502-4606 bp and 1192-1240 bp respectively (Table 1). We compared the quality of our annotations to those of six other representative species in the *Amaranthaceae* family, as well as *Arabidopsis thaliana* and *Vitis vinifera* (Supplementary Figure 4b). The distributions of CDS length and exon number of our amaranth genomes were highly consistent with those of other *Amaranthaceae, A. thaliana* and *V. vinifera*. Given the extensive validation and hence high quality of the *A. thaliana* genome (R., Zhang *et al.*, 2022), this consistency provides evidence for the quality of our annotations. Over 95% of annotated genes in the 8 parent accessions can be assigned functions using

InterProScan (v5.47-82) and BUSCO (v5.8.3) scores of >95% demonstrate the high level of completeness of the annotated gene models in the eight genomes (Table 1).

Table 1. Summary of genome assemblies and annotation for the eight *Amaranthus* MAGIC population parent accessions and *A. cruentus* reference genome (Ma et al., 2021). Ac = *A. cruentus*, Ah = *A. hypochondriacus* with * indicating accessions with genome admixture.

Sample	*A. cruentus* reference	VI044371	VI044437	VI061487*	VI061494-A*	VI050446	VI060472*	VI062433	VI063769
Species	Ac	Ac	Ac	Ac	Ac	Ah	Ah	Ah	Ah
Assembly									
Estimated genome size (Mb)	400	410.3	412.4	404.5	425.5	416.4	410.6	403.2	406.3
Total length (Mb)	370.9	410.0	404.8	413.4	416.4	420.3	404.4	410.7	416.8
No. of contigs	625	695	766	875	1039	843	735	773	903
Contig N50 (Mb)	21.7	1.8	1.6	1.1	1.1	1.3	1.4	1.5	1.5
Annotation									
Number of annotated genes	23,814	24,973	23,756	24,475	24,595	24,494	23,988	24,103	24,698
Average gene length (bp)	4606	4502	4574	4523	4525	4590	4545	4577	4543
Average mRNA length (bp)	1240	1192	1240	1215	1217	1227	1216	1224	1212
Average exon length (bp)	238	231	236	233	233	233	231	233	231
Average exons per gene	5.19	5.15	5.26	5.21	5.21	5.26	5.26	5.25	5.23
Average intron length (bp)	799	798	784	786	787	790	782	790	788
BUSCO									
Complete	94.8%	96.2%	96.2%	96.8%	98.0%	96.8%	95.5%	96.1%	97.0%
Complete and single-copy	91.9%	93.4%	93.4%	93.8%	94.9%	93.6%	92.3%	93.1%	93.8%
Complete and duplicated	2.9%	2.7%	3.4%	3.2%	3.1%	3.2%	3.0%	3.0%	3.2%
Fragmented	0.6	0.8%	0.7%	0.8%	0.8%	1.0%	0.9%	0.7%	0.8%

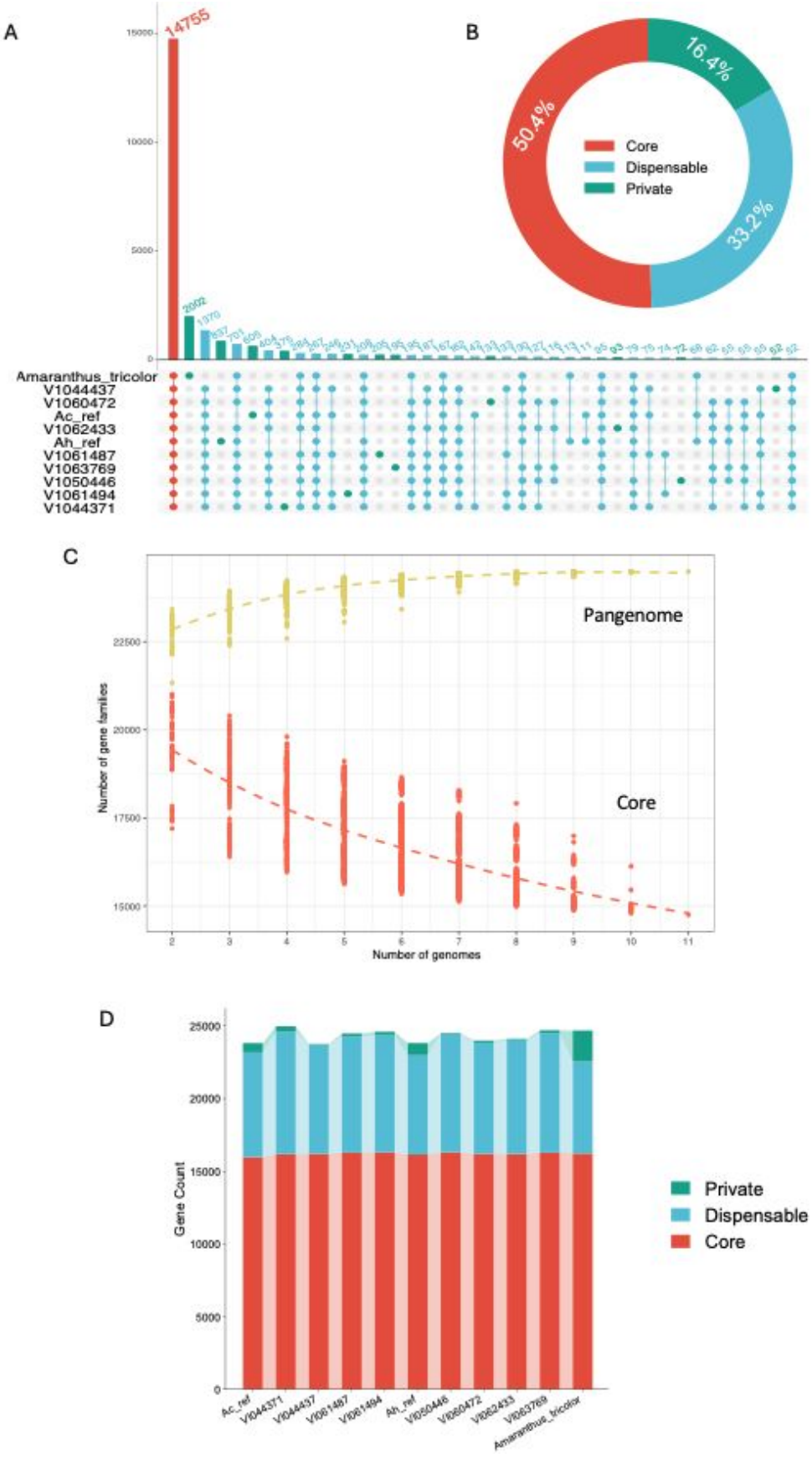
Missing	4.6%	3.1%	2.5%	2.2%	1.2%	2.2%	3.8%	3.2%	2.2%
Functional annotation (%)	97	95.7	97.2	95.1	95	96.8	96.5	96.6	96.4

411

412 A similar proportion (~57%) of the assembled sequence of each accession was composed of
413 repetitive sequences, with the highest proportion of known elements being long-terminal repeat
414 retrotransposons (LTRs)(Supplementary Dataset 5). The Copia content was consistently slightly
415 higher than Gypsy content in all amaranth accessions, which is in contrast to other
416 *Amaranthaceae* species like quinoa and beet (Ma *et al.*, 2021). The eight newly sequenced
417 amaranth accessions and the two reference genomes (Ma *et al.*, 2021; Winkler *et al.*,
418 2024) showed a high degree of pairwise co-linearity at the chromosomal scale, with
419 approximately 80% of genes in co-linear blocks (Supp Figure 5a) and good concordance in
420 gene space representation (Supp Figure 5b). Within each genome approx. 38% of genes were
421 in co-linear blocks (representing duplicated genes), apart from the *A. hypochondriacus*
422 reference genome with only 24.57% of genes (Supp Fig 5a). A phylogenetic tree based on
423 14,729 single-copy genes common to all 10 accessions separated the two species as expected,
424 and in line with the metabolite data above and GBS data (Lin et al. 2022) (Supplementary
425 Figure 5c).

426 **The genetic repertoire of an inter-specific Amaranth pangenome**

427 Using these 8 new genomes, we generated a gene-based pangenome, including the updated
428 reference *A. cruentus* (Ma *et al.*, 2021) and the latest *A. hypochondriacus* reference (Winkler *et*
429 *al.*, 2024) genomes and the genome of a third species, *A. tricolor* (Wang *et al.*, 2022) . Using
430 Orthofinder (Emms and Kelly, 2019), genes from the 11 amaranth genomes were classified into
431 24,489 assigned families and 4,764 unassigned single-copy genes unique to one accession
432 (Supplementary Dataset 6). 14,755 orthogroups were detected in all 11 genomes and defined
433 as the core gene families (50.4%), 9,698 were identified in 2 - 10 accessions, termed
434 dispensable gene families (33.2%), and 36 gene families (186 genes) were detected in only one
435 accession, which along with the single-copy genes unique to a single accession, resulted in
436 4,800 private orthogroups (16.4%)(Figure 3a, b). 86% of the core gene families (12,755) are
437 single copy genes in all accessions. The percentage of core gene families declined as more
438 amaranth accessions were added to the pangenome as expected, but did not plateau (Figure
439 3c) indicating further accessions are likely to bring additional variation in gene content.
440 Approximately 70% of the genes in each genome were part of core gene families in the final
441 pangenome of the 11 accessions (Figure 3d), with ~30% variable gene content. The highest
442 number of private genes was found in the *A. tricolor* genome, presumably reflecting the greater
443 evolutionary distance between *A. tricolor* and *A. cruentus*/*A. hypochondriacus* (which are part of
444 a species complex). Evaluating conservation of genes within the MAGIC population parent
445 species (*A. cruentus* and *A. hypochondriacus*) demonstrated a higher proportion of core genes
446 in each species sub-pangenome compared to the full pangenome (Table 2).


A
14755
15000
10000
5000
0
2002
1370
Amaranthus_tricolor
V1044437
V1060472
Ac_ref
V1062433
Ah_ref
V1061487
V1063769
V1050446
V1061494
V1044371
B
16.4%
50.4%
33.2%
Core
Dispensable
Private
C
Pangenome
Core
Number of gene families
22500
20000
17500
15000
2
3
4
5
6
7
8
9
10
11
Number of genomes
D
Gene Count
25000
20000
15000
10000
5000
0
Ac_ref
V1044371
V1044437
V1061487
V1061494
Ah_ref
V1050446
V1060472
V1062433
V1063769
Amaranthus_tricolor
Private
Dispensable
Core

447

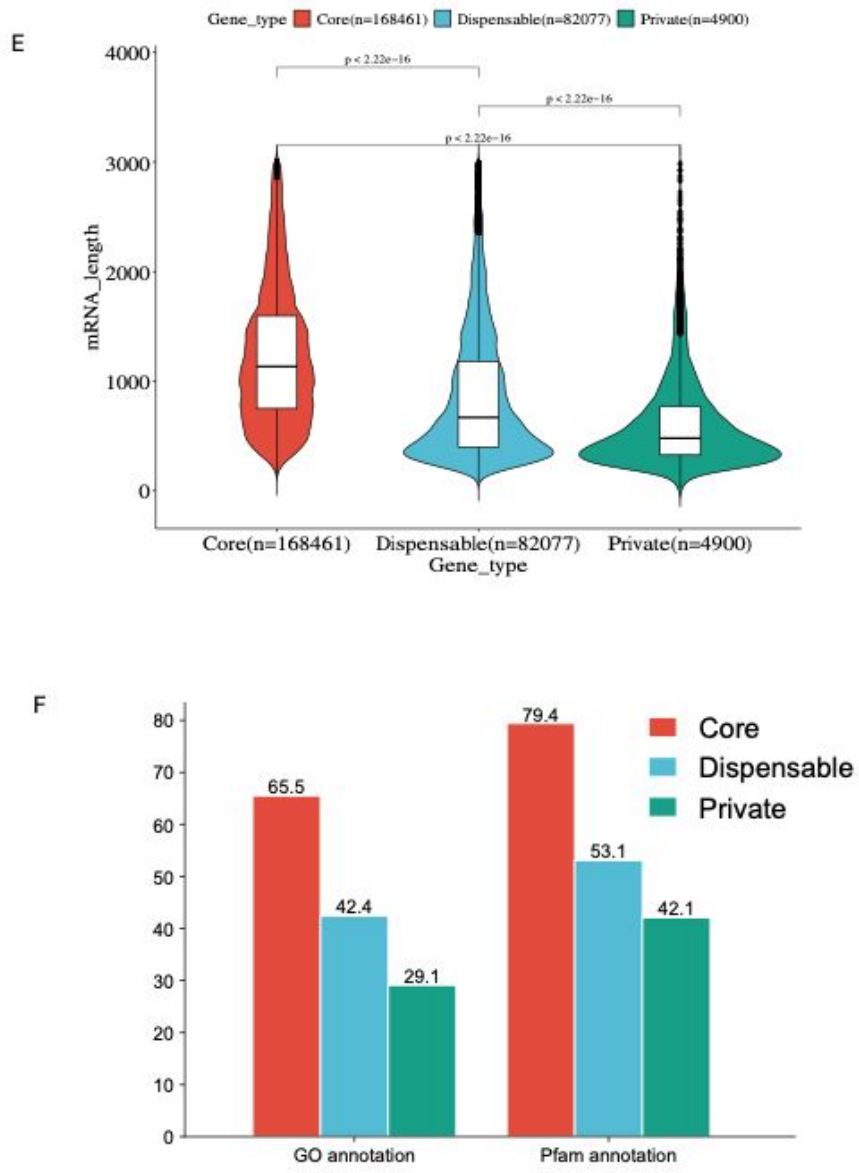


Figure 3: Analysis of pangenome components across 11 *Amaranthus* genomes (eight MAGIC population parents and three published reference genomes, *A. cruentus, A. hypochondriacus* and *A. tricolor).* **A)** The proportion of core, dispensable and private gene families (including single copy genes unique to one accession) across the 11 amaranth genomes. **B**) The number of gene families shared between the 11 amaranth genomes. **C**) The proportion of core and total gene families in the pangenome as the number of genomes compared increases. **D**) The proportion of genes in each genome that are classified as core, dispensable or private in the pangenome of 11 accessions. **E**) mRNA length of genes in the three pangenome components. **F)** The proportion of core, dispensable and private genes that are functionally annotated via Gene Ontology or the Pfam database.

