

This is a repository copy of *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.*

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/173360/>

Version: Published Version

---

**Article:**

Ma, Xiao, Vaistij, Fabian Emmanuel, Li, Yi et al. (10 more authors) (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*. pp. 613-628. ISSN 1365-313X

<https://doi.org/10.1111/tpj.15298>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:


<https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

## RESOURCE

# 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

Xiao Ma<sup>1,2,†</sup>, Fabián E. Vaistij<sup>3,†</sup>, Yi Li<sup>3,†</sup>, Willem S. Jansen van Rensburg<sup>4</sup>, Sarah Harvey<sup>3</sup>, Michael W. Bairu<sup>4</sup>, Sonja L. Venter<sup>4</sup>, Sydney Mavengahama<sup>5</sup>, Zemin Ning<sup>6</sup>, Ian A. Graham<sup>3</sup>, Allen Van Deynze<sup>7</sup>, Yves Van de Peer<sup>1,2,8,9</sup> and Katherine J. Denby<sup>3,\*</sup> 

<sup>1</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent 9054, Belgium,

<sup>2</sup>Center for Plant Systems Biology, VIB, Ghent 9054, Belgium,

<sup>3</sup>Department of Biology, Centre for Novel Agricultural Products (CNAP), University of York, Wentworth Way, York YO10 5DD, UK,

<sup>4</sup>Agricultural Research Council, Vegetable, Industrial and Medicinal Plants Research Campus, Private Bag X293, Pretoria 0001, South Africa,

<sup>5</sup>Crop Science Department, Faculty of Natural and Agricultural Sciences, North West University, P/Bag X2046, Mmabatho 2735, South Africa,

<sup>6</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK,

<sup>7</sup>Department of Plant Sciences, Seed Biotechnology Center, University of California, Davis, CA 95616, USA,

<sup>8</sup>Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa, and

<sup>9</sup>College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

Received 14 March 2021; revised 17 April 2021; accepted 21 April 2021.

\*For correspondence (e-mail katherine.denby@york.ac.uk).

†These authors are joint first authors.

## SUMMARY

Traditional crops have historically provided accessible and affordable nutrition to millions of rural dwellers but have been neglected, with most modern agricultural systems over-reliant on a small number of internationally traded crops. Traditional crops are typically well-adapted to local agro-ecological conditions and many are nutrient-dense. They can play a vital role in local food systems enhanced nutrition (particularly where diets are dominated by starch crops), food security and livelihoods for smallholder farmers, and a climate-resilient and biodiverse agriculture. Using short-read, long-read and phased sequencing technologies, we generated a high-quality chromosome-level genome assembly for *Amaranthus cruentus*, an under-researched crop with micronutrient- and protein-rich leaves and gluten-free seed, but lacking improved varieties, with respect to productivity and quality traits. The 370.9 Mb genome demonstrates a shared whole genome duplication with a related species, *Amaranthus hypochondriacus*. Comparative genome analysis indicates chromosomal loss and fusion events following genome duplication that are common to both species, as well as fission of chromosome 2 in *A. cruentus* alone, giving rise to a haploid chromosome number of 17 (versus 16 in *A. hypochondriacus*). Genomic features potentially underlying the nutritional value of this crop include two *A. cruentus*-specific genes with a likely role in phytic acid synthesis (an anti-nutrient), expansion of ion transporter gene families, and identification of biosynthetic gene clusters conserved within the amaranth lineage. The *A. cruentus* genome assembly will underpin much-needed research and global breeding efforts to develop improved varieties for economically viable cultivation and realization of the benefits to global nutrition security and agrobiodiversity.

**Keywords:** *Amaranthus cruentus*, biosynthetic gene clusters, gene annotation, genetic improvement, genome assembly, nutrition, traditional crop, underutilized crop.

## INTRODUCTION

Substantial investment in genetic improvement of crops during the Green Revolution (1960s to 1980s) led to the development of high-yield varieties of staple cereals (predominantly maize, wheat and rice) to feed the world's population. However, while these crops provide significant calories for human consumption, they are relatively low in protein (and in particular amino acids), vitamin and micronutrient content. Moreover, wide-scale cultivation of these staple crops has dramatically reduced agrobiodiversity and heightened the vulnerability of the global agricultural system. Orphan crops are typically traditional crops grown, traded and consumed within subsistence farming systems. They are not traded internationally but can play a major role in the diets and economy of low-income communities across the developing world. Importantly, they are accepted in local diets. Research and advances in orphan crops have lagged significantly behind that of staple crops, but orphan crops are often uniquely adapted to their local environments with enhanced nutritional content compared with more widely cultivated cereals, vegetables and fruits (Jamnadass *et al.*, 2020).

The genus *Amaranthus* comprises approximately 60 species of annual and short-lived perennial herbaceous plants, of which several are considered orphan crops. These species have great variability and phenotypic plasticity, and probably included introgression and hybridization between species (Sauer, 1957). This makes taxonomy based on morphological characteristics difficult. Amaranth plants are originally from the Americas where they have been grown mainly for grain for over 8000 years and once were a staple food for the Mayan, Aztec and Inca civilizations (Sauer, 1950). Amaranth plants grow widely in sub-Saharan Africa and South-East Asia where they have adapted to local conditions, are now considered native to these locations, and consumed mainly as a leafy vegetable. Amaranth is one of the most important informally traded leafy vegetables in Africa (Grubben and Denton, 2004). Traditionally, different amaranth species have been grown for grain or as a leafy vegetable: *Amaranthus hypochondriacus*, *Amaranthus caudatus* and *Amaranthus cruentus* species, with cream coloured seeds, are generally used for grain. These three species have been independently semidomesticated from the wild black-seeded *Amaranthus hybridus* (Brenner *et al.*, 2000; Stetter and Schmid, 2017; Stetter *et al.*, 2020). Three other black-seeded amaranth species, *Amaranthus blitum*, *Amaranthus dubius* and *Amaranthus tricolor* are cultivated for their leaves, with the latter being the most cultivated in South-East Asia (Andini *et al.*, 2013). However, the distinction between 'grain amaranth' and 'leafy amaranth' species is cultural rather than scientific, with the three grain species mentioned above also being used as leafy vegetables in sub-Saharan Africa

and Asia (Hauptli and Jain, 1984). Moreover, recently there have been efforts to develop *A. cruentus* as a dual-use (grain-leaf) crop (Hoidal *et al.*, 2019).

Both grain and leaf amaranth harbour high nutritional qualities (reviewed by Coelho *et al.*, 2018). The fibre and mineral content (such as iron, magnesium and calcium) of amaranth seed is typically higher than in grain of most cereals (Alvarez-Jubete *et al.*, 2009; Pedersen *et al.*, 1987), and amaranth seed contains elevated amounts of lysine, which is often a limiting amino acid in cereals. In addition, amaranth seed has a low gluten content that makes it suitable for consumption by people with coeliac disease and gluten sensitivity (Ballabio *et al.*, 2011), and a wide array of compounds with antioxidant properties (Coelho *et al.*, 2018). Amaranth leaf is a rich source of protein, vitamins (including A, C, B, E and K) and minerals at similar or higher levels than spinach and chard leaves (Venskutonis and Kraujalis, 2013). One cooked portion of *A. cruentus* leaves was estimated to contribute 89% of the daily vitamin A needs and 34% of the daily iron needs of a child (4–8 years) (van Jaarsveld *et al.*, 2014), a particularly striking example of the nutrition gap that traditional crops can help meet. Amaranth leaves also contain anti-nutrients such as oxalates, nitrates and phytates, which can sequester minerals and prevent uptake. However, analysis of Indian germplasm found no correlation between the accumulation of nutrients and anti-nutrients suggesting that new varieties can be found or developed with improved nutritional status (Prakash and Pal, 1991). As species with C<sub>4</sub> photosynthesis (Kadereit *et al.*, 2003), amaranth plants have increased water use efficiency and, hence, resilience to drought. In addition, amaranth is highly tolerant to other abiotic stresses such as salinity, heat and ultraviolet irradiance (Jamalluddin *et al.*, 2019; Omami and Hammes, 2006). All of these nutritional and physiological characteristics contribute to making amaranth a 'superfood' crop that can be cultivated on poor quality land in low-input agricultural systems (Joshi *et al.*, 2018).

There have been several efforts to develop genomic resources for amaranths to inform breeding programs producing improved varieties with enhanced agronomic and/or nutritional traits. The most detailed of these efforts describes the evolution at chromosome level of *A. hypochondriacus* compared with quinoa (*Chenopodium quinoa*) and sugar beet (*Beta vulgaris*), other members of the Amaranthaceae family (Lightfoot *et al.*, 2017). Here we present a high-quality genome assembly for *A. cruentus*, confirming the genome is  $n = 17$  in agreement with cytogenetic studies (Yssel *et al.*, 2019). We demonstrate that the whole genome duplication (WGD) seen in *A. hypochondriacus* occurred before divergence of this species and *A. cruentus*, and highlight intra- and inter-chromosomal rearrangements compared with *A. hypochondriacus*. We identify gene families that contracted and expanded in the *A. cruentus* genome, as well as

potential biosynthetic gene clusters that may play a role in the high nutritional content of this underutilized crop.

## RESULTS

### *Amaranthus cruentus* genome consists of 17 chromosomes

Genomic DNA (gDNA) was extracted from a black-seeded *A. cruentus* line 'Arusha' (Gerrano *et al.*, 2017) (Figure S1) and sequenced using Illumina paired-end (PE) reads, 10× PE reads (10×; Genomics, Pleasanton, CA, USA), Oxford Nanopore Technology (Oxford, UK) and Hi-C sequencing (Phase Genomics, Seattle, WA, USA) (Table S1). For the initial assembly, 24.5 Gb of Oxford Nanopore Technology sequencing data were assembled using WTDBG2 (Ruan and Li, 2020), with polishing using approximately 110 Gb of Illumina short-read data and Purge\_dups (Guan *et al.*, 2020) to remove haplotigs and overlaps. This assembly was combined with approximately 120 Gb of Hi-C reads using three-dimensional (3D)-DNA (Dudchenko *et al.*, 2017) to produce an initial chromosome-level assembly. Manual inspection of the Hi-C heat map and the placement of telomere signature sequences was used to adjust (and confirm) final chromosomal scaffolds. The Illumina and 10× Illumina reads were used with Pilon (Walker *et al.*, 2014) for polishing to produce the final assembly.

K-mer analysis gave a genome size estimate of 398.6 Mb (Figure S2), which is similar to the reported 376.4–403.9 Mb genome assembly size of *A. hypochondriacus* (Lightfoot *et al.*, 2017). Our the *A. cruentus* genome assembly was 370.9 Mb in length, and consisted of 625 scaffolds with a scaffold N50 of 21.7 Mb (Table 1); 98.5% of the genome assembly was anchored and oriented into 17 pseudochromosomes (Table S2, Figure S3), six of which had telomeric repeats at one end. The 17 chromosome-length scaffolds have been named corresponding to the *A. hypochondriacus* assembly (Lightfoot *et al.*, 2017), based on the degree of synteny between the two genomes. A heat map of the genome assembly (Figure S4) demonstrates the alignment of scaffolds into pseudochromosomes and the overall quality of the assembly. K-mer analysis indicated that the 'Arusha' *A. cruentus* line has very low levels of heterozygosity (0.07%, Figure S2), supported by the fact that the process to remove haplotigs and contig overlaps to produce a haploid genome sequence only removed a small fraction (5.9 Mb) from the initial assembly.

