Application of chloroplast phylogenomics to resolve species relationships within the plant genus *Amaranthus*

Erika Viljoen[1,2], Damaris A. Odeny[3], Martin P. A. Coetzee[4], *Dave K. Berger[2], David J. G. Rees[1,5]

[1]Biotechnology Platform, Agricultural Research Council, Onderstepoort, Pretoria, 0110, South Africa

[2]Department of Plant and Soil Sciences, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Hatfield, 0083, South Africa

[3]International Crops Research Institute for the Semi-Arid Tropics, Nairobi, Kenya

[4]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Hatfield, 0083, South Africa

[5]Department of Life and Consumer Sciences, College of Agricultural and Environmental Sciences, University of South Africa, Florida, 1710, South Africa

ORCID ID
Erika Viljoen: http://orcid.org/0000-0003-2149-2609
Martin P. A. Coetzee: http://orcid.org/0000-0001-7848-4111
Dave K Berger: http://orcid.org/0000-0003-0634-1407

*Corresponding Author:

Prof. Dave K. Berger
Email: dave.berger@up.ac.za

Acknowledgements

The authors wish to thank the Department of Science and Technology of South Africa, the National Research Foundation and the Professional Development Program of the Agricultural Research Council (ARC) in South Africa for providing funding for the PhD study from where this work originated. We acknowledge Dr Charles Hefer at the ARC for bioinformatics support. The authors thank Mr Willem Jansen van Rensburg and his staff at the ARC Vegetable and Ornamental Plant Institute for providing the *Amaranthus* germplasm set (SAG) as well as plant maintenance. We thank the Core Facility team at the ARC Biotechnology Platform for DNA sequencing.

Abstract

Amaranthus species are an emerging and promising nutritious traditional vegetable food source. Morphological plasticity and poorly resolved dendrograms have led to the need for well resolved species phylogenies. We hypothesized that whole chloroplast phylogenomics would result in more reliable differentiation between closely related amaranth species. The aims of the study were therefore: to construct a fully assembled, annotated chloroplast genome sequence of *Amaranthus tricolor*; to characterize *Amaranthus* accessions phylogenetically by comparing barcoding genes (*mat*K, *rbc*L, ITS) with whole chloroplast sequencing; and to use whole chloroplast phylogenomics to resolve deeper phylogenetic relationships. We generated a complete *A. tricolor* chloroplast sequence of 150,027 bp. The three barcoding genes revealed poor inter- and intra-species resolution with low bootstrap support. Whole chloroplast phylogenomics of 59 *Amaranthus* accessions increased the number of parsimoniously informative sites from 92 to 481 compared to the barcoding genes, allowing improved separation of amaranth species. Our results support previous findings that two geographically independent domestication events of *Amaranthus hybridus* likely gave rise to several species within the Hybridus complex, namely *Amaranthus dubius*, *Amaranthus quitensis*, *Amaranthus caudatus*, *Amaranthus cruentus* and *Amaranthus hypochondriacus*. Poor resolution of species within the Hybridus complex supports the recent and ongoing domestication within the complex and highlights the limitation of chloroplast data for resolving recent evolution. The weedy *Amaranthus retroflexus* and *Amaranthus powellii* was found to share a common ancestor with the Hybridus complex. Leafy amaranth, *Amaranthus tricolor*, *Amaranthus blitum*, *Amaranthus viridis* and *Amaranthus graecizans* formed a stable sister lineage to the aforementioned species across the phylogenetic trees. This study demonstrates the power of next generation sequencing data and reference-based assemblies to resolve phylogenies and facilitated the identification of unknown *Amaranthus* accessions from a local genebank. The informative phylogeny of the *Amaranthus* genus will aid in selecting accessions for breeding advanced genotypes to satisfy global food demand.

Keywords

Phylogenomics, chloroplast, *Amaranthus*, barcode

Introduction

The plant genus *Amaranthus* comprises approximately 60 species, mostly annuals of naturally open habitats and are distributed throughout the world's tropical and temperate regions (Stetter and Schmid 2017). Although the majority of amaranth species are cosmopolitan weeds, the genus also includes cultivated species used as leafy vegetables (*Amaranthus* subgenus *Albersia*) (van Rensburg *et al.* 2007), a source of grain (*Amaranthus* subgenus *Amaranthus*) (Maughan *et al.* 2009) as well as ornamental plants (Sauer 1967). The edible leafy amaranth species are rich in essential micronutrients including β-carotene (Raju *et al.* 2007; Sangeetha and Baskaran 2010), minerals (Mnkeni *et al.* 2007) and sulphur-containing amino acids (Mlakar *et al.* 2010). Grain amaranths, also referred to as "pseudo-cereals" due to their non-grass nature (Das 2011), have a nutritional advantage over more conventional cereal grains due to their lack of gluten and increased starch and lysine content compared to other cereal crops (Rastogi and Shukla 2013). The grain oil is high in squalene (7-11%) and plays an important role in the medicinal and cosmetic fields (Mlakar *et al.* 2010). Amaranth plants in general are reported to be useful for rehabilitating wastelands (Alamgir *et al.* 2011) and can be used for biofuel production (Timofte *et al.* 2009; Akond *et al.* 2013). Amaranths are therefore an ideal choice for sustainable food production, crop diversification and nutritional security in many nations across the world (Ebert 2014).

Amaranths are cultivated throughout the world, including Central America, Mexico, Eastern Africa and Asia (India, Nepal, China, Indonesia) (Peter and Gandhi, 2017). In Africa and India, *Amaranthus tricolor* is a popular leafy vegetable especially due to the high palatability of the leaves (Srivastava, 2017). It has been found that *A. tricolor* has a high tolerance to drought conditions, and especially to high salinity in the soil compared to other amaranth species (Lubbe et al., 2016). In addition, leaf extracts of *A. tricolor* show antibacterial activities against plant pathogenic bacteria such as *Xanthomonas*, *Erwinia* and *Pseudomonas*. In animal studies (as well as a few clinical trials in humans), the anti-cancer, anti-viral, antioxidant and hepatoprotective properties of leaf and root extracts from *A. tricolor* has been demonstrated (Peter and Gandhi, 2017).

Identification of members within *Amaranthus* based solely on morphological characteristics is difficult, because species are mainly separated based on small (sometimes microscopic) morphological characters. The morphology of *Amaranthus* plants are severely influenced by environmental conditions, species types, production techniques and genotypes, which in turn lead to significant differences in phenotypes between and within species groups (Srivastava 2015). Furthermore, several intermediate morphological forms exist in the wild and sporadic cross-species hybridization occurs, which can result in a mixture of morphological characters that differ from the type species (Gudu and Gupta 1988; Brenner 1990; Achigan-Dako *et al.* 2014). Due to the difficulty associated with morphological identification, the Amaranthus species is thought to belong to three main sub-genera (*Amaranthus* subspecies *Acnida*, *Amaranthus* subgenus *Albersia* and *Amaranthus* subgenus *Amaranthus*). However, the exact evolutionary history and domestication events that led to amaranth species today is still under speculation (Stetter and Schmid, 2017). To overcome problems associated with morphological classification, recent studies have moved towards molecular analysis based on nuclear and

chloroplast sequence information for effective identification and classification of amaranths (Costea *et al.* 2006; Das 2011).

Earlier studies revealed that sequencing data generated from chloroplast genes could provide sufficient information to delineate plant species and phylogenetic reconstruction of taxa (Bell *et al.* 2016; Patil *et al.* 2016; Bezeng *et al.* 2017; Braukmann *et al.* 2017). Each chloroplast is uniparentally inherited and non-recombinant; simplifying phylogenetic reconstructions (Zhang *et al.* 2012). The International Barcoding of Life Initiative (iBOL) has emphasized the taxonomic importance of one or more targeted gene sequences to confidently resolve closely related species within species complexes (Kress and Erickson 2007). Although successful sequence barcodes have been identified for animals (the mitochondrial cytochrome oxidase I gene – *COI*), bacteria (nuclear 16s rRNA) and fungi (nuclear ITS), no such single universal identifier has yet been developed for plants (Lahaye *et al.* 2008; Purty and Chatterjee 2016). It has been proposed that a multi-locus barcode would increase phylogenetic resolution between closely related plant species. Chloroplast gene regions such as *atp*F-H, *mat*K, *psb*K-I, *rbc*L, *rpo*C1, *rpo*B, *trn*H-*psb*A, *trn*L-F and the nuclear intergenic spacer region (ITS) have been evaluated extensively across different species groups (Hollingsworth *et al.* 2011). Despite receiving considerable criticism due to complications in primer design, PCR amplifications, poor performance in the resolution of closely related species and the complicated occurrence of inversions and insertions, the most valuable gene regions for barcoding are the chloroplast *mat*K, *rbc*L, *trn*H-*psb*A, the nuclear ITS and more recently, the *ycf*1 gene (Dong *et al.* 2015). However, the cost and time involved with screening large populations with different gene sets makes barcoding largely impractical.

The increasing affordability of Next Generation Sequencing (NGS) technologies have made efficient, rapid and affordable high-quality sequencing of entire genomes or plastomes possible. The high-copy nature, structural simplicity, highly conserved gene content and relatively small size make chloroplast sequences an ideal target for high throughput sequencing (Stull *et al.* 2013). As a result, tedious and expensive targeted chloroplast isolation is not necessary and low coverage sequencing is sufficient to access many phylogenetically informative characters within the chloroplast genome sequence (Chaney *et al.*, 2016). The multiplex identifier tools employed by NGS technologies further enhance output, as multiple chloroplasts can be sequenced at an adequate depth in a single experiment. As bioinformatic tools become more user-friendly, only moderate computing power and data analysis knowledge will be needed to assemble, annotate and compare full-length chloroplast genomes (Li *et al.* 2014).

Using the whole chloroplast sequence as a "super-barcode" has gained popularity in recent years (Parks *et al.* 2009; Yang *et al.* 2013; Li *et al.* 2014), as it provides considerably more sequence-based variation leading to greatly increased resolution at lower plant taxonomic levels. This method can also circumvent issues pertaining to low PCR efficiency, missing database information of less popular plant species and limited variation supplied by few gene regions. In addition, an in-depth study of a species subset can lead to the identification of genetic barcodes specific to the genus under scrutiny (Li *et al.* 2014; Dong *et al.* 2014). Several studies have employed whole chloroplast sequences with the objective to obtain highly resolved phylogenies. For example, phylogenies were obtained for species in the genera *Oryza* (Hackett *et al.* 2008; Parks *et al.* 2009; Nock *et al.* 2011; Barrett *et al.* 2013; Ma *et al.* 2014), *Bamboo* (Poaceae: Bambusoideae)

(Zhang *et al.* 2011), *Lilium* (Kim and Kim 2013), *Camellia* (Huang *et al.* 2014) and *Acacia* (Williams *et al.* 2016).

To date, eight full-length chloroplast sequences have been assembled for taxa within *Amaranthus*, including five *A. hypochondriacus* accessions, *A. cruentus*, *A. caudatus* and *A. hybridus* (Chaney *et al.*, 2016). Sequencing reads for *A. hypochondriacus* (cultivar Plainsman) were produced through long read PacBio technology, which resulted in a high-quality full-length assembly of the chloroplast. Subsequent sequencing and assemblies of related species revealed several INDEL's, polymorphic microsatellite markers and informative Single Nucleotide Polymorphic (SNP) markers, which can be used in downstream phylogenetic and genetic diversity studies of members of the genus (Chaney *et al.*, 2016).

Due to the mainly maternal inheritance of the chloroplast (and consequent lack of hybridization evidence in the sequence data), several studies are also investigating the use of nuclear SNP based markers to determine plant phylogenies. For *Amaranthus*, a phylogeny for a large set of diverse species was determined using genome wide SNP markers identified through NGS genotyping (Stetter and Schmid, 2017). Based on the results from this study, novel conclusions could be reached about the incomplete domestication syndrome of especially grain *A. caudatus*, and the data also shed light on the complex relationship between domesticated and wild species within this genus (Stetter et al., 2017, Stetter and Schmid, 2017).

In the current study, we generated and annotated the whole chloroplast sequence of *A. tricolor*, an economically important, but neglected leafy vegetable in South Africa. Phylogenetic analysis of the member species was initially undertaken using the most commonly employed barcoding genes (*mat*K, *rbc*L and ITS). Subsequently, the *A. tricolor* whole chloroplast sequence was used as a reference to assemble representative chloroplast sequences of additional 58 diverse *Amaranthus* accessions from South Africa and elsewhere. Whole chloroplast phylogenomics revealed a highly resolved phylogeny within the genus.

Methods and Materials

Growth and maintenance of germplasm accessions

Forty-five accessions representing 13 different *Amaranthus* species were provided by Dr David Brenner, North Central Regional Plant Introduction Station (NCRPIS), Ames, Iowa, United States (Table 1). Accessions were selected from the Germplasm Resources Information Network (GRIN) computer database of the USDA-ARS National plant germplasm system (hereafter referred to as the known, previously identified germplasm set – GRIN). An additional 14 *Amaranthus* accessions, which had previously been collected from different countries across the world, were obtained from the Agricultural Research Council (ARC) - Vegetable and Ornamental Plant Institute (VOPI), Pretoria, South Africa (hereafter referred to as the South African Germplasm set – SAG) (Table 1). Seeds of the 59 accessions were germinated using potting soil in a glasshouse with natural light intensity during day (25°C-35°C) and night (20°C-25°C) at ARC-VOPI. Plants were watered every day for the first three weeks and three times a week thereafter.

Genomic DNA isolation

Approximately 5 grams of young amaranth leaves were collected from seedlings of each accession and stored at -80°C until use. Genomic DNA was isolated using a DNeasy® Plant Mini DNA Isolation kit (Qiagen, Valencia CA, USA) following the protocol provided by the manufacturer. DNA concentrations were determined using the Qubit® 2.0 Fluorometer Broad Range dsDNA quantification assay (Invitrogen, Life Technologies, CA, USA). DNA integrity was evaluated by 1% agarose gel electrophoresis stained with 0.5 μg/ml ethidium bromide.

Table 1: *Amaranthus* genus germplasm sets

GRIN Germplasm				
Code	GRIN Accession	Species	Country of Collection	Species based on whole chloroplast phylogeny[b]
GRIN1	Ames 24670	*A. blitum*	Portugal	*A. blitum*
GRIN2	PI 652433	*A. blitum*	Brazil	*A. blitum*
GRIN3	Ames 13890	*A. caudatus*	China	*A. caudatus*
GRIN4	Ames 15179	*A. caudatus*	Argentina	*A. caudatus*
GRIN5	PI 669934	*A. caudatus*	India	*A. caudatus*
GRIN6	PI 481458	*A. caudatus*	Germany	*A. caudatus*
GRIN7	PI 553073	*A. caudatus*	USA	*A. caudatus*
GRIN8	Ames 2056	*A. cruentus*	Nigeria	*A. cruentus*
GRIN9	Ames 5313	*A. cruentus*	USA	*A. cruentus*
GRIN10	PI 566897	*A. cruentus*	India	*A. cruentus*
GRIN11	Ames 1967	*A. dubius*	India	*A. dubius*
GRIN12	PI 482047	*A. dubius*	Zimbabwe	*A. dubius*
GRIN13	PI 612850	*A. dubius*	USA	*A. dubius*
GRIN14	PI 641049	*A. dubius*	Nigeria	*A. dubius*
GRIN15	PI 608661	*A. graecizans*	India	*A. graecizans*
GRIN16	PI 658732	*A. graecizans*	Portugal	*A. graecizans*
GRIN17	Ames 1990	*A. hybridus*	India	*A. hybridus*
GRIN18	Ames 25409	*A. hybridus*	South Africa	*A. hybridus*
GRIN19	PI 604602	*A. hybridus*	Mexico	*A. hybridus*
GRIN20	PI 641051	*A. hybridus*	Nigeria	*A. hybridus*
GRIN21	PI 652416	*A. hybridus*	Brazil	*A. hybridus*
GRIN22	PI 667174	*A. hypochondriacus*	Zimbabwe	*A. hypochondriacus*
GRIN23	Ames 5689	*A. hypochondriacus*	Brazil	*A. hypochondriacus*
GRIN24	PI 337611	*A. hypochondriacus*	Uganda	*A. hypochondriacus*
GRIN25	PI 538322	*A. hypochondriacus*	USA	*A. hypochondriacus*
GRIN26	PI 619247	*A. hypochondriacus*	Mexico	*A. hypochondriacus*
GRIN27	PI 636187	*A. hypochondriacus*	India	*A. hypochondriacus*
GRIN28	Ames 15306	*A. powellii*	Mexico	*A. hypochondriacus*[b]
GRIN29	PI 572260	*A. powellii*	France	*A. powellii*
GRIN30	PI 604671	*A. powellii*	USA	*A. powellii*
GRIN31	AMES 15315	*A. quitensis*	Argentina	*A. quitensis*
GRIN32	PI 652421	*A. quitensis*	Brazil	*A. quitensis*
GRIN33	Ames 21767	*A. retroflexus*	China	*A. retroflexus*
GRIN34	Ames 25428	*A. retroflexus*	Pakistan	*A. retroflexus*
GRIN35	PI 572263	*A. retroflexus*	USA	*A. retroflexus*
GRIN36	Ames 2150	*A. spinosus*	Kenya	*A. spinosus*
GRIN37	PI 482058	*A. spinosus*	Zimbabwe	*A. spinosus*
GRIN38	PI 632248	*A. spinosus*	USA	*A. spinosus*
GRIN39	Ames 5110	*A. tricolor*	West Africa	*A. tricolor*
GRIN40	Ames 5134	*A. tricolor*	USA	*A. tricolor*
GRIN41	Ames 5139	*A. tricolor*	USA	*A. tricolor*
GRIN42	Ames 23271	*A viridis*	India	*A viridis*
GRIN43	Ames 25412	*A. viridis*	South Africa	*A. viridis*
GRIN44	PI 641048	*A. viridis*	Nigeria	*A. viridis*
GRIN45	PI 654388	*A. viridis*	USA	*A. viridis*
SAG - South African Germplasm				
Code	Accession	Species (preliminary)[a]	Country of Collection	Species based on whole chloroplast phylogeny[b]
SAG 1	50612	*A. bouchonii*	Unknown	*A. powellii*
SAG 3	50613	*A. caudatus*	Unknown	*A. quitensis/A. hybridus*
SAG 4	PI 477913 (Grain)	*A. cruentus*	Mexico	*A. cruentus/A. caudatus*
SAG 7	Arusha leaf	sp.	Unknown	*A. cruentus*
SAG 9	Tanzania	sp.	Tanzania	*A. cruentus/A. caudatus*
SAG 10	Botswana	sp.	Botswana	*A. praetermissus*[c]
SAG 11	W6927N	sp.	Unknown	*A. tricolor*
SAG 12	Bosbok	sp.	South Africa	*A. praetermissus*[c]
SAG 14	Local 33	sp.	South Africa	*A. hybridus*
SAG 17	Vukani Thepe	sp.	South Africa	*A. praetermissus*[c]
SAG 29	*A. tricolor*	*A. tricolor*	USA	*A. tricolor*
SAG 30	Arusha Grain	sp.	Unknown	Unknown
SAG 34	AM Fune	sp.	Unknown	*A. dubius*
SAG 36	AC7	*A. tricolor*	Unknown	*A. tricolor*