Table 2: Composition of species-specific subpangenomes. The proportion of core, dispensable and private gene families (and single copy private genes) in the species-specific sub pangenomes, and subpangenomes of the *A. cruentus* and *A. hypochondriacus* MAGIC founder accessions only (i.e. not including the *A. cruentus* or *A. hypochondriacus* reference genomes of Ma et al. 2021 and Winkler et al. 2024, respectively).

	Full pangenome	***A. cruentus* accessions only**	***A. hypo-chondriacus* accessions only**	***A. cruentus* MAGIC founder lines only**	***A. hypo-chondriacus* MAGIC founder lines only**
Core gene families	50.4%	67.1%	67.8%	80%	86%
Dispensable gene families	33.2%	30.3%	28.7%	11%	10%
Private gene families and single copy private genes	16.4%	2.6%	3.5%	9%	4%

We observed that the annotated mRNA length of core genes is significantly longer than that of dispensable and private genes (Figure 3e). Furthermore, using RNA-seq data from leaf tissue of the eight MAGIC founder accessions, we observed that overall core genes showed higher expression compared to dispensable genes but there was no obvious difference between expression of dispensable and private genes (Supp Figure 6a). Core genes also showed reduced nucleotide diversity compared to dispensable genes indicating a greater level of conservation (Supp Figure 6b). As expected, the species-specific subpangenomes had a higher proportion of core gene families than the intra-specific pangenome with a greatly reduced proportion of private and single copy genes (Table 2). However, the subpangenomes of just the *A. cruentus* or *A. hypochondriacus* MAGIC founders (not including the reference genomes) had an even higher proportion of core gene families. The lower proportion of dispensable genes indicates that both reference genomes contain a number of genes lacking from all four of the same species MAGIC parent lines. However, there is still a substantial private gene component in these MAGIC species-specific pangenomes (Table 2).

1,961 Benchmarking Universal Single-Copy Ortholog (BUSCO) genes were identified in the amaranth inter-specific pangenome and, as expected, most of these genes (78.4%, 1,538 genes) were part of core gene families (Supp Figure 6d). 20.6% (403 genes) were dispensable and 1% (20 genes) were classed as private genes.

Gene Ontology (GO) and protein domain annotation (Pfam) analysis revealed that the proportion of genes with a functional annotation decreased from core to dispensable to private genes (Figure 3f). This signal, combined with the lower nucleotide diversity and higher gene expression of core genes, is consistent with the core genes being essential. We analysed GO term enrichment among the genes assigned to different pangenome components. GO term enrichment analysis revealed differences in categories of gene function between the genes in different pangenome components. The most significantly enriched terms in the core gene families included several related to transcriptional regulation, and over 70% of the pangenome

transcription factor genes are part of core gene families (Supp Dataset 7). Other central biological processes were also enriched such as mRNA binding, chromatin binding, protein deubiquitination, microtubule binding and ATP hydrolysis. Dispensable gene families included genes with a role in different aspects of the plant defence/immune response, response to auxin, gibberellin and oxidative stress, potentially underlying plant adaptation to changing environmental conditions. GO terms uniquely enriched in the private gene families included secondary metabolite and phenylpropanoid biosynthesis, genes which could drive metabolic diversity between accessions (Supp Dataset 7).

Sequence and structural variation in the amaranth pangenome

The HiFi sequencing reads of the eight MAGIC population parent accessions were aligned to the *A. cruentus* "Arusha" reference genome (Ma *et al.*, 2021) to identify both sequence and structural variants. A total of 8,378,957 single nucleotide polymorphisms (SNPs) were identified: 1,767,497 were found in the four *A. cruentus* parent accessions, while 6,611,460 were detected in the four *A. hypochondriacus* parent accessions, indicating these genomes were less similar to the *A. cruentus* reference (Table 3). Furthermore, the two *A. cruentus* accessions showing genome admixture (VI061487 and VI061494, Li et al. 2022) contained a higher number of SNPs compared to the two other *A. cruentus* accessions, and conversely, the *A. hypochondriacus* accession with genome admixture (VI060472) had the lowest number of variants amongst the *A. hypochondriacus* parent genomes, although the difference was less pronounced. Despite the genome admixture, PCA analysis using the SNPs clearly split the MAGIC parent lines into two species-defined groups (Figure 4a) as seen with the metabolite profiling data (Figure 2a).

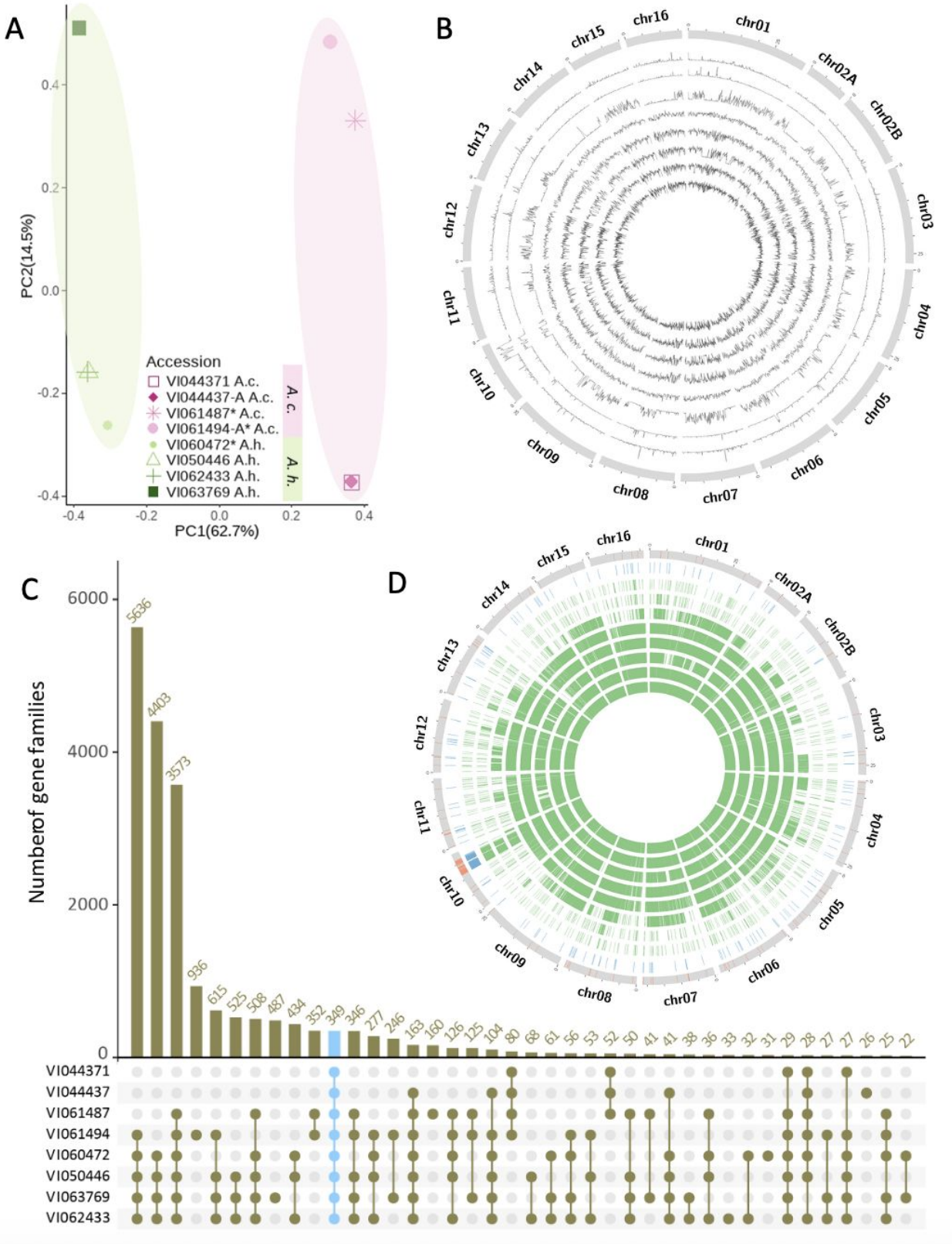


Figure 4. Single Nucleotide Polymorphisms (SNPs) between the MAGIC founder lines and reference *A. cruentus* genome. **A)** Principal component analysis using SNPs indicating clear separation of the MAGIC founder lines into two species groups. **B**) the distribution of SNPs

along the genome. From the outer to inner layer is V1044371, V1044437, V1061487, V1061494, V1050446, V1060472, V1062433 and V1063769. **C**) the number of gene families that contain SNPs in at least one of the eight MAGIC founder lines. Blue indicates (core) gene families where all MAGIC parents contain SNPs in the genes compared to the *A. cruentus* reference. **D**) the distribution of genes containing SNPs along the genome. Green is all genes with SNPs for each accession (accessions in the same order as in B). Blue indicates core genes with SNPs in all accessions and orange is single copy core genes with SNPs in all accessions.

Table 3: The number of Single Nucleotide Polymorphisms (SNPs) and structural variants (SVs) in each MAGIC parent accession genome. The variants are in comparison to the *A. cruentus* "Arusha" reference genome (Ma et al. 2021). The total number of SNPs, SNP density, SNPs in protein coding regions and the number of genes containing at least one SNP are given along with the number of presence absence variation (PAV), inversions (INV), copy number variations (CNV) and translocations (TRANS). * indicates an accession showing admixture (Li et al. 2022). Grey shading indicates the four *A. hypochondriacus* accessions.

	SNPs				SVs			
Accessions	Total #	SNP density (SNPs /kb)	# in protein coding region	# of gene loci with SNP(s)	PAV	INV	CNV	TRANS
V1044371	88,253	0.22	2511	824	22621	97	3093	188
V1044437	105,345	0.26	3224	1,142	22860	101	3060	188
V1061487*	455,149	1.10	19314	6,336	31849	110	3842	468
V1061494*	1,118,750	2.69	40584	13,542	45210	138	5362	932
V1050446	1,677,395	3.99	72845	17,734	63665	210	4833	1,160
V1060472*	1,541,975	3.81	66835	16,262	59253	200	4729	1,054
V1062433	1,718,562	4.18	74016	17,914	64177	224	4916	1,230
V1063769	1,673,528	4.02	76143	17,852	68351	208	5573	1,280
Total	8,378,957		355,472	91,606	377,986	1,288	35,408	6,500

The four *A. hypochondriacus* genomes showed a relatively even distribution of SNPs along the genome (Figure 4b), although VI060472 (genetic admixture) contains some clear regions of reduced SNP density (e.g. regions on chromosome 1, 6, 11 and 14) which may represent regions inherited from an *A. cruentus* ancestor. The four *A. cruentus* accessions showed different patterns of variation density. VI061494 (admixture) showed an even distribution of SNPs along the genome (similar to the *A. hypochondriacus* accessions) whereas the VI061487 genome (admixture) showed specific regions of high SNP density. One region of variation (on chromosome 10) was clearly shared across three of the *A. cruentus* accessions (Figure 4b), with a large proportion of SNPs in the two accessions most closely related to the reference line (Supp Figure 4c) concentrated in this region.