### Genome annotation reveals 25 477 protein-coding genes

Repetitive sequences account for approximately 58% of the 370.9 Mb *A. cruentus* genome, with 35% due to transposable elements (TEs) (Table 2). The proportion of TEs in *A. cruentus* (35%) is higher than that in its closest sequenced relative *A. hypochondriacus* (24%) and similar to the TE

**Table 1** *Amaranthus cruentus* genome assembly statistics

Statistic	<i>A. cruentus</i> (n = 17)
Assembly size (Mb)	370.9
Number of scaffolds	625
Scaffold N50 size (Mb)	21.7
Scaffold L50	8
Scaffold N90 size (Mb)	16.1
Scaffold L90	15
Longest scaffold (Mb)	34.6
Total length of pseudomolecules (Mb)	365.2
GC (%)	33.08
Total number of N	138 295
N per 100 kb	37

content in the related quinoa A- and B-genome diploid ancestors (*Chenopodium pallidicaule*, 38%; *Chenopodium suecicum*, 34%) and sugar beet (*B. vulgaris*, 31%) genomes (Table S3). Quinoa (*C. quinoa*) has a much higher content at 78%. The key difference in TE content between *A. hypochondriacus* and *A. cruentus*/diploid quinoa/beet relatives is a much lower content of long terminal repeats (LTRs) retrotransposons in *A. hypochondriacus* (12% versus 22–27%), in particular the content of Gypsy-like LTRs. Gypsy-like LTRs comprise <5% of the *A. hypochondriacus* genome versus 11% in *A. cruentus*, 10% in sugar beet, 23%

**Table 2** Repeat sequences identified in the *Amaranthus cruentus* genome assembly

Type of repeat	Subtype	Length (bp)	% of genome sequence	
DNA transposons	Classical (TIR and non-TIR)	21 832 669	5.9	
	Non-classical (rolling-circles)	1 459 238	0.4	
Retrotransposons	LTR	Ty1/Copia	51 562 759	13.9
		Gypsy/DIRS1	38 968 648	10.5
		Other	370 913	0.1
	Non-LTR	LINES	13 867 309	3.7
	SINEs	153 731	0.04	
Satellites		392 286	0.1	
Simple repeats		5 657 988	1.5	
Low complexity		800 581	0.2	
Unknown		78 967 558	21.29	
Total repetitive sequences		214 850 155	57.7	

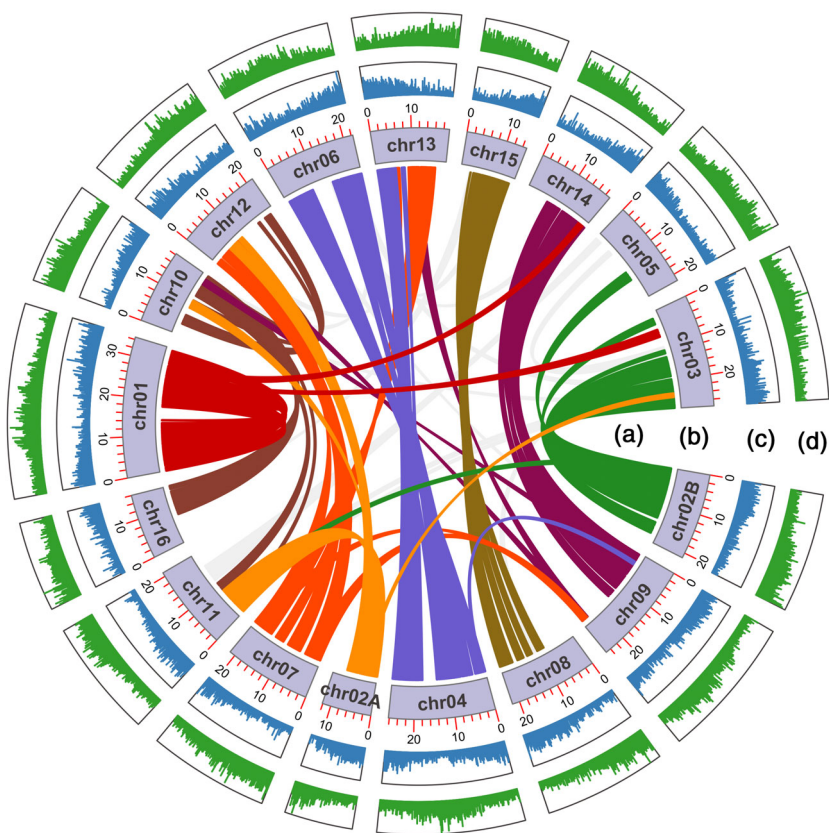
LTR, long terminal repeat.

and 19% in the quinoa ancestors and 33% in quinoa. Content of other classes of TE is similar across these related species. The distribution of TEs across the genome shows an inverse correlation to gene density, as expected from other plant genome assemblies (Figure 1).

Based on a combination of *ab initio* prediction, protein homology and RNA-sequencing (RNA-seq) evidence, we predicted 25 477 protein-coding genes in the *A. cruentus* genome. This is similar to the gene number in the *A. hypochondriacus* and *B. vulgaris* genomes and, as expected, considerably lower than the allotetraploid *C. quinoa* (Table 3). We generated full-length transcripts from 10 different *A. cruentus* tissues and/or conditions using Pacbio Isoseq technology. These full-length transcripts supported 51% of the *A. cruentus* genes with 73% supported by homology to other species. On average, protein-coding genes in *A. cruentus* are 4337 bp long and contain 4.86 exons, with these values similar to those of other sequenced Amaranthaceae species (Table 3); 94.3% of these *A. cruentus* genes could be assigned functions using InterProScan (Quevillon *et al.*, 2005), as well as protein-level homology analysis. BUSCO (v4.0.4) analysis of these predicted protein sequences showed that 90.5% of the

BUSCO gene set are found as complete genes (1461 of 1614), with another 40 genes present as incomplete sequences, bringing coverage to 93%. Of the complete BUSCO genes, 87.9% are single copy and 2.6% are found in duplicate. Seven per cent of the BUSCO gene set (113 genes) is missing.

As an additional check of the quality of the structural annotation, we compared the lengths of the coding sequences (CDS) of annotated genes in *A. cruentus* with their reciprocal best-hit orthologue in *A. hypochondriacus* (Figure S5). The vast majority of predicted protein-CDS were the same length in both species; however, a greater proportion had a longer sequence in the *A. cruentus* genome (compared with *A. hypochondriacus*) than vice versa, suggesting an improved annotation of the *A. cruentus* genome compared with *A. hypochondriacus* (i.e. with more full-length CDS). Furthermore, the distributions of transcript, CDS, exon and intron length are all very comparable with three closely related species (Figure S6). The density of protein-coding genes varies along each chromosome (Figure 1), as expected from other plant genomes, presumably reflecting reduced gene density around centromeric regions.



**Figure 1.** The landscape of genome assembly and annotation of *Amaranthus cruentus*.

Tracks from the inner to outside correspond to: a, syntenic blocks; b, pseudomolecules (with length in Mb); c, gene density; d, Transposable element density.

**Table 3** Amaranthaceae comparative gene annotation

Statistic	A. <i>cruentus</i>	A. <i>hypochondriacus</i>	B. <i>vulgaris</i>	C. <i>quinoa</i>
Protein coding genes	25 477	23 883	26 920	44 776
Mean gene length, bp	4337	4102	4302	4362
Mean CDS length, bp	1141	1066	1057	1274
Mean exon per gene	4.86	4.87	4.47	5.45
Mean exon length, bp	235	219	236	234
Mean intron length, bp	829	786	937	695

Genome statistics are derived from this study (*Amaranthus cruentus*), Lightfoot *et al.*, (2017) (*Amaranthus hypochondriacus*), Dohm *et al.*, (2014) (*Beta vulgaris*) and Jarvis *et al.*, (2017) (*Chenopodium quinoa*).

CDS, coding sequences.

In addition to protein-coding genes and repetitive sequences, we also annotated 131 microRNAs (miRNAs) from 33 families with high confidence and predicted putative target genes for 32 of these families (Table S4a,b). As expected, the predicted target genes represent a broad set of functions [61% could be associated with a gene ontology (GO) term through homology) with enrichment for genes involved in metabolism, regulatory processes, response to hormones and transport associated processes and functions (Table S4c). Nine hundred and twenty-six genes encoding transfer RNAs and 325 genes encoding ribosomal RNAs were predicted in the genome assembly. The *A. cruentus* genome and annotation information are publicly available within the African Orphan Crop Consortium portal of the ORCAE database (<https://www.bioinformatics.psb.ugent.be/gdb/amaranthus/>) (Sterck *et al.*, 2012; Yssel *et al.*, 2019).

### Comparative genomics highlights a shared WGD in the *Amaranthus* lineage

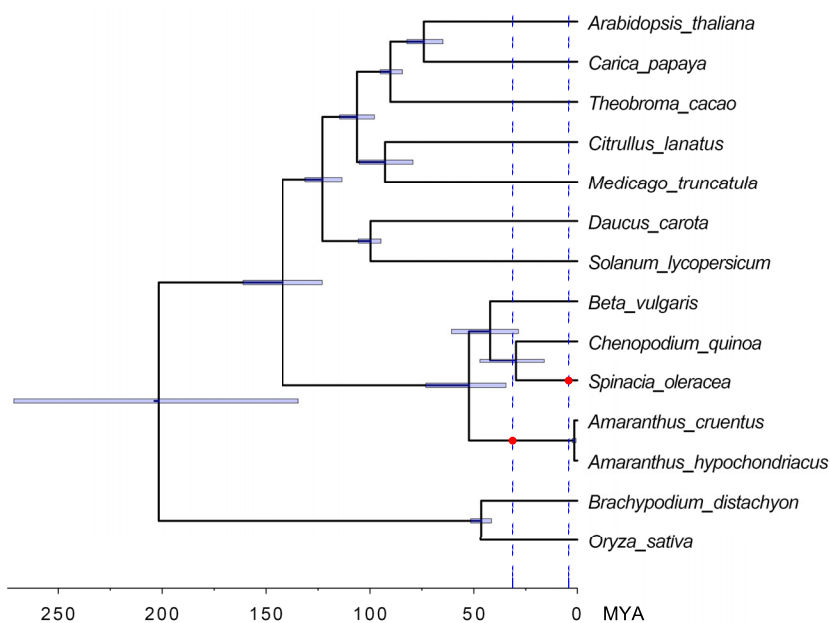
Using the *wgd* package (Zwaenepoel and Van de Peer, 2019), we inferred the distribution of the rate of synonymous substitutions per synonymous site (*Ks*) for the whole paranome and anchor pairs (homeologues) in the *A. cruentus* and *A. hypochondriacus* genomes. Using the anchor pairs, a very clear signature of WGD was present at *Ks* = 0.54 in both *A. cruentus* and *A. hypochondriacus* (Figure S7), consistent with the published results for *A. hypochondriacus* (Lightfoot *et al.*, 2017), and indicating that these two species share the same WGD event that occurred between 18 and 34 million years ago (MYA). This is confirmed by the cross-species (*A. cruentus*–*A. hypochondriacus*) *Ks* comparison indicating divergence

between these two genomes is much more recent (Figure S7). Within the Amaranthaceae, previous studies indicated a single relatively recent WGD event in *C. quinoa* that occurred between 3.3 and 6.3 MYA (Jarvis *et al.*, 2017) and no recent WGD event in *B. vulgaris*. Hence, the new *A. cruentus* genome confirms the independent WGD events in the amaranth and quinoa lineages and places speciation of *A. cruentus* and *A. hypochondriacus* significantly after the amaranth WGD event (Figure S7).