[a]Identified based on preliminary morphological analysis (data not shown)

[b]Based on whole chloroplast sequencing - Fig 5

[c]Based on *mat*K sequencing (data not shown)

Illumina library preparation and sequencing

For whole chloroplast sequencing, 5 µg of *Amaranthus tricolor* (SAG29, GenBank accession nr: KX094399) genomic DNA was used to prepare 100 bp paired-end sequencing libraries with the Nextera™ DNA Sample Prep Kit (Illumina, San Diego, USA) according to the manufacturer's protocol. Size selection was performed by excising an approximate 300 bp size fragment from a 1% agarose gel stained with 0.5 µg/ml ethidium bromide using the MinElute Gel Extraction Kit (Qiagen). The sample was sequenced on a HiScanSQ Illumina sequencer (Illumina) using TruSeq SBS v3. The sequencing was performed at the Biotechnology Platform Sequencing Facility, Agricultural Research Council, South Africa. For the remaining 58 *Amaranthus* accessions, 5 µg genomic DNA of each sample was used to prepare sequencing libraries followed by barcode-indexing (Illumina) and sequencing on one lane of a MiSeq Illumina sequencer (Illumina) at the same facility. Approximately 100-200 Mb data was generated for each accession and the resulting raw reads were de-multiplexed.

Assembly of the *Amaranthus tricolor* chloroplast genome sequence

The paired-end sequencing data (2 x 100 bp) was imported to CLC Bio Genomics Workbench v8 (CLCBio, CLC Inc., Aarhus, Denmark). Sequencing adapters and barcodes were trimmed and low-quality reads with Q-value ≤ 30 removed. Trimmed paired end reads were mapped to the chloroplast sequence of sugarbeet (*Beta vulgaris*, GenBank accession nr: EF534108.1), a close relative within the Amaranthaceae family, with default parameters. The consensus *A. tricolor* chloroplast sequence was retrieved and used as a reference for a second round of mapping of *A. tricolor* reads in order to validate the consensus *A. tricolor* chloroplast sequence. All trimmed and quality filtered sequence reads (including reads that mapped to the reference chloroplast sequence and that were used to assemble the *A. tricolor* chloroplast sequence) have been deposited in the Short Read Archive (SRA) archive of NCBI (Accession # PRJNA318736). Non-mapped reads, which are assumed to be of non-plastid origin, were excluded from further analysis. To close gaps and resolve the four junction region sequences between the large single copy, small single copy and inverted repeat regions (IR_A and IR_B), 22 primer pairs (Supplementary Table S1) were designed and used to amplify *A. tricolor* genomic DNA before sequencing on an ABI 3130XL sequencer (Applied Biosystems, CA, USA) at Inqaba Biotechnical Industries Pty. Ltd. (Pretoria, South Africa). In addition, an *A. tricolor* chloroplast sequence was assembled based on mapping reads to the recently published *A. hypochondriacus* chloroplast sequence (GenBank accession nr: KX279888.1) (Chaney *et al.*, 2016), however this was only used for comparison with the final *A. tricolor* chloroplast sequence determined by mapping to the sugarbeet chloroplast, which was used for all subsequent analyses.

***Amaranthus tricolor* chloroplast genome annotation**

The complete *A. tricolor* chloroplast genome sequence was annotated using the Dual Organellar GenoMe Annotator (DOGMA, http://dogma.ccbb.utexas.edu/). DOGMA predicts protein coding genes, ribosomal RNA and transfer RNA genes, together with start and stop codons, as well as the presence of pseudogenes. A circular diagram for the chloroplast was generated using the web-based chloroplast visualization software

GenomeVx (http://oldwolfe.gen.tcd.ie/GenomeVx/). The assembled and annotated *A. tricolor* chloroplast genome sequence was deposited at NCBI (Accession KX094399).

Mapping of the additional *Amaranthus* germplasm sets

Trimmed reads from each species sampled (45 accessions of GRIN and 13 accessions of SAG) were independently mapped to the *A. tricolor* reference (SAG29) sequence using CLC Genomics Workbench v8 with default parameters. Due to the identical nature of the inverted repeat regions of the chloroplast (IR_A and IR_B), the second inverted repeat (IR_B) was removed from the reference chloroplast sequence to simplify mapping and subsequent phylogenetic analysis, as well as to avoid inherent redundancy. By using the default mapping parameters, it was possible to assemble individual consensus chloroplast genome sequences using a small sample read set, while retaining minimal gapped regions (mapping statistics in Supplementary Table S2). The raw reads for each individual mapping was deposited to the SRA archive of NCBI (BioProject: PRJNA318736; SRA: SRS1400803-SRS1400864).

Angiosperm whole chloroplast genome phylogeny

The entire chloroplast genomes of 26 angiosperm species (Supplementary Table S3) representing the main plant family groups were downloaded from the National Centre for Biotechnology Information (NCBI) nucleotide database. Sequences of the 26 species were aligned together with that of *A. tricolor* using MUltiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm in the Molecular Evolutionary Genetics Analysis (MEGA) v6.06 software (Tamura *et al.* 2011) with default settings. Constant, variable and parsimoniously informative site analysis was performed for all sequences. jModelTest v2.1.5 (Guindon and Gascuel 2003; Darriba *et al.* 2012) was used to determine the nucleotide substitution model with the best fit for the dataset, which subsequently led to the incorporation of a General Time Reversible (GTR) model into the phylogenomic analysis. A maximum likelihood analysis was performed using RaxmlGUI v1.31 (Stamatakis 2014) and confidence for nodes determined using bootstrap analysis with 1000 replicates without partitioning the data.

***Amaranthus* barcoding and phylogenetic analysis**

Three barcoding gene regions were investigated during the phylogenetic analysis (chloroplast *mat*K, *rbc*L and nuclear ITS) as previously suggested (Dong *et al.* 2015). The *mat*K gene was amplified using primers *mat*K-F (5'-CACTATGTATCATTTGATAACCCTC-3') and *mat*K-R (5'-TATTACAATCAACATTTCAGAATAG-3') (Burgess *et al.* 2011). Primers *rbc*L-F (5'ATGTCACCACAAACAGAGACTAAAGC-3') and *rbc*L-R (5'-GTAAAATCAAGGTCCACCRCG-3') were used to amplify the *rbcL* gene (Burgess *et al.* 2011). The ITS regions were amplified with primers ITS-F (5'-TCCTCCGCTTATTGATATGC-3') and ITS-R (5'-GGAAGTAAAAGTCGTAACAAGG-3') (Xu and Sun 2001). Amplification of chloroplast and nuclear barcoding genes was performed for 45 GRIN samples (Table 1). DreamTaq PCR Mastermix (2x, DreamTaq DNA Polymerase, 2x DreamTaq Buffer, dNTPs and 4 mM $MgCl_2$) (Thermo Fisher Scientific, Waltham, MA, USA) was used to amplify each sample in a final

volume of 25 µl. The PCR was performed on a G-Storm Thermal Cycler (BioRad Laboratories, CA, USA) with the following conditions: 95°C for 1 min, 30 cycles of 94°C for 30 sec, 55°C for 1 min, 72°C for 1 min and a final extension of 72°C for 10 min. PCR products were visually assessed after electrophoresis on a 1% (w/v) agarose gel stained with 0.5 µg/ml ethidium bromide. Single amplicons obtained for each germplasm accession were sequenced using Sanger sequencing at Inqaba Biotechnical Industries Pty. (Ltd) (Pretoria, South Africa).

Phylogenetic trees were constructed for the *mat*K, *rbc*L and ITS data sets. Sequences obtained for the three barcoding genes of the 45 GRIN *Amaranthus* accessions (Table 1) have been deposited at GenBank as follows: chloroplast *mat*K (KX079543-KX079587); chloroplast *rbc*L (KX079588-KX079632); and nuclear ITS (KX079498-KX079542). Sequence data for *mat*K, *rbc*L and ITS regions for *Beta vulgaris* (used as an outgroup) was obtained from GenBank (accession numbers AY514832.1, DQ067450.1 and AY858597.1, respectively). Each individual barcoding gene was aligned for all the *Amaranthus* species using MUSCLE in MEGA v6.06. In total, there were 702 (ITS), 836 (*mat*K) and 606 (*rbc*L) base pair positions in the final dataset after primer sequences were removed. Model testing (jModelTest) was used to determine the nucleotide substitution model that best fit the sequence alignment for each gene region individually. The three regions were concatenated and phylograms were obtained using a partitioned analysis in RAxML by applying models that were unlinked for each of the genomic regions. Confidence for the nodes on the phylogram was determined using bootstrap analysis with 1000 replicates.

***Amaranthus* whole chloroplast genome phylogenetic analysis**

The large single copy region (LSC), small single copy region (SSC) and IR_A chloroplast sequence regions of 45 GRIN and 14 SAG *Amaranthus* accessions were aligned using Clustal Omega (Sievers *et al.* 2011). Sequence alignments were adjusted manually where necessary. The majority of the sites (70%) that contained missing data for two or more accessions were consequently removed using a custom python script. MEGA was used to determine whether the sites were variable, constant or parsimoniously informative. Phylogenetic analyses were performed using the consensus chloroplast genome sequences of the 45 GRIN and 14 SAG *Amaranthus* accessions. Two datasets were produced; (A) 45 GRIN accessions to confirm identities and classifications of known *Amaranthus* accessions; and (B) 45 GRIN combined with 14 unknown SAG accessions for identifications based on associations to the GRIN accessions. Analyses were conducted on both datasets by treating them as unpartitioned and partitioned in separate analyses. The *B. vulgaris* chloroplast sequence was included as the outgroup to root the trees. Phylogenetic trees were constructed based on both Maximum Likelihood and Bayesian inference of phylogenies. jModelTest was used to determine the nucleotide substitution model that best fit the dataset. Consensus trees and support for nodes were viewed using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Phylogenies based on unpartitioned datasets incorporated entire chloroplast sequences, including genic and intergenic regions. RaxmlGUI v1.31was used for maximum likelihood estimation, utilizing a GTR+G model of nucleotide substitution and 10,000 bootstrap replicates. Bayesian analysis was performed using MrBayes v3.2.4 (Huelsenbeck and Ronquist 2001). The Markov Chain Monte Carlo (MCMC) algorithm was performed

for 10,000,000 generations with trees sampled every 1,000 generations. Tree convergence was assessed by evaluating effective sample size (ESS) values in Tracer v1.6 (available from http://beast.bio.ed.ac.uk/Tracer). The first 20% of trees from the runs were discarded as burn-in and the remaining trees used to generate a consensus tree and calculate the posterior probabilities for each node.

For the partitioned analysis, each of the 119 genic regions on the chloroplast genome sequence was extracted and analysed individually with jModelTest to determine nucleotide substitution models. Phylogenetic analyses were performed using RAxML and MrBayes as described above but with data being partitioned and applying a gene-specific substitution model on each partition in the Bayesian analysis and a GTR+G model on each partition in the maximum likelihood analysis. Genic informative site analysis and model parameters are shown in Supplementary Table S4.

Identification of an alternative barcoding region for the *Amaranthus* genus

The annotated whole chloroplast alignment data was used to identify genic regions demonstrating a high level of nucleotide variation between the different amaranth accessions (Table 2). Sequences from these individual regions were subsequently concatenated to the original barcoding data set (*mat*K, *rbc*L and ITS). A phylogenetic analysis was performed to evaluate whether their addition would add significant phylogenetic signal and increase the robustness of the tree generated from the original barcoding gene data set. Partitioned phylogeny reconstruction was performed for each newly generated alignment (MEGA) using a maximum likelihood analysis with RAxML and 1,000 bootstrap analysis. The TOPological Distance/From Multiple To Single (TOPD/FMTS) (Puigbo *et al.* 2007) software was used to evaluate the differences and similarities of each new barcoding tree topology, compared to the partitioned tree obtained through whole chloroplast analysis. Parameters were set to calculate and evaluate the Split Distance (SD), using a random analysis with 1,000 repetitions.

Table 2: Additional barcoding region statistics

Barcoding region	Size (bp)	Chloroplast region	Informative sites (%)	TOPD Results	
				Disagreement (%)	Split distance
*ndh*D	1,503	SSC	1.9	43	0.58
*rpo*C2	2,584	LSC	2.9	50	0.72
*atp*E	409	LSC	1.4	52	0.65
*rpl*22	592	LSC	5.6	52	0.65
*acc*D	1,659	LSC	1.6	54	0.72
*atp*F	503	LSC	1.8	54	0.72
*psb*A	1,068	LSC	1.2	54	0.70
*rps*14	311	LSC	3.5	56	0.62
*rpo*B	3,249	LSC	1.2	57	0.67
*ndh*G	534	SSC	1.9	58	0.69
*rps*3	656	LSC	2.13	58	0.58
*psb*I-*trn*S IGR	226	LSC	27.2	59	0.74
*trn*Q	91	LSC	18.6	59	0.76
*ndh*H	1,204	IR	3.9	60	0.74
*pet*A	1,005	LSC	14	60	0.81
*psa*C	246	SSC	1.6	60	0.69
*atp*A	1,524	LSC	1.0	63	0.67
*ndh*I	514	SSC	1.16	63	0.63
*ndh*K	861	LSC	2.0	65	0.70
*rpl*16	401	LSC	2.0	65	0.67
*rpo*A	1,012	LSC	2.4	65	0.72
*rpo*B-*trn*C IGR	1,197	LSC	5.4	65	0.72
*trn*R-*atp*A	109	LSC	13.7	65	0.69
*trn*T-*psb*D IGR	1,326	LSC	5.8	65	0.72
*ycf*1	1,407	IR	0.6	65	0.72
*atp*B	1,495	LSC	1.3	67	0.74
*cem*A	702	LSC	1.13	67	0.65
*pet*B	648	LSC	1.0	67	0.65
*pet*D	544	LSC	7.0	67	0.72
*psb*M	103	LSC	7.7	67	0.74
*rps*2	708	LSC	1.0	67	0.74
*trn*S	89	LSC	11.2	67	0.72
*ccs*A	972	SSC	2.9	69	0.74
*rps*8	404	LSC	1.5	69	0.72
*atp*H-*atp*I IGR	679	LSC	4.8	70	0.69
*trn*G	83	LSC	18.0	70	0.72
*trn*N	73	IR	21.9	70	0.72
*psb*D	1,092	LSC	15.0	71	0.81
*trn*S-*rps*4	340	LSC	12.35	71	0.72
*pet*G-*trn*W IGR	136	LSC	13.2	74	0.72
*psb*C	1,476	LSC	11.9	74	0.88
*psb*I	156	LSC	17.0	74	0.74
*rps*16-*trn*Q	1,555	LSC	3.2	74	0.74
*rrn*23-*rrn*4.5 IGR	98	IR	10.2	74	0.72
*trn*C	72	LSC	6.9	74	0.72
*ndh*F	2,316	IR	4.0	76	0.81
*rps*18	309	LSC	13.9	78	0.79
*rrn*16-*trn*I IGR	302	IR	22.5	78	0.86
*rpl*16-*rps*3 IGR	1,588	LSC	10.1	85	0.83
*rrn*16	1,549	IR	18.0	87	1.00
*rrn*23	2,842	IR	6.5	87	0.97

Results

Assembly and annotation of *A. tricolor* chloroplast genome sequence

The chloroplast genome of *Beta vulgaris* (sugarbeet) was used as a reference to construct a preliminary full-length *A. tricolor* chloroplast genome sequence. We generated approximately 1.6 Gb paired-end sequencing data of *Amaranthus tricolor*, 7% of which mapped to the sugarbeet chloroplast genome, with an average coverage of 2 236x. The overall distribution of coverage across the *B. vulgaris* genome can be found in Supplementary Figure S1. Assembly of these reads produced a consensus *A. tricolor* draft chloroplast genome sequence of 149,200 bp. There were 18 gap regions with an average size of 2436 bp. The gapped regions were resolved through Sanger sequencing (Supplementary Table S1), resulting in a total *A. tricolor* complete chloroplast genome sequence of 150,027 bp (Figure 1).