Of the 8,378,957 SNPs identified, 355,473 were located within protein-coding genes, and could cause gain or loss of start or stop codons and splice donor or acceptor sites, as well as frameshift or premature stop codons. Figure 4c shows the number of gene families that contain SNPs in one or more of the amaranth accessions with 349 core gene families having SNPs in all accessions compared to the *A. cruentus* reference genome. The distribution of SNPs within protein-coding regions (Figure 4d) follows a similar distribution to that of all SNPs (Figure 4b). Variation within single copy genes could have a more significant impact on plant phenotype. Of the 12,755 single copy core gene families, 11,776 contain SNPs in one or more of the MAGIC parent lines. The apparent "hotspot" for variation on chromosome 10 is in the gene-dense (low TE) region of this chromosome (Ma *et al.*, 2021) and many of the core gene families (and single copy core gene families) containing SNPs in each accession are located here (Figure 4d).

A total of 421,182 structural variants (SVs) of up to 100 kb were identified including 377,986 presence and absence variations consisting of 231,633 insertions (presence) and 146,353 deletions (absence) with respect to the reference. In addition, 1,288 inversions, 35,408 copy number variations and 6,500 translocations were identified amongst the MAGIC founder lines (Table 3). The number of SVs in each accession is much lower than the number of SNPs as expected, but follows a similar pattern of abundance (Supp Figure 7a). Both deletions and insertions were relatively evenly distributed across the genome with a small number of "hotspots". The hotspots for deletions and insertions were generally not co-located with the exception of regions on chromosome 10, 1 and 2b (Supp Figure 8a, b). Inversions appeared to be more frequently co-located across the accessions (Supp Figure 8c), perhaps indicating a key evolutionary difference between the MAGIC population parents and the *A. cruentus* line chosen as reference.

The SVs were abundant (62-67%) in repeat regions of the genome, while 33–38% of SVs in each accession overlapped genic and flanking regions (5 kb each side of a gene), suggesting a potential impact of SVs on gene content and regulation (Supp Figure 7b). To probe this further, we identified all genes that were within deleted regions (or contained a deleted region) of the parent lines compared to the *A. cruentus* reference. The number of genes ranged from 242 to 4374 with larger numbers in the *A. hypochondriacus* accessions as expected (Supp Figure 9a). Nearly half of the genes identified were shared by just the *A. hypochondriacus* accessions, with the rest showing a variable pattern of presence/absence (Supp Figure 9b). Genes absent or containing a deletion in *A. hypochondriacus* accessions were enriched for functions in lipid metabolism, DNA repair and response to stimulus, light and hormones, potentially impacting on the plant's response to environmental conditions. In contrast, deletions in *A. cruentus* accessions showed enrichment for functions in cellulose biosynthesis and ADP metabolism, both of which could impact yield. Genes deleted in the two *A. cruentus* accessions showing admixture were also enriched for the *A. hypochondriacus*-associated terms "response to stimulus" and "lipid metabolism". Of the 5869 unique genes missing in one or more MAGIC parent lines (Supp Dataset 9), 4426 were grouped in an orthogroup with one or more Arabidopsis genes, and of these 3174 had a function associated with an Arabidopsis orthologue. Genes that were present in all four *A. cruentus* MAGIC parent lines and absent in all four *A. hypochondriacus* lines included those with functions in gibberellin biosynthesis, regulation of flowering time and most interestingly, fatty acid biosynthesis. A significant proportion of the annotated metabolites differing between the two species are fatty acyls (higher content in *A. cruentus* lines, Figure 2), hence these genes may potentially underly this

phenotypic difference. However, any real association between structural variants and phenotype will emerge from future analysis of the full MAGIC population.

Using Assemblytics (Nattestad and Schatz, 2016) to identify large structural variants (SVs) greater than 100 kb, we identified a total of 16 deletions and 136 insertions compared to the reference *A. cruentus* genome (Table 4). Curiously, the number of insertions was higher in the *A. cruentus* accessions compared to the *A. hypochondriacus* accessions, which does not reflect their phylogenetic relationships. However, this is likely to simply reflect the relatively low number of large SVs in the genomes compared to SNPs and smaller SVs.

Table 4: Large insertions and deletions detected in each MAGIC parent accession genome using Assemblytics. The deletions are in comparison to the *A. cruentus* "Arusha" reference genome (Ma et al. 2021). * indicates an accession showing admixture (Li et al. 2022). Grey shading indicates the four *A. hypochondriacus* accessions.

	Deletions				Insertions			
Accessions	**100 - 500 Kb**	**500 - 1,000 Kb**	**> 1 Mb**	**Total (>100 Kb)**	**100 - 500 Kb**	**500 - 1,000 Kb**	**> 1 Mb**	**Total (>100 Kb)**
V1044371	0	0	0	0	19	8	2	29
V1044437	1	0	0	1	18	6	1	25
V1061487*	0	0	0	0	15	7	0	22
V1061494*	1	0	1	2	14	4	3	21
V1050446	2	1	2	5	8	2	1	11
V1060472*	0	1	1	2	5	3	2	10
V1062433	1	1	2	4	6	2	2	10
V1063769	1	0	1	2	6	1	1	8
Total	**6**	**3**	**6**	**16**	**91**	**33**	**12**	**136**

An amaranth graphical pangenome was constructed from the SVs identified using Minigraph Cactus (Hickey *et al.*, 2023). To demonstrate the value of such a pangenome, we mapped whole genome sequences from 121 amaranth accessions (Stetter *et al.*, 2020) to both the linear *A. cruentus* reference genome and the graphical pangenome. The mean read mapping rate increased from 90.7% (linear reference) to 93.2% (pangenome), with a properly paired rate increase from 83.6% to 92.8%. This improvement in mapping was accompanied by a decrease in the number of mismatches per read (1.84 to 1.05). Additional alignment metrics are shown on a per accession basis in Supplementary Figure 10 and Supplementary Dataset 10.

To illustrate the use of the graphical pangenome, we identified two syntenic regions containing copy number variation, likely by tandem duplication and/or gene loss, of genes encoding

nucleotide-binding leucine-rich repeat (NLR) proteins (Figure 5). NLR proteins act as immune receptors and are often arranged in clusters and vary considerably between and within plant species (Barragan and Weigel, 2021). Information on key amaranth diseases is limited to a survey of fungal diseases (Appiah-Kubi *et al.*, 2022) and damping-off in seedlings (Sealy *et al.*, 1988). However, the eight MAGIC parent lines do differ in susceptibility to *Pythium aphanidermatum* (a causal agent of damping-off)(Thanyasiriwat *et al.*, 2025) and future analysis of the MAGIC population generated from these parent lines could enable identification of genetic variation underlying this trait. Capturing the diversity of genes and genome structure using pangenomes will aid in linking sequence and structural variation with multiple phenotypic traits and the breeding or engineering of improved crop lines.

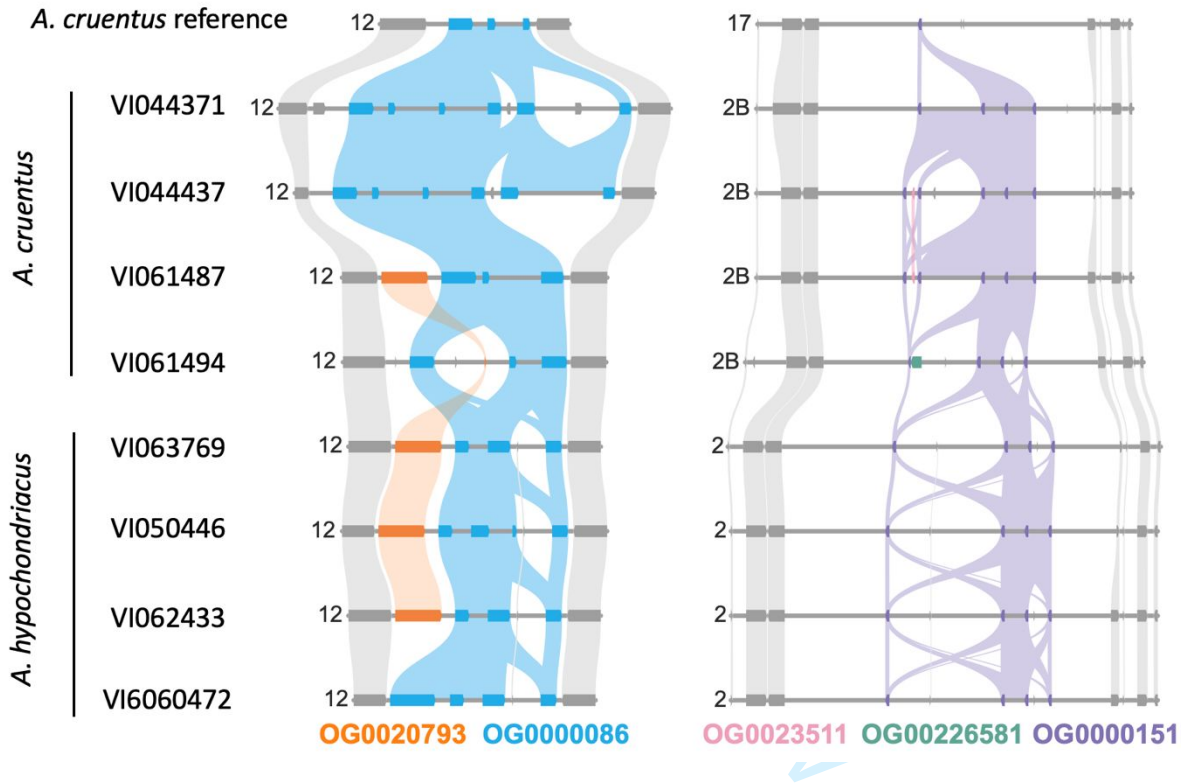


Figure 5: Variation in copy number of putative disease resistance gene orthologues in the Amaranthus pangenome. Genes encoding orthologous nucleotide binding and leucine-rich repeat (NLR) proteins in two genomic locations are indicated for each of the eight founder accessions and the *A. cruentus* reference. The corresponding orthogroup is indicated by colour, with non-NLR genes flanking the region of interest indicated in grey. The chromosomal location is indicated at the left end of each genomic region.

Discussion

The 8 new genomes presented here represent a significant advance in the genomic resources available for amaranth, and constitute the first pangenome within this genus. Given the ability of *A. cruentus* and *A. hypochondriacus* to hybridise (Lin *et al.*, 2022) we have generated an inter-specific pangenome enabling direct comparison of variants between multiple accessions of each species. These two species are thought to have been independently domesticated from *A.*

hybridus but not completely as key domestication traits are still lacking (Stetter *et al.*, 2020). Indeed, several of the accessions in our study have black seed versus the typical cream-coloured seed of domesticated lines. As has been often stated, *Amaranthus* species are difficult to distinguish reliably based on morphology and *A. cruentus* and *A. hypochondriacus* are part of a species complex along with *A. caudatus* (another species independently domesticated from *A. hybridus*). The ability for these species to hybridise is clearly of value in capturing variation for breeding purposes and has also led to admixture within accessions. A genotype-by-sequencing approach (Lin *et al.*, 2022) indicated that three of the lines used in this study (VI060472, VI061487 and VI061494A) contained genome components not typically associated with their given species classification. Consequences of this genomic admixture could be seen in the numbers of SNPs identified in each accession compared to the reference *A. cruentus* genome (Table 3) however, a PCA using the SNPs clearly distinguished the two species groups (Figure 4a). Remarkably, a PCA using metabolite profiles from leaf tissue also clearly distinguished the two species groups (Figure 2a). This raises the possibility of using metabolite profiling (potentially cheaper, quicker and easier) for amaranth species classification where genome sequencing is not possible or too costly. Metabolite profiles from a larger diversity set with accompanying genome data would be useful to evaluate the broader value of this approach beyond the accessions in this study.