Using 106 single-copy gene orthologues across 14 angiosperm species, including *A. cruentus* and *A. hypochondriacus*, we built a phylogenetic tree. As expected, this shows that *A. cruentus* is most closely related to *A. hypochondriacus* and places the divergence of the two species approximately 1.45 (0.61–2.41) MYA (Figure 2). In this analysis, *C. quinoa* appeared more closely related to spinach (*Spinacia oleracea*) than *B. vulgaris* [as expected from previous molecular analyses (Kadereit *et al.*, 2003) with an estimated divergence time of approximately 30 (16.5–47.4) MYA] (Figure 2).

### Chromosome evolution differs between *Amaranthus cruentus* and *Amaranthus hypochondriacus*

As indicated above, the amaranth and quinoa lineages have undergone independent WGD events, presumably from an ancestor with a haploid chromosome number of 9. *Chenopodium quinoa* has retained its haploid chromosome number of *n* = 18 during the 3.3–6.3 million years since genome duplication, but the haploid chromosome numbers of *A. cruentus* and *A. hypochondriacus* have been reduced to *n* = 17 and *n* = 16 respectively, over the 18–34 million years since their shared WGD event. Syntenic analysis in *A. hypochondriacus* (Lightfoot *et al.*, 2017) indicated that this reduction was likely due to the loss of one homoeologue of chromosome 5 and the fusion of the two homoeologues of chromosome 1 to produce *n* = 16. A similar pattern is evident in *A. cruentus* (Figure 1). We examined the homologous relationships among the 17 *A. cruentus* pseudochromosomes using MCScanx (Wang *et al.*, 2012) and found 4561 collinear genes. Chromosome 1 has 496 collinear genes that showed collinearity to the other half of chromosome 1 and very little collinearity to any other pseudochromosome (Figure 1), indicating the likely fusion of the original subgenome homeologues. Chromosome 1 is also the longest of the *A. cruentus* chromosomes (Table S2). Chromosome 5 has minimal similarity to the other pseudochromosomes (just one collinear block with chr02B) suggesting that (as in *A. hypochondriacus*) its homoeologous copy was lost during evolution of modern amaranths, and that the loss of chromosome 5 occurred before speciation of *A. cruentus* and *A. hypochondriacus*. Ten of the other 15 pseudochromosomes (chr02B, chr03, chr04, chr06, chr08, chr09, chr10, chr14, chr15 and chr16) have clearly identifiable one-to-one homoeologous



**Figure 2.** Time-calibrated phylogenetic tree, based on 106 gene families consisting of a single gene copy in each lineage using MCMCtree (95% confidence intervals indicated by blue horizontal bars). Red circles represent independent whole genome duplication events in the amaranth and quinoa lineages, with the dashed lines highlighting the time of these. MYA, million years ago.

relationships (chr02B–chr03, chr04–chr06, chr08–chr15, chr09–chr14, chr10–chr16), while additional rearrangements have given rise to five pseudochromosomes (chr02A, chr07, chr11, chr12 and chr13) showing substantial homology to two pseudochromosomes (chr02A–chr11 and chr12; chr07–chr12 and chr13; and chr13–chr07 and chr04) (Figure 1).

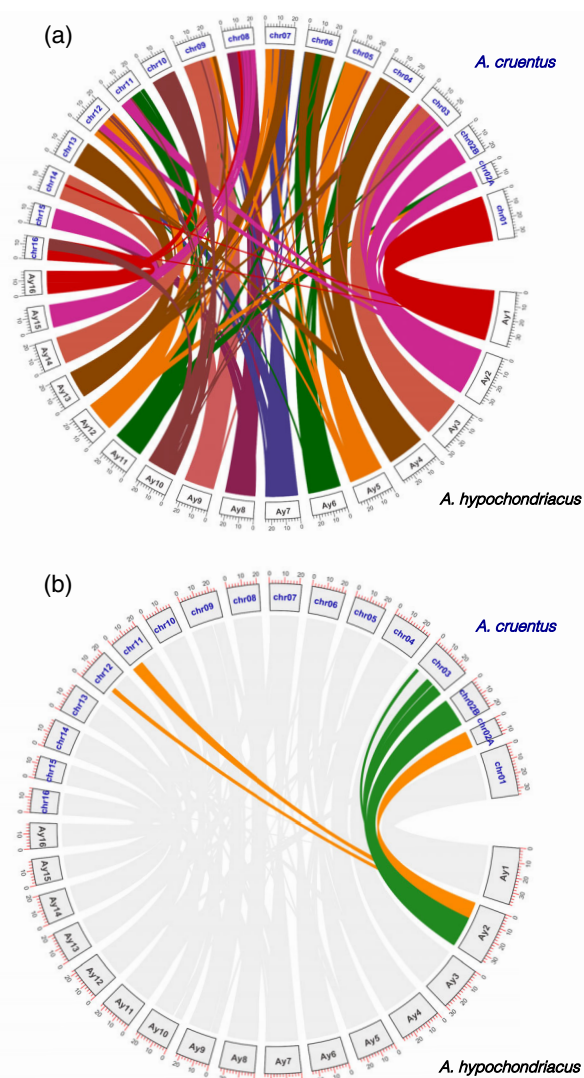
Comparative syntenic analysis of the *A. cruentus* genome ( $n = 17$ ) and *A. hypochondriacus* genome ( $n = 16$ ) showed a high level of collinearity between the two genomes with MCScanx identifying 35 242 collinear genes representing 74% and 71% of the *A. hypochondriacus* and *A. cruentus* gene complements, respectively. Most collinear blocks on *A. hypochondriacus* have one or two homoeologous copies in *A. cruentus* (and vice versa) indicating a predominant 1:1 relationship with translocations of a number of collinear blocks (Figure 3). However, the entire blocks on chromosome 02A and 02B in *A. cruentus* share collinearity with chromosome 2 in *A. hypochondriacus* (Figure 3) indicating a fission of chromosome 2 into 02A and 02B in *A. cruentus* (subsequent to the chromosome 5 copy loss and chromosome 1 fusion shared with *A. hypochondriacus*) to produce  $n = 17$ . Consistent with this, chr02A is the shortest of the 17 *A. cruentus* chromosomes (Table S2).

There was extensive synteny between the genomes of *A. cruentus* and *B. vulgaris* ( $n = 9$ ). MCScanx confirmed 13 963 collinear genes and most collinear blocks showed a 2:1 orthologous relationship (*A. cruentus*/*B. vulgaris*), resulting from the WGD event in the *A. cruentus* lineage that is not shared with *B. vulgaris*. For example, Bvchr6–chr02B and chr03; Bvchr7–chr08 and chr15; and Bvchr8–chr09 and chr14 (Figure S8). Several more complex

orthologous relationships are evident, for example, large collinear blocks on Bvchr5 are orthologous to regions on chr02A, chr07, chr11 and chr12 in *A. cruentus*, presumably due to multiple chromosomal rearrangements. However, we find only a single *A. cruentus* chromosome (chr1) with large-scale orthology to Bvchr9, supporting the previous conclusion that *A. cruentus* experienced a chromosome fusion of chromosome 1 subsequent to the last WGD event.

#### ***Amaranthus cruentus* gene family evolution**

We compared gene families across the 14 genomes presented in the phylogenetic tree (Figure 2; Table S5). Overall, 78.8% of the 454 388 genes from these genomes could be assigned to orthogroups (21 039 orthogroups) (Emms and Kelly, 2019). One-third of these orthogroups are found across all 14 species, and <4% (773 orthogroups) are species-specific. Twenty orthogroups are specific to *A. cruentus*, representing 88 genes, of which only 19 had Pfam annotations. Notably, two of these *A. cruentus*-specific genes had inositol 1,3,4-trisphosphate 5/6-kinase, ATP-grasp domain Pfam annotations. This enzyme plays a key role in the production of inositol hexaphosphate (InsP6), also known as phytic acid, an anti-nutrient that chelates minerals in the gastrointestinal tract inhibiting their uptake. *Amaranthus cruentus* is known to contain phytate, with some studies finding higher levels in Amaranths compared with cereals such as rice (Lorenz and Wright, 1984). Phytic acid also plays a role in regulating intracellular calcium ion signalling, a key component of plant stress responses, and acts as a co-factor to the jasmonic acid and auxin receptors in *Arabidopsis* (Hou *et al.*, 2016). Other *A. cruentus*-specific



**Figure 3.** Syntenic relationship between the *Amaranthus cruentus* genome and *Amaranthus hypochondriacus* genome.

Size of chromosomes is indicated in Mb around the outside with syntenic blocks of genes shown in the same colour.

(a) Synteny with colouring based on *A. hypochondriacus* chromosomes.

(b) Highlights the fission of chromosome 2 in *A. cruentus* with separate colouring of the two halves of *A. hypochondriacus* chromosome 2.

orthogroups of note with respect to plant stress responses include genes encoding a protein with an auxin response factor domain and a protein containing the NB-ARC domain, a key domain in plant disease resistance proteins.

We used the *CAFE* software (Han *et al.*, 2013) to identify gene families expanded or contracted across the same 14 genomes (Figure 4). Five hundred and three gene families were expanded in *A. cruentus* compared with the other 13 genomes, and 1099 families contracted. The expanded gene families were enriched for multiple GO terms relating to ion transport and for UDP-glycosyl transferase (UGT) activity (Figure S9). Several studies have found higher