The new assembly conformed to the expected angiosperm chloroplast topology, divided into a large single-copy region (LSC), a small single-copy region (SSC) and two identical inverted repeats (IR_A and IR_B) (Sato *et al.* 1999). Each inverted repeat region had a size of 24,345 bp, separated by 83,735 bp (LSC) and 17,598 bp (SSC). The GC content was 36.6%, similar to *Arabidopsis thaliana* (36.3%) (Sato *et al.* 1999), *Beta vulgaris* (37%) and *Spinacia oleraceae* (36.8%) (Schmitz-Linneweber *et al.*, 2001). The IR_A and IR_B regions each had a GC content of 42.7%, while the GC content of LSC and SSC were 34.5% and 30.2%, respectively. One hundred and nineteen genes were identified from the *A. tricolor* chloroplast genome (Table 3), 21 of which were duplicated in the inverted repeat regions, making the total number of genes 140. Based on the annotations, the highest percentage of genes was related to photosynthesis (28.6%, Table 3).

As a further validation step of the *A. tricolor* chloroplast sequence based on mapping to sugarbeet and subsequent gap-filling by Sanger sequencing, a second *A. tricolor* chloroplast sequence was assembled after mapping all the *A. tricolor* reads to the published *A. hypochondriacus* (KX279888.1) chloroplast sequence. A total of 8.1% *A. tricolor* reads mapped to the *A. hypochondriacus* chloroplast sequence, with an average coverage of 1 272x. The two *A. tricolor* consensus chloroplast sequences produced by mapping to *B. vulgaris* or *A. hypochondriacus* chloroplasts were 99.8% identical, and thus the *A. tricolor* chloroplast sequence based on mapping to *B. vulgaris* was used for all further analyses.

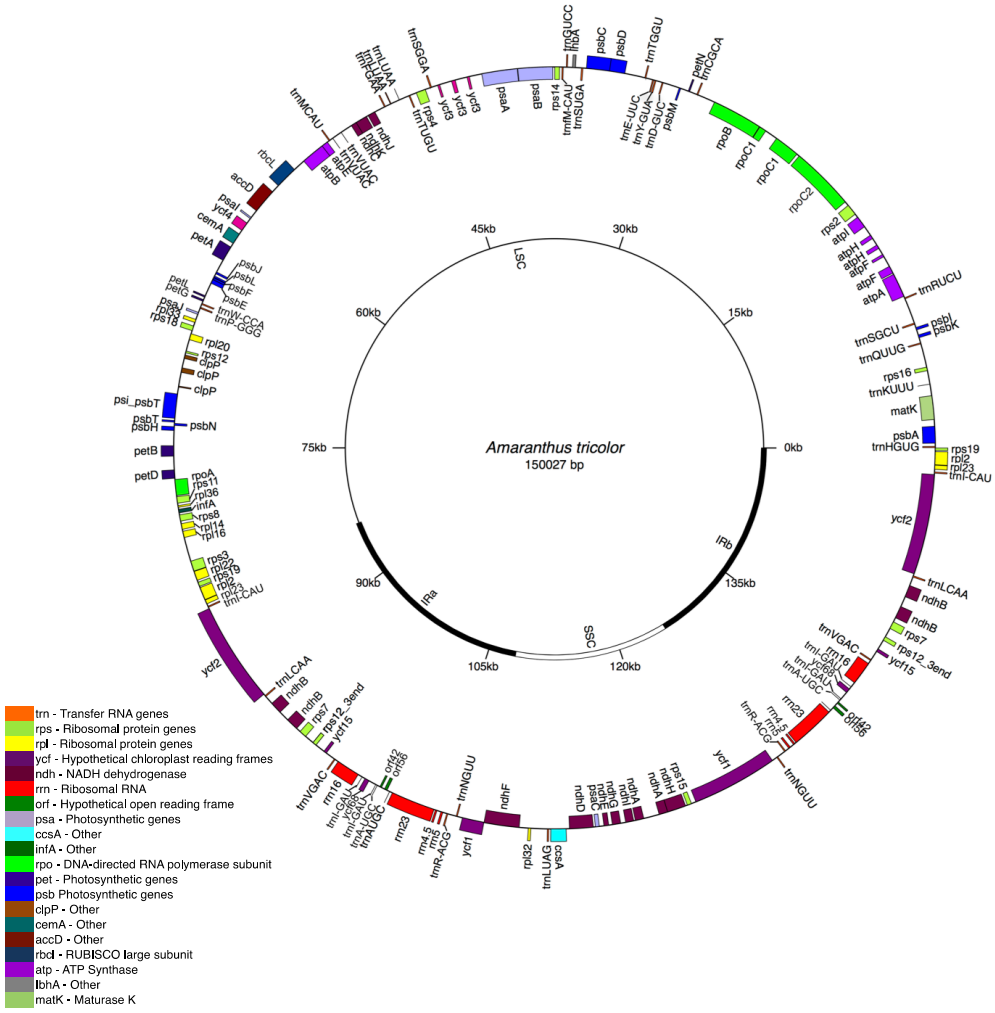

Figure 1: Circular gene map for *Amaranthus tricolor* chloroplast genome. Genes indicated on the outer region are transcribed clockwise while genes on the inside are transcribed counter-clockwise. Genes with similar functions are grouped together and colour coded

Table 3: Classification of gene regions identified on chloroplast genome

RNA genes						
Ribosomal RNA genes						
*rrn*16*	*rrn*23*	*rrn*4.5*	*rrn*5*			
Transfer RNA genes						
*trn*A-UGC*#	*trn*C-GCA	*trn*D-GUC	*trn*E-UUC	*trn*F-GAA	*trn*M-CAU	*trn*G-UCC
*trn*H-GUG	*trn*I-CAU*	*trn*I-GAU*#	*trn*K-UUU	*trn*L-CAA*	*trn*L-UAA#	*trn*L-UAG
*trn*M-CAU	*trn*N-GUU*	*trn*P-GGG	*trn*P-UGG	*trn*Q-UUG	*trn*R-ACG*	*trn*R-UCU
*trn*S-GCU	*trn*S-GGA	*trn*S-UGA	*trn*T-GGU	*trn*T-UGU	*trn*V-GAC*	*trn*V-UAC#
*trn*W-CCA	*trn*Y-GUA	*trn*fM-CAU				
Polypeptide genes						
Ribosomal protein genes						
*rpl*14	*rpl*16	*rpl*2*	*rpl*20	*rpl*22	*rpl*23*	*rpl*32
*rpl*33	*rpl*36					
*rps*11	*rps*12*α*	*rps*12_3end*	*rps*14	*rps*15	*rps*16	*rps*18
*rps*19*	*rps*2	*rps*3	*rps*4	*rps*7*	*rps*8	
*orf*42*	*orf*56*					
Transcription/Translation genes						
*rpo*A	*rpo*B	*rpo*C1#	*rpo*C2	*inf*A		
Photosynthetic genes						
*rbc*L						
*psa*A	*psa*B	*psa*C	*psa*I	*psa*J		
*psb*A	*psb*C	*psb*D	*psb*E	*psb*F	*psb*H	*psb*I
*psb*J	*psb*K	*psb*L	*psb*M	*psb*N	*psb*T	*psi_psb*T
*pet*A	*pet*B	*pet*D	*pet*G	*pet*L	*pet*N	
*atp*A	*atp*B	*atp*E	*atp*F#	*atp*H#	*atp*I	
*ycf*3§	*ycf*4					
NAHD dehydrogenase genes						
*ndh*A#	*ndh*B*#	*ndh*C	*ndh*D	*ndh*E	*ndh*F	*ndh*G
*ndh*H	*ndh*I	*ndh*J	*ndh*K			
Other protein genes						
*acc*D	*clp*P§	*ccs*A	*mat*K	*cem*A	*ihb*A	
Open Reading Frames						
*ycf*1Ψ	*ycf*15*Ψ	*ycf*2*	*ycf*68*			

*	Two copies due to inverted repeat
#	Contains an intron
§	Contains two introns
Ψ	Pseudogene
α	Gene divided into two independent transcription units

Angiosperm whole plastid genome phylogeny

To confirm the phylogenetic position of *A. tricolor* alongside other angiosperm species, an alignment of 87,064 sites was obtained of which 72% (63,184 bp) sites were variable and 52% (32,817 bp) were parsimoniously informative. The topology of the phylogenetic tree generated (Figure 2) was consistent with those from previous studies of angiosperms using a combination of nuclear and plastid genes (Soltis *et al.* 1999; Kuzoff and Gasser 2000). The chloroplast sequence of *Amaranthus tricolor* grouped within the Caryophyllales clade (bootstrap (BS) = 100%) that included *Beta vulgaris* and *Spinacea oleraceae*, which represented the Amaranthaceae family.

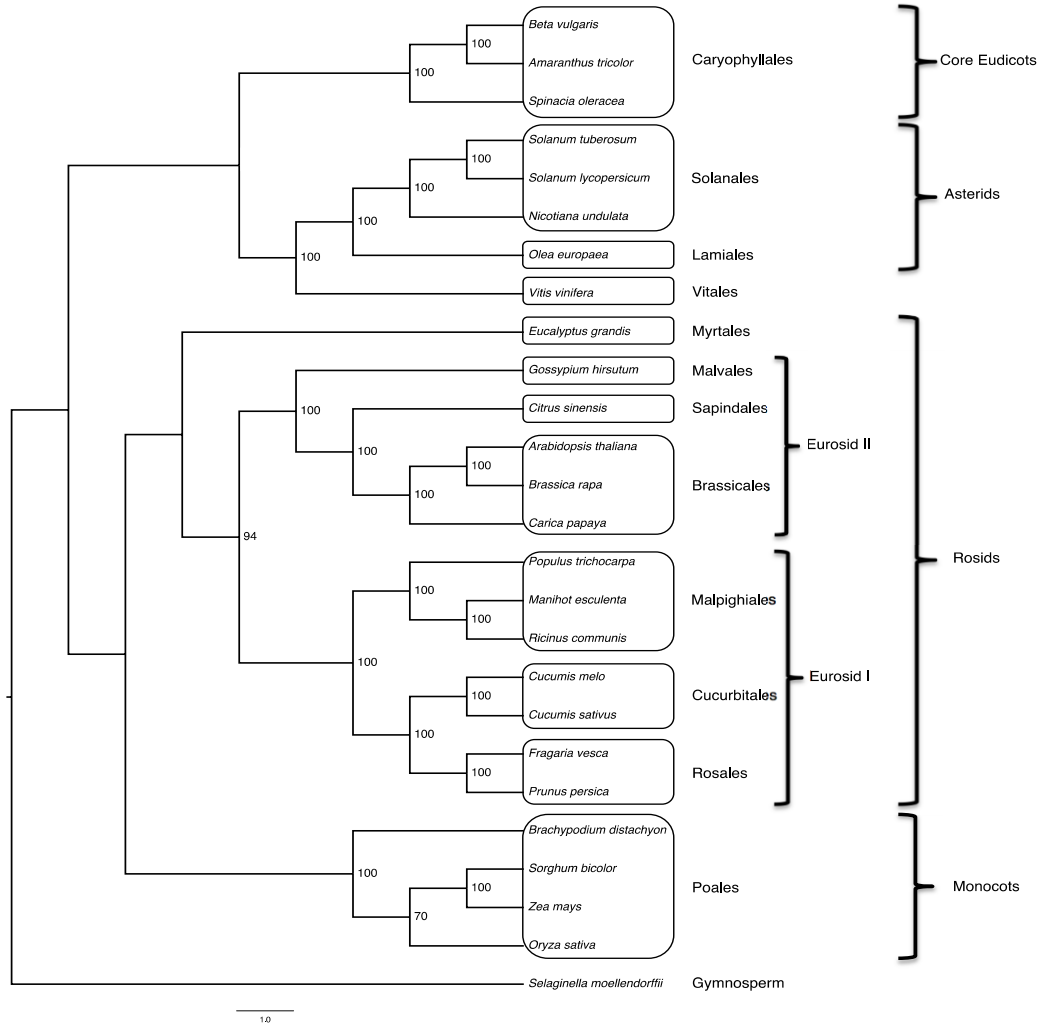

Figure 2: Classification of species within the angiosperm plant group based on maximum likelihood phylogeny (Model GTR, 1000 bootstrap replicates) of whole chloroplast genomes. Bootstrap confidence values (> 60%) are indicated at branch nodes. The addition of the newly assembled *A. tricolor* chloroplast genome reiterated its position within the Caryophyllales flowering plant order, and formed a sister branch to *Beta vulgaris* with 100% bootstrap support

Amaranthus phylogeny using barcoding genes

Chloroplast *mat*K, chloroplast *rbc*L and nuclear ITS gene region sequences from 45 GRIN accessions of the *Amaranthus* genus were determined. Alignment of the concatenated sequences (total length 2,164 bp per accession) produced 80.8% constant, 16% variable, and 2.8% parsimoniously informative sites (Table 4). Phylogenetic relationships between the 45 *Amaranthus* accessions were inferred from all nucleotide sites using the partitioned maximum likelihood method based on the GTR+G (ITS and *mat*K) and HKY (*rbc*L) evolutionary models (Table 4).

Table 4: Barcoding nucleotide site analysis

	Total length (bp, no gaps)	Constant sites (bp)	Variable sites (bp)	Parsimoniously informative sites (bp, %)		JModelTest
*Mat*K	837	723	96	18	2.5%	GTR+G
*rbc*L	625	592	26	7	1.1%	HKY
ITS	702	434	231	37	5.3%	GTR+G
Total	2164	1749	353	62	2.8%	

The *Amaranthus* phylogeny determined by the partitioned maximum likelihood analysis was divided into five clades A – E (Figure 3). Overall, clades A, B and C corresponded to species previously assigned to the *Amaranthus* subgenus *Amaranthus*, while clade D represented species within the *Amaranthus* subgenus *Albersia* (Mosyakin and Robertson, 1996). Weedy amaranths (clade A), a mix of grain and leafy amaranths (clade B and C) and leafy amaranths (clade D) formed separate groupings in this phylogeny (Figure 3).

Although only about one third of the total nodes were supported with a BS > 60%, certain clades and sub-clades could be identified (Figure 3). *Amaranthus retroflexus* (A1) shared a monophyletic origin with *A. powellii* (A2), while clade B was composed of one pure *A. hypochondriacus* subclade (B1), as well as *A. spinosus* (B2), *A. dubius* (B3) and a separate, mixed subclade composed of *A. hypochondriacus, A. hybridus* and *A. caudatus* accessions (B4). Clade B represented both leafy and grain amaranth and formed a sister group with clade C, which was composed of a mixture of species (*A. caudatus, A. quitensis* (C1) and *A. caudatus*; *A. cruentus*; *A. dubius*; *A. hybridus* and *A. hypochondriacus* (C2)). The bootstrap support values were lower than 20% within the entire subclade C2 and positive identifications based on nodal support was not possible. Clade D (a paraphyletic group with the remaining amaranth accessions) indicates a close relationship between *A. tricolor* (D1) and *A. viridis* (D2), but bootstrap support for the monophyletic grouping was not acquired. In clade E, *A. blitum* formed a basal group to clades A, B and C. The remaining accessions (GRIN15, GRIN16 – *A. graecizans*; GRIN28 – *A. powellii*; GRIN7 – *A. caudatus*; GRIN12 – *A. dubius* and GRIN18 – *A. hybridus*) did not fall within a clade and their identities remained uncertain. Overall, the phylogenetic tree generated from the barcoding genes was not well supported due to a lack of statistical confidence (above 60%) at more than 70% of the terminal nodes (Figure 3).

A Bayesian analysis of the partitioned barcoding analysis is provided in Supplementary Figure S2. Overall, the phylogeny was highly comparable to the maximum likelihood analysis, where several nodes were collapsed and had no posterior probability (PP) support for making positive identifications.