Our amaranth pangenome includes not only genomes of *A. cruentus* and *A. hypochondriacus* accessions, but also a reference genome of *A. tricolor*. It considerably expands the scope of known amaranth gene families with 16.4% of gene families (the vast majority of these just single copy genes) only detected in a single amaranth accession (Figure 3). This is a high proportion of private genes compared to many other plant pangenomes (*Raphanus*, jujube, soybean and spinach all have 1% or less (Zhang *et al.*, 2021; Liu *et al.*, 2020; She *et al.*, 2024; Guo *et al.*, 2024). This is likely due to the inclusion of three species in our pangenome, and a significant proportion of private genes come from the reference *A. tricolor* genome which is not part of the amaranth species complex of *A. cruentus*, *hypochondriacus* and *caudatus* (Table 2). Both the *A. cruentus* and *A. hypochondriacus* reference genomes contain a higher proportion of private genes compared to the 8 MAGIC parent lines. If this was due to the quality of the genomes, we would expect more private genes in the MAGIC parents, and although the N50 values are lower for the MAGIC parent genomes, the BUSCO and gene length values are similar across all genomes. Hence, we do not have a clear explanation for this and it may simply reflect genetic difference between the lines. The number of dispensable/private gene families in our pangenome appears to plateau after the inclusion of 7 genomes, whereas the proportion of core gene families keeps decreasing. This is a similar pattern to that seen in a *Brassica napus* pangenome (Hurgobin *et al.*, 2018) and may reflect additional genomes adding new gene variants rather than completely novel genes/gene families. At 50.4%, the proportion of core gene families (those present in all 11 accessions) was within the range seen in other plant pangenomes (Tao *et al.*, 2019). However, this proportion does vary widely and is not necessarily correlated with the number of accessions included, for example, 24% in a Setaria pangenome from 110 accessions (He *et al.*, 2023), 60% in an Arabidopsis pangenome from 69 accessions (Lian *et al.*, 2024), and 35% in a pangenome assembled using 8 jujube accessions (Guo *et al.*, 2024). The proportion of core gene families is still decreasing as more genomes are integrated (Figure 3c) indicating that this collection of 11 lines has not captured the full range of

amaranth diversity and it is likely that incorporation of additional genomes will reduce the proportion of core genes below 50%.

The pangenome not only expands our knowledge of amaranth genes, but highlights sequence and structural variation that is likely to impact gene function and regulation. The density of SNPs identified in the eight MAGIC founder lines (compared to the *A. cruentus* reference genome) varied from 0.22 to 4 SNPs per kb (Table 3). Although only 4% of the SNPs were in protein-coding genes, SNPs were identified in one or more of the amaranth lines in 20,221 of the 23,796 gene families in the pangenome (Figure 4c). The density of SVs was significantly lower (0.06 to 0.18 SVs per kb) and the majority of these were found in repeat regions of the genome, however, at least 1/3 of the SVs in each genome are found overlapping with coding genes and/or their 5 kb flanking regions (Supp Figure 7) suggesting these SVs would impact on presence/absence and/or gene regulation. Further sequencing of the amaranth accessions using methods that capture physical association between DNA sequences (e.g. Hi-C) would likely identify additional SVs and particularly enable us to pinpoint larger genomic rearrangements between the lines.

The 8 genomes that have been newly sequenced, assembled, annotated and compared to the reference *A. cruentus* genome, are the founders of a MAGIC population. As a community resource, our inter-specific pangenome will therefore be of immediate value for genotyping the population and facilitating the mapping of phenotypic traits. Indeed, amaranth was recently highlighted as a key “opportunity crop” for agricultural development and food system resilience (Lecy *et al.* 2025). The genetic improvement of traditional crops, with advantages for smallholder climate resilience, nutrition, dietary diversity and economics, is partly limited by a lack of genome resources and availability of markers linked to beneficial traits (Chapman *et al.*, 2022). Most underutilised crop genome studies involve a single species and single reference genome (Hu *et al.*, 2025) and there is a clear advantage of pangenomes in better representing diversity within a (or between) species. Robust linkage between traits and markers requires identification of genomic variation, with structural variants/pangenomes increasing the known variation and enabling the inclusion of genes that are missing in a single reference genome. However, the selection of accessions to use when constructing a pangenome is not always clear (Hu *et al.*, 2025). In this case as the parents of the MAGIC population there is one clear rationale and value for the pangenome. A future amaranth pangenome could benefit from additional accessions, using genotyping of germplasm collections to select the lines that are likely to bring the most new genomic diversity.

The value of a MAGIC population is dependent on the genetic and phenotypic diversity captured within the parent lines. Genetic diversity is highlighted above both in terms of variants and the presence/absence of gene families. Although several accessions exhibit genomic admixture, we anticipate this will have minimal impact on trait mapping as all accessions show very low levels of heterozygosity (Supp Dataset 4). Our study also demonstrates that the founders vary in key phenotypic traits that are relevant to the cultivation and consumption of amaranth as a leafy vegetable. These include leaf yield, as well as leaf size (length/width) and flowering time (i.e. days to flowering). Smallholder farmers in South Africa prefer fewer larger leaves (rather than more smaller leaves) for ease of harvesting. Similarly, the branching habit of VI063769 (many branches from the base of the stem rather than branches up the stem) makes manual harvesting easier and quicker. A recent study using six of the parent lines demonstrated variation in drought tolerance (assessed by leaf yield under different soil water regimes)(Khoza *et al.*, 2025) . The MAGIC founder lines differ in content of nutritionally-relevant metabolites

(Figure 2). More than 3000 non-redundant features could be detected in our metabolomic analysis, but only 169 of these could be annotated (of which 122 differed between the amaranth accessions). The annotation rate (5%) reflects a current assessment of annotation ability in plant metabolomics (Alseekh and Fernie, 2023) but is lower than studies in other underutilised crops such as jujube (Z., Zhang *et al.*, 2022). This may be due to different annotation methods; we applied a stringent annotation workflow by manually reconciling MS1 matches against a restricted plant-derived database to chemical class matches assigned using MS2 fragment data from Sirius 4.0 (Dührkop *et al.*, 2019). Some of the metabolite classes differing between the eight amaranth accessions (e.g. polyphenolics) are of importance to human nutrition but also play important roles in flavour and in defence against pathogens and herbivory (Stiller *et al.*, 2021). Both roles are of relevance to the use of amaranth in food systems, and fit with a recent call to develop "crops that nourish" incorporating nutrition and environmental resilience, not just yield (Lecy *et al.*, 2025). The MAGIC population parent lines are available from the World Vegetable Center enabling amaranth breeders around the world to use them in their own programmes and start to incorporate genomic information.

In conclusion, the inter-specific pangenome generated here from diverse amaranth accessions, expands the known genetic repertoire of amaranth, highlights variation in gene families between accessions, and provides a valuable resource for the investigation of the genetic basis underlying agronomic traits. The use of two amaranth species captures greater genetic (including gene family) diversity than one species with the potential to identify the genomic mechanisms underlying the striking species differences in metabolite profiles.

Acknowledgments

The project was funded by a Biotechnology and Biological Sciences Research Council award to KD, SV, MB and SM (BB/R020345/1) and Grand Challenges Research Fund QR Funding distributed by the University of York. We thank the African Orphan Crop Consortium (AOCC, http://africanorphancrops.org/) for their collaboration and support of the genome sequencing. YVdP acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01). MdK is funded by the BBSRC White Rose Doctoral Training Partnership grant (BB/T007222/1). WorldVeg was supported by the Ministry of Agriculture Taiwan and by the long-term strategic donors Taiwan, UK aid from the UK government, United States Agency for International Development (USAID), Australian Centre for International Agricultural Research (ACIAR), Germany, Thailand, Philippines, Korea, and Japan. We acknowledge Dr. Oanh Nguyen from the University of California, Davis Genome Center for her input and sequencing of the genomes.

Author Contributions

The study was conceived and designed by KD, AvD, RS, SM, WJvR, AG, MB, SV and YvdP. Experimental work was carried out by FV, FD, LC-y, RS, TL and MdK, with genome assembly by YL. Genome annotation, pangenome construction and comparative analyses were carried out by XM, JC, TS, SS, DJ and MdK and metabolite analysis by MdK and TL. Funding for the research was obtained by KD, SV, MB, SM, YVdP and AvD. The manuscript was written by XM, MdK and KD with input from all authors.

Data availability statement

Data presented in this paper can be found in the National Center for Biotechnology Information under Bioproject PRJNA1193012 (genome sequencing, assembly and RNA sequencing data) and in the MASSIVE database under datasets MSV000096703, MSV000096711, MSV000096712, MSV000096713, MSV000096714 and MSV000096870 (as outlined in the Supplementary Information).

All genome assemblies, annotations, protein-coding genes, sequence and structural variants, and the graphical pangenome are deposited in the ORCAE database at https://bioinformatics.psb.ugent.be/gdb/Amranthus/

Conflict of interests

The authors declare no conflicts of interest.

Supplementary information

Supplementary Methods: Detailed Metabolite Profiling methodology

Supplementary Data 1; Data from Lin et al. (2022) for the 8 MAGIC parent accessions.

Supplementary Data 2; MAGIC population parent accession field trial information and data.

Supplementary Data 3; Features detected from metabolome profiling of the 8 amaranth MAGIC parent accessions.

Supplementary Data 4; Sequencing statistics for the 8 Amaranthus accessions.

Supplementary Data 5; Repetitive element content of the 8 Amaranth MAGIC population parent genomes and the *A. cruentus and A. hypochondriacus* reference genomes.

Supplementary Data 6; Orthologous gene families and single copy genes (orthogroups) within the 8 amaranth MAGIC population parent lines and the reference *A. cruentus*, *A. hypochondriacus* and *A. tricolor* genomes.

Supplementary Data 7; Gene Ontology terms showing significant enrichment in genes from the core, dispensable and private pangenome components.

Supplementary Data 8: Custom metabolite database for annotation.

Supplementary Data 9: Genes within the *A. cruentus* reference genome which are missing in one or more MAGIC population parent lines.

Supplementary Dataset 10: Mapping statistics for alignment of the 121 amaranth whole genome sequences from Stetter *et al*. 2020 to the *A. cruentus* linear reference genome (bams) and to the graphical pangenome generated in this study (gams).

Supplementary Figure 1: Leaf trait data from the irrigated field trial of the eight MAGIC parent accessions.

Supplementary Figure 2: Accession and metabolite sample clustering based on normalised data for 3256 non-redundant detected features.

Supplementary Figure 3: Clustering of 122 annotated metabolites based on their abundance across samples from the 8 Amaranth MAGIC population founder lines.

Supplementary Figure 4: Amaranthus MAGIC population parent accession genome assembly and annotation.

Supplementary Figure 5: Co-linearity and phylogenetic relationships between the eight Amaranth MAGIC population parent genomes.

Supplementary Figure 6: Analysis of pangenome components.

Supplementary Figure 7: Structural variants in the eight MAGIC population founder lines.

Supplementary Figure 8: Density of structural variants in the eight MAGIC population founder accession genomes.

Supplementary Figure 9: Genes contained within deletion regions (or containing deleted regions) of the MAGIC parent genomes compared to the *A. cruentus* reference genome.

Supplementary Figure 10: Linear reference versus pangenome mapping statistics.

References

Abeel, T., Parys, T.V., Saeys, Y., Galagan, J. and Peer, Y.V. de (2012) GenomeView: A next-generation genome browser. *Nucleic Acids Research*, **40**, e12. Available at: https://doi.org/10.1093/nar/gkr995.

Achigan-Dako, E.G., Sogbohossou, O.E.D. and Maundu, P. (2014) Current Knowledge on Amaranthus Spp.: Research Avenues for Improved Nutritional Value and Yield in Leafy Amaranths in Sub-Saharan Africa. *Euphytica*, **197**, 303–317. Available at: https://doi.org/10.1007/s10681-014-1081-9.

Alexa, A. and Rahnenfuhrer, J. (2023) topGO: Enrichment Analysis for Gene Ontology. Available at: https://bioconductor.org/packages/topGO/.

Alonge, M., Lebeigle, L., Kirsche, M., et al. (2022) Automated Assembly Scaffolding Using RagTag Elevates a New Tomato System for High-Throughput Genome Editing. *Genome Biology*, **23**, 258. Available at: https://doi.org/10.1186/s13059-022-02823-7.

Alonge, M., Wang, X., Benoit, M., et al. (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, **182**, 145-161.e23. Available at: https://www.sciencedirect.com/science/article/pii/S0092867420306164.

Alseekh, S. and Fernie, A.R. (2023) Expanding Our Coverage: Strategies to Detect a Greater Range of Metabolites. *Curr. Opin. Plant Biol.*, **73**, 102335. Available at: https://www.sciencedirect.com/science/article/pii/S1369526622001649.