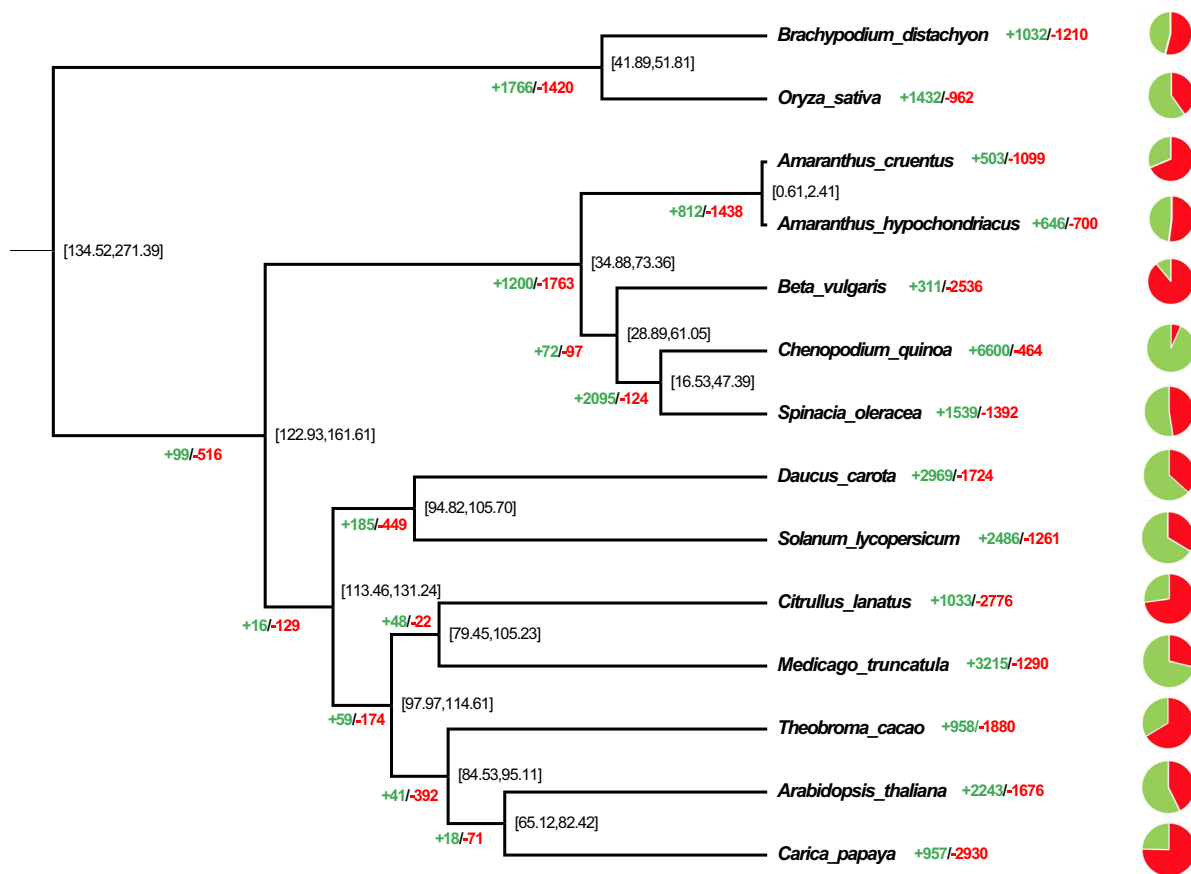
levels of minerals in Amaranths (and other traditional leafy vegetables) compared with conventional species such as spinach and kale (Odhav *et al.*, 2007), and it is possible that expanded cation transporter families contribute to this. UGTs are a family of proteins that transfer UDP-activated sugar moieties on to a variety of substrates. In plants, UGTs are involved in the production of defence compounds, the activity of phytohormones, the activity and stability of secondary metabolites, as well as driving xenobiotic detoxification (Louveau and Osbourn, 2019). With the exception of betalains (Brockington *et al.*, 2011), the scientific community is just starting to explore the range of specialized metabolites produced in amaranths (e.g. Sarker and Oba, 2020) and the relevance of an increased number of UGTs may become clearer.

A larger number of gene families are contracted in *A. cruentus* compared with the other 13 genomes, and markedly, the majority of the GO terms enriched in these families are associated with fundamental gene expression processes, e.g. messenger RNA splicing, ribosomal RNA processing and ribosome biogenesis. Why these families have been contracted in *A. cruentus* is not clear; the genome has a similar number of genes to related plant species (Table 3), and long-read RNA-seq analysis of 10 different *A. cruentus* tissues detected 51% of these. Interestingly, gene families annotated with GO terms associated with messenger RNA splicing are also contracted when comparing the two amaranth species (*A. hypochondriacus* and *A. cruentus*) with the other plant genomes in this analysis (Figure S10). The expansion of gene families annotated with GO terms relating to ion transport are also enriched in both *A. hypochondriacus* and *A. cruentus* compared with the other plant genomes in this analysis (Figure S10).

### Secondary metabolite biosynthetic gene clusters

*Amaranthus*, *Beta*, *Spinacia* and *Chenopodium* are four genera within the Amaranthaceae family that produce tyrosine-derived betalain pigments rather than anthocyanins (Brockington *et al.*, 2011; Clement and Mabry, 1996). Three key genes in betalain biosynthesis are: (i) a cytochrome P450 with tyrosine hydroxylase activity (Polturak *et al.*, 2016); (ii) L-DOPA 4,5-dioxygenase (DODA) (Christinet *et al.*, 2004) with all enzymes shown to have DODA activity to date being DODA $\alpha$  orthologues (Sheehan *et al.*, 2020); and (iii) L-DOPA oxidase (catalysed by the cytochrome P450, CYP76AD1, with again the  $\alpha$  clade seeming specific to betalain synthesis) (Brockington *et al.*, 2015; Hatlestad *et al.*, 2012). In *B. vulgaris* the betalain-specific isoforms of DODA $\alpha$ 1 and CYP76AD1 are co-located on chromosome 2, with a single gene between them (Sheehan *et al.*, 2020). Orthologues of these two genes are also co-located in *A. hypochondriacus* (chromosome 16) and in *C. quinoa*, with the grouping present on two chromosomes in this species consistent with the recent WGD (Sheehan





**Figure 4.** Gene family contraction and expansion across 14 plant genomes.

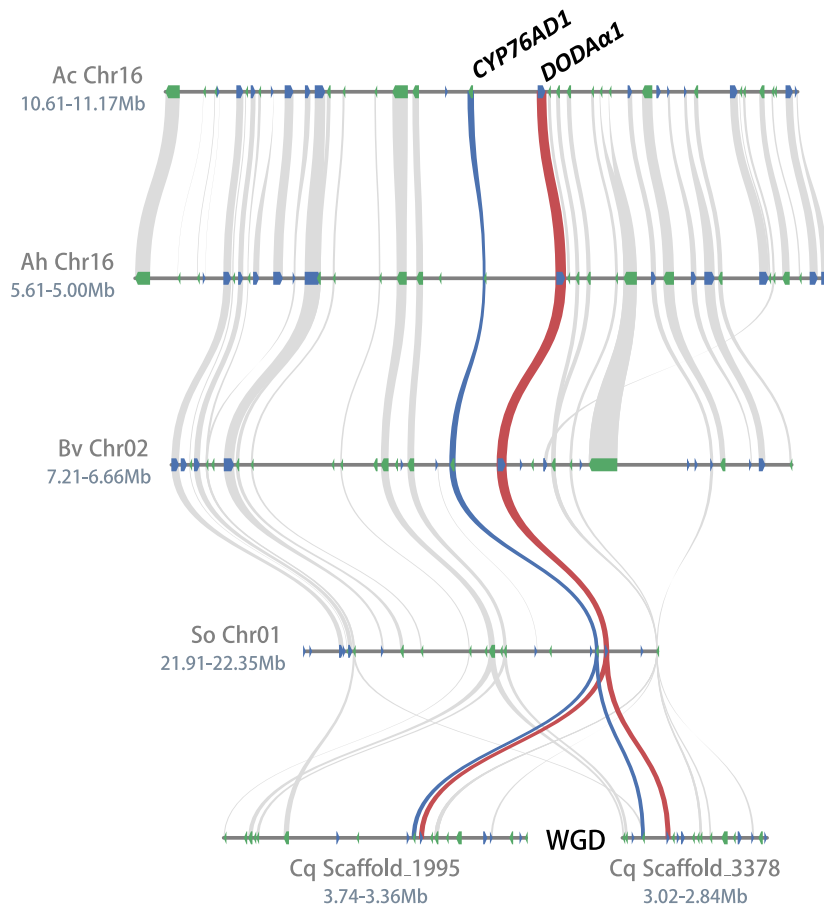
Number of expanded gene families in each species, or at each lineage branch point, is shown in green, with the number of contracted families in red. Overall proportion of expanded/contracted gene families in each species is represented by a pie diagram. *Chenopodium quinoa* has a high proportion of expanded gene families due to the recent whole genome duplication event.

*et al.*, 2020). Analysis of our *A. cruentus* genome indicated that *DODA21* and *CYP76AD1* orthologues are also co-located in *A. cruentus*, on chromosome 16, with this biosynthetic co-localization also conserved in the *S. oleracea* genome (Figure 5). Within spinach and quinoa, tandem gene duplications (in addition to the recent WGD event in the quinoa lineage) appear to have given rise to additional copies of these two enzymes, with a tandem pair of the two genes in spinach, and a tandem duplication and triplication of the two biosynthetic genes in quinoa (Figure S11).

Given the importance of secondary metabolites to the nutritional content of amaranth, particularly as a leafy vegetable, we used PLANTISMASH (Kautsar *et al.*, 2017) to identify additional potential secondary metabolite biosynthetic gene clusters in the *A. cruentus* and *A. hypochondriacus* genomes. There is increasing evidence that, as in bacteria and fungi, genes encoding secondary metabolite biosynthetic enzymes involved in the same pathway are found in clusters in plant genomes (Nützmann *et al.*, 2016). With constraints that each cluster must include at least three

biosynthetic genes of two different types (and closely related duplicate genes are counted only once), PLANTISMASH identified 22 clusters in *A. cruentus* and 23 clusters in *A. hypochondriacus* (Table S6a,b). Ten of these clusters were clearly orthologous in *A. cruentus* and *A. hypochondriacus* (containing the same core domains, located on homologous chromosomes and with similar overall cluster sizes) (Table S6c), strengthening the evidence that they are genuine biosynthetic gene clusters. These shared clusters include those likely to synthesize lignans, terpenes, alkaloids and polyketides, as well as clusters containing UGT genes for glycosylation (a gene family expanded in these two species; Figure S9) and likely synthesizing saccharide compounds. Within these clusters the presence of non-biosynthetic genes varies between the two amaranth species, and biosynthetic gene inversions and variation in the exact biosynthetic gene number is seen in several orthologous clusters demonstrating genome rearrangements that are specific to each *Amaranthus* species.

Six of these 10 biosynthetic gene clusters show conservation across other genomes in the Amaranthaceae. Three



**Figure 5.** Co-localization of the betalain biosynthetic genes across the Amaranthaceae.

DODA $\alpha$ 1 and CYP76AD1 genes are co-localized in the genomes of five Amaranthaceae species: Ac, *Amaranthus cruentus*; Ah, *Amaranthus hypochondriacus*; Bv, *Beta vulgaris*; Cq, *Chenopodium quinoa*; So, *Spinacia oleracea*. Blue and green rectangles indicate predicted gene models, with colour showing the gene orientation (blue, – strand; green, + strand). Orthologous gene pairs are linked by grey lines, with red and blue lines linking orthologous DODA $\alpha$ 1 and CYP76AD1, respectively.

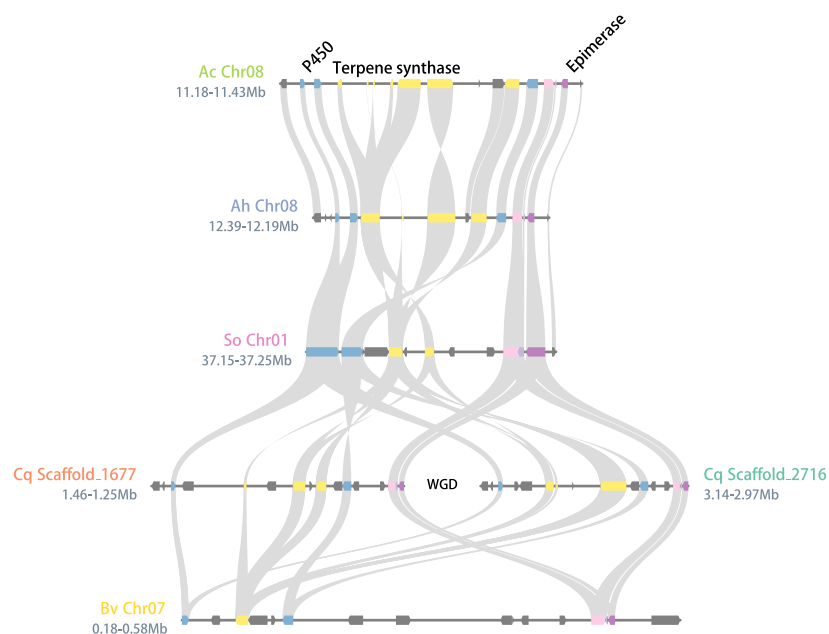
of these conserved biosynthetic gene clusters involve multiple copies of two different types of biosynthetic enzyme (Figure S12a-c).