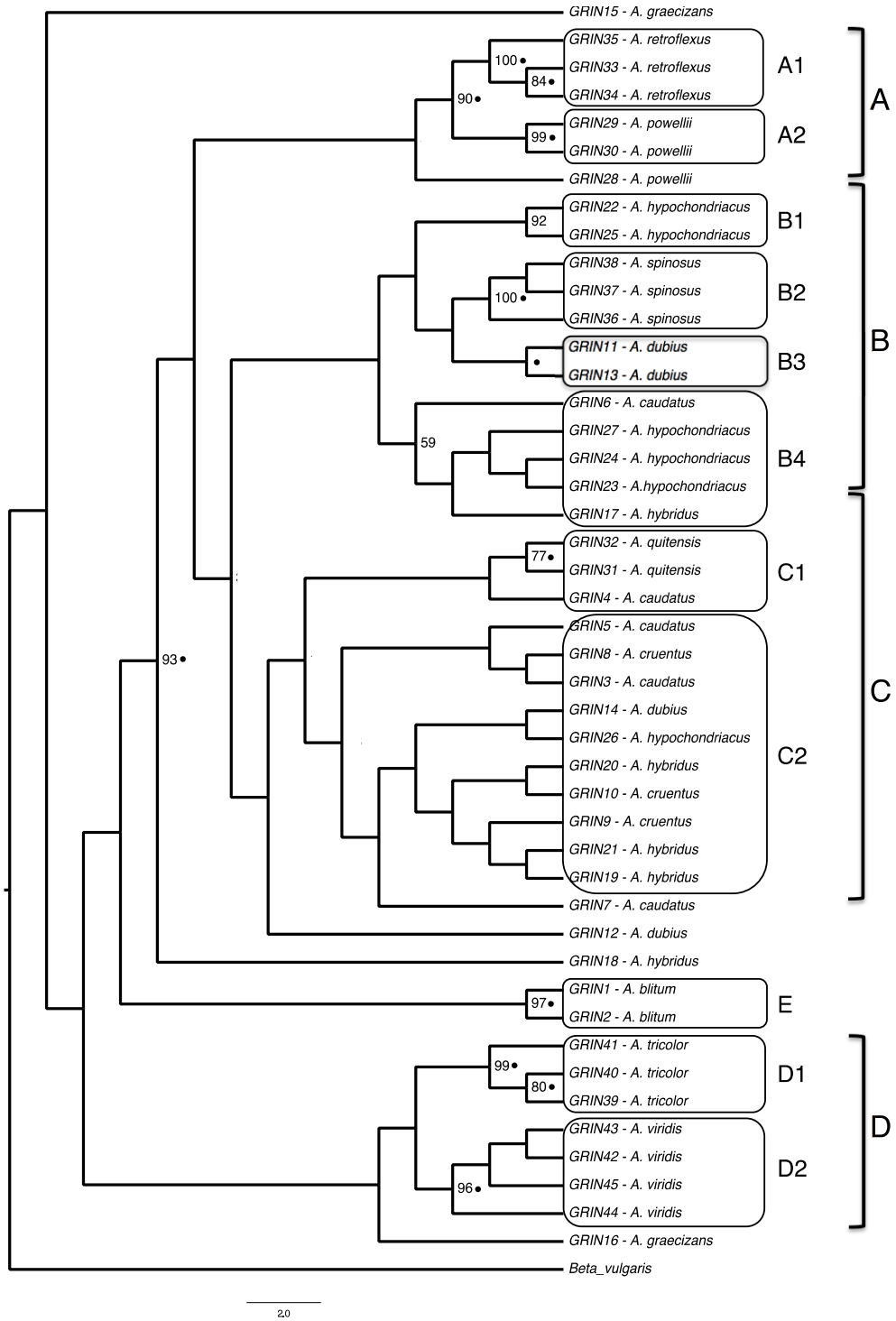

Figure 3: *Amaranthus* phylogeny based on DNA barcoding of chloroplast *mat*K, *rbc*L and nuclear ITS gene regions. The phylogeny was constructed using a partitioned maximum likelihood analysis (Model GTR+G (ITS, *mat*K) and HKY (*rbc*L), 1000 bootstrap replicates), and bootstrap confidence values (> 60%) were indicated at branch nodes. Bayesian probability values above 0.95 are indicated on nodes by (•). The genus is divided into three main subclades, broadly representing weedy amaranth (Clade A), mix of grain and leafy amaranth (Clade B and C) and leafy amaranth (Clade D and E) accessions. GRIN – Previously identified amaranth accessions (Table 1)

Amaranthus whole chloroplast genome phylogenetic analysis

The chloroplast genome sequence for leafy amaranth *A. tricolor* was successfully constructed and subsequently served as a mapping reference for the chloroplast genomes of an additional 45 GRIN and 13 unknown SAG *Amaranthus* accessions. After strict mapping, there were 3% gaps on average for most accessions with GRIN14 having the least gapped sites (0.2%) and GRIN7 having the highest number of gapped sites (21%). Gapped regions were excluded from further analysis, regardless of whether other accessions had sequence information for those nucleotide positions.

Partitioned analysis of 45 GRIN accessions (chloroplast sequences)

The 45 previously identified GRIN accessions were investigated for their genetic relatedness and revealed better resolution of clade separation through whole chloroplast phylogenomics (both partitioned and unpartitioned datasets), than for barcoding analysis. For the unpartitioned analysis, the alignment of 45 GRIN amaranth chloroplast sequences (with IR_B removed) revealed a total alignment length of 246,850 bp. In the partitioned analysis, 119 genic regions on the chloroplast genome were included in the final alignment. The site characteristics before and after gap removal are presented in Table 5. Both partitioned and unpartitioned analysis resulted in highly similar tree topologies (data not shown).

Table 5: Whole chloroplast GRIN nucleotide site analysis

	Total length (bp)	Constant sites (bp)	Variable sites (bp)	Parsimoniously informative sites (bp)	
Unpartitioned Analysis					
Chloroplast Alignment with gaps	246,850	115,389	17,831	4,296	1.7%
Chloroplast Alignment with gaps removed	76,919	72,175	4,744	809	1.0%
Partitioned Analysis					
Chloroplast Alignment with gaps	58,219	53,345	4,338	764	1.3%
Chloroplast Alignment with gaps removed	45,504	42,593	2,911	481	1.0%

Due to the highly similar tree topology resulting from both the partitioned and unpartitioned analyses, only the partitioned tree was further examined (Figure 4). Overall, the four main clades (clades A to D) obtained from the barcoding gene tree were retained. These clades reflected weedy amaranths (clade A, BS=96%, PP=1), a combination of weedy and grain amaranths (clade B, BS=89%, PP=0.97), grain amaranths (clade C, BS=90%, PP=0.95) and leafy amaranths (clade D, BS=100%, PP=1). Clade A was composed of subclade A1 (*A. powellii*) and subclade A2 (*A. retroflexus*) (Figure 4). Each of the subclades was supported with BS=100% and PP=1. The grouping of these species mirrors the monophyly obtained from the barcoding analysis (Figure 3). The grouping of accessions within clades A1 and A2 were supported with BS=100%, respectively. Clade B was comprised of three main subclades (B1+B2, B3 and B5). Subclades B1+B2 (BS=95%, PP=1) were combined (in contrast to the barcoding analysis, Figure 3) and included three accessions of *A. spinosus* and two accessions of *A. hypochondriacus*. Subclade B3 (two accessions of *A. dubius*) was separated into a highly supported clade (BS=100%) that formed a sister group to clade B1+B2. Accessions residing in sub-clade B4 from the barcoding analysis (Figure 3) had been re-distributed among the remaining clades on the tree

generated from the whole chloroplast analysis. Accessions GRIN23, GRIN24 and GRIN27 (*A. hypochondriacus*) now formed part of sub-clade C7 while GRIN 17 (*A. hypochondriacus*) fell within sub-clade C8. Accessions GRIN6 (*A. caudatus*) formed an outlier to sub-clades C7 and C6 (Figure 4). Subclade B5 represented a group not observed during the analysis of the barcoding genes, and contained a mixture of *A. quitensis, A caudatus and A. hybridus* accessions.

Four subclades could be separated in clade C based on phylogenetic analysis of the chloroplast genome (Figure 4). The subclades included in this main clade were: C5 (two accessions of *A. dubius*); C6 (two *A. cruentus*, one *A. hybridus*, one *A. hypochondriacus* and one *A. powellii* accession); C7 (three *A. hypochondriacus* and one *A. cruentus* accessions) and C8 (two accessions of *A. hybridus*).

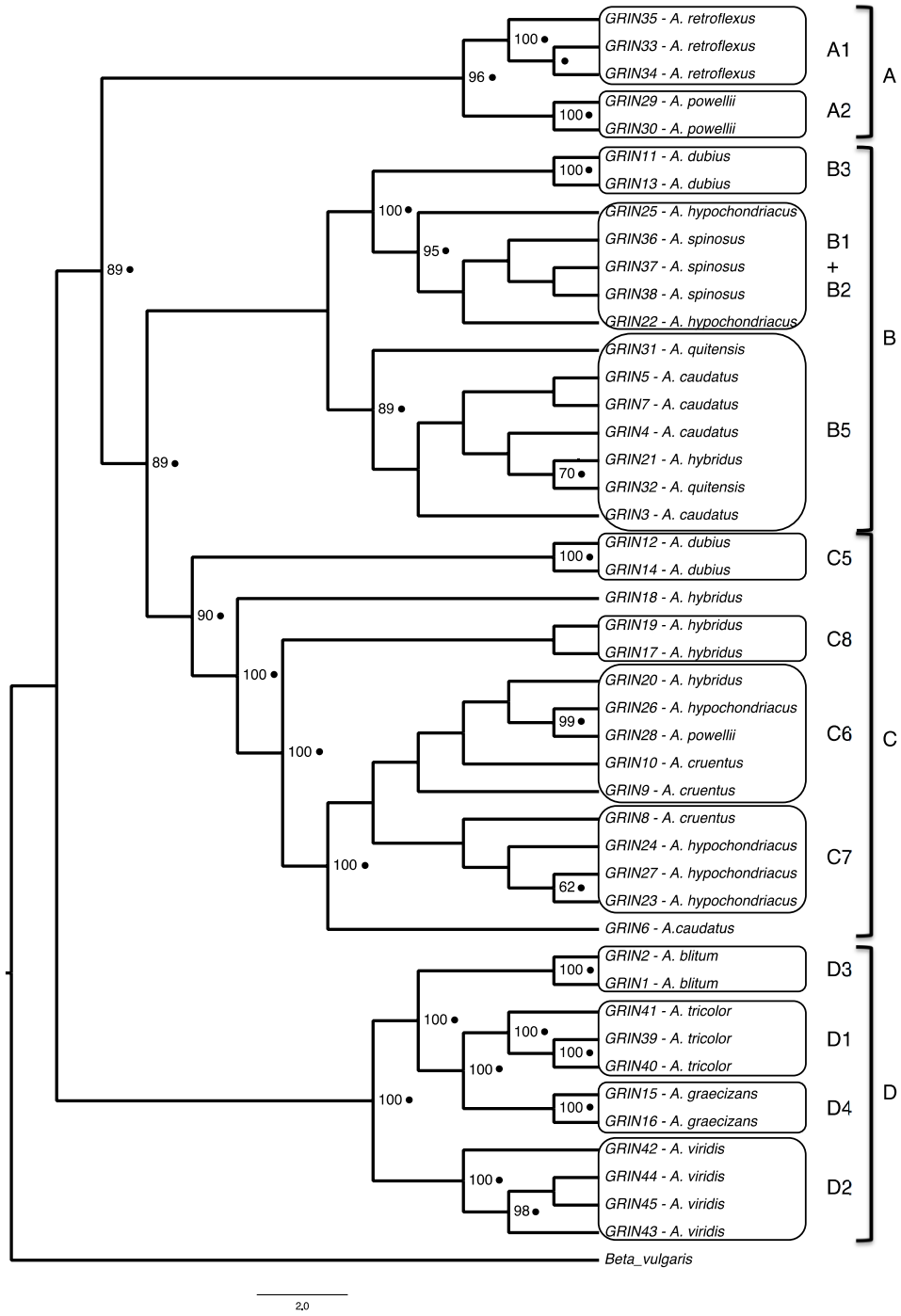

Figure 4: Phylogeny of taxonomically described *Amaranthus* accessions. The analysis was based on a partitioned maximum likelihood analysis of 45 GRIN amaranth accessions. (Model GTR+G, 1000 bootstrap replicates) of whole chloroplast genome sequences. Bootstrap confidence values (> 60%) are indicated at branch nodes. Bayesian probability values above 0.95 are indicated on nodes by (•). Four main clades are identified: (A) weedy amaranth (*A. retroflexus, A. powellii*); (B) weedy amaranth (*A. dubius, A. spinosus, A. quitensis*), grain amaranth (*A. hypochondriacus, A. caudatus*); (C) weedy amaranth (*A. dubius*), grain amaranth (*A. hypochondriacus, A. cruentus*), leafy amaranth (*A. hybridus*) and (D) leafy amaranth (*A. blitum, A. tricolor, A. graecizans, A. viridis*)

Subclades C1 and C2, which were observed in the barcoding analysis (Figure 3), were not subsequently found in the tree obtained from the whole chloroplast analysis and their nodes were re-distributed across newly formed subclades. Within subclade C6, accession GRIN8 (*A. cruentus*) could possibly have been re-classified as *A. hypochondriacus* due to the close association with three other *A. hypochondriacus* accessions, but there was little support for the node grouping these species. The *A. dubius* subclade (C5) was separated from the remaining subclades with BS=90% and PP=0.99, while *A. hybridus* (subclade C8) formed a polytomy at the node that groups clade C. Phylogenetic trees generated from this analysis indicated that subclades C6 and C7 have a monophyletic origin, but identification within each subclade remains difficult due to low internal nodal support of the majority of the branches. A mix of mainly *A. cruentus* and *A. hypochondriacus* accessions appeared in subclades C6 and C7 and conclusive identification of each individual accession remains elusive. Clade D included four subclades, each supported with BS=100% and PP=1 that represented clear species groups (Figure 4). These groups included subclades D1 (*A. tricolor*), subclade D2 (*A. viridis*), subclade D3 (*A blitum*) and subclade D4 (*A. graecizans*). *Amaranthus tricolor* (subclade D1) and *A. graecizans* (subclade D4) appeared to share a common ancestor BS=100% and PP=1. *Amaranthus blitum* (D3, previously clade E from the barcoding analysis, Figure 3) had a sister relationship (BS=100%; PP=1) with subclades D1 and D4. Clade E formerly only had a basal relationship to the D clade. *Amaranthus viridis* (subclade D2) formed a sister group with the monophyletic clade that included subclades D1, D3 and D4 (BS=100%; PP=1). Clade D exclusively represented species belonging to the leafy amaranth group. Whole chloroplast genome alignments were also subjected to Bayesian analysis (Supplementary Figure S3). For 96% of the accessions, the Bayesian analysis was identical to the RAxML result, but higher bootstrap support can be seen for 13 of the terminal nodes. Two accessions of *A. caudatus* (GRIN6 and GRIN8) formed a collapsed side branch to subclade C6, while in the RAxML analysis GRIN8 was part of subclade C7.

Partitioned analysis of 45 GRIN and 14 SAG accessions (chloroplast sequences)

Thirteen unknown SAG *Amaranthus* accessions were added to the GRIN analysis to investigate the possibility of their identification and classification. The total length of the alignment was 246,850 bp and the phylogenetic analysis was performed with RAxML and MrBayes software. Nucleotide site analysis for the alignments is presented in Table 6.

Table 6: Whole Chloroplast GRIN + SAG nucleotide site analysis

	Total length (bp)	Constant sites (bp)	Variable sites (bp)	Parsimoniously informative sites (bp)	
Unpartitioned Analysis					
Chloroplast Alignment with gaps	246,850	115,312	19,154	5,246	2.1%
Chloroplast Alignment with gaps removed	76,857	72,013	4,844	914	1.2%
Partitioned Analysis					
Chloroplast Alignment with gaps	58,792	53,289	4,202	851	1.4%
Chloroplast Alignment with gaps removed	45,305	42,541	2,764	537	1.2%

The majority of the main clades and subclades (A, B, C and D) obtained from the GRIN analysis (Figure 4) was observed for the combined analysis of GRIN and the unknown SAG accessions (Figure 5). However, the inclusion of additional amaranth individuals resulted in the formation of three new subclades; C9, C10 (sister group to clade C7) and D5 (sister group to clades D1 and D4) (BS=100%). The Bayesian analysis of this dataset revealed an identical topology and only minor differences in nodal support values for eight of the nodes could be observed (data not shown).