Anuradha, Kumari, M., Zinta, G., Chauhan, R., Kumar, A., Singh, Sanatsujat and Singh, Satbeer (2023) Genetic Resources and Breeding Approaches for Improvement of Amaranth

(Amaranthus Spp.) and Quinoa (Chenopodium quinoa). *Frontiers in Nutrition*, **10**, 1129723. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10405290/.

Appiah-Kubi, Z., Adomako, J., Dwamena, H., Obeng, E.A., Adu-Gyamfi, R., Osekre, E.A., Quian, M. and Kwoseh, C. (2022) Survey of fungal diseases associated with amaranth (Amaranthus species) in peri-urban vegetable farms in Kumasi and Tamale metropolis of Ghana. *Afr. J. Agric. Res.*, **18**, 792–802.

Bantis, F., Kaponas, C., Charalambous, C. and Koukounaras, A. (2021) Strategic Successive Harvesting of Rocket and Spinach Baby Leaves Enhanced Their Quality and Production Efficiency. *Agriculture*, **11**, 465.

Barragan, A.C. and Weigel, D. (2021) Plant NLR diversity: the known unknowns of pan-NLRomes. *Plant Cell*, **33**, 814–831.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software*, **67**, 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. and Borodovsky, M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*, **3**, lqaa108. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33575650.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J. (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.*, **38**, 5825–5829.

Chapman, M.A., He, Y. and Zhou, M. (2022) Beyond a reference genome: pangenomes and population genomics of underutilized and orphan crops for future food and nutrition security. *N. Phytol.*, **234**, 1583–1597.

Cheng, H., Jarvis, E.D., Fedrigo, O., Koepfli, K.-P., Urban, L., Gemmell, N.J. and Li, H. (2022) Haplotype-Resolved Assembly of Diploid Genomes without Parental Data. *Nature Biotechnology*, **40**, 1332–1335. Available at: https://www.nature.com/articles/s41587-022-01261-x.

Coletta, R.D., Qiu, Y., Ou, S., Hufford, M.B. and Hirsch, C.N. (2021) How the Pan-Genome Is Changing Crop Genomics and Improvement. *Genome Biology*, **22**, 3. Available at: https://doi.org/10.1186/s13059-020-02224-8.

Danecek, P., Auton, A., Abecasis, G., et al. (2011) The Variant Call Format and VCFtools. *Bioinformatics*, **27**, 2156–2158. Available at: https://doi.org/10.1093/bioinformatics/btr330.

Danecek, P., Bonfield, J.K., Liddle, J., et al. (2021) Twelve Years of SAMtools and BCFtools. *GigaScience*, **10**, giab008. Available at: https://doi.org/10.1093/gigascience/giab008.

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J. and Böcker, S. (2019) SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nature Methods*, **16**, 299–302. Available at: https://www.nature.com/articles/s41592-019-0344-8.

Edge, P. and Bansal, V. (2019) Longshot Enables Accurate Variant Calling in Diploid Genomes from Single-Molecule Long Read Sequencing. *Nature Communications*, **10**, 4660. Available at: https://www.nature.com/articles/s41467-019-12493-y.

Edwards, G.E., Franceschi, V.R. and Voznesenskaya, E.V. (2004) SINGLE-CELL C4 PHOTOSYNTHESIS VERSUS THE DUAL-CELL (KRANZ) PARADIGM. *Annual Review of Plant Biology*, **55**, 173–196. Available at: https://www.annualreviews.org/content/journals/10.1146/annurev.arplant.55.031903.141725.

Emmanuel, O.C. and Babalola, O.O. (2021) Amaranth Production and Consumption in South Africa: The Challenges of Sustainability for Food and Nutrition Security. *International Journal of Agricultural Sustainability*, **0**, 1–12. Available at: https://doi.org/10.1080/14735903.2021.1940729.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biology*, **20**, 238. Available at: https://doi.org/10.1186/s13059-019-1832-y.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F. (2020) RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families. *Proc Natl Acad Sci U S A*, **117**, 9451–9457. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7196820/.

Garrison, E., Sirén, J., Novak, A.M., et al. (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, **36**, 875–879.

Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y. and Durbin, R. (2020) Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies. *Bioinformatics*, **36**, 2896–2898. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31971576.

Guo, M., Lian, Q., Mei, Y., et al. (2024) Analyzes of Pan-Genome and Resequencing Atlas Unveil the Genetic Basis of Jujube Domestication. *Nature Communications*, **15**, 9320. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11522667/.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013) QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics*, **29**, 1072–1075. Available at: https://doi.org/10.1093/bioinformatics/btt086.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R. (2008) Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, **9**, R7. Available at: https://doi.org/10.1186/gb-2008-9-1-r7.

He, Q., Tang, S., Zhi, H., et al. (2023) A Graph-Based Genome and Pan-Genome Variation of the Model Plant Setaria. *Nature Genetics*, **55**, 1232–1242. Available at: https://www.nature.com/articles/s41588-023-01423-w.

Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Marschall, T., Li, H. and Paten, B. (2023) Pangenome Graph Construction from Genome Alignments with Minigraph-Cactus. *Nat Biotechnol*, 1–11. Available at: https://www.nature.com/articles/s41587-023-01793-w.

Hu, H., Zhao, J., Thomas, W.J.W., Batley, J. and Edwards, D. (2025) The role of pangenomics in orphan crop improvement. *Nat. Commun.*, **16**, 118.

Hurgobin, B., Golicz, A.A., Bayer, P.E., et al. (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnol. J.*, **16**, 1265–1274.

Jayakodi, M., Schreiber, M., Stein, N. and Mascher, M. (2021) Building Pan-Genome Infrastructures for Crop Plants and Their Use in Association Genetics. *DNA Research*, **28**, dsaa030. Available at: https://doi.org/10.1093/dnares/dsaa030.

Jeffares, D.C., Jolly, C., Hoti, M., et al. (2017) Transient Structural Variations Have Strong Effects on Quantitative Traits and Reproductive Isolation in Fission Yeast. *Nature Communications*, **8**, 14061. Available at: https://www.nature.com/articles/ncomms14061.

Jiang, T., Cao, S., Liu, Y., Liu, S., Liu, B., Wang, G. and Wang, Y. (2023) Regenotyping Structural Variants through an Accurate Force-Calling Method. , 2022.08.29.505534. Available at: https://www.biorxiv.org/content/10.1101/2022.08.29.505534v3.

Kadereit, G., Borsch, T., Weising, K. and Freitag, H. (2003) Phylogeny of Amaranthaceae and Chenopodiaceae and the Evolution of C4 Photosynthesis. *International Journal of Plant Sciences*, **164**, 959–986. Available at: https://www.journals.uchicago.edu/doi/abs/10.1086/378649.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O. and Grau, J. (2018) Combining RNA-seq Data and Homology-Based Gene Prediction for Plants, Animals and Fungi. *BMC Bioinformatics*, **19**. Available at: https://link.springer.com/epdf/10.1186/s12859-018-2203-5.

Khoza, L.N., Mavengahama, S., Rensburg, W.S.J. van, Araya, H., Bairu, M.W., Venter, S., Denby, K., Nyathi, M.K. and Mojapelo, M. (2025) Yield and growth of different amaranth genotypes under varying water regimes. *Acta Hortic.*, 425–436.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**, 907–915. Available at: https://www.nature.com/articles/s41587-019-0201-4.

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L. and Pertea, M. (2019) Transcriptome Assembly from Long-Read RNA-seq Alignments with StringTie2. *Genome Biology*, **20**, 278. Available at: https://doi.org/10.1186/s13059-019-1910-1.

Lecy, K.S., Gonzalez, F.A., Becker-Reshef, I., et al. (2025) Agricultural research approaches for crops that nourish by improving nutrition, soil health, resilience and prosperity. *Nat. Food*, 1–4.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19451168.

Lian, Q., Huettel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., Roux, F., Schneeberger, K. and Mercier, R. (2024) A Pan-Genome of 69 Arabidopsis Thaliana Accessions Reveals a Conserved Genome Structure throughout the Global Species Range. *Nature Genetics*, **56**, 982–991. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11096106/.

Lightfoot, D.J., Jarvis, D.E., Ramaraj, T., Lee, R., Jellen, E.N. and Maughan, P.J. (2017) Single-Molecule Sequencing and Hi-C-based Proximity-Guided Assembly of Amaranth (Amaranthus Hypochondriacus) Chromosomes Provide Insights into Genome Evolution. *BMC Biology*, **15**, 74. Available at: https://doi.org/10.1186/s12915-017-0412-4.

Lin, Y.-P., Wu, T.-H., Chan, Y.-K., family=Zonneveld, given=Maarten and Schafleitner, R. (2022) De Novo SNP Calling Reveals the Genetic Differentiation and Morphological Divergence in Genus Amaranthus. *Plant Genome*, **15**, e20206. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/tpg2.20206.

Liu, Y., Du, H., Li, P., et al. (2020) Pan-Genome of Wild and Cultivated Soybeans. *Cell*, **182**, 162-176.e13. Available at: https://www.sciencedirect.com/science/article/pii/S0092867420306188.

Ma, X., Vaistij, F.E., Li, Y., et al. (2021) A Chromosome-Level Amaranthus Cruentus Genome Assembly Highlights Gene Family Evolution and Biosynthetic Gene Clusters That May Underpin the Nutritional Value of This Traditional Crop. *The Plant Journal*, **107**, 613–628. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.15298.

Mabhaudhi, T., O'Reilly, P., Walker, S. and Mwale, S. (2016) Opportunities for Underutilised Crops in Southern Africa's Post–2015 Development Agenda. *Sustainability*, **8**, 302. Available at: https://www.mdpi.com/2071-1050/8/4/302.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018) MUMmer4: A Fast and Versatile Genome Alignment System. *PLoS computational biology*, **14**, e1005944.

Monjotin, N., Amiot, M.J., Fleurentin, J., Morel, J.M. and Raynal, S. (2022) Clinical Evidence of the Benefits of Phytonutrients in Human Healthcare. *Nutrients*, **14**, 1712. Available at: https://www.mdpi.com/2072-6643/14/9/1712.

Montgomery, J.S., Giacomini, D., Waithaka, B., et al. (2020) Draft Genomes of Amaranthus Tuberculatus, Amaranthus Hybridus, and Amaranthus Palmeri. *Genome Biology and Evolution*, **12**, 1988–1993. Available at: https://doi.org/10.1093/gbe/evaa177.

Nattestad, M. and Schatz, M.C. (2016) Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly. *Bioinformatics*, **32**, 3021–3023.

Nyonje, W.A., Yang, R.-Y., Kejo, D., Makokha, A.O., Owino, W.O. and Abukutsa-Onyango, M.O. (2022) Exploring the Status of Preference, Utilization Practices, and Challenges to Consumption of Amaranth in Kenya and Tanzania. *Journal of Nutrition and Metabolism*, **2022**, 2240724. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/2240724.

Ou, S., Chen, J. and Jiang, N. (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.*, **46**, e126–e126.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017) Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat. Methods*, **14**, 417–419. Available at: https://www.nature.com/articles/nmeth.4197.

Paysan-Lafosse, T., Blum, M., Chuguransky, S., et al. (2023) InterPro in 2022. *Nucleic Acids Research*, **51**, D418–D427. Available at: https://academic.oup.com/nar/article/51/D1/D418/6814474.

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes. *Nat Commun*, **11**, 1432. Available at: https://www.nature.com/articles/s41467-020-14998-3.

Riggins, C.W. and Mumm, R.H. (2021) Amaranths. *Curr. Biol.*, **31**, R834–R835. Available at: https://www.cell.com/current-biology/abstract/S0960-9822(21)00756-9.

Sealy, R.L., Kenerley, C.M. and McWilliams, M., E.L. (1988) Evaluation of Amaranthus accessions for resistance to damping-off by Pythium myriotylum. *Plant Disease*, **72**, 985–989.

She, H., Liu, Z., Xu, Z., Zhang, H., Wu, J., Cheng, F., Wang, X. and Qian, W. (2024) Pan-Genome Analysis of 13 Spinacia Accessions Reveals Structural Variations Associated with Sex Chromosome Evolution and Domestication Traits in Spinach. *Plant Biotechnology Journal*, **22**, 3102–3117. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/pbi.14433.

Sirén, J., Monlong, J., Chang, X., et al. (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**, abg8871.

Soneson, C., Love, M.I. and Robinson, M.D. (2016) Differential Analyses for RNA-seq: Transcript-Level Estimates Improve Gene-Level Inferences. *F1000Research*, **4**, 1521. Available at: https://f1000research.com/articles/4-1521/v2.