In two clusters, three different types of biosynthetic enzyme are conserved across *A. cruentus* and *A. hypochondriacus* but only two of these enzyme types are conserved across all five genomes (Figure S12d,e). However, one cluster is particularly striking with genes encoding three different types of biosynthetic enzymes conserved across all five Amaranthaceae genomes analysed (*A. cruentus*, *A. hypochondriacus*, *B. vulgaris*, *S. oleracea* and *C. quinoa*) (Figure 6). This cluster contains genes encoding the core domains of terpene synthase (type II squalene-hopene cyclase), epimerase and cytochrome P450 enzymes.

## DISCUSSION

Here we present a high-quality chromosome-level genome assembly of the traditional crop *A. cruentus*. This genome

is the second *Amaranthus* chromosome-level genome to be published (after *A. hypochondriacus*; Lightfoot *et al.*, 2017) and will drive molecular plant breeding initiatives for these neglected crops, which have much to offer in terms of climate-resilience, nutrition, food security and small-holder farmer livelihoods, as well as gluten-free cereal alternatives. Enhanced agrobiodiversity through new and diversified crops will be vital for the environmental, economic and social sustainability of future agricultural systems. The World Vegetable Center (WVC) has developed a number of improved lines for East Africa, which have been partially adopted, but further improvement is still much needed (Ochieng *et al.*, 2019). *Amaranthus* genetic resources at the WVC include a MAGIC population from eight parental lines: four *A. cruentus* and four *A. hypochondriacus* (R. Schafleitner, personal communications). The high-quality genome sequences of the two parental species will therefore be extremely valuable in designing markers for quantitative genetic analysis of this



**Figure 6.** Conserved biosynthetic gene cluster in the genomes of five Amaranthaceae species.

The five Amaranthaceae species were: Ac, *Amaranthus cruentus*; Ah, *Amaranthus hypochondriacus*; Bv, *Beta vulgaris*; Cq, *Chenopodium quinoa*; So, *Spinacia oleracea*. This cluster includes three different enzyme domains: epimerase, terpene synthase and cytochrome P450. Boxes indicate predicted gene models with different colours indicating different enzyme types (yellow, terpene synthase; blue, cytochrome P450; purple, epimerase) and orthologous genes linked by grey lines. Pink boxes indicate a non-biosynthetic enzyme encoding gene that is also co-located across all five genomes.

population and in determining the molecular basis of beneficial traits. *A. cruentus*, like other amaranth species, lacks strong domestication traits (Stetter *et al.*, 2020) hence, effective breeding and crop development, for both grain and leafy vegetable traits, will help realize the environmental and socioeconomic potential of this species. Our *A. cruentus* genome assembly and annotation are freely available within the AOCC portal of the ORCAE database (Sterck *et al.*, 2012; Yssel *et al.*, 2019) both to enable easy reuse of the information, as well as facilitate community-based manual annotation of the genome.

*Amaranthus cruentus* is monoecious and can self-pollinate. Amaranths are also known to be able to hybridize (Sauer, 1950), and interspecific hybridization is thought to underlie the success of *A. tuberculatus* as a weed (Trucco *et al.*, 2009). However, the line we sequenced (Arusha; Gerrano *et al.*, 2017) is a semi-domesticated accession, and as expected had a low level of heterozygosity (0.07%). Combining long-read, short-read and phased sequencing technologies, we generated a 371 Mb genome assembly, with 365 Mb assembled into 17 pseudochromosomes. This is slightly shorter than the latest *A. hypochondriacus* genome assembly of 404 Mb (Lightfoot *et al.*, 2017); however, K-mer analysis indicated a total *A. cruentus* genome length of approximately 399 Mb.

Annotation of the *A. cruentus* assembly indicated a gene number similar to *A. hypochondriacus* and closely related *B. vulgaris*, with very similar distributions of transcript, CDs, exon and intron lengths between these species (as well as *C. quinoa*). The overall proportion of TEs in the two *Amaranthus* genomes varied (35% versus 24%) with *A. cruentus* containing a higher proportion of LTRs. However,

this may reflect different repeat analysis methods across the two studies. Other types of element were found at similar proportions, apart from SINE elements, which were responsible for five times as much sequence in *A. cruentus* compared with *A. hypochondriacus*. The proportion of TEs in plant genomes varies widely, for example, 14% in *Arabidopsis thaliana*, 63% in tomato and 84% in maize (Ragupathy *et al.*, 2013). The proportion of TEs in a genome is related to genome size (increasing as genomes increase in size) (Tenailon *et al.*, 2010) with 35% in *A. cruentus* in line with what is seen in other similarly sized genomes.

Interestingly, our *A. cruentus* genome contained a larger proportion of longer protein-CDS than the published *A. hypochondriacus* genome (Lightfoot *et al.*, 2017). This is likely due to the *A. cruentus* Isoseq transcript long-read data we used to complement homology and *ab initio* gene prediction approaches. We included different tissues and/or conditions to capture a wide range of transcripts, bearing in mind many transcripts will show fairly restricted expression patterns. Fifty-one per cent of predicted genes in *A. cruentus* were present as in our long-read transcriptome data, a fairly low gene coverage but with the advantage of full-length transcripts for these genes. Generation of long-read transcriptome data from additional cell/tissue types and conditions/treatments will be needed to further improve the genome annotation and identify full-length transcripts (and importantly, transcript isoforms) for the remaining *A. cruentus* genes. Such a transcript dataset is critical for accurate RNA-seq expression analysis and investigation of differential splicing (Brown *et al.*, 2017).

Comparison of the new *A. cruentus* genome with 13 other plant genomes placed the known WGD event in *A.*

*hypochondriacus* (Lightfoot *et al.*, 2017) at approximately 30 MY before speciation of *A. cruentus* and *A. hypochondriacus*, approximately 1.45 MYA (Figure 2). It also highlighted substantial chromosomal rearrangements following the amaranth lineage WGD event, both pre- and post-speciation. The loss of one copy of chromosome 5 and the fusion of both copies of chromosome 1 were common to both *A. cruentus* and *A. hypochondriacus*, whereas the fission of one copy of chromosome 2 into two is specific to *A. cruentus*. The latter event giving *A. cruentus* a haploid chromosome number of 17, and *A. hypochondriacus*, 16. Draft genomes have been published for the agricultural weed species, *A. tuberculatus*, *A. palmerii* and *A. hybridus* (Montgomery *et al.*, 2020), but the lack of annotation restricts comparative genome rearrangement analysis. However, chromosome loss, fusion and fission after the WGD appears to be prevalent in the *Amaranthus* genus. Among 30 amaranth accessions analysed, four species have a haploid chromosome number  $n = 16$  and 26 have  $n = 17$  (Grant, 1959). As more *Amaranthus* genomes are assembled, it remains to be seen to what extent chromosome loss, fusion and fission events occurred independently in each species. Assembled *Amaranthus* genomes will also assist in understanding of hybridization between species with different chromosomal numbers, and its impact on improvement strategies for this neglected crop.

Analysis of gene families in *A. cruentus* and related genomes, highlighted two *A. cruentus*-specific genes encoding inositol 1,3,4-trisphosphate 5/6-kinase ATP-grasp domains, which may play a role in phytic acid synthesis. Phytic acid has both anti-nutrient properties, inhibiting the uptake of minerals, as well as beneficial anti-cancer and antioxidant activity. It is likely that with a well-balanced diet the beneficial properties outweigh the negative effects, although in populations with high levels of micronutrient deficiencies and/or malnourishment, the impact of phytates on mineral uptake may be significant (Schlemmer *et al.*, 2009). However, a recent study across 20 cereal and pseudocereal flakes demonstrated that amaranth and teff contained the best ratios of minerals versus dietary fibre, phytates and tannins (Kiewlicz and Rybicka, 2020). Hence consumption of these would help deliver gains in mineral nutrition (particularly for magnesium and iron). The expansion of ion transporter gene families in *A. cruentus* (and *A. hypochondriacus*) compared with other plant genomes (Figure S10) could play a role in the increased accumulation of minerals in this genus compared with cereals and pseudocereals.

Although the Arusha *A. cruentus* line sequenced here has not been seen to accumulate visible red pigments in either the inflorescence or leaves, it is one of the *Amaranthus* species producing betalain pigments rather than anthocyanins (Brockington *et al.*, 2011). Consistent with this, we detected conserved co-location of the genes encoding the key enzymes of betalain synthesis, which was previously

identified in *A. hypochondriacus*, *B. vulgaris* and *C. quinoa*. We also highlight the same conservation of co-location in the *S. oleracea* genome (Dohm *et al.*, 2014). In addition to this betalain enzyme grouping, we identified 22 potential biosynthetic gene clusters for plant secondary metabolites in the *A. cruentus* genome, with 10 of these conserved in *A. hypochondriacus*. Six of these showed conserved co-location of biosynthetic enzymes across all five Amaranthaceae genomes outlined above, with one cluster exhibiting co-location of genes encoding two cytochrome P450 enzymes, an epimerase, and multiple terpene synthases. We hypothesize this gene cluster plays a role in a terpene synthesis pathway leading to compounds that may be specific to this lineage. Co-location of biosynthetic enzymes can dramatically increase the ease of identifying biosynthetic pathway genes, particularly when combined with transcriptome analysis correlating gene expression with the presence of specific metabolites in specific plant tissues. Complementation of this genome assembly with transcriptome and metabolome analyses will further aid in the discovery and exploitation of *A. cruentus*, and drive realization of its potential in supporting a healthy and sustainable food system. Enhanced agrobiodiversity through new and diversified crops will be vital for the environmental, economic and social sustainability of future agricultural systems.

## EXPERIMENTAL PROCEDURES

### Plant growth, sampling and DNA extraction

Plants (*A. cruentus* cv. 'Arusha'; Gerrano *et al.*, 2017) were grown in the glasshouse under 12 h light (28°C)/12 h dark (24°C) in 10 cm × 10 cm × 10 cm pots with fine peat-based compost and added sand. In this sequencing project different pools of individual plants were used for gDNA extraction. For short-read Illumina a total amount of 1.0 µg DNA per sample was used as input material for a sequencing library generated using NEBNext® DNA Library Prep Kit according to the manufacturer's protocol. For long-read Nanopore sequencing, gDNA was extracted from leaf nuclei (isolated according to Workman *et al.*, 2018) using the Nanobind Plant Nuclei Big DNA kit (Circulomics, Inc., Baltimore, MD, USA). Long fragment DNA sequencing libraries were prepared using the Oxford Nanopore Technologies (Oxford, UK) ligation sequencing kit LSK-SQK109. For chromatin conformation capture (Hi-C), 4–6 week-old leaf material was cross-linked by incubation at room temperature for 1 h in 1% formaldehyde. Glycine (125 mM final concentration) was added with incubation for a further 15 min before rinsing with water, and flash-freezing in liquid nitrogen. The cross-linked material was shipped to Phase Genomics who performed subsequent gDNA extraction, chromatin isolation, library preparation and sequencing. For 10× Chromium Genome sequencing, 4–6-week-old leaf material was sent to the UC Davis Genome Center for subsequent gDNA extraction, library preparation (according to 10× Genomics) protocols and sequencing.

### Genome sequencing datasets

Long-read sequencing was carried out by the Technology Facility (University of York) using Oxford Nanopore Technologies MinION

sequencers running R9.4.1 flow cells. Data were basecalled using GUPPY v1.8.3. We recovered 24.5 Gb of long-read data comprising 7 728 433 reads with a N50 read length of 6.43 kb and with a maximum length of 356 126 bp (Table S1). Given an estimated genome size of 400 Mb (see below), this represents a coverage of approximately 60×. We used these long reads to construct the initial contig-level assembly.