The partitioned chloroplast sequence analysis enabled species identifications for most of the SAG accessions (Figure 5), and their updated identities have been included in Table 1. Firstly, the accession sequenced in this study, SAG29, together with SAG11 and SAG36 were confirmed to be *A. tricolor*, since they grouped in clade D1 with the three *A. tricolor* GRIN accessions. SAG1 was identified as an accession belonging to the *A. powellii* group with strong bootstrap support (clade A2). SAG14 was identified as *A. hybridus* since it forms a highly supported sub-clade C9 with GRIN18 (*A. hybridus*). SAG4 and SAG9 are members of the Hybridus complex, since they group with GRIN6 (*A. caudatus*) and GRIN8 (*A. cruentus*). SAG3 was identified as an accession of *A. quitensis* or *A. hybridus* based on its grouping with GRIN32 (*A. quitensis*) and GRIN21 (*A. hybridus*), and all three of these accessions originate from Brazil, which adds confidence to the identification. SAG34 was identified as *A. dubius* since it groups with two *A. dubius* accessions in sub-clade B3 with good bootstrap support. SAG7 groups with two accessions of *A. cruentus* (GRIN9 and GRIN10), with low bootstrap support, and is therefore cautiously classified as an accession of *A. cruentus*. SAG10, SAG12 and SAG17 are closely related in a newly formed sub-clade D5 and are separated from the monophyletic *A. tricolor*/*A. graecizans* clades (100% bootstrap). All three of these accessions were collected in southern Africa and BLAST sequence analysis of their *mat*K barcoding gene revealed a potential identification as *A. praetermissus* (data not shown). The accession SAG30 could not be identified since it is included in clade B with strong bootstrap support, but forms an outgroup to sub-clades B1/B2 and B3.

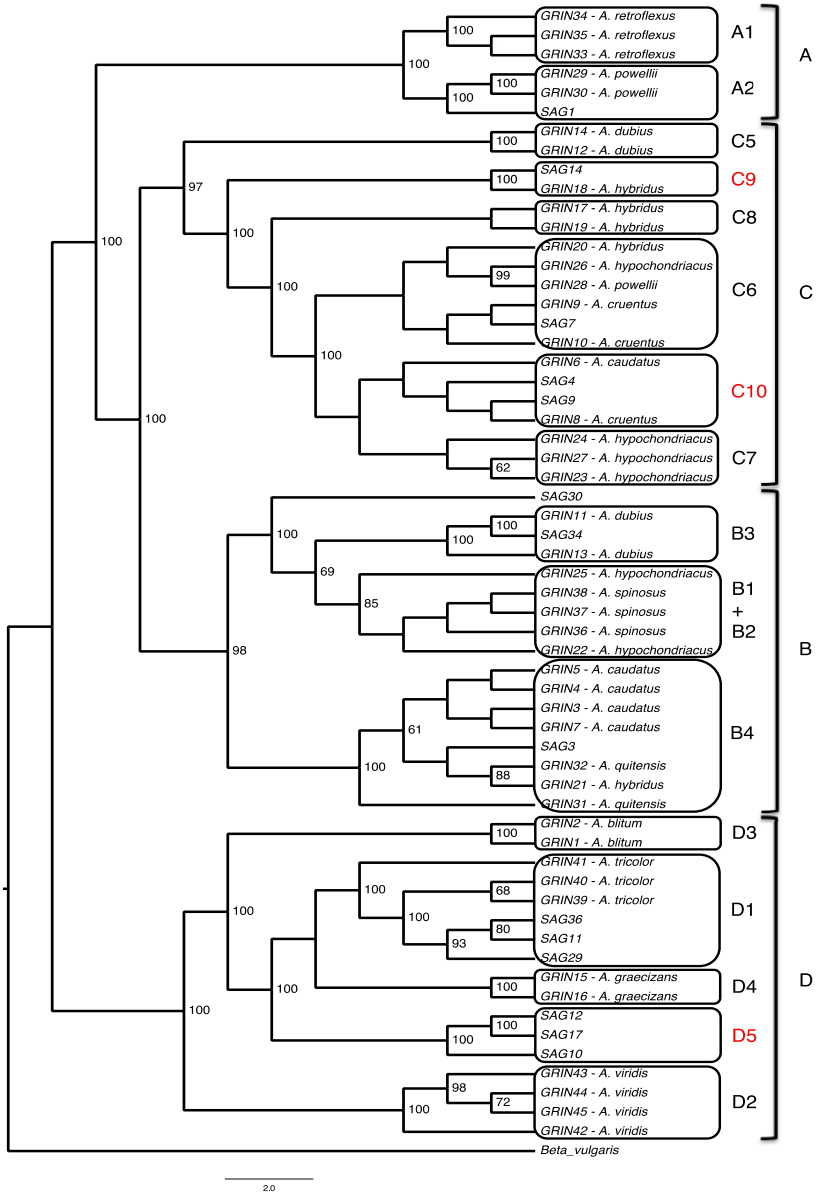

Figure 5: Phylogenetic distribution of unknown *Amaranthus* accessions in relation to previously described *Amaranthus* species. The analysis was based on a partitioned maximum likelihood analysis of 45 GRIN and 14 SAG amaranth accessions (Model GTR+G, 1000 bootstrap replicates) of whole chloroplast genome sequences. Bootstrap confidence values (> 60%) are indicated at branch nodes. Four main clades are identified: (A) weedy amaranth (*A. retroflexus, A. powellii*); (B) weedy amaranth (*A. dubius, A. spinosus, A. quitensis*), grain amaranth (*A. hypochondriacus, A. caudatus*); (C) weedy amaranth (*A. dubius*), grain amaranth (*A. hypochondriacus, A. cruentus*), leafy amaranth (*A. hybridus*) and (D) leafy amaranth (*A. blitum, A. tricolor, A. graecizans, A. viridis*). The addition of unknown amaranth accessions (SAG) results in the formation of additional sub-clades C9, C10 and D5

Investigation of genes to complement existing barcoding genes for phylogenetic analysis of *Amaranthus* species

A set of genic and intergenic regions of the whole chloroplast genome sequence were investigated for regions which could contribute to an informative phylogeny for *Amaranthus* without using whole chloroplast sequence data. Using the whole chloroplast alignment, 51 potential new barcoding regions were extracted that contained more than 1% parsimoniously informative characters across the total sequence length. Each individual region was joined to the original alignment of ITS, *mat*K and *rbc*L and a new sequence alignment was obtained. In total, 51 maximum likelihood phylogenetic trees were generated based on the newly selected gene together with the original barcoding genes; and were compared to the whole chloroplast genome phylogeny to assess the contribution of the gene in context to the overall phylogenetic signal by using the TOPD/FMTS software. Firstly, when the disagreement between the original barcoding tree (ITS, *mat*K and *rbc*L) and the whole chloroplast tree was assessed, it was found that 27 out of 46 (58.7%) terminal node positions varied between the trees (represented by blue bar in Supplementary Figure S4). The split distance (the number of splits that disagree between the two trees) was calculated as 74%, which indicated that only 26% of the bipartitions were shared between the respective trees. While evaluating the potential alternative barcoding trees, 55% was taken as a maximum level of disagreement (to create a manageable dataset for downstream analysis). For the purpose of this study, three gene phylogenies displaying the least disagreement (therefore highest congruence) to the whole chloroplast phylogenetic tree were found to be *ndh*D (43% disagreement), *rpo*C2 (50% disagreement) and *rpl*22 (52% disagreement) (Supplementary Figure S4). The addition of the *ndh*D, *rpl*22 and *atp*E gene regions to the existing barcoding dataset resulted in a 26%, 22% and 13% improvement in nodal placement compared to the original barcoding phylogenies, while 31%, 28% and 63% nodal placements remained constant between the trees. Previous studies have used *ndh*D due to the number of informative sites available, but it seems to vary between genera (1.9% in *Amaranthus* vs. 19% in *Asteraceae*) (Panero and Funk 2008; Nock *et al.* 2011; Dong *et al.* 2013; Shaw *et al.* 2014). Due to their low level of agreement, *rpl*22 and *atp*E were not considered for further analysis.

The new barcoding phylogeny based on *ndhD* as well as ITS, *mat*K and *rbc*L (Supplementary Figure S5) indicated improved clade placements (and therefore putative identifications) of five accessions compared to the original barcoding tree (GRIN31, GRIN32 – *A. quitensis* and GRIN4, GRIN5, GRIN7 – *A. caudatus*). However, the addition of the *ndh*D gene region also resulted in the placement of six accessions as outliers to defined clades and therefore hampered their identifications (GRIN8, GRIN9, GRIN10 – *A. cruentus*; GRIN19, GRIN20 – *A. hybridus* and GRIN26 – *A. hypochondriacus*). Overall, the addition of *ndh*D did not significantly improve clade resolution within the *Amaranthus* genus. As a whole, the whole chloroplast phylogeny remained superior in resolving *Amaranthus* species relationships, compared to the barcoding analysis.

Discussion

The plant genus *Amaranthus* contains many economically important species, as well as potentially useful orphan crops, which can contribute to global food security. Several earlier studies showed that the phylogenetic relationships between the *Amaranthus* species are highly intricate and difficult to resolve (Chan and Sun 1997; Xu and Sun 2001; Mandal and Das 2002; Costea *et al.* 2006; Mallory *et al.* 2008; Gerrano *et al.* 2015), especially among the grain and weedy types. Phylogenetic analysis based on data sets containing large numbers of DNA or amino acid characters can reveal increased resolution (and higher support) for clade hypotheses (Straub *et al.* 2012). In the past, phylogenetic analysis was constrained by the high cost of sequencing and was therefore limited to a few gene loci that were considered highly informative (mainly *mat*K, *rbc*L and ITS). However, with the advent of next-generation sequencing technologies, massive parallel sequencing has become the method of choice for rapid sequencing of plastid genomes (Parks *et al.* 2009) resulting in 2 086 complete eukaryote chloroplast genome sequences being available in GenBank (26 February 2018, www.ncbi.nlm.nih.gov/genomes).

The *Amaranthus tricolor* chloroplast assembly

This study is the first to report a fully assembled and annotated chloroplast sequence of the leafy vegetable, *A. tricolor*. The chloroplast sequence was assembled by mapping to the chloroplast sequence of *B. vulgaris*, which is from a different genus of the family Amaranthaceae. Comparison to an *A. tricolor* chloroplast sequence based on mapping to the *A. hypochondriacus* chloroplast revealed high identity (99.8%). This indicates that whole chloroplast sequence assembly of a plant species does not require a reference genome from the same genus, since a chloroplast genome from the same family will suffice.

Sequence identities, gene organization and relative positions of the genes in *A. tricolor* and other angiosperm species were highly similar, corroborating the conserved nature of plant chloroplast genomes (Sugiura 1992; Schmitz-Linneweber *et al.* 2001). Nine of the 119 gene regions identified on the *A. tricolor* chloroplast genome contained a single intron (*trn*A-UGC, *trn*I-GAU, *trn*L-UAA, *trn*V-UAC, *rpo*C1, *atp*F, *atp*H, *ndh*A, *ndh*B), while two genes contained two introns each (*ycf*3, *clp*P). The remaining genes contained no intronic regions. The intron that was found in *trn*K-UUU of spinach (*Spinacea oleraceae*, Amaranthaceae), another leafy vegetable, was not found in *A. tricolor* (Chaney *et al.* 2016) nor in *Arabidopsis thaliana* (Sato *et al.* 1999), *Solanum tuberosum* (Chung *et al.* 2006) and *Nicotiana tabacum* (Shinozaki *et al.* 1986). The intron sizes for *A. tricolor* were more than 80% similar to the introns found in chloroplast genomes of *S. oleraceae* (Schmitz-Linneweber *et al.* 2001), *A. thaliana* (Sato *et al.* 1999), *S. tuberosum* (Chung *et al.* 2006) and *Artemisia frigida* (Liu *et al.* 2013). The only gene with a significant difference in intron size was *trn*L-UAA, which was 50% larger in *A. tricolor* than in *S. oleraceae* (Schmitz-Linneweber *et al.* 2001).

The presence of pseudogenes was observed in the *A. tricolor* chloroplast genome. Analysis of *ycf*1 and *ycf*15 showed that both had premature stop codons within the sequence. However, it is possible for these genes to still produce functional proteins after translation (Poliseno *et al.* 2010). As in plastids of other higher plants, there were potential open reading frames (ORFs) for which no functions have yet been inferred. Particular

ORFs are conserved between different plant species (also known as hypothetical chloroplast reading frames – *ycf*) (Schmitz-Linneweber *et al.* 2001). The *A. tricolor* chloroplast genome harboured six *ycf* genes (*ycf*1, *ycf*2, *ycf*3, *ycf*4, *ycf*15 and *ycf*68) and the comparable predicted open reading frames suggested they may form similar polypeptides as *ycf* genes in other species. Dong *et al.* 2015 recently suggested that the *ycf*1 gene could be part of a protein channel present in the membrane of chloroplast cells and can also potentially be used as an additional plant barcode due to its high variability between taxa. In addition, two more open reading frames (*orf*42 and *orf*56) were identified for *A. tricolor*, which had 98% sequence identity between *A. tricolor, S. oleraceae,* and *S. tuberosum.*

Resolving the position of *A. tricolor* within the Angiosperm phylogeny

Utilizing the large collection of angiosperm plastid sequences available on GenBank, it was possible to re-construct a broad phylogeny of several different plant orders. The newly assembled chloroplast of *A. tricolor* was placed within the *Caryophyllales* together with *B. vulgari*s and *S. oleraceae*, representing the Amaranthaceae family. The correct placement of *A. tricolor* is supported by previous work on two important morphological traits shared by species within the Amaranthaceae, namely the presence of betalain pigments conferring leaf/stem/flower colours instead of anthocyanins (Cuénoud *et al.* 2002; Venskutonis and Kraujalis 2013) and the C_4 photosynthetic machinery enabling these plants to thrive in warm, arid areas (Alemayehu *et al.* 2015), both of which have been reported for *A. tricolor* (Achigan-Dako *et al.* 2014).

Whole chloroplast analysis provides better phylogenetic resolution within *Amaranthus* than DNA barcoding

During this study, the complicated nature of *Amaranthus* phylogeny was demonstrated as previously reported when using AFLP's (Xu and Sun 2001), RAPD's (Mandal and Das 2002), combined AFLP and micromorphology (Costea *et al.* 2006), microsatellite markers (Mallory *et al.* 2008) and SNP marker analysis (Stetter *et al.* 2016). Chan and Sun (1997) initially revealed a close relationship between the grain *A. cruentus* and *A. hypochondriacus*; as well as the potential ancestor *A. hybridus* using RAPD analysis. In addition, *A. caudatus* formed a close sister clade together with a weedy *A. dubius* relative in their study. This result was highly congruent with the phylogeny obtained when genome wide SNP data was used, indicating a mostly robust relationship between the grain amaranth species (*A. cruentus, A. caudatus* and *A. hypochondriacus*), their potential progenitors (*A. hybridus* and *A. quitensis*) and the wild *A. dubius* (Stetter *et al.* 2016). Our analysis corresponds in part with what was reported previously, where *A. cruentus* accessions formed a sister clade to *A. hypochondriacus* (Chan and Sun 1997). However, in contrast to the aforementioned studies, the *A. caudatus, A. hybridus* and *A. dubius* accessions did not conform to separate clades and were found scattered between and within the *A. cruentus* and *A. hypochondriacus* clades. In particular, the presence of *A. hybridus* within different clades implies that the complex had a recent split from the remaining species in the *Amaranthus* genus. Our results may in part be explained by the fact that the maternally inherited chloroplast sequences were used, and that better resolution between species that share the common ancestor of *A. hybridus*, which diverged recently, may be obtained from nuclear markers.

When more accessions of leafy and wild amaranth species are included in the phylogenies, a consistent grouping of *A. powellii* together with *A. retroflexus* was observed using microsatellite (Mallory *et al.* 2008) and SNP marker (Stetter *et al.* 2016) analysis. This was also reflected in the chloroplast analysis conducted in the present study. In all cases, the *A. powellii*/*A. retroflexus* clade form a sister clade to the Hybridus complex (overall consisting of *A. cruentus, A. hypochondriacus, A. caudatus, A. hybridus, A. quitensis* and *A. dubius*). These clades correlate well with members assigned to the *Amaranthus* subgenus *Amaranthus*, although the groupings within this subgenus are still mostly unresolved (Mosyakin and Robertson, 1996). The addition of leafy amaranth accessions *A. tricolor*/*A. viridis* (which fall within one sub-clade) and *A. blitum* (forming a close sister clade to *A. tricolor*/*A. viridis*) in the present study resulted in a highly supported separated clade from the species discussed previously, and concurs with the results obtained during RAPD (Chan and Sun 1997) and SNP (Stetter *et al.* 2016) analysis. This clade of leafy amaranth corresponds to the previously defined *Amaranthus* subgenus *Albersia* (Mosyakin and Robertson, 1996). The only species that did not show a consistent phylogeny in all the studies mentioned was *A. spinosus*, which grouped together with *A. powellii*/*A. quitensis* using RAPD markers (Chan and Sun 1997) and formed a completely separate clade during SNP analysis (Stetter *et al.* 2016). In the barcoding phylogeny obtained in our study, it grouped strongly with two *A. hypochondriacus* accessions.

Another aim of this study was to identify a universal set of chloroplast barcoding genes for *Amaranthus* species. Previous studies indicated that the *ndh*D gene, in combination with the other chloroplast barcoding genes, might have the potential to correctly delineate species (Panero and Funk 2008; Nock *et al.* 2011; Dong *et al.* 2013; Shaw *et al.* 2014). However, in *Amaranthus*, the *ndh*D gene had the same level of informativeness (1.9% parsimoniously informative sites) as the previously used chloroplast *mat*K (1.9%) and *rbc*L (1.1%) and did not enhance the barcoding phylogeny. In contrast to previous studies, a decreased number of accessions for which positive identifications could be made was observed. It is clear that future studies aiming at the development of genus/taxonomic specific barcodes in *Amaranthus* should focus on chloroplast intergenic regions, nuclear gene regions or other genetic markers.