Stetter, M.G., Vidal-Villarejo, M. and Schmid, K.J. (2020) Parallel Seed Color Adaptation during Multiple Domestication Attempts of an Ancient New World Grain. *Molecular Biology and Evolution*, **37**, 1407–1419. Available at: https://doi.org/10.1093/molbev/msz304.

Stiller, A., Garrison, K., Gurdyumov, K., Kenner, J., Yasmin, F., Yates, P. and Song, B.-H. (2021) From Fighting Critters to Saving Lives: Polyphenols in Plant Defense and Human Health. *International Journal of Molecular Sciences*, **22**, 8995. Available at: https://www.mdpi.com/1422-0067/22/16/8995.

Tang, H., Krishnakumar, V., Zeng, X., et al. (2024) JCVI: A versatile toolkit for comparative genomics analysis. *iMeta*, **3**, e211.

Tao, Y., Zhao, X., Mace, E., Henry, R. and Jordan, D. (2019) Exploring and Exploiting Pan-genomics for Crop Improvement. *Molecular Plant*, **12**, 156–169. Available at: https://www.sciencedirect.com/science/article/pii/S1674205218303836.

Thanyasiriwat, T., Kawicha, P., MacDonald, S., Sangdee, A., Tongyoo, P. and Denby, K. (2025) Pathogenicity and genome assembly of a Pythium aphanidermatum isolate causing damping-off in amaranth in controlled environment agriculture.

Venskutonis, P.R. and Kraujalis, P. (2013) Nutritional Components of Amaranth Seeds and Vegetables: A Review on Composition, Properties, and Uses. *Compr Rev Food Sci F*, **12**, 381–412. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/1541-4337.12021.

Wang, H., Xu, D., Wang, S., et al. (2022) Chromosome-Scale Amaranthus Tricolor Genome Provides Insights into the Evolution of the Genus Amaranthus and the Mechanism of Betalain Biosynthesis. *DNA Research*, dsac050. Available at: https://doi.org/10.1093/dnares/dsac050.

Wang, Y., Tang, H., DeBarry, J.D., et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*, **40**, e49. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326336/.

Wingett, S.W. and Andrews, S. (2018) FastQ Screen: A tool for multi-genome mapping and quality control. *F1000research*, **7**, 1338. Available at: https://doi.org/10.12688/f1000research.15931.2.

Winkler, T.S., Vollmer, S.K., Dyballa-Rukes, N., Metzger, S. and Stetter, M.G. (2024) Isoform-resolved genome annotation enables mapping of tissue-specific betalain regulation in amaranth. *N. Phytol.*, **243**, 1082–1100.

Zhang, R., Kuo, R., Coulter, M., et al. (2022) A High-Resolution Single-Molecule Sequencing-Based Arabidopsis Transcriptome Using Novel Methods of Iso-seq Analysis. *Genome Biol*, **23**, 149. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02711-0.

Zhang, X., Liu, T., Wang, J., et al. (2021) Pan-Genome of Raphanus Highlights Genetic Variation and Introgression among Domesticated, Wild, and Weedy Radishes. *Mol Plant*, **14**, 2032–2055. Available at: https://www.sciencedirect.com/science/article/pii/S167420522100318X.

Zhang, Z., Shi, Q., Wang, B., et al. (2022) Jujube Metabolome Selection Determined the Edible Properties Acquired during Domestication. *The Plant Journal*, **109**, 1116–1133. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.15617.

Zonneveld, M. van, Kindt, R., McMullin, S., Achigan-Dako, E.G., N'Danikou, S., Hsieh, W., Lin, Y. and Dawson, I.K. (2023) Forgotten Food Crops in Sub-Saharan Africa for Healthy Diets in a Changing Climate. *Proc National Acad Sci*, **120**, e2205794120. Available at: https://www.pnas.org/doi/10.1073/pnas.2205794120.

Zonneveld, M. van, Kindt, R., Solberg, S.Ø., N'Danikou, S. and Dawson, I.K. (2021) Diversity and Conservation of Traditional African Vegetables: Priorities for Action. *Divers. Distrib.*, **27**, 216–232. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.13188.

Figure legends

Figure 1. Leaf yield of the eight MAGIC population founder accessions in an irrigated field trial. A) Boxplots show yield measured as fresh weight of harvested leaves per plant. Yield was harvested from three replicate plots of 20 plants with yield per plant calculated to adjust for any loss of plants in a plot. Harvest number indicates each subsequent harvest and regrowth of the plants. **B)** Total leaf yield (i.e. the sum of all five harvests) per plant per accession. Letters indicate significant differences within each harvest as assessed by Tukey's honest significance test. Initials *A.c.* and *A.h.* indicate the species of each line with '*' indicating lines with admixture.

Figure 2. Variation in metabolite content between the eight Amaranthus MAGIC population parent accessions. A) Principal component analysis using the abundance of the 122 annotated metabolites significantly differing between MAGIC parent accessions. **B)** The relative abundance of the 122 metabolites across the 8 amaranth accessions organized by metabolite class. Presence absence values (PAV) based on raw data and metabolites varying significantly between the two species are indicated. '*' indicates accessions with genomic admixture.

Figure 3: Analysis of pangenome components across 11 *Amaranthus* genomes (eight MAGIC population parents and three published reference genomes, *A. cruentus, A. hypochondriacus* and *A. tricolor).* A) The proportion of core, dispensable and private gene families (including single copy genes unique to one accession) across the 11 amaranth genomes. **B)** The number of gene families shared between the 11 amaranth genomes. **C)** The proportion of core and total gene families in the pangenome as the number of genomes compared increases. **D)** The proportion of genes in each genome that are classified as core, dispensable or private in the pangenome of 11 accessions. **E)** mRNA length of genes in the three pangenome components. **F)** The proportion of core, dispensable and private genes that are functionally annotated via Gene Ontology or the Pfam database.

Figure 4. **Single Nucleotide Polymorphisms (SNPs) between the MAGIC founder lines and reference *A. cruentus* genome**. **A)** Principal component analysis using SNPs indicating clear separation of the MAGIC founder lines into two species groups. **B**) the distribution of SNPs along the genome. From the outer to inner layer is V1044371, V1044437, V1061487, V1061494, V1050446, V1060472, V1062433 and V1063769. **C**) the number of gene families that contain SNPs in at least one of the eight MAGIC founder lines. Blue indicates (core) gene

families where all MAGIC parents contain SNPs in the genes compared to the *A. cruentus* reference. **D**) the distribution of genes containing SNPs along the genome. Green is all genes with SNPs for each accession (accessions in the same order as in B). Blue indicates core genes with SNPs in all accessions and orange is single copy core genes with SNPs in all accessions.

Figure 5: Variation in copy number of putative disease resistance gene orthologues in the Amaranthus pangenome. Genes encoding orthologous nucleotide binding and leucine-rich repeat (NLR) proteins in two genomic locations are indicated for each of the eight founder accessions and the *A. cruentus* reference. The corresponding orthogroup is indicated by colour, with non-NLR genes flanking the region of interest indicated in grey. The chromosomal location is indicated at the left end of each genomic region.

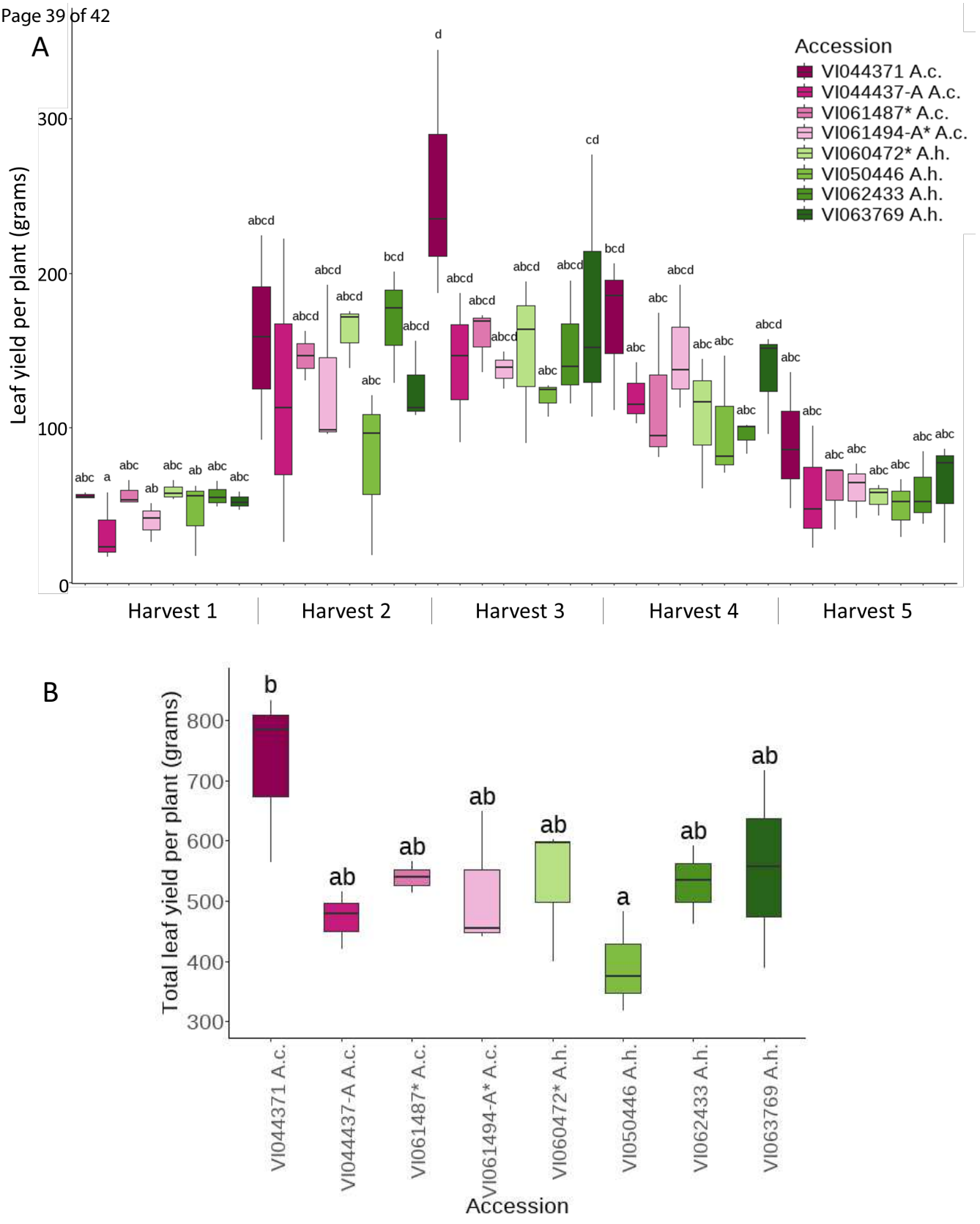
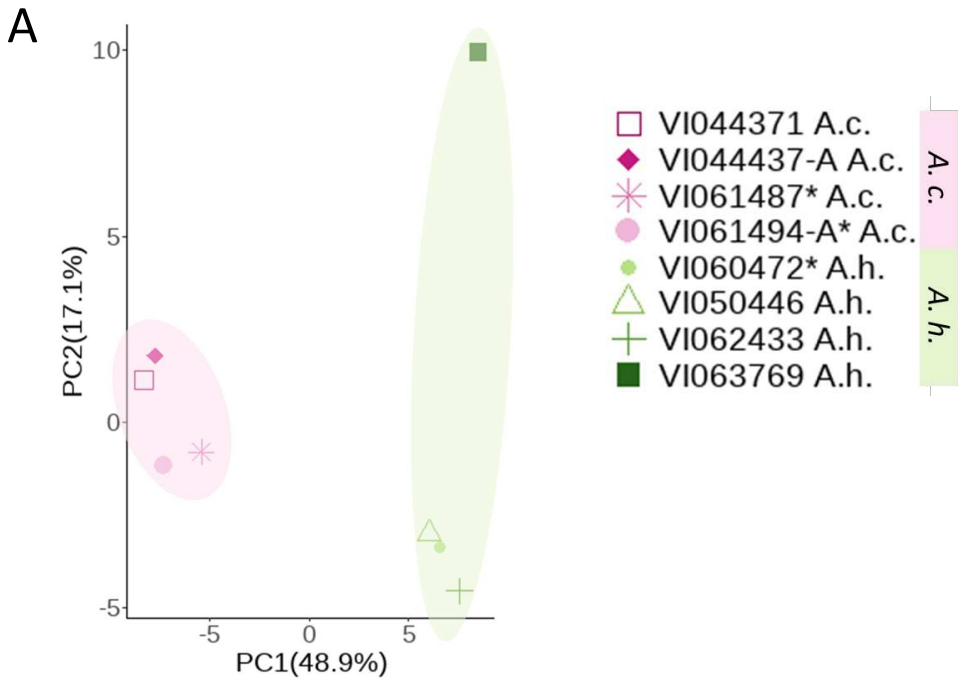


Figure 1. Leaf yield of the eight MAGIC population founder accessions in an irrigated field trial. A) Boxplots show yield measured as fresh weight of harvested leaves per plant. Yield was harvested from three replicate plots of 20 plants with yield per plant calculated to adjust for any loss of plants in a plot. Harvest number indicates each subsequent harvest and regrowth of the plants. **B)** Total leaf yield (i.e. the sum of all five harvests) per plant per accession. Letters indicate significant differences as assessed by Tukey's honest significance test. Initials *A. c.* and *A. h.* indicate the species of each line with '*' indicating lines with admixture.