We generated three different types of PE short-read sequences using the Illumina sequencing platform, including Hi-C, 10× Genomics and regular short inserts (Table S1). For Hi-C sequencing, we obtained 120 Gb (approximately 300× coverage) of *in vivo*-generated Hi-C data from Phase Genomics. The Hi-C data were used for the construction of a chromosome-scale assembly from the contig-level assembly. One hundred and twenty Gb of 10× Genomics reads and 110 Gb from Illumina PE read libraries with short inserts were used for the estimation of genome size and heterozygosity level, as well as base polishing of the genome assembly at the final stages to improve its accuracy. We used GENOMESCOPE (Vurture *et al.*, 2017) to estimate the genome size of *A. cruentus*. First, a 32-mer distribution was generated with JELLYFISH (Marçais and Kingsford, 2011) from the 120 Gb 10× Genomics short reads and used as input in the subsequent GENOMESCOPE analysis.

### Genome assembly with chromosome-length scaffolds and polishing

For the initial contig-level assembly, we ran the WTDBG2 (Ruan and Li, 2020) assembly pipelines with the Oxford Nanopore dataset producing an initial assembly of 372.9 Mb with 1948 contigs and a contig N50 of 0.98 Mb. Purge\_dups (Guan *et al.*, 2020) was used to remove haplotigs and contig overlaps to produce a haploid representation of the genome resulting in a 367.0 Mb haploid assembly with 1438 contigs and a contig N50 of 1.02 Mb.

To construct highly contiguous scaffolds from the above contig-level haploid genome, we used the 3D-DNA pipeline (Dudchenko *et al.*, 2017) with the Hi-C data first to correct misjoins in the contigs, then to scaffold and merge overlaps. The resulting assembly was adjusted and curated using (i) guidance from the chromogram of the Hi-C reads, where pixel intensity in the contact matrix indicates how often a pair of loci co-locate in the nucleus, and (ii) placement of telomere signature sequences. The final *A. cruentus* genome assembly is detailed in Table 1, and includes 17 chromosome-length scaffolds. The genome assembly metrics at different stages of the process outlined above are shown in Table S7.

To improve the base accuracy of the genome assembly further, using the PILON software tool, we polished (Walker *et al.*, 2014): two rounds after mapping the Illumina PE short reads to the working assembly with BWA (Li and Durbin, 2009), and another two rounds after mapping the reads from 10× Chromium Genome sequencing with the LONGRANGER mapping tool (10× Genomics), resulting in a polished final haploid assembly of 370.9 Mb. The quality of the genome assembly was assessed (via cumulative length and genome assembly statistics) using QUAST (Gurevich *et al.*, 2013); 96.8% of the 368 million 150-bp sequencing reads we generated aligned to the final genome assembly with 92.5% of the reads mapped as accurate pairs.

### Full-length transcript isoforms of *Amaranthus cruentus*

To assist genome annotation for the identification and validation of gene structure and coding regions, we used the PACBIO SINGLE MOLECULE, Real-Time sequencing technology and Iso-Seq analysis

pipeline to generate full-length cDNA sequences. RNA was extracted from 10 different tissues (2-week-old seedlings; 6-week-old plant root, stem and leaf; flowers; seed; 10-week-old senescent leaf; leaves from 4 week-old plants 1 h after wounding; and leaves of 4-week-old plants 24- and 48-h after *Botrytis cinerea* infection) according to standard cetyltrimethylammonium bromide protocols followed by lithium chloride precipitation. Samples were pooled and first-strand cDNA synthesized according to Pacific Biosciences (Menlo Park, CA, USA) Isoseq recommended protocol. Library preparation and sequencing were performed according to Pacific Biosciences protocols by the UC Davis Genome Center.

We analysed the highly accurate long reads (HiFi reads) generated with the Circular Consensus Sequence algorithm using the Iso-Seq bioinformatics pipeline: identifying and clustering a full-length transcript dataset at the isoform level that spans entire transcript isoforms, and polishing to create the *A. cruentus* high-quality consensus dataset containing 51 928 sequences. Cupcake, a set of supporting scripts for processing Iso-Seq data (Gordon *et al.*, 2015), was used to remove further redundancies with the parameters (MIN\_ALN\_COVERAGE: 0.95; MIN\_ALN\_IDENTITY: 0.90) producing a final dataset of unique full-length transcript isoforms containing 21 732 sequences in total.

### Genome annotation

REPEATMODELER v2.0.1 (Flynn *et al.*, 2020) was used to identify the repeat families in the genome assembly of *A. cruentus* based on the default TE Dfam database and Repbase database with the support of LTR\_Struc (McCarthy and McDonald, 2003). Furthermore, LTR\_Finder (v1.0.7), LTR\_harvest from GENOMETOOLS (v1.5.9) and LTR\_retriever (v2.9.0) were used to identify and trace the LTR elements in the *A. cruentus* genome (Ellinghaus *et al.*, 2008; Ou and Jiang, 2018; Xu and Wang, 2007). We merged the libraries from REPEATMODELER and LTR\_retriever by USEARCH (Edgar, 2010) with 80% identity as the minimum threshold for combining similar sequences in the *de novo* libraries to get the non-redundant *de novo* repeat library. Finally, we used REPEATMASKER v4.1.1 ([https://sc.icsrunch.org/resolver/RRID:SCR\\_012954](https://sc.icsrunch.org/resolver/RRID:SCR_012954), Chen, 2004) with parameter: -e rmbast -a -s -norna -xsmall -gff -lib to identify and classify repeats in the *A. cruentus* genome assembly. Gene models were predicted with a combination of *ab initio* prediction, homology search and RNA-aided annotation. BRAKER2 (Brüna *et al.*, 2021) was used for *ab initio* gene prediction using model training based on Iso-seq data from *A. cruentus* and proteins of very close homology from *A. hypochondriacus* after the annotated repeats were soft masked in the assembly. For homology prediction, protein sequences from four closely related species that belong to the same family were used as query sequences to search the reference genome using TBLASTN with different *e*-values (*A. hypochondriacus* with the *e*-value  $\leq 1e - 10$ , *B. vulgaris*, *C. quinoa* and *S. oleracea* with the *e*-value  $\leq 1e - 5$ ). Regions mapped by these query sequences were subject to Exonerate (Slater and Birney, 2005) to generate putative transcripts. For RNA-aided annotation, the full-length transcript isoforms generated from the Isoseq data were used as input for the PASA pipeline (Haas *et al.*, 2003) for gene structure predictions, with TRANSDCODER (Haas *et al.*, 2013) also used to predict open reading frames. Finally, EVIDENCEMODELER v1.1.1 (Haas *et al.*, 2008) was used to integrate all of the above evidence and BUSCO v4.0.4 (Benchmarking Universal Single-Copy Orthologs; Seppey *et al.*, 2019) to assess the quality of annotation results.

Putative gene function was identified using InterProScan with different databases, including PFAM, Gene3D, PANTHER, CDD, SUPERFAMILY, ProSite and GO. Meanwhile, functional annotation

of these predicted genes was obtained by aligning the protein sequences of these genes against the sequences in public protein databases and the UniProt database using BLASTP ( $E$ -value  $< 1 \times 10^{-5}$ ).

### miRNAs and putative target gene prediction

Known mature miRNAs were obtained from miRbase (release 22.1) and aligned to the soft-masked *A. cruentus* genome using SeqMap with no mismatches (Jiang and Wong, 2008). We extracted approximately 110 bp upstream and downstream sequence surrounding every aligned locus and discarded miRNA candidates located within protein-CDS or repetitive elements to produce the refined miRNA set. The predicted stem-loop structure and minimum free energy were analysed for each region using the RNAfold program of VIENNA RNA v2.1.1 (Lorenz *et al.*, 2011) with default settings. A miRNA target transcript prediction pipeline was developed using VMATCH (Schnable *et al.*, 2009). Mature miRNA sequences in the refined miRNA set were reverse complemented and matched against the protein-CDS of *A. cruentus*, with the parameters allowing up to three mismatches. Predicted miRNA target genes were tested for GO term enrichment using TBtools (Chen *et al.*, 2020).

### Ks distribution analysis

Ks distribution analysis was performed using the wgd package and the paranome (entire collection of duplicated genes) was obtained with 'wgd mcl' using all-against-all BlastP and MCL clustering. Ks distribution of *A. cruentus* was then constructed using 'wgd ksd' with default settings using MAFFT (Katoh and Standley, 2013) for multiple sequence alignment, codeml for maximum likelihood estimation of pairwise synonymous distances, and FastTree for inferring phylogenetic trees used in the node weighting procedure. Anchors or anchor pairs (paralogous genes lying in collinear or syntenic regions of the genome) were obtained using I-ADHORE, employing the default settings in 'wgd syn'.

### Phylogenetic divergence

The amino acid sequences of all proteins from *A. cruentus* and 13 other angiosperms were downloaded from PLAZA (Van Bel *et al.*, 2018). All-versus-all BLASTP with an  $E$ -value cut-off of  $1e - 05$  was performed and orthologous genes were clustered using ORTHOFINDER v2.1.2 (Emms and Kelly, 2019). Single-copy orthologous genes were extracted from the clustering results. MAFFT (Katoh and Standley, 2013) with default parameters was used to perform multiple alignment of protein sequences for each set of single-copy orthologous genes, and transform the protein sequence alignments into codon alignments. Poorly aligned or divergent regions were removed from the multiple sequence alignment results using TRIMAL. The resulting codon alignments from all single-copy orthologues were then concatenated to one supergene for species phylogenetic analysis. A maximum likelihood phylogenetic tree of single-copy proteins alignments and codon alignments from *A. cruentus* and 13 other angiosperms was constructed using IQ-TREE with the GTR+G model and 1000 bootstrap replicates. Divergence times between the 13 plant species were estimated using MCMCTree from the PAML package (Yang, 2007), using reference speciation times of 42–52 MYA for the divergence between *Oryza sativa* and *Brachypodium distachyon*, 63–82 MYA for that between *A. thaliana* and *Carica papaya*, 84–95 MYA for that between *A. thaliana* and *Theobroma cacao*, 98–117 MYA for that between *A. thaliana* and *Medicago truncatula*, 95–106 MYA for that between *Daucus carota* and *Solanum lycopersicum*, 111–131 MYA for that between

*A. thaliana* and *D. carota*, 22–61 MYA for that between *B. vulgaris* and *C. quinoa*, and 115–308 MYA for that between *O. sativa* and *A. thaliana*.

### Genome and gene family evolution

Syntenic analysis of genomes was performed using MCSCANX (Wang *et al.*, 2012) with parameters '-s 10' and the circos figures were drawn using TBtools (Chen *et al.*, 2020). CAFÉ (v3.1) (Han *et al.*, 2013) was used to identify the expansion and contraction of gene families following divergence predicted by the phylogenetic tree above. TBtools was also used to determine enrichment of GO terms in expanded and contracted families (Chen *et al.*, 2020).