Implication of the low phylogenetic resolution to *Amaranthus* phylogeny

The barcoding phylogenetic tree of the *Amaranthus* germplasm set generated from what is considered the 'core plant barcodes' (including *mat*K, *rbc*L and nuclear ITS gene regions) overall suffered from low resolution, making inferences regarding the phylogenetic relationships of the taxa problematic. The main significance of the resultant *Amaranthus* phylogeny is that standard plant barcodes may be best suited to substantiate existing species classifications, rather than having the power to discriminate between unknown inter- and intra-species accessions as previously reported (Hollingsworth *et al.* 2016). The low resolution obtained from the barcoding phylogenetic analyses suggested that the diversification of particularly the grain amaranths was fairly recent, and that an incomplete domestication event is observed as was reported by (Stetter *et al.*, 2017). This warrants further investigation into an integrative approach of morphology and additional informative nuclear or chloroplast gene regions to differentiate between the amaranths at interspecies level.

In our study we attempted to improve the species tree observed with the barcoding analysis by conducting a whole chloroplast phylogeny. The results revealed that the inclusion of more informative sites (and a diverse array of species) leads to much greater clade resolution and support in the genus representation. Based on entire plastomes, *A. dubius*/*A. hypochondriacus*/*A. cruentus*/*A. hybridus* (clade C) is sister to *A. caudatus*/*A. quitensis*/*A. spinosus*/*A. hybridus* (clade B); while *A. powellii*/*A. retroflexus* (clade A) forms a paraphyletic relationship to clades B and C. The aforementioned clades (A, B and C) represent members of the *Amaranthus* subgenus *Amaranthus* (Mosyakin and Robertson, 1996; Wassom and Tranel, 2005). Clade B represents the grain amaranth species *A. caudatus*, together with its putative progenitors *A. quitensis* and *A. hybridus* (representing accessions collected from South America), while the remaining two grain *Amaranthus* species (*A. cruentus* and *A. hypochondriacus*) and a putative progenitor (*A. hybridus*) are present in clade C (representing species from Central America). The split between South American and Central American species was highly supported in the phylogeny, indicating that geographical separation probably led to allopatric speciation. Grain amaranths were mostly produced through human intervention, deliberate domestication and occasional accidental wild interspecies crosses (Stetter and Schmid 2017), which is evident in the low levels of genetic diversity seen between *A. hypochondriacus*, *A. cruentus*, *A. caudatus* and *A. hybridus*. The *A. powellii* and *A. retroflexus* sister lineage (clade A) was also observed when analysing genetic diversity with nuclear SSR markers (Mallory 2008). Since clade A (*A. powellii* accessions) is separated from clade B/C (*A. hypochondriacus* accessions) with strong statistical support, the previously reported hypothesis that *A. powellii* is a progenitor to *A. hypochondriacus* becomes highly unlikely (Sauer 1967). An exception is the *A. powellii* accession from Mexico (GRIN28), however it groups with *A. hypochondriacus* GRIN26 from Mexico in clade C6 and other *A. hypochondriacus* accessions in clade C, and thus GRIN28 may be mis-identified.

The groupings observed for clade B and C closely mirror results presented in a recent study utilizing whole genome SNP marker data for different amaranth species (Stetter and Schmid 2017). The combination of *A. caudatus*, *A. cruentus*, *A. hypochondriacus*, *A. hybridus* and *A. quitensis* are often referred to as the Hybridus complex and are notoriously difficult to classify (Stetter and Schmid 2017). Clade C and clade B5 of our chloroplast phylogeny represents the Hybridus complex, since it contains accessions of all five of these species (Figure 4), and we ascribe the distinction between the two clades to geographical separation during the ongoing domestication process. In this vein, the chloroplast phylogeny supported the hypothesis that *A. hybridus* and *A. quitensis* contributed to the incomplete domestication of *A. caudatus* in a particular geographic region (Stetter *et al.*, 2017), since all three species are grouped together in clade B5, including closely related *A. hybridus* and *A. quitensis* accessions from Brazil (Figure 4). The *A. hybridus* accession GRIN19 from a different geographic region (Mexico in Central America) grouped in a different clade C8. Stetter *et al.*, (2017) also highlighted the role of geographic separation, since they observed a population of closely related *A. hybridus* and *A. quitensis* accessions (Peruvian amaranth) that was distinct from a population of *A. hybridus* and *A. quitensis* in Ecuador.

The grouping of accessions into sub-clades B3 and B1+B2 are indicative of a deeper level of identification that would not be possible based solely on morphological analysis. In this study, GRIN11 and GRIN13 (sub-clade B3) were originally identified as *A. dubius* accessions but did not cluster together with the other *A. dubius*

accessions (GRIN12 and GRIN14, sub-clade C5). Being a known allotetraploid, it has been suggested that *A. dubius* could be a hybrid to which *A. spinosus* had contributed one chromosome set and the other parent could be either *A. quitensis* or *A. hybridus* (Sauer 1967). This hypothesis was supported by the chloroplast tree, as GRIN11 and GRIN13 form a confident sister lineage to *A. spinosus*. Further studies are, however, needed to confidently confirm the identification of GRIN12 and GRIN14 as *A. dubius* (due to their placement in sub-clade C5 instead of sub-clade B3) by determining their genome size. *Amaranthus dubius* is the only allotetraploid amaranth species (Stetter and Schmid, 2017) investigated during this study; therefore, similar genome sizes of GRIN12 and GRIN14 would be indicative of these accessions being conspecific with *A. dubius*. The inclusion of two *A. hypochondriacus* accessions within sub-clade B1+B2 together with *A. spinosus* has been observed before during chloroplast and nuclear gene phylogenies and it is suspected that these may be spineless versions of *A. spinosus* rather than true *A. hypochondriacus* accessions (Waselkov 2013).

Within clade C of the whole chloroplast phylogeny, sub-clades C6, C7 and C8 formed poorly resolved clades. In these clades the species boundaries between the *A. hybridus*, *A. hypochondriacus* and *A. cruentus* accessions could not be determined with confidence. The low level of genetic diversity with sub-clades C6, C7 and C8 could indicate very recent hybridization or domestication events, possibly due to self-hybridization of *A. hybridus* as previously suggested (Sauer 1967). The low genetic diversity could also be attributed to low temporal resolution provided by chloroplast data. To further validate the genetic diversity, additional nuclear markers should be included.

Accession residing in clade D formed a distant lineage to clades A, B, and C. Within this clade, *A. tricolor* and *A. graecizans* (leafy amaranths) were monophyletic, while *A. blitum* and *A. viridis* formed close sister groups. These species are thought to be introductions of Asian and European origin (classified within *Amaranthus* subgenus *Albersia*), which would support their robust separation from the Central and South American grain/weedy amaranth types (Mosyakin and Robertson 2003). The high resolution of terminal nodes in leafy and weedy amaranths (clades A and D) indicate older, more stably integrated diversification events.

Conclusion

This study is the first report of a complete chloroplast sequence of the leafy vegetable, *A. tricolor*. Phylogenetic trees based on additional chloroplast assemblies of a diverse range of amaranth accessions confirmed a moderately resolved phylogeny for the grain amaranths and a highly resolved phylogeny for most of the weedy and leafy amaranths. The two main weedy amaranths (*A. retroflexus* and *A. powellii*) resided within the same clade based on three different phylogenetic analyses (original barcoding, new barcoding and whole chloroplast phylogeny). The *A. retroflexus* and *A. powellii* species formed a sister lineage to the commonly known Hybridus complex consisting mainly of *A. hypochondriacus, A. cruentus, A. caudatus, A. hybridus, A. quitensis* and *A. dubius*. Within the Hybridus complex, separation of species groups was still not adequate to make confident classifications. Broadly, *A. caudatus* consistently grouped with *A. quitensis* and selected *A. hybridus* accessions, while *A. cruentus* and *A. hypochondriacus* grouped with alternative *A. hybridus* accessions. This

is in line with the conclusion of Stetter *et al.*, (2017) that *A. caudatus* accessions are the result of partial domestication from *A. hybridus* including gene flow from *A. quitensis*. *A. hypochondriacus* and *A. cruentus* could be domesticated versions of different geographical isolates of *A. hybridus*, as previously reported (Stetter and Schmid 2017). The grouping of all the aforementioned species (*A. caudatus*, *A. cruentus*, *A. hypochondriacus*, *A. hybridus*, *A. quitensis*, *A. powellii* and *A. retroflexus*) conform to the initially described *Amaranthus* subgenus *Amaranthus* (Mosyakin and Robertson, 1996). The leafy amaranths seem to be stable in their genetic content, by revealing the same robust topology between the barcoding as well as the whole chloroplast analysis. Whole chloroplast sequence analysis also facilitated the identification of unknown *Amaranthus* accessions in the South African genebank (Figure 5, Table 1).

Further investigations of the whole chloroplast sequence to identify additional gene barcodes for in-depth phylogenies, however, proved unsuccessful. None of the potential "new" barcoding regions possessed adequate polymorphic content to discriminate between accessions to the level of whole chloroplast analysis. The leafy amaranth accessions (*A. tricolor*, *A. viridis*, *A. blitum* and *A. graecizans*) conform to the initially described *Amaranthus* subgenus *Albersia* (Mosyakin and Robertson, 1996), and was separated from the subgenus *Amaranthus* with high support in the current phylogeny (BS=100%).

The results of this study indicate that plastomes contain the discriminatory power to separate *Amaranthus* accessions into different species groups with a high level of confidence. However, a number of nodes in the phylogenetic trees obtained in this study suffered from low statistical support. A future step to obtain a species tree with stronger resolution (especially within the Hybridus complex) would be to incorporate nuclear (sequence and marker) data and whole mitochondrial sequence phylogeny. Nuclear data would be particularly useful to investigate incongruent species placements due to their divergent histories, especially since most of the species were collected from geographical areas with no prior known selection pressures. Previous studies have shown that the development of nuclear SNP markers is extremely useful in constructing a highly informative phylogeny of the *Amaranthus* genus, both to group the species within their respective sub-genera and to identify highly differentiated groups within each sub-genus (Stetter and Schmid, 2017). Whereas the chloroplast is mostly maternally inherited, the nuclear genome can shed light in historical recombination and hybridization events (Nikiforova *et al.* 2013). Since evidence of hybridizations and ongoing gene flow between amaranth species exist, care should be taken when investigating nuclear markers. Polyploid species (such as *A. dubius*) may be characterized by a high level of heterozygosity, and the presence of multiallelic SNPs can influence phylogenetic interpretations.

The plastome is extremely useful in phylogenetic analysis due to its relatively small size, the conserved gene order and content across different plant families, high copy number in plant cells, the absence of recombination and mostly uniparental inheritance (Davis *et al.* 2014; Hollingsworth *et al.* 2016). In addition, the intergenic regions of the chloroplast usually have higher mutation rates than the genic regions, allowing more informative phylogenies over a longer time scale (Nock *et al.*, 2011). Furthermore, the chloroplast sequences could be used to investigate unique SNP markers for each species group. Instead of a genic or intergenic barcode, a SNP barcode could be developed. By sequencing and assembling whole chloroplasts of representative species from a genus, it would be feasible to identify species-specific SNPs. As illustrated by Chaney *et al.* (2016)

and Lightfoot *et al.* (2017), long read next generation sequencing technologies such as PacBio or Oxford Nanopore provide high quality sequences in highly repetitive regions such as the chloroplast IRa and IRb regions. Consequently, phylogenomics has potential to resolve phylogenetically difficult plant families.

Future studies of the *Amaranthus* genus should also focus on the addition of more species within the Hybridus complex. These additional species should represent their native and introduced geographical ranges to increase species sampling and genetic variation associated by geographical separation. Care should also be taken to ensure that one species is not overrepresented, to fully understand the underlying genetic diversity and the complex relatedness of different species groups. In this manner, a more complete picture will be obtained of grain amaranth domestication and the role of their weedy ancestors.

This study highlights the great potential of next generation sequencing for the study of plant species evolution. The identification and classification of *Amaranthus* accessions in this study will be an important tool to provide resources in terms of positively identified breeding lines for investigating nutritional, biochemical, biotic and abiotic resistance and medicinal traits naturally found in the *Amaranthus* genus.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Achigan-Dako EG, Sogbohossou OE, 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

Akond M, Islam S, Wang X (2013) Genotypic variation in biomass traits and cell wall components among 35 diverse accessions of Amaranthaceae family. Journal of Applied Phytotechnology in Environmental Sanitation 2:37-45

Alamgir M, Kibria M, Islam M (2011) Effects of farm yard manure on cadmium and lead accumulation in Amaranth (*Amaranthus oleracea* L.). Journal of Soil Science and Environmental Management 2:237-240

Alemayehu RF, Bendevis MA, Jacobsen SE (2015) The potential for utilizing the seed crop amaranth (*Amaranthus* spp.) in East Africa as an alternative crop to support food security and climate change mitigation. Journal of Agronomy and Crop Science 201:321-329

Barrett CF, Davis JI, Leebens‐Mack J, Conran JG, Stevenson DW (2013) Plastid genomes and deep relationships among the commelinid monocot angiosperms. Cladistics 29:65-87

Bell KL, de Vere N, Keller A, Richardson RT, Gous A, Burgess KS, Brosi BJ (2016) Pollen DNA barcoding: current applications and future prospects 1. Genome 59:629-640

Bezeng B *et al.* (2017) Ten years of barcoding at the African Centre for DNA barcoding. Genome:1-10

Braukmann TW, Kuzmina ML, Sills J, Zakharov EV, Hebert PD (2017) Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. PLoS One 12:e0169515

Brenner D (1990) The grain amaranth gene pools.

Burgess KS et al. (2011) Discriminating plant species in a local temperate flora using the *rbcL*+ *mat*K DNA barcode. Methods in Ecology and Evolution 2:333-340

Chan K, Sun M (1997) Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of *Amaranthus*. Theoretical and Applied Genetics 95:865-873

Chaney L, Mangelson R, Ramaraj T, Jellen EN, Maughan PJ (2016) The complete chloroplast genome sequences for four *Amaranthus* species (Amaranthaceae). Applications in Plant Sciences 4:1600063

Chung H-J et al. (2006) The complete chloroplast genome sequences of *Solanum tuberosum* and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. Plant Cell Reports 25:1369-1379

Costea M, Brenner DM, Tardif FJ, Tan YF, Sun M (2006) Delimitation of *Amaranthus cruentus* L. and *Amaranthus caudatus* L. using micromorphology and AFLP analysis: an application in germplasm identification. Genetic Resources in Crop Evolution 53:1625-1633

Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbc*L, *atp*B, and *mat*K DNA sequences. American Journal of Botany 89:132-144

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9:772-772

Das S (2011) Systematics and taxonomic delimitation of vegetable, grain and weed amaranths: a morphological and biochemical approach. Genetic Resources in Crop Evolution 59:289-303

Davis CC, Xi Z, Mathews S (2014) Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. BMC Biology 12:11-15

Dong W, Xu C, Cheng T, Lin K, Zhou S (2013) Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. Genome Biology and Evolution 5:989-997

Dong W *et al.* (2015) *ycf*1, the most promising plastid DNA barcode of land plants. Scientific Reports 5:8348

Ebert AW (2014) Potential of underutilized traditional vegetables and legume crops to contribute to food and nutritional security, income and more sustainable production systems. Sustainability 6:319-335

Gerrano AS, van Rensburg WSJ, Adebola PO (2015) Genetic diversity of *Amaranthus* species in South Africa. South African Journal of Plant and Soil 32:39-46

Gudu S, Gupta V (1988) Male-sterility in the grain amaranth (*Amaranthus hypochondriacus* ex-Nepal) variety Jumla. Euphytica 37:23-26

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696-704

Hackett SJ *et al.* (2008) A phylogenomic study of birds reveals their evolutionary history. Science 320:1763-1768

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS One 6:e19254

Hollingsworth PM, Li D-Z, van der Bank M, Twyford AD (2016) Telling plant species apart with DNA: from barcodes to genomes. Philosophical Transactions of the Royal Society B 371:20150338

Huang H, Shi C, Liu Y, Mao S-Y, Gao L-Z (2014) Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evolutionary Biology 14:151-168

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755

Kim JS, Kim JH (2013) Comparative genome analysis and phylogenetic relationship of order Liliales insight from the complete plastid genome sequences of two Lilies (*Lilium longiflorum* and *Alstroemeria aurea*). PLoS One 8:e68180

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbc*L gene complements the non-coding *trn*H-*psb*A spacer region. PLoS One 2:e508

Kuzoff RK, Gasser CS (2000) Recent progress in reconstructing angiosperm phylogeny. Trends in Plant Science 5:330-336

Lahaye R *et al.* (2008) DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences 105:2923-2928

Liu Y *et al.* (2013) Complete chloroplast genome sequences of Mongolia medicine *Artemisia frigida* and phylogenetic relationships with other plants. PLoS One 8:e57533