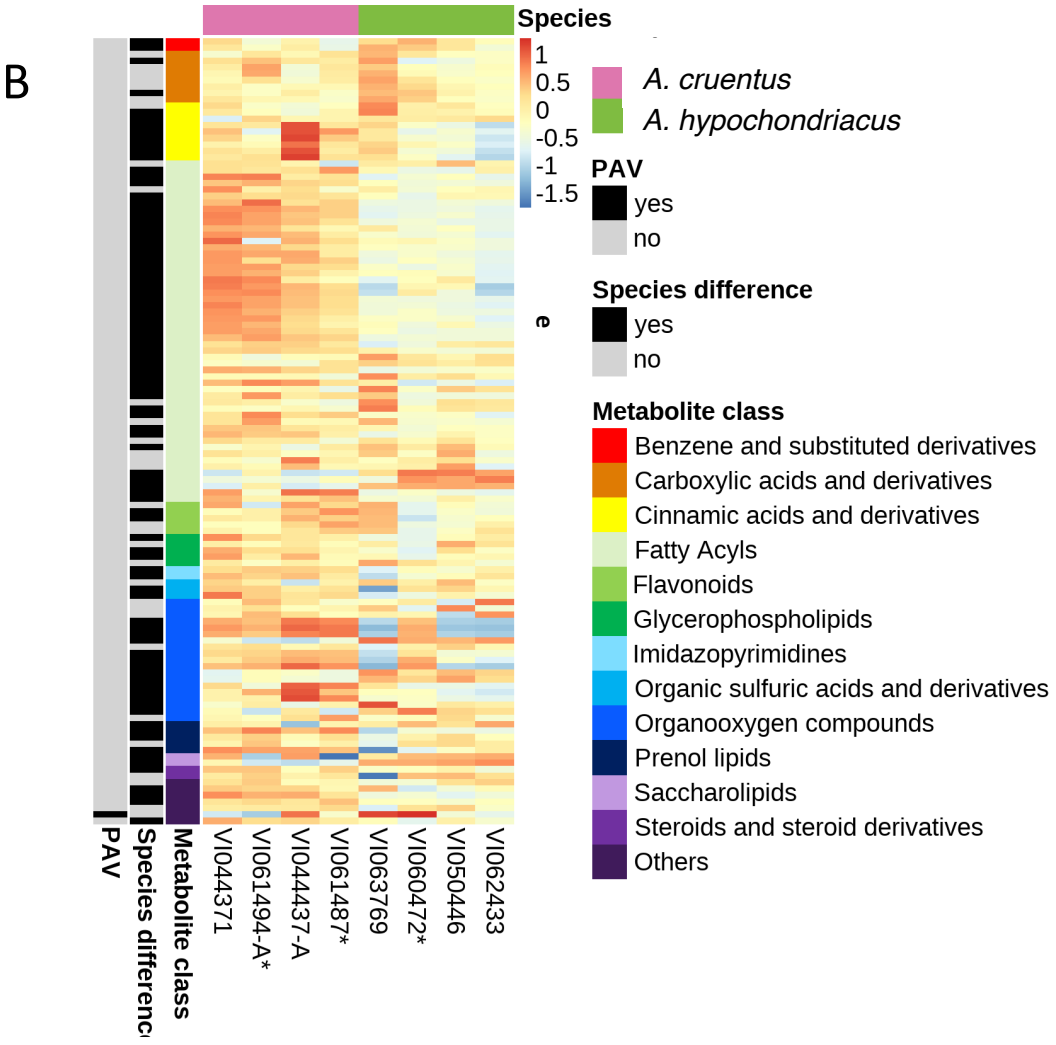
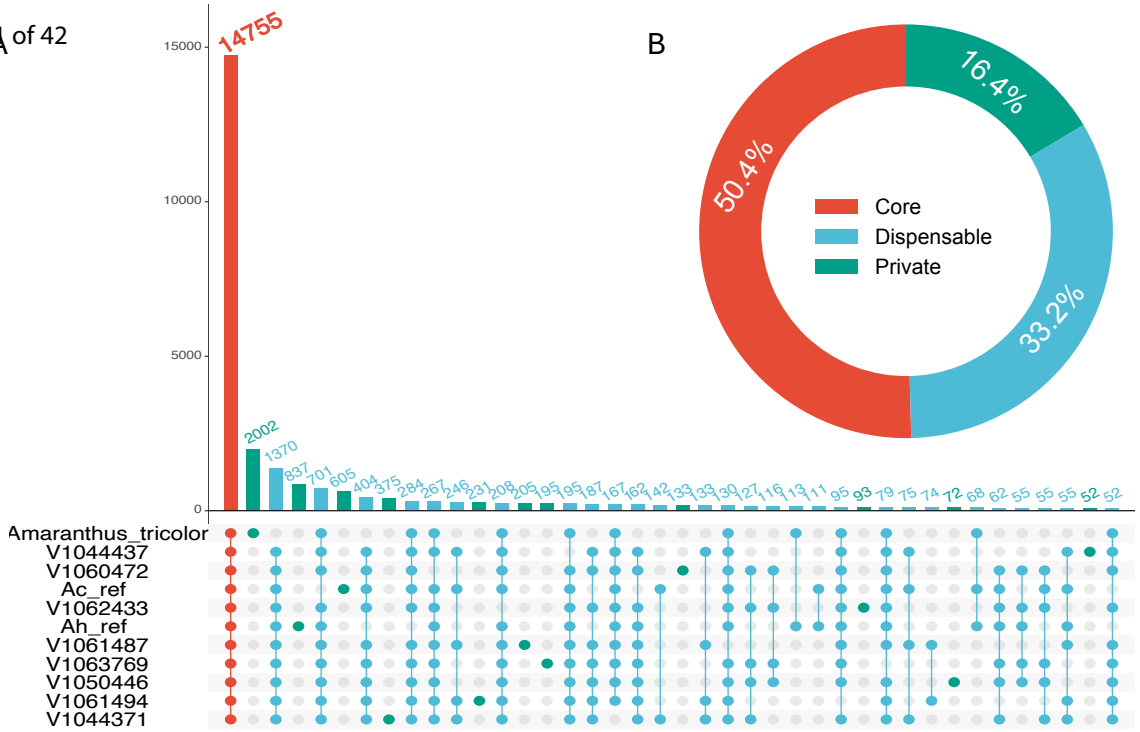


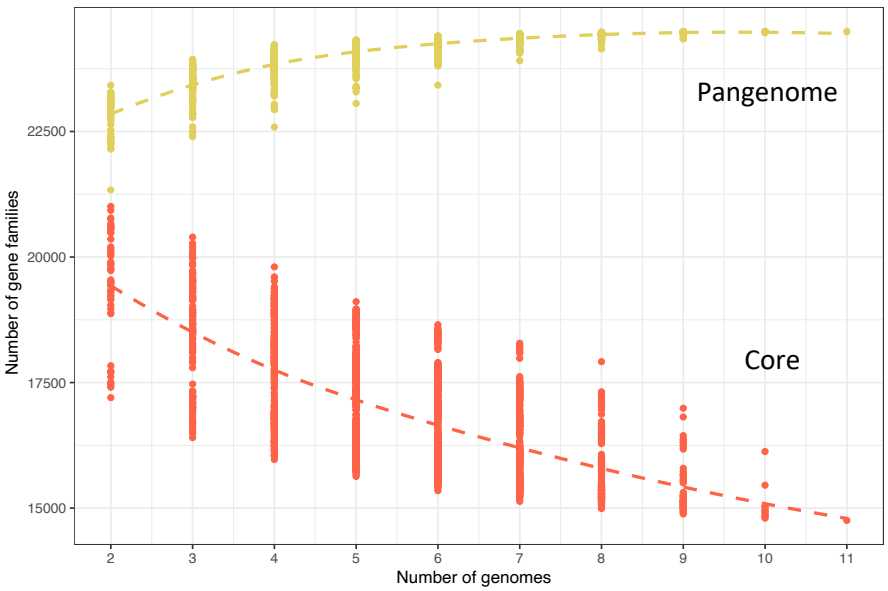
Figure 2. Variation in metabolite content between the eight Amaranthus MAGIC population parent accessions. **A)** Principal component analysis using the abundance of the 122 annotated metabolites significantly differing between MAGIC parent accessions. **B)** The relative abundance of the 122 metabolites across the 8 amaranth accessions organized by metabolite class. Presence absence values (PAV) based on raw data and metabolites varying significantly between the two species are indicated. '*' indicates accessions with genomic admixture.