### Analysis of biosynthetic gene clusters

For the analysis of betalain biosynthetic genes, DODA sequences in *C. quinoa* (Jarvis *et al.*, 2017), *S. oleracea* (Xu *et al.*, 2017), *A. hypochondriacus* (Lightfoot *et al.*, 2017) and *A. cruentus* (this study) were identified using BLAST searches with the protein sequence of the previously characterized BvDODA $\alpha$ 1 (Hatlestad *et al.*, 2012). CYP76AD1 orthologues were identified via a BLAST search using *B. vulgaris* subsp. *Cicla* sequence from the National Center for Biotechnology Information (NCBI) (accession no.: KU644144). All gene sequences were checked manually, aligned using MAFFT and a maximum likelihood gene tree based on protein sequences constructed using IQ-TREE with GTR+G model and 1000 bootstrap replicates to confirm orthology. Microsynteny of DODAA1 and CYP76AD1 orthologues was analysed using MCSCAN JCVI (Tang *et al.*, 2008) including approximately 20 genes from either side of the betalain enzymes.

PLANTISMASH (Nützmann *et al.*, 2016) was used to identify potential biosynthetic gene clusters in the newly assembled *A. cruentus* genome and that of *A. hypochondriacus* (Lightfoot *et al.*, 2017). Based on the gene families from ORTHOFINDER analysis, we identified the orthologous sequences in *A. hypochondriacus*, *B. vulgaris*, *C. quinoa* and *S. oleracea* for each member of the gene cluster in *A. cruentus* and used MCSCAN JCVI to visualize the microsynteny (Tang *et al.*, 2008).

### ACKNOWLEDGEMENTS

We would like to thank Katherine Newling and Jon Davey of the University of York Biology Technology Facility for generating the long-read sequencing data and initial processing, and Oanh Nguyen for her expertise in library construction and genome sequencing at the UC Davis Genome Center. The project was initiated by seed funding from the N8 AgriFood programme (<https://www.n8agrifood.ac.uk/>) funded through the Higher Education Funding Council for England, with the majority of the experimental work funded from a Biotechnology and Biological Sciences Research Council (BBSRC) Grand Challenges Research Fund award to KJD, IAG, SLV, MWB and SM (BB/R020345/1). We thank the African Orphan Crop Consortium (AOCC, <http://africanorphanrops.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).

### AUTHOR CONTRIBUTIONS

The study was conceived and designed by KJD, IAG, SM, WSJvR, MWB, SLV, YVdP and AVD. Experimental work was carried out by FEV, SH, WSJvR and MWB, with the

genome assembly by YL and ZN. Genome annotation and comparative analyses were carried out by XM, with the analysis of biosynthetic gene clusters by KJD and XM. Funding for the research was obtained by KJD, IAG, SLV, MWB, SM, YVdP and AVD. The manuscript was written by XM, FEV and KJD with input from all authors.

## CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

## DATA AVAILABILITY STATEMENT

All data were available either within the supporting information of this manuscript or associated with the *A. cruentus* genome in the AOCC portal of ORCAE (<https://www.bioinformatics.psb.ugent.be/gdb/amaranthus/>). In this portal, Chromosome 2B is named Chromosome 17. Raw data and the genome assembly are available at NCBI under BioProject: PRJNA713964.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** *Amaranthus cruentus* cv. Arusha. Plants were raised as seedlings in the greenhouse and planted in the field approximately 4 weeks after sowing. (a,b) Show plants approximately 9 weeks after transplanting with the plants approximately 1.2 m tall; (c) shows an inflorescence, approximately 30 cm in length; and (d) black seed with diameter approximately 1 mm.

**Figure S2.** K-mer analysis indicates a genome size of 398.6 Mb and a heterozygosity level of 0.07%.

**Figure S3.** Cumulative length of the *Amaranthus cruentus* genome assembly by number of contigs and scaffolds.

**Figure S4.** Heat map of the Hi-C paired reads indicating the 17 linkage groups. Heat map colour intensity indicates the frequency of paired read co-location.

**Figure S5.** *Amaranthus cruentus* predicted protein lengths are similar to those in *Amaranthus hypochondriacus*, with a greater proportion of longer predicted proteins. Frequency of the length ratio between reciprocal best hit orthologues is shown, with values >1 indicating longer proteins in *A. hypochondriacus*, and values <1 indicating longer proteins in *A. cruentus*.

**Figure S6.** Distribution of transcript, protein-coding sequence, exon and intron length are similar across the genomes of *Amaranthus cruentus*, *Amaranthus hypochondriacus*, *Beta vulgaris* and *Chenopodium quinoa*.

**Figure S7.** Ks distributions for paralogues within the *Amaranthus cruentus* genome and for orthologues of *A. cruentus* and related species.

**Figure S8.** Syntenic relationship between the *Amaranthus cruentus* genome and *Beta vulgaris* genome. Size of chromosomes is indicated in Mb around the outside with syntenic blocks of genes shown in the same colour.

**Figure S9.** Gene ontology (GO) term enrichment in expanded (a) and contracted (b) gene families in *Amaranthus cruentus*. Only statistically significant enrichment is shown (adjusted  $P < 0.05$ ).

**Figure S10.** GO term enrichment in expanded (a) and contracted (b) gene families in the two amaranth species (*Amaranthus cruentus* and *Amaranthus hypochondriacus*) compared with other 12

plant genomes analysed. Only statistically significant enrichment is shown (adjusted  $P < 0.05$ ).

**Figure S11.** (a) *Spinacia oleracea* and (b) *Chenopodium quinoa* genomes show evidence for tandem gene duplications of DODA $\alpha$ 1 and CYP76AD1. Ac, *Amaranthus cruentus*; Ah, *Amaranthus hypochondriacus*; Bv, *Beta vulgaris*; Cq, *C. quinoa*; So, *S. oleracea*. Rectangles indicate gene locations and colour the relative orientation (blue, – strand; green, + strand). Orthologous genes are linked by grey lines, with DODA $\alpha$ 1 orthologues linked by red lines, and CYP76AD1 orthologues by blue.

**Figure S12.** Conserved biosynthetic gene clusters in the genomes of five Amaranthaceae species: Ac, *Amaranthus cruentus*; Ah, *Amaranthus hypochondriacus*; Bv, *B. vulgaris*; Cq, *Chenopodium quinoa*; So, *Spinacia oleracea*. Each of the panel boxes indicate predicted gene models with different colours indicating different enzyme types and orthologous gene pairs linked by grey lines. Genes indicated with grey boxes are ‘other’ genes within the cluster or flanking the cluster, which do not have a predicted biosynthetic role.

**Table S1.** Statistics of raw sequence read files for the assembly and polishing of the *Amaranthus cruentus* genome. Data are the output of seqkit (<https://github.com/shenwei356/seqkit>) software command (seqkit stats -a).

**Table S2.** Chromosome statistics. Length of each pseudochromosome is shown in bp.

**Table S3.** A comparison of transposable element content of *Amaranthus cruentus* with related species. Genome information is obtained from this study (*A. cruentus*), Jarvis et al. (2017) (*Chenopodium* genomes), Lightfoot et al. (2017) (*Amaranthus hypochondriacus*) and Dohm et al. (2014) (*Beta vulgaris*). Data are the length of relevant sequence in Mb, and the percentage of the genome sequence that a type of transposable element contributes.

**Table S4.** (a) miRNA annotation in *Amaranthus cruentus* genome; (b) predicted miRNA target genes in the *A. cruentus* genome. First column is the name of the miRNA identified in this study and the second column is the predicted miRNA target gene; (c) miRNA target gene GO term enrichment. Target genes of miRNAs are from Supplementary Table 4b.

**Table S5.** Orthology across species analysis in *Amaranthus cruentus* and 13 additional plant genomes.

**Table S6.** (a) PLANTISMASH results for *Amaranthus cruentus* indicating the cluster number, chromosome the cluster is located on, type of metabolite the cluster core enzyme produces, location of the cluster in the genome, size of the cluster, key biosynthetic protein domains encoded in the cluster, the number of CD-HIT gene groups in the cluster, and the gene IDs present in the cluster (with the core domain genes in bold); (b) PLANTISMASH results for *Amaranthus hypochondriacus* indicating the cluster number, chromosome the cluster is located on, type of metabolite the cluster core enzyme produces, location of the cluster in the genome, size of the cluster, key biosynthetic protein domains encoded in the cluster, the number of CD-HIT gene groups in the cluster and the gene IDs present in the cluster (with the core domain genes in bold); (c) PLANTISMASH biosynthetic gene clusters conserved across the *A. cruentus* and *A. hypochondriacus* genomes.

**Table S7.** Metrics at different stages of the *Amaranthus cruentus* genome assembly

## REFERENCES

- Alvarez-Jubete, L., Arendt, E.K. & Gallagher, E. (2009) Nutritive value and chemical composition of pseudocereals as gluten-free ingredients. *International Journal of Food Science and Nutrition*, **60**(Suppl 4), 240–257.