Ma P-F, Zhang Y-X, Zeng C-X, Guo Z-H, Li D-Z (2014) Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). Systematic Biology 63:933-950

Mallory MA, Hall RV, McNabb AR, Pratt DB, Jellen EN, Maughan PJ (2008) Development and characterization of microsatellite markers for the grain amaranths. Crop Science 48:1098-1106

Mandal N, Das P (2002) Intra-and interspecific genetic diversity in grain *Amaranthus* using random amplified polymorphic DNA markers. Plant Tissue Culture 12:49-56

Maughan PJ, Yourstone SM, Jellen EN, Udall JA (2009) SNP Discovery via genomic reduction, barcoding, and 454-pyrosequencing in Amaranth. The Plant Genome Journal 2:260-270

Mlakar SG, Turinek M, Jakop M, Bavec M, Bavec F (2010) Grain amaranth as an alternative and perspective crop in temperate climate. Journal of Geography 5:135-145

Mnkeni A, Masika P, Maphaha M (2007) Nutritional quality of vegetable and seed from different accessions of *Amaranthus i*n South Africa. Water SA 33:377-380

Nikiforova SV, Cavalieri D, Velasco R, Goremykin V (2013) Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. Molecular Biology and Evolution 30:1751-1760

Nock CJ, Waters DL, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal 9:328-333

Panero JL, Funk V (2008) The value of sampling anomalous taxa in phylogenetic studies: major clades of the Asteraceae revealed. Molecular Phylogenetics and Evolution 47:757-782

Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology 7:84

Patil SM, Rane NR, Adsul AA, Gholave AR, Yadav SR, Jadhav JP, Govindwar SP (2016) Study of molecular genetic diversity and evolutionary history of medicinally important endangered genus *Chlorophytum* using DNA barcodes. Biochemical Systematics and Ecology 65:245-252

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465:1033-1038

Purty R, Chatterjee S (2016) DNA Barcoding: An effective technique in molecular taxonomy. Austin Journal of Biotechnology and Bioengineering 3:1059

Raju M, Varakumar S, Lakshminarayana R, Krishnakantha T, Baskaran V (2007) Carotenoid composition and vitamin A activity of medicinally important green leafy vegetables. Food Chemistry 101:1598-1605

Rastogi A, Shukla S (2013) Amaranth: a new millennium crop of nutraceutical values. Critical Reviews in Food Science and Nutrition 53:109-125

Sangeetha RK, Baskaran V (2010) Carotenoid composition and retinol equivalent in plants of nutritional and medicinal importance: Efficacy of β-carotene from *Chenopodium album* in retinol-deficient rats. Food Chemistry 119:1584-1590

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Research 6:283-290

Sauer JD (1967) The grain amaranths and their relatives: a revised taxonomic and geographic survey. Annals of the Missouri Botanical Gardens 54:103-137

Schmitz-Linneweber C, Maier RM, Alcaraz J-P, Cottet A, Herrmann RG, Mache R (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Molecular Biology 45:307-315

Shaw J, Shafer HL, Leonard OR, Kovach MJ, Schorr M, Morris AB (2014) Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. American Journal of Botany 101:1987-2004

Shinozaki K *et al.* (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO Journal 5:2043-2049

Sievers F *et al.* (2011) Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. Molecular Systematic Biology 7:539-545

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402-404

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 22:2688-2690

Stetter MG, Müller T, Schmid KJ (2016) Genomic and phenotypic evidence for an incomplete domestication of South American grain amaranth (*Amaranthus caudatus*). Molecular Ecology 26:871-886

Stetter MG, Schmid KJ (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

Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of Botany 99:349-364

Stull GW *et al.* (2013) A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. Applied Plant Sciences 1:1200497

Sugiura M (1992) The chloroplast genome. In: Schilperoort RA, Dure L (eds) 10 Years Plant Molecular Biology. Springer Netherlands, Dordrecht, pp 149-168.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28:2731-2739

Timofte I, Timofte N, Brega V (2009) Development of bioenergy in Moldova. Problemele Energericii Regionale 2:1-12

Van Rensburg WJ *et al.* (2007) African leafy vegetables in South Africa. Water SA 33:317-326

Venskutonis PR, 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

Waselkov K (2013) Population genetics and phylogenetic context of weed evolution in the genus *Amaranthus*: Amaranthaceae. Washington University, St Louis, USA

Williams AV, Miller JT, Small I, Nevill PG, Boykin LM (2016) Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in Acacia. Molecular Phylogenetics and Evolution 96:1-8

Xu F, Sun M (2001) Comparative analysis of phylogenetic relationships of grain amaranths and their wild relatives (*Amaranthus*; Amaranthaceae) using internal transcribed spacer, amplified fragment length polymorphism, and double-primer fluorescent intersimple sequence repeat markers. Molecular Phylogenetics and Evolution 21:372-387

Zhang T, Fang Y, Wang X, Deng X, Zhang X, Hu S, Yu J (2012) The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. PLoS One 7:e30531

Zhang YJ, Ma PF, Li DZ (2011) High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLoS One 6:e20596

Supplementary Data

Supplementary Table S1: Chloroplast gap-closing primer statistics

Primer	Sequence	Annealing Temp (°C)	Product Size (bp)
CSS1 F	AAAACCAACCAGATCCGATG	60	688
CSS1 R	GTCCCGAAGAGAGGGAAGAG		
CSS2 F	TCAGGATCAGTCGTGGTCTT	60	827
CSS2 R	GGTGCTCAACCTACAGAGAC		
CSS3 F	GCTGATTCTAGCCTGCTCAT	60	1600
CSS3 R	CTAGTGGTTCAGGACATCTC		
CSS4 F	AGGGCAAGTGTTCGGATCTA	60	702
CSS4 R	TGAAGCGCATAATTGGTTGA		
CSS5 F	GGAACAGCAAAAGGAAATGAA	59	700
CSS5 R	ACTCGGCCATCTCTCCTACA		
CSS6 F	AGCTAAGCGGGCTCACATAA	60	705
CSS6 R	TCTTTCATTTCGGTCGGAAC		
CSS7 F	TTTCTCACTTTCACCTCGATTTT	59	717
CSS7 R	TAATGAATCGGAGCACATGG		
CSS8 F	TTTTTCGTTGGCGATATCCT	59	706
CSS8 R	TTCATAAGAAATCGCGTGGT		
CSS9 F	ACCCGCAATGAATAGGGAAG	59	675
CSS9 R	GCCTCAACAATTGGGTTCAT		
CSS10 F	GATTTTCAATGCGCGATTTC	59	683
CSS10 R	CGTTCGGCTCGATCTATTTC		
CSS11 F	TCCGAAACAAATACGATTACCC	60	676
CSS11 R	ACTAAGGGGGTTGGTGGAAA		
CSS12 F	GACCACGATTGGGAATTGTT	59	705
CSS12 R	ACATCCCGGATCTGATGAAA		
CSS13 F	CCTCATACGGCTCCTCAAGA	60	736
CSS13 R	CGGGTTGAAACGAATTATGG		
CSS14 F	CCGCACTTTTTGGGGTACTA	59	721
CSS14 R	TGGCTAATCGAACAAGGACA		
CSS15 F	CCACTCAGCACGTTGGAATA	60	1100
CSS15 R	CGGGTTTGAAAAATCCCTTC		
CSS16 F	TCGGCCCTTTATTCGAGTA	54	1355
CSS16 R	GCCGTGGGAATCAAAAAA		
CSS17 F	GCGTGTAAACGAGGTGCTCT	60	689
CSS17 R	AAAAGACCATGCCGCTACTG		
CSS18 F	GAGGAGGAAGCTTATAAAGA	53	1298
CSS18 R	GGCACAAAAATAGAGGGA		
IRa/SSC F	CTTCTCCTTCCCTTTTTTC	52	1329
IRa/SSC R	GATGCGACCTTTTCTTAC		
SSC/IRb F	GGCTTCCTTTGTTCGGTT	55	1269
SSC/IRb R	CCTTCTCCTTCCCTTTTTT		
IRb/LSC F	GGACAAGTGGGAAATGTTGG	60	707
IRb/LSC R	GCAATGCCGTTTTCTTGTTT		
LSC/IRa	GCATAACACGGAACAAAG	54	1027
LSC/IRa	AGAAAAAACCCCACAACC		

Supplementary Table S2: Whole chloroplast mapping statistics

Accession	Number of mapped reads	Average coverage (x)	Gapped region (%)
SAG1	247,686	40	0.7
SAG3	814,670	33	1.3
SAG4	624,540	46	1.1
SAG7	7,475,938	64	1.1
SAG9	776,094	51	0.8
SAG10	204,046	20	1.3
SAG11	184,422	18	1.6
SAG12	457,482	26	0.7
SAG14	360,530	59	1.0
SAG17	1,097,686	54	3.4
SAG30	8,000,000	54	1.2
SAG34	264,576	18	2.1
SAG36	217,382	22	1.3
GRIN1	395,184	72	6.4
GRIN2	146,808	26	1.0
GRIN3	530,138	76	0.6
GRIN4	180,532	26	19.4
GRIN5	217,480	41	16.1
GRIN6	133,530	112	11.2
GRIN7	206,552	30	21.0
GRIN8	103,753	79	6.9
GRIN9	346,464	39	1.0
GRIN10	276,816	20	1.5
GRIN11	284,138	62	5.1
GRIN12	404,558	43	6.7
GRIN13	376,862	42	2.9
GRIN14	10,803,196	1,203	0.2
GRIN15	143,646	30	0.8
GRIN16	243,438	24	2.0
GRIN17	242,294	38	2.8
GRIN18	382,952	87	1.9
GRIN19	414,372	113	1.0
GRIN20	553,770	55	5.8
GRIN21	166,668	43	0.8
GRIN22	328,768	58	3.6
GRIN23	243,712	32	1.3
GRIN24	572,312	60	5.5
GRIN25	477,164	33	4.8
GRIN26	244,434	24	1.5
GRIN27	3,914,990	332	0.8
GRIN28	439,398	93	2.1
GRIN29	392,140	30	3.5
GRIN30	243,920	51	0.8
GRIN31	266,330	50	0.9
GRIN32	507,266	28	2.6
GRIN33	392,304	67	5.8
GRIN34	338,822	45	0.7
GRIN35	275,688	52	2.4
GRIN36	549,562	64	1.5
GRIN37	359,518	30	5.7
GRIN38	304,238	19	12.5
GRIN39	1,845,970	143	0.6
GRIN40	466,072	31	5.3
GRIN41	335,342	42	1.6
GRIN42	211,512	31	1.0
GRIN43	612,516	39	5.3
GRIN44	465,612	67	2.5
GRIN45	197,362	14	6.2

Supplementary Table S3: Whole angiosperm chloroplast sequences obtained from Genbank (NCBI)

Angiosperm species	Common name	NCBI Accession Nr	Size (bp)	GC content (%)	Family
Arabidopsis thaliana	Thale cress	gi|5881673|	154,478	36.3	Brassicaceae
Amaranthus tricolor	Leafy amaranth	PRJNA318736	150,027	36.6	Amaranthaceae
Beta vulgaris	Sugar Beet	gi|148607972|	149,696	37.0	Amaranthaceae
Brachypodium distachyon	Purple False Brome	gi|194033128|	135,199	38.6	Poaceae
Brassica rapa	Turnip	gi|323481965|	153,482	36.4	Brassicaceae
Carica papaya	Pawpaw	gi|166344111|	160,100	36.9	Caricaceae
Citrus sinensis	Orange	gi|113952601|	160,129	38.5	Rutaceae
Cucumis melo	Melon	gi|346578170|	156,047	36.9	Cucurbitaceae
Cucumis sativus	Cucumber	gi|67511377|	155,293	37.1	Cucurbitaceae
Eucalyptus grandis	Rose gum	gi|309322431|	160,137	37.9	Myrtaceae
Fragaria vesca	Strawberry	gi|325126844|	155,691	37.2	Rosaceae
Gossypium hirsutum	Cotton	gi|329317136|	160,256	37.2	Malvaceae
Manihot esculenta	Cassava	gi|169794052|	161,453	35.9	Euphorbiaceae
Nicotiana undulata	Tobacco	gi|347453879|	155,863	37.9	Solanaceae
Olea europaea	Olive	gi|334084552|	155,657	37.8	Oleaceae
Oryza sativa	Rice	gi|109156581|	134,496	39.0	Poaceae
Populus trichocarpa	Poplar	gi|133712039|	157,033	36.7	Salicaceae
Prunus persica	Peach	gi|313183801|	157,790	36.8	Rosaceae
Ricinus communis	Caster oil	gi|372450118|	163,161	35.7	Euphorbiaceae
Solanum lycopersicum	Tomato	gi|113531108|	155,461	37.9	Solanaceae
Solanum tuberosum	Potato	gi|82754608|	155,312	37.9	Solanaceae
Sorghum bicolor	Sorghum	gi|118614470|	140,754	38.5	Poaceae
Spinacia oleraceae	Spinach	gi|7636084|	150,725	36.8	Amaranthaceae
Vitis vinifera	Grape Vine	gi|91983971|	160,928	37.4	Vitaceae
Zea mays	Maize	gi|11994090|	140,384	38.5	Poaceae

Gymnosperm species	Common name	NCBI Accession Nr	Size (bp)	GC content (%)	Family
Selaginella moellendorffii	Lycophyte	gi|296399203|	143,775	51.0	Selaginellaceae

Supplementary Table S4: Genic information for partitioned phylogenomic analysis

Chloroplast Region	Strand	Fragment size	Pi* Sites %	Model
*acc*D	+	1,658	1.9	TVM+I
*atp*A	-	1,523	1.4	TPM1uf+G
*atp*B	-	1,492	0.5	GTR
*atp*E	-	405	1.5	TPM2uf
*atp*F	-	160	0.0	F81
*atp*I	-	754	0.9	TVM
*ccs*A	+	973	3.0	TPM1uf+G
*cem*A	+	701	1.1	TrN+I
*clp*P_Exon1	-	227	0.9	TPM1uf
*clp*P_Exon2	-	292	1.7	TPM1uf
*inf*A	-	232	0.9	TPM3uf
*mat*K	-	1,515	3.6	TVM
*ndh*A_Exon1	-	539	2.0	TIM1
*ndh*A_Exon2	-	554	0.9	TPM1uf
*ndh*D	-	1,502	2.3	TVM+G
*ndh*E	-	309	1.0	TPM2uf
*ndh*F	-	3,452	3.3	TVM+I+G
*ndh*G	-	533	1.9	TPM2uf
*ndh*H	-	1,181	0.8	TVM+I+G
*ndh*I	-	512	1.2	TPM1uf
*ndh*J	-	476	1.1	TIM3
*ndh*K	-	862	2.2	TVM+G
*pet*A	+	961	0.9	TVM+I
*pet*B	+	648	1.1	TPM1uf+G
*pet*D	+	521	1.0	TPM1uf
*pet*G	+	113	2.7	TrN
*psa*A	-	2,251	0.6	TVM+G
*psa*B	-	2,202	0.8	TVM+I
*psa*C	-	243	1.6	TrN
*psb*A	-	1,061	1.2	TIM1+I
*psb*C	+	1,456	0.3	HKY
*psb*D	+	1,061	0.8	TVM
*psb*H	+	239	2.1	TPM2uf+I
*psb*I	+	163	4.3	TPM1uf+G
*psb*K	+	184	1.6	TPM1uf
*psb*T	+	107	1.9	TPM1uf+I
*psi_psb*T	+	1,526	1.2	GTR+G
*rbc*L	+	1,426	0.7	TPM1uf+G
*rpl*14	-	365	0.5	TIM2
*rpl*16	-	399	2.3	TVM+I
*rpl*22	-	591	4.6	TVM+G
*rpl*32	+	177	6.2	F81+G
*rpl*36	-	111	1.8	TIM1
*rpo*A	-	1,012	2.4	TPM1uf
*rpo*B	-	3,214	1.4	GTR+I
*rpo*C1_Exon1	-	1,610	2.0	TVM+G
*rpo*C1_Exon2	-	447	0.7	TPM1uf
*rpo*C2_Exon1	-	2,439	1.7	TVM+G
*rpo*C2_Exon2	-	1,645	1.0	TIM1+I
*rps*11	-	457	1.3	TPM1uf
*rps*14	-	303	1.3	TPM3uf
*rps*15	-	271	2.2	HKY+I
*rps*16	-	228	0.9	TVM
*rps*18	+	411	4.9	TrN
*rps*19	-	276	1.1	TPM1uf
*rps*2	-	710	1.0	TPM1uf
*rps*3	-	655	2.3	TPM2uf+I
*rps*8	-	402	1.5	TPM1uf+G
*ycf*1	+	1,406	0.6	TPM1uf
*ycf*2_Exon1	+	2,077	0.2	TPM1uf
*ycf*2_Exon2	+	4,635	0.6	TPM1uf+I
*ycf*3_Exon1	-	152	0.0	TrNef
*ycf*3_Exon2	-	230	0.9	HKY
*ycf*4	+	555	1.4	TIM3