A

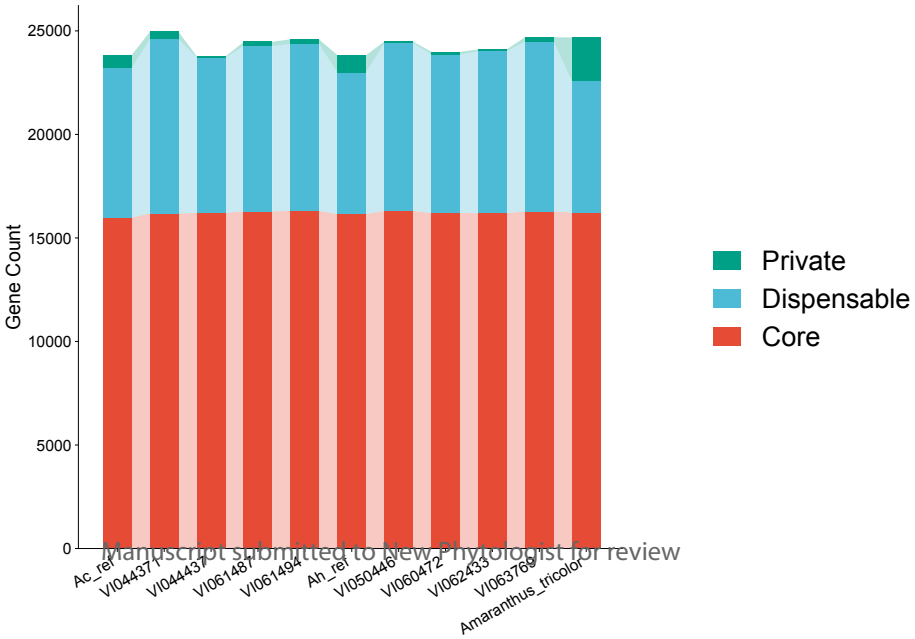


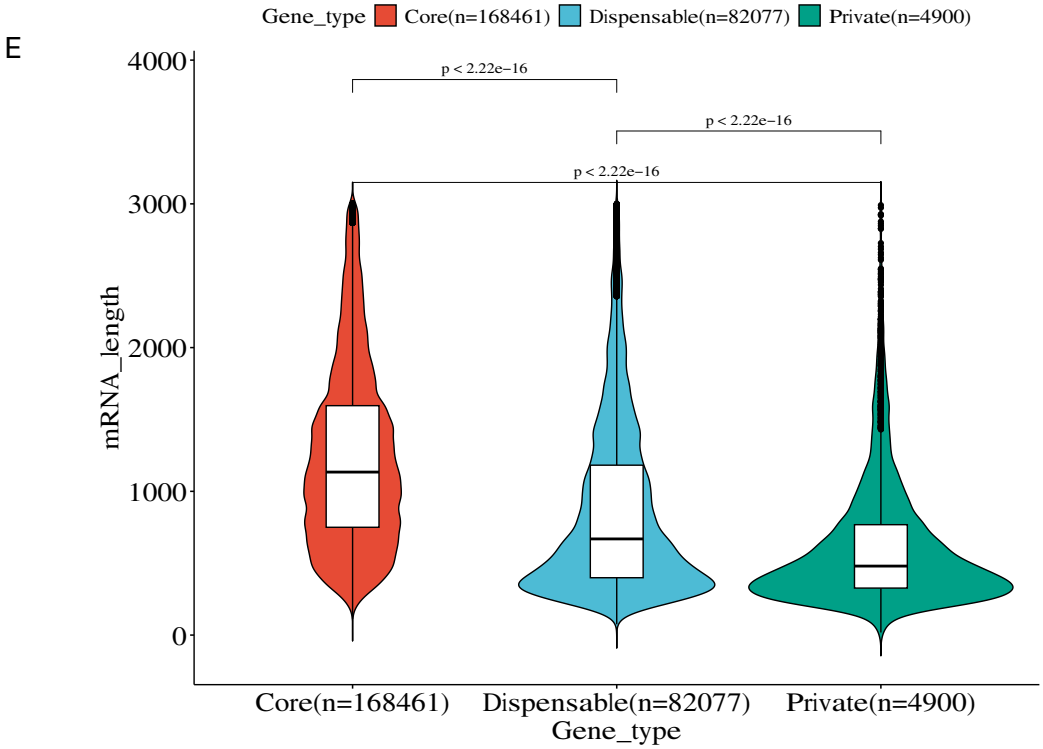
B

C



D





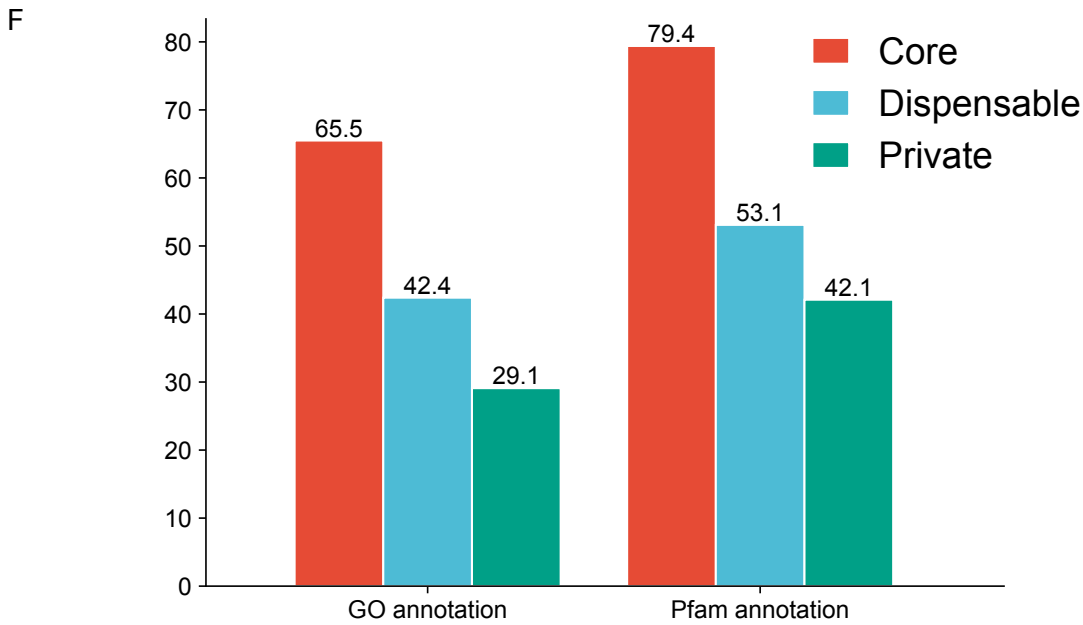


Figure 3: Analysis of pangenome components across 11 *Amaranthus* genomes (eight MAGIC population parents and three published reference genomes, *A. cruentus, A. hypochondriacus* and *A. tricolor).* **A)** The proportion of core, dispensable and private gene families (including single copy genes unique to one accession) across the 11 amaranth genomes. **B)** The number of gene families shared between the 11 amaranth genomes. **C)** The proportion of core and total gene families in the pangenome as the number of genomes compared increases. **D)** The proportion of genes in each genome that are classified as core, dispensable or private in the pangenome of 11 accessions. **E)** mRNA length of genes in the three pangenome components. **F)** The proportion of core, dispensable and private genes that are functionally annotated via Gene Ontology or the Pfam database.

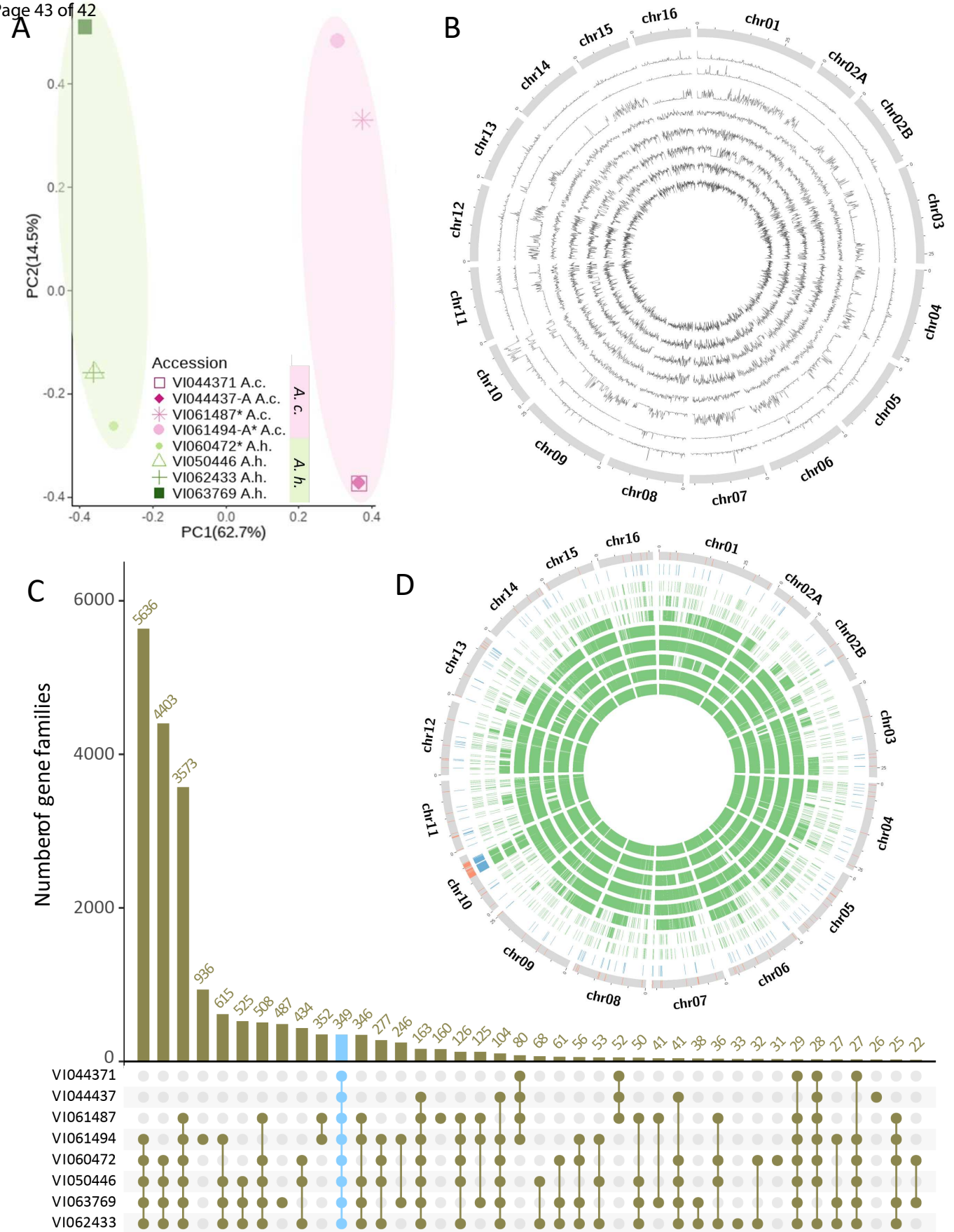


Figure 4. Single Nucleotide Polymorphisms (SNPs) between the MAGIC founder lines and reference *A. cruentus* genome. **A)** Principal component analysis using SNPs indicating clear separation of the MAGIC founder lines into two species groups. **B)** the distribution of SNPs along the genome. From the outer to inner layer is V1044371, V1044437, V1061487, V1061494, V1050446, V1060472, V1062433 and V1063769. **C)** the number of gene families that contain SNPs in at least one of the eight MAGIC founder lines. Blue indicates (core) gene families where all MAGIC parents contain SNPs in the genes compared to the *A. cruentus* reference. **D)** the distribution of genes containing SNPs along the genome. Green is all genes with SNPs for each accession (accessions in the same order as in **B)**. Blue indicates core genes with SNPs in all accessions and orange is single copy core genes with SNPs in all accessions.

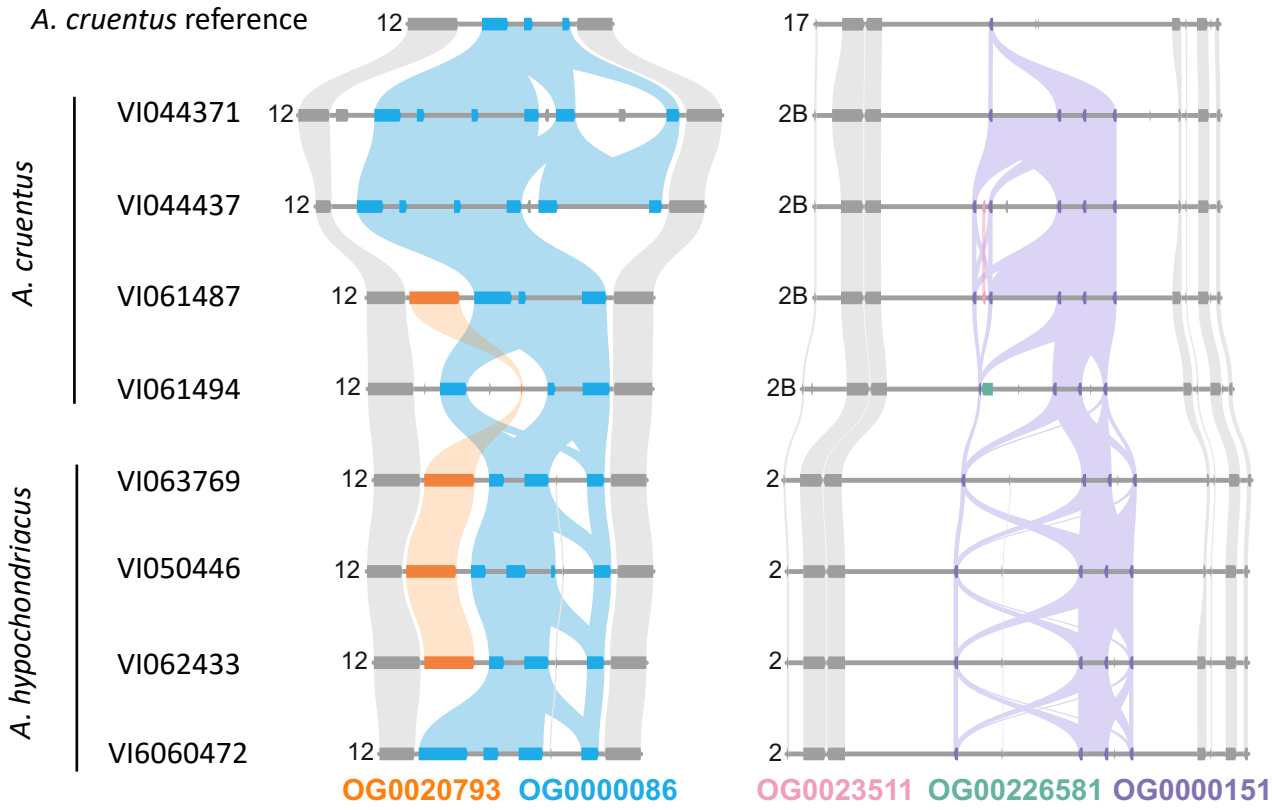


Figure 5: Variation in copy number of putative disease resistance gene orthologues in the Amaranthus pangenome. Genes encoding orthologous nucleotide binding and leucine-rich repeat (NLR) proteins in two genomic locations are indicated for each of the eight founder accessions and the *A. cruentus* reference. The corresponding orthogroup is indicated by colour, with non-NLR genes flanking the region of interest indicated in grey. The chromosomal location is indicated at the left end of each genomic region.