- Andini, R., Yoshida, S., Yoshida, Y. & Ohsawa, R. (2013) Amaranthus genetic resources in Indonesia: Morphological and protein content assessment in comparison with worldwide amaranths. *Genetic Resources and Crop Evolution*, **60**, 2115–2128.
- Ballabio, C., Uberti, F., Di Lorenzo, C., Brandolini, A., Penas, E. & Restani, P. (2011) Biochemical and immunochemical characterization of different varieties of amaranth (*Amaranthus* L. ssp.) as a safe ingredient for gluten-free products. *Journal of Agriculture and Food Chemistry*, **59**, 12969–12974.
- Brenner, D.M., Baltensperger, D.D., Kulakow, P.A., Lehmann, J.W., Myers, R.L., Slabbert, M.M. *et al.* (2000) Genetic resources and breeding of *Amaranthus*. In: Janick, J. (Ed.) *Plant breeding reviews*. Oxford: John Wiley & Sons Ltd, pp. 227–285.
- Brockington, S.F., Walker, R.H., Glover, B.J., Soltis, P.S. & Soltis, D.E. (2011) Complex pigment evolution in the Caryophyllales. *New Phytologist*, **190**, 854–864.
- Brockington, S.F., Yang, Y.a., Gandia-Herrero, F., Covshoff, S., Hibberd, J.M., Sage, R.F. *et al.* (2015) Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. *New Phytologist*, **207**, 1170–1180.
- Brown, J.W.S., Calixto, C.P.G. & Zhang, R. (2017) High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. *New Phytologist*, **213**, 525–530.
- Brüna, T., Hoff, K.J., Lomsadze, A., Stanke, M. & Borodovsky, M. (2021) BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, **3**, lqaa108. <https://doi.org/10.1093/nar/gab/lqaa108>.
- Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y. *et al.* (2020) TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant*, **13**, 1194–1202.
- Chen, N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, Chapter 4, Unit 4.10. <https://doi.org/10.1002/0471250953.bi0410s05>
- Christinet, L., Burdet, F.X., Zaiko, M., Hinz, U. & Zryd, J.-P. (2004) Characterization and functional identification of a novel plant 4,5-extradiol dioxygenase involved in betalain pigment biosynthesis in *Portulaca grandiflora*. *Plant Physiology*, **134**, 265–274.
- Clement, J.S. & Mabry, T.J. (1996) Pigment evolution in the Caryophyllales: a systematic overview. *Botanica Acta*, **109**, 360–367.
- Coelho, L.M., Silva, P.M., Martins, J.T., Pinheiro, A.C. & Vicente, A.A. (2018) Emerging opportunities in exploring the nutritional/functional value of amaranth. *Food and Function*, **9**, 5499–5512.
- Dohm, J.C., Minoche, A.E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H. *et al.* (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, **505**, 546–549.
- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C. *et al.* (2017) De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.
- Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, **26**, 2460–2461.
- Elinghaus, D., Kurtz, S. & Willhoelt, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18. <https://doi.org/10.1186/1471-2105-9-18>.
- Emms, D.M. & Kelly, S. (2019) OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 238.
- Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. *et al.* (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences United States of America*, **117**, 9451–9457.
- Gerrano, A., Jansen van Rensburg, W., Mavengahama, S., Bairu, M., Venter, S. & Adebola, P. (2017) Qualitative morphological diversity of *Amaranthus* species. *Journal of Tropical Agriculture*, **55**, 12–20.
- Gordon, S.P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z. *et al.* (2015) Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One*, **10**, e0132628.
- Grant, W.F. (1959) Cytogenetic studies in *Amaranthus*. 111. Chromosome numbers and phylogenetic aspects. *Canadian Journal of Genetics and Cytology*, **1**, 313–318.
- Grubben, G.J.H. & Denton, O.A. (2004) *Plant resources of tropical Africa 2, vegetables*. Leiden, Netherlands: Backhuys Publishers.
- Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y. & Durbin, R. (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics (Oxford, England)*, **36**, 2896–2898.
- Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013) QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
- Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Hannick, L.I. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**, 5654–5666.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidence-Modeler and the program to assemble spliced alignments. *Genome Biology*, **9**, R7. <https://doi.org/10.1186/gb-2008-9-1-r7>.
- Han, M.V., Thomas, G.W.C., Lugo-Martinez, J. & Hahn, M.W. (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*, **30**, 1987–1997.
- Hatlestad, G.J., Sunnadeniya, R.M., Akhavan, N.A., Gonzalez, A., Goldman, I.L., McGrath, J.M. *et al.* (2012) The beet R locus encodes a new cytochrome P450 required for red betalain production. *Nature Genetics*, **44**, 816–820.
- Hauptli, H. & Jain, S.B. (1984) Genetic structure of landrace populations of the New World grain amaranths. *Euphytica*, **33**, 875–884.
- Hoidal, N., Diaz Gallardo, M., Jacobsen, S.-E. & Alandia, G. (2019) Amaranth as a dual-use crop for leafy greens and seeds: Stable responses to leaf harvest across genotypes and environments. *Frontiers in Plant Science*, **10**, <https://doi.org/10.3389/fpls.2019.00817>.
- Hou, Q., Ufer, G. & Bartels, D. (2016) Lipid signalling in plant responses to abiotic stress. *Plant Cell and Environment*, **39**, 1029–1048.
- Jamalluddin, N., Massawe, F.J. & Symonds, R.C. (2019) Transpiration efficiency of Amaranth (*Amaranthus* sp.) in response to drought stress. *The Journal of Horticultural Science and Biotechnology*, **94**, 448–459.
- Jamnadas, R., Mumm, R.H., Hale, I., Hendre, P., Muchugi, A., Dawson, I.K. *et al.* (2020) Enhancing African orphan crops with genomics. *Nature Genetics*, **52**, 356–360.
- Jarvis, D.E., Ho, Y.S., Lightfoot, D.J., Schmöckel, S.M., Li, B.o., Borm, T.J.A. *et al.* (2017) The genome of *Chenopodium quinoa*. *Nature*, **542**, 307–312.
- Jiang, H. & Wong, W.H. (2008) SeqMap: Mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **24**, 2395–2396.
- Joshi, D.C., Sood, S., Hosahatti, R., Kant, L., Pattanayak, A., Kumar, A. *et al.* (2018) From zero to hero: The past, present and future of grain amaranth breeding. *Theoretical and Applied Genetics*, **131**, 1807–1823.
- Kaderit, G., Borsch, T., Weising, K. & Freitag, H. (2003) Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C4 photosynthesis. *International Journal of Plant Science*, **164**, 959–986.
- Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
- Kautsar, S.A., Suarez Duran, H.G., Blin, K., Osbourn, A. & Medema, M.H. (2017) plantSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, **45**, W55–W63.
- Kiewlicz, J. & Rybicka, I. (2020) Minerals and their bioavailability in relation to dietary fiber, phytates and tannins from gluten and gluten-free flakes. *Food Chemistry*, **305**, 125452.
- Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**, 1754–1760.
- Lightfoot, D.J., Jarvis, D.E., Ramaraj, T., Lee, R., Jellen, E.N. & 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. <https://doi.org/10.1186/s12915-017-0412-4>.
- Lorenz, K. & Wright, B. (1984) Phytate and tannin content of amaranth. *Food Chemistry*, **14**, 27–34.
- Lorenz, R., Bernhart, S.H., Höner zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P.F. *et al.* (2011) ViennaRNA package 2.0. *Algorithms for Molecular Biology*, **6**, 26. <https://doi.org/10.1186/1748-7188-6-26>.



- Louveau, T. & Osbourn, A. (2019) The sweet side of plant-specialized metabolism. *Cold Spring Harbor in Perspective Biology*, **11**, a034744.
- Marçais, G. & Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770. <https://doi.org/10.1093/bioinformatics/btr011>.
- McCarthy, E.M. & McDonald, J.F. (2003) LTR\_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362–367.
- Montgomery, J.S., Giacomini, D., Waithaka, B., Lanz, C., Murphy, B.P., Campe, R. et al. (2020) Draft genomes of *Amaranthus tuberculatus*, *Amaranthus hybridus*, and *Amaranthus palmeri*. *Genome Biology and Evolution*, **12**, 1988–1993.
- Nützmann, H.-W., Huang, A. & Osbourn, A. (2016) Plant metabolic clusters – from genetics to genomics. *New Phytologist*, **211**, 771–789.
- Ochieng, J., Schreinemachers, P., Ogada, M., Dinssa, F.F., Barnos, W. & Mndiga, H. (2019) Adoption of improved amaranth varieties and good agricultural practices in East Africa. *Land Use Policy*, **83**, 187–194.
- Odhav, B., Beekrum, S., Akula, U. & Bajjnath, H. (2007) Preliminary assessment of nutritional value of traditional leafy vegetables in KwaZulu-Natal, South Africa. *Journal of Food Composition and Analysis*, **20**, 430–435.
- Omami, E.N. & Hammes, P.S. (2006) Interactive effects of salinity and water stress on growth, leaf water relations, and gas exchange in amaranth (*Amaranthus* spp.). *New Zealand Journal of Crop and Horticultural Science*, **34**, 33–44.
- Ou, S. & Jiang, N. (2018) LTR\_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, **176**, 1410–1422.
- Pedersen, B., Hallgren, L., Hansen, I. & Eggum, B.O. (1987) The nutritive value of amaranth grain (*Amaranthus caudatus*). *Plant Foods for Human Nutrition*, **36**, 325–334.
- Polturak, G., Breitel, D., Grossman, N., Sarrion-Perdigones, A., Weithorn, E., Pliner, M. et al. (2016) Elucidation of the first committed step in betalain biosynthesis enables the heterologous engineering of betalain pigments in plants. *New Phytologist*, **210**, 269–283.
- Prakash, D. & Pal, M. (1991) Nutritional and antinutritional composition of vegetable and grain amaranth leaves. *Journal of the Science of Food and Agriculture*, **57**, 573–583.
- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. et al. (2005) InterProScan: Protein domains identifier. *Nucleic Acids Research*, **33**, W116–W120.
- Ragupathy, R., You, F.M. & Cloutier, S. (2013) Arguments for standardizing transposable element annotation in plant genomes. *Trends in Plant Science*, **18**, 367–376.
- Ruan, J. & Li, H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, **17**, 155–158.
- Sarker, U. & Oba, S. (2020) Phenolic profiles and antioxidant activities in selected drought-tolerant leafy vegetable amaranth. *Science Reports*, **10**, 18287.
- Sauer, J.D. (1950) The grain amaranths: A survey of their history and classification. *Annals of Missouri Botanical Garden*, **37**, 561–632.
- Sauer, J. (1957) Recent migration and evolution of the dioecious amaranths. *Evolution*, **11**, 11–31.
- Schlemmer, U., Frölich, W., Prieto, R.M. & Grases, F. (2009) Phytate in foods and significance for humans: Food sources, intake, processing, bioavailability, protective role and analysis. *Molecular Nutrition and Food Research*, **53**, S330–S375.
- Schnable, P.s., Ware, D., Fulton, R.s., Stein, J.c., Wei, F., Pasternak, S. et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
- Seppy, M., Manni, M. & Zdobnov, E.M. (2019) BUSCO: Assessing genome assembly and annotation completeness. In: Kollmar, M. (Ed.) *Gene prediction: Methods and protocols, methods in molecular biology*. New York, NY: Springer, pp. 227–245.
- Sheehan, H., Feng, T., Walker-Hale, N., Lopez-Nieves, S., Pucker, B., Guo, R. et al. (2020) Evolution of I-DOPA 4,5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales. *New Phytologist*, **227**, 914–929.
- Slater, G.S.C. & Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
- Sterck, L., Billiau, K., Abeel, T., Rouzé, P. & Van de Peer, Y. (2012) ORCAE: Online resource for community annotation of eukaryotes. *Nature Methods*, **9**, 1041–1041.
- Stetter, M.G. & Schmid, K.J. (2017) Analysis of phylogenetic relationships and genome size evolution of the *Amaranthus* genus using GBS indicates the ancestors of an ancient crop. *Molecular Phylogenetics and Evolution*, **109**, 80–92.
- Stetter, M.G., Vidal-Villarejo, M. & 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.
- Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. & Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.
- Tenaillon, M.I., Hollister, J.D. & Gaut, B.S. (2010) A triptych of the evolution of plant transposable elements. *Trends in Plant Science*, **15**, 471–478.
- Trucco, F., Tatum, T., Rayburn, A.L. & Tranel, P.J. (2009) Out of the swamp: Unidirectional hybridization with weedy species may explain the prevalence of *Amaranthus tuberculatus* as a weed. *New Phytologist*, **184**, 819–827.
- van Jaarsveld, P., Faber, M., van Heerden, I., Wenhold, F., Jansen van Rensburg, W. & van Averbeke, W. (2014) Nutrient content of eight African leafy vegetables and their potential contribution to dietary reference intakes. *Journal of Food Composition and Analysis*, **33**, 77–84.
- Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y. et al. (2018) PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research*, **46**, D1190–D1196.
- Venskutonis, P.R. & Kraujalis, P. (2013) Nutritional components of amaranth seeds and vegetables: A review on composition, properties, and uses. *Comprehensive Reviews in Food Science and Food Safety*, **12**, 381–412.
- Vurtule, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. et al. (2017) GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, **33**, 2202–2204.
- Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. et al. (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
- Wang, Y., Tang, H., DeBarry, J.d., Tan, X., Li, J., Wang, X. et al. (2012) MScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, **40**, e49.
- Workman, R.E., Myrka, A.M., Wong, G.W., Tseng, E., Welch, K.C. Jr & Timp, W. (2018) Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. *GigaScience*, **7**, <https://doi.org/10.1093/gigascience/giy009>.
- Xu, C., Jiao, C., Sun, H., Cai, X., Wang, X., Ge, C. et al. (2017) Draft genome of spinach and transcriptome diversity of 120 Spinacia accessions. *Nature Communications*, **8**, 15275.
- Xu, Z. & Wang, H. (2007) LTR\_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**, W265–W268.
- Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.
- Yssel, A.E.J., Kao, S.-M., Van de Peer, Y. & Sterck, L. (2019) ORCAE-AOCC: A centralized portal for the annotation of African orphan crop genomes. *genes*, **10**, 950.
- Zwaenepoel, A. & Van de Peer, Y. (2019) wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*, **35**, 2153–2155.