*Pi: Parsimoniously informative

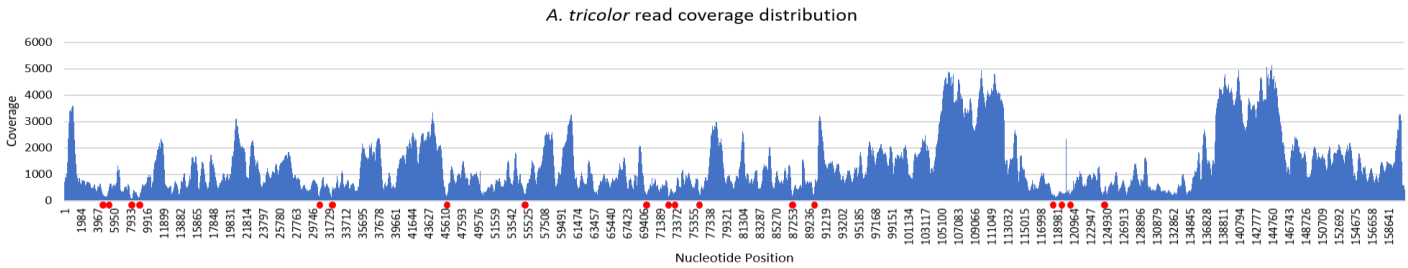

Figure S1: Mapping of raw *Amaranthus tricolor* Illumina sequencing reads to the *Beta vulgaris* chloroplast sequence resulted in a mostly uniform coverage distribution. Areas with coverage below 50x were identified (marked by red circles) and were further confirmed/resolved through primer design and Sanger sequencing

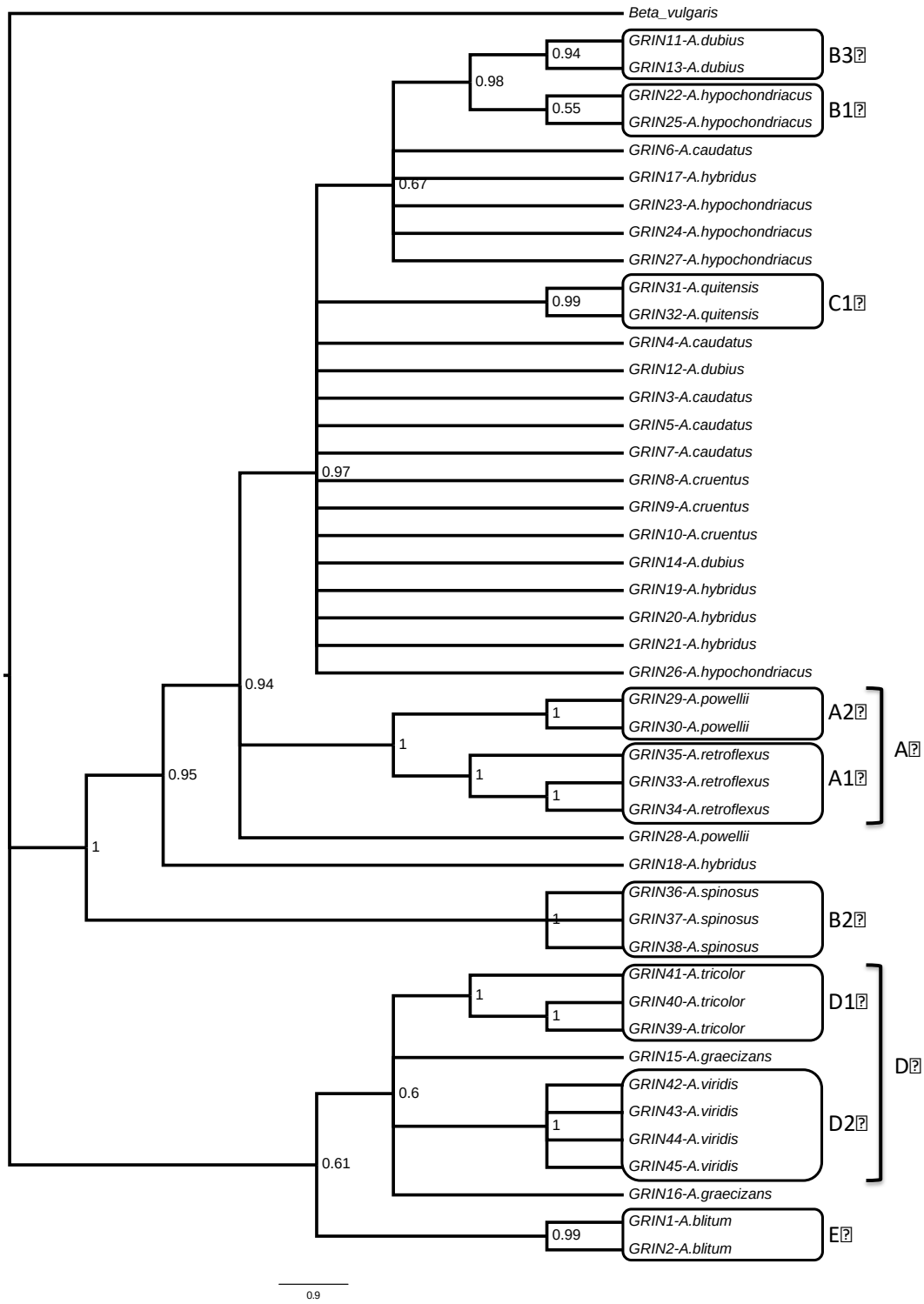

Figure S2: Amaranth phylogeny based on DNA barcoding of chloroplast *mat*K, *rbc*L and nuclear ITS gene regions using Bayesian analysis. Percentage posterior probability values are indicated on nodes. Only nine sub-clades could be identified compared to the RAxML analysis, but their probability values are higher. The remaining nodes have collapsed, and identifications become questionable

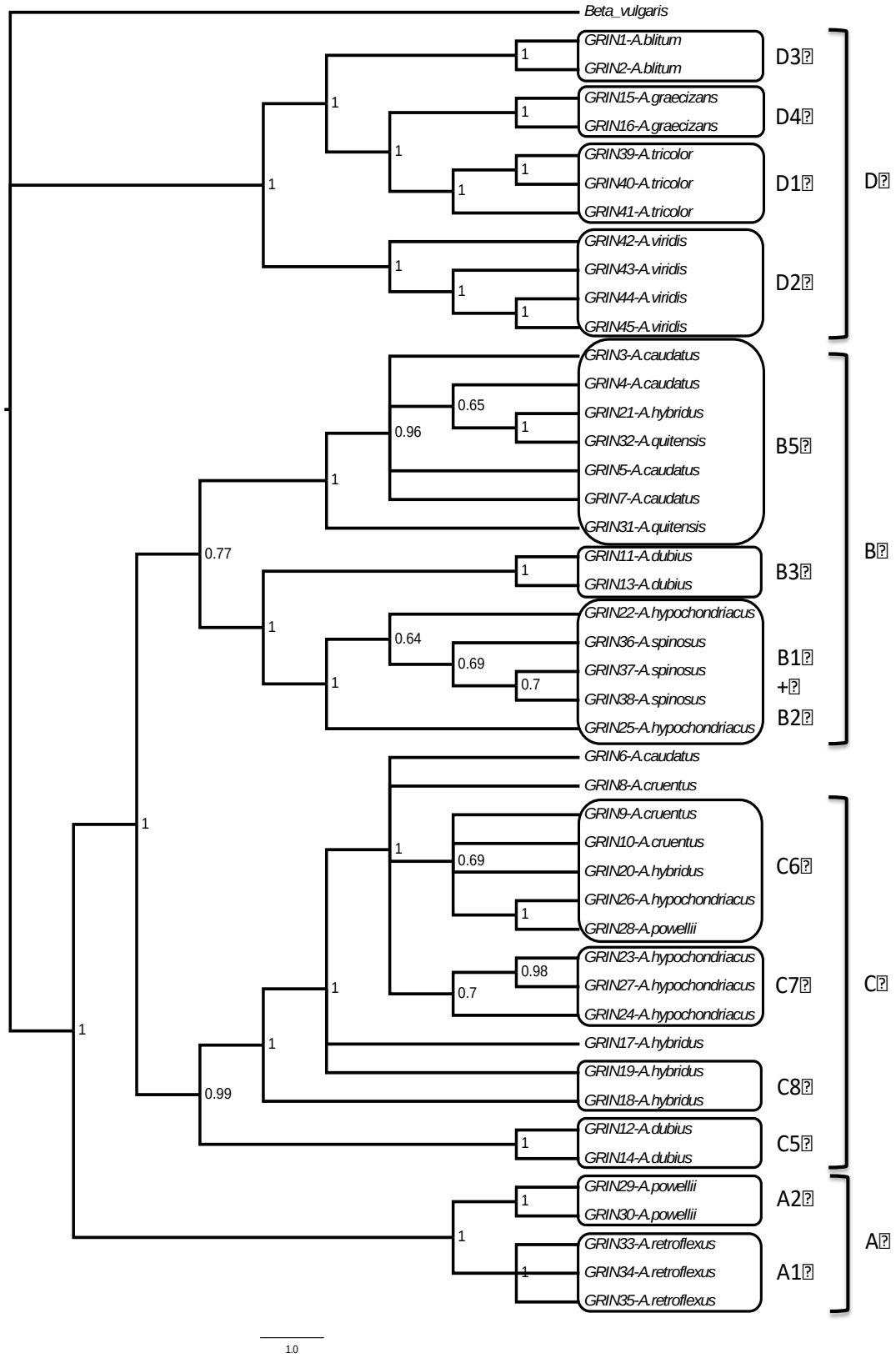

Figure S3: Amaranth phylogeny based on a Bayesian analysis of whole chloroplast genome GRIN sequences. Percentage posterior probability values are indicated on nodes. As for the RAxML analysis, four main clades could be identified: (A) weedy amaranth; (B) a combination of weedy amaranth and grain amaranth; (C) a combination of weedy amaranth, grain amaranth and leafy amaranth and (D) leafy amaranth.

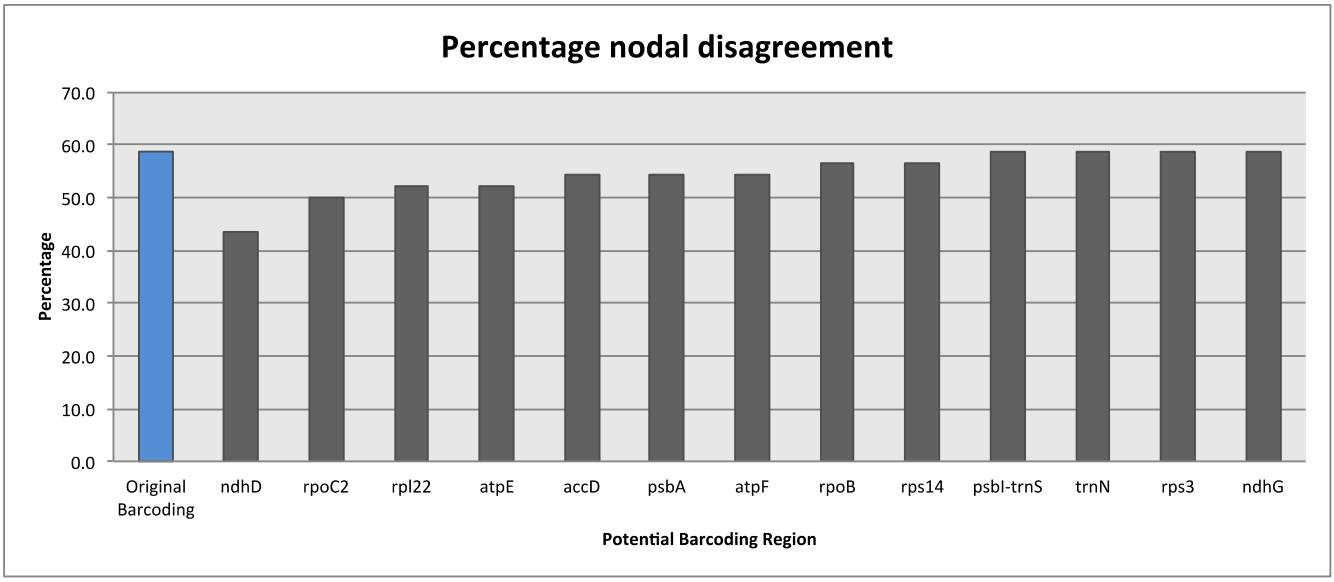

Figure S4: Bar graph indicating the percentage disagreement between a new barcoding phylogeny vs. the whole chloroplast phylogeny using the TOPD/FMTS software. The blue bar reflects the percentage disagreement between the original barcoding set (*mat*K, *rbc*L and ITS) and the whole chloroplast phylogeny. Grey bars indicate the percentage disagreement of a barcoding phylogeny with an added gene (x-axis), to the whole chloroplast phylogeny. Lower percentages indicate a larger agreement between the phylogenetic trees

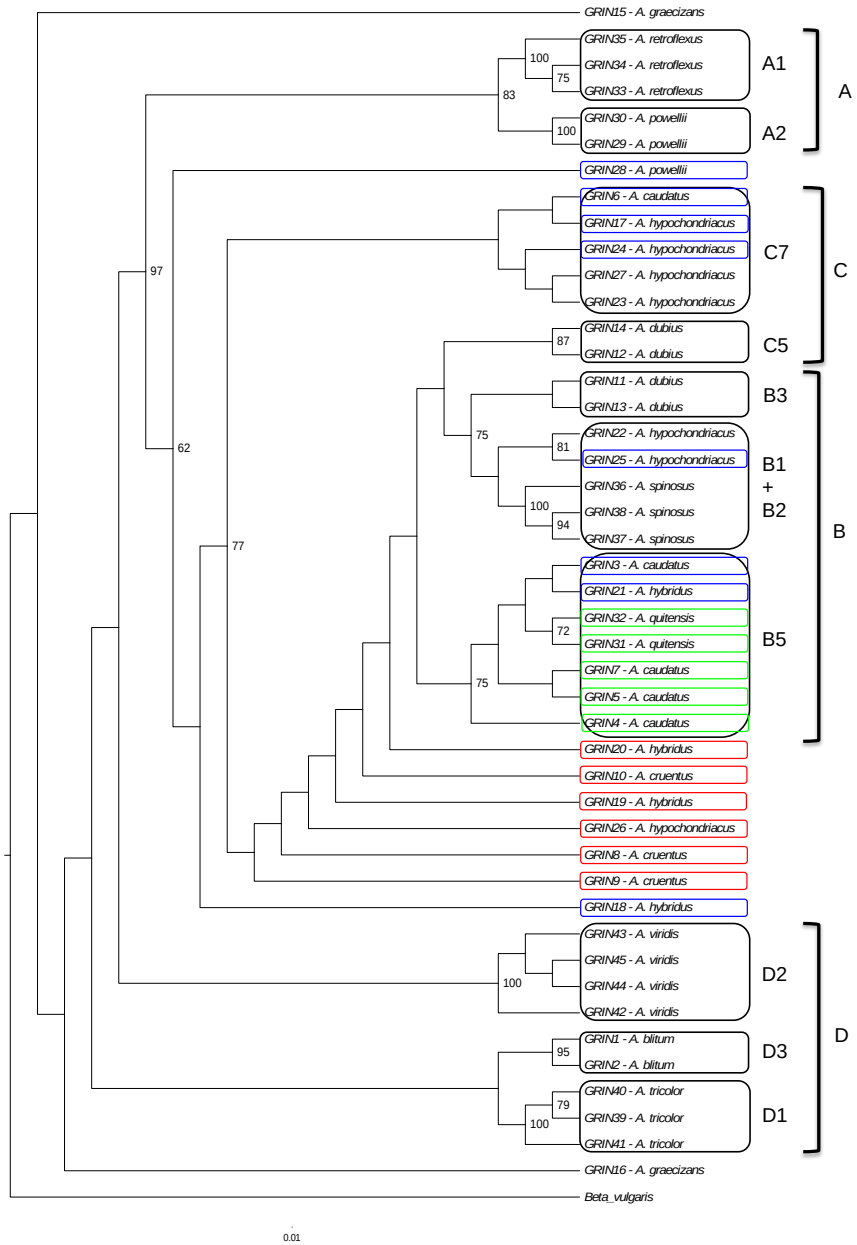

Figure S5: *Amaranthus* phylogeny based on DNA barcoding of chloroplast *mat*K, *rbc*L, *ndh*D and nuclear ITS gene regions on the entire germplasm set. The phylogeny was constructed using a partitioned maximum likelihood analysis (Model GTR+G (ITS, *mat*K, *ndh*D) and HKY (*rbc*L), 1000 bootstrap replicates), and bootstrap confidence values (> 60%) were indicated at branch nodes. The genus is divided into three main subclades, broadly representing weedy amaranth (Clade A), mix of grain and leafy amaranth (Clade B and C) and leafy amaranth (Clade D) accessions. Blue blocks represent accessions that are in accord between the new barcoding phylogeny and the whole chloroplast phylogeny. Green blocks indicate accessions that showed an improvement in their phylogenetic placement, while red blocks represent accessions to which no identity could be assigned in the new barcoding phylogeny compared to the whole chloroplast phylogeny. GRIN – Previously identified amaranth accessions (Table 5.1)